UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE FILOSOFIA E CIÊNCIAS, CAMPUS DE MARÍLIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Jorge Janaite Neto

Recuperação de Informação Textual Baseada em Cluster Conceitual

Marília – SP
2023

UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE FILOSOFIA E CIÊNCIAS, CAMPUS DE MARÍLIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Jorge Janaite Neto

Recuperação de Informação Textual Baseada em Cluster Conceitual

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências – Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, campus de Marília, como requisito parcial para obtenção do título de Doutor em Ciência da Informação

Área de concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Orientador: Prof. Dr. Edberto Ferneda

Marília – SP
2023

J33r

Janaite Neto, Jorge
Recuperação de informação textual baseada em cluster conceitual / Jorge Janaite Neto. -- Marília, 2023
128 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília
Orientador: Edberto Ferneda

1. Algoritmos de computador. 2. Recuperação da informação. 3. Indexação automática. 4. Análise por agrupamento. 5. Estruturas conceituais (Teoria da informação). I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Impacto potencial desta pesquisa

Esta tese traz uma proposta que possui o potencial de causar impactos positivos à sociedade. Conforme estabelecido pela Assembleia Geral das Nações Unidas (AGNU) no ano de 2015, os Objetivos de Desenvolvimento Sustentável (ODS) oferecem "um plano compartilhado para a paz e a prosperidade das pessoas e do planeta, agora e no futuro"[1]; este plano é composto por 17 metas que compõem um plano maior para Desenvolvimento Sustentável, chamado Agenda 2030.

O impacto potencial para a sociedade que esta tese apresenta está fortemente relacionado a três objetivos: ODS 04 (Educação de Qualidade), ODS 10 (Redução das Desigualdades) e ODS 16 (Paz, Justiça e Instituições Eficazes).

ODS 04: *Educação de Qualidade* – "Garantir uma educação de qualidade, inclusiva e equitativa, e promover oportunidades de aprendizagem ao longo da vida para todos"[2]. Recuperar informação está intimamente ligado à aprendizagem. A proposta de empregar clusters para a representação dos conceitos contidos nos documentos, além de promover uma indexação automática mais assertiva, também oferece a possibilidade de novas interfaces de busca. Isso contribui significativamente para ampliar as oportunidades de aprendizagem ao longo da vida das pessoas.

ODS-10: *Redução das Desigualdades* – "Reduzir a desigualdade de rendimentos dentro e entre os países"[3]. Existe um certo consenso de que há indícios de uma forte relação entre nível de escolaridade e renda aqui no Brasil. Este trabalho científico traz uma proposta que tem o potencial de oferecer maneiras mais intuitivas e mais assertivas para a recuperação de informação, tornando menos elitizado o acesso aos materiais, incentivando o indivíduo a explorar os textos, independentemente de quão familiarizado ele esteja com tais; com isso incentivando os estudos e trazendo como consequência, a longo prazo, uma redução desta desigualdade de rendimentos.

ODS 16: *Paz, Justiça e Instituições fortalecidas* – "Promover sociedades pacíficas e inclusivas para o desenvolvimento sustentável, proporcionar acesso à justiça para todos e construir instituições eficazes, responsáveis e inclusivas a todos os níveis"[4] . Neste aspecto, esta tese ao propor melhoria na forma de manipular os conceitos e com isso auxiliar no processo de in-

[1] "[The 17 goals are] shared blueprint for peace and prosperity for people and the planet, now and into the future". Disponível em <https://sdgs.un.org/goals>. Acesso em 01.nov.2023.

[2] "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all"

[3] "Reduce income inequality within and among countries"

[4] "Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels"

dexação automática e recuperação, favorece a criação de novas ferramentas para promover o acesso à justiça, tais como ferramentas de busca por situações semelhantes para garantir a uniformidade das decisões aplicadas.

Esta pesquisa também busca contribuir com a Ciência da Informação e com a linha de pesquisa Informação e Tecnologia, do Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Unesp de Marília, ao propor novas maneiras de representar a informação.

Potential impact of this research

This thesis presents a proposal that has the potential to have a positive impact on society. As established by the United Nations General Assembly (UNGA) in 2015, the Sustainable Development Goals (SDGs) offer "a shared blueprint for peace and prosperity for people and the planet, now and into the future"[5]; this plan is made up of 17 goals that make up a larger plan for Sustainable Development, called 2030 Agenda.

The potential impact on society that this thesis presents is strongly related to three goals: SDG 04 (Quality Education), SDG 10 (Reducing Inequalities) and SDG 16 (Peace, Justice and Effective Institutions).

SDG 04: *Quality Education* "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". Retrieving information is closely linked to learning. The proposal to use clusters to represent the concepts contained in documents not only promotes more assertive automatic indexing, but also offers the possibility of new search interfaces. This contributes significantly to expanding people's lifelong learning opportunities.

SDG-10: *Reducing Inequalities* – "Reduce income inequality within and among countries". There is a certain consensus that there is evidence of a strong relationship between educational attainment and income here in Brazil. This scientific work puts forward a proposal that has the potential to offer more intuitive and more assertive ways of retrieving information, making access to materials less elitist, encouraging individuals to explore texts, regardless of how familiar they are with them; thereby encouraging study and bringing about, in the long term, a reduction in this income inequality.

SDG 16: *Peace, Justice and Strengthened Institutions* – "Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels". In this respect, this thesis, by proposing an improvement in the way concepts are manipulated and thereby assisting in the automatic indexing and retrieval process, favors the creation of new tools to promote access to justice, such as search tools for similar situations to ensure uniformity in the decisions applied.

This research also aims to contribute to Information Science and the Information and Technology research line of the Postgraduate Program in Information Science (PPGCI) at Unesp Marília, by proposing new ways of representing information.

[5] Available at <https://sdgs.un.org/goals>. Accessed on 01.nov.2023.

Jorge Janaite Neto

Recuperação de Informação Textual Baseada em Cluster Conceitual

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências – Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, campus de Marília, como requisito parcial para obtenção do título de Doutor em Ciência da Informação

Área de concentração: Informação, Tecnologia e Conhecimento
Linha de Pesquisa: Informação e Tecnologia

Banca Examinadora

Prof. Dr. Edberto Ferneda (Orientador)
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Profa. Dra. Rachel Cristina Vesu Alves
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Prof. Dr. Cecilio Merlotti Rodas
Departamento de Ciência da Informação
Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)

Prof. Dr. Fabricio Baptista
Instituto Federal de Educação, Ciência e Tecnologia do Paraná (IFPR)
Campus Jacarezinho

Prof. Dr. Marckson Roberto Ferreira de Sousa
Departamento de Ciência da Informação
Universidade Federal da Paraíba (UFPB)

Marília, 28 de setembro de 2023.

Este trabalho é dedicado
ao meu incrível e estimado irmão
Antonio Janaite Filho

e aos meus queridos pets[6]
que ao longo destes anos foram,
literalmente, aparecendo em minha vida,
cada um acompanhado de uma história

Vocês todos são o máximo!

[6] Gatos: Sir James White, Penélope Caroline, Coquinho Cristina, Ivanov Vaquinha, Charlie Root, Leon Henrique, David Fluke (*in memoriam*), Billy (*in memoriam*), Luizão, Mike, Salém, Lucy, Dennis Yellow; Cachorros: Shoyo (*in memoriam*), Bill, Meggie, Greta, Chico, Bóris, Maverick, Ada, Ted Nelson, Layla e Max (*in memoriam*).

Agradecimentos

A elaboração de qualquer pesquisa envolve diversas vivências ao longo da jornada.

Quero agradecer, primeiramente, ao Prof. Dr. Edberto Ferneda, meu orientador, que soube com serenidade e conhecimento conduzir esse trabalho, proporcionando ótimas discussões e trazendo reflexões inusitadas a todo momento.

Agradeço à Giselli Hara que se fez presente, com dedicação e companheirismo, durante todas as etapas.

Agradeço também aos membros da banca de qualificação e defesa, pelas importantes sugestões que serviram para aprimorar este trabalho.

Meus sinceros agradecimentos a todos os envolvidos que, de alguma maneira, contribuíram com este trabalho, incluindo aqui todos aqueles que participaram desta minha jornada acadêmica desde os tempos de minha graduação até este momento.

[. . .] Espero, disse Poole a si mesmo, que essa confiança seja justificada.
Alguém disse uma vez que
qualquer tecnologia suficientemente avançada é indistinguível da magia
Encontrarei magia neste novo mundo – e serei capaz de lidar com ela?
Arthur C. Clarke (1997, p. 36)

É na prática que o homem tem de comprovar a verdade, isto é,
a realidade e o poder, a natureza interior de seu pensamento.
A disputa acerca da realidade ou não realidade do pensamento
– que é isolado da prática –
é uma questão puramente escolástica.
Karl Marx (2007, p. 533).

Resumo

A atual abundância na produção de documentos torna necessário o desenvolvimento de novos esquemas de classificação que sejam capazes de organizar o enorme volume de material produzido incessantemente. Como uma parte expressiva deste material textual é produzido e armazenado em meios digitais, isso favorece bastante o uso de sistemas de indexação automáticos.

Recuperação de informação é um processo linguístico, ao passo que a indexação automática operada por computadores é um processo estatístico, tornando necessário uma aproximação destas áreas do conhecimento. Tradicionalmente, os textos são considerados como um conjunto de palavras portadoras de uma relevância tópica proporcional à frequência de ocorrência dentro de cada documento e à frequência entre os documentos que compõe o corpus documental, sendo esta representação denominada de bag-of-words. A principal deficiência destas representações clássicas baseadas no modelo bag-of-words é o tratamento dado às palavras ambíguas: elas são descartadas ou ignoradas; isso empobrece muito a qualidade da indexação e consequentemente a qualidade da recuperação. O problema da ambiguidade terminológica é um problema linguístico: algumas palavras ortograficamente idênticas possuem significados diferentes. Se superarmos a questão terminológica e operarmos em nível conceitual, o problema da ambiguidade estaria solucionado: os conceitos são inequívocos.

O propósito desta tese é investigar e propor o uso de clustering a partir dos conceitos com o objetivo de melhorar a eficácia do processo de indexação automática e recuperação de informação, aperfeiçoando a representação dos textos que compõe o corpus documental e os representando por agrupamentos conceituais. Ao final é realizado um experimento para ilustrar a aplicação prática do algoritmo proposto bem como demonstrar os resultados promissores alcançados e lançar um base para uma futura implementação completa.

Palavras-chave: Algoritmos de computador. Recuperação da informação. Indexação automática. Análise por agrupamento. Organização da informação.

Abstract

The current abundance of document production makes it necessary to develop new classification schemes that can organize a large volume of material produced incessantly. Since a significant part of this textual material is produced and stored digitally, this greatly favors the use of automatic indexing systems.

Information retrieval is a linguistic process while automatic indexing operated by computers is a statistical process, making it necessary to bring these areas of knowledge closer together. Traditionally, texts are considered as a set of words with a topical relevance proportional to the frequency of occurrence within each document and the frequency between the documents that make up the document corpus, this representation is called bag-of-words. The main shortcoming of these classic representations based on the bag-of-words model is the treatment given to ambiguous words: they are discarded or ignored; this greatly reduces the quality of indexing and consequently the quality of retrieval. The problem of terminological ambiguity is a linguistic problem: some words that are orthographically identical have different meanings. If we overcome the terminological issue and operates at a conceptual level, the problem of ambiguity would be solved: the concepts are unambiguous.

The purpose of this dissertation is to investigate and propose the use of concept-based clustering to improve the effectiveness of the automatic indexing and information retrieval process by improving representation of the texts that make up the document corpus, representing them by conceptual groupings. At the end, an experiment is carried out to illustrate the practical application of the proposed algorithm, as well as to demonstrate the promising results achieved and lay the groundwork for a future full implementation of it.

Keywords: Computer algorithms. Information retrieval. Automatic indexing. Cluster analysis. Conceptual structures (Information theory). Information organization.

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

CLI	*Command Line Interface*
CO_2	Gás Carbônico – dióxido de carbono
CSV	*Comma separated values*. Arquivo separado por vírgulas
Docfreq	Frequência de um termo ou UT em um documento
IDF	*Inverse Document Frequency*
IPCC	Painel Intergovernamental sobre Mudanças Climáticas
ISO	*International Organization for Standardization*
IUCN	União Internacional para Conservação da Natureza
MND	Mutual Neighbor Distance
MS Excel	Software Microsoft Excel – Software planilha de cálculos
PHP	*PHP Hypertext Processing* (acrônimo recursivo)
PIN	*Perceived Information Need*
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informações
RIN	*Real Information Need*
RSLP	Removedor de Sufixos da Língua Portuguesa
SI	Sistemas de Informação
SMART	*System for the Mechanical Analysis and Retrieval of Text*
SRI	Sistema de Recuperação de Informações
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
TGT	Teoria Geral da Terminologia
UC	Unidade Conceitual
UT	Unidade Terminológica

Lista de símbolos

$\in$	Símbolo matemático de “pertence”
$\mathbb{R}$	Conjunto dos Números Reais
$\equiv$	Símbolo matemático para “equivalente”
φ	Letra grega minúscula Phi Neste trabalho, função “medida de dissimilaridade”
$=$	Símbolo matemático para igualdade
$\neq$	Símbolo matemático para não igualdade
$\approx$	Símbolo matemático para aproximadamente
α	Letra grega minúscula alfa
δ	Letra grega minúscula delta
$\sum$	Letra grega Sigma Símbolo matemático para somatória Exemplo: $\sum_{n=5}^{7} X_i$ soma os elementos $X_5 + X_6 + X_7$
π	Letra grega minúscula Pi
$\mathfrak{D}$	Letra D em escrita gótica $\mathfrak{Fraktur}$ Neste trabalho, conjunto de transações/termos que serão analisados pelo algoritmo *Apriori*; e conjunto de documentos que serão processados e agrupados pelo algoritmo *K-Means*
$\|x\|$	Operação módulo. Valor absoluto
τ	Letra grega minúscula Tau
$A \subseteq B$	Conjunto A é um subconjunto de B. A esta contido em B
$A \subset B$	Conjunto A é um subconjunto próprio de B. Condição: $B \subseteq A$ e $B \neq A$
$A \cap B$	Intersecção entre conjunto A e conjunto B
$\emptyset$	Símbolo matemático para conjunto vazio
$A \cup B$	União entre o conjunto A e o conjunto B
$++$	Símbolo matemático duplo sinal de soma. Representa incremento unitário de uma variável inteira

Sumário

1 Introdução

O constante crescimento na produção e armazenamento de informações propiciado principalmente pelas novas tecnologias torna necessário o desenvolvimento de técnicas para organizá-las a um ritmo cada vez mais acelerado para que todo esse conhecimento se torne acessível.

A produção de documentos textuais é abundante, com uma parte expressiva sendo produzida e armazenada em meios digitais, viabilizando o uso de métodos automáticos ou semiautomáticos de classificação, uma vez que o conteúdo já está representado em algum formato digital, favorecendo o processamento.

Uma possível solução para a organização da informação textual é o uso de métodos automáticos, que dispensam intervenção humana. O tratamento computacional de um texto tradicionalmente considera o conjunto de suas palavras, ponderadas com suas respectivas frequências de ocorrência em um determinado documento e nos demais textos que compõem o corpus documental. Esta representação é denominada bag-of-words (Van Rijsbergen, 1979), que desconsidera a posição que cada palavra ocupa em cada texto. Diversos modelos de recuperação, classificação e clustering de textos adotam esta maneira para representar e analisar a informação.

O modelo bag-of-words, no contexto desta discussão, apresenta dois problemas principais: (1) ambiguidade e (2) presunção de independência entre as palavras. A ambiguidade surge porque o modelo ignora o fato de que algumas palavras ortograficamente diferentes tenham o mesmo significado e que algumas palavras ortograficamente idênticas têm significados diferentes; para resolver esta ambiguidade é necessária uma análise baseada no contexto de uso, porém como as palavras são isoladas de seu contexto esta resolução de ambiguidade fica completamente prejudicada. O segundo problema é a presunção de independência entre as palavras, ignorando que as palavras não existem como unidade linguísticas isoladas, havendo sempre um relacionamento entre elas para formar estruturas mais complexas capazes de expressar uma ideia; ao interpretarmos um texto, são estas relações entre as palavras que nos auxiliam a extrair os conceitos e compreender a ideia expressa.

Ao buscar uma melhor solução para este problema de organização, tentando amenizar os problemas mencionados com o modelo bag of words, temos as técnicas de clustering. O clustering conceitual é uma técnica que analisa automaticamente as relações entre textos e os organiza em estruturas temáticas coerentes denominadas de clusters. Os textos agrupados dentro de um determinado cluster compartilham tópicos similares. Esse agrupamento temático também favorece a exploração e análise da informação, propiciando novas maneiras de visualização do corpus textual.

Esta tese propõe o uso de clustering a partir dos conceitos com o objetivo de melhorar a eficácia do processo. A definição de conceito adotada aqui é a mesma da *International Standard for Terminology Work – Principles and Methods* (ISO 704:2009, 2009) que os define como "[conceitos são] unidades de conhecimento que abstraem e representam um conjunto perceptível de objetos com as mesmas características", portanto objetos com características diferentes são abstraídos em conceitos diferentes. O termo é a expressão literal que representa o conceito, podendo um mesmo termo fazer referência a conceitos distintos de acordo com o contexto em que for empregado; por exemplo, o termo "manga" faz referência a dois conceitos distintos: o conceito de fruta e o conceito de parte de uma peça do vestuário; neste caso o termo é ambíguo, porém os conceitos não, os conceitos são inequívocos. Um sistema conceitual consiste em um conjunto de estruturas conceituais relacionadas entre si (ISO 704:2009, 2009).

O Clustering de documentos textuais é uma técnica de processamento que produz agrupamento de textos semelhantes segundo um ou mais critérios; estes agrupamentos são chamados de clusters. É utilizado frequentemente para descobrir padrões, tendências ou temas dentro de grandes conjuntos de dados textuais. A recuperação de informação baseada em clustering conceitual combina técnicas de recuperação de informação e clustering para melhorar a precisão e relevância dos resultados em sistemas de recuperação de informação. O objetivo é agrupar documentos ou informações similares com base em conceitos subjacentes, independente das correspondências exatas de palavras ou termos. O clustering conceitual busca organizar os documentos em grupos com base em características relacionadas aos conceitos presentes no conteúdo. Esses clusters representam conceitos ou tópicos, o que pode facilitar a navegação e a recuperação de informações relevantes.

Ao realizar a recuperação de informações baseada em clustering conceitual, os sistemas de busca podem levar em consideração os conceitos relacionados ao conteúdo dos documentos textuais, uma vez que os clusters representam conceitos. Isto permite uma abordagem mais intuitiva aos usuários que desejam encontrar informações específicas, mesmo desconhecendo as palavras-chave exatas porque torna possível a apresentação de interfaces de busca que visualmente demonstrem a relação conceitual existente entre documentos e a hierarquia formada entre essas relações, exemplo uma interface gráfica baseada em dendrograma dos documentos a partir da dissimilaridade conceitual entre eles.

É importante notar que a implementação dessa abordagem envolve várias etapas, incluindo a representação dos documentos e a identificação de características relevantes (denominadas *features*). O resultado obtido por meio da aplicação das técnicas de agrupamento pode ser utilizado, por exemplo, na criação de uma interface amigável para a navegação pelos resultados de busca de um sistema de recuperação de informação ou um mecanismo de busca na *Web*.

1.1 Motivação

Um sistema de Recuperação de informação textual pode ser esquematizado em dois blocos principais mediados por um terceiro bloco: (a) situado em uma das extremidades, um conjunto de documentos textuais; (b) na outra extremidade o usuário com sua necessidade de informação; (c) mediando as duas extremidades, uma função de busca, que é responsável pela seleção e ranqueamento dos documentos segundo a necessidade de informação do usuário.

Os documentos textuais são compostos por palavras que posicionadas em um contexto representam determinados conceitos. Os conceitos são diretamente inacessíveis, sendo necessário o uso de termos contextualizados representados por uma ou mais palavras dentro de uma construção sintaticamente válida formando o texto. A necessidade de informação do usuário será satisfeita apenas por determinados conceitos; portanto, o problema que a recuperação de informação busca solucionar é um problema conceitual: como representar os conceitos discutidos nos textos e como saber quais conceitos satisfarão o usuário do sistema.

Existem sistemas conceituais dos mais variados tipos, sendo alguns mais simples como vocabulário de termos controlados ou tão complexos quanto ontologias online. Cada sistema tem o seu foco específico em domínio, cobertura e abrangência. O uso destes sistemas já é consagrado dentro da Ciência da Informação, podemos citar como exemplo os vocabulários controlados utilizados em bibliotecas durante a indexação do acervo. Porém esta solução é fortemente dependente de um trabalho intelectual complexo, baseado em profissionais altamente treinados e qualificados ficando restrita a certos domínios do conhecimento.

Os sistemas automáticos de indexação e recuperação de informação, tradicionalmente, representam os textos do corpus documental como um conjunto não ordenado de termos e, durante seu uso, transformam a necessidade de informação manifesta pelo usuário em uma expressão de busca composta também por termos não ordenados que serão utilizados pela função de busca na seleção e ranqueamento dos documentos. Uma deficiência deste processo é que os conceitos quando reduzidos aos termos e removidos de contexto produzem termos por diversas vezes ambíguos prejudicando bastante a qualidade final da recuperação.

A utilização de clustering na recuperação de informação é uma abordagem que visa organizar grandes volumes de dados de forma a facilitar a sua recuperação e análise. O clustering textual envolve o agrupamento de documentos semelhantes em clusters, onde documentos pertencentes a um mesmo cluster compartilham características e tópicos semelhantes.

Essa técnica é particularmente útil na recuperação de informação, pois permite identificar grupos de documentos com característica comuns, o que pode facilitar a navegação e a busca por informações relevantes. Ao invés de apresentar uma lista linear de resultados, um sistema de recuperação de informação que utiliza clustering textual pode apresentar grupos de documentos relacionados, o que ajuda os usuários a entenderem a distribuição e a diversidade dos conteúdos disponíveis.

Existem várias abordagens para a implementação de clustering de documentos textuais, incluindo métodos baseados em palavras-chave, análise de tópicos, técnicas de aprendizado de máquina, como o *k–means*, e algoritmos de aprendizado profundo, como o *embedding-based clustering* (Dai; Bikdash; Meyer, 2017).

No contexto da recuperação de informação, clustering textual pode ser usado para melhorar a organização e a apresentação dos resultados de busca, bem como para descobrir padrões e temas presentes nos documentos. Além disso, essa técnica também pode ser aplicada em sistemas de recomendação e análise de sentimentos, entre outras aplicações.

Em vista do exposto, notamos que um sistema de indexação automática e recuperação é muito útil em diversas situações, porém a ambiguidade terminológica resultante das técnicas tradicionais prejudica o processo de recuperação de informação. Assim, considera-se que um sistema deveria operar mais em nível conceitual do que em nível terminológico, e que dada as dificuldades já expostas com os sistemas tradicionais, seria necessário que os conceitos pudessem ser extraídos de maneira automática. Isso nos leva à nossa questão de pesquisa.

1.2 Problema de pesquisa

O problema de pesquisa deste trabalho consiste em: Como organizar um acervo textual utilizando os conceitos presentes nos textos? Como extrair esses conceitos a partir do conjunto de documentos?

Em qual medida seria possível um algoritmo computacional, automaticamente, representar os conceitos expressos em um texto e utilizar essa representação conceitual no processo de indexação automática de documentos textuais?

1.3 Hipóteses

As hipóteses de pesquisa são: (1) processos estatísticos aplicados ao processamento de palavras permitem a representação de conceitos presentes nos textos por meio do agrupamento destas; (2) estes agrupamentos, quando utilizados em conjunto, servem como representação dos documentos presentes em um acervo durante o processo de indexação automática. Em síntese:

- Um conceito pode ser especificado por um conjunto de termos, que individualmente podem se mostrar genéricos ou ambíguos;
- A terminologia presente em um acervo de documentos textuais é capaz de especificar os conceitos necessários para a organização deste.

1.4 Pressupostos teóricos

- Os conceitos são expressos por um grupo de palavras, mesmo que polissêmicas, e em conjunto servem como representação de um texto. A polissemia da palavra é resolvida com contextualização;
- A terminologia presente em um conjunto de documentos textuais é capaz de especificar os conceitos necessários para a organização deste;
- É possível tratar estatisticamente a terminologia presente em um conjunto de documentos textuais, detectando padrões e extraindo contextos.

1.5 Objetivo

Devido à maneira como um sistema de Recuperação de Informação textual opera, qualquer alteração na representação dos textos impacta na qualidade do sistema, por isso, nosso objetivo geral é aprimorar o processo de recuperação de informação textual aperfeiçoando a exatidão da representação dos textos que compõe o corpus documental e os representando por agrupamentos conceituais. A partir deste objetivo geral, desdobram-se três objetivos específicos:

1. Extrair conceitos representados nos textos sem utilizar base externa de conhecimento;
2. Agrupar os conceitos sem o uso de um sistema conceitual predefinido;
3. Representar todos os textos do corpus documental por meio dos agrupamentos conceituais.

1.6 Metodologia

A pesquisa científica é um método de investigação que objetiva a solução dos problemas por meio da produção de novos conhecimentos (Barros; Lehfeld, 2002; Gil, 2002). O método científico é "[...] o conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo [...] traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista" (Marconi; Lakatos, 2003, p. 83). É um procedimento formal que consegue descobrir verdades parciais.

O presente estudo caracteriza-se como de natureza pura onde o conhecimento científico é desenvolvido sem uma preocupação direta com sua aplicação prática, sendo formalizado com o objetivo de construir teorias e leis; desenvolvido "com base em material já elaborado, constituído principalmente de livros e artigos científicos" (Gil, 2002, p. 44).

Trabalhos de cunho exploratório buscam levantar informações sobre um determinado objeto "delimitando assim um campo de trabalho, mapeando as condições de manifestações desse objeto" (Severino, 2016, p. 132).

A técnica de pesquisa empregada para o levantamento de dados é a pesquisa bibliográfica (fonte secundária) por meio de publicações científicas realizadas em livros e periódicos das áreas de text mining e recuperação de informação (Marconi; Lakatos, 2003).

1.7 Organização do trabalho

Este trabalho está organizado em seis seções, da seguinte maneira:

- Seção 1: *Introdução* — aqui apresentada, onde é discutida a motivação e a questão que a presente pesquisa pretende responder;
- Seção 2: *Conceitos básicos* — serão apresentados alguns conceitos básicos necessário para a discussão realizada em seções posteriores. Alguns conceitos são: documento, termo/conceito, necessidade de informação, relevância, features.
- Seção 3: *Clustering de documentos* — esta seção discute o conceito de clustering de documentos incluindo o conceito de cluster, medidas de similaridade, coeficientes de distância dentre outros;
- Seção 4: *Definindo um método de clustering baseado em conceito* – no qual apresentará a proposta de um método de clustering cujas característica atendam ao enunciado pela motivação deste trabalho;
- Seção 5: *Experimentação* — será descrito um experimento simples para ilustrar na prática todo o processo proposto por este trabalho;
- Seção 6: *Conclusões* — nesta seção são apresentadas as possibilidades que esta proposta abre bem como suas deficiências e sugestões para trabalhos futuros;
- Apêndices e Anexos: aqui estão alguns quadros e materiais de apoio para as seções anteriores.

2 Conceitos básicos

A temática principal deste trabalho, Recuperação de Informação e clustering, envolve três campos científicos distintos: a Ciência da Informação, a Matemática, e a Ciência da Computação, por isso poderão surgir problemas terminológicos resultantes das diferentes nomenclaturas utilizadas para um mesmo conceito.

Este trabalho preferencialmente utilizará a terminologia empregada na Ciência da Informação exceto os termos já consolidados nas outras ciências e amplamente utilizados, nestes casos a preferência será pelo termo mais comumente empregado.

2.1 Documento

Existem várias definições sobre o que é um documento. Uma definição abrangente foi dada por Suzanne Briet (1951), nos trazendo a definição de documento como uma representação de algo físico ou conceitual:

> Um documento é uma prova de um fato [..] qualquer evidência concreta ou simbólica, preservada ou registrada, com a finalidade de representar, de reconstituir ou comprovar um fenômeno físico ou intelectual[1] (Briet, 1951, p. 7, tradução nossa).

Desta forma a autora posiciona o documento com uma maneira de se obter acesso a uma evidência, elencando alguns exemplos, dentre os quais está o exemplo do antílope: Um antílope selvagem, vivendo na natureza não é um documento, mas a partir do momento em que ele é capturado, levado a algum zoológico e transformado em um objeto de estudo ele se torna uma evidência física e passa a ser um documento. Segundo Michael Buckland (1997, p. 806), as regras de Briet para determinar quando um objeto passa a ser um documento não são muito claras, ele faz quatro inferências: (1) A existência de materialidade, ou seja, apenas objetos físicos ou sinais físicos; (2) a intencionalidade, o objeto tem a intenção de ser tratado como evidência; (3) os objetos precisam ser processados, ou seja, precisa ser transformados em documento; (4) existência de um posicionamento fenomenológico, no qual o objeto é percebido como um documento.

Estendendo o conceito de documento, temos o documento digital. Michael Buckland (1998) ao discutir sobre o que é um documento digital traça um panorama sobre as definições de documento e como isso refletiu na denominação da ciência que lida com tal. Inicia com o surgimento do termo "bibliografia", afirmando que este veio da necessidade de nomear as

[1] *Un document est une preuve à l'apui d'un fait [...] tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène ou physique ou intellectuel* (Briet, 1951, p. 7)

"técnicas eficientes e confiáveis [...] para coleta, preservação (organização), representação (descrição), seleção (recuperação), reprodução (cópia) e disseminação dos documentos"[2] . Prossegue traçando um histórico desde o início do século XX onde a palavra documentação já era amplamente empregada na Europa, em substituição ao termo bibliografia e que após o ano de 1920, o termo "documentação" foi sendo gradativamente empregado para englobar bibliografia, serviços de informação acadêmica, gerenciamento de registros e trabalho de arquivo[3](Buckland, 1998). O autor conclui que essas alterações na terminologia refletem as mudanças conceituais que o documento sofreu ao longo do tempo e afirma que devemos: "definir um documento em termos de sua função ao invés de seu formato físico"[4] porque mesmo a saída produzida por algoritmos como, por exemplo, um para gerar uma tabela de logaritmos, é um documento digital dinâmico.

Em um sistema de recuperação de informação, a definição de documento como algo físico na qual a informação está contida atende perfeitamente, pois o processo de Recuperação de Informação ficará restrito à recuperação de documentos que potencialmente possuem a informação necessitada pelo usuário do sistema.

No escopo deste trabalho, documento será aquele objeto digital de característica estritamente textual escrito em alguma linguagem natural respeitando às regras ortográficas e gramaticais vigentes para aquela linguagem. Ele é o suporte material em que a informação está contida. O presente trabalho pretende tratar sobre documentos textuais escritos em língua portuguesa.

2.2 Termo e Conceito

Segundo Dahlberg (1978, p. 101)

> Desde que o homem foi capaz de pensar e de falar, empregou palavras (conjunto de símbolos) para designar os objetos de sua circunstância assim como para traduzir os pensamentos formulados sobre os mesmos. Foi também através de formas verbais que se fez entender pelos seus semelhantes.

Ainda segundo a autora, "a linguagem constitui a capacidade do homem designar os objetos que o circundam assim como de comunicar-se com os seus semelhantes". Assim, as linguagens utilizadas para a nossa convivência social são denominadas linguagens naturais. O ser humano também criou outros tipos linguagens: as linguagens artificiais ou linguagens especiais, tal como a linguagem da Química, a linguagem da Matemática, etc.

---

[2] *Efficient and reliable techniques were needed for collecting, preserving, organizing (arranging), representing (describing), selecting (retrieving), reproducing (copying), and disseminating documents. The traditional term for this activity was "bibliography"* (Buckland, 1998).

[3] *Encompass bibliography, scholarly information services, records management, and archival work* (Buckland, 1998).

[4] *It would be consistent with the trend, described above, towards a defining a document in terms of function rather than physical format* (Buckland, 1998).

Ainda segundo Dahlberg (1978), todo enunciado sobre objetos contém um elemento do respectivo conceito. Estes elementos se identificam com as chamadas características dos conceitos. Assim, a formação de um conceito se faz pela reunião e compilação de enunciados verdadeiros a respeito de determinado objeto. Para fixar o resultado dessa compilação de enunciados necessitamos de um instrumento constituído pela palavra ou qualquer signo que possa fixar essa compilação. Portanto, pode-se definir um conceito como a compilação de enunciados verdadeiros sobre um determinado objeto, fixada por um símbolo linguístico.

Mario Barité Roqueta (2000) traz a noção de conceito como um iceberg, onde a parte de baixo da linha da água é o conceito e a parte visível é a palavra, o símbolo que expressa esse conceito. Portanto a terminologia auxilia na manipulação dos conceitos. Toda conceitualização possui duas faces: uma interna e outra externa; a face interna corresponde ao conjunto de enunciados que definem o conceito enquanto a face externa corresponde à palavra ou qualquer outro símbolo que identificará o conceito

> Os termos são sim conceitualizações objetivadas, mas nunca objetos, são a expressão de artefatos abstratos de extrema complexidade que exigem uma aproximação multidisciplinar para descrevê-los e utilizá-los com excelência para nossas finalidades[5] (Barité Roqueta, 2000, p. 51, tradução nossa).

Walter Moreira (2019) define a relação entre termo e conceito da seguinte forma:

> Um termo é o resultado da equação que compreende o conceito mais sua designação verbal [...] as relações que são estabelecidas entre os termos são conceituais e não terminológicas [..] Tais relações podem ser, inclusive, orientadas por modelos de dados (Moreira, 2019, p. 19).

As relações estabelecidas entre termos e conceitos são estudadas pela terminologia. O conceito está situado em um âmbito cognitivo mais profundo e inacessível diretamente, cabendo aos termos oferecerem uma designação verbal, trazendo o conceito para o âmbito tangível e por isso comunicável. Neste trabalho adotamos a definição de Moreira (2019), pois uma vez que os termos estabelecem entre si relações conceituais, isso viabiliza delinearmos os conceitos a partir do estudo estatístico da relação de frequência e ocorrência entre conjuntos de termos, inclusive representarmos conceitos a partir destes conjuntos.

2.3 Terminologia

Terminologia é uma disciplina que estuda "o conjunto de termos de um domínio e dos conceitos (ou noções) por eles designados" (Barros, 2004, p. 34). Ao analisar os termos

[5] *Los términos son sí, conceptualizaciones objetivadas, pero nunca objetos, sino la expresión de artefactos abstractos de extrema complejidad, que exigen una aproximación multidisciplinaria para describirlos y utilizarlos con excelencia para nuestras finalidades* (Barité Roqueta, 2000, p. 51)

em contexto, é possível documentá-los e promover o seu uso correto. Este estudo pode ser limitado a uma língua ou pode cobrir mais de uma língua ao mesmo tempo (terminologia multilíngue, bilíngue, trilíngue, etc.).

Na tradução, a gestão da terminologia é um elemento central para uma boa legibilidade e correção técnica de textos traduzidos. Os tradutores profissionais administram a terminologia na forma de glossários bilíngues, usando ferramentas de controle de qualidade que fazem com que o mesmo termo técnico seja traduzido uniformemente em todo o texto.

A palavra "terminologia" pode assumir os seguintes significados:

- Uso e estudo de termos que são utilizadas em contextos específicos;
- Estudo dos termos técnicos usados em um determinado contexto ou domínio do conhecimento;
- Conjunto de termos utilizados em um contexto ou em um domínio, por uma pessoa ou em uma região geográfica;
- Estudo que identifica e delimita os conceitos característicos de uma área e a designação de cada um destes por um determinado termo.

A norma ISO 1087–1:2000 (2000, p. 10) define terminologia de duas formas:

1. conjunto de designadores pertencentes a uma linguagem especial;
2. ciência que estuda a estrutura, formação, desenvolvimento e gestão das terminologias em variados domínios, sendo "domínios" um campo especializado do conhecimento.

Cabré (1993) postula que dentro da ciência terminológica podem ser observadas três abordagens, não necessariamente excludentes: a terminologia como matéria autônoma, aquela com interesse nos sistemas de conceitos e organização do conhecimento (abordagem filosófica) e uma terceira centrada na linguística, considerando a terminologia como "um subcomponente do léxico e das linguagens especializadas" (Cabré, 1993, p. 32). A Teoria Geral da Terminologia (TGT), parte da primeira abordagem mencionada, a da terminologia como matéria autônoma que se preocupa com a natureza do conceito, as relações conceituais, as relações entre termo e conceito; e a atribuição de termos aos conceitos.

A Teoria Geral da Terminologia (TGT) surge como ciência a partir dos trabalhos de Eugen Wüster (1971). Segundo esta teoria, "os termos se definem uns em relação aos outros, formando assim um sistema" (Campos, 2001 apud Wüster, 1971, p. 68). Ela é considerada a base das correntes de estudo terminológico. Busca fixar os conceitos por meio de definições e estabelecer princípios para a criação de novos termos. Preocupa-se

em estabelecer uma comunicação mais precisa entre os especialistas de uma área do conhecimento humano.

Duas importantes características que diferenciam a TGT da Lexicografia são:

1. A lexicografia tem como escopo de trabalho a língua natural, resultado do devir histórico e repleta de polissemia, homonímia, sinonímia, etc; a TGT tem como seu escopo a língua artificial, desenvolvida e compreendida dentro de um grupo de especialistas, objetivando ser unívoca na relação entre conceito e denominação (Campos, 2001 apud Wersig, 1981, p. 67);

2. Na atividade lexicográfica, a unidade de estudo é a "palavra", enquanto na TGT o trabalho terminológico inicia no "conceito", que por definição possui uma "unidade de denominação" chamada "termo"; cabe a TGT o papel de unificá-los, exercendo uma função de natureza prescritiva. "O conceito é o significado do termo" (Campos, 2001, p. 66).

Nos estudos terminológicos existem duas grandes perspectivas para análise: análise semasiológica, que parte do termo para o conceito e análise onomasiológica, que parte do conceito para o termo. Em uma abordagem semasiológica, a terminologia estuda a verbalização do conhecimento, tratando os termos como unidades lexicais especializadas que operam dentro de um sistema linguístico próprio. Já em uma abordagem onomasiológica, existe a preocupação em se eliminar a ambiguidade da linguagem por meio da normalização terminológica, é neste tipo de abordagem que encontramos a TGT (Santos, 2010, p. 72–83).

A TGT classifica as relações em lógicas e ontológicas e dentro destas últimas existem as relações partitivas e associativas. Ela enfatiza o papel das características para a formação do conceito. Quando estas características são atribuídas é necessário estabelecer relações. Preocupa-se com os princípios que norteiam o estabelecimento destas relações (Sales, 2006, p. 72).

A terminologia como disciplina, segundo Cabré (1993, p. 100), necessita apoiar-se em três teorias: uma do conhecimento, uma da comunicação e uma da linguagem. Na situação específica da Recuperação de Informação, conforme lembra Cabré, são utilizadas ferramentas terminológicas com o objetivo de representar os conceitos contidos nos documentos e isto é feito mediante o uso dos termos, sendo comum o emprego de tesauros. Esses tesauros são basicamente "recompilações de termos relacionados semanticamente", pois a informação contida em um documento origina da disposição dos conceitos que este contém, sendo "cada conceito portador de informação, e entre os conceitos se estabelecem distintos tipos de relação" (Cabré, 1993, p. 101).

Neste trabalho utilizaremos a terminologia a partir de uma abordagem semasiológica de cunho estatístico, ou seja, partimos do termo para o conceito

assumindo que os conceitos, como entidades mais elaboradas, sempre serão compostos por mais de um termo, posicionados dentro de um contexto e que diversas palavras (sejam elas sinônimos ou formas flexionadas) remetem a um mesmo termo; também assumimos que é possível detectar esse comportamento da linguagem natural empregando-se técnicas de análise estatística sobre a representação textual.

2.4 Necessidade de Informação e Relevância

Necessidade de Informação e relevância estão relacionados ao que o usuário busca e ao que ele obtém como resposta de sua busca. Esta necessidade de informação é aquilo que o usuário possui, é uma lacuna em seu conhecimento que ele necessita preencher e para isso recorre a um sistema de recuperação. Relevância é a avaliação que o usuário faz a respeito dos resultados obtidos a partir de sua busca; é um julgamento a respeito do conteúdo dos documentos recuperados, se estes foram ou não capazes de satisfazer sua a necessidade de informação.

De acordo com Sanz Casado (1994, p. 19), o usuário da informação é o "indivíduo que necessita de informação para o desenvolvimento de suas atividades". Essa necessidade surge quando um indivíduo se depara com uma situação ou problema que requer conhecimento ou orientação. As demandas por informações podem ser influenciadas por fatores internos, como conhecimentos prévios, o desejo por saber, valores, crenças e interesses, ou externos, como situações cotidianas, problemas complexos que demandam resolução, mudanças sociais, econômicas e tecnológicas (Figueiredo, 1994).

As demandas por informação podem ser entendidas como as necessidades informacionais de um usuário frente a um problema. A compreensão das necessidades informacionais, muitas vezes, é associada às necessidades cognitivas de uma pessoa, ou seja, a falta ou deficiência de conhecimento ou de compreensão sobre um problema, que podem ser expressas em perguntas ou tópicos apresentados a um sistema ou fonte de informação. No entanto, a informação também precisa satisfazer necessidades emocionais ou afetivas, pois a busca e o uso da informação acontecem em situações sociais (Choo, 2003).

As necessidades informacionais são geradas principalmente pelo desempenho de tarefas organizacionais, como planejamento e tomada de decisões, e por fatores relacionados à personalidade do usuário. Desta forma, as necessidades emocionais, como a necessidade de conquista, de expressão e de realização, são igualmente importantes na busca pela informação (Choo, 2003). Além disto, à medida em que as circunstâncias, os interesses e os objetivos dos indivíduos evoluem ou mudam, suas necessidades de informação também se modificam e se adaptam.

Inicialmente, o indivíduo pode perceber uma sensação de intranquilidade ou inadequação com seu conhecimento, o que o leva a buscar informações. Gradualmente, o indivíduo forma uma opinião sobre a importância do problema e identifica os vazios de

informação que precisam ser preenchidos. A consciência da necessidade de informação nem sempre leva à busca, pois a pessoa pode decidir aceitar desconsiderar o problema, levando em conta a importância do assunto, seu conhecimento sobre o tema e o esforço necessário para fazer a busca (Choo, 2003).

Segundo Wilson (2006), o conceito de necessidade informacional como uma experiência subjetiva que ocorre exclusivamente na mente de cada indivíduo, sendo inacessível diretamente ao observador. Para identificar esta necessidade é necessário deduzir o problema por meio da observação do comportamento do indivíduo ou através da própria manifestação da necessidade pelo indivíduo. Os estudos de Dervin (1992) destacam o caráter cognitivo e não observável das necessidades informacionais, apontando para a existência de lacunas que podem gerar descontinuidade no conhecimento humano. Quando percebidas pelo indivíduo, estas lacunas suscitam o surgimento da necessidade informacional.

Belkin (1980) explicou o fenômeno da lacuna informacional a partir do conceito de estado anômalo do conhecimento. Belkin (1980, p. 44, tradução nossa), considera que, quando "o estado de conhecimento do usuário em relação a um tópico é de alguma forma inadequado em relação à capacidade da pessoa de alcançar algum objetivo"[6], este pode ser entendido como anômalo. Tal percepção de lacuna de conhecimento leva o usuário a buscar ativamente por informações adicionais, a fim de corrigir essa lacuna e adquirir um nível de conhecimento mais completo e sólido. O usuário reconhece que precisa preencher esta falta de informação para tomar decisões informadas e alcançar seus objetivos. Em adição a isso, Le Coadic (2004, p. 8) destacou que:

> Nosso estado (ou nossos estados) de conhecimento a respeito de determinado assunto, em determinado momento, são representados por uma estrutura de conceitos ligados por suas relações: nossa imagem de mundo. Quando constatamos uma deficiência ou anomalia desse(s) estado(s) de conhecimento, encontramos um estado anômalo de conhecimento. Tentamos obter uma informação ou informações que corrigirão essa anomalia. Disso, resultará um estado novo de conhecimento (Le Coadic, 2004, p. 8).

Le Coadic (2004, p. 40) ainda complementou, indagando:

> [...] o que leva uma pessoa a procurar informação? A existência de um problema a resolver, de um objetivo a atingir e a constatação de um estado anômalo de conhecimento, insuficiente e inadequado (Le Coadic, 2004, p. 40).

Nota-se que a ausência de conhecimento sobre um determinado assunto ou um problema específico do usuário pode desencadear o processo de busca por informações.

[6] *the user's state of knowledge with respect to a topic is in some way inadequate with respect to the person's ability to achieve some goal* (Belkin, 1980, p. 4).

Esta demanda por conhecimento exerce influência sobre o comportamento do usuário diante de seu problema informacional.

Le Coadic (2004, p. 39) indicou que as necessidades e os usos da informação são "interdependentes, se influenciam reciprocamente de uma maneira complexa que determinará o comportamento de um usuário e suas práticas". Desta forma, considerando o contexto em que a necessidade informacional se manifesta, as habilidades e competências do usuário, bem como, a disponibilidade de acesso a recursos informacionais, tem-se o que se denomina de comportamento informacional. Desta forma, o comportamento informacional se refere às atividades, às estratégias e aos processos que os indivíduos empregam para buscar, acessar, avaliar e utilizar informações que atendam às suas necessidades informacionais.

Stefano Mizzaro (1998) demonstra como ocorre o processo da representação da necessidade informação do usuário em um sistema de Recuperação de Informação, fazendo uma distinção entre a necessidade real de informação (RIN – *Real Information Need*), a necessidade percebida (PIN – *Perceived Information Need*), a requisição (*Request*) e a expressão de busca (*Query*).

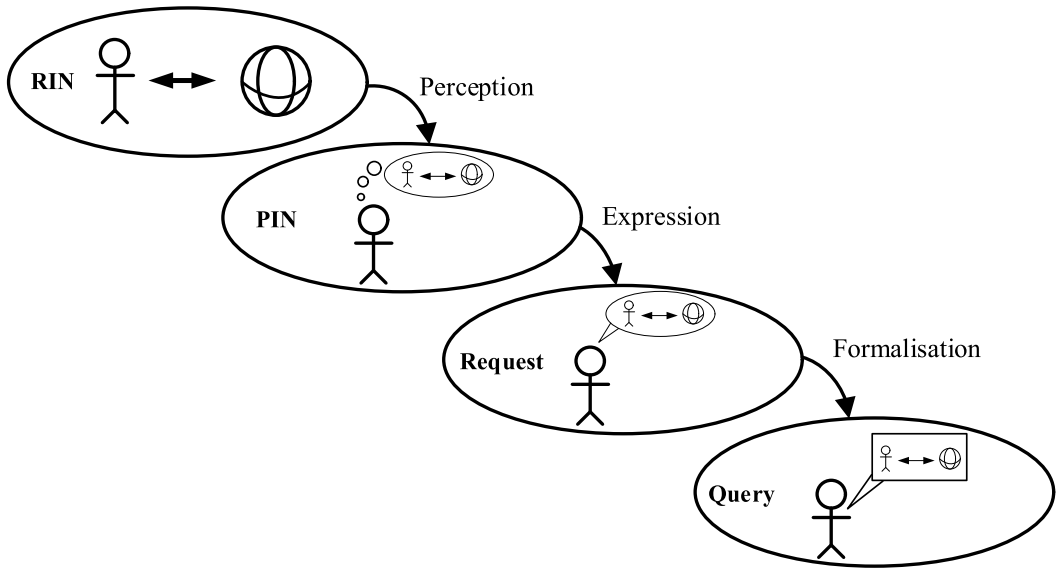
O processo de representação começa com uma lacuna no conhecimento do usuário a respeito de um determinado assunto, esta é a RIN, que ao percebê-la e tentar pensar a seu respeito torna-se PIN, ou seja, a PIN é a representação mental da RIN, não sendo necessariamente correta ou completa; a seguir esta PIN é expressa por meio de uma representação em linguagem humana, que é denominada *request* e por fim, esta *request* é formalizada em uma expressão de busca (*query*). Todo esse processo é apresentado na Figura 1.

Na Figura 1 é ilustrado o processo de representação da necessidade de informação conforme descrito por Mizzaro (1998). Nesta figura podemos observar os quatro estágios de representação mental que uma necessidade informacional passa até que seja formalizada em uma *query* e submetida a um sistema de recuperação de informação. Inicialmente existe apenas uma lacuna informacional, a *Real Information Needed* (RIN); esta lacuna, por meio do processo de percepção, emerge na mente do usuário sob a forma de uma *Perceived Information Need* (PIN); que por sua vez precisa ser expressa por meio de uma *Request*, uma expressão verbalizada da necessidade percebida; a *Request* então é formalizada em uma *Query* e será fornecida para o sistema de recuperação de informação.

Ao compreender e interpretar a informação recuperada, o usuário avalia seu potencial de relevância ou pertinência. Lima e Campos (2022) definiram relevância como:

> [...] grau de similaridade entre termos que compõem as expressões de buscas de usuários e a ocorrência em documentos da coleção ou nos termos de indexação. Já a pertinência é a relação que existe entre a informação obtida em uma busca que responde à necessidade ou demanda

Figura 1 – Representação da Necessidade de Informação



Fonte: Mizzaro (1998, p. 306)

> de informação do usuário, ou seja, a informação que é útil para o usuário (Lima; Campos, 2022, p. 2).

Relevância, porém, é um conceito bastante subjetivo e elástico, dependendo do contexto em que é avaliado. A avaliação de relevância da informação depende de uma série de fatores objetivos e subjetivos dos usuários, que devem ser considerados para o projeto e aprimoramento de Sistemas de Informação (SIs). Araújo (1995, p. 31) diz que este conceito é

> [...] indissociável dos sistemas de recuperação da informação, do usuário e dos documentos, é extremamente subjetivo e sujeito a variações de interpretação e julgamento, dependendo dos momentos e condições iniciais do sistema, dos diferentes usuários e dos documentos em suas inter-relações. Qualquer alteração nessas variáveis pode mudar drasticamente os resultados esperados. O que é relevante para um elemento do sistema (responsável, por exemplo, pela seleção, indexação e demais processos) pode ou não ser para ele próprio em um outro momento no tempo, ou para outros elementos do sistema; o que é relevante para um usuário em um tempo T pode não ser para outros usuários ou para o mesmo em outro tempo T1; documentos têm sua própria relevância, a priori do sistema, e que pode ser alterada dependendo do conjunto ao qual esses documentos venham a pertencer; há, no sistema, uma imprevisibilidade de comportamento. E o sistema como um todo é sensível às alterações que tais imprevisibilidades vão provocar em suas variáveis – alterações estas que, conforme visto, não são lineares (Araújo, 1995, p. 31).

De forma geral, a relevância da informação pode ser considerada como um atributo de qualidade, ou seja, um atributo que indica se a recuperação de uma informação em

um SI atende as necessidades e as expectativas do usuário. Isso inclui aspectos como precisão, completude, atualidade, relevância, confiabilidade e clareza da informação. Uma informação de alta qualidade é aquela considerada precisa, confiável, relevante e útil para o usuário em um contexto específico.

A percepção de relevância envolve diversos fatores além do atributo de qualidade, uma vez que os processos emocionais e lógicos envolvidos na percepção de relevância variam de acordo com as condições potenciais de uso da informação (Kuhlthau, 1991). Identificar o que é relevante para um usuário específico é, portanto, condicionado por vários fatores complexos e inter-relacionados, tanto sistêmicos quanto relacionados à apropriação do uso da informação (Araújo Júnior, 2005).

Segundo Borlund (2003), duas abordagens podem ser consideradas para o estudo da relevância em Sistemas de Recuperação de Informação (SRIs): uma abordagem orientada pelo sistema, e outra abordagem cognitiva orientada pelo usuário, cada uma com sua própria compreensão de relevância.

A abordagem orientada pelo sistema trata a relevância como um conceito estático e objetivo, enquanto a abordagem cognitiva orientada pelo usuário considera a relevância como uma experiência mental subjetiva, individualizada, que envolve reestruturação cognitiva. Diferentes tipos de relevância são identificados em ambas as abordagens, incluindo relevância do sistema ou algorítmica, relevância semelhante ao tópico, pertinência ou relevância cognitiva, relevância situacional e motivacional, e afetiva (Borlund, 2003).

A relevância algorítmica é o tipo mais comum e claro de relevância de sistema, é usada na avaliação tradicional de sistemas de RI, medindo quão bem o tópico da informação recuperada corresponde ao tópico da solicitação. Este tipo de relevância é restrito, pois lida apenas com o grau em que a representação da consulta corresponde ao conteúdo dos objetos de informação recuperado. A especificação dos tipos objetivos de relevância algorítmica pode ser rotulada como "tópico igual a conteúdo" (Borlund, 2003).

As abordagens cognitivas, por sua vez, são mais direcionadas às particularidades dos usuários e, portanto, têm maior adequação para avaliar SRIs que lidam com perfis distintos de indivíduos. A abordagem cognitiva mais comum é a relevância situacional, que leva em conta a relação individual do usuário com a informação que foi recuperada. Diferentemente de outras abordagens, a relevância situacional não se baseia apenas na relação entre uma representação de consulta e um objeto de informação recuperado. Em vez disto, esta abordagem considera a utilidade ou o valor do objeto de informação em relação à tarefa de trabalho em questão. A relevância situacional envolve aspectos motivacionais e afetivos, considera a característica de todos os tipos de relevância subjetiva, descrevendo a relação entre as intenções, objetivos e motivações do usuário, e os objetos de informação (Borlund, 2003).

Embora a relevância situacional seja o conceito mais comumente considerado no estudo das relevâncias cognitivas, é importante mencionar outras abordagens complementares, que contribuem para a compreensão da subjetividade do usuário na avaliação de informações recuperadas. Na literatura, encontram-se os conceitos de "relevância psicológica" (Harter, 1992), "relevância ostensiva" (Campbell; Van Rijsbergen, 1996) e "relevância da tarefa" (Mizzaro, 1998; Reid, 1999).

A relevância psicológica, proposta por Harter (1992), está alinhada com as ideias básicas e fundamentais do ponto de vista cognitivo: a mudança das estruturas de conhecimento do receptor pelo ato de processamento de informações. A relevância psicológica descreve um estado de efeito que existe quando o usuário recupera informações, que sugerem novas conexões cognitivas, analogias frutíferas, metáforas esclarecedoras, aumento ou diminuição na força de uma crença. Assim, a relevância psicológica é vista como o efeito de uma mudança nas estruturas de conhecimento.

Em complemento, Campbell e Van Rijsbergen (1996) propuseram o conceito de relevância ostensiva, que se refere ao grau em que as evidências do objeto de informação recuperado são representativas da necessidade atual de informação do usuário. Esse conceito leva em conta a ideia de que a necessidade de informação é dinâmica, refletida na ponderação de probabilidade no "modelo ostensivo" de RI (Campbell; Van Rijsbergen, 1996). Este modelo define a relevância como uma relação entre a consulta e o documento, em que um documento é considerado relevante se for útil para atender a necessidade de informação do usuário.

A relevância, segundo o modelo ostensivo, é uma característica do documento em relação à consulta, e não uma propriedade inerente do documento em si. Este modelo reconhece a subjetividade do usuário na avaliação da relevância e na escolha dos documentos, destacando a importância da interpretação da consulta pelo sistema, levando em consideração o contexto em que a consulta foi formulada e as necessidades do usuário (Campbell; Van Rijsbergen, 1996). Assim, o modelo ostensivo de recuperação da informação é importante para a compreensão da subjetividade do usuário na avaliação da relevância e na escolha dos documentos mais adequados para sua necessidade de informação.

Em uma outra conceituação, Mizzaro (1998) propôs um modelo de relevância definida como a relação entre um recurso de informação e a representação do problema do usuário, avaliada de acordo com tópico, tarefa e/ou contexto em um determinado momento. Ainda para Mizzaro (1998), o tipo final de relevância é a das informações recebidas para a necessidade real de informação do usuário em um determinado momento. Isto direciona a avaliação para o contexto de uso efetivo da informação, em detrimento do conceito de pertinência sugerido por Foskett (1972).

Reid (1999) complementou esta abordagem, chamando este tipo de relevância de "relevância da tarefa", embora sua definição seja idêntica à relevância situacional. Tanto Mizzaro (1998) quanto Reid (1999) estão interessados em capturar a utilidade percebida pelo usuário, os objetos de informação recuperados com referência à tarefa e necessidade real de informação.

Os autores Schamber, Eisenberg e Nilan (1990, p. 774, tradução nossa), de forma sumarizada, destacaram que a relevância da informação pode ser classificada em três categorias:

1. Relevância é um conceito cognitivo multidimensional cujo significado depende, em grande parte, das percepções de informação dos usuários e de suas próprias situações de necessidade de informação;
2. Relevância é um conceito dinâmico, que depende dos julgamentos dos usuários sobre a qualidade da relação entre a informação e a necessidade de informação em um determinado momento;
3. Relevância é um conceito complexo, mas sistemático e mensurável se abordado conceitualmente e operacionalmente a partir da perspectiva do usuário.

Ao considerar as abordagens mencionadas sobre o que pode ser considerado relevante para um usuário de SRI, é perceptível a presença de possíveis desafios para avaliar a informação recuperada, devido à subjetividade envolvida na avaliação, que inclui aspectos emocionais, cognitivos e dinâmicos da relação entre o usuário e o sistema.

Assim, a avaliação de relevância algorítmica não é adequada para fornecer resultados satisfatórios neste contexto, tornando-se uma ferramenta inadequada para avaliar sistema de informação digitais. De forma resumida, quatro pontos são percebidos como desafiadores para uma plena avaliação da relevância em sistemas de recuperação da informação: (1) a avaliação da relevância, (2) a avaliação de satisfação, (3) a atualização de conhecimento de um usuário e (4) a incerteza nas buscas por informação.

A avaliação da relevância de uma informação recuperada, por parte de um usuário, depende de certo conhecimento anterior sobre a temática ou a problemática de busca. Sem isso o usuário pode ter dificuldade em encontrar valor ou relevância na informação recuperada. Sobre isto, Hjørland (2010, p. 231) aponta que determinar quais itens são relevantes em relação a um determinado objetivo/tarefa requer conhecimento do assunto e depende de diferentes teorias/visões. Os usuários de sistemas de informação, portanto, não são automaticamente competentes para julgar a relevância Hjørland (2010, p. 231, tradução nossa)[7].

[7] *To determine which items are relevant in relation to a given goal/task requires subject knowledge and is dependent on different theories/views. Users of information systems are therefore not automatically competent to judge relevance* (Hjørland, 2010, p. 231).

Isso significa que a importância tanto da seleção quanto da produção de informações por um usuário depende do conhecimento prévio que ele tem sobre a informação em questão. No entanto, em contextos de produção de ambientes digitais com grande volume de dados, a análise de relevância se torna menos importante. Isto ocorre porque toda a produção de conteúdo em formato digital é, geralmente, armazenada pois a dinâmica das ferramentas digitais permite que usuários comuns produzam dados e informações em grandes volumes.

Durante a avaliação de satisfação do processo de recuperação de informação, é importante notar que nem sempre está relacionada à relevância da informação recuperada. Outros aspectos do processo, como desempenho do sistema de informação, apresentação da informação etc., podem influenciar na satisfação geral do usuário. Além disso, pode ser difícil, para o usuário, distinguir entre conceitos, como satisfação e relevância, e ainda, avaliar a qualidade da informação recuperada. Por fim, é importante ressaltar que a avaliação da relevância da informação recuperada nem sempre é o único fator que influencia na escolha do usuário, sobre quais registros utilizar.

Os autores Coeira e Vickland (2008) identificaram que, dependendo do processo de recuperação da informação, a avaliação da relevância pelos usuários não é um preditor forte do impacto das informações na tomada de decisão. Coeira e Vickland (2008) observam que a relação entre a relevância percebida e o impacto na tomada de decisão é complexa e deve ser investigada mais a fundo. Eles sugerem que outros fatores, como a credibilidade da fonte, a experiência do usuário e a confiança na tecnologia, também podem desempenhar um papel importante na tomada de decisões, conforme apresentado por Hjørland (2010) e Kuhlthau (1991). Schamber, Eisenberg e Nilan (1990) reforçam essa ideia:

> [...] quando a satisfação é operacionalizada como uma medida na avaliação do desempenho de sistemas de informação, pode na verdade ser uma medida composta que contém vários tipos de julgamentos, incluindo julgamentos de relevância. A relevância e outros julgamentos (às vezes, o termo relevância nem é utilizado) podem ser usados para avaliar aspectos amplamente variados de um sistema (Schamber; Eisenberg; Nilan, 1990, p. 760, tradução nossa)[8].

Também, a atualização do conhecimento de um usuário, tendo em vista a dinamicidade informacional, pode alterar sua posição de relevância sobre uma mesma busca. Usuários com conhecimentos atualizados podem não mais ver relevância em informações às quais já possuem conhecimento. Hjørland (2010) afirmou que:

> [...] o conhecimento está sempre atualizado, o próprio conhecimento muda dinamicamente e, portanto, a natureza dinâmica das "necessidades de informação" e da "relevância" é, em grande parte, causada por essa

[8] *when satisfaction is operationalized as a measure in evaluating the performance of information systems, it may actually be a composite measure that contains several kinds of judgments, including judgments of relevance. Relevance and other judgments (sometimes the term relevance is not used at all) may be used to evaluate widely varied aspects of a system* (Schamber; Eisenberg; Nilan, 1990, p. 760)

> mudança em nosso conhecimento coletivo. Na literatura da Ciência da Informação, a natureza dinâmica da "relevância" está, no entanto, frequentemente ligada ao usuário, e não ao conhecimento em si (Hjørland, 2010, p. 222, tradução nossa)[9].

Os aspectos afetivos, como confusão e incerteza nas buscas, podem também afetar o julgamento sobre a relevância de uma informação recuperada (Kuhlthau, 1991). Em muitos casos, como visto, devido ao estado anômalo de conhecimento do usuário (Coeira; Vickland, 2008), a avaliação da qualidade da informação recuperada, em termos de sua relevância, só pode ser feita após a apropriação e uso da informação. Além disso, barreiras linguísticas, cognitivas e simbólicas podem fazer com que uma informação seja descartada, devido à incapacidade do usuário de avaliar sua relevância. Isso se torna ainda mais desafiador, uma vez que sistemas de informação digitais, geralmente não são projetados para considerar os problemas dos usuários (Barlow, 2013).

A fim de mitigar esses problemas, métodos de avaliação de relevância e pertinência da informação foram propostos em pesquisas como a de Araújo Júnior (2005), Manning, Raghavan e Schütze (2008). Esses métodos visam à avaliação da resposta do usuário frente a uma informação recuperada, fornecendo feedback para o sistema, com objetivo de calibrá-lo na relação de busca e recuperação da informação.

Dentro desta discussão sobre necessidade de informação e relevância é pertinente discutir também os sistemas de informação e recuperação de informação.

2.5 Sistemas de Informação e a Recuperação de Informação

Para Cesarino (1978, p. 224), um sistema de informação é uma organização ou unidade social que procura atingir um objetivo específico, servindo de fonte intermediária entre o produtor e o consumidor da informação. Cohen (1995, p. 14) complementa tal afirmação, pontuando que um sistema de informação é um "conjunto de canais formais e informais de comunicação da informação dentro de uma organização ou de uma comunidade". Nota-se que nesse contexto um sistema de informação é aquele que atua como ferramenta comunicacional para uma comunidade, com um objetivo específico, de produtor para usuário.

Segundo Araújo, sistemas de informação:

> [...] são aqueles que, de maneira genérica, objetivam a realização de processos de comunicação. Alguns autores contextualizam sistemas de informação mais amplamente para incluir sistemas de comunicação de massa, redes de comunicação de dados e mensagens etc., independentemente

[9] *knowledge is always updated, knowledge itself changes dynamically, and therefore the dynamic nature of "information needs" and "relevance" is to a very large degree caused by this change in our collective knowledge. In the literature of information science, the dynamic nature of "relevance" is, however, often connected to the user, rather than to knowledge itself* (Hjørland, 2010, p. 222)

> da forma, natureza ou conteúdo desses dados e mensagens (Araújo, 1995, p. 1).

Tais sistemas pressupõem a existência de um arcabouço informacional que se adapta às mudanças da sociedade para as quais são pensados, sendo chamados de "estoques de informação" (Smit; Barreto, 2002). Os SIs são projetados para lidar com a movimentação de tais estoques, que armazenam registros feitos pelos usuários para uso presente ou futuro. De um modo geral, os estoques informacionais são recursos que armazenam informações e dados em diferentes formatos. Eles são importantes para a preservação, acesso e compartilhamento do conhecimento, contribuindo para o funcionamento dos sistemas de informação. Sobre isso, Smit e Barreto (2002) apontam que a:

> [...] produção da informação documentária é operacionalizada por meio de práticas bem definidas e se apoia em um processo de transformação orientado por uma racionalidade técnica específica; representa atividades relacionadas à reunião, seleção, codificação, redução, classificação e armazenamento de informação. Todas essas atividades estão orientadas para a organização de estoques de informação, de uso imediato ou futuro. Esse repositório de informação representa um estoque potencial de conhecimento e é imprescindível que exista, para que se realize a transferência de informação (Smit; Barreto, 2002, p. 4).

Embora as informações armazenadas em bancos de dados, bibliotecas, arquivos ou museus, tenham a capacidade de gerar conhecimento, isto só ocorre por meio de uma ação de comunicação mutuamente acordada entre a fonte (os estoques) e o receptor (Smit; Barreto, 2002). Desta forma, os estoques de informação dependem de processos que visam à disponibilização da informação produzida e armazenada para os indivíduos, através de um fluxo que conecta o estoque e o indivíduo.

Sobre os fluxos informacionais destes estoques, os autores destacam que:

> [...] dois critérios permeiam o fluxo da informação entre os estoques, ou espaços de informação, e os usuários: o critério da tecnologia da informação, que almeja possibilitar o maior e melhor acesso à informação disponível, e o critério da Ciência da Informação, que intervém para qualificar este acesso em termos das competências que o receptor da informação deve ter para assimilar a informação, ou seja, para elaborar a informação para seu uso, seu desenvolvimento pessoal e dos seus espaços de convivência. Não é suficiente que a mensagem esteja disponível, ela deve também poder ser apropriada pelo receptor (Smit; Barreto, 2002, p. 15–16).

As regras de seleção do que será armazenado em um estoque informacional devem levar em conta as especificidades dos usuários desse estoque, cabendo à instituição responsável pelo estoque determinar o que armazenar e como apresentar tais informações, através de sistema próprio de informação. É importante indicar que a "[...] produção dos

estoques de informação não possui um compromisso direto e final com a produção do conhecimento" (Barreto, 1994, p. 4).

A informação é criada dentro de um contexto de produção, selecionada e armazenada a partir de um critério com vista em um usuário específico, sendo apresentada para este indivíduo frente à sua demanda (Smit; Barreto, 2002). O indivíduo, em posse de uma informação, analisa sua relevância e, ao se apropriar dela operacionaliza sua ação inicial para qual empreendeu a abordagem ao estoque.

Sistemas de informação podem ser classificados de acordo com sua função, área de aplicação ou modelo de processamento de informações. Alguns tipos de sistemas de informação são: sistemas de processamento de transações, sistemas de informação gerencial, sistemas de suporte de decisão, sistemas de informação executiva, sistemas de recuperação de informações, sistemas de informação geográfica, sistemas de automação de escritório, sistemas de comércio eletrônico, sistemas de informação em saúde, entre outros (Bio, 1996). Cada um destes sistemas tem sua própria finalidade sendo usados para atender a diferentes necessidades das organizações e usuários.

Sistemas de informação podem ser classificados como digitais ou físicos. Enquanto os sistemas físicos utilizam formatos de armazenamento tangíveis, como livros e revistas, os sistemas digitais utilizam arquivos eletrônicos, como bancos de dados e imagens digitais. A acessibilidade da informação, ou seja, a capacidade de um usuário de compreender forma e conteúdo da informação que lhe é recuperada, é outra diferença importante, pois os sistemas tradicionais só podem ser acessados em locais físicos, enquanto os digitais podem ser acessados remotamente de qualquer lugar do mundo com conexão à rede de dados (geralmente internet). Sobre SRIs digitais, Vickery e Vickery (2004) apresentaram a seguinte definição:

> [...] a essência da recuperação eletrônica é que uma coleção de mensagens é armazenada em algum meio legível por computador [...] e é acessada por um software executado em um computador ao qual o armazenamento está vinculado. Um sistema pode ser pessoal [...], institucional [...], ou público (Vickery; Vickery, 2004, p. 117, tradução nossa) [10].

Os sistemas digitais possuem uma capacidade de armazenamento muito maior em comparação com os sistemas físicos, permitindo o armazenamento de um número virtualmente infinito de documentos, limitado à sua capacidade física e de escalonamento do banco de dados. Outro aspecto relevante diz respeito à forma de consultar esses recursos informacionais, uma vez que os sistemas digitais possibilitam buscas mais precisas e eficientes, graças a recursos como índices e palavras-chave, enquanto nos sistemas físicos a busca pode ser mais demorada e trabalhosa. Por fim, a interatividade é um aspecto que

[10] *[...] the essence of electronic retrieval is that a collection of messages is stored in some computer-readable medium [...] and is accessed by software run on a computer to which the store is linked. A system may be personal [...], or institutional [...] or public* (Vickery; Vickery, 2004, p. 117)

diferencia significativamente os sistemas digitais, uma vez que permitem que os usuários interajam com os documentos de várias maneiras, enquanto nos sistemas tradicionais essa interatividade é limitada.

Ambientes informacionais digitais têm se destacado, em termos tecnológicos, como os locais onde as problemáticas mais evidentes em sistemas de informação digitais são encontradas, e podem ser considerados como uma base sólida para a análise desses ambientes. Portanto, será dada prioridade a este tipo de sistema de recuperação de informações, e as descrições fornecidas serão específicas para este modelo de funcionamento do sistema de informação.

Um sistema de recuperação de informação digital é um tipo específico de sistema de informação com uma finalidade bem definida. Enquanto os sistemas de informação, em geral, visam armazenar, processar e disseminar informações de forma geral e estruturada, para que os usuários possam acessá-las e utilizá-las eficientemente, os sistemas de recuperação da informação digital têm como objetivo recuperar informações específicas dentro de um conjunto de dados ou documentos, usando técnicas de busca e recuperação da informação. Araújo Júnior (2005, p. 69) descreve a busca e a recuperação da informação

> [...] como o processo de localizar documentos e itens de informação que tenham sido objeto de armazenamento, com a finalidade de permitir o acesso dos usuários aos itens de informação, objetos de uma solicitação (Araújo Júnior, 2005, p. 69).

De forma direta, sistemas de recuperação da informação digital podem ser definidos como:

> [...] um sistema de operações interligadas para identificar, dentre um grande conjunto de informações (uma base de dados, por exemplo), aquelas que são de fato úteis, ou seja, que estão de acordo com a demanda expressa pelo usuário (Araújo Júnior, 2005, p. 77).

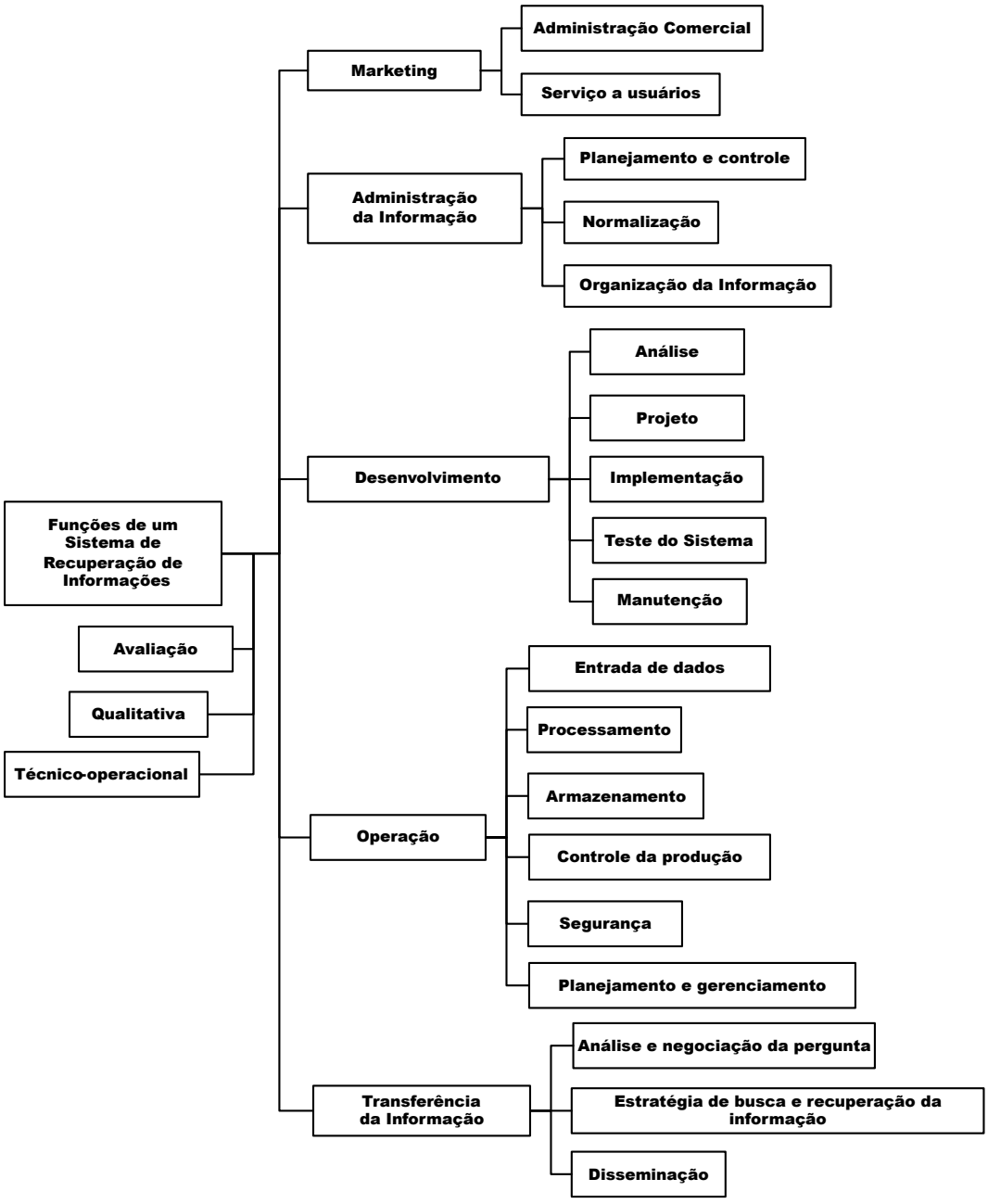
Como visto, em consonância com o apresentado sobre SIs, as preocupações com as necessidades dos usuários são mantidas, bem como, as preocupações relativas ao processamento das informações estocadas. Ao projetar e desenvolver um sistema de recuperação de informação digital é importante levar em consideração as necessidades e especificidades dos usuários potenciais. A função principal dos sistemas de recuperação da informação digital é satisfazer a necessidade de informação do usuário, levando-o ao documento correto, ou informação que atenderá sua necessidade específica de informação (Araújo Júnior, 2005).

É possível, assim, descrever que os sistemas de recuperação de informação digital têm a função de representar o conteúdo dos documentos presentes no seu estoque informacional digital e disponibilizá-los para o usuário, de forma que possibilite a seleção

rápida dos itens que atendam total ou parcialmente às suas necessidades de informação, expressas através de uma expressão de busca (Ferneda, 2003).

Araújo Júnior (2005, p. 82), na Figura 2, apresenta algumas funções possíveis para sistemas de recuperação de informações, digitais ou não, considerando as necessidades informacionais de usuários.

Figura 2 – Funções de um sistema de recuperação de informações



Fonte: Araújo Júnior (2005, p. 82)

Ingwersen (1996) apresenta um modelo de RI em ambientes computacionais, considerando o espaço cognitivo do usuário da informação, conforme apresentado na Figura 3.

Figura 3 – Modelo de SRI proposto por Ingwersen (1996)

Fonte: Ingwersen (1996, tradução nossa)

Para Ingwersen (1996), a busca de informação em um sistema de RI é focada na busca ativa, na qual o usuário aborda o sistema de recuperação de informação. No entanto, Ingwersen identifica uma série de outros elementos que devem ser considerados. Primeiro, o autor destaca que cada área do modelo (que é um modelo de comportamento de busca de informação) inclui várias entidades, como o usuário da informação, o autor do documento, o intermediário, a interface e o sistema de RI.

Cada uma dessas entidades tem uma função específica na interação do usuário com o sistema de RI, e cada função é baseada em modelos cognitivos explícitos ou implícitos do domínio da busca de informação. Desta forma, todos os elementos de um SRI devem considerar as especificidades de necessidade informacional de seus usuários, bem como, suas condições ambientais e sociais.

Ingwersen (1996) enfatizou a importância de incluir o sistema de recuperação de informações como parte do modelo abrangente de comportamento de busca de informações, entendendo a recorrência deste processo e a importância de sua relação com a realidade e o ambiente do usuário. Isto é importante, pois seu modelo considera que existem transformações cognitivas que ocorrem entre o cotidiano dos usuários e a pesquisa. Desta forma, o autor aponta para a necessidade de comunicação efetiva entre todas as entidades envolvidas no sistema de RI, em especial, aquelas que consideram a realidade do usuário.

Embora Saracevic (1996) tenha sugerido que o modelo de Ingwersen (1996) tivesse problemas para avaliação de uso, em especial, no que tange a análise de atualização a longo prazo, o modelo de Ingwersen (1996) ainda é considerado uma contribuição importante para a compreensão do comportamento de busca de informação em sistemas de recuperação da informação. Wilson (2006) argumentou que, no entanto, uma possível fraqueza remanescente do modelo é que não analisa explicitamente o comportamento de informação além da recuperação de informação.

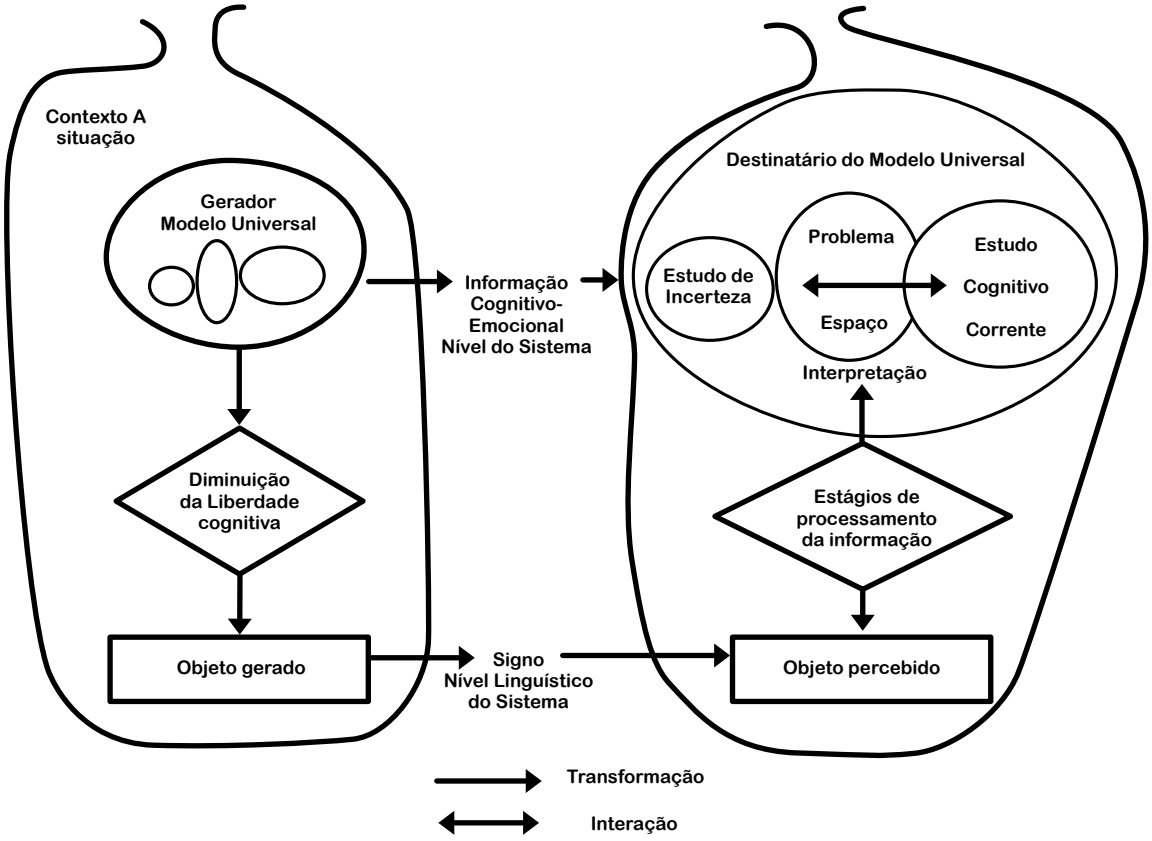
Isto significa que o modelo de Ingwersen (1996) é "incompleto" para o estudo da informação dentro deste SRI, pois não considera como os usuários chegam ao ponto de fazer a pesquisa, nem como suas estruturas cognitivas são afetadas pelos processos de decidir como e quando se mover em direção à pesquisa de informações. Desta forma, de maneira geral este modelo não se preocupa com o contexto do usuário. Estas questões podem ser discutidas em termos do ambiente social ou organizacional, mas não são explicitamente abordadas no modelo de Ingwersen (1996), de acordo com Wilson (2006).

O modelo proposto por Ingwersen (1999) na Figura 4 é uma versão atualizada do modelo anteriormente proposto pelo mesmo autor em 1996. Esse modelo leva em consideração dois aspectos cruciais: as tarefas de trabalho e a percepção cognitiva do usuário. Como parte do contexto situacional do usuário, ou seja, a combinação do ambiente físico em que ele se encontra, as tarefas ou atividades que está realizando, as restrições e recursos disponíveis, as características do sistema ou dispositivo sendo utilizado e as suas necessidades e expectativas, estão incluídas as tarefas de trabalho impostas pelo ambiente social-organizacional. Essas tarefas são percebidas pelo usuário por meio de seu estado cognitivo, como interesse, problema ou tarefa a ser realizada.

A percepção, conforme destacado por Ingwersen (1999), pode ser considerada como um componente dominante da situação problemática, sendo a causa do surgimento da necessidade de informação. Em um sentido cognitivo, a percepção do usuário durante uma tarefa de trabalho tende a ser mais estável durante a sessão de recuperação da informação do que a necessidade de informação dinâmica correspondente, ou seja, aquela que se altera ao longo do processo de recuperação da informação.

A percepção da tarefa de trabalho é, portanto, apropriada para ser utilizada, pois pode fornecer o contexto necessário para o sistema recuperar informações relevantes, ou seja,

Figura 4 – Modelo atualizado de SRI de Ingwersen (1999)



Fonte: Ingwersen (1999 apud Araújo Júnior, 2005, p. 88)

informações úteis para aquele usuário ao realizar a tarefa de trabalho. Assim, estratégias de busca podem ser refinadas, melhorando o processo de recuperação da informação.

Neste sentido, Ingwersen (1999) apresenta um modelo que se preocupa com sua atualização, no sentido de satisfazer as necessidades informacionais adaptadas aos usuários ao longo do tempo, dependendo de sua atualização informacional. De fato, esta é uma preocupação observada em SRIs, digitais ou não, em que a atualização cognitiva do usuário pode impactar no entendimento de relevância da informação recuperada.

Choo (2003, p. 76), a partir da organização das ideias de Saracevic (1975), apresentou outro modelo de fluxo de trabalho para SRIs, cabível também para SRIDs:

1. O usuário tem um problema a resolver (características do usuário, declaração do problema);
2. O usuário procura resolver o problema formulando uma pergunta e iniciando uma interação com um sistema de informação (declaração da pergunta, características da pergunta);
3. Interação de pre-investigação com um pesquisador intermediário, humano ou computador (características do pesquisador, análise da pergunta);
4. Formulação de uma busca (estratégia de busca, características da busca);
5. Atividade de busca e interações (busca);

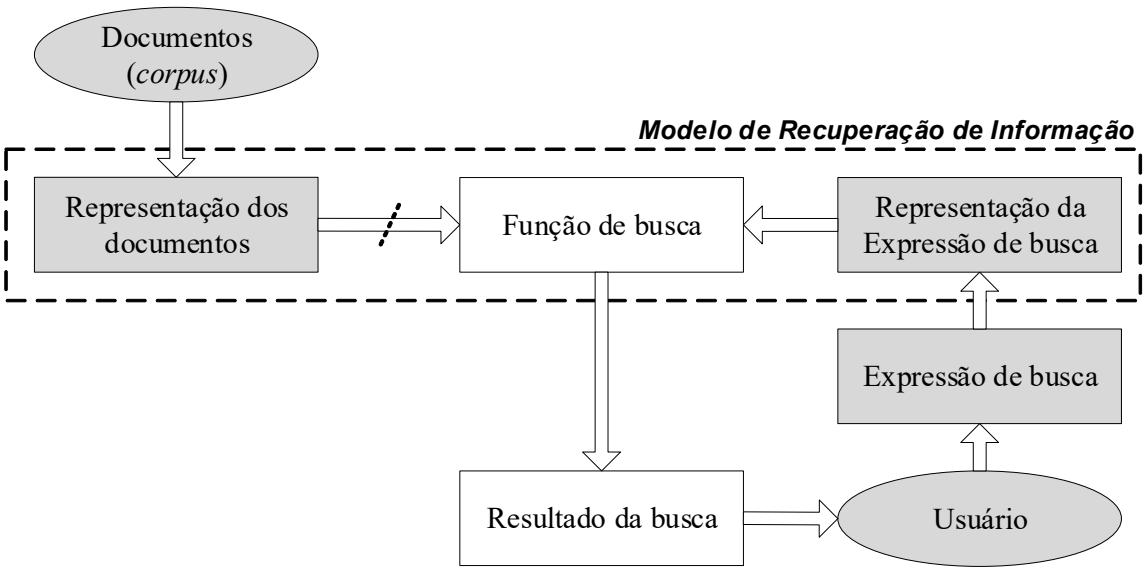
6. Entrega das respostas ao usuário (itens armazenados, formatos despachados);
7. Avaliação das respostas pelo usuário (relevância, utilidade).

Sobre a qualidade deste modelo, Choo (2003) indicou que:

> A análise dos dados empíricos mostrou que "os modelos sugeridos foram aprovados, ou seja, os elementos sugeridos pelos modelos apresentaram uma significativa relação com os resultados armazenados". Confirmou-se, por exemplo, que o contexto de uma pergunta é importante, inclusive os antecedentes que levam à formulação da pergunta e a intenção por trás do uso da informação a ser armazenada. Diferentes tipos de perguntas - classificadas de acordo com sua clareza, especificidade, complexidade, etc. - terão provavelmente diferentes níveis de desempenho no que diz respeito ao armazenamento da informação. Ciclos de busca tendem a melhorar os resultados, já que os resultados intermediários podem ser revistos e as estratégias de busca refinadas de acordo com eles (Choo, 2003, p. 76).

Isso entra em consonância com o modelo apresentado por Ingwersen (1999) sobre a operação de SIs. Em um outro modelo de operação, conforme descrito por Ferneda (2012), o processo de recuperação de informações em um sistema de recuperação de informações digital ocorre quando o usuário interage com o sistema por meio de uma expressão de busca para obter acesso às informações desejadas. O sistema, por sua vez, utiliza uma função de busca como mediação para representar os documentos presentes no estoque informacional e, em caso de correspondência com a expressão de busca, apresenta-os como resultado para o usuário. Na Figura 5 é apresentado este processo.

Figura 5 – Fluxo de um sistema de recuperação de informações digital



Fonte: Ferneda (2012, p. 14)

Em um SRI, é responsabilidade do usuário formular uma expressão de busca que descreva adequadamente sua necessidade de informação (Araújo Júnior, 2005). A busca de informação neste sistema pode ser realizada através de linguagem natural ou palavras-chave, com o objetivo de recuperar um conjunto de documentos relevantes para o usuário.

Após apresentação dos termos de busca, o sistema os relaciona com o seu estoque através de representações. A representação de documentos em sistemas de recuperação de informação digital busca descrever ou identificar cada documento do acervo através de seu conteúdo (Ferneda, 2003). Segundo Novellino (1996):

> A principal característica do processo de representação da informação é a substituição de uma entidade linguística longa e complexa - o texto do documento - por sua descrição abreviada. O uso de tal sumarização não é apenas uma consequência de restrições práticas quanto ao volume de material a ser armazenado e recuperado. Essa sumarização é desejável pois sua função é demonstrar a essência do documento. Ela funciona então como um artifício para enfatizar o que é essencial no documento considerando sua recuperação, sendo a solução ideal para organização e uso da informação(Novellino, 1996, p. 38).

Esta representação é, geralmente, realizada por meio do processo de indexação (Ferneda, 2003). A indexação é um processo essencial de representação realizado em um SRI, no qual o indexador, seja ele humano ou um sistema digital, examina o documento ou arquivo e distingue a informação relevante daquela periférica, a fim de representá-lo de maneira adequada para posterior recuperação (Lima; Campos, 2022). Segundo Lima e Campos (2022), a indexação é vista como "[. . .] a representação do conteúdo dos documentos por meio de símbolos especiais, quer retirados do texto original, quer escolhidos numa linguagem de informação ou de indexação". Ainda sobre indexação, Lima e Campos (2022, p. 2) indicaram que:

> [. . .] esse processo é realizado em duas etapas: a primeira é a da análise do documento para identificação de seu conteúdo informacional; a segunda, a de tradução dos conceitos nos termos de uma linguagem de indexação, utilizando-se os sistemas de organização do conhecimento, do tipo tesauros e sistema de classificação bibliográfico (Lima; Campos, 2022, p. 2).

Esta representação permite identificar o documento e definir seus pontos de acesso para a busca, podendo também ser utilizada como seu substituto. Este processo é descrito por Ferneda (2003) como parte importante do processo de recuperação de informação, em que a representação dos documentos é essencial para a mediação entre a expressão de busca do usuário e os resultados apresentados pelo sistema.

A partir de indexação, documentos e informações são recuperados para o usuário, que avalia sua relevância e o potencial de utilização mediante sua necessidade informacional.

Ocorre que a relevância de um documento pode ser afetada pelo contexto em que o termo aparece ou pela idade do documento, por exemplo (Ferneda, 2003) . Isso representa um desafio para os processos de RI em um sistema de recuperação de informação no que tange sua representação.

Os modelos apresentados não esgotam as possibilidades de proposição de estratégias para o projeto de sistemas de recuperação de informações clássico ou digital, mas dão uma boa ideia das possibilidades de sua operacionalização. Podemos, a partir destes modelos, afirmar que a estratégia de busca em sistemas de recuperação de informação é baseada nas necessidades de informação dos usuários, que possuem um estado anômalo de conhecimento (Belkin, 1980).

Os resultados de um sistema de busca e recuperação podem envolver não apenas documentos potencialmente relevantes, mas também a avaliação e o julgamento da informação pelos usuários, cujo conhecimento pode ser alterado durante a interação com o sistema (Araújo Júnior, 2005). Observa-se uma grande preocupação com a questão do alinhamento das informações documentadas no repositório e as necessidades dos usuários. Esse entendimento de SRIs enfatiza a importância da interação entre o usuário e o sistema, bem como, a necessidade de adaptar os mecanismos de recuperação de informação as necessidades e expectativas dos usuários.

É importante destacar, no entanto, que SRIs possuem dificuldades relativas às suas especificidades e estratégias apresentadas, conforme apontado por Ferneda (2003), Araújo Júnior (2005), Saracevic (1996), Ingwersen e Jäverlin (2004), entre outros autores. Dessa forma, os modelos apresentados, mesmo sendo considerados para o projeto de SRIs podem possuir dificuldades em alguns aspectos, em especial, sistemas de recuperação da informação digital que utilizam de algoritmos computacionais para o processamento de informações e operação do sistema, alguns pontos se elevam como mais preocupantes, tais como os descritos a seguir.

Ingwersen e Jäverlin (2004) apresentam algumas problemáticas relacionadas à recuperação de informação. São elas:

> [...] Carência do usuário e das tarefas. [...] Carência de interação e requisições dinâmicas. [...] Carência de variabilidade tática. [...] Carência da expectativa de incerteza. [...] Carência de relevância orientada ao usuário. [...] Carência de variedade dos bancos de dados. [...] Premissa de independência documental e negligência de sobreposição documental. [...] Insuficiência de recordação para precisão. [...] Excesso de média nas buscas. [...]. Apenas recuperação documental (Ingwersen; Jäverlin, 2004, p. 7–9, tradução nossa)[11].

[11] *Lack of users and tasks. [...] Lack of interaction and dynamic reaquests. [...] Lack of tactical variability. [...] Lack of uncertainty. [...] Lack of user-oriented relevance. [...] Lack of variety in collections. [Assuming document Independence and neglectin overlaping. [...] Insufficiency of recall and precision. [...] Heavy averaging. [...] Just documental retrieval* (Ingwersen; Jäverlin, 2004, p. 7–9).

Em complemento, de acordo com Vickery e Vickery (2004), os problemas fundamentais da recuperação da informação em ambientes digitais estão relacionados com natureza das mensagens armazenadas, como registros no sistema e a forma como essas mensagens se relacionam com as consultas que serão realizadas pelos usuários. A efetividade do sistema de recuperação de informação dependerá da capacidade do sistema de compreender as necessidades de informação dos usuários e de apresentar resultados relevantes e precisos. Em outras palavras, se o sistema não for capaz de capturar as necessidades dos usuários e fornecer resultados precisos, a recuperação das informações será comprometida. Para Vickery e Vickery (2004):

> [...] os problemas centrais da recuperação da informação surgem da natureza das mensagens armazenadas como registros no sistema e da relação dessas mensagens com as consultas que provavelmente serão feitas ao sistema. (Vickery; Vickery, 2004, p. 118, tradução nossa)[12].

Também, sobre isto, Vickery e Vickery (2004) expuseram que,

> [...] nos sistemas de recuperação de informações, os "valores" (por exemplo, os textos) armazenados são de variedade ilimitada, os termos de busca apresentados nas consultas são imprevisíveis, e a relação entre as mensagens armazenadas e as consultas processadas é frequentemente ambígua. (Vickery; Vickery, 2004, p. 118, tradução nossa)[13].

Ferneda (2003) complementa estas afirmações, indicando que a maior dificuldade enfrentada pelos usuários é saber quais termos utilizar para encontrar documentos que atendam às suas necessidades específicas.

É importante lembrar que a recuperação de informação pode ser um processo subjetivo, uma vez que diferentes usuários podem ter diferentes maneiras de descrever a mesma necessidade de informação. Desta forma, Ferneda (2003) destacou que é fundamental que o sistema de informação permita uma certa flexibilidade na formulação de consultas, e ofereça recursos que possam ajudar os usuários a refinar suas buscas, como sugestões de termos relacionados ou filtros de resultados.

A natureza das mensagens armazenadas no sistema e a relação dessas mensagens com as consultas dos usuários apresentam dificuldades, como a variedade ilimitada dos valores armazenados, a imprevisibilidade dos termos de pesquisa e a ambiguidade na relação entre as mensagens e as consultas. Desta forma, o projeto de sistemas de recuperação da informação clássicos e digitais deve ter uma atenção depositada no usuário e sua relação

[12] *The central problems of information retrieval arise from the nature of the messages stored as records in the system and the relation of these messages to the queries likely to be put to the system* (Vickery; Vickery, 2004, p. 118)

[13] *In information retieval systems the "values" (for example, the texts) stored are of unlimited variety, the search terms presented in queries are unpredictable, and the relationship between messages stored and queries processed are often ambiguous* (Vickery; Vickery, 2004, p. 118)

com o sistema para aumentar as chances de vinculação entre busca e recuperação de informação.

Essas questões são agravadas e ampliadas quando se trata do projeto de sistemas de informação digitais. Isto ocorre porque a seleção dos documentos que devem ser armazenados no sistema é virtualmente inexistente, o que enfraquece a conexão entre as necessidades do usuário e o estoque de informações disponível.

Quando as características de um usuário específico não são consideradas no projeto de sistemas de informação digitais, há uma desconexão entre os registros do sistema e as necessidades efetivas dos usuários. Isto agrava a possibilidade de vinculação entre busca e recuperação da informação, comprometendo a relevância da informação recuperada. O papel do usuário como sujeito informacional em regimes de informação mais amplos, também tem um impacto na relação entre documentos e informação, conforme apontado por Rabelo (2017).

De acordo com Gonzalez de Gómez (2012), é possível que mais de um regime de informação seja formado a partir da mesma combinação de tecnologia, serviços e conteúdos informacionais. Regimes informacionais dizem respeito às configurações possíveis entre perfis de indivíduos e sua relação com a informação. Em ambientes que utilizam sistemas de informação, um novo regime informacional pode influenciar a definição dos sujeitos, instituições, autoridades informacionais e os recursos preferenciais de informação.

Conforme explicado por Rabelo (2017), estes regimes são mais fluidos e resilientes, permitindo que o usuário tenha papéis de fonte e de receptor da informação, tornando-se um sujeito informacional. Nesses sistemas os usuários também têm um papel de produtor, gerando dados e informações que alimentam os novos regimes informacionais (Gonzalez de Gómez, 2012; Rabelo, 2017). Devido à quantidade, velocidade e variedade dessas produções, algoritmos de processamento são necessários e a dinâmica dos sistemas de acesso simples é alterada, tornando os usuários ativos na recepção de novas produções.

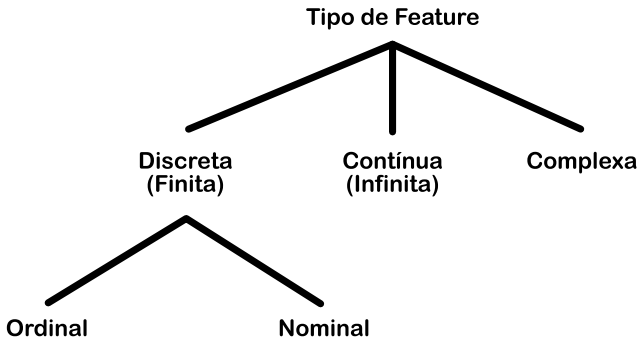
2.6 Features

A *feature* é um atributo descritivo de algo relevante para algoritmo matemático utilizado. Elas podem também ser denominadas atributos, propriedades ou características. Podemos sintetizar o conceito de *feature* como: Atributo individual de uma ou mais características ou comportamentos relevantes que pode ser representado numericamente [14]

[14] Essa definição amplia a definição tradicional para incorporar as *features* obtidas por meio da união de vários atributos em uma única *feature*. Exemplo: Atividade de clustering de imóveis disponíveis em um banco de dados com o objetivo de agrupá-los pelo critério de conforto habitacional para o morador; não é necessário tratar como *feature* independente o número de quartos e o número de suítes; poderia contar cada quarto com o valor 1 e cada suíte com o valor 3; somando tudo e resultando em uma única *feature*. Esse procedimento auxilia na computação do algoritmo porque diminui a dimensionalidade do problema e facilita no scaling das *features*. A definição tradicional é "uma *feature* é a especificação de

De acordo com a característica do valor atribuído, segundo Liu e Motoda (1998, p. 3), as features podem ser classificadas em três grandes categorias: discreta (finito), contínua (infinita) e complexa; a categoria discreta é dividida em outras duas: ordinal e nominal (veja na Figura 6).

Figura 6 – Classificação das *features*



Fonte: Liu e Motoda (1998, tradução nossa)

As *features* contínuas geralmente possuem valores situados no domínio dos números Reais ($n \in \mathbb{R}$) , como por exemplo a medição de uma temperatura em um determinado momento. *Features* discretas possuem um número limitado de valores possíveis. *Features* discretas ordinais possuem valores que representam a ordem de ocorrência de eventos, por exemplo, o posicionamento de competidores ao final de um torneio, cada competidor terá um valor único; já as *features* nominais possuem valores que representam nomes, como nomes das cores por exemplo, $vermelho \equiv 1$, $verde \equiv 2$, $azul \equiv 3$. (Liu; Motoda, 1998, p. 3).

Neste trabalho utilizaremos fundamentalmente *features* discretas, pois ao manipularmos o conteúdo dos textos (as palavras) em níveis mais profundos (termos e conceitos), toda a representação deste conteúdo será feita por valores como 1, 2 ou 3 (valores discretos) durante o processamento.

um atributo com seu valor" (A feature is the specification of an attribute and its value). – (Kohavi; Provost, 1998), ou seja, é uma variável de entrada aplicada a um modelo de machine learning.

3 *Clustering* de documentos

O *Clustering* tem como objetivo detectar grupos a partir de dados não rotulados. A diferença entre *clustering* e classification está no tipo de dado utilizado. A *classification* é um processo de classificação supervisionada, portanto utiliza dados rotulados durante a aprendizagem; enquanto o *clustering* é um processo de classificação não supervisionada, utilizado dados não rotulados (Kononenko; Kukar, 2007, p. 321).

Segundo Kononenko e Kukar (2007, p. 322), um algoritmo útil de *clustering* deve satisfazer oito critérios: (1) escalabilidade; (2) capacidade de lidar com diferentes tipos de atributos; (3) descobrir agrupamentos com formas arbitrárias; (4) necessidade de conhecimento mínimo de domínio para conseguir determinar os parâmetros de entrada; (5) capacidade de lidar com ruido, dados incompletos e *outliers*; (6) imunidade a ordenação dos registros de entrada; (7) alta dimensionalidade; e (8) interpretabilidade e usabilidade. Explicando de uma maneira mais detalhada cada um destes oito critérios temos:

1. Escalabilidade: capacidade de processar conjuntos de dados cada vez maiores com um aumento da demanda de processamento seguindo proporções lineares, ou seja, um algoritmo bom possui uma complexidade próximo de logarítmica ou, na pior hipótese, linear; infelizmente na prática isso nem sempre é possível;

2. A capacidade de lidar com diferentes tipos de atributos: significa que o algoritmo consegue trabalhar com atributos situados em faixas numéricas distintas;

3. Descobrir agrupamentos com formas arbitrárias: Neste caso, forma foi a tradução encontrada para o termo shape. Uma técnica bastante utilizada para análise de dados é a representação gráfica espacial dos dados de alguma maneira, normalmente 2D ou 3D; quando os dados são representados desta maneira, o algoritmo de *clustering* consegue formar desenhos que representam os agrupamentos; desenhos mais complexos requerem algoritmos mais refinados se comparados a desenhos mais simples;

4. Necessidade de conhecimento mínimo de domínio para conseguir determinar os parâmetros de entrada: O ideal é que o sistema não necessite de conhecimento específico sobre um determinado domínio para que consiga fazer uma boa escolha dos parâmetros que serão utilizados na entrada do algoritmo;

5. Capacidade de lidar com ruido, dados incompletos e *outliers*: *outlier* é aquele dado que difere significativamente de todas as outras observações, sendo originário de algum ruido nas medições ou mesmo uma ocorrência excepcional que não deve ser considerado;

6. Imunidade a ordenação dos registros de entrada: Esta imunidade à ordenação de entrada é importante pois dependendo da maneira como os dados são obtidos, eles serão processados assim que extraídos, não havendo qualquer tipo de ordenação do conjunto antes do processamento. No caso específico do *clustering* de conteúdo textual, significa que ao terminar a análise de todo o *corpus* textual o algoritmo deve produzir a mesma saída, independentemente da ordem em que os textos sofreram a extração de seus dados;

7. Alta dimensionalidade: é a capacidade que o algoritmo tem em lidar com um número elevado de *features*. A definição de *feature* utilizada é a de Cristopher Bishop (2006, tradução nossa): "Uma *feature* é uma propriedade mensurável individual ou característica de um fenômeno"[1];

8. Interpretabilidade e usabilidade: se refere a capacidade que o algoritmo tem em produzir *clusters* que sejam empiricamente interpretáveis, ou seja, se a amostra for devidamente compensada (*scaled*) e representada graficamente será possível visualizar os grupos que o algoritmo produziu.

3.1 Representação de documentos textuais

Representar numericamente os documentos textuais é o primeiro passo para tornar o texto processável por computadores. Podemos criar macro representações a partir da síntese de uma ou mais representações. Quando manipulamos textos, sem qualquer metadado de apoio, o único material disponível é o conjunto de palavras que naquele contexto remetem a uma série de conceitos, e serão estes conceitos que utilizaremos durante o processo de *clustering*.

Uma representação mais tradicional dos textos para processamento automático é aquela que utiliza o modelo *bag-of-words*; neste modelo todas as palavras de um texto são extraídas, é feito um pré processamento para eliminar as palavras de pouca relevância (este processo é chamado de remoção de *stopwords*), as palavras restantes são transformadas geralmente por um processo de stemming e o resultado deste é utilizado para compor o conjunto de palavras pertencentes a um determinado texto. Tradicionalmente este conjunto de palavras é utilizado para determinar cada uma das dimensões de um vetor, das quais as magnitudes são obtidas por meio de cálculos, geralmente, *Term frequency – Inverse document frequency* (TF–IDF); desta forma, cada documento é representado por meio de um vetor. Esta forma de representação foi empregada por Gerard Salton (1964) em diversos experimentos envolvendo o sistema SMART desenvolvido por ele.

O processo de associar textos a um ou mais rótulos de classes é denominado classificação. Os algoritmos de classificação de textos dividem-se em dois grandes grupos:

[1] *feature is an individual measurable property or characteristic of a phenomenon* (Bishop, 2006)

supervisionados e não supervisionados. Os algoritmos supervisionados são aqueles que necessitam de treinamento mediante o uso de um conjunto de dados etiquetados previamente que o algoritmo utilizará para aprender. Os algoritmos não supervisionados dispensam o uso do conjunto de dados etiquetados, operando a partir de regras e parâmetros pré-definidos aplicados diretamente ao próprio conjunto de entrada. O processo de *clustering* é feito utilizando um algoritmo não supervisionado.

Existem duas abordagens principais para o processo de *clustering*: *clustering* por particionamento e *clustering* aglomerativo. Enquanto a primeira abordagem está ancorada no princípio de que todas as entidades pertencem a um único grupo, que será iterativamente testado e dividido em grupos menores; a segunda abordagem assume inicialmente que todas as entidades não pertencem a qualquer grupo, e iterativamente essas entidades serão associadas aos grupos existentes ou darão origem a novos grupos; o *clustering* aglomerativo é hierárquico por natureza. Além destas duas abordagens principais existem outras, como é o caso do algoritmo *K–Means* que não é considerado pertencente a nenhuma das duas porque ele inicia já com o número de *clusters* predefinidos e este número não sofrerá alterações.

3.2 Medidas de similaridade

Medir a similaridade entre *features* é um pressuposto necessário para que qualquer algoritmo de classificação ou agrupamento consiga operar. Existem diversas fórmulas para realizar esta medição de similaridade (também chamado de coeficiente de similaridade):

A medição da similaridade nem sempre é feita de maneira binária, medindo uma *feature* de análise em relação a todas as outras *features* uma por vez. Existem medições que consideram pontos adjacentes às duas *features* em medição, estes pontos são denominados contexto. Uma métrica contextual é a *mutual neighbor distance* (MND), proposta por Gowda e Krishna em 1977 (Jain; Murty; Flynn, 1999). Existem coeficientes de similaridade probabilísticos ou apenas descritivos (Sneath; Sokal, 1973, p. 118).

A similaridade entre dois elementos é estimada usado uma quantificação de semelhança entre as *features* dos dois elementos, sintetizada no coeficiente de similaridade.

3.2.1 Fórmulas

Nas fórmulas apresentadas é assumido que as *features* das amostras estão dispostas em uma matriz de nxt dimensões. As fórmulas utilizam as letras j e k para representar duas *features* de mesma posição e amostras distintas:

matriz $n \times t$

$$\begin{bmatrix} x_{11} & \dots & x_{1t} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nt} \end{bmatrix} = \begin{cases} n = feature \\ t = amostra \end{cases}$$

$$n = rows = character$$

$$j, k = x_{ij} = \begin{cases} i = feature \\ j = amostra \end{cases}$$

Os valores das *features* presentes nesta matriz podem ser valores resultantes de medições direta ou valores transformados por meio de processos de padronização

Existem demonstrações algébricas de que a maioria dos teoremas aplicáveis ao espaço tridimensional convencional podem ser estendidos para n dimensões no Hiperespaço Euclidiano, desta maneira é possível propor maneiras de computar a distância entre duas amostras considerando cada *feature* em uma dimensão própria do hiperespaço. Esta é a base do modelo vetorial de recuperação de informação no qual cada termo pertence a uma dimensão e a distância entre os vetores é calculada por meio do *dot product* (produto escalar) das matrizes termo a termo.

3.2.2 Medidas de Dissimilaridade

As medidas de dissimilaridade são "funções que convertem coeficientes de similaridade em medidas de dissimilaridade, como complementos do valor máximo de um coeficiente de associação ou o arco cosseno de um coeficiente de correlação"[2] . Não é necessário que o espaço seja euclidiano, mas ele precisa ter sua topologia determinada por uma função métrica, para isso as medidas de dissimilaridade devem satisfazer quatro axiomas (Sneath; Sokal, 1973, p. 120):

1. $\varphi(a, b) \geq 0$ e $\varphi(a, a) = \varphi(b, b) = 0$
2. $\varphi(a, b) = \varphi(b, a)$
3. $\varphi(a, c) \leq \varphi(a, b) + \varphi(b, c)$
4. Se $a \neq b$ então $\varphi(a, b) > 0$

O primeiro axioma postula que itens idênticos são indistinguíveis enquanto os não idênticos podem ou não serem distinguidos por uma função de dissimilaridade.

O segundo axioma estabelece a relação de simetria entre a medida, sendo a similaridade entre a e b idêntica à similaridade entre b e a.

O terceiro axioma é o axioma da desigualdade triangular[3] : afirma que em um triângulo a soma do comprimento de dois lados é sempre maior que o comprimento do terceiro lado

[2] *Functions that convert similarity coefficients into measures of dissimilarity, such as complements from the maximum value of association coefficients or the arc cosine of the correlation coefficient* (Sneath; Sokal, 1973, p. 20)

[3] *Triangle inequality axiom*

O quarto axioma afirma que se a e b são diferentes, o coeficiente deve ser maior que zero. Não existe coeficiente negativo.

3.2.3 Coeficientes de distância

Coeficientes de associação e coeficientes de correlação podem ser relacionados a distância. Coeficientes de associação são funções pareadas que medem a semelhança entre pares de amostras considerando um conjunto composto por duas ou mais *features*[4](Sneath; Sokal, 1973, p. 129). A maneira mais comum de computar esse coeficiente é a partir de uma matriz binária (que utiliza apenas valores discretos 0 e 1).

Ao tratar todos os atributos como valores binários, a comparação entre os pares formados pela mesma *feature* presente em duas amostras distintas (amostra t_i e amostra t_j) resultará em quatro quantidades (Kononenko; Kukar, 2007, p. 325):

- a = número de vezes que ambas as *features* são 1;
- b = número de vezes em que $t_i = 1$ e $t_j = 0$;
- c = número de vezes em que $t_i = 0$ e $t_j = 1$;
- d = número de vezes em que ambas as *features* são 0.

Estabelecida estas definições, as duas fórmulas mais comuns para medida de dissimilaridade com *features* binárias são Equação 3.1 e Equação 3.2:

Coeficiente de *simple match*:

$$d_{ij} = \frac{a + d}{a + b + c + d} \tag{3.1}$$

Coeficiente de similaridade de Jaccard:

$$d_{ij} = \frac{a}{a + b + c} \tag{3.2}$$

Ambas as fórmulas estão contidas na fórmula generalizada de dissimilaridade:

$$d_{ij} = \frac{b + c}{\alpha a + b + c + \delta d} = \begin{cases} \alpha > 0 \\ \delta \geq 0 \end{cases} \tag{3.3}$$

$$d_{ij} = \frac{1}{a} \sum_{k=1}^{a} d_{ijk} \tag{3.4}$$

[4] Traduzido e adaptado de "*They are pair-functions that measure the agreement between pairs of OTU's over an array of two-state or multi-state*" (Sneath; Sokal, 1973, p. 129)

Exemplo, dada a seguinte matriz de documentos e *features*:

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Neste exemplo cada linha da matriz é a representação de um documento (neste caso, 3) e cada coluna de uma *feature* qualquer (neste caso, 5). Nesta matriz, a primeira linha representa o Doc_1, a segunda linha o Doc_2 e a terceira linha o Doc_3; as colunas são as *features* F_1,F_2,F_3,F_4,F_5. Ao transformarmos essa matriz em um quadro contendo "sim" para os valores unitários e "não" para os valores zero, obtemos

Quadro 1 – Exemplo para cálculo dos coeficientes de distância

	F_1	F_2	F_3	F_4	F_5
doc_1	sim	sim	não	sim	não
doc_2	não	sim	não	sim	sim
doc_3	sim	não	sim	não	não

Fonte: autor

A partir do Quadro 1, calculamos os coeficientes a, b, c e d entre cada documento. Isso é demonstrado no Quadro 2

Quadro 2 – Exemplo coeficientes a, b, c, d entre documentos Doc_1 x Doc_2

	a (ambos "sim")	***b*** ("sim" e "não")	***c*** ("não" e "sim")	***d*** (ambos "não")
$\boldsymbol{doc_1 \times doc_2}$	2	1	1	1
$\boldsymbol{doc_1 \times doc_3}$	1	2	1	1
$\boldsymbol{doc_2 \times doc_3}$	0	3	2	0

Fonte: autor

A partir dos coeficientes obtidos no Quadro 2, podemos aplicar as fórmulas descritas acima (Equação 3.1 e Equação 3.2) e calcular os coeficientes de distância, conforme demonstrado no Quadro 3 e no Quadro 4:

Quadro 3 – Exemplo: coeficiente de distância Simple Match

	$\boldsymbol{doc_1}$	$\boldsymbol{doc_1}$	$\boldsymbol{doc_3}$
$\boldsymbol{doc_1}$	0	$3/5$	$2/5$
$\boldsymbol{doc_2}$	$3/5$	0	$2/5$
$\boldsymbol{doc_3}$	$2/5$	$2/5$	0

Fonte: autor

Observe que neste exemplo a distância entre doc_2 e doc_3 segundo o coeficiente de Jaccard é 0, o que significa que ambos são completamente dissimilares; já o coeficiente

Quadro 4 – Exemplo: coeficiente de distância de Jaccard

	doc_1	doc_1	doc_3
doc_1	0	$^2/_4$	$^1/_4$
doc_2	$^2/_4$	0	0
doc_3	$^1/_4$	0	0

Fonte: autor

Simple Match indica que a dissimilaridade entre doc_2 e doc_3 possui a mesma magnitude do que aquela entre doc_1 e doc_3; em nosso caso, comparação terminológica, isso não seria adequado pois os documentos doc_1 e doc_3 possuem uma *feature* em comum, ao passo que os documentos doc_2 e doc_3 não possuem nenhuma *feature* em comum.

3.2.4 Discussão sobre o Coeficiente de Jaccard

O Coeficiente de Jaccard serve para aferir a similaridade entre amostras que possuam *features* em comum, ele não considera as *features* não presentes em ambas as amostras, tornando o mais adequado que o coeficiente *Simple Match* quando usado para aferição de similaridade entre textos.

Para ilustrar essa característica, segue um exemplo: dado dois textos T1, T2 e uma lista contendo N termos, podemos criar uma matriz binária com duas colunas, uma para cada texto; e com N linhas, uma para cada termo. Essa matriz conterá os valores 0 quando o texto não possuir o termo e 1 quando o termo estiver presente no texto

Quadro 5 – Exemplo: termo × documento

	T_1	T_2
$termo_1$	1	1
$termo_2$	0	1
$termo_2$	1	0
$termo_2$	1	1
...	0	0
$termo_n$	0	0

Fonte: autor

Empregando a convenção estabelecida (Kononenko; Kukar, 2007, p. 325), temos:

- $a = 2$ (número de vezes que ambos os textos possuem o termo)
- $b = 1$ (número de vezes que apenas o primeiro texto contém o termo)
- $c = 1$ (número de vezes que apenas o segundo texto contém o termo)
- $d = N - a - b - c$ (número de vezes que o termo não a aparece em nenhum dos textos)

Neste exemplo o Coeficiente de Jaccard entre T_1 e T_2 sempre será de $1/2$, independentemente de quantos termos sejam adicionados à matriz (desde que esses termos não estejam presente em ambos os textos). Isto é demonstrado na Equação 3.5.

$$d_{ij} = \frac{a}{a+b+c} = \frac{2}{2+1+1} = \frac{1}{2} = 0.5 \tag{3.5}$$

Compare o mesmo exemplo, com os dados do Quadro 5, aplicando o Coeficiente de *Simple Match*:

Para 4 termos ($N = 4$)

$$d_{ij} = \frac{a+d}{a+b+c+d} = \frac{2+0}{2+1+1+0} = \frac{2}{4} = 0.5 \tag{3.6}$$

Para 5 termos ($N = 5$)

$$d_{ij} = \frac{a+d}{a+b+c+d} = \frac{2+1}{2+1+1+1} = \frac{3}{4} = 0.6 \tag{3.7}$$

Para 10 termos ($N = 10$)

$$d_{ij} = \frac{a+d}{a+b+c+d} = \frac{2+7}{2+1+1+7} = \frac{9}{11} \approx 0.82 \tag{3.8}$$

Para 100 termos ($N = 100$)

$$d_{ij} = \frac{a+d}{a+b+c+d} = \frac{2+97}{2+1+1+97} = \frac{99}{101} \approx 0.98 \tag{3.9}$$

Ou seja: o coeficiente é alterado sempre que novos termos são adicionados à lista, mesmo que estes termos não estejam presentes em ambos os textos.

O exemplo acima demonstra como o *simple match* (Equação 3.6, Equação 3.7, Equação 3.8 e Equação 3.9) é influenciado pela quantidade total de termos que o sistema considera, ao passo que o Jaccard (Equação 3.5) não; observe que no primeiro caso, envolvendo apenas os 4 primeiros termos, ambos os coeficientes são idênticos.

3.3 Cluster: definição

O *Clustering* pode ser usado para encontrar grupos homogêneos representativos, procedimento denominado data reduction, ou encontrar agrupamentos naturais, ou mesmo encontrar exemplos não usuais, chamado *outlier detection*.

Existem diferenças entre *clustering* e *classification*: a classificação é um tipo de aprendizagem supervisionada onde as classes (grupos) são conhecidas previamente; enquanto o *clustering* é uma categoria de algoritmos de aprendizagem não supervisionados onde o número de grupos (*labels*) não é conhecido de antemão (Kononenko; Kukar, 2007, p. 322).

3.4 Identificação de *clusters*

A atividade de *clustering* pode ser analisada de maneira melhor se decomposta em cinco componentes (Jain; Murty; Flynn, 1999, p. 266). A partir destes cinco componentes elementares propostos, se delimitarmos o conteúdo exclusivamente ao tipo textual, adequando a terminologia empregada, teremos os seguintes componentes:

1. Representação textual;
2. Medição da similaridade;
3. Método de *Clustering*;
4. Representação do *Clustering* (abstração dos dados);
5. Validação do *Clustering* (validação da saída).

3.4.1 Representação textual

Representar o texto consiste em escolher um método capaz de, por meio de *features*, representar o conteúdo temático do texto de uma maneira ponderada numericamente. Tradicionalmente esta representação é feita utilizando o modelo espaço vetorial (Salton; Yang; Wong, 1975), onde o texto é transformando em um vetor multidimensional no qual cada dimensão com seu valor escalar corresponde a uma *feature* (palavra, conjunto de palavras, conceito, conjunto de conceitos); este valor escalar define o peso relativo que esta *feature* tem em relação ao conteúdo encontrado no texto funcionando como uma representação do grau de pertencimento que um texto específico tem em relação a *feature* em questão.

A *feature* mais comum usada para descrever o conteúdo do texto é a palavra, o que resulta em um modelo conhecido como *bag-of-words model*. Este modelo possui certas limitações, sendo as principais: 1) incapacidade em detectar e agrupar sinônimos, uma vez que a correspondência das palavras é feita de maneira ortográfica; 2) impossibilidade de lidar com palavras polissêmicas, pois o modelo descarta o contexto das palavras tornando a resolução da ambiguidade muito difícil, talvez até inviável; 3) o modelo ignora a conexão entre as palavras, assumindo que a ocorrência delas se dê de maneira independente entre si, o que sob o ponto de vista linguístico não é possível.

3.4.2 Medição de similaridade

Após definir uma maneira para representar o texto é necessário medir a similaridade entre dois textos. O objetivo desta medição é determinar, geralmente sob uma ótica temática, o quão similar ou diferente dois textos são entre si. Esta necessidade de extrair e representar os temas desenvolvidos em cada texto dialoga diretamente com a representação textual adotada e determina o critério utilizado para o *clustering*.

3.4.3 Métodos de *clustering*

Os métodos de *clustering* podem ser categorizados em três grupos dualísticos (Hartigan, 1975; Jain; Murty; Flynn, 1999):

1. Aglomerativo versus divisivo;
2. Hierárquico versus particional;
3. Hard versus Fuzzy.

Aglomerativo versus divisivo

Este aspecto está relacionado com o modo de operação do algoritmo de *clustering*. No modo aglomerativo, o algoritmo no início assume que cada pattern pertence a seu próprio cluster; então, de maneira repetida, os clusters são avaliados entre si quanto à similaridade, e caso sejam similares são unidos formando um cluster maior; então estes clusters maiores são avaliados entre si quanto à similaridade e este processo é repetido até que um critério de parada seja atingido. No modo divisivo, o algoritmo assume incialmente que todos os patterns pertencem a um único cluster, que então é dividido dando origem a dois outros clusters, que por sua vez podem ou não serem divididos novamente, e este processo é repetido até que o critério de parada seja atingido.

Hierárquico versus particional

Métodos hierárquicos produzem partições aninhadas, ou seja, todas as partições estão aninhadas de alguma maneira e compõe partições maiores; já os métodos particionais produzem partições independentes umas das outras.

Hard versus Fuzzy

É a maneira pela qual o algoritmo atribui um pattern a um determinado cluster. *Hard clustering* significa que um pattern pode pertencer exclusivamente a um único cluster; já o *Fuzzy clustering* atribui a cada pattern um índice de pertencimento a cada um dos clusters existentes, por isso neste método os patterns podem pertencer, e possivelmente pertencerão, a diversos clusters com graus de pertencimento distintos.

3.4.4 Representação do Clustering (abstração dos dados)

A abstração dos dados consiste em "extrair uma representação simples e compacta do *data set*" (Jain; Murty; Flynn, 1999, p. 267). Neste caso, o conceito de simplicidade é empregado sob duas óticas: análise automática ou humana. Observe que essas duas abordagens são costumeiramente antagônicas entre si.

Uma representação simples para análise automática, é uma otimização que busca reduzir o número de *features*, reduzindo assim a dimensionalidade do problema e com

isso diminuindo a demanda por recursos computacionais tanto em armazenamento quanto em processamento. Um exemplo de otimização é aquele que ocorre quando é utilizado centroids durante o cálculo inicial explorativo dos dados.

Uma simplificação sob uma perspectiva humana, consiste em obter uma representação mais intuitiva dos dados facilitando a compreensão das pessoas envolvidas na manipulação destes dados; geralmente essa preocupação é mais evidente em processos semiautomáticos ou mesmo manuais auxiliados por ferramentas de *software*.

3.4.5 Validação do Clustering

Todo algoritmo de *clustering* estabelece grupos a partir dos dados de entrada, mesmo que esses dados não sejam similares o suficiente para isso; em outras palavras, todo algoritmo de *clustering* produzirá grupos, mesmo que o conjunto dos dados de entrada não os tenha. Essa característica dos algoritmos remete ao seguinte problema: como saber se o conjunto de dados possui uma estrutura de grupos, sem explicitamente identificar tais grupos; esta característica é denominada *clustering* tendency.

Mesmo quando o conjunto de dados apresenta uma forte *clustering tendency*, vários algoritmos necessitam de parametrizações diretamente relacionadas ao número de agrupamentos que o algoritmo espera encontrar, e esta parametrização é definida antes da execução dos algoritmos, ou seja, antes do processo de *clustering*. Uma abordagem tradicional para solucionar este problema circular é utilizar alguma heurística qualquer para determinar os parâmetros iniciais, executar o algoritmo e mensurar a qualidade dos agrupamentos obtidos; esse procedimento é denominado *clustering validation* (Theodoridis; Koutroumbas, 2008, p. 863–864). Validar os *clusters* obtidos consiste em mensurar o quão representativos eles são.

Em síntese, a validação do *clustering* pode ser abordada de duas maneiras distintas e complementares: (1) estudos de *cluster tendency* e (2) análise de *cluster validity*. No primeiro caso, o estudo de *cluster tendency* é feito nos dados de entrada com o objetivo de detectar se existem *clusters* a serem descobertos ou não; isso se justifica porque os algoritmos de *cluster*, quando executados, sempre produzirão *clusters* independentemente destes *clusters* existirem ou não nos dados de entrada, ou seja, é possível produzir *clusters* inexistentes (dados dissimilares agrupados erroneamente dentro de um *cluster*). A segunda abordagem é a análise de *cluster validity*, feita nos dados de saída produzidos pelo algoritmo; esta análise tem como objetivo avaliar a qualidade dos agrupamentos obtidos.

Dentro da análise de *cluster validity*, existem três tipos de estudos de validação (Jain; Murty; Flynn, 1999, p. 268): (1) Externa: compara a estrutura recuperada com uma estrutura predefinida; (2) interna: tenta determinar se a estrutura é intrinsecamente apropriada para os dados; (3) relativa: compara duas estruturas e mede o seu mérito relativo.

Algoritmos fortemente dependentes de parametrização, como é o caso do *K–Means*, demandam pelo menos análise de *cluster validity* nos resultados obtidos porque suaves mudanças em alguns parâmetros produzem resultados bem distintos, por exemplo, o parâmetro k do algoritmo *K–Means* que estabelece quantos grupos serão produzidos.

4 Definindo um método de clustering baseado em conceito

Neste trabalho propomos uma forma de indexação automática de documentos textuais por meio de agrupamento conceitual dos textos. Utilizando técnicas de *clustering* baseadas em conceitos, obtemos grupos etiquetados de documentos conceitualmente próximos bem como supergrupos de termos ambíguos compartilhados entre estes grupos de documentos conceitualmente próximos.

4.1 Representação de documentos por meio de conceitos

O primeiro desafio enfrentado é obter uma representação pertinente do conteúdo presente nos documentos sem utilizar extensas bases de conhecimento extratextuais; nesta proposta os termos são extraídos e processados automaticamente, sendo transformados em **unidades terminológicas**, portanto um documento será representado por um conjunto de unidades terminológicas. Uma vez que os documentos estiverem representados de uma maneira uniforme, podemos medir a similaridade entre eles e utilizar um algoritmo de *clustering* para agrupá-los em *clusters* conceituais. Estes agrupamentos inicialmente não possuem denominação, esta denominação será obtida por meio de um processamento feito com os termos extraídos de cada documento constituinte daquele grupo, obtendo-se os termos representativos que serão atribuídos como rótulo daquele agrupamento, sendo cada um destes rótulos compostos por vários termos relacionados. Esses termos que compõem os rótulos podem se repetir entre os grupos, permitindo que eles também sejam agrupados, formando um supergrupo de termos que participam da designação de mais de um grupo conceitual; quanto mais ambíguo for o termo isolado maior será o número de grupos que compõe aquele supergrupo.

O primeiro processo ao qual todos os documentos serão submetidos é o de extração de termos, que consiste em obter todas as palavras do texto, eliminando àquelas de uso comum por meio de um dicionário negativo (*stopwords*); as demais serão submetidas a um tratamento ortográfico denominado stemming para que as palavras flexionadas sejam tratadas como idênticas; uma vez feito isso, todas essas variações das palavras serão agrupadas como uma única unidade terminológica. Essas unidades terminológicas serão mensuradas em função do número de ocorrências internas à cada documento (medida de *term frequency*) e apenas aquelas acima de um determinado limiar serão consideradas no próximo passo, que consiste na análise para estabelecer as regras de associação entre elas dentro de cada documento, aquelas mais fortemente associadas serão tratadas como uma unidade conceitual; neste passo um mesmo documento possivelmente será representado por algumas unidades conceituais. Observe que não são os termos designativos do grupo que necessariamente participaram como critério de agrupamento.

Devido à dificuldade que existe em sintetizar os termos em um único termo representativo do conceito tratado, por causa da flexibilidade do idioma escrito e da complexa relação entre termo e conceito, optamos por representar os conceitos por meio de termos agrupados. Por exemplo, um termo ambíguo como "manga" que pode significar um tipo de fruta ou uma parte de uma roupa, será representado em dois grupos conceituais distintos como: "moda, roupa, manga" e "alimento, fruta, manga", e possivelmente também aparecerá em um supergrupo "manga", composto pelos dois grupos citados.

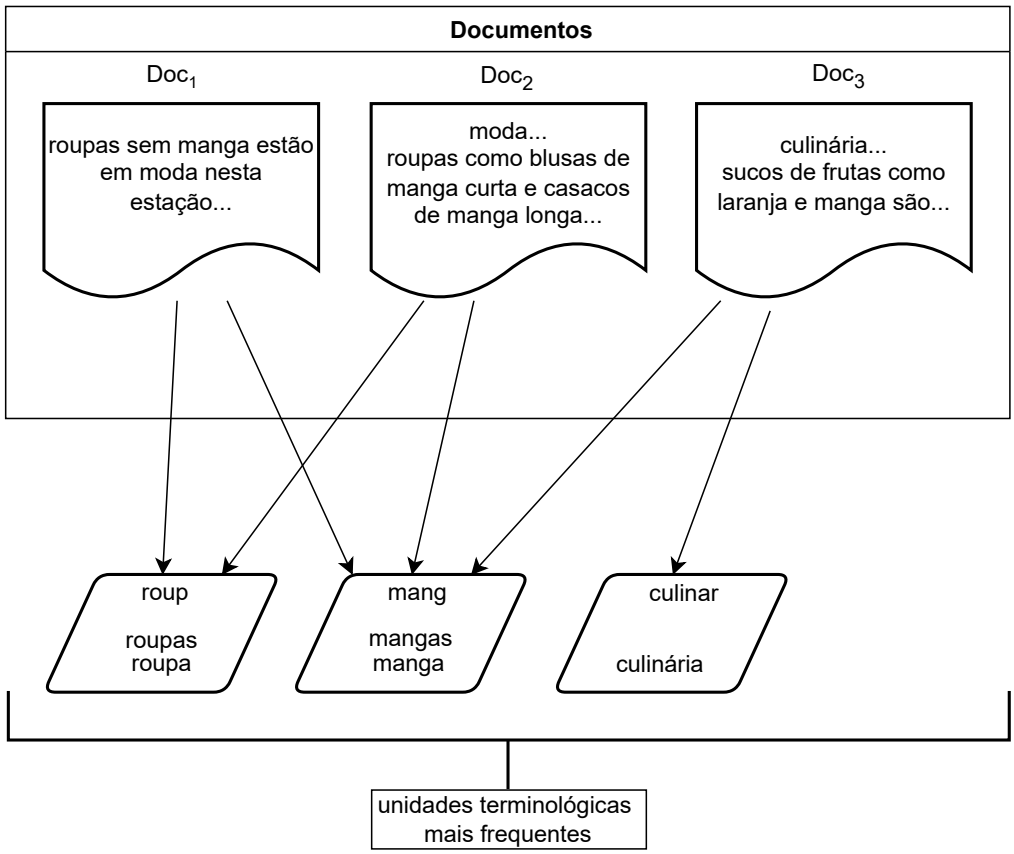
O processo de extração de termos de um documento é o passo inicial, consiste em obter todas as palavras que compõem o texto do documento eliminando apenas as pontuações. Nesta fase inicial todas as palavras serão obtidas, inclusive aquela que não possuem qualquer significado quando removidas do contexto em que estão inseridas tais como artigos, números, etc.; essas palavras serão eliminadas no passo seguinte.

O próximo passo é a remoção das *stopwords* utilizando uma tabela pequena contendo as palavras indesejadas que serão eliminadas, por exemplo: "a, ainda, além, ambas, ambos, antes, ao, aonde, aos, após, aquele, aqueles, as, assim, com, como, contra, contudo, cuja, cujas, cujo, cujos". Após essa remoção, assumimos que as palavras restantes têm alguma relação com os assuntos do texto tratado e que mesmo se tratadas de forma isolada ainda serão capazes de representar ideias.

Em seguida precisamos lidar com as variações ortográficas das palavras, principalmente com os termos flexionados, por exemplo: "apresentação, apresentado, apresentando". Essas palavras sofrem um processamento denominado stemming, sendo no exemplo anterior todas elas transformadas em "apresent". Note que uma palavra que foi radicalizada (*stemmed word*) muito provavelmente não é mais uma palavra ortograficamente válida, o que interessa nesta operação é que todas as variações sejam representadas da mesma maneira. Desta forma, quando submetidas a outros processamentos, como a contagem dos termos, todas essas variações serão tratadas como idênticas; essa característica é importante porque apesar de ortograficamente diferentes todas elas remetem ao mesmo conceito, por isso precisam ser tratadas como uma unidade. O que obtemos ao final deste processo são unidades terminológicas que possuem como entrada a palavra que foi radicalizada (*stemmed word*) e que remetem ao conjunto original de termos responsáveis; essa relação será armazenada porque no final de outros processos precisaremos destas palavras em sua grafia original. O processo descrito até este momento pode ser visto na Figura 7.

Uma vez obtidas as unidades terminológicas podemos contar as ocorrências delas dentro de cada documento (*term frequency*) e a partir de um limiar mínimo de suporte obter um novo e reduzido conjunto destas entidades que são as mais representativas de cada documento; este conjunto é a primeira forma de representação de cada documento. Com esta representação é possível mensurar a similaridade que um documento possui em relação a outro, isto é algo essencial pois todos os algoritmos de *clustering* necessitam de

Figura 7 – Obtenção das unidades terminológicas



Fonte: autor

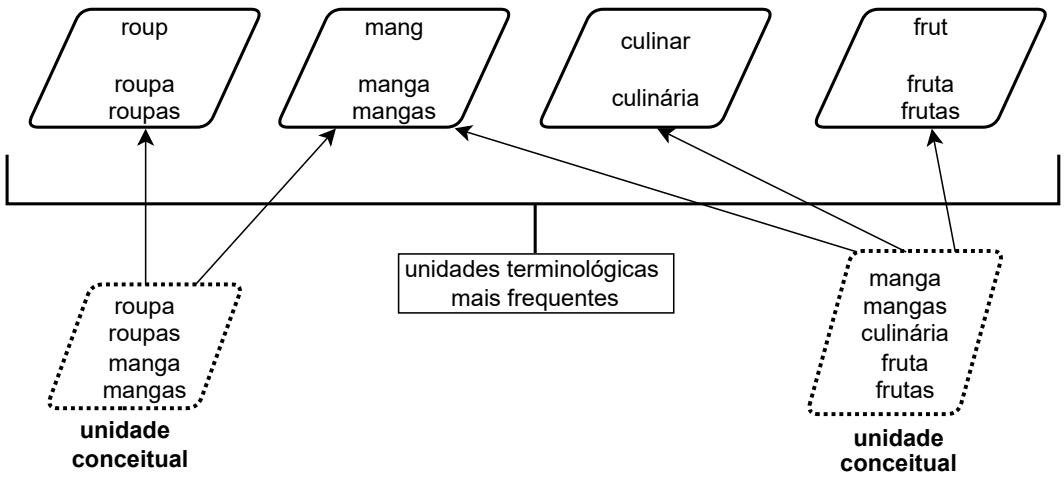
uma função capaz de aferir a similaridade entre cada documento durante o processo de agrupamento.

A seguir os documentos são processados por um algoritmo de *clustering* que produzirá grupos contendo documentos de alguma forma similares. Esses grupos não possuem etiquetas, nomes ou qualquer outro tipo de identificação; nesta fase sabemos que todos os documentos pertencentes ao mesmo grupo possuem um elevado grau de similaridade, apenas isso.

O próximo passo será processar cada grupo de documentos individualmente, estabelecendo as regras de associação entre os termos radicalizados envolvidos na medição de similaridade; termos que possuírem uma taxa elevada de associação serão considerados como uma unidade terminológica composta, que neste caso é o que denominamos de unidade conceitual, veja Figura 8. A identificação dos agrupamentos de documentos será uma lista de uma ou mais unidades conceituais; essas unidades além de possuírem o designativo radicalizado também possuem o conjunto originário de termos.

Estabelecer regras de associação é um processamento utilizado em *data mining* que busca estabelecer uma cadeia de implicações, ou seja, tenta encontrar elementos

Figura 8 – Obtenção das unidades conceituais



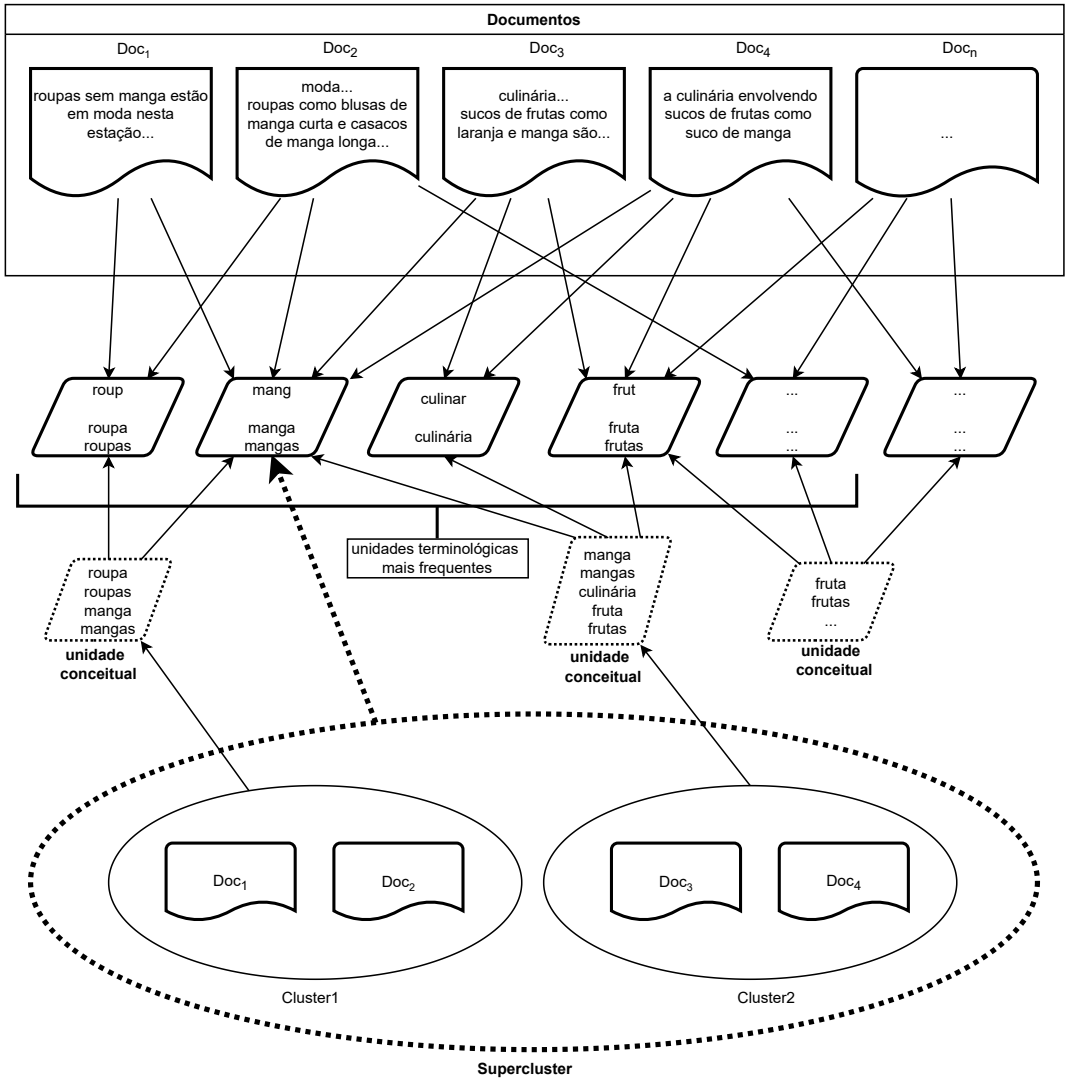
Fonte: autor

cuja presença implica na presença de outros elementos dentro de uma mesma transação. Exemplo: em um conjunto de documentos de diversos assuntos, muito provavelmente aqueles de discutem sobre "desmatamento" (primeiro elemento) também mencionarão com frequência "meio-ambiente" (segundo elemento) dentro do mesmo documento (transação), neste caso a presença de um termo implica na presença do outro no mesmo documento, ou se utilizarmos de outra nomenclatura: a presença do primeiro elemento implica na presença do segundo elemento dentro de uma mesma transação com uma probabilidade X disto ocorrer. Essa relação de implicação sempre será probabilística e a cadeia não precisa necessariamente limitar-se a dois elementos.

Como neste momento já possuímos documentos agrupados e cada grupo identificado por uma ou mais unidades conceituais, podemos obter o supergrupo que será composto pelos grupos de documentos nos quais a identificação de cada um deles possua várias unidades conceituais em comum entre si. A ideia aqui é que este supergrupo sirva para representar os termos ambíguos. A identificação deste agrupamento é obtida por meio da análise das regras de associação entre os termos radicalizados presentes nas unidades conceituais utilizadas como designativo de cada grupo envolvido em um determinado supergrupo.

Ao final do processo descrito obtemos dois níveis de agrupamento. No primeiro nível teremos documentos agrupados por semelhança conceitual, inferida a partir da semelhança terminológica e da probabilidade associativa entre os termos ao compor a cadeia terminológica representativa do conceito. No segundo nível o agrupamento ocorre entre os grupos de documentos, sendo este agrupamento representativo dos termos ambíguos presentes na representação de vários conceitos. O principal objetivo deste segundo agrupamento é lidar com situações em que o corpus textual possua documentos com assuntos diversos, sempre

Figura 9 – Ilustração da proposta

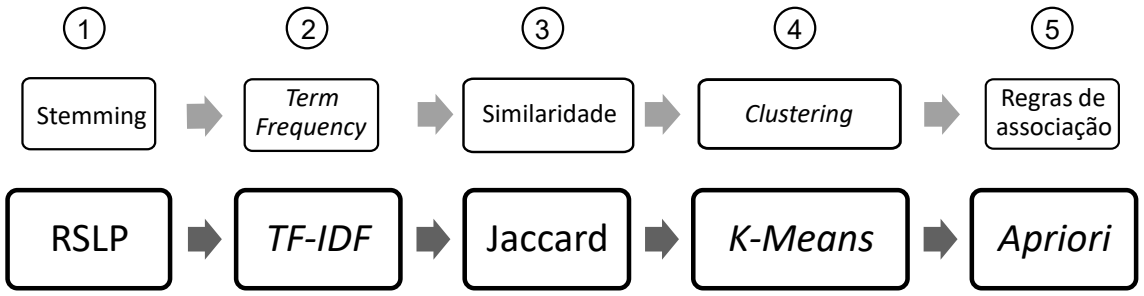


Fonte: autor

considerando que cada texto individualmente trate majoritariamente de um único assunto. Todo esse processo descrito está ilustrado na Figura 9.

4.2 Definição dos algoritmos utilizados

Figura 10 – Ilustração dos algoritmos utilizados na proposta



Fonte: autor

Uma vez definido os passos necessários precisamos escolher os algoritmos mais adequados à proposta. Precisaremos definir os seguintes algoritmos: (1) *stemming*, (2) *term frequency*, (3) similaridade, (4) *clustering*, (5) regras de associação, conforme ilustrado na Figura 10.

O algoritmo de stemming escolhido foi o Algoritmo Removedor de Sufixo da Língua Portuguesa (RSLP). Dentre os diversos algoritmos para stemming existentes, esse foi desenvolvido especificamente para o idioma português, que é o perfil de documento utilizado neste trabalho.

A estimativa das palavras mais representativas do conteúdo de cada documento é feita por meio de *term frequency – inverse document frequency*, utilizando-se a fórmula geral com compensação logarítmica, que beneficia termos que são simultaneamente frequentes dentro de um determinado documento e infrequentes entre os documentos.

Outra escolha importante durante a aplicação do algoritmo de *clustering* é a função de similaridade. Optamos por utilizar a medida de semelhança denominada Coeficiente de Similaridade de Jaccard.

Para os agrupamentos será necessário um algoritmo de *clustering* usado em dois momentos. O algoritmo escolhido para tal tarefa é o *K–Means Clustering*. Primeiramente, ele será responsável pelo agrupamento dos documentos semelhantes, sendo necessário estimar a quantidade de agrupamentos almejada antes de empregarmos o algoritmo. Posteriormente, ele produzirá os supergrupos dos termos ambíguos.

Finalmente, para viabilizar esta proposta, será necessário determinar as regras de associação; estas serão descobertas por meio do Algoritmo *Apriori*. A seguir faremos uma descrição detalhada dos algoritmos escolhidos para compor cada um dos passos descritos anteriormente.

4.3 Stemming: Removedor de Sufixo da Língua Portuguesa (RSLP)

O algoritmo Removedor de Sufixo da Língua Portuguesa (RSLP) é um algoritmo de stemming proposto em 2001 por Viviane Moreira (na época Viviane Moreira Orengo) e Christian Huyck (Orengo; Huck, 2001). Ele consiste em 8 passos, sendo cada passo composto por uma série de regras que serão testadas sequencialmente.

Dentro da cada passo, cada regra define uma precondição que deve ser atendida pela palavra em processamento e uma lista de exceções (palavras completas ou sufixos) na qual a palavra não pode constar; uma vez satisfeita essas duas condições, a regra de transformação será aplicada e o passo atual encerrado, não ocorrendo a avaliação das demais regras daquele passo; palavras que combinem com alguma da lista de exceções, encerram a regra atual e o processamento prossegue para a regra seguinte dentro do mesmo passo. Os oito passos são os seguintes:

1. redução do plural;
2. redução do feminino;
3. redução do aumentativo/diminutivo;
4. redução adverbial;
5. redução nominal;
6. redução verbal;
7. remoção de vogais temáticas;
8. remoção de acentuação.

Os passos possuem como precondição um tamanho mínimo da palavra em processamento e em alguns casos uma lista de sufixos. Por exemplo: a regra de redução do plural exige uma palavra com tamanho mínimo de três letras e sufixo “s”; a regra de redução do feminino também exige uma palavra com no mínimo três letras e sufixos “a” ou “ã”; os demais passos redutivos não exigem tamanho mínimo da palavra nem sufixo, sendo aplicados sempre.

As regras são compostas por quatro elementos: (1) o sufixo que será removido, (2) o tamanho mínimo da raiz resultante após a remoção do sufixo, (3) o novo sufixo que será aplicado à raiz e (4) uma lista de exceções composta pelas palavras que não deverão sofrer a aplicação da regra. Como as regras são analisadas sequencialmente na ordem numericamente crescente, os sufixos mais longos estão posicionados nas regras numericamente menores que aquelas que possuem sufixos mais curtos garantindo que

o *match* do sufixo mais longo ocorra primeiro. A Figura 11 apresenta o fluxograma do algoritmo.

O procedimento adotado nos primeiros sete passos consiste em utilizar parâmetros definidos em um quadro específico para cada passo (veja anexo A). Cada quadro possui cinco elementos: (1) ordem de execução, (2) sufixo original, (3) tamanho mínimo da raiz, (4) sufixo final e (5) exceções.

O processamento de todos os passos é idêntico, existindo um quadro paramétrico específico para cada passo. Cada quadro desses é composto por uma ou mais regras dispostas uma por linha, que devem ser analisadas em ordem numericamente crescente, iniciando-se sempre na regra número 01.

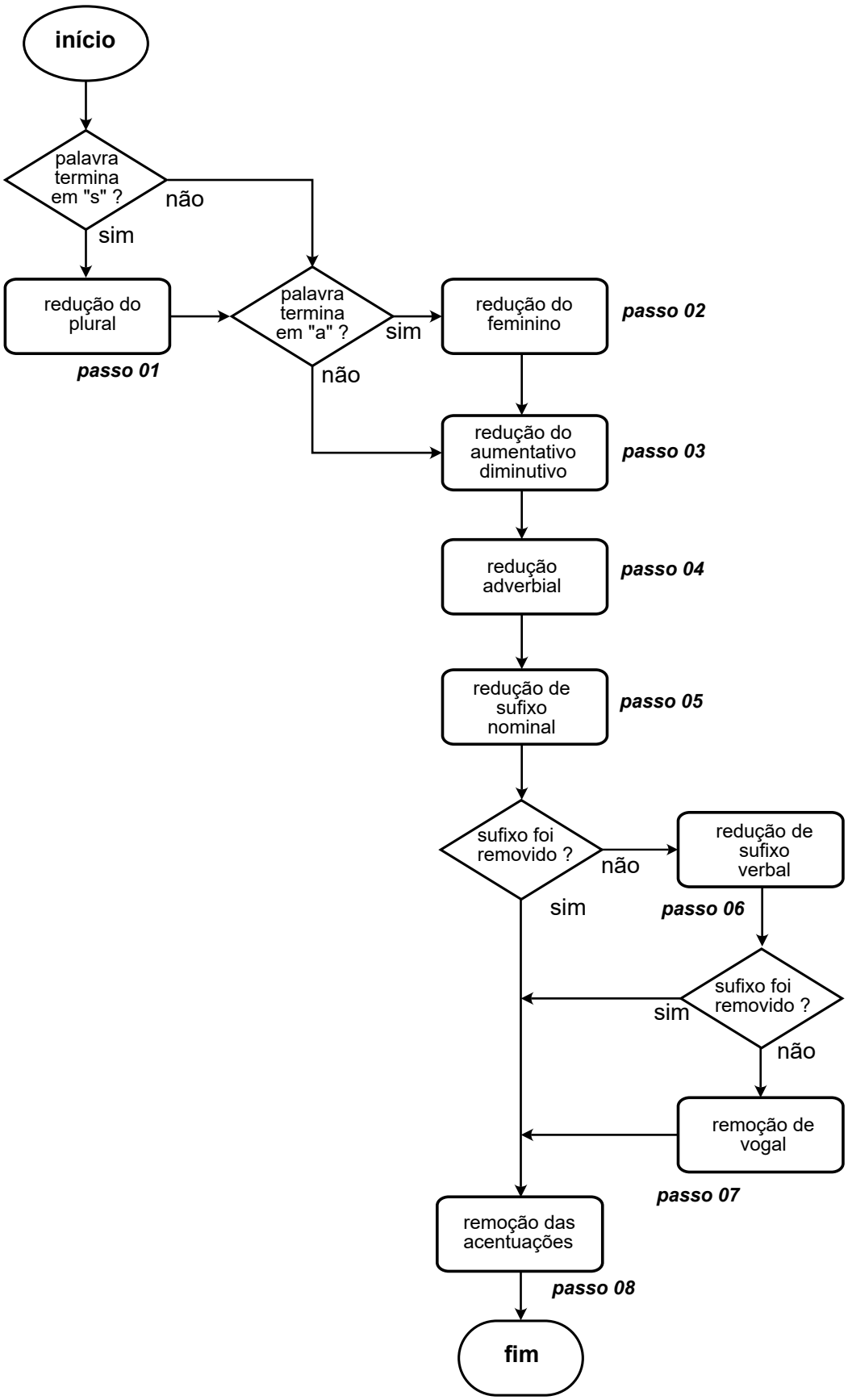
Inicialmente a palavra em processamento é decomposta em $prefixo + sufixo$ de forma que o sufixo tenha o mesmo tamanho (em quantidade de letras) que o sufixo constante na regra sendo analisada para que ambos sejam comparados em semelhança textual, ou seja, comparar sufixos significa comparar o trecho final da palavra em processamento com o trecho constante na regra atualmente em análise. O sufixo assim obtido é comparado com o sufixo da regra atual e caso não sejam idênticos, avançamos para a próxima regra repetido o processo de decomposição da palavra conforme descrito aqui. Sendo idênticos, o início da palavra em análise, desconsiderando o sufixo, será tratado como prefixo. Veja na Figura 12.

Após esta primeira comparação, caso os sufixos sejam idênticos, o prefixo ou a palavra completa (isso dependerá do passo) serão testados em relação a lista de exceções. Sendo exceção prosseguimos para a próxima regra. Se não constar na lista de exceções, é verificado se o tamanho (em letras) mínimo do prefixo é atendido, e em caso positivo, geramos uma nova palavra a partir do prefixo extraído e do novo sufixo constante na regra. Caso o tamanho do prefixo não atinja o mínimo exigido pela regra, passamos para a próxima regra.

O procedimento para a aplicação de cada regra envolvida em cada passo, conforme descritos acima, estão esquematizados no fluxograma da Figura 13. Observe que o banco de dados representado no fluxograma com o nome “RSLP regras” é composto pelas regras descritas nos quadros constantes no Anexo A (p. 110).

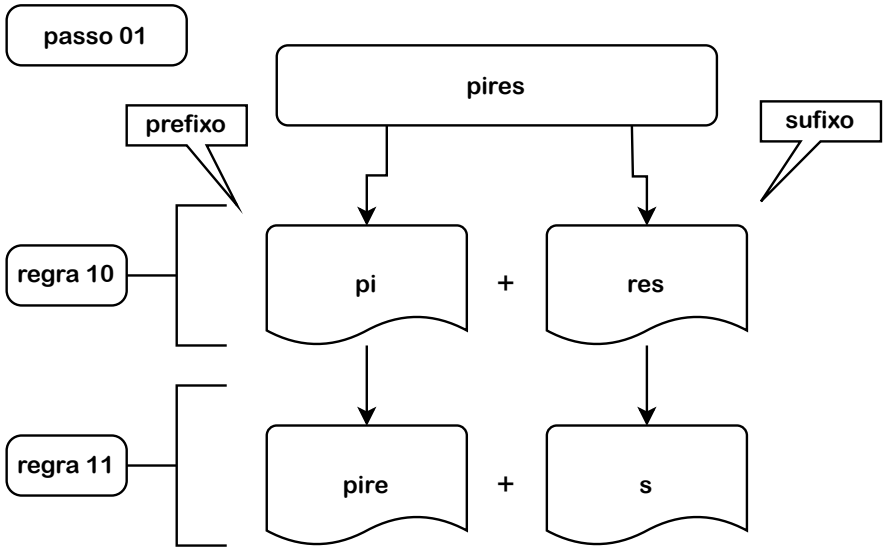
A seguir, serão exemplificados os resultados obtidos após a aplicação de cada um dos passos em determinadas palavras.

Figura 11 – fluxograma do algoritmo RSLP



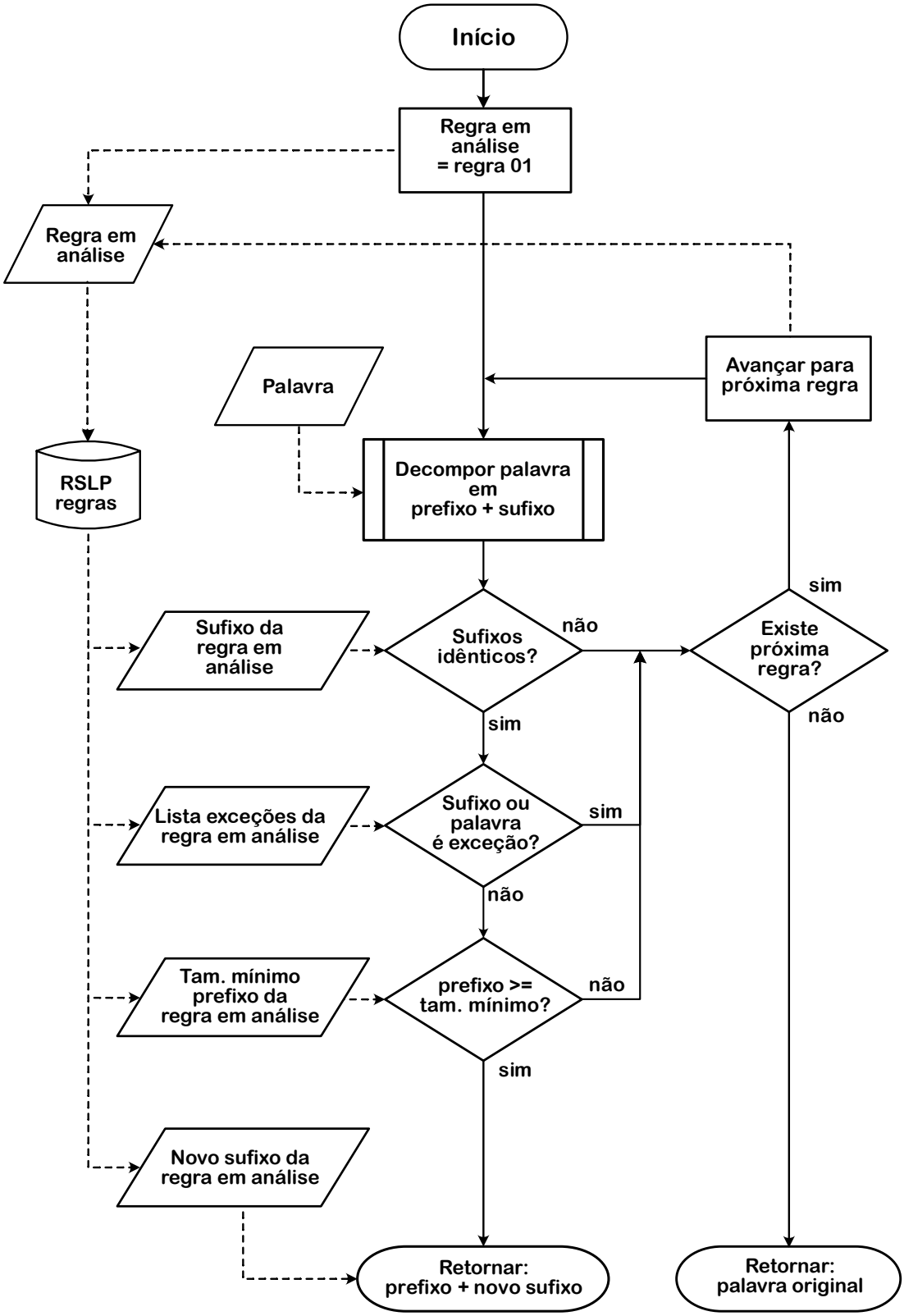
Fonte: autor

Figura 12 – Obtenção do prefixo e do sufixo



Fonte: autor

Figura 13 – Fluxograma de execução de um passo do Algoritmo RSLP



Fonte: autor

4.3.1 Passo 01: redução do plural

O primeiro passo é a redução do plural, aplicado apenas às palavras terminadas em "s". Consiste em reduzir, a partir de uma tabela contendo 11 sufixos (veja no Anexo A, Quadro 24, p. 110) dos quais a palavra sendo processada será testada conforme explicado anteriormente. Exemplos no Quadro 6. (Anexo A, Quadro 24, p. 110)

Quadro 6 – Algoritmo RSLP exemplo de redução plural

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
bons	bo + ns	aplicada a regra 01	bom
balões	bal + ões	aplicada a regra 02	balão
mais	m + ais	exceção à regra 04 (palavra listada como exceção)	mais
mancais	manc + ais	aplicada regra 04	mancal
pires	pi + res pire + s	exceção à regra 10 (raiz não possui o tamanho mínimo); exceção à regra 11 (palavra listada como exceção)	pires

Fonte: autor.

4.3.2 Passo 02: redução do feminino

Este passo é aplicado apenas aquelas palavras cuja terminação seja a letra "a". Todas as palavras que atendam a esse critério terão seu processamento realizado conforme explicado anteriormente utilizando-se como parâmetros o quadro específico deste passo (Anexo A, Quadro 25, p. 111). Exemplos no Quadro 7.

Quadro 7 – Algoritmo RSLP exemplo de redução plural

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
amora	am + ora	exceção à regra 02 (raiz não possui o tamanho mínimo)	amora
chinesa	chin + esa	aplicada a regra 05	chinês
galinha	gal + inha	aplicada regra 04	galinho
minha	m + inha	exceção à regra 04 (raiz não possui o tamanho mínimo)	minha
amorosa	amor + osa	aplicada regra 06	amoroso
prosa	pr + osa	exceção à regra 06 (palavra listada como exceção)	prosa

Fonte: autor.

4.3.3 Passo 03: redução adverbial

Este passo possui uma única regra: sufixo "mente" com raiz mínima de quatro letras (Anexo A, Quadro 26, p. 111), remoção do sufixo; exceptuando-se a palavra "experimente". Exemplos no Quadro 8 a seguir:

Quadro 8 – Algoritmo RSLP exemplo de redução adverbial

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
experimente	experi + mente	exceção à regra 01 (palavra listada como exceção)	experimente
loucamente	louca + mente	aplicada regra 01	louca
rapidamente	rapid + mente	aplicada regra 01	rapid

Fonte: autor.

4.3.4 Passo 04: redução do aumentativo/diminutivo

Este passo está exemplificado no Anexo A, Quadro 27, p. 112; veja o exemplo no Quadro 9.

Quadro 9 – Algoritmo RSLP exemplo de redução do aumentativo e do diminutivo

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
zézinho	zé + zinho	aplicada a regra 02	zé
queridíssimo	querid + íssimo	aplicada a regra 03	querid
moinho	mo + inho	exceção à regra 10 (raiz não possui o tamanho mínimo)	moinho
quentinho	quent + inho	aplicada a regra 10	quent
caminho	cam + inho	exceção à regra 10 (palavra listada como exceção)	caminho

Fonte: autor.

4.3.5 Passo 05: redução de sufixo nominal

O exemplo sobre "redução de sufixo nominal" está no Quadro 10 (parâmetros no Anexo A, Quadro 28, p. 116).

Quadro 10 – Algoritmo RSLP exemplo de redução de sufixo nominal

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
alimento	al + imento ali + mento	exceção à regra 06 (raiz não possui o tamanho mínimo); exceção à regra 07 (raiz não possui o tamanho mínimo)	alimento
alfabetizado	alfabet + izado	exceção à regra 10 (palavra listada como exceção)	alfabetizado
sonhador	sonh + ador	aplicada a regra 17	sonh

Fonte: autor.

4.3.6 Passo 06: redução de sufixo verbal

A "redução do sufixo verbal" é exemplificada no Quadro 11 (parâmetros no Anexo A, Quadro 29, p. 119).

Quadro 11 – Algoritmo RSLP exemplo de redução de sufixo verbal

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
festejassem	festej + assem	aplicada a regra 13	festej
festejando	festej + ando	aplicada regra 27	festej
faroeste	faro + este	exceção à regra 51 (palavra listada como exceção)	faroeste

Fonte: autor.

4.3.7 Passo 07: remoção de vogal

O passo reponsável pela "remoção vogal" é exemplificado no Quadro 12 (parâmetros no Anexo A, Quadro 30, p. 119).

Quadro 12 – Algoritmo RSLP exemplo de remoção de vogal

palavra original	**raiz+sufixo**	**motivo**	**palavra final**
contábil	contá + bil	aplicada a regra 01	contável
gangue	gan + gue	exceção à regra 02 (palavra listada como exceção)	gangue
estilingue	estilin + gue	aplicada a regra 02	estiling
bebê	beb + ê	exceção à regra 04 (palavra listada como exceção)	bebê
criança	crianç + a	aplicada a regra 05	crianç

Fonte: autor.

4.3.8 Passo 08: remoção das acentuações

Neste último passo, toda a acentuação é removida incluindo: cedilha, acento agudo, acento grave, til. Apenas a acentuação é removida, não a letra acentuada; portanto "ç" será substituído por "c", "ã" será substituído por "a" e assim sucessivamente. Exemplos: crianç $\Longrightarrow$ crianc; bebê $\Longrightarrow$ bebe; contável $\Longrightarrow$ contavel.

4.4 Term frequency – Inverse Document Frequency (TF-IDF)

Esta medida tenta estimar a importância que um determinado termo possui dentro de um documento a partir do número de vezes que este termo aparece dentro do texto. É uma medida estatística de frequência relativa e ponderada.

Encontramos o conceito de TF-IDF na obra de Karen Spärck Jones, que o discute a partir dos conceitos de exaustividade e especificidade, ponderando que a especificidade é

"uma propriedade semântica do termo de indexação" (Spärck Jones, 1972, p. 14, tradução nossa), sendo a especificidade quem define a capacidade de discriminação de um termo, se contrapondo à exaustividade, que é uma propriedade das descrições obtidas a partir do processo de indexação, sendo mais exaustiva a indexação quanto mais termos descritores forem relacionados. A tese da autora é que "[a especificidade] deve ser interpretada como uma propriedade estatística em vez de semântica dos termos de indexação"[1] (Spärck Jones, 1972, p. 13, tradução nossa), ou seja, a especificidade de um termo é em função da frequência de uso deste termo, fundamentando com isso a abordagem estatística da extração de termos.

A fórmula para o cálculo do IDF de um termo k (Salton; McGill, 1983, p. 63) é definida na Equação 4.1 e simplificada na Equação 4.2, sendo idênticas:

$$IDF_k = \log_2 \frac{n}{Docfreq_k} + 1 \tag{4.1}$$

$$\log_2(n) - \log_2(Docfreq_k) + 1 \tag{4.2}$$

Os parâmetros destas equações (4.1 e 4.2) são: n é o número de documentos totais que compõe o corpus; $Docfreq_k$ é o número de documentos em que o termo k aparece pelo menos uma vez. A função $\log_2$ serve para que termos muito raros não sejam muito beneficiados pela função, obtendo um índice IDF_k muito baixo, próximo de seu valor mínimo que é 1.

A partir da fórmula de IDF_k, Salton e McGill (1983, p. 63) criaram a função para cálculo do peso que cada termo extraído tem para aquele documento, remetendo à ideia de quão representativo aquele termo é para aquele documento e em relação a outros documentos. O peso $WEIGHT$ de um termo k em um documento i é definido como:

$$WEIGHT_{ik} = FREQ_{ik} * IDF_k \tag{4.3}$$

Na Equação 4.3, $FREQ_{ik}$ é o *Term Frequency* (TF) daquele termo k especificamente dentro do documento i, ou seja, é a contagem do número de ocorrências do termo k dentro do documento i. O nome mais popular para $WEIGHT_{ik}$ é *Term Frequency Inverse Document Frequency* (TF-IDF).

Utilizando este índice TF IDF, podemos estabelecer um patamar mínimo para que uma palavra extraída seja considerada um termo, por causa de seu poder discriminatório, e com isso participar do algoritmo aqui proposto.

[1] *It should be interpreted as a statistical rather than semantic property of index terms.*(Spärck Jones, 1972, p. 13)

4.5 Similaridade entre documentos: Coeficiente de Jaccard

Este coeficiente afere a similaridade terminológica entre os documentos envolvidos, considerando apenas aqueles que possuem em comum os termos envolvidos, com isso, quanto mais próximo de zero for este valor, mais semelhantes são os documentos. O cálculo é feito empregando a seguinte fórmula (4 1):

$$S_{ij} = \frac{b+c}{a+b+c} \tag{4.4}$$

Os parâmetros da Equação 4.4 são:

- a = número de vezes que ambos os termos estão presentes (valor 1)
- b = número de vezes em que $t_i = 1$ e $t_j = 0$
- c = número de vezes em que $t_i = 0$ e $t_j = 1$

Exemplo: cálculo do índice de similaridade de Jaccard

$$doc_1 = \begin{Bmatrix} laranja \\ pera \\ banana \\ maca \end{Bmatrix} \; doc_2 = \begin{Bmatrix} laranja \\ banana \\ limao \end{Bmatrix} \; doc_3 = \begin{Bmatrix} pera \\ uva \\ limao \\ ameixa \\ maca \end{Bmatrix}$$

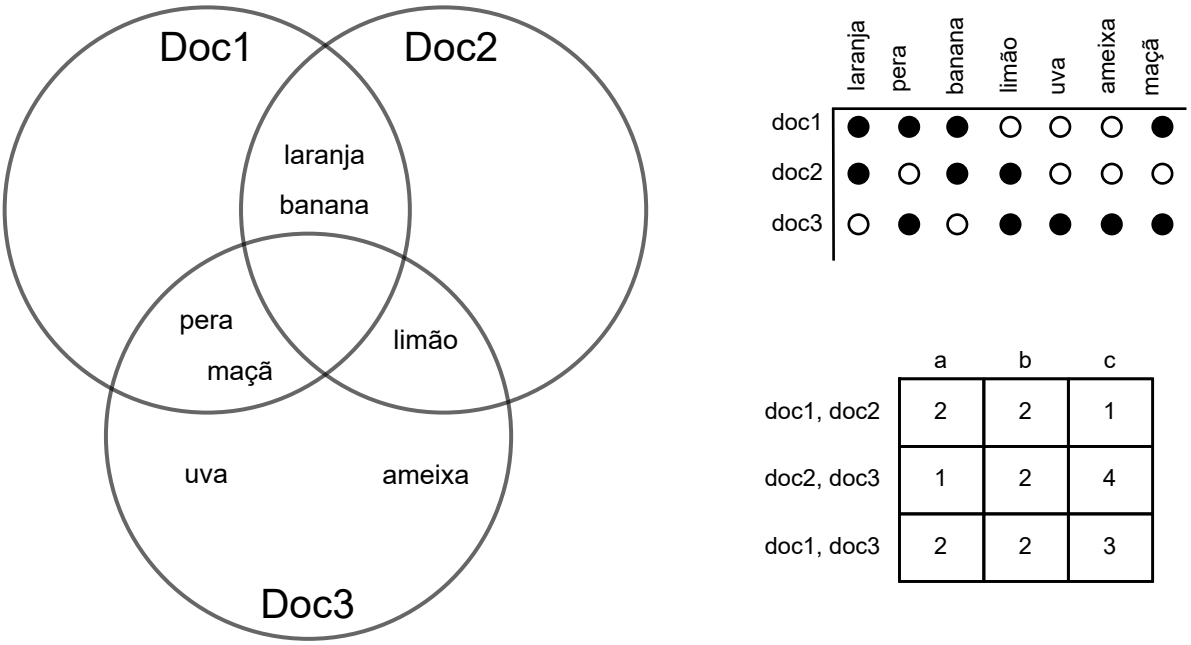
Neste exemplo temos três documentos (doc_1, doc_2, doc_3) que possuem palavras em comum entre si e que estão representados graficamente na Figura 14 por meio de um Diagrama de Venn, bem como por três outras tabelas. Estas tabelas sintetizam as três maneiras de contabilizar os termos presentes, ao considerarmos os documentos aos pares. Estas maneiras são: (a) palavras presentes em ambos os documentos, (b) palavras presentes unicamente no primeiro documento e (c) palavras presentes unicamente no segundo documento. Estes são os três parâmetros que serão utilizados na fórmula de similaridade de Jaccard (Equação 4.4).

Quadro 13 – matriz de similaridade entre os documentos doc1, doc2, doc3

{laranja, pera, banana, limão, uva, ameixa, maçã}	doc_1	doc_2	doc_3
$Doc_1 = \{1,1,1,0,0,0,1\}$	0	$(2+1)/(2+2+1) = 3/5 = 0.6$	$(2+3)/(2+2+3) = 5/7 \approx 0.71$
$Doc_2 = \{1,0,1,1,0,0,0\}$	0.60	0	$(2+4)/(1+2+4) = 6/7 \approx 0.86$
$Doc_3 = \{0,1,0,1,1,1,1\}$	0.71	0.83	0

Fonte: autor.

Figura 14 – exemplo de similaridade de Jaccard



Fonte: autor.

Após os cálculos, a ordem de semelhança dos documentos, indo do mais semelhante para o menos semelhante é: (doc_1, doc_2), (doc_1, doc_3), (doc_2, doc_3).

4.6 Algoritmo de *Clustering*: *K–Means Clustering*

O algoritmo *K–Means Clustering* é do tipo não supervisionado, ou seja, não necessita de treinamento prévio utilizando um conjunto de dados etiquetados; o único parâmetro necessário é o quantitativo de agrupamentos desejados ao final da execução do algoritmo, denominado parâmetro k. Uma característica peculiar deste algoritmo é que ele sempre produzirá agrupamentos na quantidade k solicitada independentemente destes agrupamentos existirem empiricamente, portanto, é necessário cautela ao definir este único parâmetro porque se ele não for condizente com a realidade do conjunto de dados, obteremos agrupamentos incorretos. Existem métodos estatísticos para se determinar esse valor, um desses métodos que será utilizado é o chamado *Silhouette*.

O algoritmo *k–means clustering* é definido pela Equação 4.5 proposta por Wu (2012, p. 7–8):

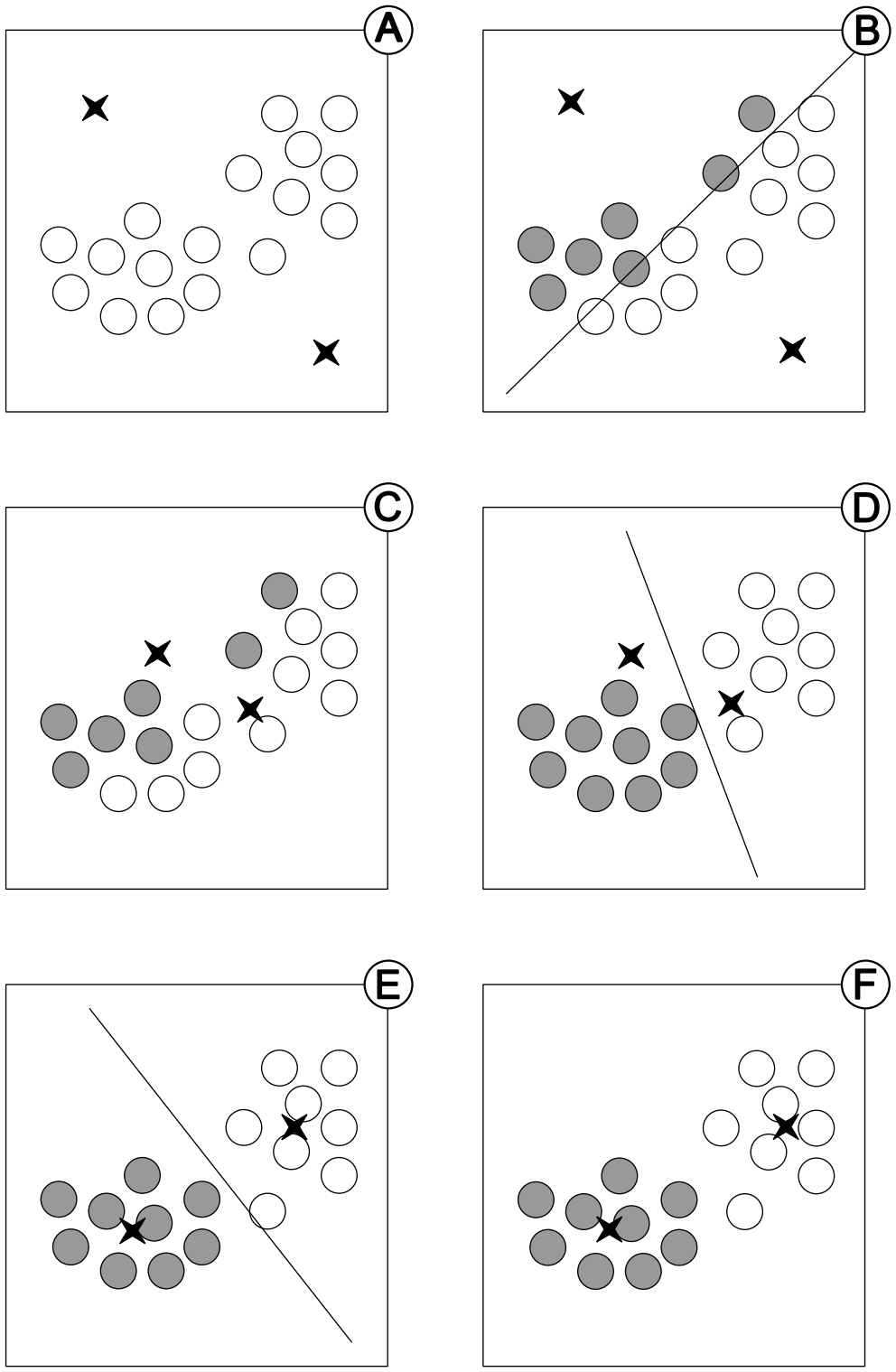
$$\min_{\{m_k\},1\leq k\leq K} \sum_{k=1}^{K} \sum_{x\in C_k} \pi_x.dist(x, m_k) \tag{4.5}$$

A Equação 4.5 assume o seguinte: que existe um $\mathfrak{D} = \{x_1, \ldots, x_n\}$ sendo o conjunto de dados que serão agrupados, π_x é o peso de x, n_k é o número de objetos associados ao *cluster* C_k, $m_k = \sum_{x\in C_k} \frac{\pi_x X}{n_k}$ é o centróide do *cluster* C_k; k é o número de *clusters* definido pelo usuário; e a função *dist*calcula a distância entre o objeto x e o centroide $m_k, 1 \leq k \leq K$. (Wu, 2012, p. 7–8)

A aplicação desta fórmula (4.5) é feita da seguinte maneira, conforme ilustrado na Figura 15. Nesta figura é apresentado um exemplo com dois centroides, representados graficamente por duas estrelas de quatro pontas (polígonos côncavos de oito vértices):

A entrada do algoritmo é feita com uma quantidade n de amostras, representadas como um conjunto de pontos $x_1, \ldots, x_n$, onde cada *feature* f do modelo de dados será representada por uma dimensão dentro de um espaço f dimensional. A quantidade de agrupamentos solicitada é definida por k com cada um dos agrupamentos possuindo um centroide chamado C, portanto existe um conjunto de centroides $\{C_1, \ldots, C_k\}$.

Figura 15 – Exemplo do algoritmo *K–Means Clutering*



Fonte: autor

O funcionamento do algoritmo obedece aos seguintes passos:

1. todos os centroides $\{C_1, \ldots, C_k\}$ serão posicionados randomicamente – Figura 15 (A);
2. uma função *dist* capaz de medir a distância entre dois pontos no espaço f dimensional é aplicada entre todos os pontos $\{x_1, \ldots, x_n\}$ versus todos os centroides $\{C_1, \ldots, C_k\}$ com o objetivo de encontrar e atribuir a cada ponto x_i o centroide mais próximo a ele – Figura 15 (B);
3. uma vez obtido o conjunto de pontos pertencente a cada centroid $C_1 = \{x_{a1}, x_{b1}, \ldots\}, C_2 = \{x_{a2}, x_{b2}, \ldots\}, \ldots, C_k = \{x_{ak}, x_{bk}, \ldots\}$, cada centroid é deslocado para a posição média de todos os pontos pertencentes aquele centroide – Figura 15 (C); isto é feito para todos os centroides;
4. o processo é repetido a partir do passo 2 – Figura 15 (D) e Figura 15 (E).
5. a repetição do passo anterior será finalizada quando após iterações consecutivas os pontos não mudarem de centroide e, consequentemente, os centroides pararem de se deslocar – Figura 15 (F)

4.6.1 Qualidade do *Cluster*: método *Silhouette*

O algoritmo *K–Means*, discutido na seção 4.6 (página 81), exige que o número de agrupamentos desejados seja informado antes de sua utilização, sendo necessário um método para estimar esse parâmetro. Um destes métodos é o *Silhouette*, que permite aferir a qualidade dos clusters obtidos a partir de um determinado quantitativo de grupos solicitado ao *k–means*, bastando aplicar o *k–means* com diversos quantitativos de grupos e escolher aquele quantitativo que ofereça um melhor índice de coesão e separação.

O método *Silhouette* foi proposto por Peter J. Rousseeuw (1987), ele permite medir dentro de um *cluster* sua coesão (o quão similares são os objetos pertencentes a um mesmo *cluster*) e sua separação (o quanto objetos diferentes estão associados a *clusters* diferentes). Ao final o método fornece um valor na faixa entre -1 e +1, indicando o quão adequado está um determinado objeto dentro do *cluster* que lhe fora atribuído, valores maiores indicam maior assertividade. O método é fundamentado em três funções matemáticas: $a(i)$,$b(i)$,$s(i)$.

Qualquer métrica para aferir distância pode ser utilizada, em nosso caso utilizaremos a distância de Jaccard. Nas fórmulas a seguir, a aferição de distância é representada pela função $d(i, j)$, sendo i e j os dois *data points* cuja distância queremos medir.

Inicialmente, é definida a função $a(i)$ (Equação 4.6), que medirá a distância média entre i e todos os *data points* pertencentes ao mesmo *cluster*. Na fórmula a seguir, $|C_I|$ é o número de data points pertencentes ao cluster C_I. Esta função mede a coesão do *data point* i dentro de seu *cluster*.

$$a(i) = \frac{1}{|C_I|} \sum_{j \in C_I, i \neq j} d(i,j) \tag{4.6}$$

A seguir, é definida a função $b(i)$ (Equação 4.7) que medirá a dissimilaridade entre um *data point* i e algum outro *cluster* C_J, calculando a distância média entre i e todos os *data points* pertencentes ao *cluster* C_J (sendo $C_J \neq C_I$ sempre). Como esta função utiliza um operador min, ela retornará o *cluster* com a menor dissimilaridade média possível, este *cluster* é denominado *neighboring cluster*.

$$b(i) = \min_{J \neq I} \frac{1}{|C_I|} \sum_{j \in C_J} d(i,j) \tag{4.7}$$

Por fim, a partir das funções $a(i)$ e $b(i)$, podemos definir o índice do *silhouette* para um determinado *data point* i (Equação 4.8).

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{se } a(i) < b(i) \\ 0, & \text{se } a(i) = b(i) \\ \frac{b(i)}{a(i)}, & \text{se } a(i) > b(i) \end{cases} \tag{4.8}$$

4.7 Regras de Associação: Algoritmo *Apriori*

Este algoritmo foi proposto por Agrawal e Srikant em 1994 (Agrawal; Srikant, 1994) como uma solução para o problema de descobrir regras de associação entre itens de um banco de dados contendo diversas transações, ou seja, encontrar elementos cuja presença implica na presença de outros elementos dentro de uma mesma transação.

Segundo Agrawal e Srikant (1994), o problema pode ser definido formalmente da seguinte maneira: dado $\tau = \{i_1, i_2, \ldots, i_m\}$ como sendo um conjunto de literais chamados itens, $\mathfrak{D}$ sendo um conjunto de transações, onde cada transação τ é um conjunto de itens, de maneira que $\tau \subseteq J$ e associado a cada transação existe um identificador único chamado TID; Uma transação τ contém X, um conjunto de alguns itens de I, se $X \subseteq \tau$. A partir desta definição, temos que uma regra de associação é uma implicação na forma $X \implies Y$, onde $X \subset J$, $Y \subset J$ e $X \cap Y = \emptyset$. Nesta definição, a regra $X \implies Y$ tem um suporte s no conjunto de transações $\mathfrak{D}$ se uma porcentagem $s\%$ das transações em $\mathfrak{D}$ contiverem $X \cup Y$. A definição do algoritmo é feita na Figura 16.

O único parâmetro informado é percentual mínimo de suporte (minsup), que estabelece o menor quantitativo aceitável que uma determinada transação se faz presente para que ela participe do conjunto resposta da cadeia daquele comprimento e também seja candidata a formação da próxima cadeia mais longa.

Figura 16 – Algoritmo *Apriori*

$L_1 = \{large, 1 - itemsets\}$;
for $(k = 2; L_{k-1} \neq \emptyset; k++)$ **do**
  $C_k = apriorigen(L_{k-1})$; // New candidates
  for all *transactions* $\tau \in D$ **do**
    $C_t = subset(C_k, t)$; // Candidates contained in t
    for all *candidates* $c \in C_t$ **do**
      $c.count++$;
    end for
  end for
  $L_k = \{c \in C_k | c.count \geq minsup\}$
end for
$answer = U_k L_k$

Fonte: Agrawal e Srikant (1994).

4.7.1 Exemplo do algoritmo *Apriori*

Neste exemplo será utilizado um banco de dados contendo apenas 7 transações:

$$D = \{ABCD, ABD, AB, BCD, BC, CD, BD\}$$

Este banco de dados está representado no Quadro 14; também foi estabelecendo um suporte mínimo de 43%. Durante a execução do algoritmo, ele efetuará o processamento da seguinte maneira:

Quadro 14 – Exemplo: Algoritmo *Apriori* – *dataset*

	A	**B**	**C**	**D**
#01	•	•	•	•
#02	•	•		•
#03	•	•		
#04		•	•	•
#05		•	•	
#06			•	•
#07		•		•

Fonte: autor.

Incialmente os itens serão contados individualmente para estabelecer a frequência de participação de cada item nas transações. Aqueles que atingirem o patamar mínimo entrarão no próximo passo. (Quadro 15).

Estabelecido um suporte mínimo de 43%, todos os itens serão testados no próximo passo, que consiste na combinação em pares dos itens e contagem da frequência dos pares (Quadro 16).

Quadro 15 – Exemplo: Algoritmo *Apriori* – passo 01

item	**freq.**	**%**	**> 43%**
A	3/7	43%	•
B	6/7	86%	•
C	4/7	57%	•
D	5/7	71%	•

Fonte: autor.

Quadro 16 – Exemplo: Algoritmo *Apriori* – passo 02

item	**freq.**	**%**	**> 43%**
AB	3/7	43%	•
AC	1/7	14%	
AD	2/7	29%	
BC	3/7	43%	•
BD	4/7	57%	•
CD	3/7	43%	•

Fonte: autor.

Apenas os pares que atingiram o índice mínimo de suporte serão utilizados no próximo passo, que consiste em agrupar os itens em 3 elementos (Quadro 17).

Quadro 17 – Exemplo: Algoritmo *Apriori* – passo 03

item	**freq.**	**%**	**> 43%**
ABC	1/7	14%	
ABD	2/7	29%	
BCD	2/7	29%	

Fonte: autor.

Neste passo (Quadro 17), todas as combinações não atingiram o mínimo necessário; neste caso o algoritmo será encerrado . Observe que ACD não foi considerado porque o par AC foi descartado no passo anterior.

Quadro 18 – Exemplo: Algoritmo *Apriori* – Resultado final

item	**freq.**	**%**
A	3/7	43%
B	6/7	86%
C	4/7	57%
D	5/7	71%
AB	3/7	43%
BC	3/7	43%
BD	4/7	57%
CD	3/7	43%

Fonte: autor.

Após a finalização do algoritmo (Quadro 18), descobrimos que o par de itens BD possui uma alta frequência de ocorrência, aparecendo juntos em 57% das transações.

5 Experimentação

A seguir, faremos a descrição detalhada de um experimento realizado com o objetivo de demonstrar a viabilidade da proposta desta pesquisa (indexação automática de documentos textuais por meio de agrupamento conceitual dos textos).

Inicialmente, utilizando-se de uma ferramenta denominada *Bard* da empresa *Google*[1], foi gerado seis textos com a as seguintes temáticas:

- Texto 1: o impacto das mudanças climáticas e a produção de carbono
- Texto 2: o impacto das mudanças climáticas na economia mundial
- Texto 3: o impacto das mudanças climáticas na fauna e flora do mundo
- Texto 4: o impacto das mudanças climáticas no degelo polar
- Texto 5: o impacto das mudanças climáticas na vida do ser humano
- Texto 6: o impacto das mudanças climáticas na computação

Todos os textos gerados têm como temática principal as mudanças climáticas relacionadas à quatro assuntos distintos: meio-ambiente, economia, ecologia e computação; a intenção ao utilizar uma grande temática comum é trazer um conjunto de palavras ambíguas que certamente o modelo bag-of-word falharia no processo de desambiguação e com isso demonstrar a viabilidade da proposta apresentada neste trabalho.

O passo seguinte foi submeter esses textos ao processo de extração de palavras que consiste em:

1. extrair uma palavra do texto, descartando-a caso ela seja uma stopword conforme lista apresentada no Anexo C;
2. submeter essa palavra ao processo de stemming denominado RSLP, conforme descrito no trabalho (seção 4.3, página 70);
3. agrupar estas palavras radicalizadas e contar sua frequência dentro de cada texto, descartando aquelas com frequência inferior a um determinado limiar;

[1] "É uma tecnologia experimental com o modelo PaLM2 do *Google* para a colaboração entre usuário e a IA generativa", sendo o modelo PaLM2 "modelo de linguagem [que] 'lê' trilhões de palavras, identifica padrões que compõem a linguagem humana e aprende com isso". Ref: <https://bard.google.com/faq>, acesso em 01.set.23

Neste experimento foi utilizado como parâmetro a frequência mínima de quatro ocorrências. Para isso foi desenvolvida uma ferramenta em linha de comando (CLI) em linguagem de programação PHP 8 que gera os resultados em um arquivo separado por vírgula (CSV).

A partir desta contagem inicial das palavras extraídas e radicalizadas (veja Apêndice A, p. 107) obtemos o Quadro 19. Este conjunto, formado pela palavra radicalizada mais uma lista contendo uma ou mais palavras extraídas, é denominado neste trabalho de unidade terminológica (UT). Este quadro é resultado da importação do arquivo CSV gerado no passo anterior dentro de uma planilha MS Excel 365 v2310.

As unidades terminológicas juntamente com sua contagem de ocorrências serão submetidas inicialmente ao algoritmo de *silhouette* (subseção 4.6.1) para determinar o número ideal de *clusters* e posteriormente, este mesmo conjunto de dados será submetido ao algoritmo de *clustering K–Means* (seção 4.6).

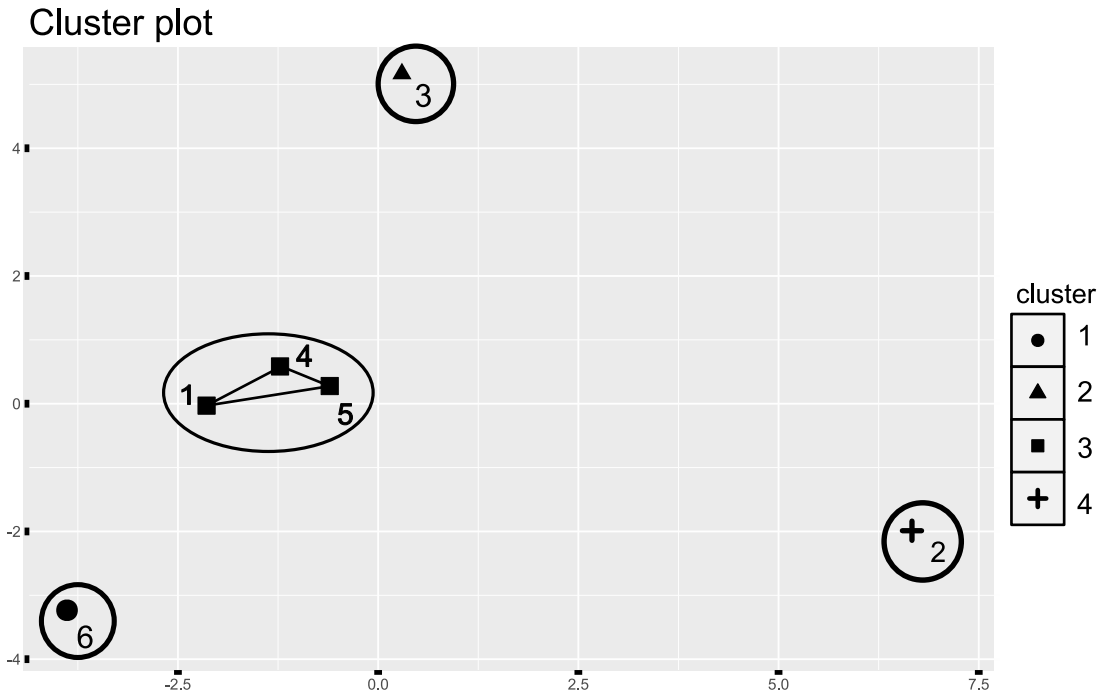
Neste experimento foi utilizada a implementação padrão do algoritmo *silhouette* e *k–means* disponível na Linguagem R versão 4.3.1; o primeiro algoritmo (*silhouete*) indicou 4 *clusters* como sendo o valor ideal e o segundo produziu o resultado presente na Figura 17. Nesta figura fica evidente que os textos 2, 3 e 6 se diferenciam entre si e que os textos 1, 4 e 5 são semelhantes.

Quadro 19 – Formação das Unidades Terminológicas

palavra radicalizada	**palavra(s) extraída(s)**	T_1	T_2	T_3	T_4	T_5	T_6	$\sum$
adapt	adaptar, adaptação	0	4	1	1	1	0	7
atmosf	atmosfera	4	0	1	2	0	0	7
aument	aumento, aumentar	7	6	4	3	3	6	29
carbon	carbono	9	0	0	0	0	0	9
caus	causas, causando	2	6	2	3	1	2	16
cent	centers	0	0	0	0	0	8	8
cibern	cibernéticos, cibernética	0	0	0	0	0	5	5
clima	climáticas, climáticos	7	22	14	9	11	8	71
comput	computação	0	0	0	0	0	4	4
consum	consumo, consumidores	1	0	0	1	1	4	7
dat	data	0	0	0	0	0	8	8
degel	degelo	0	0	0	6	0	0	6
desenvolv	desenvolvimento, desenvolvidos, desenvolver	0	6	2	0	0	0	8
econom	economia, econômicos, econômicas	0	8	0	0	0	0	8
efeit	efeito, efeitos	1	7	3	2	2	2	17
efici	eficiência, eficientes	1	0	0	1	1	4	7
energ	energia, energias	2	0	0	2	2	8	14
especi	espécies	0	0	6	2	0	0	8
faun	fauna	0	0	7	0	0	0	7
flor	flora	0	0	7	0	0	0	7
gel	gelo	0	0	0	4	0	0	4
glob	global, globais	3	1	2	3	3	4	16
habitat	habitat, habitats	0	0	4	2	1	0	7
impact	impacto, impactos	3	5	5	4	6	4	27
mar	mar, marinho, marinhos	2	2	1	6	1	1	13
mudanç	mudança, mudanças	7	21	16	8	12	7	71
muit	muitas	0	0	4	0	0	0	4
mund	mundo	0	5	3	0	0	1	9
natur	naturais	1	0	4	0	0	0	5
pais	países	0	6	0	1	3	0	10
perd	perda	1	4	1	3	1	2	12
pol	polar, polares	1	0	0	8	0	0	9
prejuiz	prejuízos	0	5	0	0	0	0	5
reduz	reduzir	3	2	4	3	3	3	18
set	setores	0	4	0	0	0	0	4
tend	tendo, tende, tendem	1	4	1	1	1	1	9

Fonte: autor.

Figura 17 – Resultado da aplicação do K-Means Cluster



Fonte: autor

A partir destes conjuntos de documentos obtidos no passo anterior, após o processamento com o algoritmo k-means, temos a seguinte definição dos conjuntos: $C_1 = \{T_1, T_4, T5\}$; $C_2 = \{T_2\}$; $C_3 = \{T_6\}$; $C_4 = \{T_3\}$ todas as unidades terminológicas que compõem os textos pertencentes a cada conjunto serão analisadas com o algoritmo apriori com o objetivo de descobrir regras de associação entre elas, sendo escolhida a(s) cadeia(s) mais longa(s).

A seleção das UTs que participarão do próximo passo é feita a partir da importância estimada de cada uma. Incialmente calculamos a importância que cada unidade terminológica possui em cada texto, considerando a métrica TF–IDF. Os valores computados para TF, Docfreq e IDF de cada UT são apresentados na Tabela 1 (página 100). O cálculo foi realizado empregando a fórmula, apresentada por Salton e McGill (1983, p. 63), obtida a partir da discussão realizada por Spärck Jones (1972) conforme apresentada na seção 4.4 (página 77 - Equação 4.1 e Equação 4.3) e reproduzidas a seguir:

$$\log_2(n) - \log_2(Docfreq_k) + 1 \tag{5.1}$$

$$WEIGHT_{ik} = FREQ_{ik} * IDF_k \tag{5.2}$$

Na Equação 5.1 a variável n é o número total de documentos presentes no corpus; k é uma das UTs; i é um documento do corpus.

Na Tabela 1, o TF informado foi ajustado pelo tamanho do documento, consistindo basicamente no quociente entre o número de vezes que a UT ocorre no documento pela quantidade total de UTs do documento. Docfreq é uma medida que considera o número de documentos nos quais a UT aparece pelo menos uma vez. IDF é uma medida logarítmica, que indica o quão comum (valores baixos) ou incomum (valores altos) a UT é dentro do corpus documental analisado. Todos os valores calculados na Tabela 1 foram obtidos a partir de uma planilha MS Excel 2023 v. 2310 preparada com as fórmulas mencionadas e preenchida com os dados provenientes dos arquivos CSV obtidos anteriormente.

Obtidos os coeficientes TF e IDF, calculamos o valor composto TF-IDF de cada UT, bem como o TF-IDF dos UTs que compõem o conjunto C1 (documentos 1, 4 e 5). O TF-IDF do conjunto C1 de documentos é obtido mediante somatória dos TF-IDF dos documentos envolvidos e posteriormente esse valor será normalizado empregando-se a fórmula min-max scalar (Equação 5.3):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5.3}$$

A partir destes valores, o TF-IDF obtido e normalizado será empregado como critério de seleção para a escolha dos UTs candidatos ao próximo passo (Tabela 2).

Tabela 1 – TF, $Docfreq$ e IDF das Unidades Terminológicas

UT	**tf (t,d1)**	**tf (t,d2)**	**tf (t,d3)**	**tf (t,d4)**	**tf (t,d5))**	**tf (t,d6)**	**Docfreq**	**IDF**
adapt	0.000	0.034	0.011	0.013	0.019	0.000	4	1.585
atmosf	0.071	0.000	0.011	0.027	0.000	0.000	3	2.000
aument	0.125	0.051	0.043	0.040	0.057	0.073	6	1.000
carbon	0.161	0.000	0.000	0.000	0.000	0.000	1	3.585
caus	0.036	0.051	0.022	0.040	0.019	0.024	6	1.000
cent	0.000	0.000	0.000	0.000	0.000	0.098	1	3.585
cibern.	0.000	0.000	0.000	0.000	0.000	0.061	1	3.585
clima	0.125	0.186	0.152	0.120	0.208	0.098	6	1.000
comput	0.000	0.000	0.000	0.000	0.000	0.049	1	3.585
consum	0.018	0.000	0.000	0.013	0.019	0.049	4	1.585
dat	0.000	0.000	0.000	0.000	0.000	0.098	1	3.585
degel	0.000	0.000	0.000	0.080	0.000	0.000	1	3.585
desenvolv	0.000	0.051	0.022	0.000	0.000	0.000	2	2.585
econom	0.000	0.068	0.000	0.000	0.000	0.000	1	3.585
efeito	0.018	0.059	0.033	0.027	0.038	0.024	6	1.000
efici	0.018	0.000	0.000	0.013	0.019	0.049	4	1.585
energ	0.036	0.000	0.000	0.027	0.038	0.098	4	1.585
especi	0.000	0.000	0.065	0.027	0.000	0.000	2	2.585
faun	0.000	0.000	0.076	0.000	0.000	0.000	1	3.585
flor	0.000	0.000	0.076	0.000	0.000	0.000	1	3.585
gel	0.000	0.000	0.000	0.053	0.000	0.000	1	3.585
glob	0.054	0.008	0.022	0.040	0.057	0.049	6	1.000
habitat	0.000	0.000	0.043	0.027	0.019	0.000	3	2.000
impact	0.054	0.042	0.054	0.053	0.113	0.049	6	1.000
mar	0.036	0.017	0.011	0.080	0.019	0.012	6	1.000
mudanç	0.125	0.178	0.174	0.107	0.226	0.085	6	1.000
muit	0.000	0.000	0.043	0.000	0.000	0.000	1	3.585
mund	0.000	0.042	0.033	0.000	0.000	0.012	3	2.000
natur	0.018	0.000	0.043	0.000	0.000	0.000	2	2.585
pais	0.000	0.051	0.000	0.013	0.057	0.000	3	2.000
perd	0.018	0.034	0.011	0.040	0.019	0.024	6	1.000
pol	0.018	0.000	0.000	0.107	0.000	0.000	2	2.585
prejuiz	0.000	0.042	0.000	0.000	0.000	0.000	1	3.585
reduz	0.054	0.017	0.043	0.040	0.057	0.037	6	1.000
set	0.000	0.034	0.000	0.000	0.000	0.000	1	3.585
tend	0.018	0.034	0.011	0.013	0.019	0.012	6	1.000

Fonte: autor

Com um limiar de 0,3 (valor escolhido experimentalmente) aplicado ao valor normalizado da somatória, foram obtidos para o conjunto C1, as seguintes UTs: atmosf, aument, carbon, clima, degel, gel, impact, mudanç e pol. O resultado é apresentado no Quadro 20.

Quadro 20 – Unidades Terminológicas selecionadas (TF-IDF normalizado >= 0,3)

		Frequência nos textos					
Palavra radicalizada	**Palavra(s) extraída(s)**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
atmosf	atmosfera	4	0	1	2	0	0
aument	aumento, aumentar	7	6	4	3	3	6
carbon	carbono	9	0	0	0	0	0
clima	climáticas, climáticos	7	22	14	9	11	8
degel	degelo	0	0	0	6	0	0
gel	gelo	0	0	0	4	0	0
impact	impacto, impactos	3	5	5	4	6	4
mudanç	mudança, mudanças	7	21	16	8	12	7
pol	polar, polares	1	0	0	8	0	0

Fonte: autor

As UTs obtidas serão submetidas ao processamento *apriori*, com o objetivo de encontrar regras de associação. O resultado é apresentado no Quadro 21.

Tabela 2 – $TF * IDF$ das Unidades Terminológicas e conjunto C_1

UT	**(t,d1)**	**(t,d2)**	**(t,d2)**	**(t,d2)**	**(t,d2)**	**(t,d6)**	**(t,C1)**	**min-max (t,C1)**
adapt	0.000	0.054	0.017	0.021	0.030	0.000	0.051	0.089
atmosf	0.143	0.000	0.022	0.053	0.000	0.000	0.196	0.341
aument	0.125	0.051	0.043	0.040	0.057	0.073	0.222	0.385
carbon	0.576	0.000	0.000	0.000	0.000	0.000	0.576	1.000
caus	0.036	0.051	0.022	0.040	0.019	0.024	0.095	0.164
cent	0.000	0.000	0.000	0.000	0.000	0.350	0.000	0.000
cibern.	0.000	0.000	0.000	0.000	0.000	0.219	0.000	0.000
clima	0.125	0.186	0.152	0.120	0.208	0.098	0.453	0.785
comput	0.000	0.000	0.000	0.000	0.000	0.175	0.000	0.000
consum	0.028	0.000	0.000	0.021	0.030	0.077	0.079	0.138
dat	0.000	0.000	0.000	0.000	0.000	0.350	0.000	0.000
degel	0.000	0.000	0.000	0.287	0.000	0.000	0.287	0.498
desenvolv	0.000	0.131	0.056	0.000	0.000	0.000	0.000	0.000
econom	0.000	0.243	0.000	0.000	0.000	0.000	0.000	0.000
efeito	0.018	0.059	0.033	0.027	0.038	0.024	0.082	0.143
efici	0.028	0.000	0.000	0.021	0.030	0.077	0.079	0.138
energ.	0.057	0.000	0.000	0.042	0.060	0.155	0.159	0.275
especi	0.000	0.000	0.169	0.069	0.000	0.000	0.069	0.120
faun	0.000	0.000	0.273	0.000	0.000	0.000	0.000	0.000
flor	0.000	0.000	0.273	0.000	0.000	0.000	0.000	0.000
gel	0.000	0.000	0.000	0.191	0.000	0.000	0.191	0.332
glob	0.054	0.008	0.022	0.040	0.057	0.049	0.150	0.261
habitat	0.000	0.000	0.087	0.053	0.038	0.000	0.091	0.158
impact	0.054	0.042	0.054	0.053	0.113	0.049	0.220	0.382
mar	0.036	0.017	0.011	0.080	0.019	0.012	0.135	0.234
mudanç	0.125	0.178	0.174	0.107	0.226	0.085	0.458	0.795
muit	0.000	0.000	0.156	0.000	0.000	0.000	0.000	0.000
mund	0.000	0.085	0.065	0.000	0.000	0.024	0.000	0.000
natur	0.046	0.000	0.112	0.000	0.000	0.000	0.046	0.080
pais	0.000	0.102	0.000	0.027	0.113	0.000	0.140	0.243
perd	0.018	0.034	0.011	0.040	0.019	0.024	0.077	0.133
pol	0.046	0.000	0.000	0.276	0.000	0.000	0.322	0.559
prejuízo	0.000	0.152	0.000	0.000	0.000	0.000	0.000	0.000
reduz	0.054	0.017	0.043	0.040	0.057	0.037	0.150	0.261
set	0.000	0.122	0.000	0.000	0.000	0.000	0.000	0.000
tend	0.018	0.034	0.011	0.013	0.019	0.012	0.050	0.087

Fonte: autor

Quadro 21 – Processamento *Apriori*

$$C_1 = \begin{cases} T_1 = \{atmosf, aument, carbon, clima, impact, mudanç, pol\} \\ T_4 = \{atmosf, aument, clima, degel, gel, impact, mudanç, pol\} \\ T_5 = \{aument, clima, impact, mudanç\} \end{cases}$$

$$apriori(C_1, 3) = \begin{cases} \{aument\ \} \\ \{clima\ \} \\ \{impact\ \} \\ \{mudanç\} \\ \{aument,\ clima\} \\ \{aument,impact\} \\ \{aument,mudanç\} \\ \{clima,impact\} \\ \{clima,mudanç\} \\ \{impact,mudanç\} \\ \{aument,clima,impact\} \\ \{aument,clima,mudanç\} \\ \{aument,impact,mudanç\} \\ \{clima,impact,mudanç\} \\ \{aument,clima,impact,mudanç\} \end{cases}$$

$$\max comprimento(apriori(C_1, 3)) = \{aument,clima,impact,mudanç\}$$

Fonte: autor

A partir desta resposta concluímos que a unidade conceitual UC_1 é composta pelas unidades terminológicas: *aument*, *clima*, *impact*, *mudança*. Este resultado é apresentado no Quadro 22.

Quadro 22 – Unidade Conceitual

palavra radicalizada	**palavra(s) extraída(s)**
aument	aumento, aumentar
clima	climáticas, climáticos
impact	impacto, impactos
mudanç	mudança, mudanças

Fonte: autor

Neste exemplo desenvolvido, como foi utilizado um conjunto pequeno de textos, os outros conjuntos obtidos contém apenas um único texto, por isso não será possível rotulá-los aqui.

O objetivo desta seção foi ilustrar como seria a aplicação prática do procedimento proposto neste trabalho para a obtenção das Unidades Conceituais a partir das palavras contidas nos textos do corpus documental. A indexação automática pode ser feita a partir destas Unidades Conceituais obtidas, utilizando técnicas tradicionais, como o modelo vetorial, discutido no trabalho Recuperação de Informação Baseada em Ontologia: Uma proposta utilizando o Modelo Vetorial (Janaite Neto, 2018).

Mesmo realizando esse experimento com um conjunto reduzido de textos curtos obtivemos o resultado esperado, demonstrando a viabilidade do método proposto. Para uma aplicação em maior escala, é necessário que seja feita uma implementação completa do algoritmo proposto, automatizando o fluxo dos dados entre cada passo.

6 Conclusões

O processo de Recuperação de Informação textual ocorre por meio da comparação entre a representação da necessidade de informação do usuário e às representações dos documentos do acervo. Será a semelhança entre ambas as representações que determinará a probabilidade de um documento satisfazer ou não a necessidade de informação do usuário.

A qualidade do processo de recuperação é fortemente dependente da maneira como os documentos e as necessidade são representados bem como a forma como a semelhança será aferida. Um grande problema enfrentado quando se tenta representar documentos textuais é que a linguagem escrita representa os conceitos por meio dos termos, e estes termos são por diversas vezes ambíguos, necessitando de um contexto fornecido pelo conjunto de termos para resolver essa ambiguidade e traçar uma relação única entre o termo delimitado por aquele contexto e o conceito.

Os algoritmos de *clustering* fornecem uma possibilidade bem interessante de organizar o conteúdo para posterior recuperação. É viável extrair conceitos a partir das relações existentes dentro de todos os textos envolvidos, sendo dispensável o uso de sistemas conceituais externos. Essas relações entre a palavras dentro dos textos são extraídas por meio de algoritmos especializados em extração de regras de associação que conseguem obter as relações de implicação entre palavras que remetem indiretamente aos conceitos discutidos. Ao dispensar o uso de uma base conceitual externa, a proposta se torna extremamente flexível e adaptável aos mais diversos tipos de conteúdo.

O presente método também permite recuperação de informação por *browsing*, ou seja, por navegação nos emaranhados conceituais fazendo o denominado *Visual Information Retrieval*.

Toda a cadeia de transformação proposta neste trabalho, na qual as palavras são transformadas em termos e estes em conceitos, pode ser vista como uma transformação de texto em dados. Ao reduzirmos um termo a um conceito estamos trabalhando com dados, portanto, após as devidas transformações, esses dados serão processados utilizando técnicas e algoritmos convencionais, neste caso, estamos operando com *clustering* de dados; nada impede que outros algoritmos de *data mining* e *machine learning* sejam utilizados. Em síntese: **quando o termo é reduzido a um conceito estamos transformando o termo em dados**.

6.1 Contribuições

Este trabalho traz as seguintes contribuições de pesquisa:

- Novos modelos de representação de texto que usam conceitos em vez de palavras;
- Método de transformação de texto em dados significativos para processamento;
- Ao representar os textos por meio dos conceitos discutidos nos próprios textos, vislumbramos a possibilidade de agrupar conceitualmente os documentos e com isso permitir a implementação de novos paradigmas de recuperação do tipo *browsing*, trazendo a possibilidade de novas formas de elaboração da *query* bem como de visualização dos resultados obtidos a partir da ideia de *clustering*.

6.2 Trabalhos futuros

Como sugestão de trabalhos futuros, uma questão que não foi explorada são as palavras compostas sem hífen. Esses casos são complexos, um exemplo seria a composição "não governamental", neste caso existe um agrupamento de conceitos definido de maneira não muito detalhada a partir da negação de outro conceito.

Uma outra situação não discutida são os sinônimos usados esporadicamente. Caso exista uma distribuição uniforme dos sinônimos entre os textos, inevitavelmente eles serão detectados como implicação e formarão cadeias semelhante que serão agrupadas em um único conceito; o problema está nos sinônimos utilizados de maneira excepcional, em um ou outro texto, nestes casos precisariam de um tratamento específico, pois se utilizados desta forma eles serão estatisticamente irrelevantes e por isso não serão detectados pelos algoritmos propostos.

Outra questão são as anáforas e outras figuras de linguagem. A técnica de análise de frequência e ocorrência de palavras não consegue lidar com figuras de linguagem. A busca de uma solução passaria por técnicas e métodos de Processamento de Linguagem Natural (PLN), algo que devido ao recorte adotado neste trabalho não foi abordado.

Percebemos que, como toda proposta, o presente trabalho precisou de um recorte temático e de um escopo bem delimitado, deixando de discutir vários problemas correlatos. Um grande obstáculo para a indexação automática está situado na linguagem escrita, que é extremamente flexível e adaptável, sendo muito difícil tratá-la como um artefato matemático processável por computadores, todas as tentativas buscam representar a linguagem para manuseá-la e como ocorre com toda representação, o resultado será uma simplificação do objeto original. A qualidade da indexação automática e, consequentemente, da recuperação da informação estão diretamente atreladas a melhores maneiras de lidar computacionalmente com essas representações dos textos. A proposta elaborada neste trabalho, quando testada em um protótipo, se mostrou promissora abrindo novas possibilidades.

Referências

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. Fast algorithms for mining association rules. In: BOCCA, Jorge B.; JARKE, Matthias; ZANIOLO, Carlo (Ed.). **Proceedings of 20th International Conference of Very Large Data Bases (VLDB'94)**. Santiago de Chile, Chile: Morgan Kaufmann, 1994. p. 487–499. ISBN 1-55860-153-8. Citado 2 vezes nas páginas 84 e 85.

ARAÚJO JÚNIOR, Rogério Henrique. **Precisão no Processo de Busca e Recuperação da Informação**. Tese (Doutorado) — Universidade de Brasília (UnB), Brasília, 2005. Citado 7 vezes nas páginas 34, 38, 41, 42, 45, 47 e 48.

ARAÚJO, Vania Maria Rodrigues Hermes de. Sistemas de informação: nova abordagem teórico conceitual. **Ciência da Informação**, v. 24, n. 1, p. 1–39, 1995. Disponível em: <https://revista.ibict.br/ciinf/article/view/610>. Citado 2 vezes nas páginas 33 e 39.

BARITÉ ROQUETA, Mario G. Los conceptos y su representación: una perspectiva terminológica para el tratamiento temático de la información. **Scire**, v. 6, n. 1, p. 31–53, jan–jun 2000. Citado na página 27.

BARLOW, Mike. **The Culture of Big Data**. Sebastpol: O'Reilly Media, 2013. Citado na página 38.

BARRETO, Aldo de Albuquerque. A questão da informação. **São Paulo em Perspectiva**, v. 8, n. 4, p. 3–8, 1994. Citado na página 40.

BARROS, Aidil de Jesus Paes de; LEHFELD, Neide Aparecida de Souza. **Projeto de pesquisa**: propostas metodológicas. Rio de Janeiro: Vozes, 2002. Citado na página 23.

BARROS, Lídia Almeida. **Curso básico de Terminologia**. São Paulo: EdUSP, 2004. Citado na página 27.

BELKIN, Nicholas J. Anomalous states of knowledge as a basis for information retrieval. **Canadian Journal of Information Science**, v. 5, n. 1, p. 133–143, 1980. Citado 2 vezes nas páginas 31 e 48.

BIO, Sergio Rodrigues. **Sistemas de Informação**: um enfoque gerencial. São Paulo: Atlas, 1996. Citado na página 40.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. 738 p. Citado na página 53.

BORLUND, Pia. The concept of relevance in IR. **Journal of the Association for Information Science and Technology**, v. 54, n. 10, p. 913–925, 2003. Citado na página 34.

BRIET, Suzanne. **Qu'est-ce que la documentation?** Paris: Édit éditions documentaires industrielles et techniques, 1951. Citado na página 25.

BUCKLAND, Michael K. What is a "document". **Journal of the American Society of Information Science**, v. 48, n. 9, p. 804–809, 1997. Citado na página 25.

BUCKLAND, Michael K. What is a "digital document"? **Document Numérique**, Paris, v. 2, n. 2, p. 221–230, 1998. Citado 2 vezes nas páginas 25 e 26.

CABRÉ, Maria Teresa. **La terminología**: Teoría, metodología, aplicaciones. Tradução Carles Tebé. Barcelona: Ed. Antártida, 1993. 521 p. Citado 2 vezes nas páginas 28 e 29.

CAMPBELL, Iain.; VAN RIJSBERGEN, Keith. The ostensive model of developing information needs. In: **Proceedings of CoLIS 2, second international conference on conceptions of library and information science**: Integration in perspective. Copenhagen: Royal School of Librarianship, 1996. p. 251–268. Citado na página 35.

CAMPOS, Maria Luiza de Almeida. **Linguagem Documentária**: teorias que fundamentam sua elaboração. Niterói: EdUFF, 2001. 133 p. Citado 2 vezes nas páginas 28 e 29.

CESARINO, Maria Augusta da Nóbrega. Bibliotecas especializadas, centros de documentação, centro de análise da informação: apenas uma questão de terminologia? **Revista da Escola de Biblioteconomia da UFMG**, v. 7, n. 2, p. 218–241, 1978. Citado na página 38.

CHOO, Chun Wei. Como ficamos sabendo – um modelo de uso da informação: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. In: _____. **A organização do Conhecimento**. São Paulo: Editora SENAC, 2003. p. 61–120. Citado 4 vezes nas páginas 30, 31, 45 e 46.

CLARKE, Arthur C. **3001: The Final Odyssey**. 1. ed. New York: Ballantine Pub. Group, 1997. 263 p. Citado na página 9.

COEIRA, Enrico. W.; VICKLAND, Victor. Is relevance relevant? user relevance ratings may not predict the impact of internet search on decision outcomes. **Journal of the American Medical Informatics Association**, v. 15, n. 4, p. 542–545, 2008. Citado 2 vezes nas páginas 37 e 38.

COELHO, Alexandre Ramos. **Stemming para a língua portuguesa**: estudo, análise e melhoria do algoritmo RSLP. Porto Alegre, 2007. Citado na página 110.

COHEN, Diana Micheline. **O consumidor da informação documentária**: o usuário de sistemas documentários visto sob a lente da análise documentária. Tese (Doutorado) — Universidade de São Paulo (USP), São Paulo, 1995. Citado na página 38.

DAHLBERG, Ingetraut. Teoria do conceito. **Ciência da Informação**, v. 7, n. 2, p. 101–107, 1978. Citado 2 vezes nas páginas 26 e 27.

DAI, Xiangfeng; BIKDASH, Marwan; MEYER, Bradley. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In: **SoutheastCon 2017**. Charlotte, NC, USA: IEEE, 2017. p. 1–7. ISSN 1558-058X. Citado na página 22.

DERVIN, Brenda. From the mind's eye of the user: the sense-making qualitative – quantitative methodology. In: _____. **Qualitative Research in Information Management**. Englewood, NJ, USA: Libraries Unlimited, 1992. p. 61–84. Citado na página 31.

FERNEDA, Edberto. **Recuperação de Informação**: Análise sobre a contribuição da ciência da computação para a ciência da informação. Tese (Doutorado) — Universidade de São Paulo (USP), São Paulo, 2003. Citado 4 vezes nas páginas 42, 47, 48 e 49.

FERNEDA, Edberto. **Introdução aos Modelos Computacionais de Recuperação de Informação**. Rio de Janeiro: Ciência Moderna, 2012. Citado na página 46.

FIGUEIREDO, Nice Menezes. **Estudos de Uso e Usuários da Informação**. Brasília: IBICT, 1994. Citado na página 30.

FOSKETT, Douglas John. A note on the concept of "relevance". **Information Storage Retrieval**, v. 8, n. 1, p. 77–78, 1972. Citado na página 35.

GIL, Antônio Carlos. **Como elaborar Projetos de Pesquisa**. 4. ed. São Paulo: Atlas, 2002. Citado 2 vezes nas páginas 23 e 24.

GONZALEZ DE GÓMEZ, Maria Nélida. Regime de informação: construção de um conceito. **Inf. & Soc.:Est.**, João Pessoa, v. 22, n. 3, p. 43–60, set./dez. 2012. Citado na página 50.

HARTER, Stephen P. Psychological relevance and information science. **Journal of the American Society for Information Science**, v. 43, n. 9, p. 602–615, october 1992. Citado na página 35.

HARTIGAN, John A. **Clustering Algorithms**. New York: John Wiley and Sons, 1975. 351 p. Citado na página 61.

HJØRLAND, Birger. The foundation of the concept of relevance. **Journal of the American Society for Information Science**, v. 61, n. 2, p. 217–237, 2010. Citado 3 vezes nas páginas 36, 37 e 38.

INGWERSEN, Peter. Cognitive perspectives of information retrieval interaction. **Journal of Documentation**, v. 52, n. 1, p. 3–50, 1996. Citado 2 vezes nas páginas 43 e 44.

INGWERSEN, Peter. Cognitive information retrieval. **Annual Review of Information Science and Technology**, v. 34, p. 3–52, 1999. Citado 3 vezes nas páginas 44, 45 e 46.

INGWERSEN, Peter; JÄVERLIN, Kalervo. Information seeking research needs extension towards tasks and technology. **Information Research**, v. 10, n. 1, p. 1–16, 2004. Citado na página 48.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 1087(1):2000**: Terminology work — vocabulary (part 1): Theory and application. Geneva, Switzerland, 2000. 42 p. Citado na página 28.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 704:2009**: Terminology work — principles and methods. Geneva, Switzerland, 2009. Citado na página 20.

JAIN, Anil Kumar.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: A review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, sep. 1999. Citado 4 vezes nas páginas 54, 60, 61 e 62.

JANAITE NETO, Jorge. **Recuperação de Informação Baseada em Ontologia**: Uma proposta utilizando o modelo vetorial. 105 p. Dissertação (Mestrado) — Universidade Estadual Paulista (UNESP), Marília, 2018. Citado na página 97.

KOHAVI, Ron; PROVOST, F. Glossary of terms. machine learning: Special issue on applications of machine learning and the knowledge discovery process. **Machine Learning**, Kluwer Academic Publishers, Boston, USA, v. 30, p. 271–274, 1998. Citado na página 51.

KONONENKO, Igor; KUKAR, Matjaž. Chapter 12: Cluster analysis. In: _____. **Machine Learning and Data Mining**. USA: Woodhead Publishing Limited, 2007. p. 321–358. Citado 4 vezes nas páginas 52, 56, 58 e 59.

KUHLTHAU, Carol Collier. **Information search process**. Boca Raton: [s.n.], 1991. (Encyclopedia of Library and Information Sciences). Citado 3 vezes nas páginas 34, 37 e 38.

LE COADIC, Yves François. **A ciência da informação**. Brasília: Briquet de Lemos Livros, 2004. Citado 2 vezes nas páginas 31 e 32.

LIMA, Gersina Ângela de; CAMPOS, Maria Luiza Almeida. Sistema de armazenamento e recuperação da informação: uma análise do impacto das variáveis e medidas visando a organização e recuperação de informação centrado no usuário. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 20, p. 1–23, 2022. Citado 3 vezes nas páginas 32, 33 e 47.

LIU, Haun; MOTODA, Hiroshi. **Feature Selection for Knowledge Discovery and data Mining**. New York: Springer Science Business Media, 1998. 214 p. Citado na página 51.

MANNING, Cristopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Himrich. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008. Citado na página 38.

MANYIKA, James. **An overview of Bard**: an early experiment with generative ai. [S.l.], s.d. 9 p. Disponível em: <https://ai.google/static/documents/google-about-bard.pdf>. Acesso em: out. 2023. Citado na página 120.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de Metodologia Científica**. 5. ed. São Paulo: Atlas, 2003. Citado 2 vezes nas páginas 23 e 24.

MARX, Karl. Teses ad feuerbach (1845). In: _____. **A Ideologia Alemã**. São Paulo: Boitempo, 2007. Citado na página 9.

MIZZARO, Stefano. How many relevances in information retrieval? **Interacting with Computers**, v. 10, n. 3, p. 303–320, 1998. Citado 4 vezes nas páginas 32, 33, 35 e 36.

MOREIRA, Walter. Tesauros e ontologias como modelos de sistemas de organização do conhecimento. **Brazilian Journal of Information Science: research trends**, v. 13, n. 1, p. 15–20, mar. 2019. Citado na página 27.

NOVELLINO, Maria Salet Ferreira. Instrumentos e metodologias de representação da informação. **Informação & Informação**, v. 1, n. 2, p. 37–45, 1996. Citado na página 47.

ORENGO, Viviane Moreira; HUCK, Christian R. A stemming algorithm for the portuguese language. In: **Proceedings of 8th INternational Symposium on String Processing and Information Retrieval (SPIRE)**. Laguna de San Raphael, Chile: IEEE Computer Society, 2001. p. 183–193. ISBN 0-7695-1192-9. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/SPIRE.2001.10024>. Acesso em: 14 set. 2017. Citado 2 vezes nas páginas 70 e 110.

RABELO, Rodrigo. Sujeito e agência informacional: comportamento, prática e ação. In: _____. **Informação: agentes e intermediação**. Brasília: IBICT, 2017. p. 101–152. Citado na página 50.

REID, Jane. A new task-oriented paradigm for information retrieval: Implications for evaluation of information retrieval systems. In: **Proceedings of CoLIS 3, 3rd international conference on the conceptions of library and information science**. Lokve, Croatia: Naklada Benja, 1999. p. 97–108. Citado 2 vezes nas páginas 35 e 36.

ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, n. 20, p. 53–65, 1987. Citado na página 83.

SALES, Luana Faria. **Ontologias de Domínio**: estudo das relações conceituais e sua aplicação. 141 p. Dissertação (Mestrado) — Universidade Federal Fluminense (UFF), Niterói, Rio de Janeiro, 2006. Citado na página 29.

SALTON, Gerard. A document retrieval system for man-machine interaction. In: **Proceedings of 1964 Annual Meeting, Association for Computing Machinery (ACM)**. New York, USA: Association for Computing Machinary, 1964. p. L2.3–1 – L2.3–20. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/800257.808923>. Citado na página 53.

SALTON, Gerard; MCGILL, Michael J. **Introduction to Modern Information Retrieval**. New York: Mcgraw Hill Computer Science Series, 1983. 448 p. Citado 2 vezes nas páginas 78 e 91.

SALTON, Gerard; YANG, Chung-Shu; WONG, Anita. A vector space model for automatic indexing. **Communications of the ACM**, v. 18, n. 11, p. 613–620, nov. 1975. Citado na página 60.

SANTOS, Claudia da Silva Amaral. **Terminologia e Ontologias: metodologias para representação do conhecimento**. Tese (Doutorado) — Universidade de Aveiro, Aveiro, Portugal, 2010. Citado na página 29.

SANZ CASADO, Elias. **Manual de estudios de usuarios**. Madri: Fundación Germán Sánchez Ruipérez, 1994. Citado na página 30.

SARACEVIC, Tefko. RELEVANCE: A Review of and a Framework for the Thinking on the Notion in Information Science. **Journal of the American Society for Information Science**, p. 321–343, 1975. Citado na página 45.

SARACEVIC, Tekfo. Modeling interaction in information retrieval. In: **Proceedings of the Annual Academy Meeting of American Society for Information Science (ASIS)**. Baltimore, USA: Wiley, 1996. v. 33, p. 3–9. ISSN 0044-7870. Citado 2 vezes nas páginas 44 e 48.

SCHAMBER, Linda; EISENBERG, Michael B.; NILAN, Michael S. A re-examination of relevance: toward a dynamic, situational definition. **Information Processing and Management**, v. 26, n. 6, p. 755–776, 1990. Citado 2 vezes nas páginas 36 e 37.

SEVERINO, Antonio Joaquim. **Metodologia do trabalho científico**. São Paulo: Cortez, 2016. Citado na página 24.

SMIT, Johanna Wilhelmina; BARRETO, Aldo de A. Ciência da informação: base conceitual para a formação do profissional. In: _____. **Formação do profissional da informação**. São Paulo: Polis, 2002. p. 9–23. Citado 2 vezes nas páginas 39 e 40.

SNEATH, Peter Henry Andrews; SOKAL, Robert Reuven. **Numerical Taxonomy**: The principles and practice of numerical classification. 1. ed. San Francisco: W. H. Freeman, 1973. Citado 3 vezes nas páginas 54, 55 e 56.

SPÄRCK JONES, Karen. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11–21, 1972. Disponível em: <https://doi.org/10.1108/eb026526>. Acesso em: 28 jul. 2017. Citado 2 vezes nas páginas 78 e 91.

THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. **Pattern Recognition**. 4. ed. San Diego, CA, USA: Academic Press, 2008. 984 p. Citado na página 62.

VAN RIJSBERGEN, Cornelis Joost. **Information Retrieval**. Londres: Butterworths, 1979. Citado na página 19.

VICKERY, Alina; VICKERY, Brian C. **Information Science in Theory and Practice**. Munique: K.G. Saur, 2004. Citado 2 vezes nas páginas 40 e 49.

WERSIG, Gernot. Procédés de la recherche terminologique. In: _____. **Textes choisis de terminologie: Fondéments théoriques de la terminologie**. Québec: Groupe interdisciplinaire de recherche scientifique et appliqué én terminologie (GISTERM), 1981. p. 283–300. ISBN 9–7829–2024–2005. Citado na página 29.

WILSON, Tom. On user studies and information needs. **Journal of Documentation**, v. 62, n. 6, p. 658–670, 2006. Citado 2 vezes nas páginas 31 e 44.

WU, Junjie. **Advances in K means Clustering**: A data mining thinking. Originalmente apresentada como tese de Ph.D. Berlin: Springer Verlang, 2012. (Springer Theses: Recognizing Outstanding Ph.D. Research). Citado na página 81.

WÜSTER, Eugene. Begriffs-und themaklassifikationen: Unterschiede in ihrem wesen und ihrer anwendung. **Nachrichtenfuer dokumentation**, v. 22, n. 3, p. 98–104, 1971. Serie INFOTERM 2-71F. Citado na página 28.

Apêndices

APÊNDICE A – Contagem de palavras radicalizadas

O Quadro 23 contém o resultado do algoritmo Removedor de Sufixos da Lingua Portuguesa (RSLP)[1] aplicado a todas as palavras dos textos constantes no Anexo B (exceptuando-se aquelas consideradas stopwords listadas no Anexo C). Também foi estabelecido um limite mínimo de 4 ocorrências para que uma determinada palavra radicalizada apareça nos quadros mencionados.

Quadro 23 – palavras radicalizadas e contagem de frequência

arquivo	palavra radicalizada	frequência
Texto1	carbon	9
Texto1	mudanç	7
Texto1	clima	7
Texto1	aument	7
Texto1	atmosf	4
Texto2	clima	22
Texto2	mudanç	21
Texto2	econom	8
Texto2	efeit	7
Texto2	aument	6
Texto2	caus	6
Texto2	pais	6
Texto2	desenvolv	6
Texto2	impact	5
Texto2	mund	5
Texto2	prejuiz	5
Texto2	tend	4
Texto2	perd	4
Texto2	set	4
Texto2	adapt	4
Texto3	mudanç	16
Texto3	clima	14
Texto3	faun	7
Texto3	flor	7
Texto3	especi	6
Texto3	impact	5

[1] seção 4.3

Quadro 23 – palavras radicalizadas e contagem de frequência (continuação ...)

arquivo	**palavra radicalizada**	**frequência**
Texto3	Aumente	4
Texto3	Habitat	4
Texto3	Muit	4
Texto3	natur	4
Texto3	reduz	4
Texto4	clima	9
Texto4	mudanç	8
Texto4	pol	8
Texto4	degel	6
Texto4	mar	6
Texto4	impact	4
Texto4	gel	4
Texto5	mudanç	12
Texto5	clima	11
Texto5	impact	6
Texto6	clima	8
Texto6	energ	8
Texto6	dat	8
Texto6	cent	8
Texto6	mudanç	7
Texto6	aument	6
Texto6	ciberne	5
Texto6	impact	4
Texto6	comput	4
Texto6	glob	4
Texto6	consum	4
Texto6	efici	4

Fonte: autor

Anexos

ANEXO A – Parâmetros do Algoritmo Removedor de Sufixos da Língua Portuguesa (RSLP)

Seguem abaixo os quadros utilizados para controle e parametrização das transformações realizadas pelo algoritmo de stemming RSLP. Todos os dados constantes nos quadros abaixo foram obtidos a partir dos anexos de *A Stemming Algorithm for the Portuguese Language* (Orengo; Huck, 2001) e *Stemming para a língua portuguesa: estudo, análise e melhoria do algoritmo RSLP*(Coelho, 2007).

Quadro 24 – Redução Plural

Algoritmo RSLP: Redução Plural				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	ns	1	m	
02	ões	3	ão	
03	ães	1	ão	mães
04	ais	1	al	cais, mais
05	éis	2	el	
06	eis	2	el	
07	óis	2	ol	
08	is	2	il	lápis, cais, mais, crúcis, biquínis, pois, depois, dois, leis
09	les	3	l	
10	res	3	r	
11	s	2		aliás, pires, lápis, cais, mais, mas, menos, férias, fezes, pêsames, crúcis, gás, atrás, moisés, através, convés, ês, país, após, ambas, ambos, messias

Fonte: autor.

Quadro 25 – Redução do Feminino

Algoritmo RSLP: Redução do Feminino				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	ona	3	ão	abandona, lona, iona, cortisona, monótona, maratona, acetona, detona, carona
02	ora	3	or	
03	na	4	no	carona, abandona, lona, iona, cortisona, monótona, maratona, acetona, detona, guiana, campana, grana, caravana, banana, paisana
04	inha	3	inho	rainha, linha, minha
05	esa	3	ês	mesa, obesa, princesa, turquesa, ilesa, pesa, presa
06	osa	3	oso	mucosa, prosa
07	íaca	3	íaco	
08	ica	3	ico	dica
09	ada	2	ado	pitada
10	ida	3	ido	vida
11	ída	3	ido	recaída, saída, dúvida
12	ima	3	imo	vítima
13	iva	3	ivo	saliva, oliva
14	eira	3	eiro	beira, cadeira, frigideira, bandeira, feira, capoeira, barreira, fronteira, besteira, poeira
15	ã	2	ão	amanhã, arapuã, fã, divã

Fonte: autor.

Quadro 26 – Redução Adverbial

Algoritmo RSLP: Redução Adverbial				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	mente	4	experimente	

Fonte: autor.

Quadro 27 – Redução Aumentativo/Diminutivo

Algoritmo RSLP: Redução Aumentativo/Diminutivo				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	díssimo	5		
02	abilíssimo	5		
03	íssimo	3		
04	ésimo	3		
05	érrimo	4		
06	zinho	2		
07	quinho	4		
08	uinho	4		
09	adinho	3		
10	inho	3		caminho, cominho
11	alhão	4		
12	uça	4		
13	aço	4		antebraço
14	aça	4		
15	adão	4		
16	idão	4		
17	ázio	3		topázio
18	arraz	4		
19	zarrão	3		
20	arrão	4		
21	arra	3		
22	zão	2		coalizão
23	ão	3		camarão, chimarrão, canção, coração, embrião, grotão, glutão, ficção, fogão, feição, furacão, gamão, lampião, leão, macacão, nação, órfão, orgão, patrão, portão, quinhão, rincão, tração, falcão, espião, mamão, folião, cordão, aptidão, campeão, colchão, limão, leilão, melão, barão, milhão, bilhão, fusão, cristão, ilusão, capitão, estação, senão

Fonte: autor.

Quadro 28 – Redução do Sufixo Nominal

Algoritmo RSLP: Redução do Sufixo Nominal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	encialista	4		
02	alista	5		
03	agem	3		coragem, chantagem, vantagem, carruagem
04	iamento	4		
05	amento	3		firmamento, fundamento, departamento
06	imento	3		
07	mento	6		firmamento, elemento, complemento, instrumento, departamento
08	alizado	4		
09	atizado	4		
10	tizado	4		Alfabetizado
11	izado	5		organizado, pulverizado
12	ativo	4		pejorativo, relativo
13	tivo	4		relativo
14	ivo	4		passivo, possessivo, pejorativo, positivo
15	ado	2		grado
16	ido	3		cândido, consolido, rápido, decido, tímido, duvido, marido
17	ador	3		
18	edor	3		
19	idor	4		ouvidor
20	dor	4		ouvidor
21	sor	4		assessor
22	atoria	5		
23	tor	3		benfeitor, leitor, editor, pastor, produtor, promotor, consultor
24	or	2		motor, melhor, redor, rigor, sensor, tambor, tumor, assessor, benfeitor, pastor, terior, favor, autor
25	abilidade	5		

Quadro 28 – Redução do Sufixo Nominal (continuação ...)

Algoritmo RSLP: Redução do Sufixo Nominal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
26	icionista	4		
27	cionista	5		
28	ionista	5		
29	ionar	5		
30	ional	4		
31	ência	3		
32	ância	4		ambulância
33	edouro	3		
34	queiro	3	c	
35	adeiro	4		desfiladeiro
36	eiro	3		desfiladeiro, pioneiro, mosteiro
37	uoso	3		
38	oso	3		precioso
39	alizaç	5		
40	atizaç	5		
41	tizaç	5		
42	izaç	5		organizaç
43	aç	3		equaç, relaç
44	iç	3		eleição
45	ário	3		voluntário, salário, aniversário, diário, lionário, armário
46	atório	3		
47	rio	5		voluntário, salário, aniversário, diário, compulsório, lionário, próprio, stério, armário
48	ério	6		
49	ês	4		
50	eza	3		
51	ez	4		
52	esco	4		
53	ante	2		gigante, elefante, adiante, possante, instante, restaurante
54	ástico	4		eclesiástico
55	alístico	3		

Quadro 28 – Redução do Sufixo Nominal (continuação ...)

Algoritmo RSLP: Redução do Sufixo Nominal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
56	áutico	4		
57	êutico	4		
58	tico	3		político, eclesiástico, diagnostico, prático, doméstico, diagnóstico, idêntico, alopático, artístico, autêntico, eclético, crítico, critico
59	ico	4		tico, público, explico
60	ividade	5		
61	idade	4		autoridade, comunidade
62	oria	4		categoria
63	encial	5		
64	ista	4		
65	auta	5		
66	quice	4	c	
67	ice	4		cúmplice
68	íaco	3		
69	ente	4		freqüente, alimente, acrescente, permanente, oriente, aparente
70	ense	5		
71	inal	3		
72	ano	4		
73	ável	2		afável, razoável, potável, vulnerável
74	ível	3		possível
75	vel	5		possível, vulnerável, solúvel
76	bil	3		vel
77	ura	4		imatura, acupuntura, costura
78	ural	4		
79	ual	3		bissexual, virtual, visual, pontual
80	ial	3		
81	al	4		afinal, animal, estatal, bissexual, desleal, fiscal, formal, pessoal, liberal, postal, virtual, visual, pontual, sideral, sucursal

Quadro 28 – Redução do Sufixo Nominal (continuação . . .)

Algoritmo RSLP: Redução do Sufixo Nominal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
82	alismo	4		
83	ivismo	4		
84	ismo	3		cinismo

Fonte: autor.

Quadro 29 – Redução Sufixo Verbal

Algoritmo RSLP: Redução Sufixo Verbal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	aríamo	2		
02	ássemo	2		
03	eríamo	2		
04	êssemo	2		
05	iríamo	3		
06	íssemo	3		
07	áramo	2		
08	árei	2		
09	aremo	2		
10	ariam	2		
11	aríei	2		
12	ássei	2		
13	assem	2		
14	ávamo	2		
15	êramo	3		
16	eremo	3		
17	eriam	3		
18	eríei	3		
19	êssei	3		
20	essem	3		
21	íramo	3		
22	iremo	3		
23	iriam	3		

Quadro 29 – Redução Sufixo Verbal (continuação ...)

Algoritmo RSLP: Redução Sufixo Verbal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
24	iríei	3		
25	íssei	3		
26	issem	3		
27	ando	2		
28	endo	3		
29	indo	3		
30	ondo	3		
31	aram	2		
32	arão	2		
33	arde	2		
34	arei	2		
35	arem	2		
36	aria	2		
37	armo	2		
38	asse	2		
39	aste	2		
40	avam	2		agravam
41	ávei	2		
42	eram	3		
43	erão	3		
44	erde	3		
45	erei	3		
46	êrei	3		
47	erem	3		
48	eria	3		
49	ermo	3		
50	esse	3		
51	este	3		faroeste, agreste
52	íamo	3		
53	iram	3		
54	íram	3		
55	irão	2		
56	irde	2		
57	irei	3		admirei

Quadro 29 – Redução Sufixo Verbal (continuação ...)

Algoritmo RSLP: Redução Sufixo Verbal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
58	irem	3		adquirem
59	iria	3		
60	irmo	3		
61	isse	3		
62	iste	4		
63	iava	4		ampliava
64	amo	2		
65	iona	3		
66	ara	2		arara, prepara
67	ará	2		alvará
68	are	2		prepare
69	ava	2		agrava
70	emo	2		
71	era	3		acelera, espera
72	erá	3		
73	ere	3		espere
74	iam	3		enfiam, ampliam, elogiam, ensaiam
75	íei	3		
76	imo	3		reprimo, intimo, íntimo, nimo, queimo, ximo
77	ira	3		fronteira, sátira
78	ído	3		
79	irá	0		
80	tizar	4		alfabetizar
81	izar	5		organizar
82	itar	5		acreditar, explicitar, estreitar
83	ire	3		adquire
84	omo	3		
85	ai	2		
86	am	2		
87	ear	4		alardear, nuclear
88	ar	2		azar, bazar, patamar
89	uei	3		

Quadro 29 – Redução Sufixo Verbal (continuação ...)

Algoritmo RSLP: Redução Sufixo Verbal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
90	uía	5	u	
91	ei	3		
92	guem	3	g	
93	em	2		alem, virgem
94	er	2		éter, pier
95	eu	3		chapeu
96	ia	3		estória, fatia, acia, praia, elogia, mania, lábia, aprecia, polícia, arredia, cheia, ásia
97	ir	3		freir
98	iu	3		
99	eou	5		

Fonte: autor.

Quadro 30 – Remoção de Vogal

Algoritmo RSLP: Remoção de Vogal				
ordem execução	**sufixo original**	**tamanho mínimo da raiz**	**sufixo final**	**exceções**
01	bil	2	vel	
02	gue	2	g	gangue, jegue
03	á	3		
04	ê	3		bebê
05	a	3		ásia
06	e	3		
07	o	3		ão

Fonte: autor.

ANEXO B – Textos utilizados no experimento

Todos os textos abaixo foram produzidos por uma ferramenta de inteligência artificial denominada Bard[1].

Texto 1 – O impacto das mudanças climáticas e a produção de carbono

As mudanças climáticas são um dos maiores desafios que a humanidade enfrenta hoje. O aumento das temperaturas globais, o aumento do nível do mar e as mudanças nos padrões de precipitação estão tendo um impacto significativo no planeta.

Uma das principais causas das mudanças climáticas é a produção de carbono. O carbono é um elemento químico que está presente em todos os seres vivos. Quando os seres vivos morrem, seu corpo se decompõe e libera carbono na atmosfera. No entanto, a atividade humana também está liberando grandes quantidades de carbono na atmosfera. Isso ocorre principalmente devido à queima de combustíveis fósseis, como o carvão, o petróleo e o gás natural.

A queima de combustíveis fósseis libera dióxido de carbono (CO2) na atmosfera. O CO2 é um gás de efeito estufa, o que significa que ele retém o calor do sol na atmosfera. Isso faz com que a temperatura global aumente.

O aumento das temperaturas globais está causando uma série de problemas, como:

- Inundações, secas e outros eventos climáticos extremos
- Derretimento das calotas polares e aumento do nível do mar
- Alterações nos padrões de precipitação
- Perda de biodiversidade
- Aumento do risco de doenças
- Conflitos e instabilidade política

O aumento da produção de carbono é um dos principais desafios que a humanidade enfrenta hoje. É necessário reduzir as emissões de carbono para mitigar os impactos das mudanças climáticas.

Há uma série de coisas que podem ser feitas para reduzir as emissões de carbono, como:

[1] MANYIKA, James. **An overview of Bard**: an early experiment with generative ai. [S.l.], s.d. 9 p. Disponível em: <https://ai.google/static/documents/google-about-bard.pdf>. Acesso em: out. 2023.

- Investir em energias renováveis, como a energia solar e a eólica
- Melhorar a eficiência energética dos edifícios e dos veículos
- Reduzir o consumo de carne
- Plantar árvores
- Apoiar organizações que trabalham para combater as mudanças climáticas

Todos nós temos um papel a desempenhar na luta contra as mudanças climáticas. Ao tomarmos pequenas ações no nosso dia a dia, podemos fazer a diferença.

Texto 2 – O impacto das mudanças climáticas na economia mundial

As mudanças climáticas já estão tendo um impacto significativo na economia mundial e esse impacto só tende a aumentar no futuro. Os efeitos das mudanças climáticas já estão sendo sentidos em todo o mundo, em forma de eventos climáticos extremos, aumento do nível do mar, mudanças na produtividade agrícola e perda de biodiversidade. Esses efeitos estão causando prejuízos econômicos em uma ampla gama de setores, incluindo agricultura, turismo, seguros e infraestrutura.

Um estudo do Painel Intergovernamental sobre Mudanças Climáticas (IPCC) estima que as mudanças climáticas já estão causando perdas econômicas globais de até US$ 500 bilhões por ano. Esse valor é esperado para aumentar para US$ 2 trilhões por ano até 2050, se nada for feito para mitigar as mudanças climáticas.

Os países em desenvolvimento são os mais vulneráveis aos efeitos das mudanças climáticas. Isso ocorre porque eles têm menos recursos para se adaptar aos efeitos das mudanças climáticas e são mais dependentes de setores que são vulneráveis às mudanças climáticas, como agricultura e turismo.

Os países desenvolvidos também estão sendo afetados pelas mudanças climáticas, mas em menor grau do que os países em desenvolvimento. Isso ocorre porque os países desenvolvidos têm mais recursos para se adaptar às mudanças climáticas e são menos dependentes de setores que são vulneráveis às mudanças climáticas.

As mudanças climáticas são uma ameaça significativa à economia mundial. Se nada for feito para mitigar as mudanças climáticas, os prejuízos econômicos só tendem a aumentar no futuro. Os países precisam trabalhar juntos para reduzir as emissões de gases de efeito estufa e desenvolver estratégias de adaptação às mudanças climáticas.

Aqui estão alguns exemplos específicos do impacto das mudanças climáticas na economia mundial:

- Eventos climáticos extremos, como furacões, tempestades e secas, estão causando prejuízos bilhões de dólares em danos à propriedade e infraestrutura.
- O aumento do nível do mar está inundando áreas costeiras, causando prejuízos à agricultura, turismo e pesca.
- As mudanças climáticas estão afetando a produtividade agrícola, levando a uma redução na produção de alimentos.
- As mudanças climáticas estão causando a perda de biodiversidade, o que está levando ao colapso de ecossistemas e perda de empregos.

Os efeitos das mudanças climáticas estão causando prejuízos econômicos em uma ampla gama de setores e esse impacto só tende a aumentar no futuro. Os países precisam trabalhar juntos para reduzir as emissões de gases de efeito estufa e desenvolver estratégias de adaptação às mudanças climáticas.

Texto 3 – O impacto das mudanças climáticas na fauna e flora do mundo

As mudanças climáticas estão tendo um impacto significativo na fauna e flora do mundo. O aumento da temperatura global, o aumento do nível do mar e as mudanças nos padrões de precipitação estão levando à perda de habitat, à diminuição da biodiversidade e ao aumento das doenças.

Algumas das principais consequências das mudanças climáticas para a fauna e flora incluem:

- Extinção de espécies: muitas espécies não conseguem se adaptar às mudanças climáticas e estão entrando em extinção. Um estudo da União Internacional para a Conservação da Natureza (IUCN) estima que 1 milhão de espécies estão ameaçadas de extinção, e muitas delas estão sendo ameaçadas pelas mudanças climáticas.
- Mudanças nos hábitos alimentares e reprodutivos: as mudanças climáticas estão forçando muitas espécies a mudar seus hábitos alimentares e reprodutivos. Por exemplo, algumas espécies de aves estão migrando para latitudes mais altas para encontrar alimentos, e outras estão se reproduzindo mais cedo no ano.
- Doenças: as mudanças climáticas estão facilitando a propagação de doenças entre as plantas e os animais. Por exemplo, o fungo que causa a ferrugem da soja está se espalhando para novas áreas devido ao aumento da temperatura global.

As mudanças climáticas são uma ameaça significativa para a fauna e flora do mundo. Se nada for feito para mitigar as mudanças climáticas, o impacto sobre a biodiversidade será devastador.

Aqui estão algumas ações que podem ser tomadas para reduzir o impacto das mudanças climáticas na fauna e flora:

- Reduzir as emissões de gases de efeito estufa: a principal causa das mudanças climáticas é a queima de combustíveis fósseis, que liberam gases de efeito estufa na atmosfera. Reduzir as emissões de gases de efeito estufa é a melhor maneira de mitigar as mudanças climáticas.
- Proteger os habitats naturais: os habitats naturais são importantes para a sobrevivência de muitas espécies. Proteger os habitats naturais é fundamental para reduzir o impacto das mudanças climáticas na fauna e flora.
- Investir em pesquisa e desenvolvimento: é importante investir em pesquisa e desenvolvimento para encontrar novas maneiras de mitigar as mudanças climáticas e proteger a fauna e flora.

Texto 4 – O impacto das mudanças climáticas no degelo polar

As mudanças climáticas estão tendo um impacto significativo no degelo polar. O aumento da temperatura global está fazendo com que as calotas polares e o gelo marinho derretam em um ritmo acelerado. Isso está causando uma série de problemas, incluindo o aumento do nível do mar, a mudança dos padrões climáticos e a perda de habitat para animais e plantas.

O aumento do nível do mar é uma das consequências mais preocupantes do degelo polar. Quando o gelo derrete, ele libera água na atmosfera, que faz com que o nível do mar suba. Isso está causando inundações em áreas costeiras, acúmulo de sal no solo e a destruição de ecossistemas marinhos.

A mudança dos padrões climáticos é outra consequência do degelo polar. Quando o gelo derrete, ele deixa a superfície mais exposta à luz solar. Isso está causando o aquecimento dos oceanos e da atmosfera, que está levando a uma série de eventos climáticos extremos, como ondas de calor, secas, furacões e tufões.

A perda de habitat para animais e plantas é outra consequência do degelo polar. As calotas polares e o gelo marinho são o lar de uma grande variedade de espécies, que estão sendo forçadas a se adaptar ou a migrar para outras áreas. Isso está levando à perda de biodiversidade e à extinção de espécies.

O degelo polar é um problema global que exige uma resposta global. Os países precisam trabalhar juntos para reduzir as emissões de gases de efeito estufa e para mitigar os impactos das mudanças climáticas.

Aqui estão algumas ações que podem ser tomadas para reduzir as emissões de gases de efeito estufa e para mitigar os impactos das mudanças climáticas:

- Investir em energias renováveis, como a energia solar e a eólica.
- Melhorar a eficiência energética dos edifícios e dos veículos.
- Reduzir o consumo de carne.
- Plantar árvores.
- Apoiar organizações que trabalham para combater as mudanças climáticas.

Todos nós temos um papel a desempenhar na luta contra as mudanças climáticas. Ao tomarmos pequenas ações no nosso dia a dia, podemos fazer a diferença.

Texto 5 – O impacto das mudanças climáticas na vida do ser humano

As mudanças climáticas estão tendo um impacto significativo na vida do ser humano. O aumento da temperatura global, o aumento do nível do mar e as mudanças nos padrões de precipitação estão causando uma série de problemas, incluindo:

- Inundações, secas e outras eventos climáticos extremos
- Perda de habitat para animais e plantas
- Aumento da incidência de doenças
- Alterações na produção agrícola e pecuária
- Migração forçada de pessoas
- Conflitos e instabilidade política

Os impactos das mudanças climáticas são sentidos de forma mais aguda em países pobres e vulneráveis. Esses países têm menos recursos para se adaptar às mudanças climáticas e estão mais expostos aos riscos.

As mudanças climáticas são um problema global que exige uma resposta global. Os países precisam trabalhar juntos para reduzir as emissões de gases de efeito estufa e para mitigar os impactos das mudanças climáticas.

Aqui estão algumas ações que podem ser tomadas para reduzir as emissões de gases de efeito estufa e para mitigar os impactos das mudanças climáticas:

- Investir em energias renováveis, como a energia solar e a eólica.
- Melhorar a eficiência energética dos edifícios e dos veículos.

- Reduzir o consumo de carne.
- Plantar árvores.
- Apoiar organizações que trabalham para combater as mudanças climáticas.

Todos nós temos um papel a desempenhar na luta contra as mudanças climáticas. Ao tomarmos pequenas ações no nosso dia a dia, podemos fazer a diferença.

Além das ações mencionadas acima, é importante também conscientizar a população sobre as mudanças climáticas e seus impactos na vida do ser humano. A educação ambiental é uma ferramenta essencial para a mudança de comportamento e para a construção de um futuro mais sustentável.

Texto 6 – O impacto das mudanças climáticas na computação

As mudanças climáticas estão tendo um impacto significativo na computação. O aumento da temperatura global, o aumento do nível do mar e as mudanças nos padrões de precipitação estão causando uma série de problemas, incluindo:

- Aumento do consumo de energia: os data centers são um dos maiores consumidores de energia do mundo. O aumento da temperatura global está fazendo com que os data centers precisem consumir mais energia para manter seus equipamentos funcionando. Isso está levando a um aumento nas emissões de gases de efeito estufa.
- Perda de infraestrutura: as mudanças climáticas estão causando eventos climáticos extremos, como furacões, tufões e inundações. Esses eventos podem danificar ou destruir data centers, o que pode levar à perda de dados e serviços.
- Aumento da vulnerabilidade: as mudanças climáticas estão tornando os data centers mais vulneráveis a ataques cibernéticos. Isso ocorre porque os data centers estão localizados em áreas que são mais propensas a eventos climáticos extremos.

As mudanças climáticas são um problema global que exige uma resposta global. A indústria de computação precisa trabalhar para reduzir seu impacto ambiental. Algumas das ações que podem ser tomadas para reduzir o impacto ambiental da computação incluem:

Investir em energias renováveis: os data centers podem ser alimentados por energias renováveis, como a energia solar e a eólica. Isso ajudaria a reduzir as emissões de gases de efeito estufa.

Melhorar a eficiência energética: os data centers podem ser mais eficientes no consumo de energia. Isso pode ser feito através do uso de equipamentos mais eficientes, da

implementação de medidas de otimização de energia e do uso de técnicas de resfriamento mais eficientes.

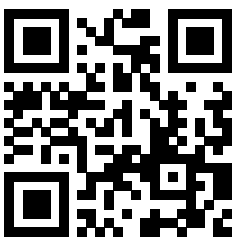
Investir em segurança cibernética: os data centers podem ser mais seguros contra ataques cibernéticos. Isso pode ser feito através do uso de medidas de segurança cibernética mais avançadas e do treinamento dos funcionários em segurança cibernética.

Todos nós temos um papel a desempenhar na luta contra as mudanças climáticas. Ao tomarmos pequenas ações no nosso dia a dia, podemos fazer a diferença.

ANEXO C – *Stopwords* utilizadas no experimento

- **A** — a, à, às, acerca, adeus, agora, ainda, além, algumas, algo, algumas, alguns, ali, além, ambas, ambos, ano, anos, antes, ao, aonde, aos, apenas, apoio, apontar, apos, após, aquela, aquelas, aquele, aqueles, aqui, aquilo, área, as, assim, através, atrás, até, aí
- **B** — baixo, bastante, bem, boa, boas, bom, bons, breve,
- **C** — cada, caminho, catorze, cedo, cento, certamente, certeza, cima, cinco, coisa, com, como, comprido, conhecido, conselho, contra, contudo, corrente, cuja, cujas, cujo, cujos, custa, cá,
- **D** — da, daquela, daquelas, daquele, daqueles, dar, das, de, debaixo, dela, delas, dele, deles, demais, dentro, depois, desde, desligado, dessa, dessas, desse, desses, desta, destas, deste, destes, deve, devem, deverá, dez, dezanove, dezessete, dezoito, dia, diante, direita, dispoe, dispoem, diversa, diversas, diversos, diz, dizem, dizer, do, dois, dos, doze, duas, durante, dá, dão, dúvida,
- **E** — e, é, és, ela, elas, ele, eles, em, embora, enquanto, entao, entre, então, era, eram, éramos, essa, essas, esse, esses, esta, estado, estamos, estar, estará, estas, estava, estavam, este, esteja, estejam, estejamos, estes, esteve, estive, estivemos, estiver, estivera, estiveram, estiverem, estivermos, estivesse, estivessem, estiveste, estivestes, estivéramos, estivéssemos, estou, está, estás, estávamos, estão, eu, exemplo,
- **F** — falta, fará, favor, faz, fazeis, fazem, fazemos, fazer, fazes, fazia, faço, fez, fim, final, foi, fomos, for, fora, foram, forem, forma, formos, fosse, fossem, foste, fostes, fui, fôramos, fôssemos,
- **G** — geral, grande, grandes, grupo,
- **H** — ha, haja, hajam, hajamos, havemos, havia, hei, hoje, hora, horas, houve, houvemos, houver, houvera, houveram, houverei, houverem, houveremos, houveria, houveriam, houvermos, houverá, houverão, houveríamos, houvesse, houvessem, houvéramos, houvéssemos, há, hão,
- **I** — iniciar, inicio, ir, irá, isso, isto,
- **J** — já,
- **L** — lado, lhe, lhes, ligado, local, logo, longe, lugar, lá,
- **M** — maior, maioria, maiorias, mais, mal, mas, me, mediante, meio, menor, menos, meses, mesma, mesmas, mesmo, mesmos, meu, meus, mil, minha, minhas, momento, muito, muitos, máximo, mês,

- **N** — na, nada, nao, naquela, naquelas, naquele, naqueles, nas, nem, nenhuma, nessa, nessas, nesse, nesses, nesta, nestas, neste, nestes, no, noite, nome, nos, nossa, nossas, nosso, nossos, nova, novas, nove, novo, novos, num, numa, numas, nunca, nuns, não, nível, nós, número,
- **O** — o, obra, obrigada, obrigado, oitava, oitavo, oito, onde, ontem, onze, os, ou, outra, outras, outro, outros,
- **P** — para, parece, parte, partir, poucas, pegar, pela, pelas, pelo, pelos, perante, perto, pessoas, pode, podem, poder, poderá, podia, pois, ponto, pontos, por, porque, porquê, portanto, posição, possivelmente, posso, possível, pouca, pouco, poucos, povo, primeira, primeiras, primeiro, primeiros, proprio, proprios, própria, próprias, próprio, próprios, próxima, próximas, próximo, próximos, puderam, pôde, põe, põem,
- **Q** — quais, qual, qualquer, quando, quanto, quarta, quarto, quatro, que, quem, quer, quereis, querem, queres, quero, questão, quieto, quinta, quinto, quinze, quais, quê,
- **R** — relação,
- **S** — sabe, sabem, saber, se, segunda, segundo, sei, seis, seja, sejam, sejamos, sem, sempre, sendo, ser, serei, seremos, seria, seriam, será, serão, seríamos, sete, seu, seus, sexta, sexto, sim, sistema, sob, sobre, sois, somente, somos, sou, sua, suas, são, sétima, sétimo, só,
- **T** — tal, talvez, tambem, também, tanta, tantas, tanto, tarde, te, tem, temos, tempo, tendes, tenha, tenham, tenhamos, tenho, tens, tentar, tentaram, tente, tentei, ter, terceira, terceiro, terei, teremos, teria, teriam, terá, terão, teríamos, teu, teus, teve, tinha, tinham, tipo, tive, tivemos, tiver, tivera, tiveram, tiverem, tivermos, tivesse, tivessem, tiveste, tivestes, tivéramos, tivéssemos, toda, todas, todo, todos, trabalhar, trabalho, treze, três, tu, tua, tuas, tudo, tão, tém, têm, tínhamos,
- **U** — último, um, uma, umas, uns, usa, usar,
- **V** — vai, vais, valor, veja, vem, vens, ver, verdade, verdadeiro, vez, vezes, viagem, vindo, vinte, você, vocês, vos, vossa, vossas, vosso, vossos, vários, vão, vêm, vós,
- **Z** — zero.



<http://www.janaite.net>