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

Claudio Roberto de Oliveira Conceição

A UTILIZAÇÃO DE *CRAWLER* FOCADO E TÉCNICAS DE *CLUSTERING* COMO RECURSOS COMPLEMENTARES NO PROCESSO DE RECUPERAÇÃO DE INFORMAÇÃO

MARÍLIA – SP
2022

Claudio Roberto de Oliveira Conceição

A UTILIZAÇÃO DE *CRAWLER* FOCADO E TÉCNICAS DE *CLUSTERING* COMO RECURSOS COMPLEMENTARES NO PROCESSO DE RECUPERAÇÃO DE INFORMAÇÃO

Dissertação de Mestrado 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 Mestre em Ciência da Informação.

ÁREA DE CONCENTRAÇÃO: Informação, Tecnologia e Conhecimento.

LINHA DE PESQUISA: Informação e Tecnologia.

ORIENTADOR: Prof. Dr. Edberto Ferneda

MARÍLIA – SP
2022

Conceição, Claudio Roberto de Oliveira.
C744 A Utilização de Crawler Focado e Técnicas de Clustering como Recursos Complementares no Processo de Recuperação de Informação / Claudio Roberto de Oliveira Conceição. — Marília, 2022.
72 f. : 30 cm.

Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2022.
Bibliografia: f. 66-72
Orientador: Prof. Dr. Edberto Ferneda.

1. Recuperação de Informação 2. Mecanismos e Busca 3. Web Crawler 4. Clustering. I. Título.

CDD 029.7.

Claudio Roberto de Oliveira Conceição

A UTILIZAÇÃO DE *CRAWLER* FOCADO E TÉCNICAS DE *CLUSTERING* COMO RECURSOS COMPLEMENTARES NO PROCESSO DE RECUPERAÇÃO DE INFORMAÇÃO

Dissertação para obtenção do título de Mestre no Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências, Universidade Estadual Paulista - UNESP - Campus de Marília, na área de concentração Produção e Organização da Informação

BANCA EXAMINADORA

Prof. Dr. Edberto Ferneda (Orientador)
Faculdade de Filosofia e Ciências (FFC)
Universidade Estadual Paulista (UNESP) – Campus de Marília

Prof. Dr. Fabrício Baptista (Membro externo)
Instituto Federal do Paraná (IFPR) – Campus Jacarezinho

Profa. Dra. Rachel Cristina Vesu Alves (Membro interno)
Faculdade de Filosofia e Ciências (FFC)
Universidade Estadual Paulista (UNESP) – Campus de Marília

Marília, 11 de maio de 2022

Dedicatória

Ao meu irmão, Pedro, que me incentivou a voltar estudar depois dos 38 anos de idade.

À minha amada mãe que sempre acreditou em minha força de vontade.

À minha família, esposa, e filhas, que sempre me apoiaram.

Agradecimentos

À Deus pela dádiva da vida, saúde e paz.

Ao meu orientador, Professor Dr. Edberto Ferneda, agradeço por ser uma pessoa doce, de coração enorme, pela sua clareza e domínio nos ensinamentos, e brilho que emana ao exercer sua dadiva de ensinar.

Aos membros da banca examinadora Profa. Rachel Cristina Vesu Alves (UNESP) e Prof. Dr. Fabrício Baptista (IFPR), que gentilmente aceitaram participar e colaborar com essa dissertação, agradeço pelas importantíssimas breves conversas.

À minha família, esposa e filhas, que sempre me acompanham com incentivos e admiração pelos meus esforços.

À minha amada esposa Eliete, que cuida de mim e me levanta nos momentos de desânimo me traz a tona a capacidade de seguir que está dentro de mim.

À CAPES, o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito.
Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes”

(Marthin Luther King)

Resumo

Em vista da rápida expansão e dinamicidade da Web, os mecanismos de busca assumiram um papel essencial para a recuperação de informação nesse imenso repositório. Com o grande número de páginas sendo constantemente adicionados e modificadas, a eficiência dos mecanismos de busca torna-se fundamental. Um *crawler* é o elemento principal de um mecanismo de busca. Sua função é navegar pela estrutura hipertextual da Web de forma sistemática afim obter e indexar páginas, formando um acervo documental utilizados pelo mecanismo de busca. Os Web *Crawlers* de propósito geral, utilizados pelos mecanismos de busca como o Google e o Bing, funcionam exaustivamente, procurando coletar e indexar o maior número de documentos possível. Um Web *crawler* focado é um tipo de *crawler* que coleta páginas contendo informações sobre um determinado tema ou assunto, gerando um conjunto de documentos qualificado e contextualizado, permitindo aumentar a eficiência de um mecanismo de busca. Este trabalho propõe a utilização de Web *crawlers* focados juntamente com técnicas de *clustering*. Técnicas de *clustering* (agrupamento) têm sido usadas na recuperação de informações para muitos propósitos diferentes tais como expansão de consulta, agrupamento de documentos, indexação de documentos e visualização de resultados de busca. A partir de uma pesquisa exploratória e descritiva, fundamentada em bibliografia específica, este trabalho propõe a utilização conjunta de Web *crawler* focado e técnicas de *clustering* como recursos complementares no processo de recuperação de informação. Inicialmente o *crawler* focado fornece um conjunto de documentos (páginas Web) restrito a um assunto ou tema. A partir do corpus temático fornecido pelo *crawler*, o processo classificatório dos algoritmos de *Clustering* podem então gerar grupos (*clusters*) de documentos relacionados às especificidades ou detalhamentos do tema. A base teórica apresentada neste estudo possui o potencial de tornar-se uma proposta para a implementação um mecanismo de busca experimental, demonstrando a sua aplicabilidade e contribuindo para o campo de pesquisa.

Palavras-Chave: Recuperação de Informação, Mecanismos e Busca, Web Crawler, Clustering.

Abstract

In view of rapid expansion and dynamism of the Web, search engines have assumed an essential role in the retrieval of information in this immense repository. With the large number of pages constantly being added and modified, the efficiency of search engines becomes critical. A crawler is the core element of a search engine. Its function is to systematically navigate the hypertextual structure of the Web in order to obtain and index pages, forming a collection of documents used by the search engine. General purpose Web crawlers, used by search engines like Google and Bing, work extensively, seeking to collect and index as many documents as possible. A focused Web crawler is a type of crawler that collects pages containing information about a certain theme or subject, generating a set of qualified and contextualized documents, allowing to increase the efficiency of a search engine. This work proposes the use of focused Web crawlers together with clustering techniques. Clustering techniques have been used in information retrieval for many different purposes such as query expansion, document grouping, document indexing, and viewing search results. From an exploratory and descriptive research, based on specific bibliography, this work proposes the joint use of focused Web crawler and Clustering techniques as complementary resources in the information retrieval process. Initially Focused Crawler provides a set of documents (Webpages) restricted to a subject or theme. From the thematic corpus provided by Crawler, the classification process of the Clustering algorithms can then generate groups of documents related to the specifics or details of the theme. The theoretical basis presented in this study has the potential to become a proposal for the implementation of an experimental search engine, demonstrating its applicability and contributing to the research field.

Keywords: Information Retrieval, Search Engines, Web Crawler, Clustering

Lista de Figuras

Sumário

1. Introdução

A demanda por soluções para o problema do tratamento da informação, evidenciado pela rápida expansão do volume de publicações científicas ocorrida após a Segunda Guerra, deu origem às pesquisas em *Information Retrieval* (Recuperação da Informação). Métodos e técnicas computacionais de recuperação de informação vêm sendo desenvolvidos desde o final da década de 1950. Porém, a partir de 1990 com o surgimento e a rápida expansão da Web, tornaram-se urgentes as pesquisas por métodos que possibilitassem uma recuperação eficiente e eficaz nesse imenso repositório de informação. O surgimento da Web proporcionou um rápido direcionamento nos esforços de pesquisa para os problemas relacionados à recuperação de informação. Porém, essa urgência para o desenvolvimento de soluções para o ambiente Web não deve prescindir das ideias, teorias e técnicas que foram desenvolvidas durante mais de 50 anos de pesquisas em Recuperação de Informação, anteriores à sua criação.

De acordo com o relatório "*Tendências da Internet*" da Bond Capital[1] , o número de usuários da Internet chegou a cerca de cerca de 3,8 bilhões em 2018, o que representava mais da metade da população mundial. Considerando esse número de usuários, pode-se ter uma noção do grande volume de informações que é disponibilizado diariamente na Internet. Essa quantidade de informações em constante crescimento e mutação faz aumentar a importância do aperfeiçoamento de técnicas para o tratamento, organização e recuperação das informações disponíveis na Web.

[1] https://www.bondcap.com/#internettrends

Diferentemente das bibliotecas, as páginas Web não estão descritas ou classificadas segundo um padrão determinado. Nesse ambiente, encontrar uma determinada informação depende principalmente da eficiência das ferramentas de busca. Essa eficiência é em grande parte dependente de seus *crawlers* (rastreadores) para a coleta e indexação de páginas Web. Também conhecidos como ***s**pider* ou *bot*, um *crawler* é um programa de computador que navega pela estrutura hipertextual da Web de maneira metódica e automatizada, obtendo documentos e seguindo links de uma página à outra (DHENAKARAN; SAMBANTHAN, 2011).

Kausar, Dhaka e Singh (2013) listam três principais técnicas usadas no desenvolvimento de Web *Crawlers*: *crawler* de propósito geral, *crawler* focado e *crawler* distribuído. Os Web *Crawlers* de propósito geral, utilizados por mecanismos de busca como o Google e o Bing, iniciam o seu rastreamento a partir de um ou mais endereços de páginas (URLs), indexam as páginas associadas a essas URLs, extraem todos os hiperlinks contidos nelas e continuam recursivamente esse processo. Esses *crawlers* funcionam exaustivamente, vasculhando e indexando a Web, procurando coletar o maior número de documentos possíveis (OLSTON; NAJORK, 2010). *Web crawler* focado é um tipo de *crawler s*eletivo, que coleta links considerados relevantes para um determinado tópico ou assunto previamente estabelecido. Baseado em contexto, os *Crawlers* focados utiliza o texto ao redor de cada link para classificar a página apontada pela URL (CALISKAN; OZCAN, 2013). Um Web *crawler* distribuído é caracterizada pela presença de muitos processos para a realização do rastreamento. Um servidor central comanda e sincroniza os *crawlers* enquanto estão geograficamente dispersos em máquinas e talvez regiões diferentes, acessando a internet e navegando através de URLs.

Este trabalho propõe a utilização de Web *crawlers* focados juntamente com técnicas de *Clustering* de documentos a fim de restringir a abrangência temática da base documental de um mecanismo de busca, com o objetivo de melhorar a precisão dos resultados obtidos.

Técnicas de *Clustering* de documentos são geralmente utilizadas para agrupar documentos semelhantes em classes não conhecidas *a priori*. Segundo Van Rijsbergen (1979), "dois documentos similares têm maior probabilidade de serem relevantes para uma mesma pesquisa". Portanto, documentos em um mesmo grupo (*cluster*) possuem grau de relevância semelhantes em relação à uma determinada necessidade de informação.

Técnicas de *clustering* têm sido usadas na recuperação de informações para muitos propósitos diferentes, como expansão de consulta, agrupamento e indexação de documentos, visualização de resultados de busca.

Neste trabalho propõe-se a utilização de técnicas de *clustering* em um conjunto de documentos relacionados a um determinado tema ou assunto, resultante do rastreamento realizado por um ou mais Web *crawlers* focados. Nesse contexto, chegamos ao seguinte problema de pesquisa: ***Como os algoritmos de Crawler Focado e técnicas de Clustering podem operar em conjunto a fim de prover maior eficácia e eficiência aos sistemas de recuperação de informação no ambiente Web?***

1.1 Hipótese de Pesquisa

Crawler focado e técnicas de *clustering* podem ser utilizados conjuntamente como recursos complementares no processo de recuperação de informação. Inicialmente o *crawler* focado fornece um conjunto de documentos (páginas Web) restrito a um assunto ou tema. A partir do *corpus* temático fornecido pelo *crawler focado*, o processo classificatório das técnicas de *clustering* pode então gerar grupos (*clusters*) de documentos relacionados às especificidades ou detalhamentos do tema, colaborando para uma recuperação da informação mais eficiente, fornecendo resultados mais relevantes e precisos.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo geral deste trabalho é propor formas de utilização de Web *crawlers* focados conjuntamente com técnicas de *clustering* a fim de demonstrar as vantagens para sistemas de recuperação de informação na busca por informações relevantes.

1.2.2 Objetivos Específicos

a. abordar questões sobre recuperação da informação, particularmente no ambiente Web;

b. apresentar o processo e as tecnologias utilizadas no desenvolvimento dos mecanismos de busca;

c. apresentar e analisar o funcionamento de Web *crawlers*, especificamente do Web *crawler* focado;

d. apresentar as técnicas de *clustering*;

e. demonstrar tecnologias envolvidas no desenvolvimento de Web *crawlers* focados com as técnicas de *clustering* e suas vantagens para recuperação de informação.

1.3 Procedimentos Metodológicos

Os procedimentos metodológicos utilizados no presente trabalho caracterizam-na quanto à sua natureza como uma pesquisa exploratória e descritiva. O caráter exploratório teve o papel de proporcionar uma visão panorâmica da temática, com novas percepções sobre o objeto de pesquisa (Web *crawlers* focados e técnicas de *clustering*) e estudar as formas de utilização conjunta dessas técnicas no intuito de contribuir para o esclarecimento dos objetivos propostos. O caráter descritivo da pesquisa buscou expor as diversidades de variações que formalizam o objeto de estudo dessa pesquisa. Trata-se de uma pesquisa qualitativa, quanto a abordagem do problema, que irá buscar, analisar e demonstrar as contribuições do uso dessas ferramentas e o quanto pode ser vantajoso (ou não), o uso colaborativo dos *crawlers* focados com técnicas de *clustering* no processo de busca e recuperação da informação.

Em relação aos procedimentos técnicos e metodológicos para coleta de dados, a pesquisa caracteriza-se como bibliográfica (SILVA; MENEZES, 2001). Dessa forma, o levantamento bibliográfico foi realizado a partir dos seguintes critérios:

- **Temas de pesquisa**: recuperação de informação através de mecanismos de busca, Web *crawler* focados, rastreadores focados, rastreadores focados e Web semântica, métodos e técnicas de *clustering*;
- **Tipos de materiais bibliográficos**: livros, periódicos, anais de congressos, teses e dissertações, sites especializados sobre o tema, base de dados nacionais e internacionais;

- **Em relação à seleção de materiais**: foi considerado os idiomas inglês e português; os principais autores que se destacam na área temática do trabalho; documentos dos últimos 10 anos, com exceção de alguns documentos clássicos até 20 ou mais.
- **Levantamento bibliográfico**: busca de materiais bibliográficos relacionados ao tema de pesquisa, de acordo com os critérios apontados anteriormente;
- **Leitura e análise dos textos**: após a seleção inicial dos textos foi realizada a leitura dos materiais e uma análise das principais características identificadas na literatura, para criar o referencial teórico da pesquisa e esclarecer o problema estabelecido;
- **Organização sistêmica do estudo exploratório**: agrupar características e fenômenos identificados na literatura para esclarecer o problema de pesquisa e atingir os resultados por meio dos objetivos propostos;
- **Elaboração do texto da dissertação** onde foram reunidas, analisadas e interpretadas as informações provenientes e evidenciadas pelas literaturas sobre o tema e sua sistematização sob coordenação do orientador.

1.4 Justificativa

Com o crescimento da Web, a busca por informações relevantes está se tornando uma tarefa cada vez mais árdua, representando um desafio para os mecanismos de busca que utilizam os *crawlers* de uso geral.

Os Web *crawlers* são um dos principais componentes dos mecanismos de busca. São os responsáveis por criam um acervo de páginas Web e sua indexação, permitindo que os usuários façam busca e encontrem as páginas que correspondem às suas necessidades de informação. O principal objetivo de um Web *crawler* é reunir de forma rápida e eficaz o maior número possível de páginas da Web, juntamente a estrutura de links que as interconecta (DHENAKARAN; SAMBANTHAN, 2011).

Devido à expansão da Web, nenhum mecanismo de pesquisa é capaz de indexar mais de um terço de toda a Web. Essa é a principal razão para os Web *Crawlers* de uso geral terem um baixo desempenho (PANT; SRINIVASAN, 2005). Assim, advoga-se que é necessário dar mais

ênfase na utilização de *Web crawlers* focados, pois esse modelo usa uma técnica de rastreamento que navega dinamicamente na internet escolhendo modos específicos que maximizam a probabilidade de descobrir páginas relevantes a partir de uma consulta de pesquisa específica pelo usuário. Prever a relevância do documento antes de ver seu conteúdo é uma das características centrais do *crawler* focado porque pode economizar uma quantidade significativa de recursos, tanto físicos como de software. (CHAKRABARTI; VAN DEN BERG; DOM, 1999).

Por essa razão justifica-se a presente pesquisa que abordará também técnicas de *clustering*, como essas técnicas podem trazer contribuições quando usadas em conjunto com algoritmos de Web *crawler* focado, demonstrando assim quais vantagens esse conjunto de técnicas pode oferecer para melhorar o Web *crawler* focado para extração de dados. Serão demonstrados os resultados obtidos sobre funções da união dessas técnicas no estudo apresentado, e se esse processo pode ser utilizado sem prejudicar a estrutura e informações das páginas dos sites, tornando sua utilização importante no campo da Recuperação de Informação. Além disso, essas informações que irão compor os resultados, poderão ser anonimamente utilizadas para estudos, estatísticas, documentações e normas que poderão contribuir para o campo da Ciência da Informação.

1.5 Terminologia adotada

Embora os termos “Web” e “Internet” refiram-se a conceitos distintos, comumente tais termos são utilizados como sinônimos. Utilizamos o termo “Web” para referenciarmos à *World Wide Web* e deixamos o termo “Internet” para referenciar a infraestrutura tecnológica criada a partir do final dos anos de 1950, que permitiu a criação e o desenvolvimento da Web na década de 1990.

O termo “recuperação de informação” pode referir-se tanto a um processo como também a uma área de pesquisa. Para diferenciarmos, convencionou-se que quando se tratar de uma área de pesquisa a grafia será feita em letras maiúsculas: “Recuperação de Informação”. Quando referirmos ao processo utilizamos todas as letras minúsculas: “recuperação de informação”.

Os termos "mecanismo de busca", "ferramenta de busca" ou "buscador" são considerados sinônimos. Podem ser definidos como um tipo sistema de recuperação de informação desenvolvido para tratar informação disponibilizada em ambiente Web.

O tema principal deste trabalho, Recuperação de Informação, envolve pelo menos dois campos científicos: a Ciência da Informação e a Ciência da Computação. Surge daí problemas terminológicos decorrentes das diferentes nomenclaturas utilizadas para referenciar um mesmo conceito. Por vezes será utilizado o termo "rastreador" como tradução direta do termo em inglês "*crawler "*. Desta forma, quando se tratar do verbo que indica a ação de coletar documentos no ambiente Web, será utilizado preferencialmente o termo "rastreamento". No texto desse trabalho será dado preferência à utilização dos *cluster* e *clustering*, em inglês. O primeiro refere-se a um substantivo que designa um *grupo* de elementos ou objetos. O segundo (*clustering*) pode se referir tanto ao substantivo "agrupamento" como o ato (verbo) de agrupar um conjunto de objetos ou quaisquer elementos: agrupar, agrupando, etc.

Neste trabalho utilizaremos o termo “sistema de recuperação de informação” para nos referirmos aos sistemas desenvolvidos antes do surgimento da Web. Utilizamos os termos “motor de busca”, “mecanismo de busca”, “ferramenta de busca”, “buscador”, “sistema de recuperação de informação Web” como sinônimos, para referenciar sistemas disponíveis no ambiente Web

1.6 Organização do Trabalho

Capítulo 1: Apresenta introdução, objetivos, justificativa e metodologia utilizados no trabalho;

Capítulo 2: Apresenta os conceitos da Recuperação da Informação e alguns componentes que essa área dispõe;

Capítulo 3: Apresenta os conceitos e as técnicas relacionados aos *Web crawlers* e Web *crawlers* focados: seus tipos, seu funcionamento e seus conceitos principais;

Capítulo 4: Apresenta a técnica de *clustering*, os conceitos que representam essa técnica, seu comportamento dentro do campo da Recuperação da Informação, abordando questões relevantes sobre sua importância para a qualidade no processo de recuperação de informação;

Capítulo 5: Nesse capitulo tratamos do assunto principal desse trabalho, abordando questões persistentes e inerentes ao processo de funcionamento do crawler focado e das técnicas de *clustering*, e por fim demonstramos a utilização dos mesmos trabalhando em conjunto, para

tentar aprimorar o processo de RI, apresentamos um conjunto de prováveis processos que poderão, no futuro, serem implementados e testados.

2.
Recuperação de Informação

O termo *Information Retrieval* (Recuperação de Informação) foi criado em 1951 por Calvin Mooers. Segundo Mooers, a recuperação de informação é um processo por meio do qual o usuário de um sistema consegue converter sua necessidade de informação em uma lista de citações de documentos. “Engloba tanto os aspectos intelectuais da descrição de informações e suas especificidades para a busca, como quaisquer sistemas, técnicas ou máquinas empregadas para o desempenho da operação” (MOOERS, 1951, p.25).

De uma forma mais ampla, a recuperação de Informação diz respeito a todas as atividades relacionadas com a organização, processamento e acesso à informação. Um sistema de recuperação de informação permite às pessoas se comunicarem com um sistema a fim de encontrarem as informações que necessitam. Na realidade, a maioria desses sistemas recuperam documentos nos quais pretensamente esteja a informação que o usuário necessita. Segundo Lancaster (1968), um sistema de recuperação de informação não informa o usuário diretamente sobre o assunto de sua indagação; ele apenas informa a existência (ou não) de documentos relacionados ao tema ou assunto de tal indagação.

Embora nascida no contexto da Ciência da Computação, para Saracevic (1996) a Recuperação de Informação pode ser considerada uma área interdisciplinaridade que possui uma relação bastante próxima com a Ciência da Informação. Enquanto a Ciência da Informação dedica-se à busca de “conhecimentos relacionados à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação, e utilização da informação” (BORKO,

1968, p.3), a Ciência da Computação atua no desenvolvimento de técnicas computacionais e algoritmos que viabilizam a aplicação desses conhecimentos.

Segundo Saracevic (1996, p.50):

> A Ciência da Computação trata de algoritmos que transformam informações enquanto a Ciência da Informação trata da natureza da informação e sua comunicação para uso pelos humanos. Ambos os objetos são interrelacionados e não competidores, mas complementares.

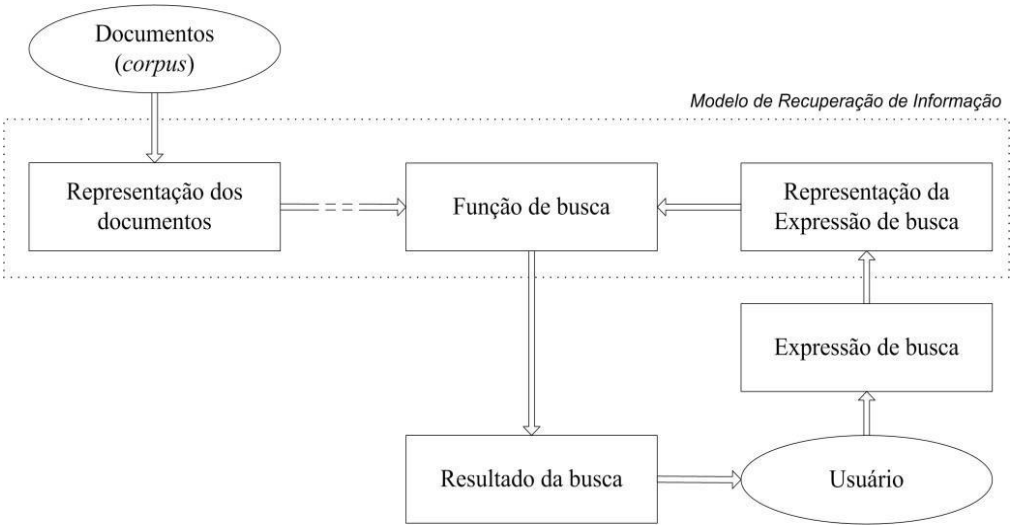
Teorias, métodos e técnicas de Recuperação de Informação vêm sendo desenvolvidas desde a década de 1950, com contribuições significativas de pesquisadores como Hans Peter Luhn, Eugene Garfield, Philip Bagley, Allen Kent, Gerard Salton, Karen Spark-Jones, Calvin Mooers, entre outros (BAEZA-YATES; RIBEIRO-NETO, 2013).

A expansão da Web, no início dos anos 1990, possibilitou um progresso significativo na área, atraindo inúmeros pesquisadores. Como consequência, problemas relacionados à Recuperação de Informação tornaram-se amplamente discutidos no âmbito da Ciência da Informação (SARACEVIC, 1995, 1996).

2.1 O Processo de Recuperação de Informação

O processo de recuperação de informação pode ser representado conforme a Figura 1. Possui basicamente três elementos: a **representação dos documentos** pertencentes ao *corpus*, a **representação da expressão de busca** e a **função de busca**, que determinará o grau de similaridade entre os dois itens anteriores.

Figura 1 - Representação do processo de recuperação de informação



Fonte: Ferneda (2013, p.48)

Recuperar uma informação consiste em identificar, em um acervo documental, quais aqueles que satisfazem total ou parcialmente a uma determinada necessidade de informação do usuário.

Buckland (1991) define o conceito de "informação como coisa*"* e argumenta que os acervos dos sistemas de informação seriam registros relacionados a coisas ou objetos. Nesses sistemas informação está vinculada ao objeto que a contém. Por sua vez, o termo documento, entendido como coisa informativa, incluiria objetos, artefatos, imagens, sons etc.

Independente dos tipos de documentos gerenciados por um sistema de informação, a eficiência da recuperação é dependente da forma como esses documentos estão representados. A representação de um documento tem por objetivo identificar e descrever resumidamente o seu conteúdo informacional, ao mesmo tempo em que define seus pontos de acesso para a busca em um sistema de informação. A tarefa de representar os documentos é feita em um tempo anterior à execução de qualquer busca. A constituição de um acervo documental e a representação de cada documento desse acervo é uma condição para que seja possível realizar a recuperação de informação

Em um sistema de recuperação de informação o usuário expressa sua necessidade de informação por meio de uma expressão de busca, composta geralmente por um conjunto de termos que representa linguisticamente sua necessidade de informação. A principal dificuldade está em

predizer os termos foram usados para representar os documentos que satisfarão sua necessidade, e ao mesmo tempo evitar a recuperação de documentos não relevantes. Para isso é necessário que o usuário tenha um relativo conhecimento do vocabulário ou da terminologia do domínio de conhecimento de seu interesse.

Para que seja possível uma comparação entre a expressão de busca e cada um dos documentos do *corpus* é necessário que tal expressão seja representada de forma similar à representação dos documentos. Diversos recursos podem ser oferecidos por um sistema a fim de facilitar o usuário na especificação de sua expressão de busca. Porém, a expressão busca (consulta) deverá ser representada em um formato similar ao formato utilizado nas representações dos documentos. Se as representações dos documentos estão no formato XML, por exemplo, as buscas terão de ser convertidas preferencialmente para esse formato.

No centro do processo de recuperação de informação está a função de busca, que compara as representações dos documentos com a representação da expressão de busca e recupera os itens que supostamente fornecerão informações relevantes ao usuário. De forma geral, a função de busca calcula o grau de similaridade entre a expressão de busca e cada um dos documentos do *corpus*. Na maioria dos sistemas, a similaridade é definida por um valor numérico que pretensamente define o quão relevante é um documento para satisfazer a necessidade de informação do usuário. O resultado de uma busca é geralmente composto por uma lista de referências a documentos, ordenada pelo grau de similaridade calculada pela função de busca.

Um modelo de recuperação de informação é a especificação formal de três elementos: a *representação dos documentos*, a *representação da expressão de busca* e a *função de busca* (FERNEDA, 2012, p.20). De maneira mais formal, Baeza-Yates e Ribeiro-Neto (2011, p.58) definem um modelo genérico de recuperação de informação como uma quadrupla:

[**D**, **Q**, $\mathcal{F}$, $R(q_i, d_j)$].

D é um conjunto composto por visões lógicas (representações) dos documentos no *corpus*;

Q é um conjunto composto de visões lógicas das necessidades de informação dos usuários;

$\mathcal{F}$ é um *framework* para a modelagem de representações dos documentos, consultas e seus relacionamentos;

R(q_i, d_j) é uma função de ordenamento (*ranking*) que atribui um número real à relação entre uma representação da consulta q_i de **Q** e a representação de um documento d_j de **D**.

Apesar de alguns dos modelos de recuperação de informação terem sido criados entre as décadas de 1960 e 1970, as suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais e nos mecanismos de busca da Web.

2.2 Recuperação de Informação na Web

Os sistemas de recuperação de informação na Web têm como principal objetivo "minimizar as dificuldades do usuário em localizar a informação requisitada, ou seja, diminuir o tempo gasto em um processo de busca até que a informação desejada possa ser acessada" (CRISTOVÃO; DUQUE; SERQUEIRA, 2012, p.5-6). Também são responsáveis por "recuperar todos os documentos que são relevantes à necessidade de informação do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes" (BAEZA-YATES; RIBEIRO-NETO, 2013, p.4). São de grande importância para a sociedade contemporânea e organizações em geral, que dependem de documentos e informações para a tomada de decisão, resolução de problemas e construção de novos conhecimentos (BEPPLER, 2008).

Apesar da sua importância, as ferramentas de busca disponíveis atualmente ainda estão distantes de atender às expectativas dos usuários. Os problemas relacionados à recuperação de informação podem ser observados a partir de inúmeras perspectivas: relacionados à qualidade e variedade dos dados, relacionados à interação dos usuários com os sistemas e relacionados ao comportamento de busca. Entre os inúmeros aspectos que geram dificuldades e requerem atenção especial dos pesquisadores, Huang (1999), Baeza-Yates e Ribeiro-Neto (2013) e Nayak, Prasad e Senapati (2015) destacam:

- grande volume de dados: a quantidade de informações disponibilizada é muito ampla, heterogênea e se modifica com muita frequência;
- dados heterogêneos: os dados são disponibilizados em uma grande variabilidade de formatos (textual, imagens, vídeos, etc.) e idiomas (inglês, português, etc.). Além disso,

os conteúdos presentes em um documento podem também serem heterogêneos, podendo conter texto, áudio, vídeo e imagens;

- dados redundantes: a quantidade de documentos duplicada é volumosa, ou seja, a mesma informação pode ser encontrada em diversas fontes e elas, na maioria das vezes, não estão correlacionadas;
- dados não estruturados: grande parte dos dados da Web são disponibilizados e duplicados de forma livre, sem a adoção de qualquer tipo de estrutura, linguagem ou padrão de interoperabilidade;
- alto percentual de dados voláteis: as informações disponibilizadas podem ser adicionadas, removidas ou alteradas com grande facilidade, gerando problemas com *links* suspensos (ou quebrados), mudança de nomes de arquivos e relocação de domínios;
- qualidade dos dados: muitos dados disponibilizados na Web apresentam problemas relacionados à qualidade, podendo ser imprecisos, obsoletos, inválidos, mal escritos (erros ortográficos, gramaticais), inocentes ou maliciosos.

Além dos problemas relacionados à qualidade e variedade das informações disponibilizadas, a recuperação de informação na Web também apresenta dificuldades referentes à interação dos usuários com os sistemas, que se manifestam principalmente em dois aspectos: na elaboração da expressão de busca e na interpretação dos resultados recuperados (BAEZA-YATES; RIBEIRO-NETO, 2013). A recuperação de informação é um processo de comunicação que ocorre entre o usuário e um acervo documental com o objetivo principal de satisfazer as necessidades informacionais do usuário (SARACEVIC, 1997; ARAUJO JUNIOR, 2007).

Para que esta comunicação aconteça de forma satisfatória é necessário que o usuário estabeleça um diálogo com o sistema a fim de expressar a sua necessidade de informação por meio de sua expressão de busca. A eficiência do processo de recuperação de informação está relacionada às coincidências terminológicas entre os termos utilizados na expressão de busca e os termos utilizados na representação dos documentos. A eficiência desse processo depende do conhecimento que o usuário possui da terminologia da área ou assunto que está pesquisando. Quanto maior o seu

conhecimento em relação ao domínio de interesse, mais opções de vocabulário o usuário terá à sua disposição para enriquecer as buscas (ANTONIOU; HARMELEN, 2004; BEPPLER, 2008).

O número excessivo de resultados que normalmente são retornados de uma busca torna praticamente impossível a análise de todos documentos recuperados. Em função disso, geralmente os usuários restringem-se em avaliar apenas as primeiras páginas de resultado, na confiança de que elas apresentam as informações que estão necessitando (HUANG, 1999; SPINK et al., 2001).

Outro problema do processo de recuperação de informação refere-se à contextualização da expressão de busca. Por ser um ambiente fortemente influenciado pelos problemas linguísticos os usuários muitas vezes apresentam dificuldades para traduzir suas necessidades de informação em um conjunto de termos de busca. Segundo Bräscher (2002), problemas relacionados ao significado das palavras afetam a qualidade da recuperação de informação e interferem diretamente nos resultados apresentados. Os termos utilizados pelo usuário na formulação das buscas podem não ser suficientes para contextualizar adequadamente a sua necessidade de informação.

> Efetuar buscas em grandes repositórios de informações não estruturadas e não padronizadas torna-se uma tarefa árdua, levando muitas vezes a ambiguidades, a conteúdo fora de contexto e até mesmo a não recuperação da informação desejada. As ferramentas de busca, ao fazerem uso da linguagem natural, necessitam de conhecimento sobre o significado das expressões utilizadas e das relações entre elas, o que possibilita contextualização e tratamento de fenômenos linguísticos que afetam a qualidade da recuperação (ALTOUNIAN; GOMES, 2016, p.32).

Bräscher (2002) afirma que a ambiguidade ocorre quando palavras ou frases podem gerar mais de uma interpretação de seu significado. O campo da semântica, tradicionalmente explorado na linguística e na filosofia, é responsável pelos estudos referentes ao significado ou ao processo de significação. De acordo com Bregagnol e Lara (2012) e Altounian e Gomes (2016), quatro elementos ou propriedades afetam diretamente as buscas realizada pelos usuários em um sistema de recuperação de informação:

- **Sinonímia:** um dos temas mais abordados no campo da semântica e estuda a relação entre duas palavras ou mais que apresentam significados iguais ou semelhantes, ou seja, os sinônimos. Exemplos: ensinamento, lição, experiência / garota, menina / débil, fraco, frágil / distante, afastado, remoto;

- **Homonímia:** refere-se à relação entre duas ou mais palavras que, apesar de possuírem significados diferentes, possuem a mesma pronúncia e grafia semelhante. Alguns exemplos são: concerto (sessão musical), conserto (corrigir) / cessão (ceder), sessão (reunião), seção (repartição).
- **Paronímia:** é a relação existente entre duas ou mais palavras que possuem significados diferentes, mas são muito parecidas na pronúncia e na escrita. Exemplos: eminente (alto), iminente (próximo) / absolver (perdoar), absorver (assimilar) / geminada (duplicada), germinada (que germinou) / descrição (descrever), discrição (discreto).
- **Polissemia:** propriedade que uma mesma palavra tem de apresentar vários significados. Exemplos: manga (fruta), manga (parte de uma peça do vestuário) / letra (símbolo utilizado na escrita), letra (música oucanção) / banco (local de realização de operações financeiras), banco (assento).

Os problemas linguísticos apresentados são fatores de interferência na recuperação de informação, uma vez que as características do pesquisador (nível de conhecimento, cultura, localização geográfica, estrato social) e a ambiguidade presente na língua escrita (sinonímia, homonímia, paronímia, polissemia e variações linguísticas) tendem a influenciar as terminologias adotadas na escrita dos documentos, no processo de indexação (quando realizado de forma manual), na elaboração da expressão de busca e na interpretação dos resultados apresentados. Isto significa que muitos documentos relevantes podem não ser recuperados em função da não coincidência entre os termos utilizados na busca e na indexação ou por limitação do vocabulário adotado pelo usuário.

Em contrapartida, outras questões ainda podem ser levantadas que também causam complicações para o processo de elaboração das expressões de busca: a) falta de interatividade que as ferramentas atuais oferecem; b) uso de poucas palavras para formular consultas (aproximadamente 2,2 termos); c) resistência e falta de conhecimento para a correta utilização de expressões booleanas; d) dificuldade de traduzir uma necessidade de informação por meio de termos ou expressões de busca; e) obrigatoriedade de conhecimento do usuário sobre o domínio pesquisado, uma vez que, para realizar uma busca utilizando palavras-chave (ou expressões)

significativas, é necessário que ele tenha conhecimento sobre o contexto de interesse (ANTONIOU; HARMELEN, 2004; BEPPLER, 2008; NAYAK; PRASAD; SENAPATI, 2015).

Lopes (2002) destaca que, apesar dos problemas observados, os sistemas de recuperação de informação na Web têm potencial para o desenvolvimento de recursos que podem contribuir para o aperfeiçoamento do processo de busca. Neste sentido, foram observadas sugestões de melhorias que envolvem questões relacionadas à elaboração das expressões de busca, indexação do *corpus* e interpretação dos resultados.

Em relação à elaboração de expressões de busca, Jackson e Moulinier (2002), Woods (2004), Baeza-Yates e Ribeiro-Neto (2013) e Shneiderman et al. (2017) apresentam as seguintes sugestões para facilitar a interação do usuário com o sistema:

- desenvolvimento de novas metáforas visuais para a elaboração de expressões de busca e visualização de resultados, com condições de reduzir o trabalho adicional do usuário para a combinação da estrutura e do conteúdo nas consultas;
- possibilidade de formulação de consultas baseadas em conceitos;
- possibilidade de formulação de consultas baseadas em conteúdo;
- ferramentas de busca devem ser capazes de lidar com relacionamentos semânticos;
- desenvolvimento de recursos que permitam a execução de múltiplas tarefas de recuperação de informação, tais como: selecionar documentos de uma base, agrupá-los em categorias ou classes e deles extrair determinados conjuntos de informações;
- novas opções devem ser exploradas nas ferramentas de busca, tais como filtragem, filtragem colaborativa, criação de sentido e análise visual.

Baeza-Yates e Ribeiro Neto (2013) trazem duas sugestões importantes para a melhoria dos sistemas de recuperação de informação, relacionadas à indexação (e ranqueamento) dos documentos:

- desenvolvimento de melhores esquemas de ranqueamento, que explorem tanto o conteúdo quanto a estrutura dos documentos;

- construção de estruturas de indexação específicas que explorem a estrutura dos documentos.

Do ponto de vista da análise e interpretação de resultados, Branski (2004), Hogan et al. (2011) e Baeza-Yates e Ribeiro-Neto (2013), trazem algumas sugestões de contribuições que podem aperfeiçoar as buscas realizadas pelos usuários:

- as ferramentas de busca devem procurar ir além da mera identificação, a partir de um termo ou conceito, dos conteúdos que serão apresentados ao usuário;
- a partir de um termo ou conceito oferecido pelo usuário, a interface deve ser capaz de identificar citações relacionadas e, a partir delas, extrair outros termos ou conceitos que indiquem novas estratégias de busca;
- é importante incluir recursos que melhorem a navegação, tais como: melhor exploração dos *links*, popularidade de páginas Web, similaridade de conteúdo, colaboração, 3D e realidade virtual, com atenção especial para abordagens de busca/navegação híbridas;
- as ferramentas de busca devem incorporar tarefas de coleta de informação acerca de um grupo de sites ou documentos. Nas ferramentas atuais, os usuários devem agregar manualmente as informações que precisam de vários locais heterogêneos, cada qual com suas características de formatação e navegação.

Saracevic (1997) afirma que o sucesso ou fracasso de qualquer sistema interativo ou tecnologia depende de como os problemas dos usuários e os fatores humanos são abordados. Portanto, conhecer as características dos diferentes tipos de usuários, as atividades que desenvolvem e de que forma eles buscam, utilizam e transferem a informação em diferentes contextos (GASQUE; COSTA, 2010) pode ser fator determinante para o projeto de sistemas de busca mais eficientes e para a redução dos problemas existentes.

3.
Web *Crawler* Focado

Com o crescimento acelerado da Web, a busca por informações relevantes está se tornando uma tarefa cada vez mais complexa, representando um enorme desafio para os Web *crawlers* e os mecanismos de busca de uso geral. Nesse contexto, grandes desafios estão sendo enfrentados por pesquisadores em todo o mundo, tendo em vista a necessidade do aprimoramento de técnicas e tecnológicas advindas de diversas áreas das atividades humanas.

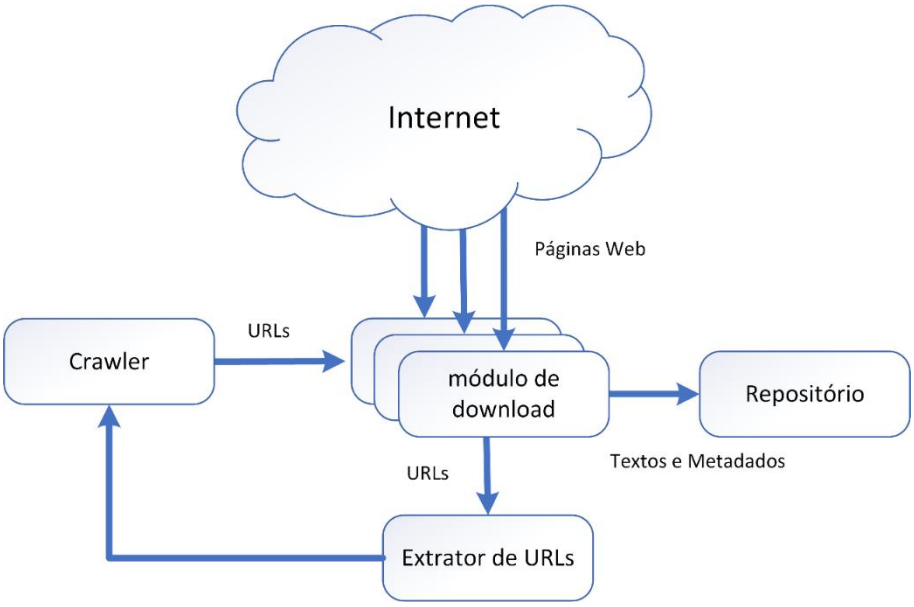
> Os sistemas de informação e de comunicação hoje permeiam e mesmo viabilizam virtualmente todas as atividades humanas, e não mais se pode conceber a sociedade sem a acentuada imbricação com as tecnologias de informação que nela surgem e que a modificam. Acompanhando o desenvolvimento dessas tecnologias, os repositórios de informações, que são produzidos no desempenho das inúmeras atividades humanas, vêm migrando para o ambiente online, cada vez mais em formatos digitais, acessíveis através de redes e sistemas de computadores. Tem-se assistido em paralelo a criação destes espaços (a Internet e a Web, as intranets empresariais, os portais corporativos, as bibliotecas digitais etc.) o surgimento de ferramentas e sistemas de recuperação de informações, para suprir a necessidade de localizar as informações criadas continuamente em ritmos vertiginosos (SOUZA, 2006, p.162).

Os motores de busca têm cada vez mais a necessidade de se especializar de forma que consiga trazer para o usuário as melhores e mais relevantes informações de acordo com sua necessidade informacional. Esse desafio vem sendo enfrentado e o assunto vem sendo pesquisado por empresas especialistas que tem se dedicado em tempo integral desenvolvendo melhorias e tentando criar técnicas e métodos que possa satisfazer os usuários, no sentido de dar mais precisão

no momento que esses sistemas de busca são utilizados. A começar pelos navegadores da Web, a otimização de busca tem sido constantemente usada e aprimorada nesse sentido.

O Web *crawler* é um algoritmo responsável por baixar as informações disponíveis na Web para o repositório do mecanismo de busca, para posterior processamento. O indexador gera um índice de termos, informações de arquivo e algumas outras características importantes das informações baixadas pelo algoritmo. O mecanismo de busca é responsável por processar uma consulta do usuário com uma ou mais palavras e combinações de curingas e conectores lógicos. O classificador é responsável por ordenar a maioria das entradas do indexador. O repositório de páginas costuma ser um meio físico para hospedar uma versão das páginas rastreadas, em um formato útil para o mecanismo de pesquisa (CHAITRA, P. G. et al, 2020, p.2). A Figura 2 ilustra o processo de funcionamento do algoritmo de *crawler*, e seus componentes básicos.

Figura 2 – Processo de um *Web Crawler*



Fonte: Adaptado de CHAITRA et al (2020, p.2)

O algoritmo faz *download* de páginas Web para um repositório. O *crawler* reserva uma lista de URLs para visitar, iniciando por uma URL semente, que pode ser definida pelo usuário. O trabalho do *crawler* começa a partir dessa URL semente. As páginas que correspondem a URL inicial são colhidas na Web e a lista de URLs não visitados da página e armazenado no crawler. O módulo de *download* extrai as páginas correspondente a URL. O repositório da Web é onde as páginas estão armazenadas e gerenciadas. Ele reserva apenas páginas HTML. Reserva as páginas

separadas em arquivos, e o gerenciador de armazenamento mantém todas as páginas acessadas pelos crawlers atualizadas (CHAITRA, P. G. et al, 2020).

A quantidade e a diversidade de documentos disponíveis na Web impõem importantes desafios na busca e recuperação de informações. Existem várias limitações que cercam os instrumentos desenvolvidos para facilitar a recuperação de informações na Internet. Para atualizar o conteúdo e indexá-lo, os mecanismos de busca usam o Web *crawler* (BHATT; VYAS; PANDYA, 2015a).

Os Web *crawlers* universais suportam mecanismos de pesquisa universais, ou seja, motores de busca que fazem suas pesquisas por toda a extensão da Web. Eles navegam na internet de forma metódica e automatizada criando um índice dos documentos encontrados. O *Web crawler* universal faz o primeiro download de uma página e em seguida, ele passa pelo HTML (*Hypertext Markup Language*) ou linguagem de marcação de hipertexto, dentro dessa página, até encontrar os links. Quando uma *tag* de link é encontrada, o *Web crawler* adiciona o link para a lista de links numa fila que ele planeja rastrear. Assim, como o *Web Crawlers* universal rastreia todas as páginas encontradas, esse rastreamento gera um custo maior pois é incorrido em muitas consultas. O *Web crawler* universal é comparativamente caro, conforme relatado em (BHATT; VYAS; PANDYA, 2015b).

Um Web *crawler* é geralmente livre para mover-se para qualquer site no ambiente virtual da Web, seguindo os links nas várias páginas. O Web *crawler* navega através dos sites, tendo em seu campo de visão sempre um próximo link para um próximo site. O Web *crawler* usa os links para selecionar seu próximo movimento, que o levará para várias outras páginas, onde novamente serão encontrados uma série de novos links, que o guiarão de forma metódica, coletando esses links e armazenando em um repositório para que possa disponibilizá-los na lista, para serem utilizados pelo mecanismo de busca. Além disso ele coleta estatísticas sobre a Web e faz indexação de aplicativos para mecanismos de pesquisa, podendo também ser usado para realizar acessibilidade e verificações de vulnerabilidade nos aplicativos, já que o *crawler* também tem a tarefa de atualizar constantemente as bases de dados desses aplicativos (CHAITRA, P. G. et al, 2020).

Os Web *crawlers* precisam se comunicar entre si para explorar as informações já registradas sobre conteúdo da Web. Ao se comunicar, esses *crawlers* usam regras predefinidas que determinam as rotas ideais em seu campo de visão. A comunicação ocorre por meio de um agente gerenciador

de regras que dita as regras de todos os *crawlers*, de acordo com a necessidade da busca para esse *crawler*. Cada *crawler* pode iniciar a comunicação para propor novas regras ao gerenciador de regras ou solicitar novas regras a qualquer momento (BATZIOS et al., 2008).

O Web *crawler* pode ser considerado um programa altamente proficiente que navega na internet de maneira deliberada e mecanizada. O seu principal objetivo é buscar as páginas da Web armazená-las em um repositório. Os Web *crawler* são usados essencialmente para seguir todas as páginas visitadas, baixando essas páginas que serão tratadas posteriormente por um mecanismo de pesquisa que indexará as páginas baixadas para auxiliar na pesquisa rápida. O trabalho dos mecanismos de busca é manter informações sobre várias páginas da Web. O programa automatizado, recupera essas páginas da Web e seus links associados. é um algoritmo cuja execução se inicia a partir de um conjunto de URLs iniciais. Depois de buscar as páginas associadas a uma URL, essa URL é removida da fila de trabalho. O Web *crawler* então analisa o documento, extrai as URLs relacionados aos links encontrados nesse documento e as adiciona à lista de URLs iniciais. Este processo continua iterativamente até todos os conteúdos acessíveis a partir das URLs semente serem alcançados (MIRTAHERI et al., 2014).

O principal objetivo de um Web *crawler* é fornecer dados atualizados para um mecanismo de busca. Eles são usados principalmente para criar uma cópia de todas as páginas rastreadas para que posteriormente possam ser processadas por um mecanismo de pesquisa após serem indexadas para fornecer resultados mais rapidamente, sem a necessidade do mecanismo de busca ter que percorrer toda a Web para isso. Os objetivos de um Web *crawler* ideal são sua escalabilidade fácil, sua capacidade de determinar qual conteúdo pode ser baixado e qual deve ser descartado (SARMIENTO; SALINAS, 2013).

O conceito de trabalho de um Web *crawler* e sua operação pode ser considerado um simples processo nas seguintes etapas: 1) inicia-se com uma URL de semente (ou identificador de um site) de uma determinada lista; 2) faz o download de suas páginas HTML; 3) faz a classificação dos links encontrados nessas páginas; 4) através desses links inicia nova pesquisa com cada documento vinculado. E assim cada novo documento vinculado é classificado novamente, e o processo segue novamente. Os metadados encontrados nos documentos que são criados pelo gerenciador de downloads, que tem o papel de examinar o conteúdo de um site, são armazenados em um repositório. Ele procura no site por mais links ou URLs, que são enviados para uma fila de espera

para processamento posterior. Por outro lado, existe um módulo denominado "escalonador", que se encarrega de retirar os links da fila de espera para enviá-los novamente e com ele realizar um novo processo, denominado varredura de segundo nível. No entanto, devido ao número de sites e de páginas que cada um deles possui, o Web *crawler* deve considerar uma maneira rápida de selecionar as páginas para download e uma maneira ideal de verificar quais mudanças essas páginas tiveram ao longo do tempo (OLSTON; NAJORK, 2010).

Um dos tipos de Web *crawlers* a serem estudados são os chamados Web *crawlers* focados, que tem como principal objetivo não coletar todas as páginas da Web, mas sim focar naqueles relevantes ou importantes para um conjunto predefinido de tópicos antes de começar a rastrear.

Os Web *crawlers* preferenciais são os rastreadores baseados em tópicos. Eles são seletivos no caso de páginas da Web. Eles são construídos para recuperar páginas dentro de um determinado tópico. Web *crawlers* focados ou *Web crawlers* de tópicos são tipos de rastreadores preferenciais que pesquisam por um tópico específico e rastreiam apenas um subconjunto predefinido de páginas de toda a Web. Os algoritmos para o Web *crawlers* de tópicos ou Web *crawlers* focados começaram a serem pesquisados em maior profundidade. Do ano de 2015 até o presente, uma variedade desses algoritmos pode ser encontrada na literatura (BHATT; VYAS; PANDYA, 2015a).

3.1 Web Crawlers: um breve histórico

Os primeiros Web *crawlers* foram escritos em 1993: *World Wide Web Wanderer, Jump Station, World Wide Web Worm* e *RBSE spider* (MCBRYAN, 1994). Esses quatro *Crawlers* coletavam principalmente informações e estatísticas sobre a Web usando um conjunto de URLs iniciais. Os primeiros Web *crawlers* que baixavam iterativamente e atualizavam seu repositório de URLs através das páginas da Web baixadas. No ano seguinte, 1994, surgiram dois novos Web *crawlers*: Web *Crawler* e *MOMspider*. Além de coletar estatísticas e dados sobre o estado da Web, esses dois Web *crawlers* introduziram conceitos de transparências e listas negras aos tradicionais Web *crawlers*. Foi considerado o primeiro Web *crawler* paralelo que baixava 15 links simultaneamente. A partir do Web *crawler World Wide Web Worm*, o número de páginas indexadas aumentaram de 110.000 para 2 milhões. Nos anos seguintes alguns rastreadores comerciais da Web se tornaram disponíveis: *Lycos, Infoseek, Excite, AltaVista e HotBot* (MIRTAHERI et al., 2014).

Em 1998, abordou-se a questão da escalabilidade, introduzindo um Web *crawler* em grande escala chamado Google. O Google abordou o problema de escalabilidade de várias formas: em primeiro lugar, alavancou muitas otimizações de baixo nível para reduzir o tempo de acesso ao disco por meio de técnicas como compactação e indexação. Em segundo lugar, e em um nível superior, o Google calculou a probabilidade de um usuário visitar uma página por meio de um algoritmo chamado *PageRank*, que calcula essa probabilidade levando em consideração o número de links que apontam para a página, bem como o estilo desses links. Tendo essa probabilidade, o Google simulou um usuário arbitrário e visitou uma página com a mesma frequência que o usuário. Essa abordagem otimiza os recursos disponíveis para um Web *crawler* reduzindo a taxa de visitas do rastreador da Web. Por meio dessa técnica, o Google alcançou alta performance. Arquitetonicamente, o Google usou uma arquitetura mestre-escravo com um servidor mestre (chamado URLServer) de envio de URLs para um conjunto de nós escravos. Os nós escravos recuperam as páginas atribuídas baixando-as da Web. No seu pico, a primeira implementação do Google atingiu downloads de 100 páginas por segundo (BRIN; PAGE, 2003).

A questão da escalabilidade foi posteriormente abordada por Allan Heydon e Marc Najork em uma ferramenta chamada Mercator em 1999. Além disso, Mercator tentou resolver o problema de extensibilidade de Web *Crawlers*. Para tratar da extensibilidade ele tirou proveito de uma estrutura modular baseada em Java. Esta arquitetura permitiu que componentes de terceiros fossem integrados no Mercator. Para resolver o problema de escalabilidade, Mercator tentou resolver o problema de URL-Seen(visto). O problema URL-visto responde à questão de uma URL ser ou não visto antes. Este problema aparentemente trivial torna-se muito demorado à medida que o tamanho da lista de URLs aumenta. Mercator aumentou a escalabilidade do URL-visto por verificações de disco em lote. Neste modo, hashes de URLs descobertos são armazenados na RAM. Quando o tamanho desses hashes cresce além de um certo limite, a lista foi comparada com os URLs armazenados no disco e a própria lista no disco foi atualizada. Usando esta técnica, a segunda versão do Mercator rastreou 891 milhões de páginas. Mercator foi integrado ao AltaVista em 2001 (HEYDON; NAJORK, 2003).

A IBM introduziu o *WebFountain* em 2001. *WebFountain* foi um Web *crawler* totalmente distribuído e seu objetivo não era apenas para indexar a Web, mas também criar uma cópia local disso. Esta cópia local era incremental, o que significa que uma cópia da página foi mantida

indefinidamente no espaço local, e esta cópia era atualizada com a mesma frequência que o *WebFountain* visitava a página. Dentro *WebFountain*, os principais componentes, como o programador foram distribuídos e o rastreamento foi um processo contínuo onde a cópia local das páginas aumentava. Esses recursos também como implantação de tecnologias eficientes, como o *Message Passing* Interface (MPI), traduzido como Interface de Passagem de Mensagens, tornou o *WebFountain* um Web crawler escalável com alta taxa de atualização. Em uma simulação, *WebFountain* conseguiu escalar com uma Web simulada que originalmente tinha 500 milhões de páginas e aumentou para o dobro de seu tamanho a cada 400 dias (SHKAPENYUK; SUEL, 2007).

Em 2002, Polybot abordou o problema de escalabilidade do URL-vista aprimorando a técnica de verificação de disco em lote, usou a árvore *Red-Black* para manter as URLs e quando a árvore cresce além de um certo limite, é fundida com uma lista classificada na memória principal. Usando esta estrutura de dados para lidar com Teste de URL-vista, *Polybot* conseguiu escanear 120 milhões de páginas (SHKAPENYUK; SUEL, 2002).

No mesmo ano, *UbiCrawler* lidou com o problema de URL-vista com uma abordagem diferente, mais ponto a ponto (P2P). O *UbiCrawler* usou *hashing* consistente para distribuir URLs entre nós do Web *Crawler*. Neste modelo, nenhuma unidade centralizada calcula se uma URL foi vista ou não antes, mas quando uma URL é descoberta, é passado para o nó responsável por responder ao teste. O nó responsável por fazer este cálculo é detectado por pegar o *hash* da URL e mapeá-la para a lista de nós. Com cinco PCs de 1 GHz e cinquenta threads, o *UbiCrawler* alcançou uma taxa de download de 10 milhões de páginas por dia (BOLDI et al., 2004).

Em 2005, foi proposto um modelo com partição aumentada de URLs entre um conjunto de nós de rastreamento em uma arquitetura P2P levando em consideração as informações geográficas dos servidores. Tal um aumento reduziu o tempo geral do rastreamento em alocar servidores de destino para um nó fisicamente mais próximo a eles (EXPOSTO et al., 2005).

Em 2008, um Web *crawler* extremamente escalável chamado *IRLbot* funcionou por 41,27 dias em um *AMD Opteron* 2.6 e rastreou mais de 6,38 bilhões de páginas da Web. O *IRLbot* abordou principalmente o problema de URL-vista por dividindo-as em três subproblemas: verificar, atualizar e verificar + atualizar. Para resolver esses subproblemas, o *IRLbot* introduziu uma estrutura chamada *Disk Repository with Update Management* (DRUM), traduzido como Repositório de Disco com Gerenciamento de Atualizações. DRUM otimiza o acesso ao disco,

segmentando o disco em vários depósitos de disco (barris). Para cada barril criado, DRUM também aloca um barril correspondente na RAM. Cada URL é mapeada para um intervalo. No início, uma URL era armazenada em seu barril de RAM. Assim que um barril na RAM estiver cheio, o depósito de disco correspondente é acessado no modo em lote. Este acesso ao modo em lote, bem como o agrupamento de dois estágios no sistema usado, permitiu que DRUM armazenasse um grande número de URLs no disco, de modo que seu desempenho não seja degradado conforme o número de URLs aumenta (LEE et al., 2008).

Em 2011, foi apresentada a versão inicial da primeira estratégia de rastreamento baseada em modelo: a estratégia de hipercubo. A estratégia faz previsões, inicialmente assumindo que o modelo do aplicativo é uma estrutura de hipercubo. A implementação inicial tinha desvantagens de desempenho que impediu que a estratégia fosse prática mesmo quando o número de eventos no estado inicial eram apenas vinte. Essas limitações foram removidas posteriormente (BENJAMIN et al., 2011).

Em 2012, Choudhary introduziu outra estratégia baseada em modelo, chamada estratégia Menu. Esta estratégia foi otimizada para os aplicativos que têm o mesmo evento sempre levando ao mesmo estado, irrelevante para o estado de origem. A estratégia de rastreamento do modelo de menu é uma estratégia de exploração baseada em eventos. A estratégia subjacente do modelo de menu é o processo de categorizar os eventos ativados em todos os estados descobertos em conjuntos de prioridades diferentes. A prioridade do conjunto é definida pela contagem de execução dos eventos presentes nesse conjunto. A contagem de execução do evento representa a contagem total de instâncias de execução do evento de diferentes estados. Por exemplo, a estratégia de rastreamento define um conjunto de eventos globais não executados que contém todos os eventos que não foram executados em qualquer lugar do aplicativo descoberto até agora. Da mesma forma, define outro conjunto chamado eventos de menu que contém todos os eventos na aplicação que seguem a hipótese na forma como a sua execução a partir estados diferentes levaram ao mesmo estado resultante (CHOUDHARY, 2012).

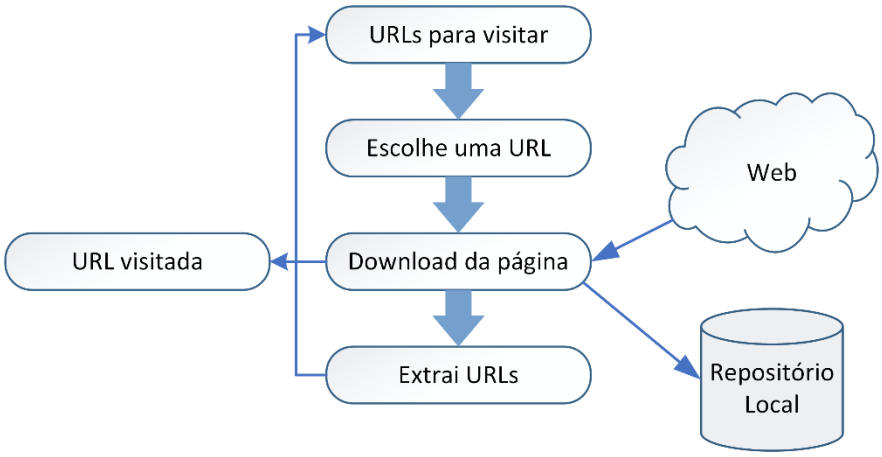
Como já mencionado, um *Web crawler* é a parte principal de um mecanismo de pesquisa que coleta informações da internet para que o indexador possa criar um índice desses dados. O *Web crawler* inicia o processo de rastreamento a partir de uma única URL ou um conjunto de URLs iniciais. Conforme o *Web crawler* visita uma URL, ele adiciona todos os hiperlinks da página da

Web a uma lista de URLs a serem visitados posteriormente. O objetivo do rastreamento é coletar o máximo possível de páginas úteis no menor tempo possível. O *Web crawler* prioriza a ordem em que as URLs são visitadas devido à grande coleção de dados existentes na Web. O grande tamanho e variedade da de páginas da Web, torna difícil para qualquer que rastreador recupere todos os dados relevantes. Assim, tornou-se uma área de pesquisa ativa e várias variantes de *Web crawlers* surgiram.

Nesse sentido pode-se dizer tecnicamente, que os Web *Crawlers* são as ferramentas para aquisição de dados nos motores de busca, que usa como entrada um conjunto de URLs sementes, e exercendo sua função, tem como saída um conjunto de páginas da Web rastreadas.

Na Figura 3 podemos apresentar a arquitetura de um simples Web *Crawler*, seguindo seu funcionamento como um Web *crawlers* de uso geral.

Figura 3 – Arquitetura de um *Web Crawler*



Fonte: (KUMAR; BHATIA; RATTAN, 2017, p.2)

A partir de uma URL visitada, o *crawler* obtém as URLs encontradas nessa página e as adiciona a uma lista de URLs a serem visitadas posteriormente. A próxima URL é escolhida e a página da Web correspondente é baixada no repositório local. Dessa página baixada, as URLs são extraídas e adicionadas novamente à lista de URLs a serem visitados.

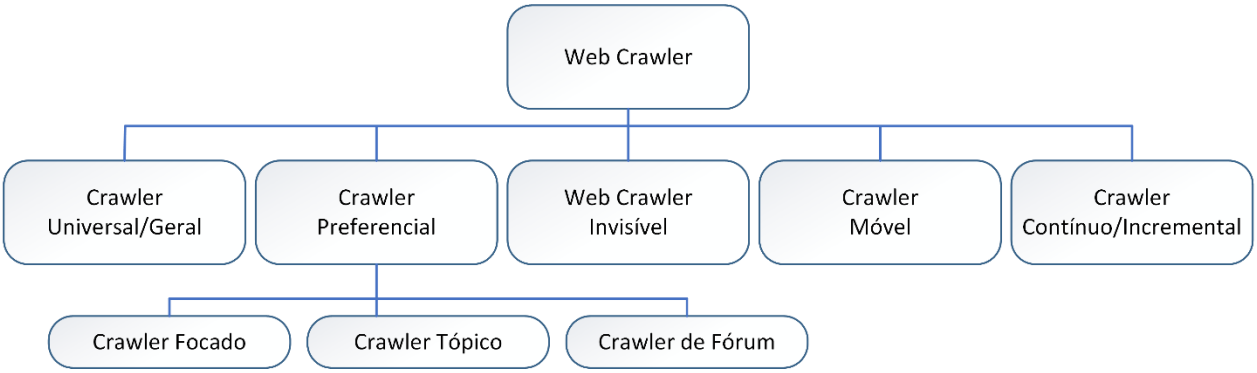
Essa arquitetura básica pode ser modificada de acordo com as necessidades da aplicação de rastreamento. Este processo é repetido recursivamente até que a lista de URLs a serem visitadas esteja vazia. Nessa arquitetura, mais componentes podem ser adicionados, ou componentes

existentes podem ser combinados conforme os requisitos. Um crawler básico precisa no mínimo ser formado por esses componentes: um conjunto de URLs para iniciar o processo de rastreamento, um modulo de download para baixar páginas Web para o repositório local; extrator de URL para obter as URLs dessas páginas (KUMAR; BHATIA; RATTAN, 2017, p.2).

3.2 Tipos de Web Crawlers

Segundo Kumar (2017), não existe uma taxonomia padrão para os Web *Crawlers* disponíveis na literatura. Vários pesquisadores têm Web *Crawlers* categorizado em sua própria maneira. A Figura 4 fornece uma ampla taxonomia do Web *Crawler*.

Figura 4 - Taxonomia de *Web Crawlers*.



Fonte: Adaptado de (KUMAR; BHATIA; RATTAN, 2017, p.3)

***Crawler* universal ou de uso geral** (*universal/broad crawler*)**:** este tipo de Web *crawler* não está limitado a páginas Web de um determinado tópico ou domínio. Ele segue links indefinidamente em todas as páginas que encontrarem.

***Crawler* preferencial** (preferencial *crawler*)**:** esta categoria de *Web crawlers* não rastreia todos os links que encontra. O usuário envia uma condição ou tópico de interesse que irá guiar o c*rawler*. Os Web *Crawlers* preferenciais podem ser categorizados como focado e de tópico. O *crawler* tópico é utilizado para rastrear informações relacionadas a algum assunto específico. O *crawler* focado permite também a utilização um conjunto páginas-exemplo como ponto de partida para rastrear páginas Web semelhantes. *Crawler de fórum* é um tipo específico de crawler que trata apenas do rastreamento de conteúdos relacionados a fóruns online na Web.

Web *crawler* oculto ou invisível (*hiden* Web *crawler)*: categoria especial de rastreadores que trata do rastreamento da chamada *Deep Web*, onde existe uma quantidade significativa de informações que não podem ser acessadas diretamente seguindo os hiperlinks nas páginas da Web.

***Crawler* móvel** (*mobile crawler)*: É um método de rastreamento no qual a seleção e a filtragem das páginas Web são realizadas no servidor, e não no mecanismo busca, reduzindo a carga de rede causada pelo rastreador tradicional da Web. Sua principal vantagem é que o processo de rastreamento é feito localmente onde os dados residem, em vez de remotamente pelo mecanismo de busca.

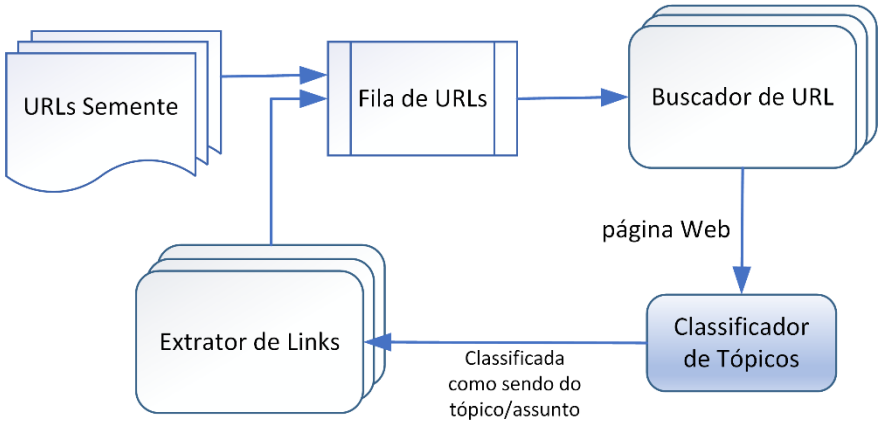
Web *crawler* contínuo ou incremental (*continuous/incremental crawler)*: a Web é dinâmica e os dados nas páginas estão mudam com frequência. Essa categoria de *crawlers* são usados para manter o banco de dados de índice do mecanismo de busca atualizado, podendo detectar as mudanças em páginas Web.

3.3 Tipos e características do Web *crawler* focado

Os Web *crawlers* focados são resultado de esforços de pesquisa para aumentar a eficiência da recuperação de informação na Web. Esse tipo de *crawler* tem a função de encontrar e indexar páginas relacionadas apenas a uma área ou a um determinado assunto.

O rastreamento começa com um conjunto de URLs iniciais e URLs a serem buscadas armazenada em uma estrutura de fila, chamada "fila URL". Dessa forma, vários segmentos são executados simultaneamente. Cada segmento recebe o URL da fila e busca as páginas da Web correspondentes no servidor. Mais tarde, esta página é analisada para extrair links e esses links são anexados à fila de URLs para serem buscados posteriormente, tornando muito mais ágil na detecção de alterações nas páginas dentro de seus limites, do que um rastreador que está rastreando toda a Web (BHATT; VYAS; PANDYA, 2015b). A Figura 5 demonstra o processo básico executado pelo *crawler* focado.

Figura 5 – Estrutura de um Web crawler focado



Fonte: adaptado de (CALISKAN; OZCAN, 2013).

O Web *crawler* focado é guiado por um classificador que aprende reconhecer a relevância dados incorporados em um tópico de categorias. Uma estratégia mais refinada é proposta usando o contexto textual limitado em torno da URL para classificação. Esse tipo Web *crawler* focado é chamado de rastreador de link, o qual é baseado em contexto no próprio documento (PANT; SRINIVASAN, 2006).

O componente vital em um Web *crawler* focado é o módulo classificador. Isso afeta diretamente a eficácia do *crawler* e usa mecanismos de pesquisa verticais (mecanismo que permite um focar em conteúdos voltados a assuntos específicos, como por exemplo imagens, vídeos, notícias, entre outros, promovendo maior qualidade nos termos retornados) para rastrear as páginas específicas de um tópico (PANT; SRINIVASAN, 2006).

Huo; Liu; Yan (2010) relata que a única diferença do Web *crawler* focado em relação ao Web *crawler* de uso geral é o classificador de tópicos que o torna mais preciso. Cada página obtida é classificada para um destino predefinido. Se a página estiver relacionada ao conteúdo do tópico, então seus links são extraídos e são adicionados à fila de URLs. Caso contrário o processo de rastreamento não procede desta página. Este tipo de *Web crawler* focado é chamado de *Web crawler* focado em "*página inteira*", pois classifica o conteúdo completo da página. Em outras palavras, o contexto de todos os links na página é o conteúdo da página inteira em si.

O Web *crawler* focado implica na busca apenas das páginas relevantes para um determinado tópico ou assunto. Diferentes abordagens foram propostas para a estratégia de foco. Em geral, todos elas são baseadas na suposição de que as páginas da Web sobre um determinado

assunto sejam mais prováveis para serem vinculadas no mesmo tópico. A partir de um conjunto de páginas da Web que representam o tópico especificado, o Web *crawler* focado segue seus links e é restringido por palavras de conteúdo nas páginas da Web de saída ou a estrutura gráfica, ou tokens de URL, ou a combinação desses critérios. Eles são usados para impedir o rastreamento de sites não relacionados, o que resultaria em um desvio do tópico específico (MEDELYAN *et al*., 2006).

O primeiro Web *crawler* focado foi descrito por Chakrabarti *et al*. (1999), depende do modelo que representa a estrutura de ligação da Web, ou seja, se uma página da Web aponta para uma página específica do tópico, outras páginas apontadas provavelmente pertencerão ao mesmo tópico. Como uma restrição adicional de foco, eles usam classificadores probabilísticos criados a partir de um conjunto de páginas representativas.

Vários estudos posteriores sobre rastreamento focado (DILIGENTIT *et al*., 2000); (AGGARWAL; AL-GARAWI; YU, 2001a); (SOMBOONVIWAT; KITSUREGAWA; TAMURA, 2005); (DILIGENTIT *et al*., 2000) constroem gráficos de contexto manualmente, em um conjunto avaliado de páginas da Web, em que páginas da Web de origem estão vinculadas.

Essa estrutura é analisada para extrair dependências de tópicos hierárquicos. Dados esses classificadores hierárquicos, Diligenti *et al*., (2000) projetaram um *Web crawler* focado no contexto que supera o padrão focado em termos de etapas necessárias até que uma página da Web específica de tópico seja encontrada. No entanto, os autores admitem que sua abordagem depende da natureza de uma categoria, ou seja, ela somente é útil para categorias que possuem um modo padrão de hierarquia.

Aggarwal *et al*., (2001b) tentam superar o problema de estruturas de links dependentes de categoria combinando-as com outros recursos, como palavras de conteúdo exibidas em páginas da Web, tokens de URL etc.

Todos esses autores supracitados têm em comum assumirem a existência de estruturas gráficas semelhantes na Web. Esse certamente é o caso de grandes sites da Web, como Yahoo, Ebay ou Amazon. No entanto, o restante dos usuários da Web não construía sites de uma maneira estruturada especial, simplesmente porque não havia diretrizes amplamente aceitas para isso (KHARE *et al*., 2004).

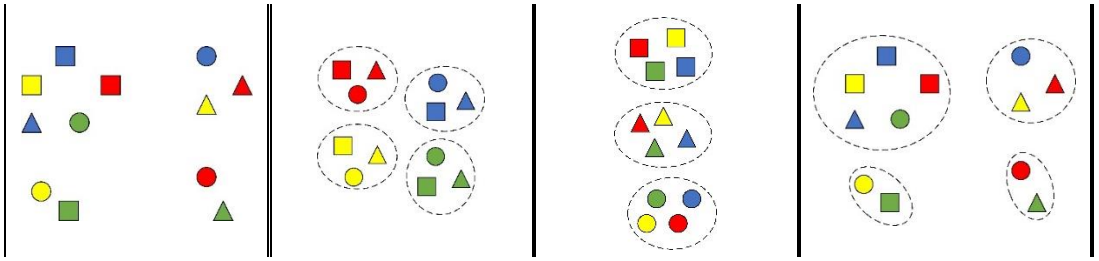
4. *Clustering*

"Um ser inteligente não pode tratar cada objeto que vê como uma entidade única diferente de qualquer outra coisa no universo. Ele precisa colocar os objetos em categorias para que possa aplicar ao objeto em mãos seu conhecimento duramente conquistado sobre objetos semelhantes, encontrados no passado"

(Steven Pinker)

Uma das capacidades mais básicas do ser humano envolve o reconhecimento de semelhanças entre objetos, criando assim grupos (*clusters*) que compartilham determinadas características ou certas propriedades. O processo de agrupamento (*clustering*) se assemelha à classificação, referindo-se à agregação de objetos "semelhantes" e à segregação de objetos distintos. No entanto, a classificação pressupõe a existência *a priori* de uma estrutura de classificação, na qual cada classe deve estar devidamente caracterizada. A partir da caracterização de cada classe, a tarefa é identificar a que classe um objeto pertence. As técnicas de agrupamento podem ser consideradas como uma etapa anterior à classificação, pois a estrutura classificatória não está disponível. Portanto, o objetivo das técnicas de *clustering* é identificar os objetos que possuem algo em comum, agrupando-os em subconjuntos de objetos similares. Nesse processo, a medida de *semelhança* é dada por uma ou mais características desses objetos. A Figura 6 apresenta um conjunto de objetos (**a.**) e as possíveis formas de agrupamento de tais objetos (**b.**, **c.**, **d.**).

Figura 6 – Agrupamento de objetos por diferentes características



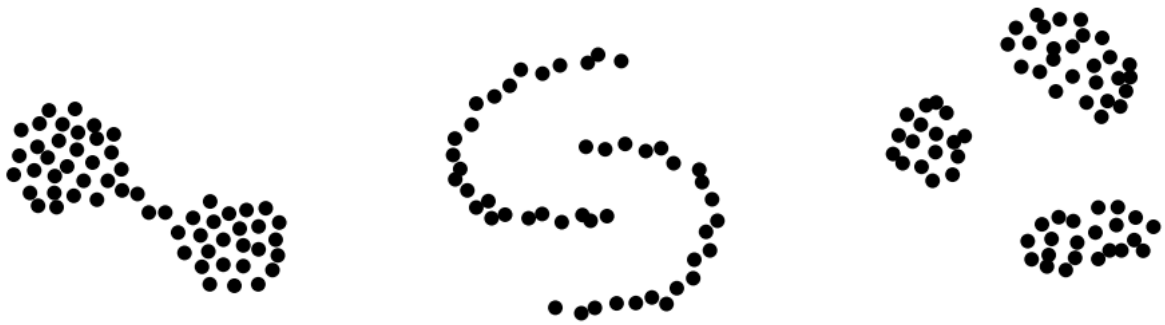
a. **b.** **c.** **d.**

Fonte: o Autor

Na ilustração da Figura 6, o conjunto inicial de objetos distribuído em um plano (**a.**) pode ser agrupado utilizando como critério a cor desses objetos (**b.**), a forma destes (**c.**) ou a distância entre eles (**d.**).

Definir formalmente o conceito de *cluster* é uma tarefa complexa. Cormack (1971) e Gordon (1999), tentam definir *cluster* em termos de coesão interna (homogeneidade) e isolamento externo (separação). Os *clusters* presentes na Figura 7 são visualmente evidentes para a maioria dos observadores, sem que seja necessária uma definição formal do termo. Por esse motivo, tentativas de tornar os conceitos de homogeneidade e separação matematicamente precisos nos remetem a diversos critérios.

Figura 7 – *Clusters* - coesão interna e isolamento externo



Fonte: (EVERITT et al., 2011, p. 7)

Segundo Everitt (1974), um *cluster* pode ser definido como:

a. Um *cluster* é um conjunto de entidades semelhantes. Entidades pertencentes a *clusters* distintos são diferentes;

b. Um *cluster* é uma agregação de pontos no espaço tal que a distância entre os pontos em um mesmo *cluster* é menor que a distância entre pontos de *clusters* diferentes;
c. Os *clusters* podem ser descritos como regiões conexas de um espaço multidimensional que contém uma grande densidade relativa de pontos. As regiões estão separadas uma das outras por regiões de baixa densidade relativa de pontos.

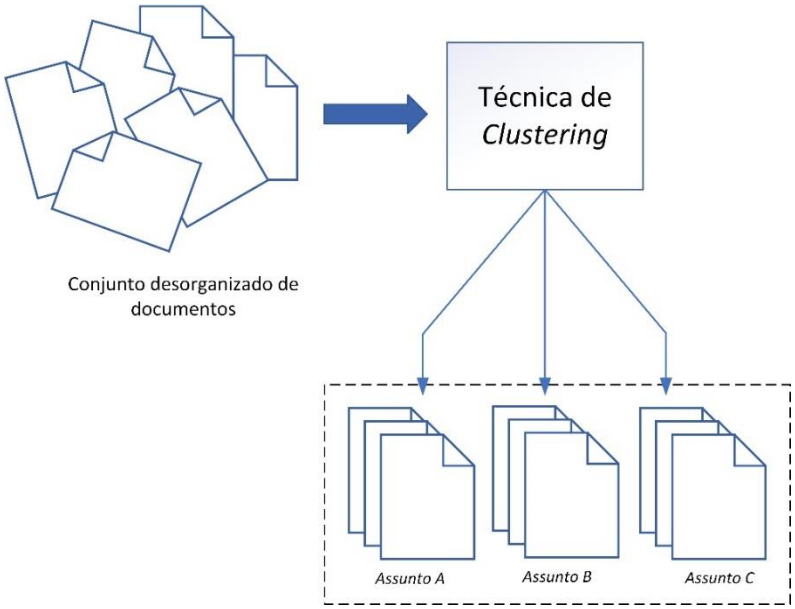
Segundo Jiawei Han (1996), na informática o agrupamento de objetos/informações é uma técnica de *Descoberta de Conhecimento* e *Mineração de Dados* estudada pela área de *Inteligência Artificial*. Muitos algoritmos de agrupamento de objetos já foram implementados, estudados e aplicados em diversas áreas do conhecimento. Atualmente, as áreas de marketing e de economia têm despertado grande interesse pelas técnicas de agrupamento com o objetivo de obter conhecimento sobre os padrões de consumo. Na Recuperação de Informação os algoritmos de agrupamento têm sido utilizados para a organização de conjuntos de documentos textuais, tornando a recuperação mais eficiente quando estes estão organizados em um mesmo bloco físico.

4.1 *Clustering* de informações textuais

O agrupamento de informações textuais tem por objetivo separar um conjunto desorganizado de documentos em uma série de subconjuntos de documentos de assuntos similares, como ilustrado na Figura 8. Para isso, parte-se da *Cluster Hypothesis* (RIJSBERGEN, 1979) que diz que objetos semelhantes e relevantes a um mesmo assunto tendem a permanecer em um mesmo grupo (*cluster*), pois possuem atributos em comum. Formalmente, o *clustering* de textos pode ser enunciado como (BAEZA-YATES; RIBEIRO-NETO, p.283):

> **Clustering de textos**: dada uma coleção D de documentos, um método de clustering textual automaticamente separa esses documentos em *K* clusters de acordo com algum critério predefinido.

Figura 8 – Funcionamento das técnicas de *clustering* de documentos



Fonte: o Autor

A maioria das técnicas de *clustering* foi desenvolvida para serem utilizadas em dados estruturados, geralmente armazenados em Sistemas Gerenciadores de Banco de Dados, mais fáceis de serem tratados computacionalmente. Porém, com o surgimento da Web, dados não estruturados como imagens, textos, gráficos, vídeos ganharam a atenção dos pesquisadores. Técnicas de agrupamento de documentos são muito utilizados em ferramentas de recuperação de informação para identificar documentos similares e armazená-los em blocos contíguos (*clusters*). Quando um dos documentos for recuperado em resposta a uma consulta, todos os outros documentos do mesmo *cluster* são também considerados relevantes. Assume-se, portanto, que todos os documentos de um *cluster* tratam de assunto similar; *clusters* diferentes tratam de assuntos diferentes. Desse modo, a informação permanece agrupada de acordo com o assunto, facilitando sua localização e manipulação.

Segundo Cooper (1969), o objetivo inicial de introduzir o processo de *clustering* de documentos na Recuperação de Informação era aumentar a eficiência dos sistemas: após o agrupamento de documentos em *clusters*, é necessário apenas encontrar os *clusters* mais similares a uma determinada consulta e não procurar por documentos individuais.

Os objetos em uma coleção a serem agrupados precisam ser descritos por atributos que permitam o cálculo da similaridade entre eles. No caso de agrupamento de documentos são

geralmente as palavras-chave que constituem os atributos. Portanto, o agrupamento é aplicado à representação dos documentos, obtidos através de indexação (HAMILL; ZAMORA, 1980).

O processo de c*lustering* pode ser totalmente automático, pois é independente do domínio, dependendo apenas do conteúdo dosdocumentos, sem necessitar do conhecimento especializado. Sua utilização é geralmente bem-sucedida na identificação grupos temáticos em coleções relativamente heterogêneas (HEARST, 1999).

4.2 *Clustering* de documentos

Segundo Willet (1988), os métodos de agrupamento (*clustering*) estão todos de alguma forma baseados em alguma medida de similaridade entre objetos. Portanto, para que os objetos possam ser agrupados, é necessário identificar as semelhanças existentes entre eles. Para identificar tais semelhanças é necessário identificar as características dos objetos e agrupá-los de acordo com a quantidade de características em comum (ou similares) que estes objetos possuem.

Nas técnicas de *clustering* tradicionais, os objetos já possuem atributos (características) bem definidos. Em um banco de dados relacional, por exemplo, os campos/atributos indicam as características das entidades. Nos textos, porém, não é fácil identificar um atributo, simplesmente porque não há qualquer sinal ou local predeterminado onde estes serão encontrados. Pode ocorrer, ainda, o fato de um atributo ser identificado em um documento e nem ao menos existir em outro. Em decorrência deste fato, as características relevantes dos objetos, podem ser variadas e muitas vezes ambíguas.

No processo de *clustering* de documentos é necessário estabelecer um método que identifique as características mais relevantes de cada objeto textual. Para esse tipo de objeto, as palavras do texto são os principais elementos para sua caracterização. Portanto, nas técnicas de agrupamento de informações textuais geralmente os termos resultantes do processo de indexação são utilizados como atributos que caracterizam documento.

O agrupamento de documentos exige uma forma de analisar todos os objetos com o objetivo de identificar semelhanças entre eles. O grau de semelhança entre dois objetos é dado, em geral, por uma fórmula ou função de similaridade. Esta fórmula analisa todas as

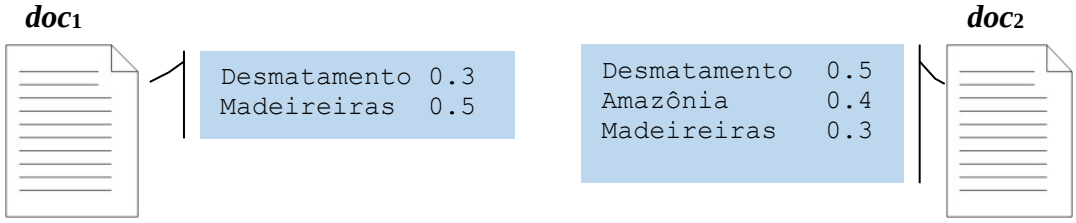
características (termos) semelhantes que os objetos possuem e retorna um valor, um grau, indicando o quanto estes objetos são similares.

Existem várias classes de funções que servem para calcular o quanto um objeto está relacionado com outro. Algumas destas funções são consideradas binárias, pois somente utilizam valores que indicam se um determinado termo está presente ou não na representação do documento. Neste caso, se um termo de indexação está presente utiliza-se o valor um (1), caso contrário, zero (0).

Algumas funções são capazes de utilizar valores que informam quanto determinada característica influi no objeto, ou seja, a quão discriminativa ela é. Essas funções geralmente levam em conta o número de ocorrências no termo no documento e em alguns casos levam em conta também o número de documentos em que um determinado termo aparece. Este valor geralmente é normalizado no intervalo entre zero (0) e um (1), onde zero diz respeito a variáveis com nenhuma influência (ou importância) para o objeto, e um (1) significando uma variável na qual o objeto é totalmente dependente (pois o caracteriza significativamente).

Na área da Recuperação de Informação, o modelo Vetorial possui característica que favorecem a implementação de *clustering* de documentos. Os documentos são representados por um conjunto de termos e seus respectivos pesos (vetor), resultantes do processo de indexação automática. Os pesos demonstram a importância de cada termo para a representação do conteúdo informacional do documento. A Figura 9 apresenta ilustração de dois documentos (**doc_1** e **doc_2**) indexados por um conjunto de termos com seus respectivos pesos.

Figura 9 – Representação Vetorial de documentos



Fonte: o Autor

Utilizando a representação vetorial é possível calcular o grau de semelhança ou similaridade entre documentos. A similaridade (***sim***) entre dois documentos é calculada utilizando-se a seguinte fórmula (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 78):

$$sim(x, y) = \frac{\sum_{i=1}^{t} w_{i,x} \times w_{i,y}}{\sqrt{\sum_{i=1}^{t} w_{i,x}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,y}^2}}$$

onde ***t*** é o número de termos (dimensão) dos vetores que representam os documentos; $w_{i,x}$ é o peso do termo ***i*** associado ao documento ***x*** ; e $w_{i,y}$ é o peso do termo ***i*** do documento ***y***.

Aplicando essa fórmula às representações dos documentos exemplificados na Figura 9 é possível calcular o grau de similaridade entre eles:

$$\boldsymbol{doc_1 = (0.3, 0.0, 0{,}5)} \qquad \boldsymbol{doc_2 = (0.5, 0.4, 0.3)}$$

$$sim(doc_1, doc_2) = \frac{(0.3 \times 0.5) + (0.0 \times 0.4) + (0.5 \times 0.3)}{\sqrt{0{,}3^2 + 0{,}0^2 + 0{,}5^2} \times \sqrt{0{,}5^2 + 0{,}4^2 + 0{,}3^2}}$$

$$\boxed{sim(doc_1, doc_2) \cong \mathbf{0.728}}$$

Portanto, o grau de similaridade/semelhança entre $\boldsymbol{doc_1}$ e $\boldsymbol{doc_2}$ é de **0.728**.

Existem diversos métodos e fórmulas de se calcular a similaridade entre documentos. Subhashini e Kumar (2010) apontam que a medida de similaridade *cosseno*, utilizada no modelo Vetorial, é superior às demais.

4.2.1 *Clustering* por partição total (*flat partition*)

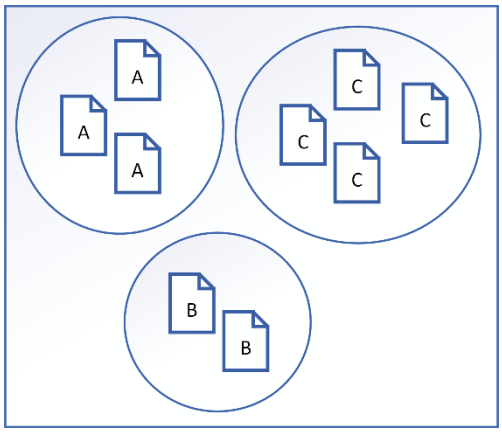
O método de partição total consiste em agrupar os documentos em um número predeterminado de *clusters*. Os documentos são agrupados de tal forma que todos os elementos de um mesmo *cluster* possuem um grau mínimo de similaridade.

Apesar de constituir *clusters* distintos, este método permite alocar um mesmo documento em mais de um *cluster*. Quando os documentos são atribuídos a um único *cluster* diz-se que o processo é *disjunto*. Caso um documento seja atribuído a mais de um *cluster* diz-se que o processo é *não-disjunto*. Algoritmos *disjuntos* inserem um determinado documento ao *cluster*

de maior similaridade. Porém, essa restrição pode ser quebrada em alguns casos. De qualquer modo, mesmo que um documento seja atribuído a mais de um *cluster*, os grupos continuam sendo considerados isolados.

Na Figura 10 os *clusters* são representados pelos círculos, que agrupam um número diverso de documentos. Como pode ser visto, os *clusters* não possuem ligações entre si, sendo totalmente isolados (disjuntos).

Figura 10 – Resultado de um agrupamento por partição total e disjunta

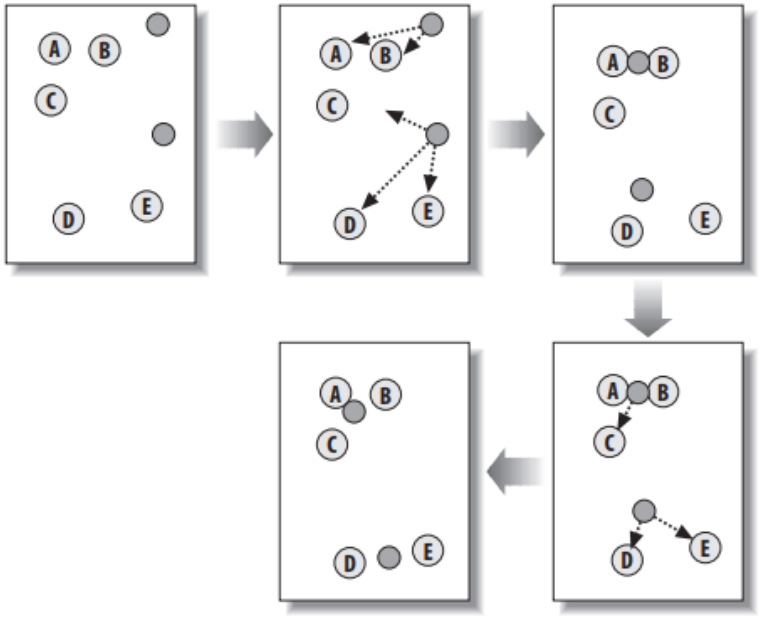


Fonte: o Autor

Na ilustração da Figura 10, as letras se referem à característica utilizada para formar cada *cluster*. Em um conjunto de documentos representados por metadados, as letras poderiam representar, por exemplo, a autoria do documento. Assim teríamos o grupo de documentos do autor **A**, o *cluster* do autor **B** e do autor **C**. As letras poderiam representar termos com os quais os documentos foram indexados: o *cluster* dos documentos indexados pelo termo **A**; o *cluster* dos documentos com o termo B e o grupo dos documentos indexados por **C**. É possível também calcular a similaridade entre pares de documentos a fim de agrupá-los por semelhança.

Um dos métodos mais utilizados no processo de *clustering* por particionamento é denominado *K-means* (MacKay, 2003). Nesse método é informado antecipadamente quantos agrupamentos distintos devem ser gerados. O processo de *clustering* K-mens é ilustrado na Figura 11.

Figura 11 – Agrupamento por particionamento *K-means*



Fonte: SEGARAN (2007, p.43).

Na Figura 11, o primeiro quadro possui dois centróides[2] (mostrados como círculos escuros) que são colocados aleatoriamente. O segundo quadro mostra que cada um dos itens é atribuído ao centróide mais próximo - nesse caso, A e B são atribuídos ao centróide superior e C, D e E são atribuídos ao centróide inferior. No terceiro quadro, cada centróide foi movido para a localização média dos itens que lhe foram atribuídos. Quando as atribuições são calculadas novamente, verifica-se que C está agora mais próximo do centroide superior, enquanto D e E permanecem mais próximos do centroide inferior. Assim, o resultado final é alcançado com A, B e C em um *cluster* e D e E no outro.

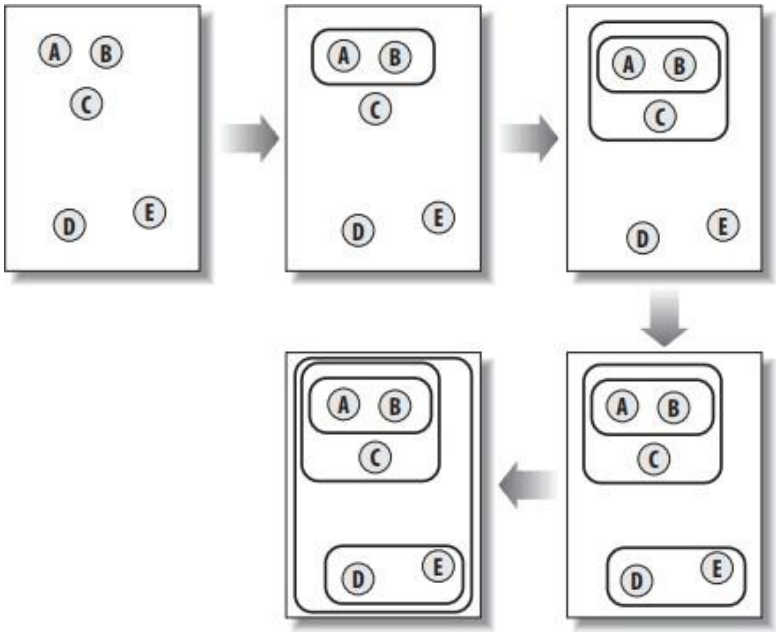
A maior desvantagem desse método é o fato de não haver uma estrutura de navegação que indique características mais abrangentes ou mais específicas, nem tampouco os inter-relacionamentos entre eles. Neste caso, pode ser difícil para o usuário localizar os documentos de que necessita.

[2] Um centróide é a representação de um documento que não existe; um conjunto de metadados que não representa um documento, servindo apenas para definir um ponto de referência em um conjunto de documentos. No modelo Vetorial um centróide é representado por um Vetor, da mesmo forma que um documento ou uma expressão de busca.

4.2.2 *Clustering* por partição hierárquica

O agrupamento hierárquico constrói uma hierarquia de grupos ao mesclar continuamente os dois grupos mais semelhantes. Cada um desses grupos começa como um único item e em cada iteração é calcula as distâncias entre cada par de grupos, sendo que os mais próximos são mesclados para formar um novo grupo. Isso é repetido até que haja apenas um grupo. A Figura 12 ilustra esse processo.

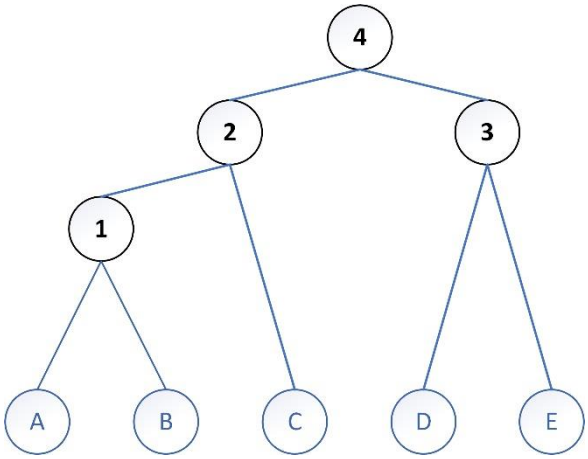
Figura 12 – Agrupamento por particionamento hierárquico



Fonte: SEGARAN (2007, p.33).

Na Figura 12, a similaridade dos itens é representada por suas localizações relativas. Quanto mais próximos dois itens mais semelhantes/similares eles são. No início, temos apenas itens (documentos) individuais. Na segunda etapa, os itens A e B fundiram-se para formar um novo *cluster*. Na terceira etapa, este novo grupo é mesclado com C. Uma vez que D e E são agora os dois itens mais próximos, eles formam um novo grupo. A etapa final unifica os dois grupos restantes. A representação hierárquica final do processo apresentado na Figura 12, pode ser ilustrado como na Figura 13

Figura 13 – Representação do processo de partição hierárquica



Fonte: o Autor

Os métodos de *clustering* hierárquico podem ser classificados como *Divisivo* ou *Aglomerativo*. No método divisivo, um único *cluster* inicial é dividido em grupos cada vez menores de documentos, encontrando diferenças entre os documentos dentro dos *clusters.* Nos métodos aglomerativos, a estrutura do *cluster* é construída por sucessivas fusões de *clusters*, como no caso da ilustração da Figura 12 (MURESAN, 2002).

Segundo experimentos realizados por Willett (1988), embora o *clustering* hierárquico de documentos exija muito recurso computacional, ele pode aumentar a eficácia da recuperação.

4.3 Algoritmos de *Clustering*

Clustering tornou-se um tema cada vez mais importante na área de mineração de dados e recuperação de informações com a explosão de informações disponíveis na internet. A organização dessa vasta quantidade de dados em grupos ou *clusters* relacionados pode ajudar os usuários a acessar um volume cada vez maior de dados mais efetivamente.

De modo geral, os métodos descrevem algoritmos como etapas de forma genérica, sem referência a detalhes de implementação, como os dados da estrutura empregada, ou o uso de memória e armazenamento de computador. Cada método pode ser implementado por uma variedade de algoritmos, que diferem em termos com a estrutura de dados que pode ser uma estrutura de dados especializada podendo melhorar a velocidade de acesso, mas aumentar a

complexidade do código e diminuir a flexibilidade e adaptabilidade do o algoritmo para coleções diferentes ou outras condições.

O uso de memória também pode ser considerado um diferencial pois menor complexidade computacional e maior velocidade pode ser obtidas se toda a estrutura de dados for armazenada na memória e se a memória também for usada para resultados intermediários. A desvantagem é a necessidade de computadores com grande memória ou, alternativamente, uma limitação severa no tamanho da coleção a ser agrupada. Ainda nesse âmbito podemos destacar com diferencial o uso de armazenamento que é geralmente equilibrado em relação ao uso de memória. No entanto, para coleções muito grandes, é inviável armazenar arquivos totalmente invertidos ou outras estruturas na memória, portanto, o acesso rápido aos dados no disco é essencial. O uso de estruturas de dados apropriadas, de armazenamento em cache e de compressão podem fazer uma grande diferença (ZOBEL; MOFFAT, 2006).

No intuito de combinar vantagens específicas que apresentam, vários algoritmos podem ser combinados, especialmente quando nenhum método pode ser aplicado com resultados satisfatórios. Para recuperação mediada, tal caso é a estruturação de coleções especializadas muito grandes em vista da exploração. Os métodos de agrupamento hierárquico são inviáveis devido à sua complexidade. Por outro lado, os métodos de particionamento produziriam muitos ou grandes *clusters* para viabilizar a exploração.

Em Croft (1977) podemos observar a aplicação de um algoritmo de *clustering* não hierárquico de passagem única rápida para particionar a coleção em uma série de grandes *clusters* nos quais um método de agrupamento hierárquico pode ser aplicado posteriormente. Uma desvantagem é que os métodos de *clustering* de passagem única são heurísticas e dependem da ordem de entrada dos documentos, portanto podem degradar a qualidade do *clustering*.

Jardine e Van Rijsbergen (1971) mostram um *clustering* principal, em um subconjunto de amostra de documentos, seguido pela atribuição dos outros documentos para os *clusters* resultantes, usando uma correspondência de *cluster* de documento e uma estratégia de função de busca descendente. Alguns problemas surgiram ao tentar encontrar uma amostra grande o suficiente para representar o *cluster* principal e o fato de adicionar documentos para usar uma estratégia de pesquisa heurística podem degradar a qualidade do *clustering*.

Em Van Rijsbergen (1979), os arquivos invertidos geralmente construídos durante a indexação dos documentos, oferecem pistas sobre quais valores de similaridade não precisam ser calculados: a similaridade entre dois documentos que não têm termos em comum é obviamente zero.

A ideia de usar o arquivo invertido para calcular uma matriz de similaridade esparsa, proposta por Croft (1977), funciona bem em termos de eficiência quando são utilizadas descrições curtas de documentos. No caso da exaustividade da indexação, quando os documentos são representados por um grande número de termos, um grande número de coeficientes de valor diferente de zero é calculado repetidamente, como pares de documentos estão na lista de postagem associada a cada termo de índice que eles compartilham, com um aumento substancial no tempo de execução (HARDING; WILLETT, 1980).

Peter Willett (1980) melhorou este algoritmo, evitando o cálculo de valores de similaridade redundantes. Croft (1977) também mostrou que, ao ignorar as listas de postagem mais longas no arquivo invertido contendo os termos de índice mais comuns, com baixo valor de discriminação, o número de os valores de similaridade calculados pode ser reduzido significativamente. Obviamente que tal ação introduz um erro, que foi demonstrado em experiencia, não afetar o resultado do *clustering*. Uma ação semelhante seria alterar o algoritmo acima no sentido de definir o valor zero para a semelhança entre cada par de documentos cujo número de documento está abaixo de um certo limite (CROFT, 1977).

Uma redução adicional do número de valores de similaridade a serem calculados é considerar apenas os ***k*** vizinhos mais próximos de cada documento e ignorar as outras similaridades. O chamado *clustering* de vizinhos mais próximos, tem sido amplamente empregado no *clustering* de documentos e mostraram que esta abordagem traz uma grande melhoria na eficiência, sem perda de eficácia (SMEATON *et al.*, 1998)

Peter Willett (1996) foi mais longe, defendendo o uso de *Clustering* de vizinhos mais próximos com boa eficiência em comparação com métodos de *Clustering* aglomerativo hierárquico e eficácia comparável e complementar à pesquisa convencional de melhor correspondência.

Outra questão é a matriz de similaridade, contendo a similaridade entre cada par de *clusters*. Conceitualmente, essa matriz é diferente da matriz de similaridade entre documentos. No entanto,

para métodos de *clustering* aglomerativo, uma estrutura de dados matricial é suficiente. Inicialmente, contém as semelhanças computadas entre os documentos, que são exatamente as semelhanças entre os *clusters* diferentes, no início do processo de agrupamento. Quando dois *clusters*, K e L, são fundidos pelo algoritmo de *clustering*, a entradas na matriz de similaridade correspondentes ao novo *cluster* KL podem sobrescrever aqueles correspondendo a K e L, portanto, nenhum armazenamento adicional é necessário durante o *cluster*. Em segundo lugar, as entradas para KL podem ser calculadas a partir das entradas para K e L, com base em fórmulas, sem recurso à semelhança do documento original matriz (WILLETT, 1996).

5.
Crawler Focado e Técnicas de *Clustering* na Recuperação de Informação

A recuperação de informação pode ser vista como um processo de comunicação entre um conjunto de documentos e os usuários que buscam por informações registradas nesses documentos. Esse processo é intermediado por um sistema cuja finalidade é interpretar a necessidade de informação do usuário e apresentar a ele um conjunto de documentos que possivelmente trará a informação desejada. Assim como todo processo de comunicação, a recuperação de informação é um processo fortemente influenciado por problemas linguísticos que afetam a qualidade dos resultados. Tais problemas poderiam ser reduzidos por meio de recursos que permitam a contextualização e a redução do campo semântico dos termos utilizados nesse processo.

Atualmente, os principais mecanismos de busca são de propósito geral, com pretensões de proporcionar acesso a toda e qualquer informação disponível na Web. Os *crawlers* utilizados por esses mecanismos tem por objetivo criar um acervo documental o mais abrangente possível, abarcando todos os assuntos e temas veiculados nas páginas Web. Essa pretensão à universalidade afeta diretamente na eficiência e na eficácia de tais ferramentas, refletindo negativamente na precisão de seus resultados.

Uma das premissas que norteiam esta dissertação é que os Web *crawlers* focados permitem a criação de uma base documental qualificada, contextualizada, restrita a um determinado tema ou

assunto. Esse conjunto de documentos restrito a uma temática específica tem o potencial de oferecer a um usuário interessado no tema resultados mais relevantes quando comparados com os resultados que seriam oferecidos por mecanismos de busca genéricos.

A Figura 13 apresenta uma ilustração do resultado da utilização de um Web *crawler* focado. A área em verde representa o conjunto de documentos sobre um determinado tópico, um assunto específico ou uma área do conhecimento.

Figura 14 – Ilustração do resultado de um Web *Crawler* Focado

Fonte: o Autor

A utilização de Web *crawlers* focados permite a criação de acervos documentais restritos a um tema ou assunto, possibilitando o desenvolvimento de mecanismos de busca temáticos, nos quais os documentos e as necessidades de informação dos usuários ficam contextualizados *a priori*. Em tais mecanismos, muitos problemas linguísticos, inerentes ao processo de recuperação de informação, podem ser minimizados ou dirimidos.

Atualmente, os mecanismos de busca utilizam diversos *crawlers* que rastreiam diferentes áreas ou domínios da Web. O uso de vários *Crawlers* focados em diferentes temas ou assuntos poderiam rastrear a Web para criar acervos documentais temáticos, permitindo o desenvolvimento de mecanismos de busca no qual o usuário escolhe e delimita o assunto de seu interesse, dirimindo ou minimizando problemas linguísticos inerentes ao processo de recuperação de informação. Assim, seria possível o desenvolvimento de mecanismos de busca no qual o usuário seleciona o

tema ou assunto de sua necessidade de informação, antes de executar a sua busca. Utilizando a tela de busca do Google a Figura 15 ilustra como seria um mecanismo de busca com tais característica.

Figura 15 – Ilustração de um mecanismo de busca temático

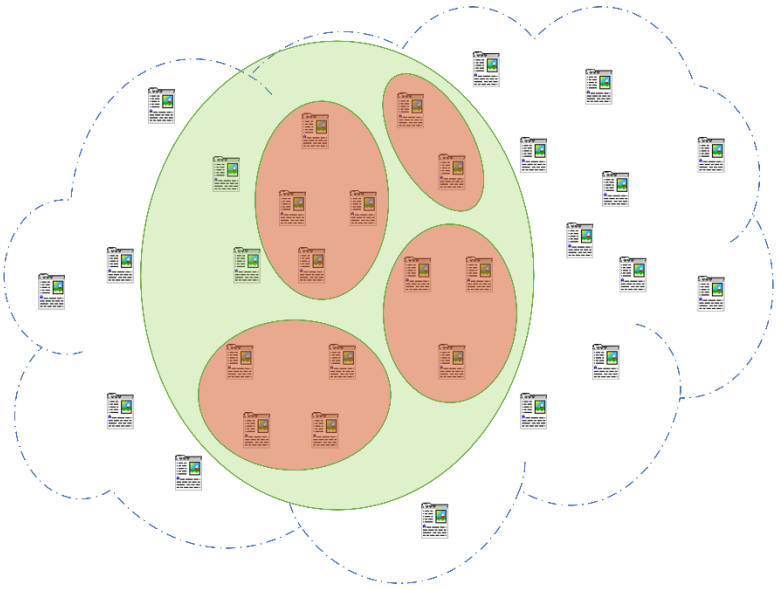
Fonte: adaptado de google.com.br

Devido à quantidade de informação disponível na Web, mesmo utilizando um conjunto de documentos restrito a um tema ou assunto, fornecido por um *crawler* focado, uma busca pode ainda resultar um grande número de documentos. Dentre esses resultados, um usuário poderá ter dificuldades em selecionar os documentos que possivelmente irão suprir a sua necessidade de informação. Assim, este trabalho propõe a utilização de técnicas de *clustering* para a organização dos documentos resultantes de uma busca, considerando um acervo documental cuja temática já está delimitada pela utilização de um c*rawler* focado.

Os grupos de documentos são formados de acordo com as associações entre suas características comuns. Esses grupos permitem ao usuário obter uma compreensão sobre a subcoleção de documentos recuperados após uma consulta.

A Figura 16 apresenta uma ilustração de grupos de documentos em um acervo documental restrito a um determinado assunto ou tema. A área em verde representa documentos que versam sobre um determinado tema, criado a partir do processamento de Web *crawlers* focados. As áreas em vermelho representam os grupos (*clusters*) de documentos que foram destacados no conjunto de documentos resultantes de uma determinada busca.

Figura 16 – Ilustração do resultado do processo de *clustering*



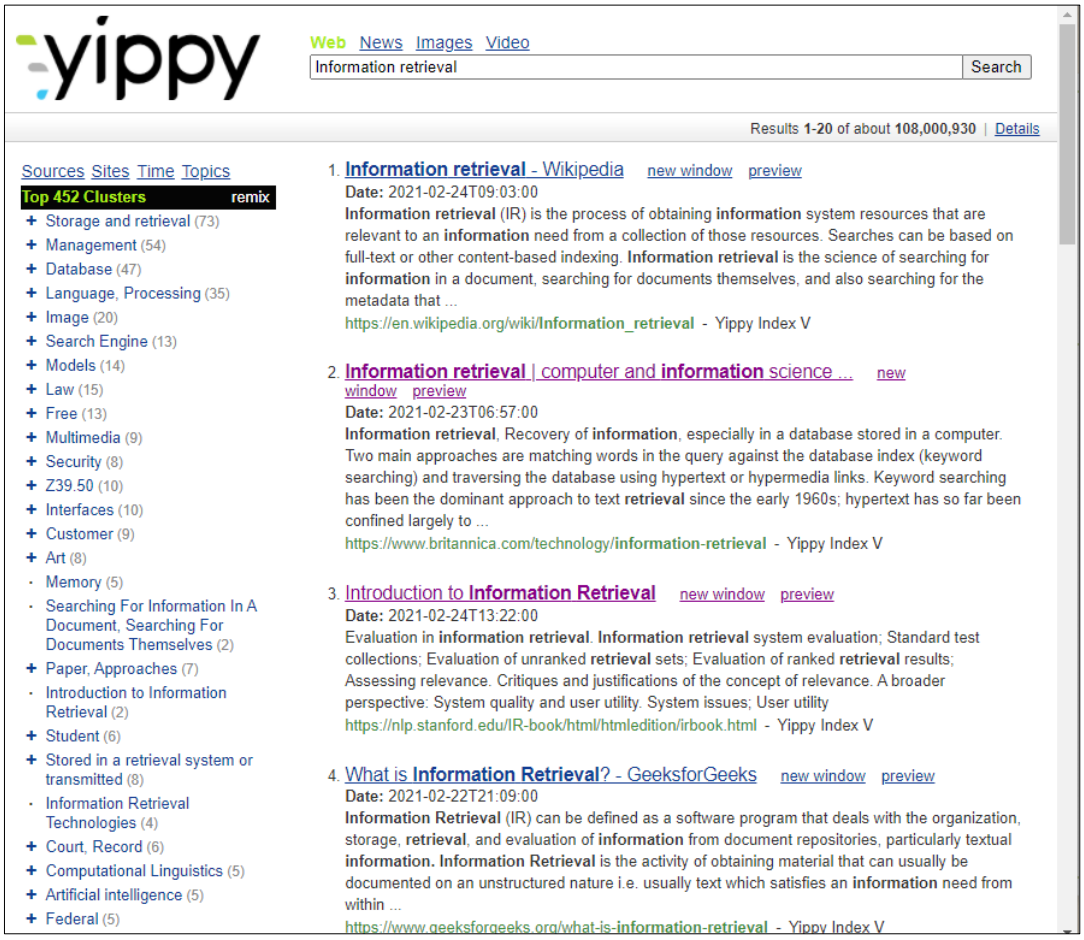
Fonte: o Autor

É importante observar que os grupos (*clusters*) são formados por documentos similares, relacionados a um determinado tema ou assunto previamente estabelecido no processamento do *crawler* focado. Assim, os *clusters* representam os diversos subtemas, detalhamentos ou especificidades do tema ou assunto geral da busca realizada pelo usuário.

Os algoritmos de agrupamento requerem uma maneira de determinar quão semelhantes (ou quão diferentes) dois itens são. No *clustering* textual, os documentos são geralmente representados por um conjunto de termos com seus respectivos pesos. A semelhança (ou a diferença) entre dois documentos é medida por uma função de similaridade, como exemplificado na Seção 4.2.

A abordagem padrão para agrupamento em sistemas de recuperação de informação tem sido particionar uma coleção em um grande número de pequenos *clusters* com o objetivo de direcionar a expressão de busca para o *cluster* mais semelhante. A partir de uma interface apropriada, um sistema pode prover ao usuário uma forma de visualizar e de navegar por entre os grupos. Um exemplo de mecanismo de busca que utilizava uma interface baseada em *clusters* é o buscador YIPPY (Figura 17).

Figura 17 – Mecanismo de Busca YIPPY



Fonte: https://yippy.com/

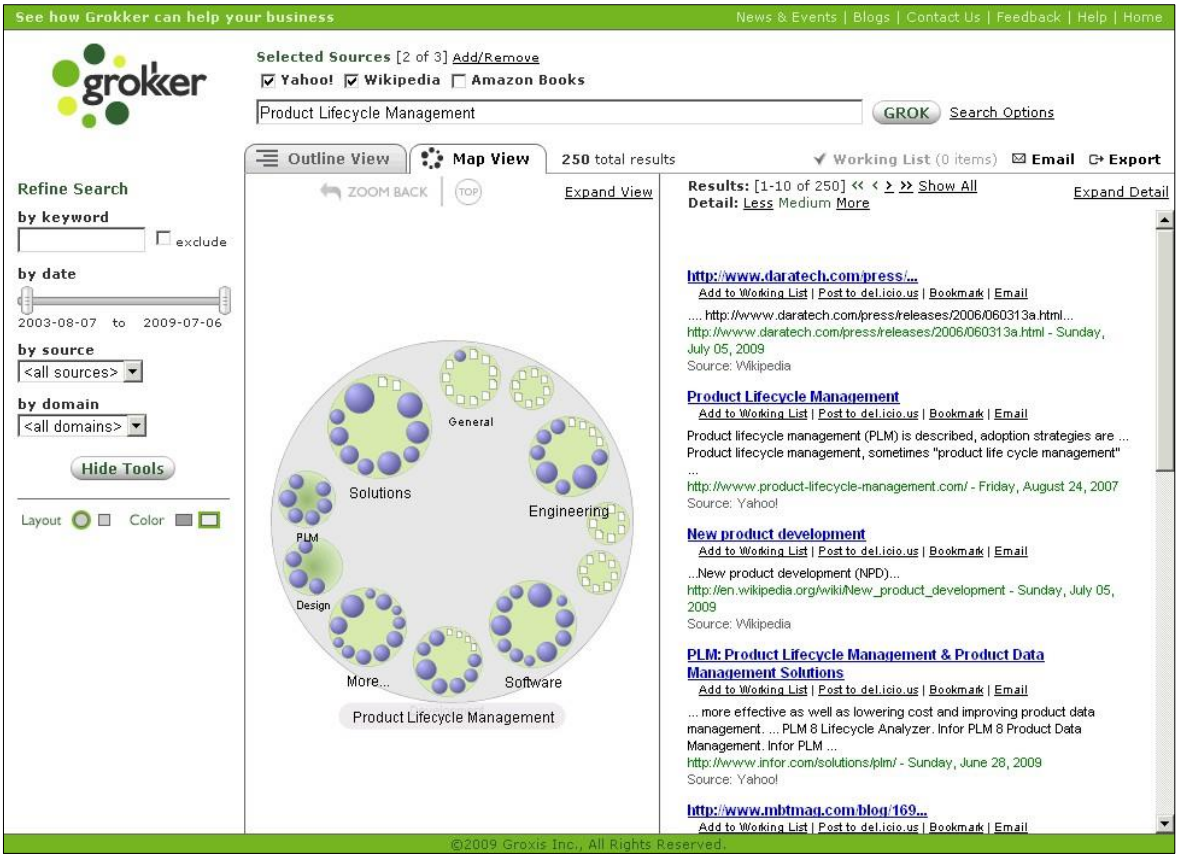
O Yippy (Figura 17) era um metabuscador que agrupava os resultados em *clusters*. Ele foi originalmente desenvolvido e lançado em 2004 sob o nome Clusty. Foi adquirido pela IBM e vendida em 2010 para uma empresa agora chamada Yippy Inc. Em agosto de 2019, a página principal do Yippy afirmava que suas buscas utilizavam o IBM Watson. No final de abril de 2021, o site **yippy.com** começou a ser redirecionado para o mecanismo de busca denominado DuckDuckGo. O destino e o futuro do site Yippy são desconhecidos.

Como pode ser observado na Figura 16, os resultados da busca "information retrieval" é apresentado em forma de lista, como nos principais mecanismos de busca. Porém, no lado esquerdo é apresentado uma lista de *clusters* de assuntos similares ao que foi pesquisado.

Um outro exemplo de metabuscador que utilizava uma interface baseada é cluster é o Grokker (Figura 18). O Grokker (KOSHMAN, 2006) foi uma metabuscador que oferecia uma

representação gráfica dos seus resultados. O mecanismo de agrupamento apresenta os documentos como uma série de círculos aninhados. Os usuários podem navegar na hierarquia clicando em um círculo, que era então ampliado, tornando o seu conteúdo visível.

Figura 18 - Mecanismo de busca Grokker



Fonte: (KOSHMAN, 2006)

O Grokker (Figura 18) foi projetado para ser um meta-buscador que apresenta os resultados de busca de uma forma gráfica. Ele executa as buscas nas principais ferramentas de busca e agrupa seus resultados, apresentando os grupos de forma visual. A empresa Groxis, desenvolvedora do sistema, enfatizava que o Grokker é uma ferramenta que facilita uma abordagem holística para entender informações complexas e desconectadas, permitindo a descoberta de relações inicialmente não conhecidas. Embora tenha recebido diversos prêmios de indústria dos Estados Unidos, a Groxis, fundada em 2001, encerrou suas atividades em março de 2009.

Percebe-se que tanto o Yippy quanto o Grokker apresentam um grande número de resultados, necessitando por parte do usuário um esforço significativo para encontrar os

documentos que considera relevantes. Tal dificuldade é resultado da inexistência de um "filtro" que restrinja a quantidade de documentos que os mecanismos de busca devem processar. A utilização de *crawler* focado permitiria uma filtragem nessa grande quantidade de documentos, restringindo-se àqueles cujo assunto ou cuja temática estejam em conformidade com o que foi estabelecido ou escolhido pelo usuário. Essa redução da quantidade de documentos permitiria uma maior agilidade nos mecanismos de busca e permitiria um direcionamento de recursos computacionais para tratar da forma de apresentação dos resultados, particularmente o agrupamento (*clustering*) de documentos resultantes da consulta do usuário.

6.
Considerações Finais

A expansão acelerada da Web e sua inerente dinamicidade impõe significativos desafios no desenvolvimento e no aprimoramento de recursos tecnológicos para a sua efetiva utilização como fonte de informação e conhecimento. Embora os mecanismos de busca tenham assumido um papel importante para a recuperação de informação nesse ambiente, diversas ideias, conceitos e resultados de pesquisas realizadas muito antes do surgimento da Web não podem ser negligenciadas ou relegadas ao esquecimento.

Este trabalho propõe a utilização conjunta de um recurso tecnológico surgido no contexto da Web, os *crawlers*, com uma área de pesquisa surgida desde os primórdios da Recuperação de Informação como área de pesquisa, o agrupamento ou *clustering*. Esses dois campos de pesquisa são utilizados de forma complementar. Essa simbiose permite a proposição de novos modelos e técnicas de recuperação de informação para o ambiente Web.

Para a consecução desse objetivo, alguns passos foram necessários. Primeiramente foi realizado um estudo sobre os métodos e técnicas utilizados no desenvolvimento de mecanismos de busca, tais como a indexação, a classificação e alguns alguns detalhes de implementação dos diversos tipos de Web *crawlers*. Em um segundo momento foi apresentado as principais ideias e técnicas relacionadas às pesquisas sobre *clustering*. Por fim foi proposto a utilização conjunta das tecnologias relacionadas aos Web *crawlers* focados com as ideias, métodos e técnicas advindas das pesquisas sobre *clustering*.

A partir de uma pesquisa exploratória, este trabalho analisou a utilização de *Web Crawlers* focados juntamente com técnicas de *clustering* como recursos complementares no processo de recuperação de informação. O *Crawler* focado fornece um *corpus* temático no qual o processo classificatório dos algoritmos de *clustering* podem gerar grupos de documentos relacionados às especificidades ou detalhamentos do tema.

No Capítulo 2 foi apresentado o processo de recuperação de informação e abordado tal processo no contexto da Web, dando resposta aos objetivos específicos **a.** e **b.**

No Capítulo 3, em resposta ao objetivo **c.**, abordou o funcionamento dos web crawlers, especificamente dos *crawlers* focados.

No Capítulo 4 são apresentadas as principais técnicas de *clustering*, como desenvolvimento do objetivo **d.** .

Por fim, no Capítulo 5 foi apresentada e demonstrada as funcionalidades e as vantagens na utilização de *crawlers* focados e técnicas de clustering no processo de recuperação de informação.

Acredita-se que a base teórica apresentada neste trabalho se configura como uma proposta para a implementação futura de um mecanismo de busca com recursos diferenciados, contribuindo assim para as pesquisas em Recuperação de Informação.

Referências

AGGARWAL, C. C.; AL-GARAWI, F.; YU, P. S. Intelligent Crawling on the World Wide Web with Arbitrary Predicates. **Proceedings of the 10th International Conference on World Wide Web, WWW 2001**, p. 96–105, 2001a.

AGGARWAL, C. C.; AL-GARAWI, F.; YU, P. S. On the design of a learning crawler for topical resource discovery. **ACM Transactions on Information Systems (TOIS)**, v. 19, n. 3, p. 286–309, 2001b.

ALTOUNIAN, M. M. de A.; GOMES, B. P. de M. A recuperação semântica da informação no contexto do controle externo. **Revista do TCU**, n. 137, p. 31-41, 2016.

ANTONIOU, G.; HARMELEN, F. Van. **A semantic Web primer.** Cambridge: MIT Press, 2004.

ARAUJO JUNIOR, R. H. Precisão no processo de busca e recuperação da informação. Brasília: Thesaurus, 2007.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**: the concepts and technology behind search. 2nd ed. New York: Addison Wesley, 2011

BAEZA-YATES, R.; RIBEIRO-NETO, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. Porto Alegre:Bookman, 2013.

BATZIOS, A. et al. BioCrawler : An intelligent crawler for the semantic Web. v. 35, p. 524–530, 2008.

BENJAMIN, K. et al. **A Strategy for Efficient Crawling of Rich Internet Applications**. International Conference on Web Engineering. **Anais**...Paphos, Cyprus: 2011

BEPPLER, F. D. **Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico**. 2008. 123 f. Tese (Doutorado em Engenharia e Gestão do Conhecimento) – Universidade Federal de Santa Catarina, Florianópolis, 2008.

BHATT, D.; VYAS, D. A.; PANDYA, S. Focused Web Crawler . **Advances in Computer Science and Information Technology (ACSIT)**, v. 2, n. 11, p. 1–6, 2015.

BOLDI, P. et al. UbiCrawler : a scalable fully distributed Web crawler . v. 726, n. March, p. 711–726, 2004.

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3-5, 1968.

BRANSKI, R. M. Recuperação de informações na Web. **Perspectivas da Ciência da Informação**, Belo Horizonte, v. 9, n. 1, p. 70-87, jan./jun. 2004.

BRÄSCHER, M. A ambigüidade na recuperação da informação. **DataGramaZero: Revista de Ciência da Informação**, v. 3, n. 1, fev. 2002.

BREGAGNOL, C. S.; LARA, L. Z. **Uma análise do tratamento da semântica nos livros didáticos**. 2012. Monografia (Especialização em Gramática Ensino de Língua Portuguesa) - Universidade Federal do Rio Grande do Sul, Instituto de Letras, Porto Alegre, 2012.

BRIN, S.; PAGE, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. p. 1994–2000, 2003.

BUCKLAND, Michael Keeble. Information as thing. **Journal of the American Society for Information Science** (JASIS), v.45, n.5, p.351-360, 1991

CALISKAN, K.; OZCAN, R. Comparing classification methods for link context based focused crawlers. **2013 International Conference on Electronics, Computer and Computation, ICECCO 2013**, p. 143–146, 2013.

CASTRILLO-FERNÁNDEZ, O. Web Scraping : Applications and Tools. **European Public Sector Information Platform Topic Report No. 2015 / 10**, p. 1–31, 2015.

CHAITRA, P. G. et al. A study on different types of Web crawlers. In: **Intelligent communication, control and devices**. Springer, Singapore, 2020. p. 781-789.

CHAKRABARTI, S.; VAN DEN BERG, M.; DOM, B. Focused crawling: A new approach to topic-specific Web resource discovery. **Computer Networks**, v. 31, n. 11, p. 1623–1640, 1999.

CHOUDHARY, S. **M-Crawler : Crawling Rich Internet Applications Using Menu Meta-Model**. [s.l.] Université d'Ottawa/University of Ottawa, 2012.

COOPER, W. S. Gerard Salton," Automatic Information Organization and Retrieval"(Book Review). **The Library Quarterly**, v. 39, p. 276, 1969.

CRISTOVÃO, H. M.; DUQUE, C. G.; SERQUEIRA, L. D. Recuperação de informação: uma aplicação na criação e configuração automáticas de cursos virtuais a distância. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 13., 2012, Rio de Janeiro. **Anais…** Rio de Janeiro: Fiocruz, 2012.

CROFT, W. B. Clustering Large Files of Documents Using the Single-Link Method. **Journal of American Society for Information Science**, n.28, p.341-44, 1977.

CUTTING, D. R. *et al*. A cluster-based approach to browsing large document collections. In: **ACM SIGIR Forum. New York, NY, USA: ACM**, p. 148–149, 2017.

DHENAKARAN, S. S.; SAMBANTHAN, K. T. Web crawler - an Overview. **International Journal of Computer Science and Communication**, v. 2, n. 1, p. 265–267, 2011.

DILIGENTIT, M. et al. Focused crawling using context graphs. Proceedings of the 26th International Conference on Very Large Data Bases, VLDB'00, p. 527–534, 2000.

EXPOSTO, J. et al. **Geographical Partition for Distributed Web Crawling**. Proceedings of the 2005 workshop on Geographic information retrieval. **Anais**...New York, NY, USA: 2005

FERNEDA, E. Introdução aos Modelos Computacionais de Recuperação de Informação. Rio de Janeio: Editora Ciência Moderna. 2012.

FERNEDA, E. **Ontologia como recurso de padronização terminológica de um sistema de recuperação de informação**. 2013. 96 f. Relatório de Pesquisa (Pós-Doutorado em Ciência da Informação) – Universidade Federal da Paraíba, João Pessoa, 2013.

GASQUE, K. C. G. D.; COSTA, S. M. de S. Evolução teórico-metodológica dos estudos de comportamento informacional de usuários. **Ciência da Informação**, Brasília, v. 39, n. 1, p. 21-32, jan./abr. 2010.

HAMILL, K.; ZAMORA, A. The use of titles for automatic document classification. **Journal of the American Society for Information Science**, v. 31, p. 396–402, 1980.

HARDING, A. F; WILLETT, P. Indexing exhaustivity and the computation of inter-document similarity matrices. **Journal of the American Society for Information Science**, n. 31, p. 298-300, 1980.

HATI, D.; SAHOO, B.; KUMAR, A. **Adaptive focused crawling based on link analysis**. 2010 2nd International Conference on Education Technology and Computer. **Anais**...IEEE, jun. 2010.

HEARST, M. A. THE USE OF CATEGORIES AND CLUSTERS FOR ORGANIZING RETRIEVAL RESULTS. In: **Natural language information retrieval**. p. 333–374, 1999.

HEYDON, A.; NAJORK, M. Mercator : A Scalable , Extensible Web Crawler . **World Wide Web**, v. 2, n. Springer, p. 1–14, 2003.

HOGAN, A. et al. Searching and browsing Linked Data with SWSE: the semantic Web search engine. **Web Semantics**: Science, Services and Agents on the World Wide Web, v. 9, n. 4, p. 365-401, dez. 2011.

HUANG, L. **A survey on Web information retrieval technologies**. New York, University of New York, 1999.

HUO, L. Y.; LIU, B.; YAN, F. Similarity computation of Web pages of focused crawler . **Proceedings - 2010 International Forum on Information Technology and Applications, IFITA 2010**, v. 2, p. 70–72, 2010.

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications**: text retrieval, extraction and categorization. John Benjamins, Amsterdam, Philadelphia, 2002.

JARDINE, N.; VAN RIJSBERGEN, C. J. Cornelis Joost. The use of hierarchic clustering in information retrieva. **Information storage and retrieval,** v. 7, p. 217–240, 1971.

JONES, K SPARCK NEEDHAM, R. M. Automatic term classifications and retrieval. **Information Storage and Retrieval**, v. 4, n. 0020–0271, p. 91–100, 1968.

KAUSAR, Mohammad Abu Kausar; DHAKA, Vijaypal Singh; SINGH, Sanjeev Kumar. Web Crawler : A Review. **International Journal of Computer Applications**, v.63, n.2, p.31-36, 2013.

KOSHMAN, S. Visualization-based information retrieval on the Web. **Library & Information Science Research**. 28. p.192-207, 2006.

KISHIDA, Kazuaki. Techniques of document clustering: A review. **Library and Information Science**, n. 49, p. 33-75, 2003.

KHARE, R.; CUTTING, D.; SITAKER, K; RIFKIN, A. Nutch: A flexible and scalable open-source Web search engine. Oregon State University, 2004.

KUMAR, M.; BHATIA, R.; RATTAN, D. A survey of Web crawlers for information retrieval. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 77, n. Wiley Online Library, p. e1218, 2017.

LANCASTER, F.W. **Information Retrieval System**. New York: John Wiley, 1968.

LEE, H. et al. IRLbot : Scaling to 6 Billion Pages and Beyond Presented by Xiaoming Wang • Introduction and challenges. **ACM Transactions on the Web (TWEB)**, v. 3, n. ACM New York, NY, USA, p. 1–39, 2008.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. **Ciência da informação**, Brasília, v. 31, n. 2, p. 60-71, maio/ago. 2002.

MCBRYAN, O. A. **GENVL and WWWW : Tools for Taming the Web**. Proceedings of the first international world wide Web conference. **Anais**...Citeseer: 1994

MEDELYAN, O. et al. Language specific and topic focused Web crawling. **Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006**, n. 2000, p. 865–868, 2006.

MIRTAHERI, S. M. et al. A Brief History of Web Crawlers. **arXiv preprint arXiv:1405.0749**, 2014.

MOOERS, C. Zatocoding applied to mechanical organization of knowledge**. American Documentation**, v.2, n.1, p.20-32, 1951.

MURESAN, G. Using document clustering and language modelling in mediated information retrieval. Robert Gordon University, PhD thesis, 2002.

NAYAK, T.; PRASAD, S.; SENAPATI, M. R. A survey on Web text information retrieval in text mining. **Research Journal of Applied Sciences, Engineering and Technology**, v. 10, n. 10, p. 1164-1174, 2015.

OLSTON, C.; NAJORK, M. Web crawling. **Foundations and Trends in Information Retrieval**, v. 4, n. 3, p. 175–246, 2010.

PANT, G. et al. **Panorama : Extending Digital Libraries with Topical**. Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004. **Anais**...2004

PANT, G.; SRINIVASAN, P. Learning to crawl: Comparing classification schemes. **ACM Transactions on Information Systems**, v. 23, n. 4, p. 430–462, 2005.

PANT, G.; SRINIVASAN, P. Link contexts in classifier-guided topical crawlers. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 1, p. 107–122, 2006.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v. 1, n. 1, p. 41-62, jan./jul. 1996.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**, v.24, n.1, p.36-41, 1995.

SARACEVIC, T. The stratified model of information retrieval interactions: extension and applications. **American Society for Information Science**, p. 313-327, 1997.

SARMIENTO, Fernando Iván Camargo; SALINAS, Sonia Ordóñez. Evolución y tendencias actuales de los Web crawlers. **Ingeniería**, v. 18, n. 2, 2013.

SEGARAN, Toby. Programming collective intelligence: building smart Web 2.0 applications. " O'Reilly Media, Inc.", 2007.

SHKAPENYUK, V.; SUEL, T. **Design and Implementation of a High-Performance Distributed Web Crawler** . Proceedings 18th International Conference on Data Engineering. **Anais**...San Jose, CA, USA: IEEE, 2002

SHNEIDERMAN, B. et al. **Designing the user interface**: strategies for effective human- computer interaction. 6. ed. Hoboken: Pearson, 2017.

SILVA, E. L. DA; MENEZES, E. M. Metodologia da Pesquisa e Elaboração de Dissertação. **Portal**, v. 29, n. 1, p. 121, 2001.

SMEATON, A.F.; BURNETT, M.; KELLEDY, F; QUINN, G. An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts. In Proceedings of the 20th BCS-IRSG Colloquium, Grenoble, France, Springer-Verlag Workshops in Computing, April 1998

SOMBOONVIWAT, K.; KITSUREGAWA, M.; TAMURA, T. **Simulation Study of Language Specific Web Crawling**. 21st International Conference on Data Engineering Workshops (ICDEW'05). **Anais**...IEEE, 2005.

SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na Web: panorama atual e tendências. **Perspectivas em ciência da informação**, v. 11, p. 161-173, 2006.

SPINK, Amanda *et al*. Searching the Web: the public and their queries. **Journal of the American Society for Information Science**, v. 52, n. 3, p. 226-234, 2001.

SUBHASHINI, R.; KUMAR, V. J. S. Evaluating the performance of similarity measures used in document clustering and information retrieval. In: **First international conference on Integrated intelligent computing (ICIIC)**, p. 27–31, 2010.

VAN RIJSBERGEN, C. J. **Information retrieval**. 2nd ed. London: Butterworths, 1979.

WILLETT, P. Document clustering using an inverted file approach. **Journal of Information Science**, v.2, n.5, 1980.

WILLETT, P. Recent Trends in Hierarchical Document Clustering: a critical review. **Information Processing & Management**, v. 24, n.5, p. 577-597,1988.

WILLETT, P. Molecular diversity techniques for chemical databases. Information Research, v. 2, n. 3, 1996.

WOODS, W. A. **Searching versus finding**: why systems need knowledge to find what you really want. Sun Microsystems Laboratories, 2004

ZOBEL, Justin; MOFFAT, Alistair. Inverted files for text search engines. ACM Computing Surveys, New York, v. 38, n. 2, p.1-56, 2006.