UNIVERSIDADE ESTADUAL PAULISTA
Faculdade de Filosofia e Ciências de Marília
Departamento de Ciência da informação
Curso de Arquivologia

JUNIOR CRISTINO DA SILVA

Técnicas para o Tratamento da Subjetividade na Recuperação da Informação

Marília

2023

JUNIOR CRISTINO DA SILVA

Técnicas para o Tratamento da Subjetividade na Recuperação da Informação

Trabalho de Conclusão de Curso (TCC) apresentado ao Conselho de Curso de Arquivologia da Faculdade de Filosofia e Ciências – Unesp – Campus de Marília, para a obtenção do título de Bacharel em Arquivologia.

Linha de Pesquisa: Informação e Tecnologia

Orientador(a): Prof. Dr Edberto Ferneda

Marília
2023

JUNIOR CRISTINO DA SILVA

TÉCNICAS PARA O TRATAMENTO DA SUBJETIVIDADE NA RECUPERAÇÃO DA INFORMAÇÃO

Trabalho de Conclusão de Curso (TCC) apresentado ao Conselho de Curso de Arquivologia da Faculdade de Filosofia e Ciências – Unesp – Campus de Marília, para a obtenção do título de Bacharel em Arquivologia.

BANCA EXAMINADORA

Orientador: Profº Drº Edberto Ferneda (Orientador)
Departamento de Ciência da Informação – UNESP-Marília

Profª Drª Daniela Pereira dos Reis
Departamento de Ciência da Informação Universidade Estadual Paulista – UNESP-Marília

Profª Drª Natália Marinho do Nascimento
Departamento de Ciência da Informação Universidade Estadual Paulista – UNESP-Marília

Marília, dezembro de 2023.

S586t

Silva, Junior Cristino da

Técnicas para o Tratamento da Subjetividade na Recuperação de Informação / Junior Cristino da Silva. -- Marília, 2023

43 p. : il.

Trabalho de conclusão de curso (Bacharelado - Arquivologia) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília

Orientadora: Edberto Ferneda

1. Recuperação da Informação. 2. Sistemas de Recuperação da Informação. 3. Arquivologia. I. Título.

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

Essa ficha não pode ser modificada.

Agradecimentos

Agradeço primeiramente a Deus por ter me sustentado durante todo o processo da realização do curso da graduação, bem como toda minha família onde me deu todo o apoio para seguir fortemente até a conclusão do mesmo.

Agradeço também a todos os meus professores do curso, sobretudo a Profª. Drª. Marcia Pazin, que passaram todo o seu conhecimento, excelência e sempre me apoiaram durante todo o período, sem vocês nada disso seria possível.

Agradeço especialmente o meu orientador Prof. Dr. Edberto Ferneda pelo incentivo, dedicação e por ter me mostrado todos os caminhos para a realização do presente trabalho de conclusão de curso, sou grato pela sua bagagem de conhecimentos e por acreditar em mim.

RESUMO

O processo de recuperação de informação pode ser visto como um processo de comunicação entre um ser humano e um acervo de documentos, mediado por um sistema computacional. Como todo processo comunicacional, a recuperação de informação enfrenta problemas inerentes à comunicação, tais como a imprecisão, a ambiguidade e a subjetividade. Na Recuperação de Informação essas características estão inseridas e são representadas pelo conceito de "relevância". O presente trabalho tem como objetivo apresentar algumas técnicas utilizadas para tratar ou minimizar os problemas causados pela subjetividade dos processos e atores envolvidos na recuperação de informação. Trata-se, portanto, de um levantamento bibliográfico de natureza básica, com uma abordagem qualitativa e exploratória de algumas técnicas que foram criadas ao longo da história de pesquisa da Recuperação de Informação. Dentre essas técnicas, destacamos o processo de *Relevance Feedback*, que propõe um método de interação entre usuário e sistema a fim de se alcançar um resultado satisfatório. Outra técnica abordada é a *expansão de consultas*, que visa melhorar a eficiência da recuperação de informação por meio do aprimoramento das consultas realizadas pelos usuários. Em um contexto mais atual, na busca para o tratamento da subjetividade inerente ao processo de recuperação de informação, são apresentadas algumas propostas para a incorporação de diálogo como recurso de interação do usuário com o sistema. Sendo assim, a área de Recuperação de informação tem tentado diminuir essa subjetividade, através das novas tecnologias que estão sendo desenvolvidas e apresentando novas formas para que as pessoas possam ter informações úteis para elas através desse dialogo entre o usuário e o sistema.

Palavras-chave: Recuperação de Informação; Subjetividade; *Relevance Feedback*; Expansão de Consultas; *Chatbot*.

ABSTRACT

The information retrieval process can be seen as a communication process between a human being and a collection of documents, mediated by a computer system. Like every communication process, information retrieval faces problems inherent to communication, such as imprecision, ambiguity, and subjectivity. In Information Retrieval, these characteristics are represented by the concept of "relevance". The present work aims to show some techniques used to treat or minimize the problems caused by the subjectivity of the processes and actors involved in information retrieval. It is, therefore, a bibliographical survey of a basic nature, with a qualitative and exploratory approach to some techniques that have been created throughout the history of Information Retrieval research. Among these techniques, we highlight the Relevance Feedback process, which proposes a method of interaction between user and system in order to achieve a satisfactory result. Another technique addressed is Query Expansion, which aims to improve the efficiency of information retrieval by improving the queries carried out by users. In a more current context, in the search for treating the subjectivity inherent in the information retrieval process, some proposals are presented for the incorporation of dialogue as a resource for user interaction with the system. Therefore, the area of Information Retrieval has tried to reduce this subjectivity, through new technologies that are being developed and presenting new ways for people to have information that is useful to them through this dialogue between the user and the system.

Keywords: Information Retrieval; Subjectivity; Relevance Feedback; Query Expansion; Chatbot.

Lista de Figuras

Lista de Quadros

SUMÁRIO

1
Introdução

O processo de recuperação da informação envolve basicamente três elementos: por um lado, um conjunto de documentos que deve ser representado por expressões linguísticas que resumem seu conteúdo informacional. Por outro lado, seres humanos que tentam descrever linguisticamente as suas necessidades de informação a fim de obterem documentos que satisfaçam tais necessidades. O terceiro elemento se refere a alguma forma de comparação entre esses dois primeiros elementos.

A representação de um determinado documento inclui os elementos descritivos que o identificam e o caracterizam em um acervo, assim como os elementos indicativos de seu conteúdo informativo, os assuntos por ele tratados. A Ciência da Informação, tendo a informação como seu objeto de pesquisa, possui um conjunto de métodos e técnicas voltadas à representação documental que permitem tornar a informação acessível. Porém, tais metodologias levam em consideração um produto intelectual acabado, cuja descrição se evidencia por suas características físicas (título, autor, editora, etc) e seu conteúdo informacional pode ser sintetizado de forma manual (intelectual) ou por métodos automatizados.

Portanto, em todo sistema de recuperação de informação o acervo documental é constituído *a priori*, anterior a qualquer busca, sendo passível de ser processado por técnicas automatizadas tais como indexação automática, mineração de textos, entre outras. Por outro lado, a necessidade de informação do usuário só é percebida após a sua enunciação, por meio de uma expressão de busca. Somente a partir de sua enunciação, a expressão de busca pode ser

utilizada em processos interativos que visem a resolução de possíveis ambiguidades ou o seu enriquecimento semântico.

A necessidade de informação do usuário é um elemento tipicamente subjetivo, cuja interpretação é dificultada pelo tamanho (número de termos) reduzido das expressões de busca, não permitindo inferir automaticamente o contexto ou a intensão do usuário. Sendo assim, diversas técnicas e métodos interativos podem ser utilizados com a finalidade de auxiliar o usuário na tradução de sua necessidade de informação em uma expressão de busca que melhor a represente. Tais métodos assumem importância significativa na medida em que buscam contornar toda a subjetividade inerente ao processo de recuperação da informação, buscando representar com maior fidelidade a real necessidade do usuário.

Toda subjetividade do processo de recuperação de informação pode ser resumida no conceito de "relevância". A relevância é crucial na Recuperação de Informação, sendo muitas vezes utilizado na própria enunciação dos objetivos dessa área.

O presente trabalho tem como objetivo apresentar algumas técnicas utilizadas para tratar ou minimizar os problemas causados pela inerente subjetividade dos processos envolvidos na recuperação de informação. Nesse contexto, a subjetividade refere-se às interpretações pessoais, preferências e contextos individuais que podem influenciar a forma como as informações são buscadas, avaliadas e utilizadas.

O termo “subjetivo” é definido como o “que pertence ao sujeito pensante e a seu íntimo”; “pertinente a ou característico de um indivíduo; individual, pessoal, particular” [1]. Subjetivo é tudo aquilo que é próprio do sujeito ou a ele relativo. É algo que está baseado em uma interpretação individual.

Visando abordar a problemática, quais são as técnicas utilizadas para minimizar os problemas causados pela subjetividade no processo de recuperação da informação na arquivologia? O interesse por essa temática surgiu a partir das aulas na disciplina de Recuperação da Informação ministradas pelo Prof. Dr. Edberto Ferneda em 2021, especificamente pelo artigo “Recuperação da Informação e seus potenciais modelos” elaborado para o trabalho final da disciplina.

[1] FERREIRA, Aurélio Buarque de Holanda. Novo Dicionário da Língua Portuguesa.

O presente trabalho justifica-se pelas lacunas existentes na exploração bibliográfica-acadêmica sobre as técnicas para o tratamento da subjetividade na recuperação de informação, principalmente estudos na língua portuguesa.

Na história das pesquisas em Recuperação de Informação, diversos métodos e técnicas foram desenvolvidos para o tratamento da subjetividade inerente ao processo de recuperação de informação. Nesse sentido, o objetivo geral dessa pesquisa é apresentar as técnicas para o tratamento da subjetividade na Recuperação de Informação. Os objetivos específicos são, a) levantar e apresentar as técnicas desenvolvidas para automatizar tarefas subjetivas que compõem o processo de recuperação da informação e; b) avaliar quais as técnicas contribuem mais para automatizar tarefas subjetivas que compõem o processo de recuperação da informação no que tange os processos arquivísticos.

A metodologia utilizada compreendeu uma pesquisa básica, com uma abordagem qualitativa e de caráter exploratório, a partir de uma pesquisa bibliográfica, apresentando uma visão geral sobre técnicas para o tratamento da subjetividade na recuperação de informação.

1.1 Procedimentos metodológicos

Esse estudo tem por finalidade realizar uma pesquisa de natureza básica, uma vez que gera conhecimento, focando em teorias científicas já existentes.

Para alcançar os objetivos propostos, foi utilizada uma abordagem qualitativa. Na abordagem qualitativa, a pesquisa tem o ambiente como fonte direta dos dados. O pesquisador mantém contato direto com o ambiente e o objeto de estudo em questão, necessitando de um trabalho mais intensivo de campo (Prodanov; Freitas, 2013, p. 70).

Com intuito de conhecer a problemática sobre a área de estudo foi realizada uma pesquisa exploratória. Segundo Prodanov e Freitas (2013, p. 127), a pesquisa exploratória “visa proporcionar maior familiaridade com o problema, tornando-o explicito ou construindo hipóteses sobre ele”.

Para obtenção uma bibliografia básica, foi realizada pesquisas durante os meses de junho, julho e agosto de 2023. Com base no levantamento de dados bibliográficos presentes nos acervos de bases de dados BRAPCI e Repositório Institucional UNESP com as seguintes expressões de busca: Subjetividade AND "Recuperação da Informação"; Subjetividade AND "Recuperação de Informação" a partir da produção científica feitas nos últimos 10 anos (2013

– 2023) obteve-se um resultado de 7 (sete) trabalhos na base de dados BRAPCI, 81 no Repositório Institucional UNESP. Além dos trabalhos selecionados nessas bases, foram utilizadas diversas Teses e Dissertações e variados livros sobre Recuperação de Informação.

1.2 Organização do trabalho

A seção 1 apresentou uma contextualização desse trabalho, abordando o problema de pesquisa, a justificativa pessoal e acadêmica; os objetivos e os procedimentos metodológicos.

A seção aborda o conceito e origem sobre o termo “Recuperação de Informação” e descreve o processo de recuperação de informação. São apresentados os três modelos ditos "clássicos”, o modelo *booleano*, modelo vetorial e o modelo probabilístico.

A seção 3 aborda o conceito e a técnica denominada *Relevance Feedback* e apresenta os seus diferentes tipos.

A seção 4 aborda o conceito e as técnicas de expansão de consulta (*query expansion*), que visam melhorar a eficiência da recuperação de informação por meio da adição de novos termos à consulta inicialmente formulada pelo usuário.

A seção 5 aborda sobre a recuperação com base nos diálogos, onde o diálogo é usado como interface de busca em um processo de recuperação de informação e pode ser pensado como um recurso natural de se obter informações.

A seção 6 é apresentada uma análise a aplicação das técnicas de tratamento da subjetividade no contexto da Arquivologia e no sistemas de arquivo.

A seção 7, por fim, traz as considerações finais desse trabalho.

2
Recuperação de Informação

O termo Recuperação de Informação (*Information Retrieval*), foi criado por Calvin Mooers no início da década de 1950. Segundo Mooers (1951, p. 25, tradução nossa):

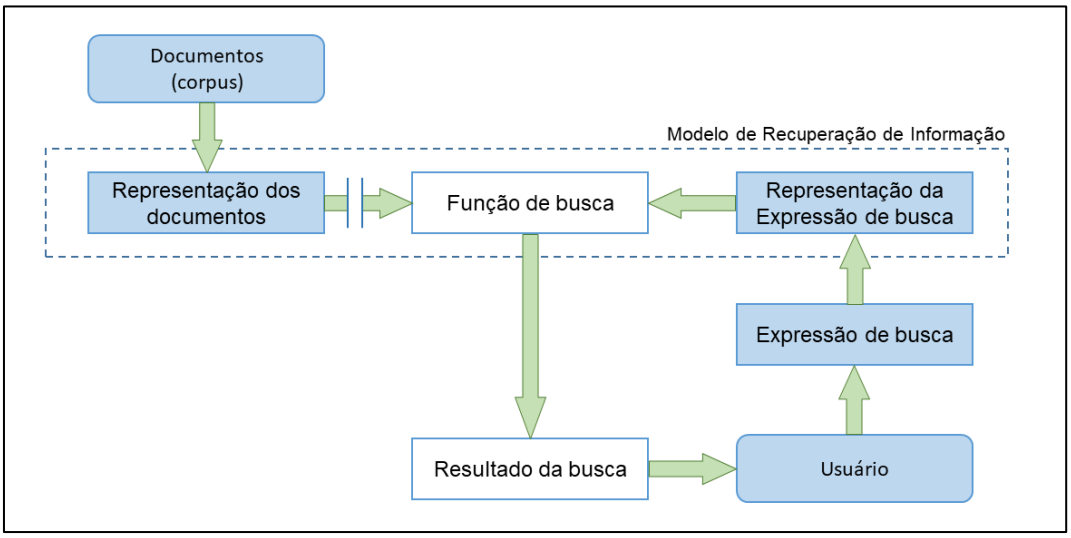
> Recuperação de informação é o nome dado ao processo ou método pelo qual um usuário de informação é capaz de converter a sua necessidade de informação em uma lista de citações a documentos em um acervo contendo informações úteis para ele.
> [...]
> Recuperação de informação abrange os aspectos intelectuais da descrição da informação e sua especificação para a busca, e também quaisquer sistemas, técnicas ou máquinas que são utilizadas para realizar a operação.
> [...]
> O assunto de cada documento ou outra unidade de informação é caracterizado ou descrito por meio de um conjunto de "descritores" tirado de um vocabulário formal de tais termos. Uma “lista de cabeçalho de assuntos” remeterá a uma aproximação grosseira do seu significado. (Mooers, 1951, p. 25).

Para Saracevic (1999), a Recuperação de Informação pode ser considerada a vertente tecnológica da Ciência da Informação e é resultado da relação desta com a Ciência da Computação.

Recuperar informação consiste em identificar num conjunto de documentos aqueles que satisfazem a uma determinada necessidade de informação. Um sistema de recuperação de informação é um elemento mediador nesse processo. Envolve, por um lado, um acervo documental onde cada documento é representado por expressões linguísticas que resumem o seu conteúdo informacional. Por outro lado, os utilizadores desse sistema tentam descrever

linguisticamente as suas necessidades de informação a fim de obterem documentos relevantes que venham a satisfazer tais necessidades. A recuperação de informação se realiza por meio da comparação entre a representação de cada documento do acervo e a representação da necessidade de informação do usuário. Um documento é recuperado se a sua representação coincidir total ou parcialmente com a representação da necessidade do usuário. É possível observar a partir da Figura 1 um esquema do processo de recuperação de informação.

Figura 1 - Processo de Recuperação de Informação



Fonte: Ferneda, 2012, p. 14.

A representação do processo de Recuperação de Informação de Ferneda (2012) é composta, por um lado, pelos documentos e suas representações. Por outro pelo usuário com suas necessidades informacionais traduzidas em uma expressão de busca e sua respectiva representação. Em uma posição central a função de busca, responsável pela comparação entre as representações e apresentação das coincidências entre elas com o resultado da busca.

A seguir serão apresentados resumidamente cada um dos elementos apresentados no diagrama da Figura 1.

2.1 Documentos (*Corpus*)

Os Documentos *(Corpus)* é o elemento principal de todo sistema de recuperação de informação. Le Coadic (2004, p.5) conceitualiza documento, como:

> Documento é o termo genérico que designa os objetos portadores de informação. Um documento é todo artefato que representa ou expressa um objeto, uma ideia ou uma informação por meio de signos gráficos e icônicos

> (palavras, imagens, diagramas, mapas, figuras, símbolos), sonoros e visuais (gravados em suporte de papel ou eletrônicos).

Esta definição corrobora com a ideia de "informação como coisa" apresentada por Buckland (1991). Segundo Buckland, a palavra informação é usada na maioria das vezes com vínculos a um objeto contendo informações: um documento.

2.2 Representação dos documentos

O processo de **representação dos documentos** tem por objetivo identificar e descrever cada documento do *corpus* por meio de seu conteúdo. Essa tarefa é geralmente realizada utilizando procedimentos de indexação. Durante a indexação são extraídos conceitos do documento a partir da análise de seu conteúdo e traduzidos em termos de uma linguagem de indexação, tais como cabeçalhos de assunto, tesauros etc.

Segundo Novellino (1996):

> O processo de representação da informação envolve dois passos principais:
>
> 1. Análise de assunto de um documento e a colocação do resultado desta análise numa expressão linguística;
> 2. Atribuição de conceitos ao documento analisado.
>
> A realização desta fase pressupõe uma linguagem documentária, instrumento de padronização da indexação, a qual visa garantir que indexadores de um mesmo sistema usem os mesmos conceitos para representar documentos semelhantes. Ela também é um instrumento de comunicação ao permitir que indexadores e usuários partilhem um mesmo vocabulário.

A análise de um documento pode envolver uma interpretação de seu conteúdo com a finalidade de agregar assuntos que não estão diretamente explicitados em sua superfície textual, mas que podem ser facilmente inferidos por um indexador humano. A indexação de um documento é efetuada tendo em vista a sua recuperação, com a preocupação de tornar o seu conteúdo visível para os usuários de um sistema de informação.

2.3 Usuário

O **Usuário** interage com o sistema em busca de documentos relevantes. Os utilizadores do sistema de informação devem traduzir as suas necessidades de informação em expressões de pesquisa utilizando a linguagem fornecida pelo sistema. Normalmente, uma expressão de pesquisa consiste em um conjunto de palavras que tentam expressar a semântica da necessidade

de informação de um usuário. A subjetividade do processo de recuperação de informação significa que grande parte da responsabilidade pela sua eficácia passe para o usuário.

2.4 Expressão de Busca (consulta)

Expressão de busca (consulta) é a tradução linguística da necessidade de informação do usuário cujo objetivo é comunicar a necessidade de informação a um sistema de recuperação. Nesse processo comunicativo entre usuário e sistema é fundamental a escolha criteriosa dos termos de busca para se recuperar documentos relevantes e ao mesmo tempo evitar itens não relevantes. Embora importante para uma recuperação eficiente, a especificação da busca é dependente do usuário. Além disso, geralmente as buscas dos usuários são expressas por meio de um número reduzido de termos ou palavras, não permitindo uma interpretação exata e inequívoca da necessidade de informação do usuário (Ferneda, 2003. p. 16).

2.5 Representação da Expressão de Busca

Um dos objetivos de um sistema de recuperação de informação é fornecer ao usuário uma forma fácil de expressar a sua necessidade de informação por meio de uma expressão de busca. Sistemas de recuperação de informação podem oferecer ao usuário recursos que o auxiliem na tarefa de traduzir a sua necessidade de informação em uma expressão de busca.

Porém, independentemente dos recursos oferecidos pelo sistema, é necessário que a expressão de busca seja representada de forma similar à utilizada na representação dos documentos. Essa homogeneidade permitirá a comparação entre a busca e todos os documentos do *corpus* do sistema por meio da função de busca.

2.6 Função de Busca

De forma geral, a **função de busca** compara as representações dos documentos com a representação da expressão de busca e recupera os itens que supostamente fornecerão a informação que o usuário procura.

A função de busca calcula o grau de similaridade entre a expressão de busca e cada um dos documentos do *corpus*. O grau de similaridade pretensamente define o quão relevante é um determinado documento e é utilizado para ordenar os documentos resultantes da busca.

2.7 Resultado da Busca

O **Resultado da Busca** em um sistema de recuperação de informação consiste de um conjunto de documentos que atendam às necessidades de informação do usuário. Esses resultados podem ser apresentados de diversas formas, na grande maioria dos sistemas eles são apresentados em forma de lista, onde os documentos mais similares com a expressão de busca (mais relevantes) são apresentados no topo da lista.

2.8 Modelo de Recuperação de Informação

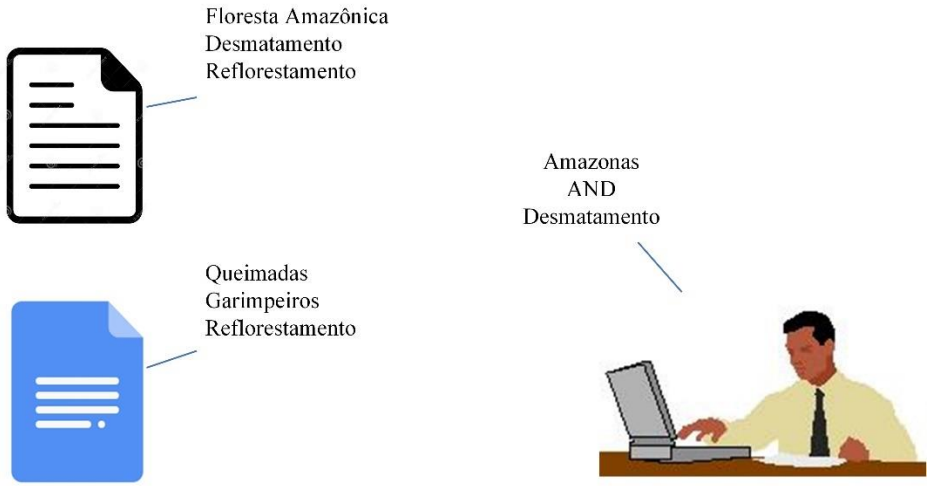
Um modelo de recuperação de informação é a especificação formal de três elementos: a **representação dos documentos**, a representação da necessidade de informação por meio de uma **expressão de busca** e como estes dois elementos serão comparados, a **função de busca**.

A seguir serão apresentados os três modelos ditos "clássicos". Apesar desses modelos terem sido criados nos anos 60 e 70, as suas principais ideias ainda estão presentes na maioria dos sistemas atuais e nos mecanismos de busca da Web.

2.8.1 Modelo Booleano

No Modelo *Booleano* (Ferneda, 2012, cap.3) um documento é representado por um conjunto de termos de indexação e as buscas são formuladas por meio de uma expressão *booleana*, composta por termos ligados através dos operadores lógicos *AND*, *OR* e *NOT*. A Figura 2 apresenta uma ilustração das representações dos documentos e da expressão de busca (consulta) do usuário.

Figura 2 – Modelo *Booleano*: representação dos documentos e da consulta do usuário



Fonte: O Autor (2023)

O resultado de uma busca é um conjunto de documentos cuja representação satisfaz as restrições lógicas da expressão de busca. No modelo *booleano*, após uma busca o conjunto de documentos do *corpus* é particionado naqueles que atendem à expressão de busca nos documentos que não atendem à busca.

Hiemstra (2009) e Baeza-Yates e Ribeiro-Neto (2011, p. 66) destacam algumas vantagens referentes ao modelo *booleano*, tais como o formalismo claro por trás do modelo e sua simplicidade. Em função disso, permite que os usuários tenham um maior controle sobre o sistema, uma vez que possibilita saber exatamente por que um documento foi recuperado para uma determinada consulta.

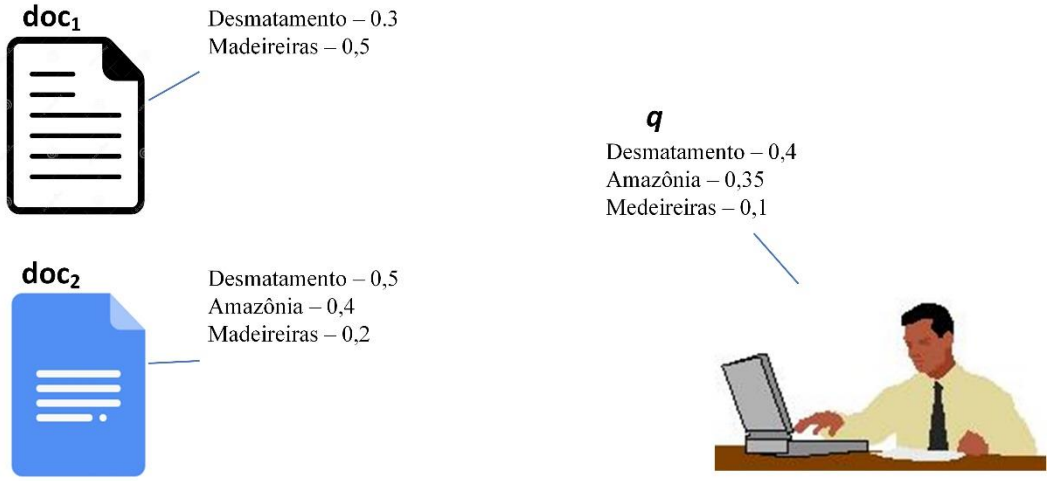
Por outro lado, o modelo possui algumas limitações, sendo que a principal delas é o fato de no modelo não existir um mecanismo que permita a atribuição de relevância relativa a cada documento resultante. Portanto, não é possível realizar um ordenamento (ranqueamento) do conjunto de documentos resultantes de uma busca.

2.8.2 Modelo Vetorial

Segundo Salton e Buckley (1998) no modelo vetorial um documento é representado por um vetor onde cada elemento representa o peso, ou relevância, do respectivo termo de indexação para o documento. Cada elemento do vetor (peso) é geralmente normalizado de forma a assumir valores entre zero e um. Os pesos mais próximos de **1** indicam termos com maior importância para a descrição do documento.

Uma expressão de busca (consulta) também é representada por um vetor numérico onde cada elemento representa a importância (peso) do respectivo termo na representação da necessidade de informação do usuário, substanciada na expressão de busca. A Figura 3 apresenta uma ilustração das representações dos documentos e das consultas.

Figura 3 – Modelo Vetorial: representação dos documentos e da consulta



Fonte: O Autor (2023)

A utilização de uma mesma forma de representação tanto para os documentos como para as expressões de busca permite calcular a similaridade entre uma expressão de busca e cada um dos documentos. Em um espaço vetorial contendo N dimensões, a similaridade (sim) entre um documento $\boldsymbol{d_j}$ e uma expressão de busca $\boldsymbol{q}$ pode ser calculada utilizando a seguinte fórmula como afirma Ferneda (2012):

$$\mathrm{sim}(q, \mathrm{doc_j}) = \frac{\sum_{i=1}^{N}(w_i \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

$\boldsymbol{w}_{i,j}$ é o peso do i-ésimo termo do documento $\mathbf{doc}_j$ e $\boldsymbol{w}_{i,q}$ é o peso do i-ésimo termo da expressão de busca $\boldsymbol{q}$.

Dada uma consulta (q), é calculado o grau de similaridade de cada documento do corpus (doc1, doc2) em relação à consulta. Seguindo os dados da Figura 3, temos:

$$\mathrm{sim}(q, \mathrm{doc_1}) = \frac{(0.2 \times 0.3) + (0.35 \times 0.0) + (0.1 \times 0.5)}{\sqrt{0.2^2 + 0.35^2 + 0.1^2} \times \sqrt{0.3^2 + 0.0^2 + 0.5^2}} \cong 0{,}457$$

$$\mathrm{sim}(q, \mathrm{doc_1}) = \frac{(0.2 \times 0.5) + (0.35 \times 0.4) + (0.1 \times 0.3)}{\sqrt{0.2^2 + 0.35^2 + 0.1^2} \times \sqrt{0.5^2 + 0.4^2 + 0.3^2}} \cong 0{,}92$$

Corroborando ainda com a ideia abordada por Ferneda (2012) em seu livro “Introdução aos Modelos Computacionais de Recuperação de Informação” a expressão de busca ***q*** possui um grau de similaridade de 0,457 (0,45%) com o documento **doc1** e de 0,92 (92%) com o documento **doc2**.

Os valores da similaridade entre uma expressão de busca e cada um dos documentos do *corpus* são utilizados no ordenamento dos documentos resultantes. Portanto, no modelo vetorial o resultado de uma busca é um conjunto de documentos ordenados pelo grau de similaridade entre cada documento e a expressão de busca.

O modelo vetorial destaca-se por ser a base da maioria dos sistemas de recuperação de informação, principalmente os voltados para a Internet, que agregam ainda outras técnicas para determinar o *ranking* dos documentos. No entanto, Ferneda (2003) destaca que uma importante limitação do modelo vetorial está no fato de não permitir a formulação de consultas *booleanas*, o que pode restringir consideravelmente a sua flexibilidade.

2.8.3 Modelo Probabilístico

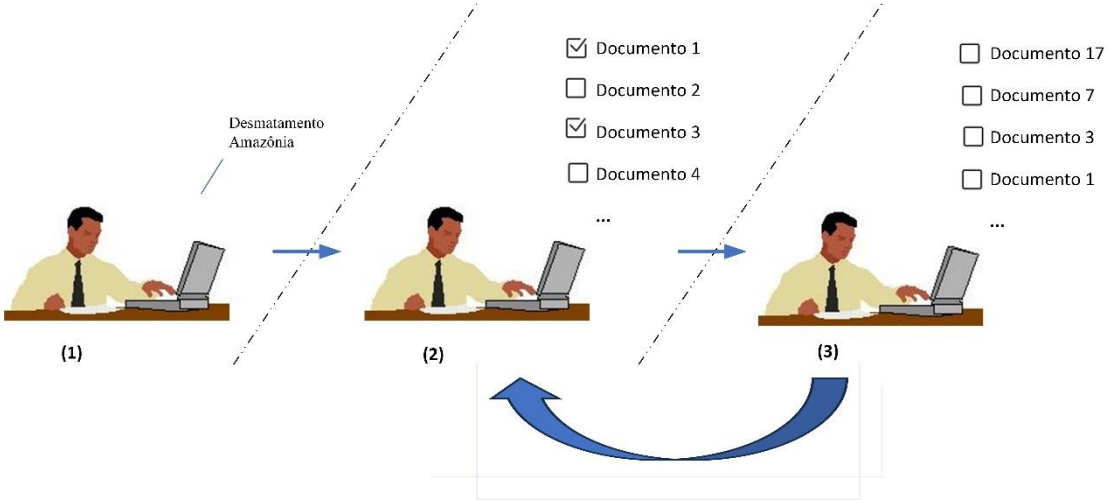
O Modelo Probabilístico foi proposto inicialmente por Maron e Kuhns (1960) e posteriormente explorado por diversos outros pesquisadores, tais como Robertson e Jones (1976). A ideia é tratar o processo de recuperação de informação como um processo probabilístico, já que ele é caracterizado por seu grau de incerteza. Assim, é mais realista pensar em uma probabilidade de relevância do que em uma pretensa relevância exata, como a utilizada nos modelos *booleano* e vetorial.

Os documentos e a necessidade de informação do usuário são representados por um conjunto de termos. Por meio de cálculos de probabilidade o sistema calcula para cada documento um valor numérico que representa a provável relevância do documento para a necessidade do usuário. Esse valor é utilizado para ordenar os resultados da busca.

Esse modelo permite a interação do usuário, permitindo julgar a relevância dos resultados que lhe foram apresentados. Por meio de interações sucessivas do usuário busca-se alcançar, gradativamente, resultados mais relevantes para a necessidade de informação do usuário. Esse processo de interação é denominado *Relevance Feedback*, traduzido muitas vezes por "retroalimentação de relevância", ou ainda "realimentação de relevância".

A Figura 4 ilustra do processo de *Relevance Feedback* do modelo probabilístico.

Figura 4 – Modelo Probabilístico: o processo de *Relevance Feedback*



Fonte: O Autor (2023)

Seguindo a ilustração da Figura 4, o usuário representa a sua necessidade de informação por meio da especificação de um conjunto de termos de busca e o envia para o sistema **(1)**; O sistema retorna uma primeira lista de documentos pertinentes à consulta **(2)**. Tendo um primeiro conjunto de documentos resultantes, o usuário pode selecionar (marcar) alguns deles que considera relevantes **(2)**. O conjunto de documentos marcados pode ser então submetido novamente ao sistema. Com esse *feedback* do usuário, o sistema pode então recalcular a relevância de cada documento e apresentar um novo conjunto de documentos pretensamente mais relevantes ou pertinentes em relação aos resultados anteriores **(3)**. Esse processo pode ser repetido até que o usuário se sinta satisfeito com os resultados obtidos.

Uma virtude do modelo probabilístico está em reconhecer que a atribuição de relevância é uma tarefa subjetiva, sendo que o usuário é o elemento mais competente para realizá-la. O Modelo Probabilístico é o único modelo que incorpora explicitamente o processo de *relevance feedback* como base para a sua operacionalização.

3
Relevance Feedback

Para Manning, Raghavan e Schütze (2008), o processo de recuperação de informação é inerentemente incerto. Os usuários podem não ter uma ideia bem elaborada da informação que necessitam, e não serem capazes de traduzir em palavras tal necessidade.

A ideia de *Relevance Feedback* [2] é envolver o usuário no processo de recuperação de informação de forma a melhorar o conjunto de resultados finais alcançados. Em particular, o usuário dá seu *feedback* sobre a relevância dos documentos num conjunto inicial de resultados. Após uma primeira consulta, e a obtenção de um primeiro conjunto resultante de documentos, o usuário marca os documentos que considera relevantes para suas necessidades e submete essas informações ao sistema. O sistema pode então usar essas informações para recuperar mais documentos semelhantes aos documentos considerados relevantes. Resumindo o procedimento básico:

- O usuário emite uma consulta (curta e simples).
- O sistema retorna como resultado um conjunto inicial de documentos.
- O usuário marca alguns desses documentos como relevantes (ou não relevantes) e submete novamente ao sistema.
- Com base nos documentos considerados relevantes, o sistema seleciona um novo conjunto de documentos que possivelmente melhor represente a necessidade de informação do usuário.

[2] Na literatura pesquisada, a o termo *Relevance Feedback* é traduzido formas variadas: "retroalimentação de relevância", "realimentação de relevância", ou ainda "feedback de relevância". Nesse trabalho será utilizado o termo original em inglês.

- O sistema exibe um conjunto revisado de resultados.

O processo de *Relevance Feedback* pode passar por uma ou mais iterações desse tipo. Esse processo explora a ideia de que pode ser difícil formular uma boa consulta quando não se conhece bem a forma como estão representados os documentos do *corpus*; todavia é fácil julgar a relevância de documentos específicos.

Segundo Manning, Raghavan; Schütze (2008), a recuperação de imagens fornece um bom exemplo da importância do processo de *Relevance Feedback*. Este é um domínio onde um usuário geralmente tem dificuldade em traduzir em palavras a imagem que deseja, mas tendo a visualização de um conjunto de imagens pode facilmente indicar quais delas são relevantes ou não.

3.1 Tipos de *Relevance Feedback*

Relevance Feedback pode ser classificado em dois tipos principais: explícito e implícito. O ***feedback* explícito** exige que o usuário indique explicitamente quais documentos são relevantes ou não relevantes para sua consulta. O ***feedback* implícito** infere as preferências do usuário a partir de seu comportamento, como clicar, visualizar ou qualquer ação realizada durante o uso do sistema de recuperação de informação.

3.1.1 *Feedback* explícito

Envolve o usuário no fornecimento do *feedback* direto sobre a relevância dos documentos resultantes de uma busca. Isso pode ser feito através de avaliações, como "relevante" ou "não relevante", dadas pelo usuário a alguns documentos retornados pela busca. Com base nesse *feedback*, o sistema pode ajustar e melhorar os futuros resultados, aumentando a precisão e a relevância das respostas fornecidas.

É, portanto, uma forma de refinamento do sistema de recuperação, utilizando a contribuição direta dos usuários para aprimorar a qualidade dos futuros resultados de busca. Uma das vantagens desta estratégia é a simplicidade e alta confiança nas informações dos *feedbacks* recebidos.

3.1.2 *Feedback* Implícito

Esse tipo de *feedback* é inferido a partir do comportamento do usuário durante a interação com os resultados da busca. A análise de cliques, rastreamento dos olhos (*eye tracking*), tempo gasto em uma página, rolagem das páginas, entre outros, pode ser usado para inferir a relevância do conteúdo para o usuário.

Quando alguém faz uma busca e interage com os resultados (clicando, lendo, ignorando), o sistema pode usar esses padrões de interação para entender a relevância desses resultados em relação à consulta original. Isso pode ser feito automaticamente, sem exigir que o usuário forneça qualquer tipo de informação ou mesmo saiba que esses dados estão sendo coletados pelo sistema.

Com base nessas interações, o sistema pode ajustar os resultados futuros para aquela consulta específica ou consultas semelhantes, melhorando assim a relevância das respostas apresentadas aos usuários.

A principal vantagem do uso de *feedback* implícito é o fato de não ser necessário uma interação explícita dos usuários. Ou seja, os usuários devem utilizar o sistema normalmente e o sistema se responsabiliza de inferir a relevância dos documentos a partir do comportamento dos usuários. A principal deficiência desta estratégia é o baixo nível de confiança sobre as informações obtidas desses *feedbacks*.

3.1.3 Pseudo *Feedback*

Esse método, também chamado de *blind feedback*, consiste em considerar como relevantes os primeiros documentos da lista de resultados da consulta. Com base nesses documentos, um novo conjunto de termos de busca é criado para refinar a busca, tentando melhorar os resultados.

A principal vantagem dessa abordagem é ser completamente automática, não dependendo de nenhuma interação anterior dos usuários. Uma desvantagem é a dependência do método na qualidade da expressão de busca (consulta) formulada pelo usuário. Caso a consulta possua, por exemplo, termos ambíguos, o resultado será uma piora na qualidade dos resultados do sistema (Manning; Raghavan; Schütze, 2008).

4
Expansão de Consultas

Um sistema de recuperação de informação é um elemento mediador entre os usuários e um determinado acervo documental. O usuário interage com o sistema a fim de comunicar a sua necessidade de informação e obter documentos que possam satisfazer tal necessidade. Na maioria dos sistemas essa comunicação é feita por meio da especificação de termos que representam a necessidade do usuário. Nesse processo comunicativo entre usuário e sistema é fundamental a escolha criteriosa dos termos de busca para se recuperar documentos relevantes e ao mesmo tempo evitar itens não relevantes. Porém, sem um conhecimento de como foram representados (indexados) os documentos do *corpus*, é difícil ao usuário predizer os termos que resultem em um conjunto de documentos que efetivamente atenderão à sua necessidade.

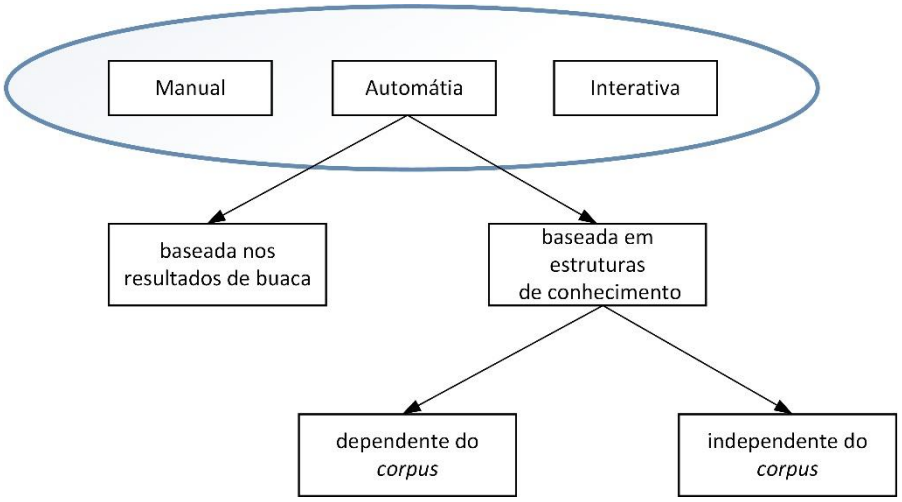
Embora importante para uma recuperação eficiente, a especificação da busca (consulta) é dependente do usuário, com toda a variabilidade inerente ao ser-humano. Além disso, geralmente as buscas dos usuários são expressas por meio de um número reduzido de termos ou palavras, não permitindo uma interpretação exata e inequívoca da necessidade de informação do usuário.

A importância e as dificuldades do processo de especificação de buscas fizeram surgir na área de Recuperação de Informação um nicho de pesquisa denominado "Expansão de Consulta" (*query expansion*). Expansão de consulta é o termo utilizado para referenciar os métodos e processos que visam melhorar a eficiência da recuperação de informação baseados no pressuposto de que as consultas definidas pelos usuários muitas vezes não refletem suas reais necessidades de informação. O objetivo principal é adicionar novos termos à consulta inicialmente formulada pelo usuário a fim de melhorar os resultados obtidos.

Muitas vezes os mecanismos de expansão de consultas podem ser aplicados para auxiliar o usuário na formulação da sua consulta inicial e adicionalmente ou alternativamente fazer uso de tais técnicas em etapas subsequentes, reformulando as consultas até que sejam satisfeitas suas necessidades de informação.

Efthimiadis (1996) distingue três modos diferentes de expansão de consulta, como representado na Figura 5. Uma reformulação é considerada **manual** (intelectual) sempre que o próprio usuário altera a sua consulta inicial por meio da adição de novos termos. A expansão é considerada **automática** quando o sistema gera os termos de expansão e adiciona à consulta original. Segundo Efthimiadis, para ser considerado automático o processo de expansão de consulta não pode ser influenciado pelo usuário nem tampouco ele pode estar ciente de sua existência. No modo **interativo** o usuário seleciona os termos que serão utilizados na expansão da consulta a partir de uma lista apresentada pelo sistema.

Figura 5 – Métodos de expansão de consulta



Fonte: Adaptado de Efthimiadis (1996)

Métodos de expansão de consulta podem variar ainda na forma como são gerados os termos da expansão. Como mostrado na Figura 5, estes termos podem ser originados dos **resultados de busca** ou de **estruturas de conhecimento**. Os métodos baseados nos resultados da busca selecionam os termos de expansão a partir dos documentos resultantes da consulta inicial. Nesse caso, a eficácia da expansão da consulta depende fortemente da qualidade da consulta original. Essa dependência não existe nos modelos de expansão **baseados em estruturas de conhecimento**.

As estruturas de conhecimento podem ser **dependentes do *corpus*** ou **independentes do *corpus***. Mecanismos **dependentes do *corpus*** analisam os documentos do acervo

documental a fim de selecionar os termos que serão utilizados para a expansão da consulta. Métodos **independentes do *corpus*** contam com estruturas de conhecimento que não apresentam relação com os documentos. São exemplos dessas estruturas: léxicos, glossários, dicionários, tesauros e ontologias.

4.1 Expansão de consultas baseada nos resultados da busca

A expansão de consultas baseadas nos resultados da busca está relacionada ao processo de *Relevance Feedback*. Este processo parte da ideia de que embora seja difícil formular uma primeira consulta eficiente, é fácil julgar a relevância dos documentos recuperados.

Embora os mecanismos de *Relevance Feedback* provem ser eficazes para melhorar resultados da recuperação, estes sofrem de diversos inconvenientes. Eles são eficazes somente se os documentos na coleção contiverem uma determinada quantidade de texto dos quais os termos da expansão possam ser obtidos. Este não é o caso em bases de dados de referências, por exemplo, onde os metadados dos documentos geralmente compreendem somente uma quantidade pequena de texto livre. Além disso, *Relevance Feedback* é somente aplicável se a consulta original do usuário resultar em um conjunto com um número razoável de documentos. Além disso, métodos de *Relevance Feedback* não podem ser aplicados na formulação da consulta inicial.

Os mecanismos de *Relevance Feedback* são dependentes da voluntariedade dos usuários em fornecer o seu parecer sobre a relevância dos documentos recuperados. Segundo Spink *et al.* (2000), na maioria das vezes os usuários são relutantes em fazer isso.

4.2 Expansão de consultas baseada em estruturas de conhecimento dependentes do *corpus*

Em vez de se obter termos de expansão a partir dos documentos resultantes de uma busca, os métodos baseados em estruturas de conhecimento **dependentes** do *corpus* utilizam a estrutura de todos os documentos para identificar tais termos. Para este propósito, dependências estatísticas entre termos são calculadas por meio da aplicação de cálculos de co-ocorrência. Uma forma simples de utilizar dados de co-ocorrência é identificar nos documentos termos de indexação que se assemelham aos termos de uma determinada consulta com o objetivo de utilizá-los como termos de expansão.

Comparado com os mecanismos de *Relevance Feedback*, métodos de expansão baseados em uma estrutura terminológica proveniente dos documentos do *corpus* possuem a vantagem de poderem ser aplicados na formulação da consulta inicial.

4.3 Expansão de consultas baseada em estruturas de conhecimento independentes do *corpus*

Estruturas de conhecimento constituem fontes especialmente promissoras para a geração dos termos de expansão nos casos em que mecanismos de *Relevance Feedback* e de co-ocorrência não são aplicáveis. Bhogal *et al.* (2007) salientam que as estruturas de conhecimento **independentes** do *corpus* são especialmente úteis se o número de documentos for pequeno ou se os seus documentos contiverem pouco texto livre. Neste caso, os mecanismos de *Relevance Feedback* e os mecanismos dependentes do *corpus* provavelmente não serão muito eficazes. A aplicabilidade dos métodos de expansão de consulta baseados em estruturas de conhecimento, por outro lado, independe do tamanho do *corpus*.

Outra vantagem do uso de estruturas de conhecimento para a expansão da consulta é sua disponibilidade a qualquer momento no processo de busca. Ao contrário dos mecanismos baseados em *Relevance Feedback*, a consulta inicial pode já se beneficiar desse tipo de expansão, pois os termos não são derivados de resultados da busca inicial. No entanto, o desenvolvimento de estruturas de conhecimento adequadas para fins de expansão de consulta pode ser um processo de alto custo. Como afirmado por Harman (1988) e Greenberg (2001), o desenvolvimento de mecanismos de expansão de consulta independentes do *corpus* muitas vezes é dificultada pela disponibilidade limitada de tesauros, ontologias ou qualquer outro tipo de estrutura terminológica.

5
Recuperação Baseada em Diálogo

A recuperação de informação (RI) pode ser vista como um processo de comunicação. Segundo Meadow *et al.* (2007), RI é um meio pelo qual autores e criadores de registros se comunicam com os leitores. Para Vieira (1994), a recuperação de informação é um processo em que emissor e receptor interagem para atender a uma necessidade de informação. "Essa interação só é viável por meio do uso da linguagem".

Segundo Crestani e Pasi (2003, tradução nossa):

> A subjetividade é uma propriedade intrínseca de qualquer sistema de RI. Está relacionado ao próprio conceito de relevância. É um fato bem conhecido que o mesmo documento pode ser totalmente relevante para um usuário e totalmente irrelevante para outro usuário, embora ambos façam a mesma consulta ao mesmo sistema de RI. Somente o usuário é o juiz final da relevância de um documento para uma necessidade de informação.

Mooers (1950), enfatiza o caráter linguístico (não numérico) do processo de recuperação de informação:

> [...] A recuperação da informação é um problema não numérico em parte porque a maior parte da comunicação humana é verbal, mas mais importante porque a maioria das ideias ou conceitos não podem ser mapeados em um espaço Euclidiano de 3 ou mais dimensões. Embora haja valores de escala para a representação de algumas informações, estas são relativamente poucas e sem importância. Conceitos espaciais e métricos não se aplicam à maioria das informações, pelo menos não nos níveis mais simples. No entanto, embora o problema de recuperação de informação não seja numérico, não parece haver alternativa ao uso de técnicas digitais para sua solução. Sistemas de recuperação de informação digital que empregam máquinas já estão operando,

> e seu grau de sucesso parece indicar que esta é a direção do progresso. (Mooers, 1950, p. 3, tradução nossa).

O processo de recuperação de informação possui os mesmos problemas inerentes ao processo comunicacional: a subjetividade, a imprecisão e a ambiguidade. Nesse contexto comunicativo, o diálogo é um meio natural de obtenção de informações e a forma cotidiana de dirimir ambiguidades e imprecisões inerentes à linguagem humana. Assim, o diálogo como interface de busca em um processo de recuperação de informações pode ser pensado como um recurso natural de se obter informações.

O desenvolvimento tecnológico dos últimos anos viabilizou a criação de novas interfaces e novas formas de interação entre pessoas e computadores. Atualmente, uma série de novos produtos e serviços passaram a ser disponibilizados por grandes empresas para acesso à informação. Por vez essas tecnologias são descritas como ferramentas de computação cognitiva ou simplesmente produtos da Inteligência Artificial.

Nesse cenário, diversas pesquisas vêm sendo desenvolvidas na tentativa de incorporar aos sistemas de recuperação de informação novas formas de interação. Dentre essas pesquisas, estão propostas da utilização do conceito de "assistente virtual" ou *chatbots* na construção de interfaces de busca e recuperação de informação.

Os sistemas de diálogo baseados em texto ou voz, também chamados de agentes de conversação, *chatbots* ou *chatterbots*, estão se tornando populares, não só por parte das grandes empresas, mas também pelos usuários de Internet e de dispositivos móveis. Uma das razões dessa popularidade é a sua capacidade de interagir de forma inteligente, que vem melhorando muito nos últimos tempos devido aos avanços significativos nas tecnologias de *hardware* e inteligência artificial (Bartl; Spanakis, 2017).

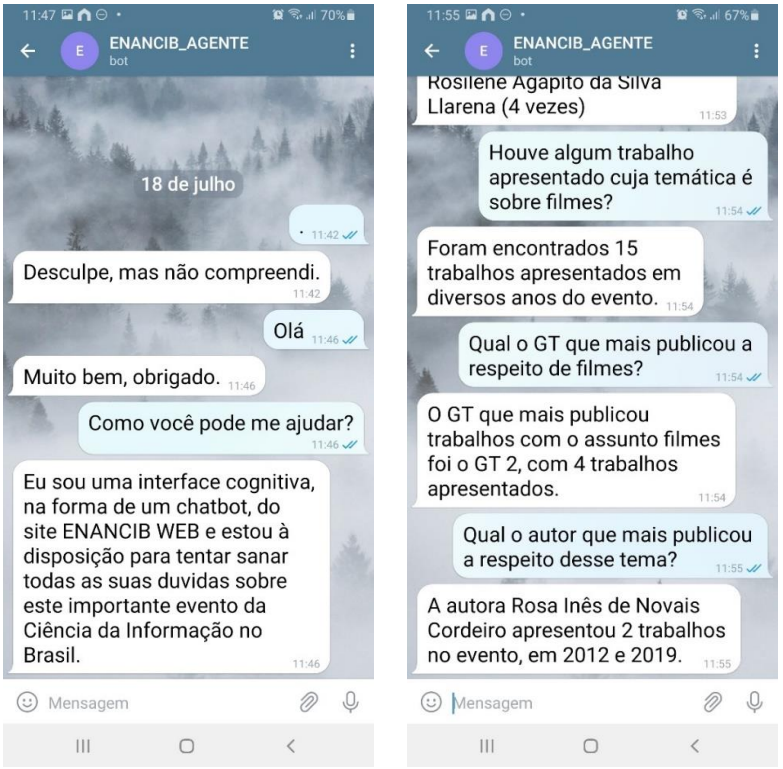
O termo *chatbot* é formado pela junção das palavras da língua inglesa: *chat* (bate-papo) e *bot*, abreviação da palavra *robot* (robô). É um programa de computador que possui a capacidade de imitar uma conversa inteligente entre um humano e um computador, permitindo ainda, examinar e até mesmo influenciar o comportamento do usuário desse sistema. Por meio da construção de perguntas e respostas, ele funciona como um mediador/interface na comunicação de um sistema computacional com os seus usuários na busca de sanar as necessidades de informações destes. O sistema tem a capacidade de formular perguntas de maneira inteligente para o usuário por meio de linguagem natural. Essa comunicação pode ser

de forma textual ou falada. O sistema é capaz de processar os termos do diálogo e construir respostas e perguntas (Abdul-Kader; Woods, 2015).

Os *chatbots* desenvolvidos por grandes empresas como *Amazon*, *Apple*, *Microsoft*, *Facebook*, utilizam os últimos avanços em sistemas de aprendizagem de máquina para aplicá-los em sistemas de recuperação de informações. Alguns modelos utilizam técnicas de tradução automática de estatísticas para tentar traduzir as frases de entrada em respostas de saída ou mesmo a utilização de redes neurais que codificam e decodificam entradas em respostas (Cahn, 2017).

Carvalho (2022) apresenta um modelo conceitual de recuperação de informação por meio de um *chatbot* ("agente conversacional") no qual a construção dos diálogos é enriquecida por termos provindos dos metadados da base de dados de um repositório. O protótipo desenvolvido nessa tese foi testado utilizando um repositório formado por dados oriundos de trabalhos completos e resumos expandidos de todas as edições do Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (ENANCIB). A Figura 6 apresenta a imagem de algumas telas do aplicativo/protótipo ENANCIB *Web*.

Figura 6 – Imagens de telas do aplicativo ENANCIB *Web*



Fonte: Carvalho (2022, p. 133 e 136)

Nguyen *et al*. (2022) propõem um modelo, denominado Inter-Rela, para integração de múltiplos domínios de conhecimento e o gerenciamento da base de conhecimento de um *chatbot* "educacional inteligente". Este *chatbot* pode atuar como um tutor, fornecendo informações e conhecimento aos alunos de uma universidade.

Höppner e Fredriksson (2019) apresentam um *chatbot* que recupera informações de uma fonte de dados constituída por texto não estruturado em linguagem natural. Segundo os autores, o *chatbot* responde corretamente às perguntas com uma precisão de 72% de acordo com testes com 25 documentos.

Meyer von Wolff *et al.* (2020) conduziram pesquisa baseada em questionário entre 166 estudantes de diversas disciplinas e níveis educacionais de uma universidade alemã. O objetivo era fazer um levantamento dos requisitos para implementar um *chatbot*, bem como tópicos relevantes e questões que os *chatbots* deveriam abordar. Como resultado, essa pesquisa indicou que os *chatbots* são adequados para o contexto universitário e que muitos estudantes estão dispostos a utiliza-los.

6
Análise e Resultados

Dentre os artigos selecionados onde continham um conteúdo compatível com o tema proposto nesse trabalho, foi feito pelo autor uma prévia leitura dos resumos. Hoje em dia, se encontram muitos trabalhos com essa temática em língua inglesa, mas são poucos os disponíveis na língua portuguesa.

Dessa forma, os principais periódicos que se enquadraram com essa abordagem, tendo em vista a subjetividade e o processo de recuperação da informação, foram utilizados para realização do presente trabalho. Com base nos critérios de seleção estabelecidos, foram selecionados alguns trabalhos, sendo eles os encontrados na BRAPCI e Repositório Institucional da UNESP, pois poucos trabalhos efetivamente abordam sobre o tema.

Os resultados demonstrados aqui, a partir da criação de um quadro para sintetizar a coleta dos dados obtidos, da revisão bibliográfica, organizado de forma cronológica, *relevance feedback*, expansão de consulta e recuperação baseada em diálogo, como foi abordada cada técnica nesse trabalho, assim elaborando um banco de dados sobre as respectivas técnicas para o tratamento da subjetividade no processo de recuperação da informação no que tange a arquivologia, pois acredita-se que desta forma auxilie no entendimento das mesmas.

No quadro 1 a seguir, apresenta-se as técnicas abordadas e as possíveis utilizações na arquivologia.

Quadro 1 – Tratamento da Subjetividade na Arquivologia

Técnicas de tratamento da subjetividade	Possíveis utilizações em sistemas de arquivo
Relevance Feedback	Em um sistema de recuperação de informação de um acervo/arquivo, as técnicas de *relevance feedback* podem ser de grande efetividade, pois quando um usuário elabora uma expressão de busca, mas acaba não se expressando de forma adequada para encontrar um determinado documento ou informação que ele deseja, essa técnica pode ser útil, porque o usuário pode selecionar dentre os resultados da busca da sua consulta inicial quais arquivos são relevantes e submeter isso ao sistema novamente, assim, o sistema lhe dará melhores resultados dos arquivos que sejam úteis para a necessidade de informação do usuário. Portanto, com essa interação entre o usuário de um arquivo e o sistema de recuperação da informação é fácil o usuário ter uma informação/documento relevante, a partir do feedback que o usuário passa para o sistema.
Expansão de Consulta	Com a expansão de consulta em um sistema de recuperação de informação de um arquivo, os usuários podem ter a ajuda do sistema para formular sua expressão de busca. De uma forma interativa, através de uma lista que o sistema oferece, o usuário pode escolher quais termos ou palavras podem ser adicionadas a sua expressão de busca; ou automaticamente, o sistema pode adicionar novos termos a expressão de busca inicialmente formulada pelo usuário, para obter arquivos, documentos e informações que seja relevante para ele, assim, fazendo uma expansão de consulta.
Recuperação baseada em Diálogo	Com o desenvolvimento da tecnologia, a recuperação da informação baseada em diálogo entre humanos e computadores é real. A inteligência artificial pode ajudar no processo para recuperar arquivos/documentos através de *chatbots*, o assistente virtual do sistema pode fazer buscas em arquivos/documentos que interesse ao usuário durante o atendimento, onde o usuários do sistema de arquivo pode interagir de forma inteligente e o sistema pode formular perguntas e respostas para os seus usuários afim de sanar suas necessidades informacionais e, esse assistente pode fornecer informações úteis de forma rápida para seus usuários.

Fonte: Elaborado pelo autor, 2023.

Assim sendo, como todos os usuários de um arquivo estão sujeitos à subjetividade no processo de encontrar documentos e informações, essas técnicas apresentadas podem ser incorporadas em um sistema de recuperação de informação de um arquivo, pois os documentos, em uma instituição arquivística, precisam ser recuperados e o acesso é uma "função arquivística destinada a tornar acessíveis os documentos e a promover sua utilização" (ARQUIVO NACIONAL, 2005, p.19).

Como mencionado por Smit (2013):

> A introdução dos recursos tecnológicos trouxe novos desafios para os arquivos, além de uma maravilhosa possibilidade de atingir um número muito maior de usuários, remotos ou virtuais, quando as referências aos documentos, ou, quando os recursos tecnológicos o permitem, os próprios documentos são disponibilizados pela rede.

Portanto, os recursos tecnológicos disponíveis hoje em dia podem facilitar e muito o acesso a documentos e informações que interessem a usuários de arquivos, independente do objetivo da pesquisa.

Os autores Ramos e Munhoz (2011) destacam que para um sistema “ideal” deve ocorrer recuperação por partes de palavras e por sinônimos... É necessário um estudo avançado do perfil de cada usuário. Sendo assim, elaborar indexações mais direcionadas aos usuários, desse modo aumentando a recuperação de informação e melhorando os documentos a serem recuperados.

7
Considerações Finais

A recuperação da informação envolve um conjunto de documentos representado geralmente por expressões linguísticas (resumo e/ou índice), pessoas que procuram descrever textualmente as suas necessidades de informação e um elemento de comparação entre as representações dos documentos e a representação da busca (consulta).

A representação de um documento pode ser produzida automaticamente por meio de métodos computacionais. No entanto, a necessidade de informação de um usuário só passível de tratamento após a sua enunciação, obtida durante o processo de busca. Sua correta interpretação só é possível por meio de técnicas que permitam uma certa interação do usuário com o sistema de recuperação. Tais técnicas tentam contornar a subjetividade dando ênfase na importância da representação da necessidade do usuário. Tenta-se assim contornar ou minimizar os efeitos da subjetividade, oferecendo ao usuário recursos para melhor descrever os seus interesses, objetivos e desejos informacionais descritos e representados na sua expressão de busca.

Historicamente, um primeiro esforço para enfatizar a importância do usuário no processo de recuperação de informação está evidenciado no Modelo Probabilístico, pois reconhece que o elemento mais competente para atribuir relevância aos documentos resultantes de uma busca é o próprio usuário que a fez. O Modelo Probabilístico incorpora, como base de seu funcionamento, o processo denominado de *Relevance Feedback*.

Relevance Feedback é uma técnica que busca envolver o usuário no processo de recuperação de informação, permitindo a sua participação contínua no julgamento de relevância dos documentos recuperados após a execução de uma busca. *Relevance Feedback* nos remete

ao conceito de "seção de busca", na qual o usuário está envolvido em um processo cíclico, onde após a execução de cada busca ele é acionado para julgar a relevância dos documentos recuperados, até que esteja satisfeito com os resultados apresentados pelo sistema.

A enunciação de uma expressão de busca (consulta) é dependente do usuário, com toda a variabilidade inerente ao ser-humano. As dificuldades na especificação de consultas que reflitam a real necessidade de informação do usuário fizeram surgir uma sub-área de pesquisa denominada "Expansão de Consulta" (*query expansion*). Tem por objetivo a criação de técnicas que visam melhorar a eficiência de um sistema de recuperação de informação auxiliando o usuário na elaboração de suas consultas.

Como ressaltado anteriormente, o processo de recuperação de informação pode ser visto como um processo de comunicação. Visto por esse ponto de vista, a interação e o diálogo é um meio natural de resolver ambiguidades e mal entendidos, frutos das linguagens humanas. Sendo assim, pode-se pensar na incorporação do diálogo como um recurso nos sistemas de recuperação de informação. Nesse contexto, algumas pesquisas propõem a utilização do conceito de "assistente virtual" ou *chatbots* na construção de interfaces de busca e recuperação de informação.

Os objetivos do estudo foram alcançados, visto que fazendo o levantamento bibliográfico para levantar e apresentar as técnicas desenvolvidas para automatizar tarefas subjetivas que compõem o processo de recuperação da informação, foi validado satisfatoriamente.

Com isso, foi possível constatar que as técnicas de *relevance feedback*, expansão de consulta, recuperação baseada em diálogos e seus potenciais modelos utilizados para minimizar os problemas causados pela subjetividade no processo de recuperação da informação, auxilia e facilita na busca por informações e documentos úteis.

Recentemente, com a popularização da Inteligência Artificial (IA), novas perspectivas se abriram para o desenvolvimento de interfaces interativas para a Recuperação de Informação. Algumas técnicas e tecnologias provenientes a IA podem agora ser pensadas e se tornarem uma realidade. Estre essas tecnologias ou métodos estão:

- **Processamento de Linguagem Natural (PLN):** Ajuda na compreensão de consultas escritas em linguagem humana, permitindo que os sistemas de recuperação entendam melhor a necessidade do usuário;

- **Aprendizado de Máquina:** Utilizado para aprimorar os algoritmos de busca, permitindo que o sistema aprenda com interações passadas e melhore continuamente a precisão dos resultados;
- **Redes Neurais:** Modelos de redes neurais são empregados para identificar padrões complexos nos dados, o que pode melhorar a relevância dos resultados de busca;
- **Sistemas de Recomendação:** A IA pode analisar o comportamento do usuário e fornecer sugestões personalizadas, ajudando na recuperação de informações relevantes.

A subjetividade é uma característica inerente ao processo de recuperação de informação, pois envolve um ser humano na enunciação de sua necessidade de informação e no julgamento da relevância dos documentos resultantes de uma busca. Historicamente, as pesquisas na área de Recuperação de Informação vêm tentando minimizar essa característica, e atualmente, com os avanços das tecnologias, estão sendo apresentadas novas maneiras de se obter as informações que uma pessoa necessita por meio de novos recursos de interação.

Referências

ABDUL-KADER, Sameera A.; WOODS, J. C. Survey on chatbot design techniques in speech conversation systems. **International Journal of Advanced Computer Science and Applications**, v. 6, n. 7, 2015.

ARQUIVO NACIONAL. **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro: Arquivo Nacional, 2005.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval**. 2ª ed. Addison-Wesley, 2011.

BHOGAL, J.; MACFARLANE, A.; SMITH, P. A review of ontology-based query expansion. **Information Processing and Management**, v. 43, n. 4, p. 866-886, 2007.

BARTL, Alexander; SPANAKIS, Gerasimos. A retrieval-based dialogue system utilizing utterance and context embeddings. In: **2017 16th IEEE International Conference on Machine Learning and Applications** (ICMLA). IEEE, 2017. p. 1120-1125.

BUCKLAND, Michael Keeble. **Information and Information Systems**. New York: Greenwood, 1991.

CAHN, Jack. CHATBOT: Architecture, design, & development. University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, 2017.

CARVALHO, Ricardo César de. **Chatbot aplicado à recuperação de informação**: um modelo orientado a metadados. Universidade Estadual Paulista (Unesp), 2022.

CRESTANI, Fabio; PASI, Gabriella. Handling Vagueness, Subjectivity, and Imprecision in Information Access: An Introduction to the Special Issue. **Information Processing & Management**, v. 39, n. 2, p. 161-165, 2003.

EFTHIMIADIS, E. N. Query expansion. *In*: WILLIAMS, M.E. **Annual Review of Information Science and Technology-ARIST**. Medford, N.J.: Information Today, 1996.

FERNEDA, Edberto. **Recuperação de informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. 2003. Tese (Doutorado em Ciência da Informação e Documentação) — Universidade de São Paulo, São Paulo, 2003. Disponível em: https://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf. Acesso em 10/11/2023.

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

GREENBERG, J. Automatic query expansion via lexical-semantic relationships. **Journal of the American Society for Information Science and Technology**, v.52, n.5, 2001.

HARMAN, D. (1988): Towards interactive query expansion. In: **Proceedings 11th annual international ACM Conference on Research and Development in Information Retrieval**, Grenoble, France, 1988.

HIEMSTRA, D. Information retrieval models. In: GOKER, A.; DAVIES, J. **Information retrieval**: searching in the 21st Century. Wiley, 2009. p. 1–17.

HÖPPNER , Falk; FREDRIKSSON, Joakim. **Chatbot for Information Retrieval from Unstructured Natural Language Documents**. Bachelor Thesis for the Computer Science and Engineering Programme. Halmstad University, 2019. Disponivel em: http://www.diva-portal.org/smash/get/diva2:1349769/FULLTEXT02.pdf. Acesso em 20/11/2023.

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

SMIT, J. W. **Recuperação, acesso e uso dos documentos arquivísticos**. , . DOI: 10.18225/ci.inf..v42i1.1391 Acesso em: 20 dez. 2023.

LE COADIC, Yves-François. **A Ciência da Informação**. 2.ed. Brasília: Briquet de Lemos, 2004.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge University Press, 2008.

MARON, Melvin Earl; KUHNS, John Larry. On relevance, probabilistic indexing and information retrieval. **Journal of the ACM (JACM)**, v. 7, n. 3, p. 216-244, 1960.

MEADOW, Charles T. *et al*. **Text Information Retrieval System**. 3rded. London UK: Elsevier, 2007.

MEYER VON WOLFF, R., NÖRTEMANN, J., HOBERT, S., SCHUMANN, M. Chatbots for the Information Acquisition at Universities – A Student's View on the Application Area. In: Følstad, A., *et al*. **Chatbot Research and Design**. CONVERSATIONS 2019. Lecture Notes in Computer Science, vol 11970. Springer, Cham., 2020. https://doi.org/10.1007/978-3-030-39540-7_16

MOOERS, Calvin N. **The theory of digital handling of non-numerical information and its implications to machine economics**. Zator Company, 1950.

MOOERS, Calvin N. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, v. 2, n. 1, p. 20-32, 1951. Wiley-Blackwell. Disponível em: http://dx.doi.org/10.1002/asi.5090020107. Acesso em 10/11/2023.

NGUYEN, Hien D.; NGUYEN, Tuan-Vi; PHAM, Xuan-Thien; HUYNH, Anh T.; PHAM, Vuong T., NGUYEN, Diem. **Design Intelligent Educational Chatbot for Information Retrieval based on Integrated Knowledge Bases**. IAENG International Journal of Computer Science, v. 49, n. 2, June 2022. Disponível em: https://www.iaeng.org/IJCS/issues_v49/issue_2/IJCS_49_2_28.pdf.

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

PRODANOV, C. C.; FREITAS, E. C. Metodologia do Trabalho Científico: métodos e técnica da pesquisa e do trabalho acadêmico. 2. ed. – Novo Hamburgo: Feevale, 2013.

RAMOS, C. R.; MUNHOZ, D. P. **A subjetividade da relevância na recuperação da informação: análise a partir de imagens representativas.** , p. 69-79, . Disponível em: http://hdl.handle.net/20.500.11959/brapci/24086. Acesso em: 25 nov. 2023.

ROBERTSON, Stephen E.; JONES, K. Sparck. Relevance weighting of search terms. **Journal of the American Society for Information science**, v. 27, n. 3, p. 129-146, 1976.

SARACEVIC, T. Information science. **Journal of the American Society for Information Science**, v.50, n.12, p.1051, 1999.

VIEIRA, Simone Bastos. **La recuperación automática de información jurídica**: metodología de análisis lógico-sintáctico para la lengua portuguesa.1994. 382 f. Tese (Doutorado em Ciência da Informação) - Universidad Complutense de Madrid, Madrid, 1994.