UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Avaliação de um sistema de analítica visual para recuperação de informação em coleções de documentos

Sherlon Almeida da Silva

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Sherlon Almeida da Silva

Avaliação de um sistema de analítica visual para recuperação de informação em coleções de documentos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
Janeiro de 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A447a

Almeida da Silva, Sherlon
Avaliação de um sistema de analítica visual para recuperação de informação em coleções de documentos / Sherlon Almeida da Silva; orientadora Maria Cristina Ferreira de Oliveira. -- São Carlos, 2021.
117 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. RECUPERAÇÃO DA INFORMAÇÃO. 2. VISUALIZAÇÃO. 3. INTERAÇÃO USUÁRIO-COMPUTADOR. 4. PROCESSAMENTO DE TEXTO. I. Ferreira de Oliveira, Maria Cristina, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Sherlon Almeida da Silva

Evaluation of a visual analytics system for information retrieval in document collections

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
January 2022

Este trabalho é dedicado à meus pais,
os quais são meus eternos companheiros
e principais incentivadores.

AGRADECIMENTOS

Dedico este trabalho especialmente aos meus pais, pelo constante apoio e incentivo durante todas as etapas desta pesquisa e de minha vida. Sem vocês nada disso seria possível e por vocês tudo isso vale a pena!

Agradeço imensamente à minha orientadora Maria Cristina Ferreira de Oliveira por todo o suporte fornecido desde o início deste projeto. Suas incontáveis contribuições foram de extrema importância para o desenvolvimento deste trabalho e para o meu desenvolvimento pessoal. Hoje a senhora é minha maior referência e inspiração no meio acadêmico! Sou muito grato por sua ajuda, paciência e compreensão! Cada email trocado e cada conversa foram muito significativos e de grande valor para minha formação. Serei eternamente grato!

Agradeço também à banca, pelas contribuições, por sua disponibilidade e aceite em avaliar meu trabalho. Aos colaboradores desta pesquisa, como os professores Evangelos E. Milios (Dalhousie University), Osvaldo Novais de Oliveira Junior (IFSC-USP), Rosane Minghim (ICMC-USP) e a desenvolvedora do sistema TRIVIR, Amanda Dias (Microsoft), pelo tempo dedicado à reuniões e por suas dicas valiosas para o desenvolvimento deste trabalho.

Ressalto minha enorme gratidão a todos os participantes que se voluntariaram a participar dos estudos realizados. Agradeço também aos membros do Laboratório de Visualização, Imagens e Computação Gráfica (VICG), com os quais compartilhei momentos de conversa, trabalho, jogatinas e muito café, especialmente aos amigos Thales Oliveira Gonçalves e Eric Macedo Cabral! Agradeço aos professores do ICMC, que transmitiram seus conhecimentos sempre com dedicação durante as disciplinas que cursei, e todos os funcionários da USP, os quais tornaram esta experiência incrível.

Sou grato à minha parceira Isadora Ferrão, que esteve ao meu lado durante todos os momentos desta etapa, sempre proferindo palavras de tranquilidade e incentivo. Só nós sabemos o quanto precisamos abrir mão de alguns momentos juntos para que nossos trabalhos fossem desenvolvidos simultaneamente, no entanto cada passo tem sido planejado para estarmos crescendo lado a lado, sempre com companheirismo. Sou muito grato por sua ajuda e todas as suas dicas, ainda mais por confiar absolutamente na tua competência. Ter o teu apoio tornou este percurso mais suave. Obrigado por tudo mesmo, desde aquele "bora trabalhar", cafés da manhã surpresa para iniciar o dia bem, até aqueles convites de última hora para passeios e falar da vida. Quero que saiba que tua dedicação tanto pessoal quanto profissional me inspira! Teu apoio me motiva de uma forma incrível! Obrigado por fazer parte da minha vida! Obrigado por ter estado comigo sempre!

Gostaria de agradecer ao grupo de bolsistas CAPES do Facebook, o qual me auxiliou nos momentos mais difíceis deste percurso mesmo que indiretamente, seja por postagens engraçadas ou de incentivo. Pessoal, vocês nem devem ter noção do bem que esse grupo me fez, principalmente porque sou um daqueles "anônimos" que não interage muito nas postagens. No entanto, cada troca de experiências e comentários positivos, mesmo que em posts de outras pessoas, me deram forças para continuar a escrita de cada página. Todos sabemos o quanto a pós-graduação é desafiadora e que tem dias produtivos e outros que a vontade é desistir, mas o apoio de vocês fez a diferença. Obrigado!

Finalmente, agradeço o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por ter fornecido minha bolsa de mestrado e tornado possível minha dedicação exclusiva à este projeto de pesquisa. Agradeço também à Universidade de São Paulo, por sua contribuição à ciência em todas as àreas, sua qualidade de ensino, pesquisa e extensão e por ter me proporcionado momentos inesquecíveis nestes últimos anos.

“A imaginação muitas vezes nos leva a mundos que nunca sequer existiram.
Mas sem ela, não vamos a lugar nenhum.”
(Carl Sagan)

RESUMO

DA SILVA, S. A. **Avaliação de um sistema de analítica visual para recuperação de informação em coleções de documentos**. 2022. 114 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

A recuperação de informações de coleções de documentos é necessária em muitos contextos, por exemplo, pesquisadores desejam recuperar artigos sobre um tópico de pesquisa, médicos procuram prontuários de pacientes relacionados a uma determinada condição, investigadores de polícia buscam relações em relatórios criminais. Em comum a esses cenários, os usuários precisam identificar informações textuais relevantes em uma coleção de documentos. A tarefa é desafiadora, especialmente quando os usuários esperam por um processo de recuperação que não perca nenhum ou poucos documentos relevantes. Abordagens de *Visual Analytics* (VA) são frequentemente defendidas para apoiar tarefas de recuperação de documentos. VA depende da integração de visualizações interativas e algoritmos de aprendizado de máquina para que um especialista no domínio possa gradualmente conduzir um sistema para identificar os documentos relevantes. Como exemplo, o TRIVIR é um sistema do estado da arte que permite explorar um *corpus* enquanto fornece *feedback* a um classificador que sugere documentos potencialmente relevantes a um documento de consulta de referência. Avaliar as estratégias de recuperação de informações com suporte de VA também é um desafio, pois o uso desses sistemas geralmente envolve muitos aspectos conceituais e práticos e as tarefas de recuperação de texto podem exigir um esforço cognitivo considerável. Neste trabalho, são apresentados resultados de estudos observacionais sobre Recuperação de Informação (RI) de texto apoiada por VA. Foram conduzidas sessões com alunos de pós-graduação e pesquisadores usando o TRIVIR para explorar artigos científicos para fins de revisão de literatura. Um primeiro estudo permitiu recolher opiniões e identificar alguns problemas de usabilidade e limitações práticas da implementação disponível. Depois de tratar alguns problemas críticos observados no nível da interface, foi conduzida uma segunda rodada de sessões para coletar mais opiniões de usuários sobre um processo de recuperação auxiliado por VA. Concluiu-se que a maioria dos usuários tem uma visão muito positiva da usabilidade do sistema e da sua capacidade de facilitar as tarefas de recuperação. No entanto, também observou-se que uma introdução adequada aos diferentes elementos da interface é muito importante, e que pode ser difícil transmitir o modelo conceitual subjacente e suas limitações. Observou-se uma variação significativa na avaliação das funcionalidades específicas por diferentes usuários, e alguns deles podem enfrentar dificuldades práticas para utilizar o sistema adequadamente, de forma autônoma.

Palavras-chave: Recuperação de Informação Textual, Visualização de Texto, Análise Visual, Coleções de Documentos.

ABSTRACT

DA SILVA, S. A. **Evaluation of a visual analytics system for information retrieval in document collections**. 2022. 114 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Retrieving information from document collections is necessary in many contexts, for example, researchers wish to retrieve papers on a research topic, physicians search for patient records related to a certain condition, police investigators seek for relationships in criminal reports. Common to these scenarios are users in need of identifying relevant textual information in a document collection. The task is challenging, especially when users hope for a retrieval process that misses none or very few of the relevant documents. Visual Analytics (VA) approaches are often advocated to support document retrieval tasks. VA relies on integrating interactive visualizations and machine learning algorithms so that a domain expert can gradually steer a system into identifying the relevant documents. As an example, TRIVIR is a state-of-the-art system that allows exploring a corpus while providing feedback to a classifier that suggests potentially relevant documents to a reference query document. Assessing VA-supported Information Retrieval (IR) strategies is also challenging, as using these systems typically involves many conceptual and practical aspects and text retrieval tasks can demand considerable cognitive effort. In this work, we present results from observational studies on VA-supported text information retrieval. We conducted sessions with graduate students and researchers using TRIVIR to explore scientific papers for purposes of literature review. A first study allowed us to collect opinions and identify some usability issues and practical limitations of the available implementation. After handling some critical issues observed at the interface level, we conducted a second round of sessions in order to collect further user opinions regarding a retrieval process assisted with VA. We concluded that most users have a very positive view of the system's usability and its ability to facilitate their retrieval tasks. Nonetheless, we also learnt that a proper introduction to the role of the interface elements is important and that conveying the underlying conceptual model and its limitations can be difficult. We observed considerable variation in user assessment of the specific functionalities and some users may face practical difficulties in using the system autonomously in an optimal way.

Keywords: Textual Information Retrieval, Text Visualization, Visual Analysis, Document Collections.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AL	*Active Learning*
AM	Aprendizado de Máquina
BoW	*Bag-of-Words*
CAL	*Continuous Active Learning*
CBOW	*Continuous Bag-of-Words*
Corpus	Coleção de Documentos
FDP	*Force-Directed Placement*
HCML	*Human-Centered Machine Learning*
HLTM	*Hierarchical Latent Tree Model*
IA	Inteligência Artificial
IB	*Information Bottleneck*
IDF	*Inverse Document Frequency*
IHC	Interação Humano-Computador
iKmeans	*Interactive Kmeans*
KNN	*K Nearest Neighbor*
LDA	*Latent Dirichlet Allocation*
LDA	*Linear Discriminant Analysis*
LDC	*Lexical Double Clustering*
LLE	*Locally-Linear Embedding*
LSP	*Least Square Projection*
MDS	*Multidimensional Scaling*
MM	Modelo baseado em Molas
NMF	*Non-Negative Matrix Factorization*
PC	*Principal Components*
PCA	*Principal Component Analysis*
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informação
SVM	*Support Vector Machine*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
TAR	*Technology Assisted Review*
TCLE	Termo de Consentimento Livre e Esclarecido

TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
VA	*Visual Analytics*
VSM	*Vector Space Model*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Identificar informações relevantes em grandes bases de documentos textuais é uma tarefa desafiadora, tendo em vista as vastas coleções facilmente acessíveis a qualquer pessoa com um dispositivo eletrônico e acesso à internet. Embora seja relativamente simples coletar uma coleção representativa de documentos candidatos, avaliar quais satisfazem uma determinada necessidade de informação pode ser difícil e demorado, normalmente exigindo uma inspeção cuidadosa do conteúdo textual em uma base de documentos individual.

Técnicas de Visualização de Informação e Aprendizado de Máquina (AM) (MADAAN; BHATIA, 2020; PATHAK; PATHAK, 2020; ABDELGHAFAR; DARWISH; HASSANIEN, 2020) são frequentemente defendidas por seu potencial para mitigar os desafios envolvidos no processamento e análise de dados altamente complexos, incluindo dados textuais (ALENCAR; OLIVEIRA; PAULOVICH, 2012). Essas técnicas visam apoiar as tarefas de análise de dados, auxiliando os usuários na interpretação dos dados e na identificação das informações relevantes (KEIM, 2002). O campo de pesquisa em *Visual Analytics* (VA) investiga como integrar a capacidade de identificação de padrões de AM com a intuição proporcionada por representações visuais (KEIM *et al.*, 2008; THOMAS; COOK, 2006; THOMAS; COOK, 2005; OLIVEIRA; LEVKOWITZ, 2003). A suposição subjacente é que um especialista humano e algoritmos automatizados desempenham papéis complementares em um processo de investigação de dados e sua integração eficaz é essencial para lidar com problemas que envolvem dados complexos.

Um exemplo é o problema de recuperar informações de coleções de documentos. Esta é uma situação frequente em cenários variados, por exemplo, pesquisadores que estudam a literatura desejam recuperar artigos sobre um ou mais tópicos de pesquisa, médicos procuram registros de pacientes relacionados a uma determinada condição para identificar potenciais tratamentos, investigadores de polícia procuram relacionamentos em ocorrências criminais diversas. Comum a esses cenários são usuários que precisam identificar materiais relevantes em uma grande coleção que pode conter documentos potencialmente relevantes, partindo de uma

consulta inicial que expressa a sua necessidade de informação. Identificar os reais documentos de interesse é reconhecidamente uma tarefa complexa, pois as necessidades de informação podem mudar conforme o conhecimento do usuário evolui, motivando o desenvolvimento de sistemas de VA para facilitar essa tarefa.

Um exemplo é o TRIVIR, um sistema de VA que integra aprendizado de máquina e visualização interativa para recuperação de informação em documentos textuais assistida pelo usuário, com alta revocação – do inglês, *High Recall* (DIAS; MILIOS; OLIVEIRA, 2019). Os usuários podem explorar múltiplas visualizações que descrevem diferentes aspectos de uma coleção para identificar documentos relevantes às suas necessidades de informação, inicialmente expressas por um documento de consulta representativo do que o usuário procura. Os usuários interagem com as visualizações para fornecer *feedback*, o qual é usado no treinamento de um classificador para sugerir novos documentos potencialmente relevantes, em um ciclo iterativo de treinamento e aprendizagem.

1.1 Motivação e objetivo

Não basta apenas desenvolver os sistemas de VA, é necessário também investigar como eles são empregados em tarefas reais de RI, a fim de verificar se as abordagens de VA realmente atendem às necessidades dos usuários. Neste contexto, este projeto de mestrado tem como objetivo principal contribuir para o desenvolvimento de sistemas de VA voltados a tarefas de recuperação de informação, por meio da avaliação sistemática de um sistema de analítica visual para recuperação de informação em coleções de documentos. Para isto, foram conduzidos estudos observacionais com potenciais usuários do sistema TRIVIR, a fim de coletar dados qualitativos de usuários reais engajados em uma tarefa significativa de recuperação de documentos, simultaneamente coletando infomações qualitativas e quantitativas sobre a usabilidade e utilidade do sistema. Foram considerados cenários individuais de pesquisadores executando revisões de literatura e dispostos a usar um sistema de VA para o apoio a tarefas de RI .

1.2 Contribuições

As principais contribuições desta pesquisa são:

1. Investigação da utilização de um sistema de VA em tarefas reais de RI;
2. Realização de um estudo qualitativo com usuários;
3. Coleta e análise de avaliações qualitativas e quantitativas sobre usabilidade e utilidade do sistema TRIVIR;
4. Desenvolvimento e validação de melhorias para o sistema, com base nas avaliações dos usuários;

5. Identificação de dificuldades associadas ao uso de sistemas de VA em cenários práticos de recuperação de informação.

1.3 Organização do Documento

O restante deste texto está organizado como segue.

1. **Capítulo 2 - Fundamentação Teórica:** são apresentados conceitos essenciais para o desenvolvimento deste trabalho, como tendências consolidadas e atuais na área de *Visual Analytics*, Recuperação de Informação e Interação Humano-Computador (IHC).

2. **Capítulo 3 - Visualização como Suporte à Análise e Recuperação de Informação em Coleções de Documentos Textuais:** são apresentados e discutidos trabalhos relevantes relacionados ao tema deste trabalho.

3. **Capítulo 4 - Avaliação e Validação do Sistema TRIVIR:** é apresentada a metodologia desenvolvida para atingir os objetivos da proposta de pesquisa deste trabalho, bem como são apresentados e discutidos os resultados obtidos por meio de estudos observacionais com usuários.

4. **Capítulo 5 - Conclusões:** é destacada a conclusão deste trabalho, juntamente com a introdução de problemas de pesquisa em aberto e trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

O presente capítulo tem como objetivo introduzir conceitos técnicos para dar suporte à compreensão de todas as etapas deste trabalho. Na Seção 2.2 são abordados conceitos sobre recuperação de informação em coleções de documentos textuais. A seguir, a Seção 2.3 descreve a etapa de pré-processamento de dados textuais em tarefas de mineração e visualização de textos. São apresentadas na Seção 2.4 maneiras de representar os textos processados para que características descritivas sejam capturadas. Na Seção 2.5 são apresentadas técnicas de redução de dimensionalidade bem como a importância de seu papel em tarefas de análise de coleções de documentos. A identificação de documentos textuais relevantes pode ser aprimorada ao obter recomendações a partir da classificação de documentos por um algoritmo de aprendizado de máquina, e algumas técnicas relacionadas a isto são apresentadas na Seção 2.6. Por fim, são apresentadas técnicas de visualização computacional para auxílio na exploração de coleções de documentos (Seção 2.7), abordagens utilizadas em IHC (Seção 2.8) e as considerações finais do capítulo (Seção 2.9).

2.2 Recuperação de Documentos Textuais

O ato de recuperar informação refere-se à capacidade de obter conteúdo a partir de uma consulta (AN; HUANG; WANG, 2020; LUZ; CONEGLIAN; SEGUNDO, 2019). No entanto, ao realizar esta consulta o usuário está modelando conceitualmente a sua necessidade de busca, ou seja, a informação recuperada poderá atender às intenções de busca do usuário, podendo ou não ser considerada como relevante.

A recuperação de informação em coleções de documentos textuais possui diferentes aplicações (ZHANG *et al.*, ; RASTOGI; VERMA; KUMAR, 2020; SI *et al.*, 2020) e uma delas

está diretamente relacionada ao avanço da ciência. Por exemplo, considere um pesquisador realizando uma revisão da literatura em busca de trabalhos relacionados ao seu tema de pesquisa com o intuito de identificar lacunas no estado da arte e então definir um problema de pesquisa para ser solucionado. Logo, cada um dos trabalhos recuperados poderá ou não ser considerado relevante para a pesquisa. Geralmente, os repositórios de trabalhos acadêmicos como o Web Of Science [1], o Google Scholar [2], Scopus [3], IEEE Xplore [4] e ACM Digital Library [5] permitem consultas a partir de termos chave e os documentos são recuperados e apresentados como listas ranqueadas. Porém, o usuário precisa analisar os documentos retornados individualmente, em um processo que demanda esforço e tempo consideráveis.

Neste caso, assim como em outros cenários, técnicas de VA podem ser aplicadas para o desenvolvimento de ferramentas visuais e interativas (CHUANG; MANNING; HEER, 2012; PLEUSS; RABISER; BOTTERWECK, 2011), a fim de capturar a intenção de busca do usuário durante o processo de investigação. Assim, pode-se identificar e recomendar documentos possivelmente relevantes, reduzindo a carga de trabalho do usuário através da otimização do processo de busca.

2.2.1 Continuous Active Learning

Em aprendizado de máquina, é comum a utilização de dados disponíveis para o treinamento de um algoritmo. Na subárea de Aprendizado Supervisionado, estes dados contêm uma descrição de seu conteúdo, também conhecido como rótulo, o qual serve para indicar ao algoritmo a existência de categorias com determinadas características. Desta forma o algoritmo irá capturar padrões intrínsicos a cada categoria para realizar a tarefa a que se propõe. Por exemplo, em tarefas de classificação, a informação extraída de dados rotulados pode aprimorar a precisão da recomendação (VERMEULEN, 2020).

Em cenários de recuperação de informação em que existem muitos dados que não são rotulados, o *Active Learning* (AL) é frequentemente utilizado, com o objetivo de alcançar alta precisão nas recomendações de documentos relevantes, ao mesmo tempo em que minimiza o custo de obtenção de dados rotulados (SETTLES, 2009). Ou seja, AL em RI permite ao usuário modelar sua intenção de busca e se familiarizar com a coleção de documentos, ao mesmo tempo em que rotula alguns documentos alimentando o algoritmo de classificação, o qual molda a recomendação de acordo com a intenção de busca do usuário (RUBENS *et al.*, 2015).

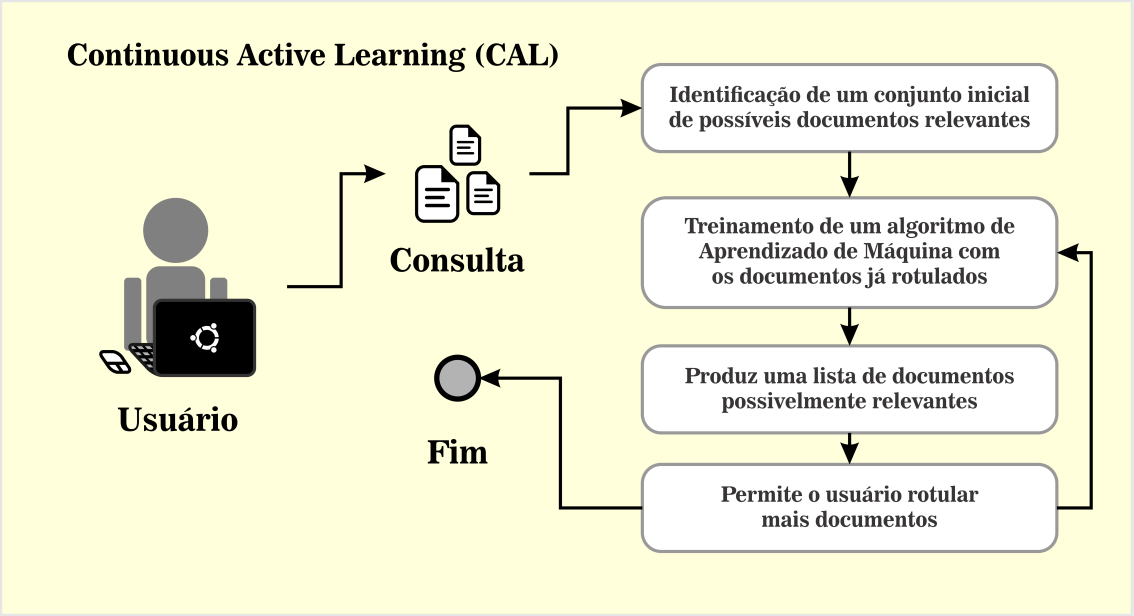
Durante o processo de investigação da coleção, o usuário recebe em lotes documentos não rotulados para identificar como relevantes ou não relevantes. Para auxiliar neste processo de investigação e recuperação de documentos relevantes, foi definido o procedimento de *Technology*

[1] Web of Science: <http://www.isiknowledge.com>
[2] Google Scholar: <https://scholar.google.com/>
[3] Scopus: <http://www.scopus.com>
[4] IEEE Explore: <http://ieeexplore.ieee.org>
[5] ACM Digital Library: <http://portal.acm.org>

Assisted Review (TAR) (CORMACK; GROSSMAN, 2014). O TAR trabalha com um conjunto de treinamento, o qual permite identificar termos importantes que caracterizam a relevância dos documentos, enquanto um algoritmo de classificação prevê documentos ainda não rotulados que podem ser relevantes (GROSSMAN; CORMACK, 2016c). Alguns exemplos de classificadores efetivos na identificação de documentos relevantes são: *Support Vector Machine* (SVM); Regressão Logística; *K Nearest Neighbor* (KNN); e Naive Bayes. Estes classificadores podem ser associados a diferentes protocolos TAR (GROSSMAN; CORMACK, 2016b), por exemplo o *Continuous Active Learning* (CAL) adotado no sistema TRIVIR (DIAS; MILIOS; OLIVEIRA, 2019).

O protocolo CAL compreende quatro etapas, ilustradas na Figura 1. **a)** A primeira etapa consiste na identificação de um conjunto inicial de possíveis documentos relevantes. **b)** Na segunda etapa um algoritmo de aprendizado de máquina é treinado com os documentos rotulados, o qual retornará uma classificação inicial. **c)** Durante a terceira etapa, o usuário rotula mais documentos informando *feedbacks* positivos e negativos, estendendo o conjunto de treinamento e reforçando assim o aprendizado do algoritmo. **d)** Na quarta etapa, as fases b) e c) podem ser repetidas, até que todos os documentos sejam rotulados, ou o usuário encerre o processo (GROSSMAN; CORMACK, 2016b).

Figura 1 – *Continuous Active Learning*.



Fonte: Elaborada pelo autor.

2.3 Pré-Processamento de Documentos Textuais

Em sistemas computacionais, a informação é codificada em linguagem binária, em que cada caractere, seja letra, número ou símbolo, é representado por uma sequência de *bits* (AMIRI;

NIKOUKAR, 2017). Por exemplo, para armazenar um documento de texto, uma mesma palavra escrita em letras maiúsculas ou em minúsculas será interpretada como palavras diferentes, pois possuem representação binária distinta. Esta e outras questões associadas à representação computacional dos textos devem ser resolvidas de maneira padronizada, permitindo o seu processamento computacional de forma coesa. Primeiramente é realizada uma etapa de pré-processamento de texto, na qual são utilizadas técnicas de conversão de palavras em letras minúsculas (BLUMENSTEIN; VERMA; BASLI, 2003), *tokenização* (GAMALLO; GARCIA, 2013), remoção de *stop-words* (SILVA; RIBEIRO, 2003) e a lematização (SINGH; GUPTA, 2016), como descrito a seguir.

Inicialmente, todo o conteúdo do texto é transformado em letras minúsculas (do inglês, *Lowercasing*) para que as palavras sejam representadas de maneira padronizada. Por exemplo, considere A = "Casa", B = "CASA"e C = "casa"; após aplicar esta técnica, A, B e C serão consideradas a mesma palavra, podendo ser interpretadas da mesma forma em tarefas de mineração. Posteriormente, é realizada uma etapa de segmentação de documentos textuais, em que há a remoção de espaços em branco e pontuações do texto. A partir deste momento, o documento passa a ser interpretado como um conjunto de palavras individuais. Esta etapa é fundamental para que o texto possa ser representado utilizando os modelos adotados pela comunidade de Processamento de Linguagem Natural (PLN), os quais serão apresentados na Seção 2.4.

Na construção de um texto, existem palavras que são semanticamente menos importantes do que outras, como é o caso de preposições, conjunções e artigos. Estas palavras, as quais são chamadas de *stop-words*, não trazem contribuição significativa para a análise de documentos textuais e consequentemente são removidas (KAUR; BUTTAR, 2018; RAVICHANDRAN; MOHANTA; NALINI, 2018; EL-KHAIR, 2017; SILVA; RIBEIRO, 2003).

A lematização busca identificar o radical de uma palavra para considerar suas possíveis conjugações como uma mesma palavra (TEER, 2018). Por exemplo, o sentido da palavra "Pensar" pode ser encontrado em várias outras palavras de um texto, como "Pensando", "Pensado", "Pensativo", entre outras. Assim, ao identificar o radical "Pens", todas estas palavras serão consideradas como a mesma, uma vez que trazem o mesmo significado semântico.

2.4 Representação de Documentos Textuais

Uma vez que os documentos foram pré-processados, a coleção de documentos agora encontra-se pronta para as etapas posteriores de mineração ou visualização. As comunidades que trabalham em mineração de textos, PLN e Visualização adotam modelos de representação dos documentos textuais em formato vetorial (MELO; MARTINS, 2017). A representação mais comum, conhecida como *Bag-of-Words* (BoW), considera uma simples contagem de palavras ocorrentes em um texto, em que cada documento é representado como um vetor de palavras e suas respectivas frequências de ocorrência no texto.

A coleção de documentos pode ser representada por uma matriz esparsa, na qual as linhas representam os documentos da Coleção de Documentos (Corpus), enquanto as colunas representam as palavras que ocorrem no corpus (Vocabulário). Esta estratégia de representação dos documentos em um espaço vetorial é conhecida como *Vector Space Model* (VSM) (HASHIMOTO *et al.*, 2016). Esta técnica é frequentemente utilizada e eficaz em diversas tarefas de processamento de informação, como classificação (AL-ANZI; ABUZEINA, 2018), ranqueamento (MITRA *et al.*, 2016) e agrupamento de documentos (DIAS; MILIOS; OLIVEIRA, 2019; MANNING; RAGHAVAN; SCHÜTZE, 2008).

Nas próximas subseções serão introduzidos alguns modelos difundidos para aprimorar a representação de documentos textuais e capturar as informações intrínsecas ao texto.

2.4.1 TF-IDF

Contabilizar apenas a frequência com que cada palavra aparece nos documentos captura apenas uma informação local acerca das palavras da coleção. Esta métrica pode ser aperfeiçoada ao utilizar *Term Frequency - Inverse Document Frequency* (TF-IDF)[6], uma medida estatística capaz de capturar a importância de cada palavra no corpus, a partir de uma análise global da coleção (CHEN *et al.*, 2016).

As definições formais do TF-IDF podem ser observadas nas Equações 2.1, 2.2 e 2.3. Esta medida considera o produto entre dois termos, o *Term Frequency* (TF) e o *Inverse Document Frequency* (IDF).

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D) \tag{2.1}$$

Primeiramente, é contabilizada a frequência d_t de um termo t em um documento d, a qual é normalizada pela quantidade total de termos d_k do documento d.

$$TF(t,d) = \frac{d_t}{\sum_k d_k} \tag{2.2}$$

Em seguida é calculada a frequência inversa dos documentos (IDF), dada pelo logarítmo do número total de documentos no corpus $|D|$ dividido pelo número de documentos do corpus que contém o termo t.

$$IDF(t,D) = log\frac{|D|}{|\{d_i \in D \mid t \in d_i\}|} \tag{2.3}$$

A motivação para utilização desta abordagem é que palavras em uma coleção possuem diferentes níveis de importância. O TF-IDF é capaz de representar esta importância numericamente como "pesos" obtidos a partir de uma análise global da coleção, aprimorando a representação do

[6] TF-IDF: <http://www.tfidf.com/>

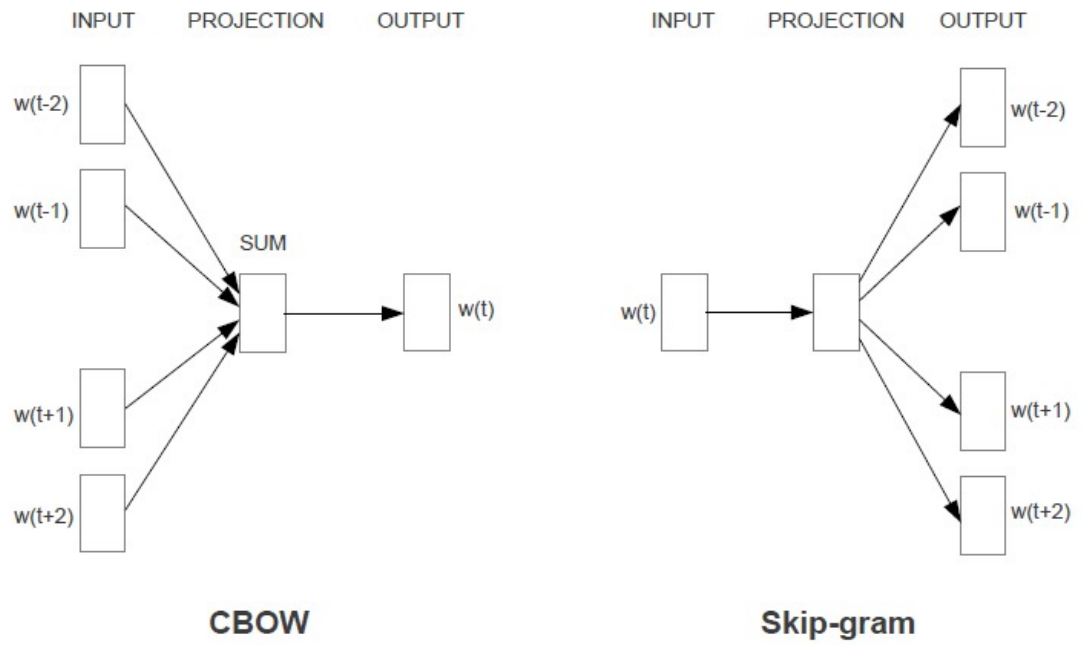
corpus para posterior identificação de documentos relevantes. Embora existam outras abordagens para atribuição de pesos ao modelo BoW, TF-IDF é o mais popular (TURNEY; PANTEL, 2010).

2.4.2 Word Embbedings

A representação de documentos utilizando o modelo BoW apresenta limitações, uma vez que não considera a ordem e o contexto em que as palavras se encontram no texto. Com isso, um novo modelo de representação textual, intitulado *Word Embbedings* foi introduzido (MIKOLOV *et al.*, 2013). Este modelo utiliza redes neurais para gerar uma representação vetorial das palavras do corpus a partir de um treinamento realizado em grandes coleções textuais. Este treinamento utilizando uma grande quantidade de informação permite ao modelo capturar o contexto em que palavras ocorrem, contornando as limitações do modelo BoW.

Mikolov *et al.* (2013) apresentaram novas arquiteturas para representações vetoriais de palavras (Figura 2). A primeira, à esquerda na figura, é o *Continuous Bag-of-Words* (CBOW), que prediz uma palavra alvo considerando as demais palavras que a cercam. O segundo modelo, à direita na figura, é o Skip-gram, o qual também é uma rede neural, porém esta prevê as palavras mais prováveis que aparecem na vizinhança de uma palavra dada.

Figura 2 – Modelo de Representação CBoW - Skipgram.



Fonte: Mikolov *et al.* (2013).

Ao utilizar este modelo para capturar o contexto em que as palavras ocorrem em grandes coleções de documentos, relações inerentes entre as palavras são capturadas. Por exemplo, a seguinte operação algébrica entre a representação vetorial das palavras w("Paris") - w("France") +

w("Italy") produz um resultado que é próximo à representação vetorial da palavra w("Rome") (MIKOLOV; YIH; ZWEIG, 2013).

O trabalho de Mikolov *et al.* (2013) foi o precursor, consolidando a representação de *Word Embbedings* e inspirando o aperfeiçoamento da técnica e a elaboração de novas propostas em trabalhos posteriores, como: *Word2Vec* (MIKOLOV *et al.*, 2013); Doc2Vec (LE; MIKOLOV, 2014); GloVe (PENNINGTON; SOCHER; MANNING, 2014); FastText (BOJANOWSKI *et al.*, 2016; JOULIN *et al.*, 2016a; JOULIN *et al.*, 2016b); BERT (DEVLIN *et al.*, 2018); entre outros.

2.5 Projeção de Dados Multidimensionais

Em tarefas de recuperação de informação, ter uma visão geral do conteúdo da coleção é importante para identificar relações entre os dados (LEHMANN; THEISEL, 2016; XU *et al.*, 2017). Estas relações, quando a recuperação de informação atua sobre dados textuais, são capturadas ao analisar o modelo de representação dos documentos em busca de padrões.

Os dados são representados por vetores de características extraídos da coleção de documentos, os quais podem ter milhares de dimensões. Visto que para um usuário é difícil investigar espaços de características com mais de três dimensões (NONATO; AUPETIT, 2018), existem técnicas que permitem reduzir a dimensionalidade do espaço de características, consequentemente, reduzindo a complexidade da representação e potencialmente da análise dos dados (KOSS *et al.*, 2020; REINBOLD; KUMPF; WESTERMANN, 2019; NYMAN, 2019; NILASHI; IBRAHIM; BAGHERIFARD, 2018; SHAO *et al.*, 2018; WANG; FANG; WANG, 2016). Estas técnicas de redução de dimensionalidade operam sobre dados em um espaço de características de alta dimensionalidade e projetam em um espaço de características de baixa dimensionalidade algum tipo de relação capturada (UZNAŃSKI, 2020; GURU *et al.*, 2020; JANAKIRAMAIAH *et al.*, 2020).

Os algoritmos de redução de dimensionalidade representam as relações entre os dados por meio de uma métrica de similaridade, por exemplo, a dissimilaridade entre os dados pode ser aproximada pela distância Euclidiana (CANTINI *et al.*, 2020). No entanto, a relação entre documentos usualmente é medida pela similaridade do cosseno (JADHAV; HOLAMBE, 2008), visto que opera sobre ângulos entre vetores multidimensionais, sendo capaz de mitigar o ruído intrínseco a dados esparsos, como é o caso de características provenientes de dados textuais.

Uma vez que a relação entre os dados é obtida, torna-se possível visualizar esta relação utilizando técnicas de visualização computacional (GALLAGHER *et al.*, 2020; DRAGAN *et al.*, 2019; JENTNER; KEIM, 2019). Por exemplo, pode-se optar por uma nuvem de pontos bidimensional, onde os eixos são coordenadas geradas por um algoritmo de redução da dimensionalidade e projeção bidimensional (JARROUSH *et al.*, 2019). Alguns algoritmos que trabalham com dados de alta dimensionalidade são apresentados e discutidos nas subseções a seguir.

2.5.1 Principal Component Analysis (PCA)

A técnica *Principal Component Analysis* (PCA) (JOLLIFFE, 2002; JOLLIFFE, 1986) reduz a dimensionalidade dos dados ao identificar combinações lineares ortogonais das características originais, conhecidas como componentes principais (*Principal Components* (PC)) (WANG *et al.*, 2020; AÏT-SAHALIA; XIU, 2019; CUI; LI; ZHANG, 2019). Em um espaço de características N dimensional, onde N é o número de atributos do vetor de características (Dimensões), os componentes principais representam aqueles atributos com maior variabilidade (ARMENI *et al.*, 2019). Em outras palavras, existem N componentes principais ordenados em ordem decrescente, onde o primeiro (PC1) representa o atributo cuja variação é a maior, ou seja, aquele que é o mais descritivo no espaço de características, para representar o conteúdo da base de dados da qual foi extraído.

2.5.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

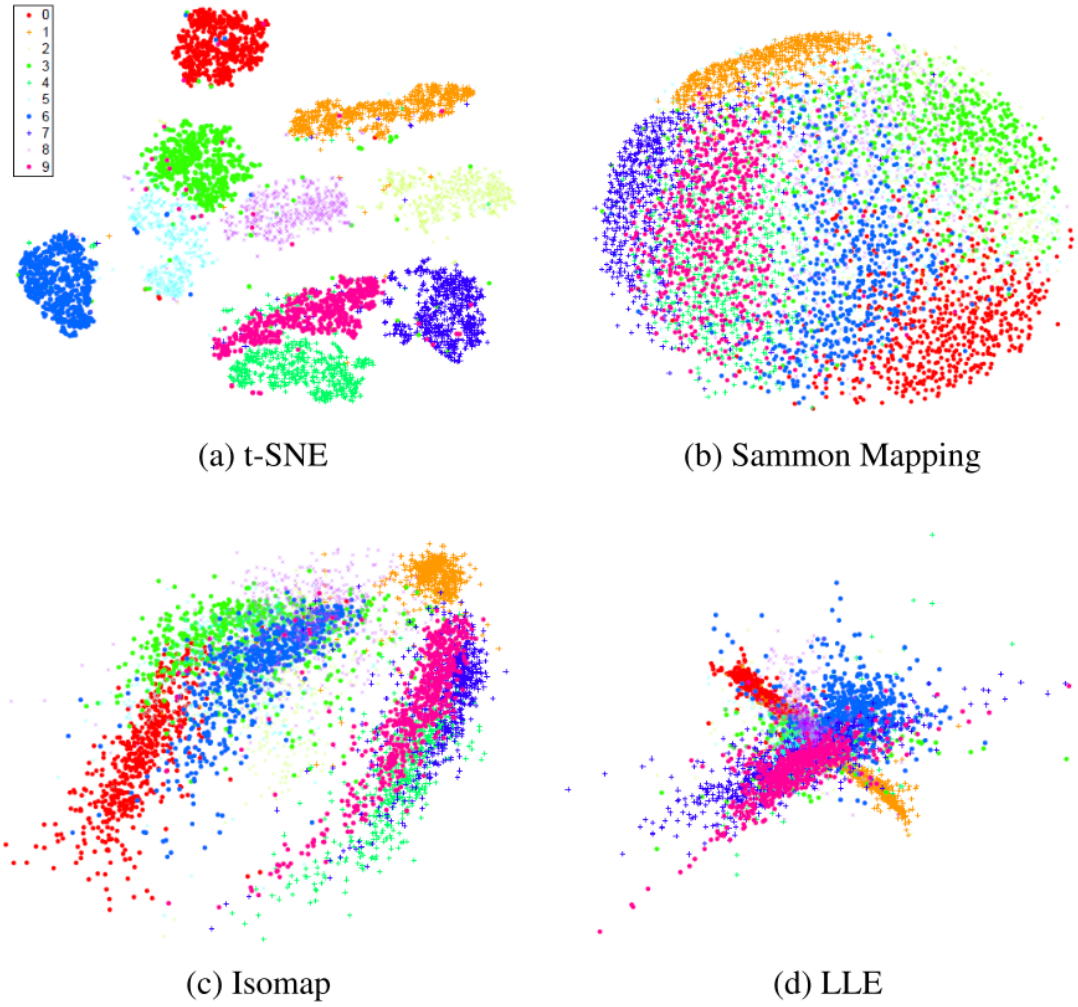
Uma das técnicas de redução da dimensionalidade dos dados amplamente difundida e adotada pela comunidade de VA é a *t-Distributed Stochastic Neighbor Embedding* (t-SNE) (MACKAY; KUSALIK, 2020; ZHOU; JIN, 2020; LOPES; NETO; MARTINS, 2020; LINDERMAN; STEINERBERGER, 2019; MAATEN; HINTON, 2008). Esta técnica tenta minimizar a divergência entre os dados no espaço de características original e no espaço de características reduzido. Ou seja, tem por objetivo encontrar um conjunto de pontos em baixa dimensão que represente o mais fielmente possível as relações entre os dados originais (RITZ *et al.*, 2020). Diferentemente de técnicas lineares, como o PCA, o qual busca maximizar a variância e preservar uma ampla variabilidade do espaço de características, o t-SNE (MAATEN; HINTON, 2008) é não linear e procura preservar pequenas distâncias em pares de pontos, ou seja, identificar similaridades locais nos dados (LINDERMAN *et al.*, 2019; KOBAK; LINDERMAN, 2019).

Em estudos de validação realizados por Maaten e Hinton (2008), apresentados na Figura 3, foi comprovada a efetividade do t-SNE para a segregação visual de grupos no espaço dimensional reduzido, obtendo melhores resultados do que outras técnicas de redução da dimensionalidade, como Sammon Mapping (SAMMON, 1969), Isomap (TENENBAUM; SILVA; LANGFORD, 2000; TENENBAUM, 1998) e *Locally-Linear Embedding* (LLE) (ROWEIS; SAUL, 2000; SAUL; ROWEIS, 2000).

2.5.3 Least Square Projection (LSP)

Paulovich *et al.* (2008) apresentaram a técnica *Least Square Projection* (LSP), a qual propõe preservar relações de vizinhança entre os dados. Esta técnica realiza a projeção multidimensional a partir de alguns passos. Primeiramente, são definidos pontos de controle representativos da variabilidade na distribuição dos dados. Em seguida, estes pontos de controle selecionados são projetados utilizando a técnica *Multidimensional Scaling* (MDS) (MACHADO; LOPES,

Figura 3 – Comparação de técnicas de redução de dimensionalidade em um espaço projetado 2D aplicadas ao dataset MNIST.

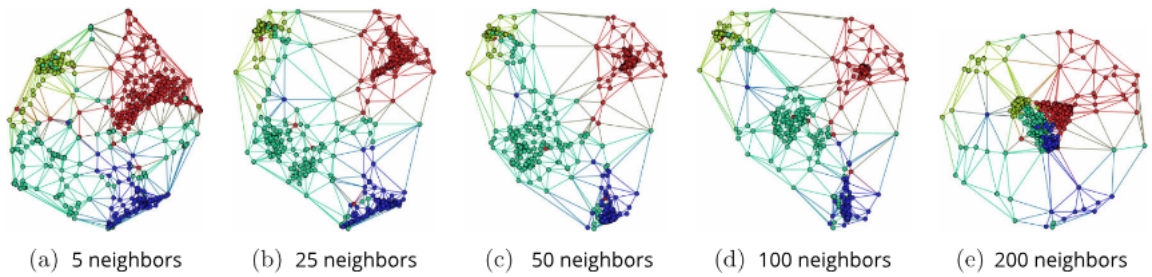


Fonte: Maaten e Hinton (2008).

2020; TORGERSON, 1952), conhecida por ser precisa, porém computacionalmente custosa.

Dada a projeção destes pontos de controle, a projeção dos demais dados é obtida por interpolação, resolvendo um sistema linear esparso para estabelecer a disposição dos pontos nas proximidades de seus vizinhos mais próximos (CABRAL, 2020). Um dos parâmetros ajustáveis desta técnica é o tamanho da vizinhança a ser preservada, o qual está relacionado à densidade dos grupos formados na projeção, isto é, quanto maior o número de vizinhos, mais densos serão os grupos formados no espaço projetado, como é possível observar na Figura 4. A LSP é uma técnica de alta precisão com custo computacional reduzido, especialmente no caso de dados não lineares definidos em espaços esparsos de alta dimensionalidade, como é o caso de representações de documentos (DIAS, 2019).

Figura 4 – Variação do número de vizinhos na técnica LSP aplicado a um corpus de 675 documentos relacionados à Ciência da Computação.



Fonte: Paulovich *et al.* (2008).

2.5.4 Force-Directed Placement (FDP)

A técnica *Force-Directed Placement* (FDP) (CHEONG; SI, 2020; FRUCHTERMAN; REINGOLD, 1991) pode ser usada para modelar os dados do corpus como um grafo, em que cada nó e aresta por sua vez representam, respectivamente, um documento e a relação, que pode indicar similaridade entre os documentos a partir de sua vizinhança. A avaliação é com um Modelo baseado em Molas (MM) (FRICK; LUDWIG; MEHLDAU, 1994), o qual tem por objetivo reduzir o número de arestas sobrepostas ao empregar forças de atração e repulsão entre os nós do grafo, ao mesmo tempo em que diminuem e aumentam a distância entre documentos similares e dissimilares, respectivamente (CHEONG; SI, 2020; SCHÖNFELD; PFEFFER, 2019).

Inicialmente, os nós do grafo são posicionados aleatoriamente. Em seguida o modelo, descrito pela Equação 2.4, tenta minimizar a "energia" de todos os nós do grafo de forma iterativa, aproximando-se cada vez mais de um estado global de equilíbrio. A força referente a cada nó do grafo é descrita pela Equação 2.5 de forma que: $\Gamma(i)$ representa o conjunto de nós vizinhos ao nó i; X constitui o conjunto de nós; K e C denotam parâmetros para adaptar as forças de atração e repulsão (CABRAL, 2020; XU; YANG; GOU, 2018; HU, 2005).

$$Energy(X,K,C) = \sum_{i \in |X|} f^2(i,X,K,C) \tag{2.4}$$

$$f(i,X,K,C) = \sum_{i \neq j} -\frac{CK^2}{\left\|x_i - x_j\right\|^2}(x_j - x_i) + \sum_{j \in \Gamma(i)} \frac{\left\|x_i - x_j\right\|}{K}(x_j - x_i) \tag{2.5}$$

Para que um modelo baseado em molas seja utilizado como uma estratégia de projeção multidimensional, as forças são calculadas de forma proporcional à diferença entre as dissimilaridades $\gamma(x_i,x_j)$ entre pontos no espaço de características original e as distâncias $d(y_i,y_j)$ entre os pontos no espaço gerado (PAULOVICH *et al.*, 2008).

2.6 Classificação de Documentos Textuais

A classificação de documentos é uma abordadem de aprendizado supervisionada, ou seja, leva em consideração os rótulos de documentos já classificados para inferir um modelo capaz de prever as classes de novos documentos ainda não rotulados (BELKEBIR; GUESSOUM, 2015; GUO; KORHONEN; POIBEAU, 2011; CARUANA; NICULESCU-MIZIL, 2006). Isto significa que para conseguir realizar previsões precisas, o algoritmo de classificação deve aprender as características que diferenciam os documentos pertencentes a cada classe, na etapa de treinamento.

No entanto, tarefas de recuperação de informação em coleções com poucos documentos rotulados não serão capazes, em um primeiro momento, de gerar uma classificação precisa. Para contornar esta limitação, pode-se aplicar o CAL (CORMACK; GROSSMAN, 2015a; CORMACK; GROSSMAN, 2015b), o qual conduzirá o usuário na investigação dos documentos e utilizará cada novo documento rotulado como informação complementar para estender o conjunto de treinamento e aprimorar o resultado do classificador.

Visto que inicialmente haverá pouca informação sobre documentos potencialmente relevantes, e também muitas vezes os usuários não conseguem modelar uma consulta suficientemente representativa, a primeira etapa do treinamento pode ser realizada a partir da recomendação dos K vizinhos mais próximos a um ou mais documentos que o usuário já sabe que são relevantes. Além disso, as demais recomendações poderão dar um peso maior para documentos rotulados pelo usuário. Ou seja, o algoritmo irá extrair características destes documentos passados como referência e irá recomendar ao usuário primeiramente aqueles que possuem maior similaridade. Assim, o classificador se torna mais preciso à medida em que o usuário identifica trabalhos relevantes e refina sua perspectiva acerca da sua intenção de busca (CORMACK; GROSSMAN, 2015a; CORMACK; GROSSMAN, 2015b).

2.7 Visualização de Documentos Textuais

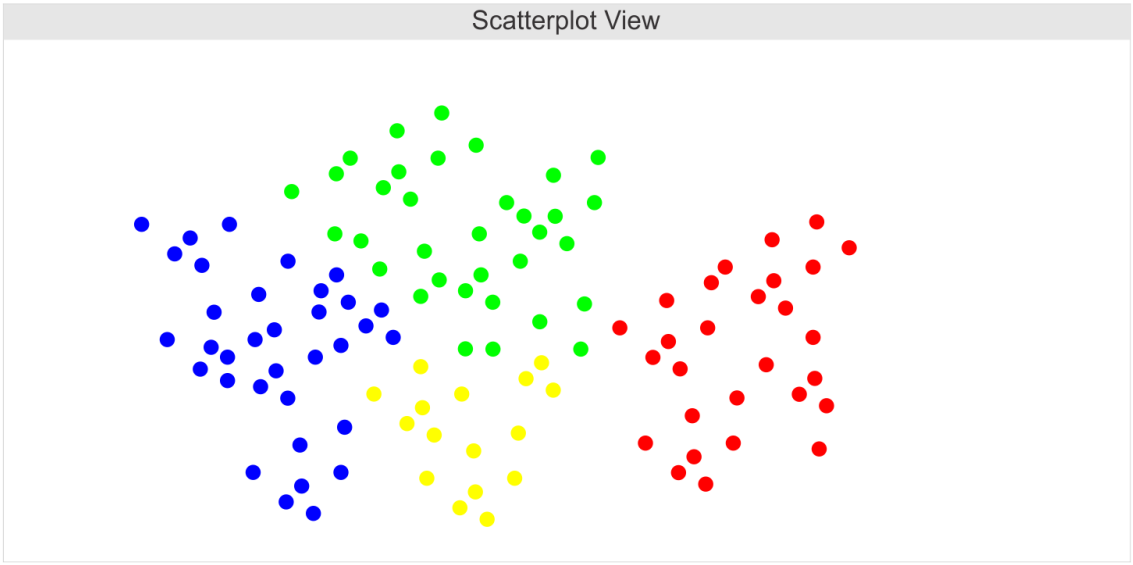
As técnicas de visualização computacional auxiliam o usuário no processo de investigação de coleções de documentos (HURTER, 2018; SCHIRRMEISTER *et al.*, 2017; KUCHER; KERREN, 2014; HOULDING, 2012; ALENCAR; OLIVEIRA; PAULOVICH, 2012). A seguir são apresentadas algumas delas.

2.7.1 Nuvem de Pontos

A representação vetorial de documentos textuais pode conter milhares de atributos. Para facilitar a análise são aplicadas técnicas de redução de dimensionalidade dos dados (CHEN, 2019; WANG; SUN, 2015), e para o suporte visual a esta tarefa pode ser utilizada uma nuvem de pontos, como ilustrado na Figura 5.

Cada glifo, neste caso círculos, representa um documento da coleção, enquanto que a sua disposição bidimensional é calculada por algoritmos de redução de dimensionalidade, como o t-SNE ou LSP. Esta visualização apresenta diversas informações sobre a coleção de documentos, como a similaridade entre documentos, representada como a distância entre os círculos e a formação natural de grupos. É possível apresentar cores com diferentes significados, como a indicação de grupos de documentos similares ou os rótulos provenientes da classificação de relevância, como no sistema TRIVIR (DIAS; MILIOS; OLIVEIRA, 2019).

Figura 5 – Visualização de uma Nuvem de Pontos, em que cada círculo representa um documento do corpus.



Fonte: Elaborada pelo autor.

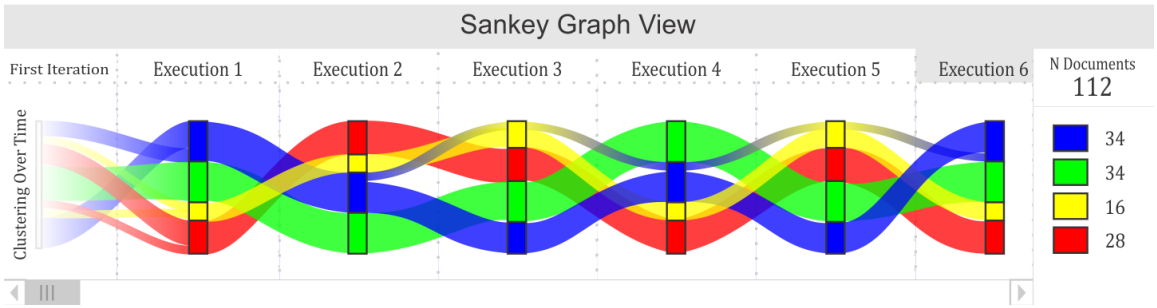
2.7.2 Sankey Graph

A técnica *Sankey Graph* (RIEHMANN; HANFLER; FROEHLICH, 2005) apresenta um fluxo temporal, permitindo capturar informações sobre o processo evolutivo dos grupos de documentos similares ou das classes de relevância obtidas por uma classificação. A Figura 6a ilustra a evolução temporal dos grupos de documentos em um sistema de agrupamento interativo, como em Sherkat *et al.* (2018). Este tipo de visualização permite explorar tanto o comportamento do corpus, como de cada grupo ou documento individualmente (Figura 6b).

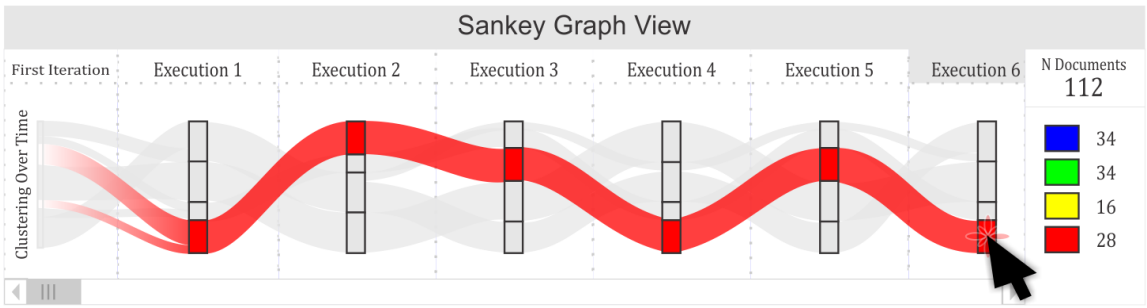
2.7.3 Wordclouds

Na análise de documentos textuais é comum encontrar a técnica de *Wordcloud* compondo ferramentas no estado da arte (PHILIP, 2020; WANICHAVORAPONG; YUSOF *et al.*, 2019; SONI *et al.*, 2019; MCGOWAN; CHANEY, 2019; HUANG; WANG; YE, 2019; DRAGAN *et al.*, 2019; JEMISON *et al.*, 2018; CHERAPANUKORN; CHAROENKWAN, 2018; JAYASHANKAR; SRIDARAN, 2017). Isto deve-se ao fato de ser um recurso visual intuitivo, visto que

Figura 6 – Visualização *Sankey Graph* para visualização temporal de agrupamentos.



(a) Visualização Temporal de diversos agrupamentos.



(b) Visualização Temporal de um agrupamento selecionado.

Fonte: Elaborada pelo autor.

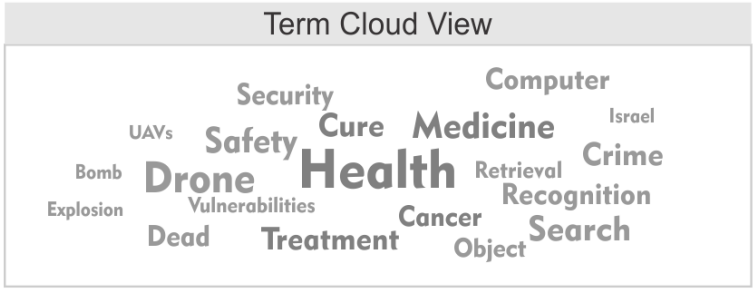
apresenta palavras mais importantes em um corpus, grupo(s) ou documento(s) e codifica a importância de cada uma por tamanho e/ou cores das palavras. Ao contrário da nuvem de pontos, na nuvem de palavras o posicionamento das palavras não codifica informação. *Wordclouds* são utilizadas para uma análise primária do texto, permitindo obter informações sobre o seu conteúdo rapidamente (PAREJO *et al.*, 2021).

Na Figura 7 é apresentado um exemplo de *wordcloud* aplicada a uma coleção de documentos contendo como mais importantes as palavras "Health", "Medicine", "Drone" e "Cure". No entanto, outras palavras como "Bomb", "Explosion"e "Dead"estão presentes. Isto informa ao usuário que possivelmente o corpus contempla diferentes conteúdos e que eles estão relacionados a saúde e atentados terroristas. Esta análise pode ser realizada apenas observando a frequência das palavras, sem requerer do usuário que analise cada documento individualmente.

2.8 Interação Humano-Computador (IHC)

Para a garantia de qualidade de um sistema, para que ele se torne amplamente aceito, efetivamente usado e atinja o maior número possível de usuários, é importante considerar os conhecimentos da área de IHC. Segundo Rocha e Baranauskas (2003), para o sucesso de um sistema, o seu *design* deve ser adequado a todas as pessoas, de modo que seja projetado considerando as necessidades e capacidades dos usuários alvo. Ainda segundo as autoras, esses usuários não devem ser obrigados a pensar como o sistema funciona. De acordo com Moggridge

Figura 7 – Visualização baseada em uma Nuvem de Palavras (*Wordcloud*) representativas em um corpus, grupo(s), ou documento(s).



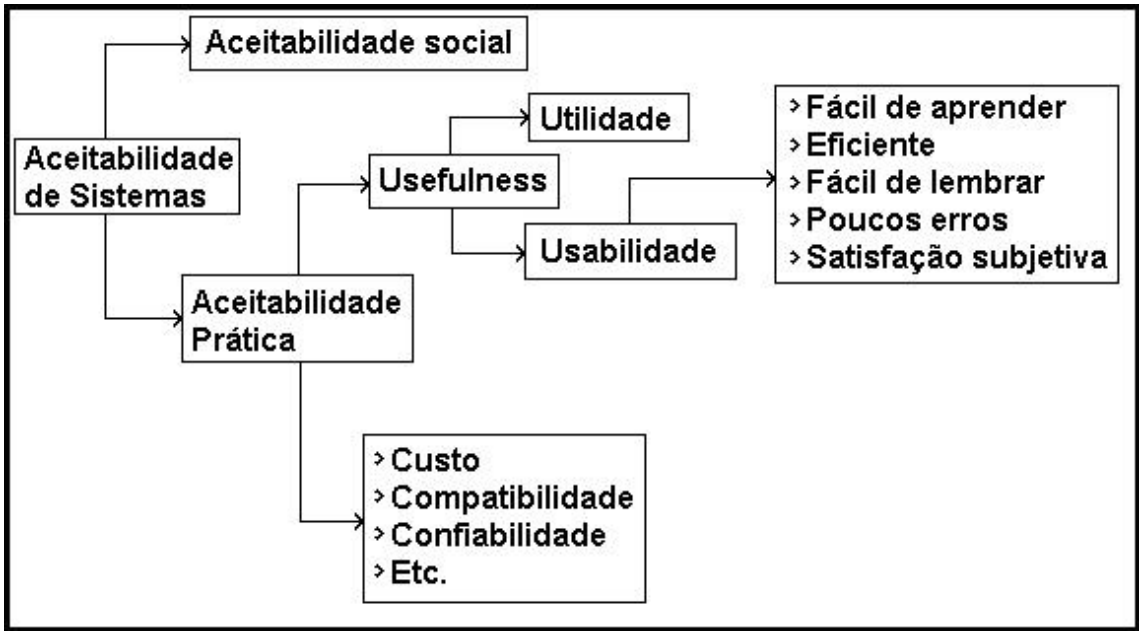
Fonte: Elaborada pelo autor.

e Atkinson (2007), a interação humano-computador é a área que investiga o projeto, avaliação e implementação de sistemas de computação interativos para uso humano e o estudo dos principais fenômenos que os cercam. Vale ressaltar que o termo IHC engloba todos os aspectos relacionados à interação entre usuários e computadores, não focando apenas em *design* de interfaces (ROCHA; BARANAUSKAS, 2003).

Os principais objetivos de IHC, segundo Rocha e Baranauskas (2003) são produzir sistemas usáveis, seguros e funcionais. Ainda segundo as autoras, esses objetivos podem ser resumidos em desenvolver ou melhorar a segurança, utilidade, efetividade e usabilidade de sistemas computacionais. Para isso, a IHC busca a aceitabilidade social e prática do sistema, conforme é ilustrado na Figura 8. De acordo com Rocha e Baranauskas (2003), a aceitabilidade geral é a combinação de sua aceitabilidade social e prática. Por exemplo, pode-se analisar os dois tipos de aceitabilidade em sistemas de controle de portas de entrada em bancos. Apesar de serem benéficos socialmente, pois tentam impedir situações de assalto, não são aceitos socialmente pois levam a que qualquer pessoa que queira entrar no banco tenha que esbarrar na porta trancada por inúmeras vezes até se desfazer de todo e qualquer objeto suspeito. Ainda de acordo com as autoras, a aceitabilidade prática trata dos tradicionais parâmetros de custo, confiabilidade e compatibilidade com sistemas existentes, como também da categoria denominada "*usefulness*". Segundo Nielsen (1995), o termo "*Usefulness*" é utilizado para atingir um objetivo pré-determinado através da combinação de **utilidade** e **usabilidade** (Figura 8). A utilidade deve verificar se a funcionalidade do sistema faz o que deve ser feito, e a usabilidade significa o quão bem os usuários podem usar a funcionalidade definida.

No contexto de IHC, a acessibilidade é uma qualidade relativa que depende da conexão entre as capacidades funcionais dos usuários com o *design* do sistema. Segundo Ferraz (2017), as pessoas com deficiências podem perceber e entender sistemas, navegar por eles e interagir com eles, além de poderem contribuir para a evolução destes sistemas. Já a usabilidade é uma importante característica a ser considerada no desenvolvimento de sistemas para tornar mais fácil a utilização por diferentes usuários. De acordo com Moumane, Idri e Abran (2016), a usabilidade é uma característica que corresponde a qualidade de um sistema computacional,

Figura 8 – *Atributos de aceitabilidade*.



Fonte: Adaptada de Nielsen (1995).

altamente relacionado com a sua eficiência, eficácia e satisfação do usuário na utilização.

2.8.1 Métodos de avaliação de usabilidade

Através dos estudos de usabilidade é possível avaliar a facilidade de aprendizado e uso, satisfação do usuário, flexibilidade, utilidade, segurança de uso, entre outros. No que tange aos métodos de avaliação de usabilidade de interfaces, existem diferentes categorias, que serão detalhadas adiante:

- **Métodos analíticos e de inspeção:** avaliação heurística e percurso cognitivo (OLIVEIRA; SILVA, 2017; MAZZA, 2009);
- **Métodos empíricos ou testes com usuários:** teste de usabilidade e percurso pluralístico (RODRIGUES *et al.*, 2012);
- **Outras formas:** questionários (BOUCINHA; TAROUCO, 2013).

2.8.1.1 Métodos analíticos

De acordo com Mack e Nielsen (1995), os métodos de avaliação analíticos são aqueles nos quais os avaliadores examinam ou inspecionam aspectos de interfaces de usuário relacionados a usabilidade. Vale ressaltar que os avaliadores geralmente são especialistas em usabilidade. São exemplos de métodos analíticos as avaliações heurísticas e o percurso cognitivo.

A avaliação heurística visa identificar uma série de problemas de usabilidade, conforme o conjunto de heurísticas proposto por Nielsen e Molich (1990) e ocorre com um pequeno grupo

de avaliadores que são responsáveis por examinar a interface e avaliar de acordo com critérios de usabilidade. As dez heurísticas propostas por Nielsen e Molich (1990), conforme tradução por Rocha e Baranauskas (2003), são:

- **Visibilidade do status do sistema:** o sistema precisa manter os usuários informados sobre o que está acontecendo, fornecendo um *feedback* adequado dentro de um tempo razoável.
- **Compatibilidade do sistema com o mundo real:** o sistema precisa falar a linguagem do usuário, com palavras, frases e conceitos familiares ao usuário, ao invés de termos orientados ao sistema. Seguir convenções do mundo real, fazendo com que a informação apareça numa ordem natural e lógica.
- **Controle do usuário e liberdade:** os usuários frequentemente escolhem por engano funções do sistema e precisam ter claras saídas de emergência para sair do estado indesejado sem ter que percorrer um extenso diálogo. Prover funções *undo* e *redo*.
- **Consistência e padrões:** os usuários não precisam adivinhar que diferentes palavras, situações ou ações significam a mesma coisa. Seguir convenções de plataforma computacional.
- **Prevenção de erros:** melhor que uma boa mensagem de erro é um *design* cuidadoso o qual previne o erro antes dele acontecer.
- **Reconhecimento ao invés de relembrança:** tornar objetos, ações e opções visíveis. O usuário não deve ter que lembrar informação de uma para outra parte do diálogo. Instruções para uso do sistema devem estar visíveis e facilmente recuperáveis quando necessário.
- **Flexibilidade e eficiência de uso:** usuários novatos se tornam peritos com o uso. Prover aceleradores de forma a aumentar a velocidade da interação. Permitir a usuários experientes "cortar caminho"em ações frequentes.
- **Estética e *design* minimalista:** diálogos não devem conter informação irrelevante ou raramente necessária. Qualquer unidade de informação extra no diálogo irá competir com unidades relevantes de informação e diminuir sua visibilidade relativa.
- **Ajudar os usuários a reconhecer, diagnosticar e corrigir erros:** mensagens de erro devem ser expressas em linguagem clara (sem códigos) indicando precisamente o problema e construtivamente sugerindo uma solução.
- ***Help* e documentação:** embora seja melhor um sistema que possa ser usado sem documentação, é necessário prover *help* e documentação. Essas informações devem ser fáceis de encontrar, focalizadas na tarefa do usuário e não muito extensas.

Dentre os pontos positivos destes métodos destaca-se a utilização de princípios de usabilidade e a possibilidade de adoção já no início do ciclo de desenvolvimento. No entanto, é

pouco adequado para identificar as necessidades não conhecidas dos usuários (MACIEL *et al.*, 2004).

O percurso cognitivo simula passo-a-passo o comportamento de um usuário em uma tarefa (MATERA; RIZZO; CARUGHI, 2006; CHI *et al.*, 2003). Em cada um dos passos, os avaliadores respondem quatro perguntas relativas a aspectos de usabilidade, que abrangem aspectos cognitivos (sobre como fazer a tarefa), motores (sobre se o usuário conseguirá executar a ação referente à tarefa) e sensoriais (como perceber se há *feedback* do sistema). Este método é efetivo para a identificação de problemas decorrentes da interação com o sistema e na habilidade de ajudar na definição dos objetivos e ações do usuário, no entanto, um dos problemas é que os resultados dos testes podem ficar enviesados pela visão do especialista de UX que fez a avaliação (BONIFÁCIO *et al.*, 2010). Além disso, por ser um método focado em tarefas, outra possível limitação é que apenas as tarefas efetivamente analisadas tenham os seus problemas revelados.

2.8.1.2 Métodos empíricos

Os métodos de avaliação empíricos permitem analisar os fatores que caracterizam a usabilidade de um sistema no que diz respeito à sua flexibilidade, satisfação do usuário, facilidade de uso e aprendizado, eficiência de uso e segurança. Com esses métodos é possível obter medidas quantificáveis dos critérios estabelecidos por meio da determinação de limites máximos, mínimos e observação direta do usuário (GONÇALVES *et al.*, 2011)

Os estudos de usabilidade ocorrem por meio da observação da interação do usuário com o sistema, com o uso de técnicas que permitam deixar claro suas decisões e desejos perante o sistema. A participação de usuários nos estudos permite confirmar suposições sobre o sistema, em outras palavras, nada melhor do que observar um usuário interessado utilizando o sistema na prática. No entanto, este tipo de estudo exige um esforço de preparação do ambiente e recrutamento de usuários.

O percurso pluralístico (BIAS, 1991) é semelhante ao percurso cognitivo mencionado na subseção anterior, com a diferença de que um grupo misto de usuários de diferentes áreas (por exemplo, designer, desenvolvedores, etc) avalia em conjunto e negocia as respostas.

2.8.1.3 Outros tipos de estudos

Nos questionários, um conjunto de questões é apresentado aos usuários, normalmente com opções de múltipla escolha ou perguntas abertas. Este tipo de estudo identifica facilmente as preferências, satisfações e ansiedade dos usuários e são amplamente utilizados para análise estatística. Apesar desses pontos positivos, os questionários estão sujeitos a discrepâncias entre o que está subjetivo na pergunta e na resposta e as reais ações e percepções dos usuários.

2.8.2 Métodos qualitativos e quantitativos em sistemas computacionais

Com o maior número de usuários utilizando sistemas computacionais e a tecnologia cada vez mais presente no cotidiano das pessoas, surgiu a necessidade de pesquisadores da área de computação estudarem aspectos não mensuráveis associados aos usuários, como os seus hábitos, comportamentos, emoções, valores, atitudes, o contraste entre o comportamento estável de sistemas e a imprevisibilidade dos usuários que o utilizam, entre outros aspectos (SILVA; SOBRINHO; VALENTIM, 2020; SOUZA *et al.*, 2019; VASCONCELOS; ANDRADE, 2018; FRANÇA; TEDESCO, 2017). De acordo com Leitão e Prates (2017), há uma crescente adoção de métodos de pesquisa qualitativos em computação, sejam aqueles tradicionalmente usados em pesquisas nas ciências humanas e sociais ou em métodos de base qualitativa criados no interior da área. Neste sentido, esta subseção tem como objetivo principal apresentar as principais características dos métodos qualitativos e as suas principais diferenças quando comparado aos métodos quantitativos, conforme é apresentado na Tabela 1.

Tabela 1 – Paradigmas quantitativo-experimental e qualitativo

	Paradigma Quantitativo Experimental	**Paradigma Qualitativo**
Pressuposto sobre os fenômenos em exame	Estabilidade e previsibilidade dos fenômenos Fenômenos abstraídos de seu contexto de ocorrência	Ocorrência não previsível dos fenômenos Fenômenos vinculados ao seu contexto de ocorrência
Tipo de problema/questão	Elaboração e teste de hipóteses a partir de algum conhecimento sobre o problema	Exploração contextualizada por meio de questões abertas, sem hipóteses prévias
Raciocínio	Hipotético-dedutivo	Interpretativo-indutivo
Ação do investigador	Manipulação de variáveis Análise estatística	Exploração de significados Análise de conteúdo/discurso
Postura do investigador	Concepção de neutralidade	Envolvimento interpretativo Análise de impactos éticos
Tipo de resultados	Abrangentes Padrões, generalizações e replicáveis	Em profundidade Framework interpretativo baseado em rede de significados relacionados ao contexto de investigação

Fonte: Leitão e Prates (2017).

Os métodos quantitativos possuem resultados quantificáveis matematicamente, já os métodos qualitativos não possuem resultados quantificáveis, mas sim com relação aos aspectos não mensuráveis ligados aos usuários. De acordo com Leitão e Prates (2017), além de um tipo de método trabalhar com números e o outro não, existem diferenças de paradigma. Em outras palavras, há diferenças de modelos de geração de conhecimento, seja este conhecimento de natureza científica, seja de natureza prática em contextos profissionais. Segundo as autoras,

o paradigma qualitativo pressupõe que os fenômenos a serem examinados são irreplicáveis, complexos, imprevisíveis e sempre relativos a um contexto de ocorrência, sendo impossível elencar e isolar todas as suas variáveis para conhecê-los através do controle experimental. De um lado, no paradigma quantitativo, enfatiza-se a possibilidade de prever o comportamento dos fenômenos na medida em que se identifica, conhece e controla suas variáveis e espera-se, consequentemente, a replicação dos fenômenos estudados.

Com relação à investigação dos resultados nesses tipos de métodos, existem diferentes categorias de ações e raciocínios. Por exemplo, nos métodos quantitativos, a ação é voltada para um modelo matemático, cujo paradigma é quantitativo-experimental e parte do raciocínio hipotético-dedutivo, em que são manipuladas variáveis operacionalmente definidas e da análise estatística (JAPIASSÚ; MARCONDES, 1996). Já os métodos qualitativos possuem um raciocínio possível de ser induzido e interpretado, o qual é realizado a partir da observação e análise das variáveis envolvidas (JAPIASSÚ; MARCONDES, 1996). De acordo com Leitão e Prates (2017), na perspectiva qualitativa, o conhecimento produzido não é um produto replicável, mas um processo de compreensão e interpretação reutilizável.

2.9 Considerações Finais

Neste capítulo foram apresentadas técnicas que constituem todas as etapas envolvidas na representação, visualização e recuperação de informação em coleções de documentos textuais, bem como foram introduzidos métodos e aspectos de avaliação de usabilidade de sistemas, com o propósito de oferecer uma visão geral de conceitos necessários para a compreensão deste trabalho.

CAPÍTULO

3

VISUALIZAÇÃO COMO SUPORTE À ANÁLISE E RECUPERAÇÃO DE INFORMAÇÃO EM COLEÇÕES DE DOCUMENTOS TEXTUAIS

3.1 Considerações Iniciais

Este capítulo possui como objetivo principal apresentar uma revisão da literatura com foco em técnicas de visualização computacional como suporte à análise e organização do conteúdo de coleções de documentos textuais, recuperação de informação de interesse de um usuário nessas coleções e avaliação de interfaces desse tipo de sistemas.

Os trabalhos apresentados foram organizados em duas categorias. Na Seção 3.2 são discutidos trabalhos que envolvem tarefas de **análise e organização de coleções de documentos**, as quais consideram a necessidade do usuário de explorar o conteúdo da coleção. Além disso, são apresentados trabalhos que oferecem métodos interativos de organização dos documentos da coleção em suas respectivas áreas temáticas. Na Seção 3.3 são discutidos trabalhos que abordam métodos para apoiar usuários em tarefas de **recuperação de informação em coleções de documentos** de acordo com as suas necessidades. Em seguida, a Seção 3.4 apresenta detalhes do **sistema TRIVIR**, o qual é objeto de estudo deste trabalho. Por fim, na Seção 3.5, **aspectos de avaliação de interfaces em sistemas de VA** são apresentados, seguindo para as considerações finais do capítulo na Seção 3.6.

3.2 Análise e Organização de Coleções de Documentos

A relação entre os documentos em uma coleção, na qual o conteúdo possui artigos científicos, pode ser baseada em como as citações ocorrem, uma vez que trabalhos correlacionados geralmente são referenciados cuidadosamente durante um processo de escrita de um artigo

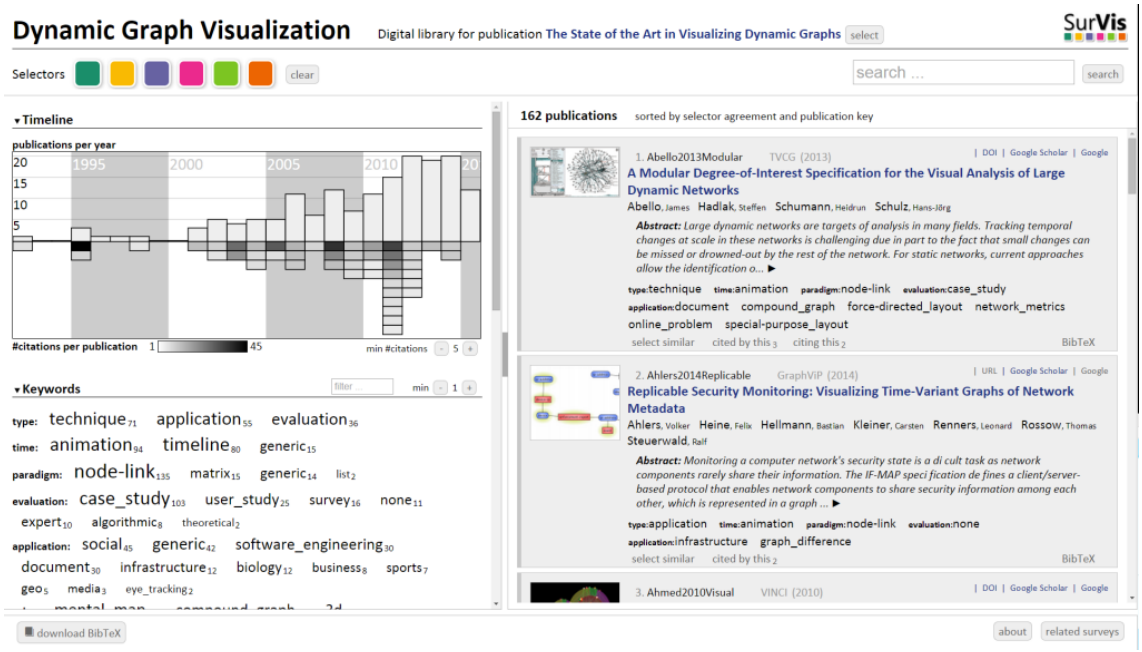
(WALLACH; GONSALVES; ROSS, 2018). Entretanto, ao invés de apenas observar a lista de trabalhos referenciados em cada artigo, é mais interessante explorar as relações do artigo e de suas citações com o apoio de técnicas visuais interativas (PHAM, 2018; PIENTA *et al.*, 2017; ALENCAR; OLIVEIRA; PAULOVICH, 2012), as quais permitem investigar com mais detalhes a coleção de documentos e também a fácil identificação de trabalhos similares potencialmente relevantes para a publicação de um *survey* da área (TOMINSKI *et al.*, 2017; WU *et al.*, 2016).

Vários sistemas de VA foram introduzidos para auxiliar a exploração interativa em coleções de documentos. Alguns se concentram em tarefas específicas, como agrupamento, identificação de tópicos ou recuperação, em que os usuários interagem diretamente com representações visuais para obter *insights*, enquanto dirigem um algoritmo de AM. Um exemplo é o sistema SurVis (BECK; KOCH; WEISKOPF, 2016), um sistema web desenvolvido para ajudar autores a conduzirem uma revisão de literatura por meio da estruturação e análise de seu banco de dados de literatura. Além disso, o sistema oferece suporte aos leitores da pesquisa para obter visões gerais e recuperar publicações específicas usando o conceito de um "seletor", sendo que diferentes seletores permitem filtrar e navegar na coleção. Por exemplo, é possível realizar pesquisa textual, filtrar por palavras-chave ou metadados, agrupar artigos semelhantes ou rastrear *links* de citações. Na Figura 9 é possível observar a interface do sistema SurVis. À esquerda tem-se um histograma, que representa o número de publicações por ano e simultaneamente apresenta abaixo uma sequência vertical de retângulos. A ordem destes retângulos, de cima para baixo, representa a relevância do trabalho, enquanto que tons de cinza mais escuros representam um maior número de citações. Abaixo há uma lista de termos identificados e à direita uma lista de trabalhos para ser explorada pelo usuário, em que também é possível realizar uma pesquisa por similaridade.

Para formular os requisitos do sistema SurVis, os autores estudaram o processo de raciocínio analítico e trabalharam com pesquisadores envolvidos na realização de pesquisas de literatura durante o desenvolvimento do sistema. A validação foi realizada por meio de estudos com usuários. Os autores recrutaram participantes com experiência em VA, IHC e/ou aspectos de usabilidade, e preferencialmente familiarizados com uma versão anterior do sistema. De 37 participantes convidados por e-mail, 14 atenderam ao critério de especialização para participar. As sessões de usuário ocorreram durante uma semana, e após cada sessão o participante respondia a um questionário expressando sua opinião sobre aspectos de usabilidade do sistema e sua conformidade com os requisitos. Embora as respostas às questões tenham sido dadas em escala Likert, o número de participantes é relativamente pequeno para uma análise estatística e a análise focou principalmente nos aspectos qualitativos expressos nas opiniões dos participantes. Os resultados sugerem que o SurVis cumpre os requisitos iniciais. No entanto, embora tenham relatado a necessidade de melhorias gerais, os participantes não identificaram problemas específicos de usabilidade.

Yang, Yao e Qu (2017) apresentam o sistema VisTopic, o qual utiliza modelagem hierár-

Figura 9 – Interface do Sistema SurVis.



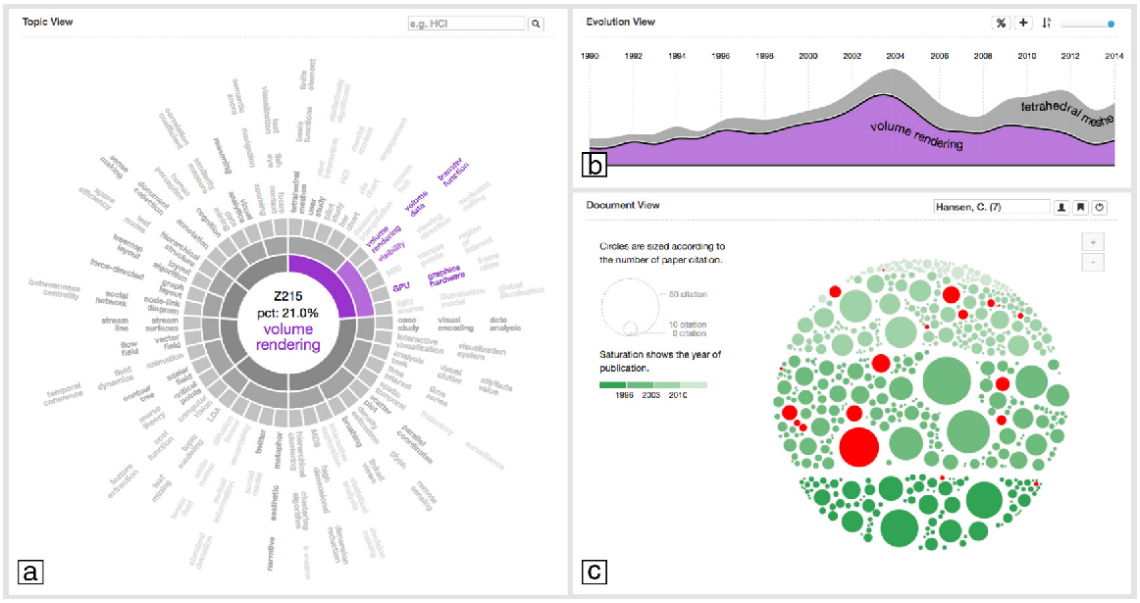
Fonte: Beck, Koch e Weiskopf (2016).

quica de tópicos e técnicas de visualização para apoiar a investigação exploratória de coleções de documentos textuais. A clássica técnica de visualização *Theme River* (HAVRE; HETZLER; NOWELL, 2000) é empregada para transmitir a evolução temporal dos tópicos presentes na coleção, enquanto que o diagrama *Sunburst* (RAMSAY; WAMPLER, 2015) e a *Tag Cloud* (KASER; LEMIRE, 2007) apresentam a hierarquia de tópicos em formato radial, aproximando os tópicos mais representativos da área central, como pode ser observado na Figura 10. Ademais, a técnica *Bubble Chart* (GÖRTLER *et al.*, 2017) também é empregada no VisTopic para apresentar a importância dos documentos a partir do tamanho da bolha, utilizando como referência a quantidade de citações de cada artigo. Com relação à sua coloração, as cores fortes representam o ano de publicação, enquanto as cores fracas representam as publicações antigas e recentes, respectivamente. Ou então, as cores podem ser atribuídas aos agrupamentos obtidos de coleções de dados já rotuladas. Os recursos do sistema VisTopic podem ser observados na Figura 10.

O desenvolvimento do sistema VisTopic seguiu um processo de *design* centrado no usuário, com requisitos formulados com o auxílio de um especialista em Visualização de Informação e um especialista em AM. Para validação, os autores apresentaram um estudo de caso em que empregam o sistema para explorar e analisar o corpus da conferência IEEE VIS, ilustrando como é possível revelar padrões interessantes. Apesar dos esforços no desenvolvimento e validação do sistema, ele não foi avaliado com potenciais usuários.

Agrupamento, ou *Clustering* (JARDINE; RIJSBERGEN, 1971), é uma tarefa frequentemente encontrada em grandes coleções de documentos. Alguns autores argumentam que é improvável que um algoritmo de agrupamento totalmente automatizado produza resultados que

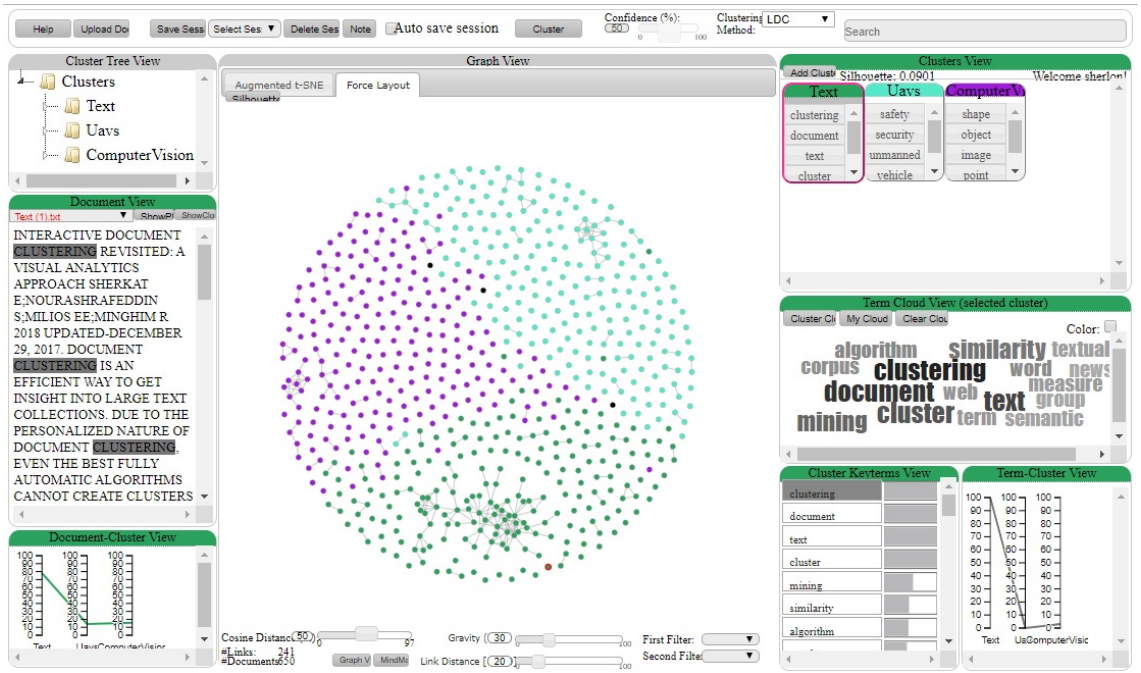
Figura 10 – Interface do Sistema VisTopic.



Fonte: Yang, Yao e Qu (2017).

correspondam exatamente às expectativas do usuário final e o agrupamento seria melhor tratado como um processo orientado pelo usuário (BAE *et al.*, 2020). Essa foi a motivação para o desenvolvimento do sistema de agrupamento de documentos interativo Vis-KT (SHERKAT; MILIOS; MINGHIM, 2019; NOURASHRAFEDDIN *et al.*, 2018; SHERKAT *et al.*, 2018). A interação do usuário e o agrupamento são baseados em termos-chave, que são palavras representativas para revelar os principais temas da coleção. O sistema combina várias visualizações integradas, incluindo um mapa de similaridade de documentos, como ilustrado na Figura 11. Para a geração inicial e interativa dos grupos, os autores implementaram e validam técnicas de agrupamento como o *Lexical Double Clustering* (LDC) e o *Interactive Kmeans* (iKmeans), esta última sendo uma nova abordagem desenvolvida pelos autores, a qual é uma adaptação do algoritmo k-means original aplicado a cenários em que os agrupamentos são interativos. Além disso, a representação visual da coleção é realizada com técnicas de projeção multidimensional como o t-SNE estendido, o qual também foi proposto pelos autores. Esta implementação reúne características do t-SNE tradicional e também propriedades de atração e repulsão do *Force-Layout*, visando assim, reduzir a sobreposição das arestas entre os documentos similares. A visualização da coleção é na forma de uma nuvem de pontos, em que cada documento é representado por um círculo e sua cor identifica o agrupamento (CUTTING *et al.*, 2017). A Figura 11 apresenta uma visão geral do sistema Vis-KT, na qual é possível observar a representação de uma base de dados com 3 grupos já identificados e exibidos com o algoritmo *Force Layout* (NOACK, 2009). Ademais, vale ressaltar que a identificação dos grupos apresentados na Figura 11 foi aprimorada interativamente a partir da utilização de técnicas como *Wordcloud* e *Coordenadas Paralelas* (HEINRICH; WEISKOPF, 2013; INSELBERG; DIMSDALE, 1990; INSELBERG, 1985), as quais enfatizam a frequência e importância dos termos da coleção de documentos.

Figura 11 – Interface do Sistema Vis-KT.



Fonte: Adaptada de Sherkat *et al.* (2018).

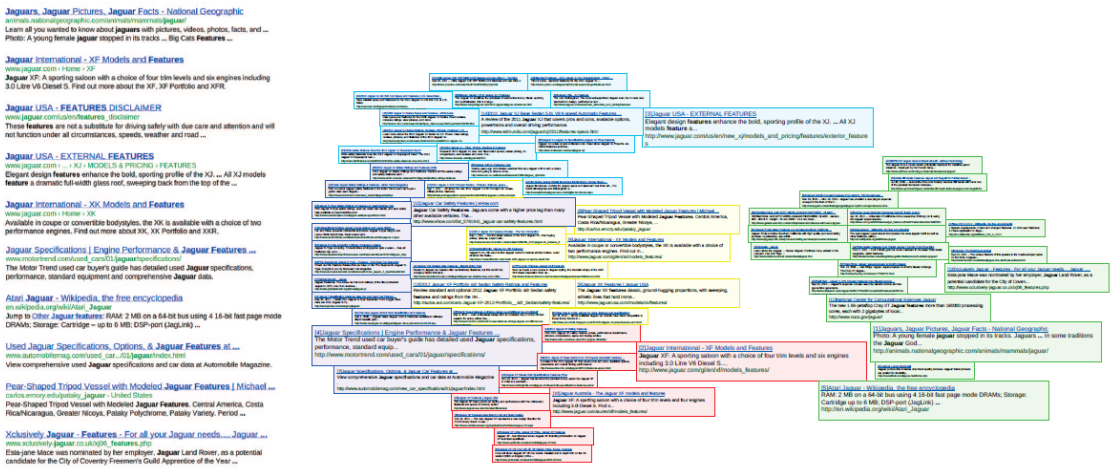
A validação dos aspectos de usabilidade e aprendizagem das visualizações do Vis-KT foi realizada pelos desenvolvedores por meio de estudos com usuários, com dezoito participantes, todos com graduação em Ciência da Computação e algum conhecimento em agrupamentos de documentos. Os participantes realizaram duas tarefas. A primeira envolveu o agrupamento de uma coleção de documentos por tópicos relacionados, em uma sessão de 30 minutos. Os resultados demonstram que a supervisão do usuário melhorou a qualidade dos agrupamentos. A segunda tarefa explorou o impacto das visualizações na obtenção de *insights*. Os participantes foram divididos em dois grupos e solicitados a explorar grupos de documentos em dois modos, um usando a interface de visualização e a alternativa usando um modo *baseline*, sem a visualização. Ambos os grupos responderam a perguntas sobre o corpus, e o grupo que utilizou as visualizações obteve melhores resultados. Os resultados sugerem que o Vis-KT atende os objetivos dos autores.

3.3 Recuperação de Informação em Coleções de Documentos

Em Gomez-Nieto *et al.* (2014) é apresentada uma estratégia denominada ProjSnippet para organização visual dos resultados retornados por uma busca textual, comumente utilizada por ferramentas de busca como o *Google Search* (TRAVIS, 2009; GVILY, 2006). Esta técnica visa aprimorar a interpretação de um resultado de busca, em situações em que a disposição do conteúdo recuperado como uma lista ranqueada é limitada (XIAO *et al.*, 2015). Essa técnica tem por objetivo organizar o conteúdo textual recuperado de uma consulta e apresentá-lo ao

usuário de uma forma visual, a qual codifica espacialmente as relações de similaridade entre os documentos textuais recuperados. O ProjSnippet utiliza a metáfora de preservação de vizinhança de técnicas de projeção multidimensional e também utiliza um algoritmo de agrupamentos para facilitar ao usuário a identificação dos grupos. Por fim, o ProjSnippet otimiza a disposição dos retângulos que denotam os *snippets* evitando a sobreposição de conteúdo, ao mesmo tempo que preserva as vizinhanças e reduz os espaços em branco não ocupados. Uma visão geral da técnica pode ser observada na Figura 12, na qual à esquerda é apresentada a estratégia usual de lista ranqueada, enquanto que à direita o mesmo conteúdo é apresentado utilizando a visualização criada pelo ProjSnippet.

Figura 12 – Estratégia ProjSnippet.

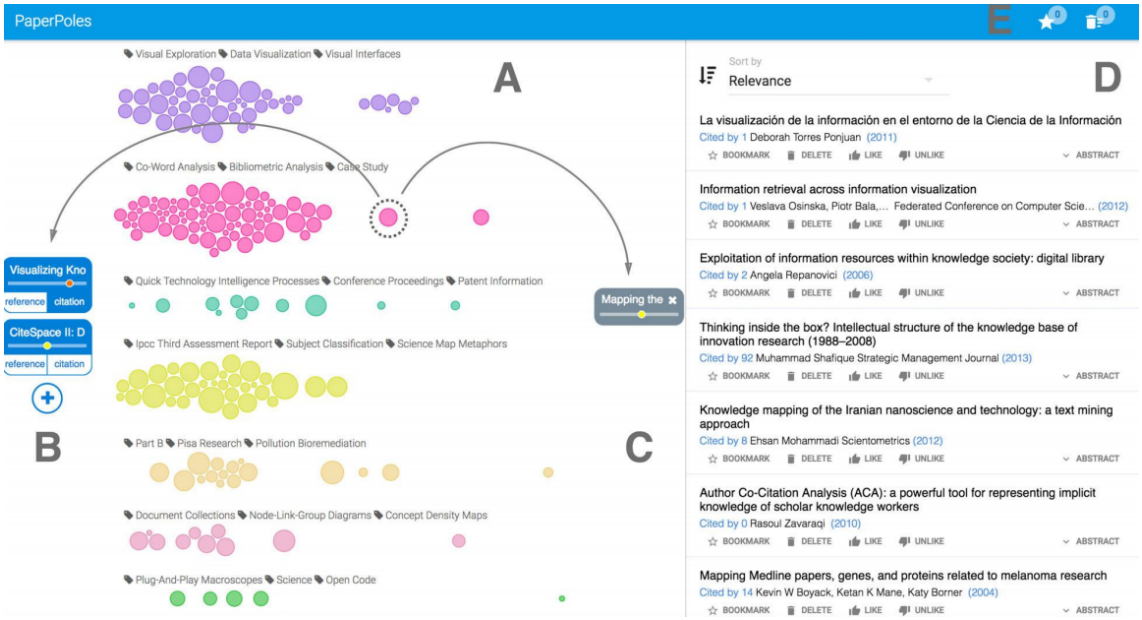


Fonte: Gomez-Nieto *et al.* (2014).

He *et al.* (2019) apresentam o sistema Paperpoles, cujo objetivo é auxiliar usuários a recuperar informação de coleções de artigos científicos, com foco em apoiar a revisão da literatura. Ou seja, dado um corpus de artigos científicos, o usuário deseja encontrar aqueles trabalhos que, além de estarem relacionados à sua pesquisa, são também relevantes. O sistema utiliza uma representação encadeada de citações entre os documentos, a qual permite inferir que aqueles documentos com alto índice de citações entre si estão correlacionados. Portanto, o usuário pode definir documentos exemplo os quais já são relevantes para sua busca. Desta forma, a recomendação do sistema será baseada nas citações referenciadas por tais documentos exemplo. É possível observar na Figura 13 que na visualização do sistema PaperPoles são criados grupos visuais mostrados verticalmente de forma isolada, evitando assim a desordem causada por uma grande variedade de documentos na coleção, permitindo ao usuário explorar cada grupo separadamente.

Os documentos são representados por uma visualização do tipo *Bubble-Based*, em que o tamanho do círculo representa o número de citações. Além disso, os círculos são dispostos por um algoritmo *Force-Based*, no qual os documentos mais relevantes no grupo serão atraídos à esquerda, enquanto que os menos relevantes são repelidos para a direita. Com estas estratégias

Figura 13 – Interface do Sistema PaperPoles.



Fonte: He *et al.* (2019).

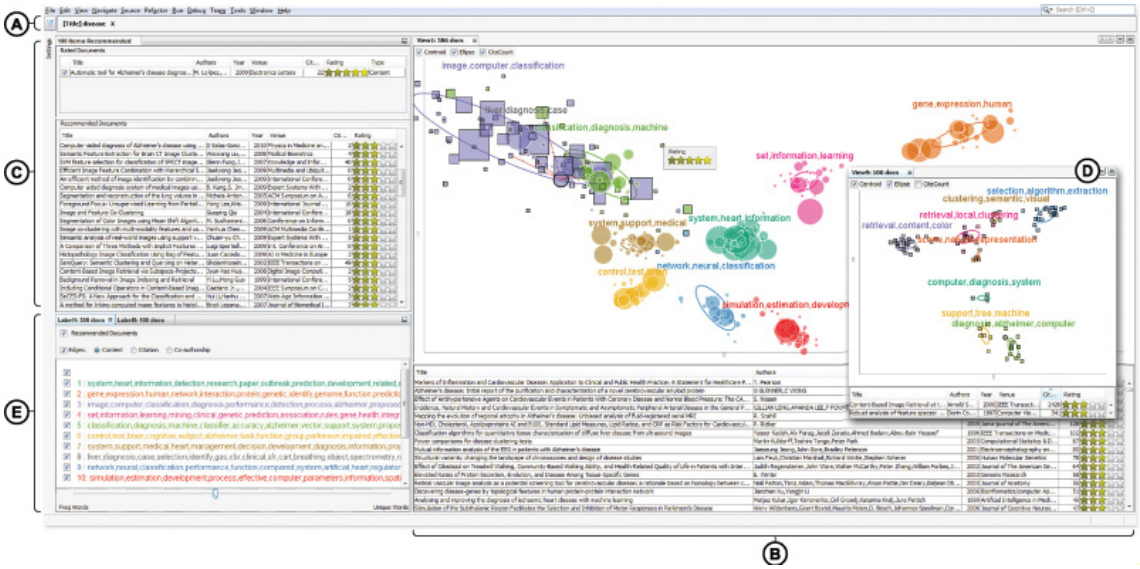
visuais, o sistema permite ao usuário explorar a coleção começando por trabalhos mais citados, ou então por trabalhos mais relevantes. Outra funcionalidade importante do sistema é a utilização de *feedback* positivo e negativo oferecido pelo usuário, o qual pode inserir documentos exemplo que considera relevantes ou não relevantes. Os usuários também podem atribuir um peso para a relevância dos documentos exemplo. As funcionalidades do Paperpoles permitem ao usuário visualizar uma organização dos documentos com os agrupamentos e identificar os trabalhos relevantes e não relevantes baseando-se em suas citações e no *feedback* do usuário durante o processo.

Reconhecendo que as necessidades de informação tendem a mudar conforme o pesquisador avança em sua investigação da literatura, o sistema PaperPoles implementa um ciclo de exploração visual adaptativo que permite reformular gradualmente uma pesquisa conforme a exploração prossegue. As necessidades de informação do usuário são expressas por meio de consultas positivas e negativas e os resultados da pesquisa são agrupados e exibidos em uma exibição de agrupamentos. Os autores realizaram uma avaliação empírica comparando o PaperPoles com uma interface baseada em lista como *baseline*. Eles recrutaram 28 participantes, que foram divididos em dois grupos equilibrados e designados para o Paperpoles ou o *baseline* para realizar duas tarefas de pesquisa acadêmica de complexidades distintas. Em ambas as tarefas, os usuários deveriam identificar 10 artigos que satisfizessem um determinado critério; por exemplo, na tarefa 2 (mais complexa do que a tarefa 1) os participantes deveriam encontrar artigos relatando o uso de técnicas de visualização para mapeamento científico. Métricas foram coletadas, como tempo de conclusão; precisão na identificação de informações relevantes; e eficácia, medida como um *trade-off* entre recuperação e exploração. Os resultados indicam que o sistema melhorou a

precisão da pesquisa em ambas as tarefas, reduziu o tempo de conclusão da tarefa de pesquisa mais complexa e também melhorou a eficácia da tarefa complexa. Uma análise subjetiva indica que executar a tarefa mais simples com o sistema *baseline* foi significativamente mais fácil, mas a melhoria de desempenho na tarefa complexa não foi significativa.

O sistema VisIRR (CHOO *et al.*, 2018) visa auxiliar o usuário na identificação de documentos relevantes em uma base de dados de publicações com mais de 400 mil artigos científicos. Para isto, utiliza informações como o conteúdo e o número de citações para identificar documentos potencialmente relevantes, bem como modelagem por tópicos para recuperar os termos mais importantes da coleção e de determinados grupos, e algoritmos de agrupamento para identificar documentos similares. A visualização baseia-se em uma nuvem de pontos, desta forma, os documentos são representados como círculos e quadrados projetados em um espaço bidimensional. Uma visão geral do sistema VisIRR é apresentada na Figura 14. Inicialmente, o sistema projeta e exibe na forma de círculos parte dos documentos recuperados pela consulta do usuário, sendo que o tamanho representa o número de citações. Posteriormente, por meio de um processo interativo com a nuvem de pontos, o usuário pode explorar informações de cada documento recuperado, bem como observar a lista de documentos recomendados pelo sistema.

Figura 14 – Interface do Sistema VisIRR.



Fonte: Choo *et al.* (2018).

Por fim, é possível identificar os tópicos relacionados a um determinado grupo, visando explorar tais termos e identificar se tal grupo está relacionado à busca do usuário. Além disso, o sistema permite ao usuário realizar mais consultas e compará-las simultaneamente. A recomendação leva em consideração um mecanismo de *feedback*, em que o usuário classifica os documentos recuperados entre 1 e 5 estrelas, no qual quanto mais estrelas, maior a relevância do documento. Ademais, além de o sistema permitir uma recomendação de documentos baseada em trabalhos similares, é possível identificar os trabalhos mais citados.

Seguindo no contexto de sistemas de visualização e recuperação, o TRIVIR suporta a recuperação de documentos assistida pelo usuário com *high recall* (DIAS; MILIOS; OLIVEIRA, 2019). O TRIVIR foi utilizado como ferramenta de estudo deste mestrado, portanto, será explanado com mais detalhes na próxima seção.

3.4 O Sistema TRIVIR

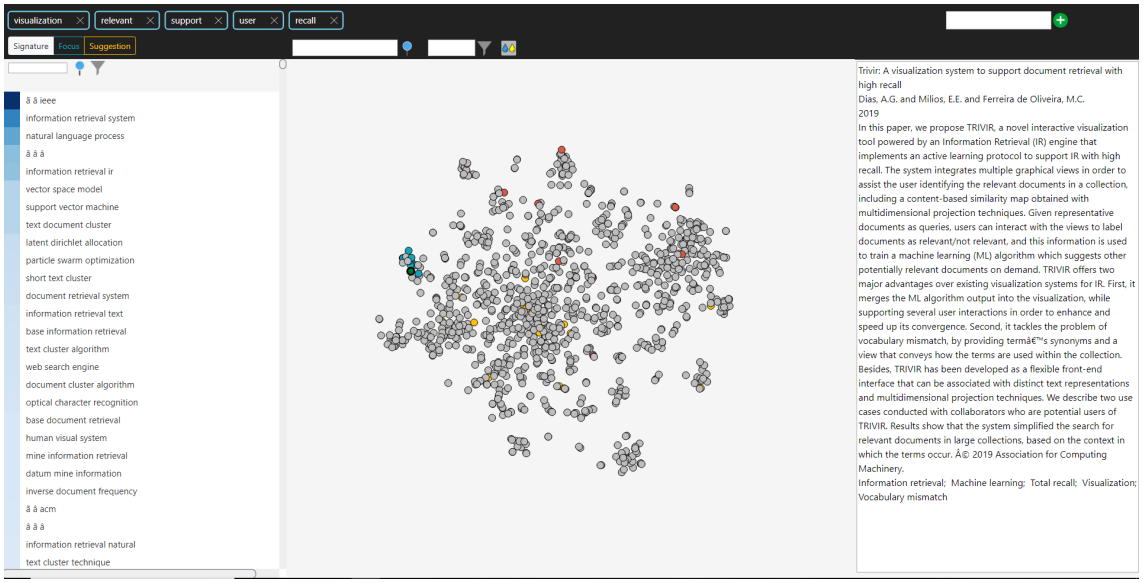
O sistema TRIVIR oferece suporte a tarefas de recuperação de informação relevante em coleções de documentos textuais, com *high recall*. Este sistema é adequado em cenários nos quais o usuário não é capaz de definir claramente uma consulta representativa da sua necessidade de informação. Assim, é permitido ao usuário realizar sua busca a partir de documentos exemplos, os quais já são considerados relevantes para a sua pesquisa e/ou sejam representativos do que está buscando. A partir deste mecanismo de busca, o sistema pode extrair informações dos documentos informados como entrada e aprender a sugerir novos documentos como potencialmente relevantes.

O TRIVIR combina várias visualizações integradas, incluindo um mapa de similaridade de documentos semelhante ao empregado no sistema Vis-KT e também implementa um protocolo CAL de aprendizagem (GROSSMAN; CORMACK, 2016a; CORMACK; GROSSMAN, 2016; CORMACK; GROSSMAN, 2015a; JI; KRISHNAPURAM; CARIN, 2006) com *feedback* do usuário para recuperação de informação (DIAS; MILIOS; OLIVEIRA, 2019). A interface do sistema TRIVIR pode ser vista na Figura 15. A visualização do sistema apresenta uma nuvem de pontos na qual os documentos, representados como círculos, são posicionados no espaço bidimensional em coordenadas determinadas por técnicas de projeção multidimensional, no caso a t-SNE ou a LSP. As cores dos pontos refletem a recomendação do sistema e também o *feedback* do usuário, indicando os artigos relevantes (Azul), não relevantes (Vermelho), sugeridos como relevantes pelo sistema (Amarelo), documentos *Seed* – ou semente – de consulta (Verde) e os documentos ainda não classificados (Cinza).

O sistema TRIVIR também oferece outras funcionalidades, como filtros que permitem ao usuário observar a coleção de documentos por outra perspectiva. Por exemplo, há uma opção para visualizar trigramas representativos da coleção, com o intuito de capturar o contexto de ocorrência das palavras no texto. Há também filtros por palavras-chave e por número de vizinhos mais próximos (KNN) (SHAHABI *et al.*, 2020; PETERSON, 2009), objetivando simplificar a visualização ao reduzir o número de pontos projetados.

Sua arquitetura conceitual é descrita na Figura 16. Uma vez que uma coleção de documentos é carregada e um documento informado como representativo do que o usuário está procurando (o documento de "consulta"), o sistema exibe múltiplas visualizações integradas que permitem explorar a coleção. O usuário pode interagir para explorar o conteúdo dos documentos e seus relacionamentos, a fim de identificar informações relevantes. Ele pode fornecer *feedback*

Figura 15 – Interface do Sistema TRIVIR.



Fonte: Adaptada de Dias, Milios e Oliveira (2019).

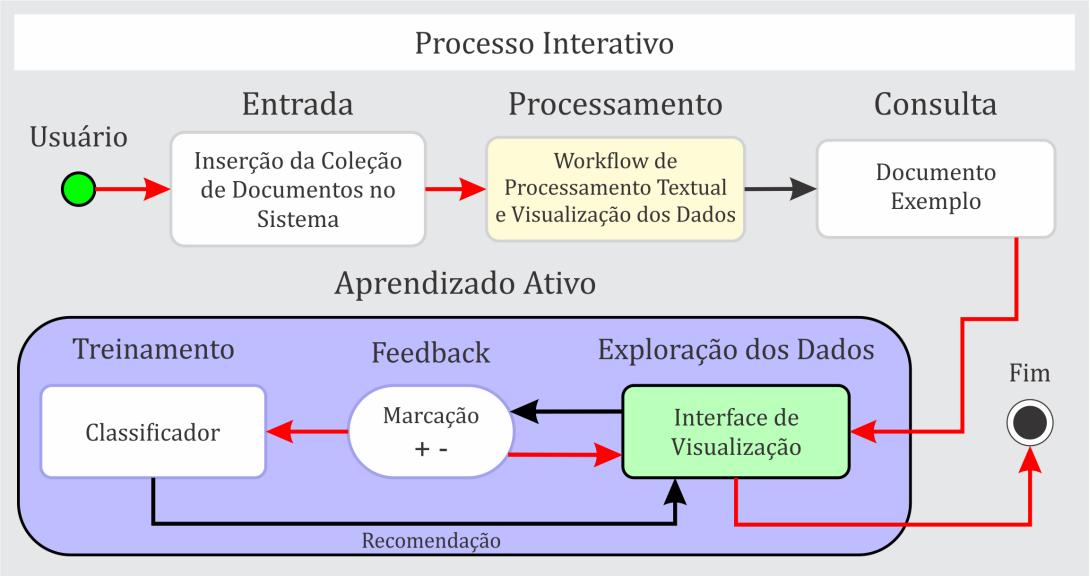
rotulando documentos como relevantes ou não relevantes, o que será considerado para treinar um classificador para sugerir documentos adicionais como potencialmente relevantes. Depois que o classificador é treinado, as visualizações são atualizadas, permitindo que novas explorações se iniciem. Este ciclo continua até que o usuário esteja satisfeito ou que todos os documentos tenham sido rotulados. Os documentos são classificados como relevantes ou não relevantes com a orientação do usuário, cujas ações na interface (*front-end*) são ilustradas na figura por setas vermelhas, enquanto as setas pretas indicam processos realizados pelo sistema no *back-end*.

A Figura 17 ilustra as principais etapas do fluxo de processamento do TRIVIR, as quais contemplam: 1) Entrada de conteúdo e pré-processamento de texto; 2) Criação da representação interna dos documentos; e 3) Computação do *layout* do mapa de similaridade e suas visualizações complementares. Embora ilustre as técnicas utilizadas no sistema TRIVIR original, em vermelho este *workflow* indica as novas técnicas em sua versão 2.0, desenvolvida neste trabalho e posteriormente apresentada no Capítulo 4 - Subseção 4.2.2.

O pré-processamento consiste nas operações típicas exigidas antes de obter uma representação de texto, ou seja, converter em letras minúsculas, tokenização, remover pontuação, caracteres especiais, *stop-words* e lematização. O sistema admite duas configurações alternativas para a representação textual: pode-se criar uma representação VSM com o esquema de ponderação TF-IDF (JONES, 1972), ou alternativamente, pode obter *embeddings* de texto utilizando a biblioteca FastText [1](BOJANOWSKI *et al.*, 2017; JOULIN *et al.*, 2016a; JOULIN *et al.*, 2016b). Os documentos são comparados usando a similaridade do cosseno. Os espaços de características representados por estes *embeddings* e a similaridade do cosseno são usados no classificador (também FastText) e para calcular o mapa de similaridade representado na visualização do *Scat-*

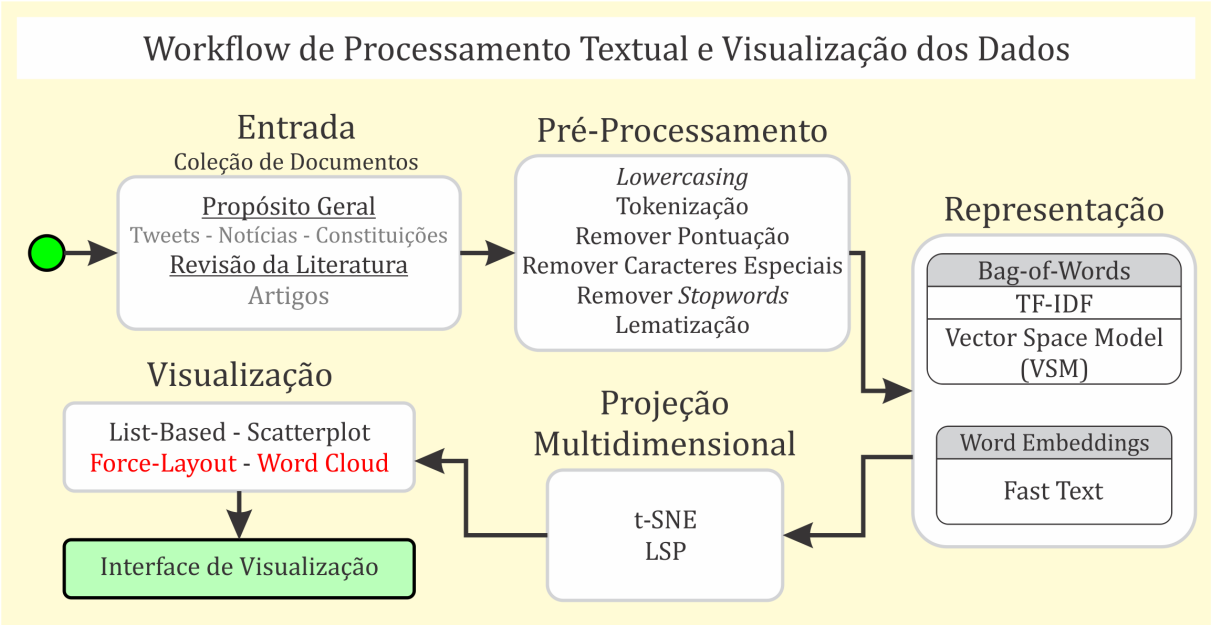
1 <https://fasttext.cc/>

Figura 16 – Protocolo de Aprendizagem Ativa (CAL) para recuperação de documentos no sistema TRIVIR. Os documentos são classificados como relevantes/não relevantes com a orientação do usuário. As setas vermelhas indicam ações do usuário na interface (*front end*), enquanto as setas pretas indicam processos realizados pelo sistema (*back end*).



Fonte: Elaborada pelo autor.

Figura 17 – Workflow de pré-processamento textual e visualização no sistema TRIVIR.



Fonte: Elaborada pelo autor.

terplot View. O documento de consulta fornece as informações iniciais para o classificador. Os 10 documentos mais semelhantes a ele são identificados, e esses documentos são rotulados como relevantes, e representados como círculos azuis na visualização do *Scatterplot View*; enquanto os 10 documentos menos semelhantes são rotulados como não relevantes, e exibidos como círculos vermelhos. Esses documentos formam o conjunto de treinamento inicial definido para o classificador FastText, que irá sugerir 20 documentos adicionais como potencialmente relevantes, os quais serão exibidos como círculos amarelos no mapa.

Um usuário pode, então, interagir com a visualização do *Scatterplot View* e suas visualizações de suporte. Com base em suas descobertas enquanto explora o corpus, o usuário pode (re) rotular quaisquer documentos como relevantes (Azul), não relevantes (Vermelho) ou indicar documentos de consulta adicionais (*Seeds* em Verde). Sempre que o usuário identifica uma *Seed* adicional indicando que encontrou outro documento representativo, os 10 documentos mais semelhantes a este são automaticamente definidos como relevantes. Sempre que o usuário rotula um documento como não relevante, ele pode escolher definir os 10 documentos mais semelhantes também como não relevantes. O classificador pode ser treinado novamente a qualquer momento com as informações atualizadas para aprender gradualmente com o *feedback* do usuário e recomendar novos documentos possivelmente relevantes para investigação posterior. A expectativa é que cada nova recomendação seja mais precisa do que a anterior, e este processo iterativo de *feedback* e classificação continua até que o usuário opte por interromper a exploração. Mais detalhes sobre a arquitetura do TRIVIR, funcionalidades e opções de implementação podem ser encontrados nos trabalhos de Dias, Milios e Oliveira (2019) e Dias (2019).

Uma validação do sistema foi realizada com três voluntários considerando cenários de revisão da literatura. Dois professores de Física e um aluno de Mestrado em Ciência da Computação utilizaram o sistema, com auxílio, para buscar artigos relevantes a fim de apoiar um esforço de levantamento bibliográfico. Uma sessão preliminar com um dos professores, durante o desenvolvimento do sistema, contribuiu para o ajuste fino da interface visual e de seus requisitos funcionais. O *feedback* qualitativo foi obtido nas sessões de validação, quando os colaboradores foram incentivados a relatar seus *insights* e opiniões. O *feedback* foi altamente positivo, pois os três pesquisadores reconheceram que identificaram publicações relevantes que possivelmente teriam perdido de outra forma.

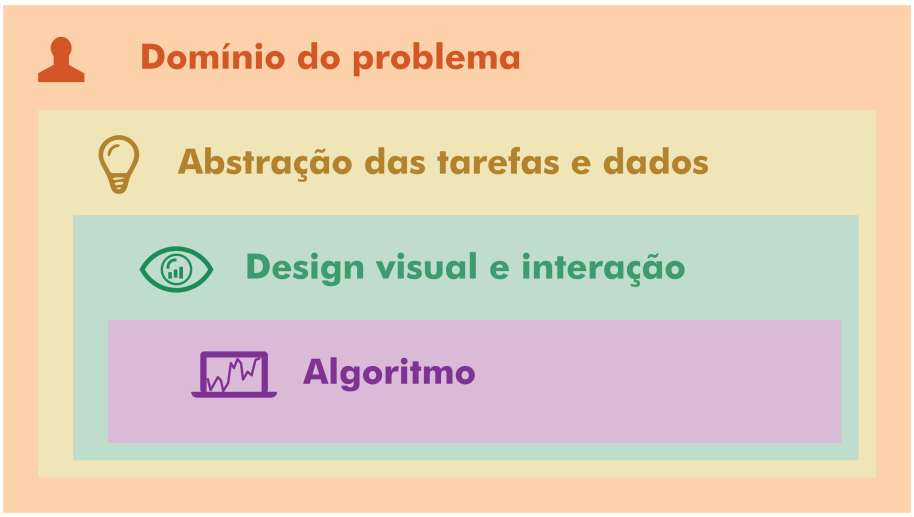
3.5 Avaliação de Interfaces e Visualização

Os estudos empíricos são importantes na avaliação de técnicas e sistemas de visualização, sendo que existe uma diversidade de práticas e cenários a serem considerados (LAM *et al.*, 2011; ISENBERG *et al.*, 2013). Munzner (2009) apresenta um modelo composto por quatro camadas aninhadas para o *design* e validação de visualizações. A validação busca garantir que um sistema satisfaça seus requisitos definidos, enquanto a avaliação diz respeito à capacidade do sistema

apoiar a execução das tarefas e à experiência do usuário.

No modelo de Munzner apresentado na Figura 18, a camada mais externa consiste em uma caracterização das tarefas e dados no domínio do problema; uma segunda camada interna define uma abstração das tarefas e dados do domínio em operações e tipos de dados; uma terceira camada corresponde ao *design* da codificação visual e técnicas de interação; e a camada mais interna consiste nos algoritmos criados para executar as técnicas com eficiência. O aninhamento se justifica porque o resultado de cada camada afeta a anterior, o que contribui para tornar a tarefa de validação tão desafiadora. Munzner (2009) ressalta que a validação deve ocorrer em cada camada, exigindo abordagens metodológicas distintas: validar algoritmos é diferente de validar a tarefa de análise do domínio, ou a caracterização das tarefas e dados, ou o *design* da codificação visual e das técnicas de interação. O esforço deste trabalho pode ser descrito como uma tentativa direcionada à terceira camada do modelo de Munzner, buscando validar o *design* das técnicas de codificação/interação visual, i.e., verificar se eles são eficazes para o usuário executar a tarefa.

Figura 18 – As quatro camadas de *design* do *Nested Model*.



Fonte: Adaptada de Munzner (2014).

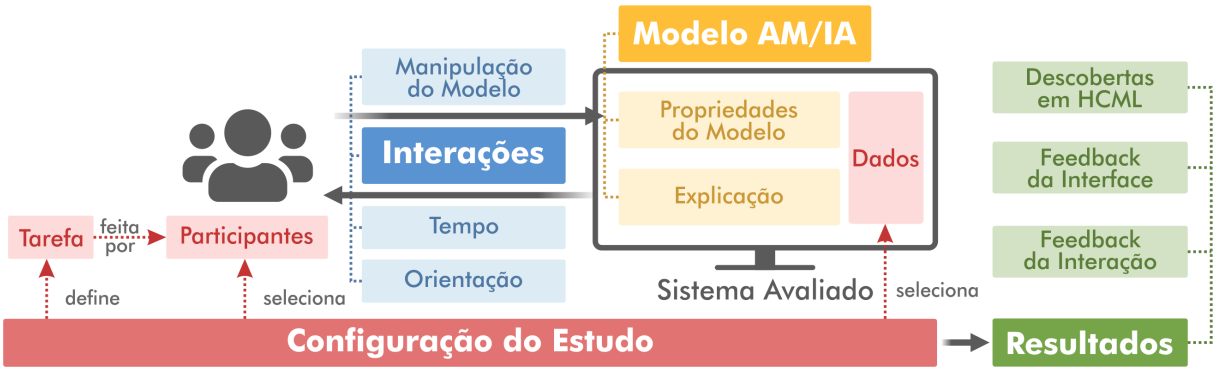
Cada camada do modelo de Munzner apresenta desafios a serem considerados, o que pode dificultar a validação da efetividade de um sistema. Por exemplo, é possível que as necessidades do usuário não tenham ficado claras na análise do domínio do problema, a apresentação das informações a serem investigadas não seja a melhor, o *design* visual não seja representativo e de fácil interpretação para a tarefa alvo, ou até mesmo que os algoritmos implementados sejam lentos. Embora este trabalho tenha focado principalmente na terceira camada deste modelo, os desafios ressaltados por Munzner em todas as camadas foram considerados a fim de identificar e mitigar problemas associados.

Avaliar e validar sistemas interativos traz consigo questões decorrentes da participação de usuários no processo, os quais são envolvidos ativamente desde as etapas iniciais de prototipação

até a validação. Modelos interativos de aprendizado de máquina são definidos por Amershi *et al.* (2014) como uma forma de incluir o usuário final no processo, para obtenção de *feedback* e agilizar o desenvolvimento do modelo, facilitando a criação de sistemas de AM adequados aos propósitos e necessidades dos usuários. Em *visual analytics*, Endert *et al.* (2014) afirma que a inclusão do usuário no processo iterativo, conhecido como "*human in the loop*", permite transmitir a perspectiva do usuário como informação útil ao refinamento do modelo explorado, e ressalta a importância de um processo iterativo centrado ao usuário. Sperrle *et al.* (2021) define *Human-Centered Machine Learning* (HCML) como o campo de pesquisa que considera humanos e máquinas como atores igualmente importantes no *design*, treinamento e avaliação de cenários de AM co-adaptativos. O sistema TRIVIR implementa esse conceito: os usuários direcionam um processo de classificação dos documentos quanto à relevância, oferecendo documentos representativos de sua necessidade de informação e *feedback* sobre as sugestões do classificador.

Endert *et al.* (2014) também afirma que a interação é o fator que integra a análise, a visualização e o usuário, o que levanta perguntas importantes: como esta interação deve ser definida e como a interface deve ser validada? Sperrle *et al.* (2021) ressalta que essa é uma tarefa complexa, uma vez que diversos fatores precisam ser considerados. Os aspectos de avaliação centrados no usuário apresentados pelos autores estão sumarizados na Figura 19.

Figura 19 – Os quatro principais aspectos (configuração do estudo, modelos de AM/IA, interações, e resultados) de avaliações centradas ao humano em HCML e dimensões relacionadas.



Fonte: Adaptada de Sperrle *et al.* (2021).

Esses aspectos capturam propriedades da configuração do estudo, dos participantes e da análise das tarefas e tipos de dados utilizados no estudo, portanto são fundamentais para a validação dos sistemas. Os quatro principais aspectos de avaliações são: A) configuração do estudo; B) modelos de AM/Inteligência Artificial (IA); C) interações; e D) resultados. Cada um, por sua vez, apresenta sub tópicos importantes, destacados a seguir. Configurações para cada aspecto a ser avaliado podem ser observadas no trabalho de Sperrle *et al.* (2021).

1. Avaliação da Configuração

   a) **Configuração do Estudo:** Tipo de Estudo, Processamento dos Resultados, Fase de Aprendizagem, Tempo Necessário;

b) **Participantes:** Experiência no domínio e conjunto de dados, Experiência em AM/IA, Idade, Gênero, Número de participantes;

c) **Tarefas e Dados:** Análise das Tarefas, Tipos de Dados.

2. Propriedades do Modelo e Explicações

a) **Propriedades do Modelo:** Qualidade, Qualidade Percebida, Transparência, Interpretabilidade, Confiabilidade, Controlabilidade;

b) **Explicações:** Transparência, Confiabilidade, Efetividade, Fidelidade.

3. Interações e Orientação

a) **Manipulação do Modelo:** Objetividade, Tipo de Interação, Impacto;

b) **Cronometragem:** Fase, Frequência;

c) **Orientação:** Lacuna de conhecimento, Grau, Adaptação.

4. Resultados

a) **Resultados:** Principais descobertas em HCML, *Feedback* da Interface, *Feedback* da Interação.

Embora suas vantagens potenciais para apoiar tarefas de exploração e recuperação em coleções de documentos sejam frequentemente anunciadas, há um número reduzido de relatos de avaliações de sistemas VA em configurações realistas. As tarefas de análise e recuperação de texto exigem grande esforço cognitivo, e os usuários de soluções que utilizam VA normalmente enfrentam uma curva de aprendizado acentuada. Realizar avaliações com usuários reais é desafiador, tendo em vista a complexidade e diversidade de tarefas que demandam tempo e esforço, além da disponibilidade restrita de especialistas. Frequentemente, as tarefas propostas em estudos de usuários são muito mais simples do que as tarefas-alvo reais, e uma análise da usabilidade do sistema como um todo tende a ser negligenciada no esforço de desenvolvimento.

3.6 Considerações Finais

Os trabalhos apresentados ao longo do capítulo foram categorizados em **análise e organização de coleções de documentos** e **recuperação de informação em coleções de documentos**. Enquanto a primeira apresenta trabalhos que lidam com a necessidade do usuário em explorar o conteúdo da coleção de documentos, a segunda reúne trabalhos que lidam com a necessidade de recuperação de informação relevante em coleções de documentos. O estado da arte apresenta sistemas interativos com suporte de técnicas de visualização computacional visando inserir o usuário neste processo de exploração e busca em uma coleção de documentos.

No Apêndice A são listadas as técnicas identificadas no estado da arte, bem como mapeadas em um *workflow* de sistemas de RI e VA interativos.

A interação do usuário diretamente com a coleção de documentos permite ao mesmo obter *insights* enquanto auxilia o sistema a identificar resultados que refletem a sua perspectiva. Embora sistemas e técnicas apresentados nas duas categorias supracitadas possam ser vistos separadamente como ferramentas de suporte ao usuário em tarefas distintas, também é possível unificar os dois propósitos em um único sistema que permitirá ao usuário analisar, organizar e recuperar documentos, tornando o processo de exploração de coleções e identificação de trabalhos relevantes mais robusto.

CAPÍTULO

4

AVALIAÇÃO E VALIDAÇÃO DO SISTEMA TRIVIR

Diversos cenários demandam recuperação de informação em coleções de documentos textuais, e comum a todos é um usuário interessado em identificar informações relevantes em um conjunto potencialmente grande de documentos. A tarefa pode ser desafiadora, principalmente em situações em que o usuário tem dificuldade de expressar de maneira concisa e objetiva a sua necessidade de informação e que seja necessário uma recuperação com *high recall*. Estratégias devem ser pensadas para maximizar a recuperação de documentos relevantes ao mesmo tempo em que tentam garantir ao usuário que a sua recuperação de trabalhos relacionados é abrangente.

Dias (2019) introduziu o sistema TRIVIR para auxiliar usuários na recuperação de documentos relevantes com *high recall*, integrando recursos de VA com um classificador de documentos quanto à sua relevância para o usuário. Buscando avançar na compreensão da adequação de técnicas de VA existentes às necessidades e objetivos dos seus potenciais usuários, este trabalho teve por objetivo fazer uma avaliação do sistema, identificando limitações conceituais e práticas associadas ao seu uso. Para isto, foram conduzidos estudos com usuários para obtenção de informações quantitativas e qualitativas a partir de uma tarefa de recuperação de documentos, mensurando aspectos da usabilidade e utilidade do sistema TRIVIR.

Antes da realização dos estudos, foram identificados dois pontos acerca da condução das sessões com cada usuário. O primeiro diz respeito à apresentação do sistema e de suas funcionalidades. O segundo ponto (B) diz respeito a como registrar e mensurar a experiência de uso, como as informações e contribuições do estudo seriam quantificadas. Inicialmente, foi definido que a introdução ao sistema seria conduzida a partir de uma apresentação de *slides*, na qual constariam os objetivos do sistema e exemplos de como o usuário poderia interagir com as funcionalidades disponíveis. Após esta apresentação, a qual está disponível no GitHub [1], na pasta "*Start Here - Presentation*", seria iniciada a interação do usuário com o sistema, sendo este

[1] <https://github.com/SherlonAlmeida/TrivirV2.0>

orientado a realizar anotações ao longo da sessão, sobre a sua experiência e dificuldades. Ao terminar a sessão, o usuário seria convidado a responder a um questionário, o qual foi elaborado para tentar capturar quantitativamente a experiência de uso do sistema durante a sessão.

Foram conduzidas três etapas de levantamento de informações sobre a perspectiva de potenciais usuários do sistema: a) um Estudo Piloto; b) um primeiro Estudo de Investigação (S1); e c) um segundo Estudo de Validação (S2). A mesma estratégia foi adotada nos três estudos, com eventuais ajustes no material de apoio. O **a) Estudo Piloto**, do qual participaram 5 voluntários, foi realizado como uma investigação informal para verificação dos procedimentos e ajuste dos materiais de apoio. Posteriormente, o **b) Estudo de Investigação (S1)** teve por objetivo conduzir um processo sistemático de coleta de informações, visando observar como os usuários interagem com o sistema e possíveis gargalos. Tanto durante a condução do estudo **Piloto** como do estudo de **Investigação (S1)** foram identificadas limitações do sistema TRIVIR, o que levou a reflexões sobre maneiras de melhorar a interface. Por fim, uma vez identificadas limitações e implementadas as melhorias propostas, o **c) Estudo de Validação (S2)** foi realizado para obter novamente a perspectiva de usuários sobre questões de usabilidade, com o intuito de validar as soluções apresentadas neste trabalho.

Este capítulo apresenta a metodologia adotada para fins de avaliação do sistema em cenários práticos, em um contexto de revisão da literatura, e relata e discute os resultados obtidos em estudos com usuários em potencial.

4.1 Estudos com os usuários: metodologia

Os estudos empíricos observacionais foram realizados com alunos de pós-graduação e pesquisadores utilizando o sistema TRIVIR para explorar uma coleção de artigos científicos de seu interesse, para fins de revisão de literatura. De acordo com Nielsen (1994), sempre se deve considerar as dificuldades dos usuários ao mexer no sistema, portanto, nosso objetivo foi compreender as dificuldades envolvidas na utilização de um sistema de VA em uma tarefa complexa e como os usuários percebem e avaliam o suporte fornecido. O planejamento e execução dos estudos com usuários foram apreciados e aprovados pelo Comitê de Ética em Pesquisa com Seres Humanos da Escola de Educação Física e Esporte de Ribeirão Preto, EEFERP-USP.

Os participantes do estudo se voluntariaram após serem convidados por meio de contato pessoal ou e-mail. Convidamos alunos de pós-graduação e pesquisadores que trabalham em laboratórios de pesquisa, no ICMC e em outras várias instituições, que se qualificariam como potenciais usuários do TRIVIR. Como um fator motivacional, a participação no estudo foi apresentada como uma oportunidade de usar uma ferramenta para apoiar um processo de revisão de literatura, de modo que cada participante trabalharia em um corpus de sua escolha. Eles poderiam fornecer uma *string* de busca ou então interagir com o pesquisador para criar uma

string de pesquisa representativa, a fim de coletar um corpus de seu interesse.

Inicialmente, foi realizada uma análise preliminar com cinco pessoas convidadas a usar o TRIVIR, após uma breve apresentação de suas principais funcionalidades. Este estudo piloto foi importante para planejar e organizar as sessões de usuário subsequentes e desenvolver os materiais de apoio. Naquele estágio da pesquisa, era necessário obter estimativas realistas da duração da sessão, decidir sobre a melhor forma de introduzir o sistema, definir quais perguntas incluir nos questionários e como formulá-las, etc. Assim, foi preparada uma apresentação detalhada do TRIVIR incluindo uma explicação ilustrada de todas as suas funcionalidades. Também foi elaborado o questionário para coletar as opiniões dos participantes, que incluiu o questionário SUS e a formulação de questões específicas sobre cada funcionalidade do sistema, detalhadas na Seção 4.2. Todos os participantes convidados possuíam habilidades de compreensão da língua inglesa, no entanto os questionários foram desenvolvidos em inglês com as respectivas traduções em português após cada pergunta, a fim de mitigar problemas de interpretação. Neste trabalho são apresentadas apenas as versões em inglês dos questionários.

Uma vez que cada participante trabalharia com sua própria coleção de artigos, a aquisição do corpus para as sessões foi feita individualmente. Cada sessão foi agendada de acordo com a disponibilidade do participante e, antes da sessão, foram tomadas as providências para a obtenção do corpus. Primeiramente, foram solicitados aos participantes que indicassem alguns artigos que consideravam representativos em relação ao seu interesse e sugerissem uma *string* de busca para consultar os repositórios pertinentes, como Web Of Science (WoS), Scopus, IEEE Explore ou outros. Em alguns casos, foram aplicados filtros para garantir que a coleção tivesse um tamanho razoável, em torno de 2.000 artigos, no máximo.

Antes de iniciar uma sessão, o corpus correspondente foi inserido no TRIVIR e definido o documento *seed* inicial (conforme informado pelo participante), a fim de evitar retardar a sessão devido ao tempo de espera decorrente do treinamento inicial do classificador. Foi utilizada a configuração padrão em todas as sessões, que adota VSM / TF-IDF como representação de texto e t-SNE para o mapa de similaridade. Cada sessão foi iniciada com a apresentação do Termo de Consentimento Livre e Esclarecido (TCLE), esclarecendo o objetivo do estudo e os direitos do participante. Seguiu-se a introdução ao sistema e, em seguida, o participante foi incentivado a explorar o corpus. O participante foi convidado a descrever e fazer anotações sobre pontos fortes e fracos percebidos e falar em voz alta durante a interação. Com base no estudo piloto, previa-se que a sessão durasse no máximo três horas, com estimativa de aproximadamente 30 minutos para introdução, 2 horas para exploração dos dados e 30 minutos para resposta ao questionário. De todo modo, os participantes foram encorajados a explorar o tempo que desejassem e enquanto se sentissem confortáveis. As sessões duraram em geral de 2 a 3 horas cada.

A navegação no TRIVIR que foi proposta aos usuários consiste em explorar os documentos do corpus utilizando as técnicas de visualização, interagindo com os filtros disponíveis e demais funcionalidades para identificar os documentos relevantes. Quando o usuário identifica

Tabela 2 – Informações sobre as duas rodadas de estudos observacionais, identificadas como S1 e S2, respectivamente.

ID	#doc S1	#doc S2	Tema do corpus
1	277	350	Safety e Security para veículos aéreos não-tripulados
2	574	622	S1: Recuperação de informação textual interativa S2: Identificação de documentos em grandes corpora para revisão da literatura
3	909	1515	Graph Convolutional Network para predição de crimes
4	554	669	Questionários para mensurar experiência de usuários em ambientes de realidade virtual
5	499	576	Privacidade e Identificação em hospitais usando biometria
6	857	1045	Técnicas de tomada de decisão para veículos autônomos
7	2003	2197	S1: Biossensores para detecção de câncer de pâncreas e COVID-19 S2: Biossensores para detecção de câncer de pâncreas
8	667	2197	Biossensores para detecção de câncer de próstata
9	546	—	Previsão da trajetória de agentes externos em veículos autônomos
10	—	808	Fake News em mídias sociais relacionadas a eleições presidenciais
11	—	1090	Blockchain: comunicação e algoritmos
12	—	430	Recuperação de informação textual interativa

um documento relevante ou não relevante, ele pode rotulá-lo explicitamente como tal. Após algumas rotulações, o usuário pode decidir treinar novamente o classificador do sistema para obter novas recomendações. Em outras palavras, o usuário está envolvido em um processo interativo, no qual fornece informações úteis como *feedbacks* positivos ou negativos para o treinamento do classificador, que poderá fornecer novas recomendações baseadas nesse *feedback*.

Foram realizadas duas rodadas de estudos, identificadas como S1 e S2 (ver Tabela 2), em dois momentos diferentes e usando versões diferentes do sistema TRIVIR, conforme explicado na Seção 4.2. Na primeira rodada (S1) os participantes utilizaram a versão original TRIVIR 1.0 disponível no GitHub [2] (DIAS, 2019). Após coletar e analisar as opiniões, foram identificados alguns problemas de usabilidade e limitações práticas da implementação disponível. Em seguida, foram implementadas algumas modificações no sistema para tratar alguns dos problemas críticos observados ao nível da interface, criando uma nova versão, o TRIVIR 2.0. Empregamos o TRIVIR 2.0 em uma segunda rodada de sessões (S2), a fim de coletar mais opiniões sobre o processo de recuperação assistido com VA. Oito usuários participaram de ambas as rodadas, conforme indicado nas Tabelas 2 e 3; um usuário que participou de S1 não participou de S2 e três voluntários adicionais participaram de S2 que não participaram de S1. Assim, tem-se um total de 12 participantes, sendo que 9 e 11 sessões observacionais foram conduzidas em S1 e S2, respectivamente. O tamanho dos corpora estudados variou de algumas centenas a cerca de 2.000 documentos. Uma vez que a segunda rodada de estudos foi realizada alguns meses após a primeira, os corpora foram atualizados para incluir publicações recentes. As Tabelas 2 e 3 mostram os tópicos explorados por cada participante em sua sessão correspondente, o tamanho do corpus em cada sessão, a formação acadêmica do participante e se ele tem experiência com Visualização ou IHC. Assim que os participantes terminaram a exploração, eles foram convidados a responder ao questionário e expressar seus pensamentos e impressões sobre o sistema.

[2] <https://github.com/amandagdias/TRIVIR>

Tabela 3 – Informações sobre os participantes, os quais foram questionados sobre sua familiaridade com Visualização e IHC.

ID	Formação	Familiaridade
1	Ciência da Computação	Sim
2	Ciência da Computação	Sim
3	Matemática + Engenharia Elétrica	Sim
4	Ciência da Computação	Sim
5	Ciência da Computação	Não
6	Engenharia de Controle e Automação	Não
7	Química	Não
8	Ensino de Ciências e Física	Não
9	Ciência da Computação	Não
10	Publicidade e Propaganda	Não
11	Ciência da Computação	Não
12	Ciência da Computação	Sim

4.2 Resultados

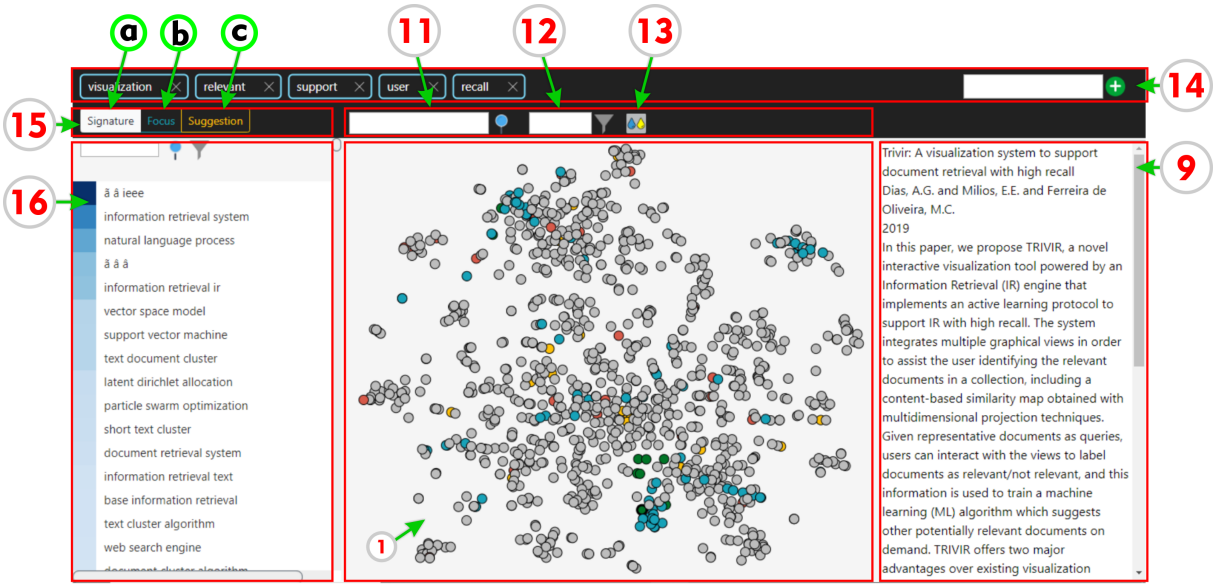
Nesta seção, os resultados dos estudos observacionais são apresentados e discutidos. Na Seção 4.2.1 são analisadas as respostas às questões sobre a utilidade de funcionalidades específicas do sistema obtidas na primeira rodada de estudos (S1). Ao fazer isso, foi possível observar que a aceitação foi alta, mas também foram identificados problemas enfrentados pelos participantes e outras limitações que impactaram o andamento das sessões. Em seguida, algumas modificações foram implementadas, descritas na Seção 4.2.2, com o objetivo de abordar alguns dos problemas mais críticos. Na Seção 4.2.3 são discutidos os resultados da segunda rodada de estudos (S2), conduzida usando a nova versão (TRIVIR 2.0), novamente considerando apenas as questões sobre as funcionalidades específicas. Finalmente, na Seção 4.2.4 são discutidas as respostas relativas à usabilidade percebida do sistema, tanto para a rodada S1 quanto para a S2.

4.2.1 Avaliando o sistema TRIVIR

Os componentes da interface do sistema TRIVIR 1.0 são apresentados na Figura 20. O componente principal é o *Scatterplot View* (1) que descreve um mapa de similaridade dos documentos em um corpus, onde cada círculo representa um documento. As cores dos círculos indicam rótulos de documentos relacionados à relevância para o usuário: Consulta/Semente (verde), Relevante (azul), Não Relevante (vermelho), Sugerido como Relevante (amarelo) ou ainda Não Rotulado (cinza). O *Scatterplot View* e outras visualizações de suporte permitem que um usuário investigue os documentos e seus relacionamentos: filtrar o mapa por uma determinada palavra (11); filtrar o mapa para mostrar os K documentos mais semelhantes à consulta inicial (12), ou para mostrar apenas os documentos potencialmente relevantes (Semente, Relevante ou Sugerido) (13); a *Document View* (9) mostra o conteúdo de um documento selecionado; a *Terms View* (14) exibe termos e sinônimos frequentes; as três visualizações baseadas em lista (15)

compartilham uma área comum (16): a *Signature* (15.a) lista trigramas representativos do corpus; a *Focus* (15.b) lista os documentos interessantes (Semente e Relevantes); e a *Suggestion* (15.c) relaciona os documentos atualmente sugeridos como Relevantes pelo classificador; finalmente há um botão que permite re-treinar o classificador após o *feedback* do usuário, localizado na aba *Suggestion* (15.c). Destaca-se que a numeração das funcionalidades apresentadas na Figura 20 segue uma padronização com a Figura 22, em que a numeração representa as mesmas funcionalidades em ambas versões do sistema. Para obter mais detalhes sobre o TRIVIR 1.0, consulte a Seção 3.4 ou o trabalho de Dias (2019).

Figura 20 – TRIVIR 1.0 - Interface original.



Fonte: Elaborada pelo autor.

A Tabela 4 lista as questões que constituem a Parte 2 do questionário na primeira rodada de estudos observacionais (S1). O questionário cobriu cada funcionalidade do sistema, solicitando uma classificação de sua utilidade como percebida pelo usuário, em uma escala numérica [0-10].

A Figura 21 resume as notas atribuídas às funcionalidades pelos participantes do estudo. Na Figura 21a os *boxplots* representam a dispersão das pontuações em cada questão - a marca representada por um '**x**' em vermelho indica a pontuação média. O mapa de calor na Figura 21b apresenta as avaliações individuais, cujas linhas indicam os participantes e as colunas suas notas correspondentes em cada questão, com tons mais escuros de azul indicando avaliações melhores (mais altas). Os participantes que declararam familiaridade com visualização ou IHC são identificados com um sinal de asterisco '*'. Observa-se que a maioria das funcionalidades recebeu avaliações consideravelmente altas, com algumas exceções. Em particular, as pontuações de cinco questões específicas, nomeadamente Q4, Q7, Q8, Q9 e Q10 foram mais baixas, com maior dispersão.

A Questão Q4 refere-se à funcionalidade de filtrar a exibição do *Scatterplot View* representando o mapa de similaridade do documento para mostrar apenas os k vizinhos mais

Tabela 4 – Questionário Parte 2 (S1) com o objetivo de avaliar a utilidade percebida de cada funcionalidade. Avaliação por meio de notas no intervalo [0-10], do menos útil (0) ao mais útil (10), exceto para a questão 11, em que o usuário identifica em uma lista as funcionalidades que considera que precisam ser melhoradas, na sua opinião.

Questionário Parte 2 (S1)
1) Considering the whole system, how do you rate it?
2) Considering the Scatterplot View (Similarity Map of Documents), how do you rate it?
3) Considering the functionality that allows the user to filter documents based on specific terms, how do you rate it?
4) Considering the filter of K most similar documents to help the user to reduce the search space, how do you rate it?
5) Do you think that reduce the point clutter (Filter by seed, relevant and suggested documents) can be useful to improve the visualization? How do you rate it?
6) Considering the Document View, which helps the user in the reading process, how do you rate it?
7) Considering the Terms View, which helps the user to identify synonyms in the documents, how do you rate it?
8) The Signature List View shows to the user 3-Grams which appears in the same context in the corpus. How do you rate it?
9) The Focus List View shows the current Relevant Documents. How do you rate it?
10) The Suggestion List View shows the documents suggested by the machine learning algorithm. How do you rate it?
11) In your opinion, what functionalities need to be improved?

próximos do documento de consulta (área (12) na Figura 20). Esta funcionalidade recebeu avaliações ruins dos participantes U2 e U4, ambos familiarizados com interfaces de visualização. Eles consideraram uma limitação o fato de que o filtro só poderia ser aplicado em relação ao documento de consulta inicial.

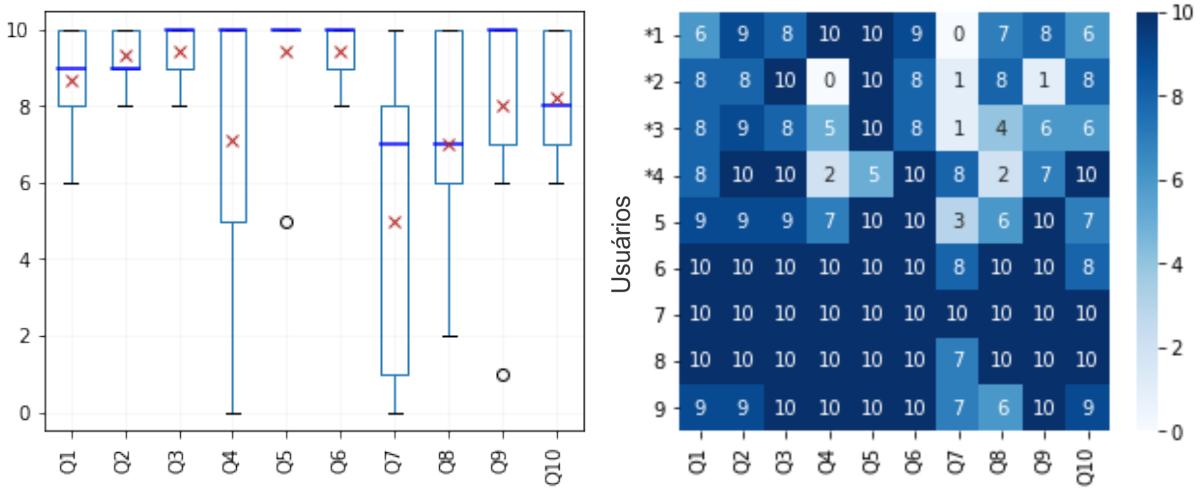
A questão Q7 refere-se ao *Terms View*, identificado como (14) na Figura 20. O *Terms View* mostra cinco palavras consideradas importantes no corpus na área superior. Quando uma palavra é clicada, sinônimos são mostrados, na expectativa de auxiliar o usuário a identificar palavras adicionais que ele possa considerar representativas, e que podem ser adicionadas ao *Terms View*. As palavras incluídas inicialmente são as mais frequentes (com base na ponderação TF-IDF) no documento de consulta inicial e, em combinação com a *Signature List View*, fornecem algumas dicas sobre o conteúdo do corpus. Este recurso foi mal avaliado em geral, e os participantes U1, U2, U3, U5 atribuíram as piores pontuações. Eles não consideraram os termos incluídos, nem os sinônimos sugeridos, representativos do vocabulário do corpus. Na verdade, os sinônimos sugeridos não são obtidos do vocabulário do corpus, mas de uma fonte externa geral (DIAS; MILIOS; OLIVEIRA, 2019). Além disso, em PLN e na recuperação de texto em geral, inferir a importância do termo apenas a partir da sua frequência de ocorrência é conhecido por ser problemático e, muitas vezes, equivocado.

As questões Q8, Q9 e Q10 referem-se às três listas (15a, b, c) que compartilham a área

(16) na interface, ver Figura 20. A *Signature List* (Q8) mostra os trigramas mais frequentes do vocabulário, sugerindo tópicos frequentes e potencialmente importantes. A *Focus List* (Q9) mostra os títulos de documentos atualmente categorizados como Semente ou Relevantes, e a *Suggestion List* (Q10) mostra os títulos de documentos atualmente recomendados pelo classificador como potencialmente relevantes. A categorização do documento em Relevante / Não Relevante também é codificada por cores nos círculos que representam os documentos no *Scatterplot View*. Além disso, como as listas ocupam a mesma área da tela, o usuário precisa alternar entre as três. O fato de as funcionalidades *Suggestion List* e *Focus List* fornecerem informações redundantes, uma vez que a mesma informação está representada por cores no *Scatterplot View*, pode explicar por que alguns participantes, por exemplo, U2, U3, atribuíram baixas pontuações. No entanto, as informações na *Signature List View* não estão codificadas no mapa e também receberam algumas pontuações baixas, por exemplo, de U3, U4 e U5.

Foram identificados dois padrões de comportamento dos usuários na atribuição de pontuações. Há um grupo de 6 usuários críticos, por exemplo, U1 a U5 e U9, enquanto os participantes U6, U7 e U8 atribuíram as melhores pontuações a quase todas as perguntas. Curiosamente, 4 dos 6 participantes críticos declararam experiência anterior com interfaces de visualização e 5 deles possuem graduação em Ciência da Computação, ao contrário dos 3 participantes extremamente positivos nas suas avaliações. Consideramos o nível de satisfação com a maioria dos recursos do sistema por parte de usuários sem familiaridade com aspectos de visualização e IHC um tanto surpreendente, em vista da complexidade do sistema. Isso levanta uma questão sobre se estes participantes realmente compreenderam todas as possibilidades oferecidas pelos recursos do sistema e suas respectivas funções e limitações. Talvez essa mesma complexidade dificulte para eles identificar deficiências ou inconsistências, que são mais prováveis de serem identificadas por participantes mais experientes.

Figura 21 – Boxplot e heatmaps das avaliações no estudo S1.



(a) Distribuição das avaliações.

(b) Avaliações por participante.

Fonte: Elaborada pelo autor.

No geral, observamos uma visão bastante otimista do potencial do TRIVIR para auxiliar na busca por material relevante na literatura, com 8 de 9 participantes declarando sua disposição de usar o sistema com frequência (Q1). No entanto, embora não esteja refletido diretamente nas pontuações da pergunta correspondente (Q2), foram identificados vários problemas com o *Scatterplot View* que apresenta o mapa de similaridades dos documentos, que é a visualização mais proeminente no TRIVIR.

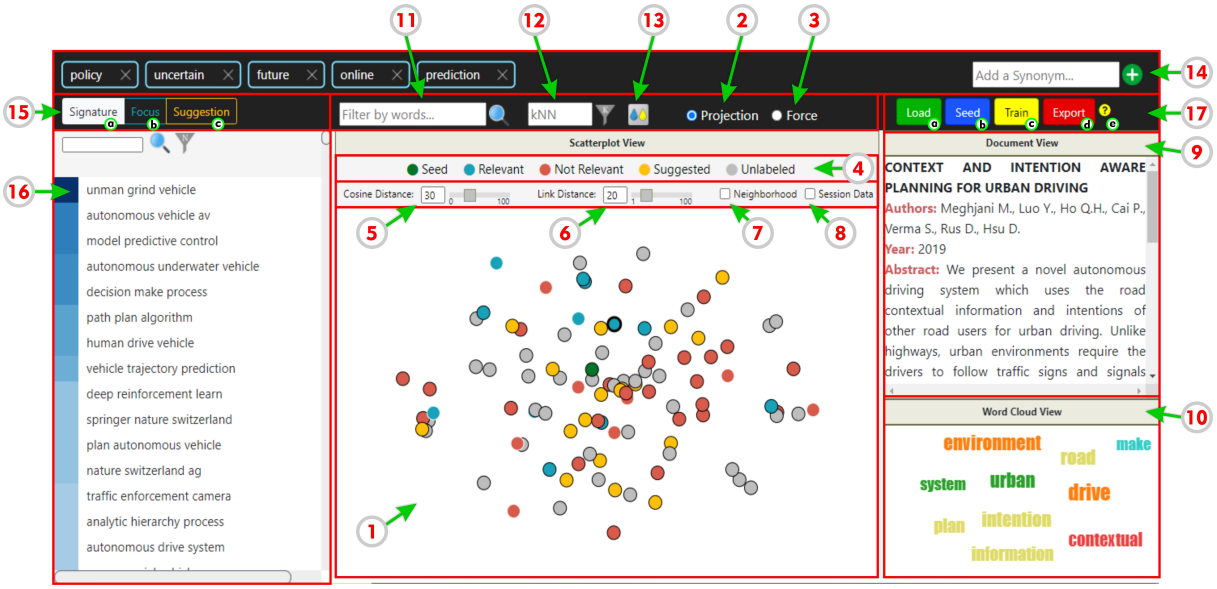
A visualização consiste de uma nuvem de pontos criada com uma técnica de projeção multidimensional - os usuários podem escolher entre t-SNE (MAATEN; HINTON, 2008) ou LSP (PAULOVICH *et al.*, 2008). As técnicas de projeção multidimensional criam uma representação bidimensional do espaço de alta dimensão, no qual os documentos são representados, procurando preservar as vizinhanças originais até certo ponto. Como tal, a interpretação do mapa representado no *Scatterplot View* é que documentos (círculos) espacialmente mais próximos são provavelmente mais semelhantes em conteúdo, enquanto círculos espacialmente distantes correspondem a documentos menos semelhantes. A representação 2D mostrada no mapa é uma aproximação abstrata e com perdas no que diz respeito à precisão da informação, de como os documentos estão organizados no espaço de alta dimensão, o que esperançosamente reflete sua similaridade em termos de conteúdo. Percebe-se que o mapa de documentos é um artefato complexo para usuários comuns não familiarizados com os conceitos subjacentes, o que foi confirmado nas sessões observacionais. Embora geralmente tenha uma pontuação alta (Q2), entende-se que os participantes tiveram dificuldade para usá-lo de forma eficaz. A sobreposição de elementos é um problema, mas outros problemas relacionados à compreensão dos conceitos subjacentes podem ter passado despercebidos.

Na rodada S1 também foram identificados vários problemas específicos de natureza prática que poderiam impactar a apreciação do usuário sobre o sistema. Em particular, entendemos que seria útil tornar a interface mais informativa em relação às diferentes visualizações e suas funções. Além disso, a recuperação dos resultados de uma sessão só era possível acessando um diretório interno, o que não era prático. Os participantes também identificaram problemas menores, por exemplo dificuldades para ler os detalhes do conteúdo dos documentos no *Document View* (texto não formatado); dificuldades de interagir com o mapa quando o grande número de documentos produz sobreposição dos círculos, o que pode ser amenizado com filtragem.

4.2.2 Introduzindo melhorias no sistema TRIVIR

Com base nas avaliações efetuadas, foram propostas e implementadas melhorias no sistema, sendo a nova versão denominada TRIVIR 2.0. Os componentes da interface do sistema TRIVIR 2.0 são apresentados detalhadamente na Figura 22 com a numeração descritiva consistente com a adotada na Figura 20, apresentada na Subseção 4.2.1. Visto que algumas funcionalidades originais permaneceram inalteradas, a seguir são apresentadas apenas as novas modificações inseridas. O componente principal permanece o *Scatterplot View* (1), no entanto

Figura 22 – TRIVIR 2.0 - Interface adaptada



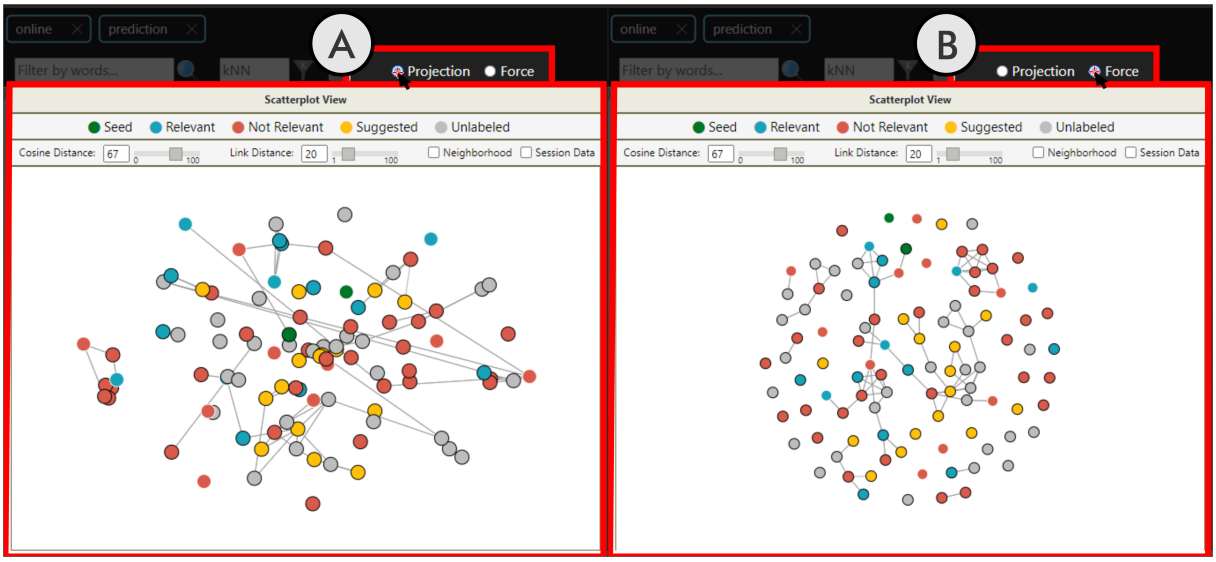
Fonte: Elaborada pelo autor.

com outras funcionalidades de suporte à exploração, como: filtrar o mapa pelo rótulo do documento (4) ou por uma determinada palavra (11), a qual agora é destacada no *Document View* (9); alternar entre a projeção multidimensional (2) ou uma nova visualização interativa baseada em força (3); definir controles do *Scatterplot View*, como distância do cosseno (5), distância da ligação (aresta) (6) ou limites de vizinhança (7); apresentar informações da sessão atual (8); o *Document View* (9) agora é formatado de acordo com os metadados provenientes do conteúdo dos documentos, como Título, Autores, Ano e Resumo; a *Wordcloud View* (10) mostra as palavras frequentes de um documento selecionado; finalmente os botões (17) são para carregar um corpus (17.a), indicar um documento de consulta (17.b), re-treinar o classificador (17.c) após o *feedback* do usuário; exportar os resultados da sessão (17.d), e iniciar um tutorial guiado do sistema (17.e).

Foram consideradas alternativas para melhor transmitir os conceitos subjacentes envolvidos no mapa de similaridade representado no *Scatterplot View*. O sistema Vis-KT (SHERKAT *et al.*, 2018), discutido anteriormente, oferece duas alternativas para criar mapas de similaridade para agrupamento interativo de documentos: a representação 2D pode ser criada com a técnica t-SNE ou com um algoritmo de *Force Layout* (CHEONG; SI, 2020; FRUCHTERMAN; REINGOLD, 1991). Inspirados por essa solução, adicionamos um mapa de similaridade alternativo obtido com um algoritmo de *Force Layout*, supondo que o conceito subjacente é mais intuitivo de entender. Além disso, o *Force Layout* oferece opções de interação adicionais, por exemplo, é possível ajustar o posicionamento do documento de forma dinâmica manipulando os pesos relativos das forças de atração e repulsão. À esquerda na Figura 23 encontra-se a Projeção Multidimensional (A), enquanto à direita encontra-se o *Force Layout* (B). Também foi incorporada nos mapas uma nova informação sobre os relacionamentos entre os documentos, exibindo arestas entre cada documento e seus vizinhos mais próximos. As arestas agregam informação que pode

ter sido perdida no processo de redução da dimensionalidade, por exemplo indicando que círculos espacialmente distantes podem corresponder a documentos vizinhos (i.e, similares) no espaço multidimensional. Esta informação fica mais intuitiva no *Force Layout*, que favorece a formação de grupos visuais dados por vizinhanças entre documentos similares. A Figura 24 mostra dois mapas de documentos de um corpus particular obtido no TRIVIR 2.0 com a projeção t-SNE e o *Force Layout*, respectivamente. Nota-se que os grupos de documentos são mais explícitos na visualização do *Force Layout* do que na visualização da projeção.

Figura 23 – *Scatterplot View:* A) Projeção Multidimensional; e B) Force Layout.



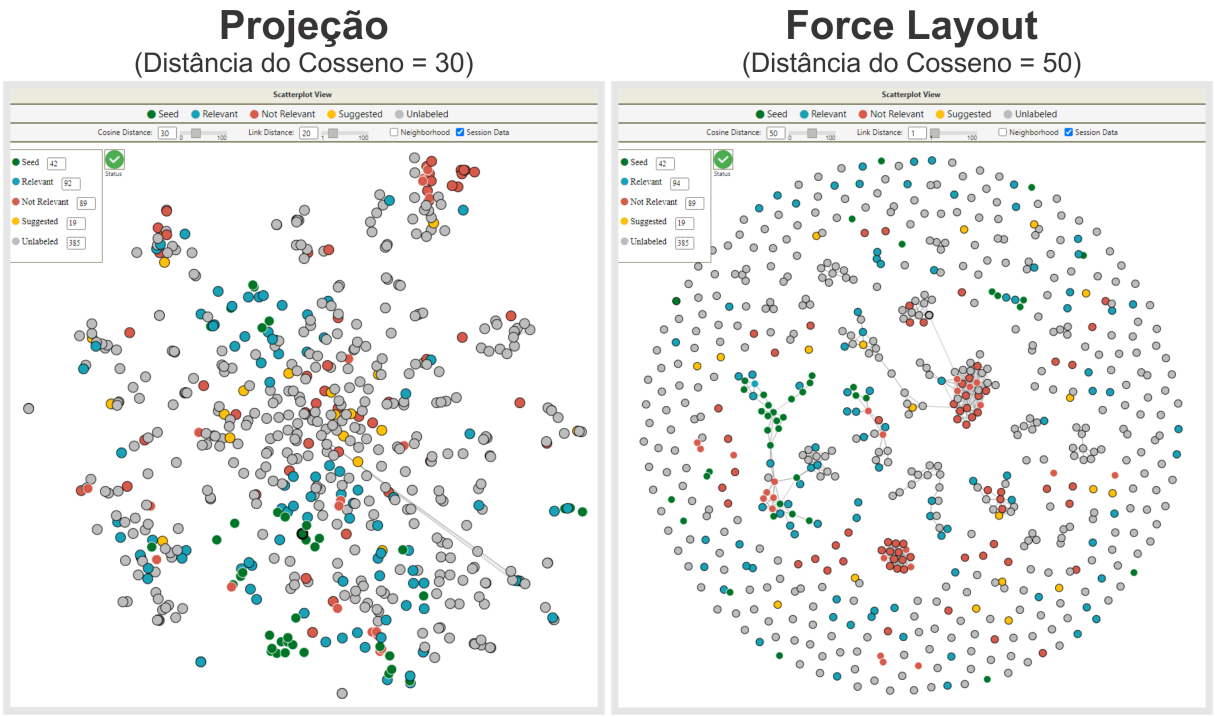
Fonte: Elaborada pelo autor.

Com o intuito de aprimorar a interação do usuário com o mapa de similaridade, foram introduzidos parâmetros com os quais é possível segregar melhor as vizinhanças do documento investigado. Dois parâmetros permitem definir limiares que afetam os grupos observados, a saber: um limiar de distância, identificado como (5) na Figura 22, e um limiar que afeta o componente de força de atração no *Force Layout*, identificado como (6) na mesma figura.

As Figuras 25 e 26 ilustram a finalidade destes parâmetros em uma base de dados de artigos científicos relacionados ao tema de RI e VA. Na Figura 25 a interação com a funcionalidade *Cosine Distance* é apresentada a partir da sua variação em quatro valores definidos por $d \in \{30, 40, 50, 60\}$ pertencentes a um intervalo [0, 100]. Horizontalmente o valor de d é variado, enquanto verticalmente ambas as visualizações são apresentadas. Uma vez que a projeção multidimensional é estática, a diferença visual se dá por meio do acréscimo de arestas entre os relacionamentos, sem modificações no posicionamento dos círculos. No entanto, o *force layout* é dinâmico, portanto observa-se tanto a adição de novas arestas, quanto a mudança de posicionamento dos círculos de acordo com o grau de similaridade.

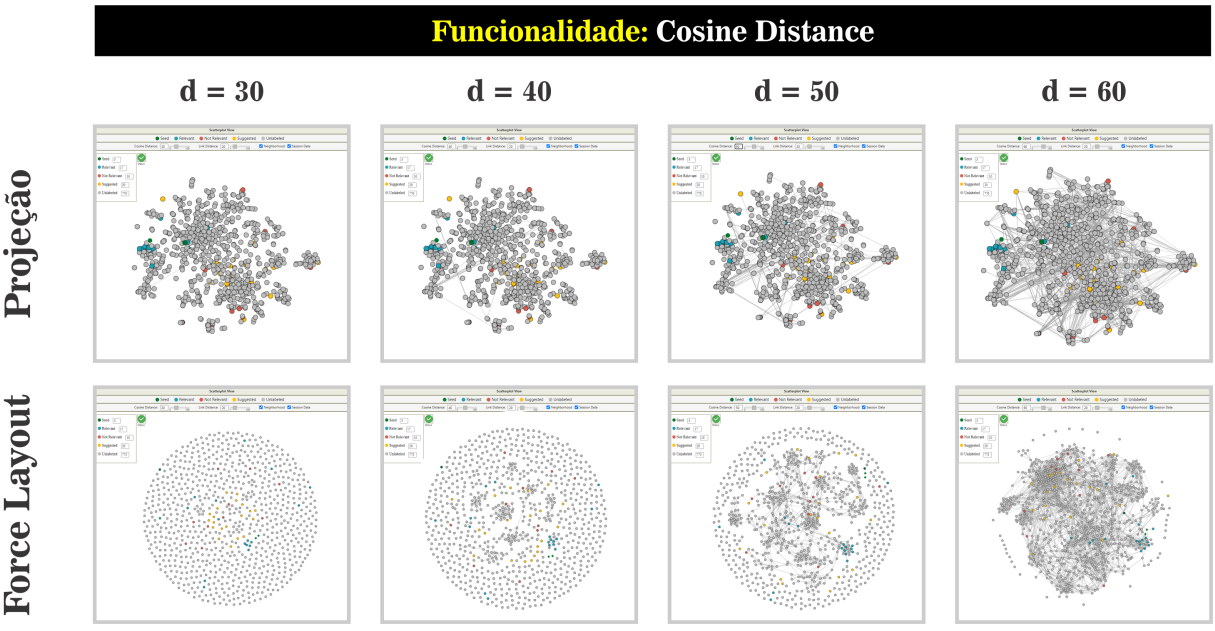
Na Figura 26 a interação com a funcionalidade *Link Distance* é apresentada a partir da variação de $l \in \{1, 10, 30, 50\}$ pertencentes a um intervalo [1, 100], em que ao diminuir o valor

Figura 24 – *Scatterplot View* mostrando mapas de similaridade de documentos para uma coleção, obtidos com a projeção t-SNE (esquerda) e o algoritmo *Force Layout* (direita).



Fonte: Elaborada pelo autor.

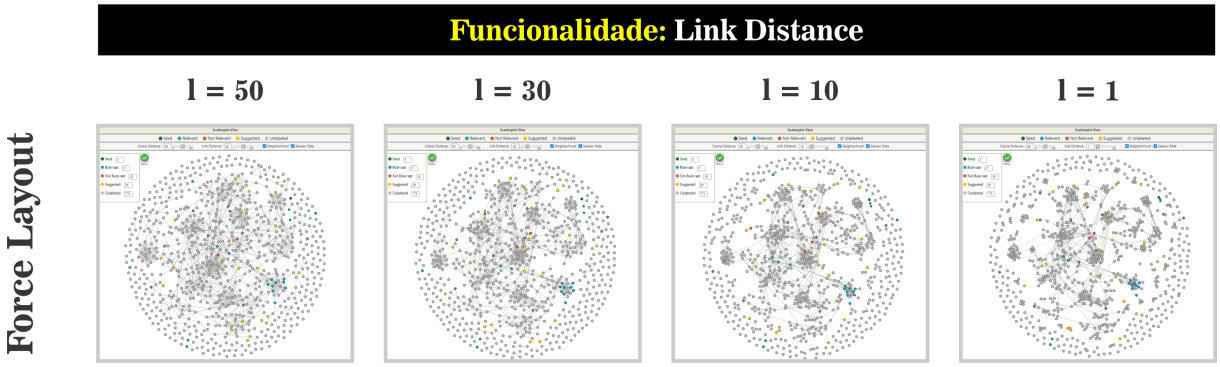
Figura 25 – A funcionalidade *Cosine Distance* permite destacar relacionamentos entre documentos a partir da similaridade entre eles.



Fonte: Elaborada pelo autor.

atribuído a l, menores são as distâncias entre os círculos, proporcionando maior segregação visual dos grupos formados. Ambas as funcionalidades permitem ao usuário investigar a melhor combinação de valores d e l para suas respectivas coleções, uma vez que podem oferecer diferentes perspectivas para a análise.

Figura 26 – A funcionalidade *Link Distance* possibilita aumentar/diminuir o tamanho da aresta entre os documentos.



Fonte: Elaborada pelo autor.

Um problema adicional com o mapa é quando o tamanho do corpus produz uma sobreposição dos círculos, o que dificulta a visibilidade e as interações, como a seleção de documentos. O problema pode ser atenuado filtrando os documentos exibidos com base em determinados critérios. No TRIVIR 1.0 é possível filtrar o mapa por palavra (11), por vizinhança em relação ao documento de consulta inicial (k-vizinhos mais próximos) (12), ou por relevância do documento (Semente, Relevante **E** Sugerido como Relevante) (13). No TRIVIR 2.0 foi adicionada a possibilidade de filtrar por qualquer categoria de documento (semente, relevante, sugerido como relevante, não relevante **OU** ainda não rotulado) (4). Esta funcionalidade foi desenvolvida para complementar a investigação das classes dos documentos de maneira seletiva pelo usuário, uma vez que qualquer classe pode ser selecionada, diferentemente da versão 1.0 em que a única opção era o filtro de relevância (13). Veja um exemplo de uso desta funcionalidade na Figura 27 a seguir.

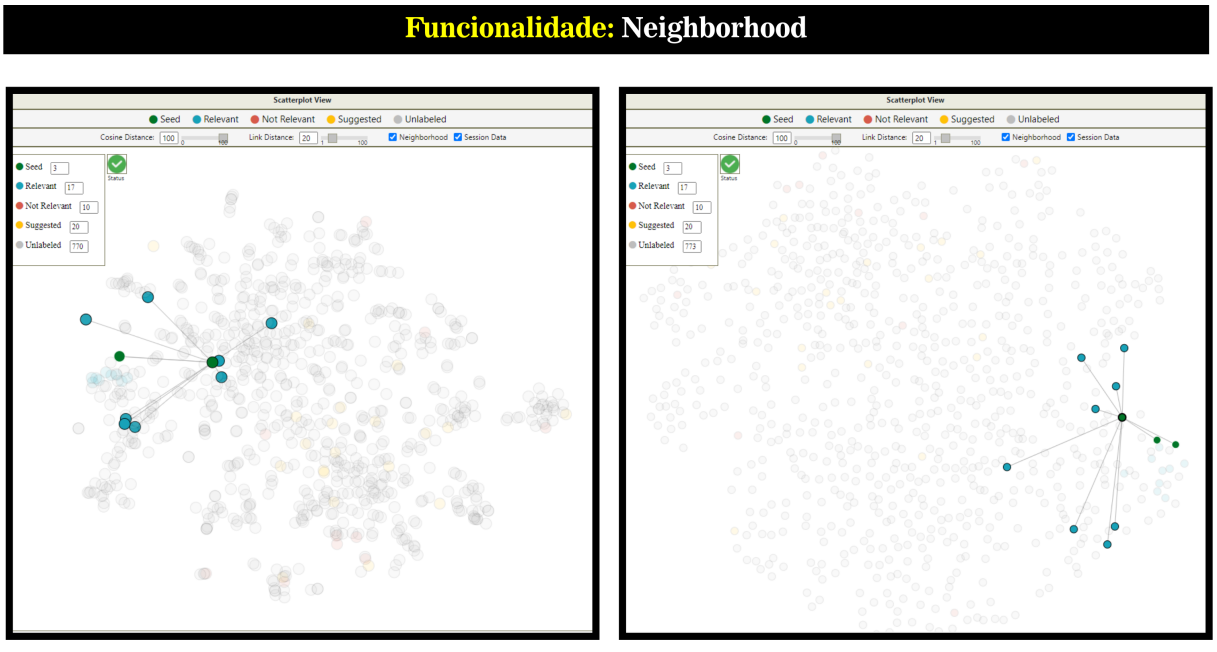
Com o objetivo de melhorar o *feedback* visual do mapa de similaridade, independentemente de ser gerado por algoritmos de projeção ou força, foi inserida a possibilidade de traçar uma linha para conectar explicitamente cada documento aos seus k vizinhos mais próximos, para um k definido pelo usuário (até k = 10), conforme ilustrado na Figura 28. À esquerda desta figura é apresentado um exemplo de vizinhança de um documento semente selecionado na projeção, enquanto à direita da figura o mesmo documento é selecionado, porém no *force layout*. Desta forma, um dos problemas identificados em S1 foi abordado, que é fornecer informação de vizinhança relativa a um documento arbitrário, em vez de apenas em relação ao documento de consulta inicial.

A exibição alternativa do *Scatterplot View* e seus parâmetros adicionais são as adições mais significativas ao TRIVIR 2.0. Também foram introduzidas modificações visando outras

Figura 27 – Filtro por classe dos documentos.

Fonte: Elaborada pelo autor.

Figura 28 – A funcionalidade *Neighborhood* destaca os relacionamentos de um documento selecionado.

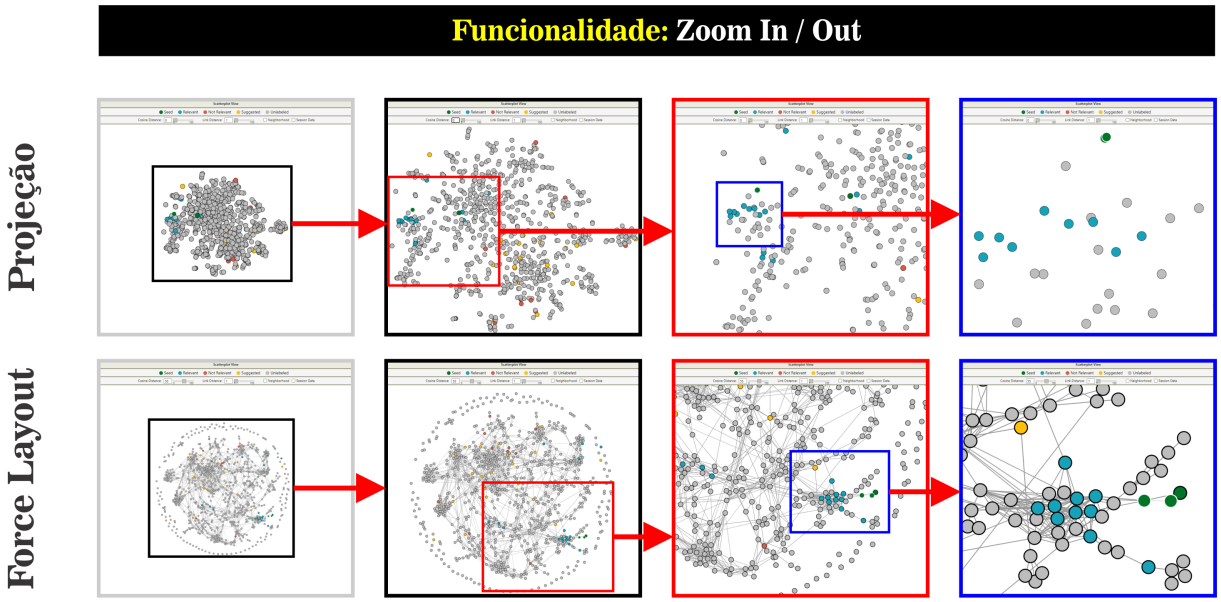


Fonte: Elaborada pelo autor.

questões específicas, a saber: modificação de escala interativa dos círculos quando *Zoom In/Out* é aplicado; adaptação dos ícones do sistema; adição de textos informativos em caixas textuais, como em (11, 12 e 14); inserção de cabeçalhos para as abas *Scatterplot View*, *Document View* e *Wordcloud View*; uma opção para exibir informações sobre o estado atual da sessão (quantos documentos por categoria); adicionado botões explícitos para carregar ou salvar uma sessão; incluído *feedback* visual sobre o status de processamento do servidor; modificação visual de um documento, caso o usuário já tenha o classificado (Lido); melhorado o *layout* do *Document View*, para enfatizar metadados como Título, Autores, Ano e Resumo e destacar os termos pesquisados; adicionada a *Wordcloud View* para representar as palavras importantes de um documento selecionado; e, finalmente, incluído um tutorial guiado dos recursos do sistema, acessível a qualquer momento.

Durante os estudos realizados, o usuário U4 relatou a necessidade da modificação do tamanho dos círculos quando um processo de aproximação/afastamento (*Zooming*) é realizado pelo usuário, visto que no TRIVIR 1.0 o *Zooming* apenas dispersava os círculos, sem a modificação de escala. De acordo com U4, que é um professor e pesquisador na área de IHC, esta modificação é importante tanto por questões de acessibilidade, pois aumentar o tamanho dos círculos facilita a exploração por usuários com baixa visão, quanto pela intuitividade no mundo real de objetos próximos parecerem maiores, enquanto objetos distantes parecem menores. A Figura 29 exemplifica o novo processo de *Zooming* em ambas as visualizações, agora com a modificação de escala dos círculos.

Figura 29 – A funcionalidade *Zoom In/Out* permite aproximar/afastar regiões do mapa de similaridade.

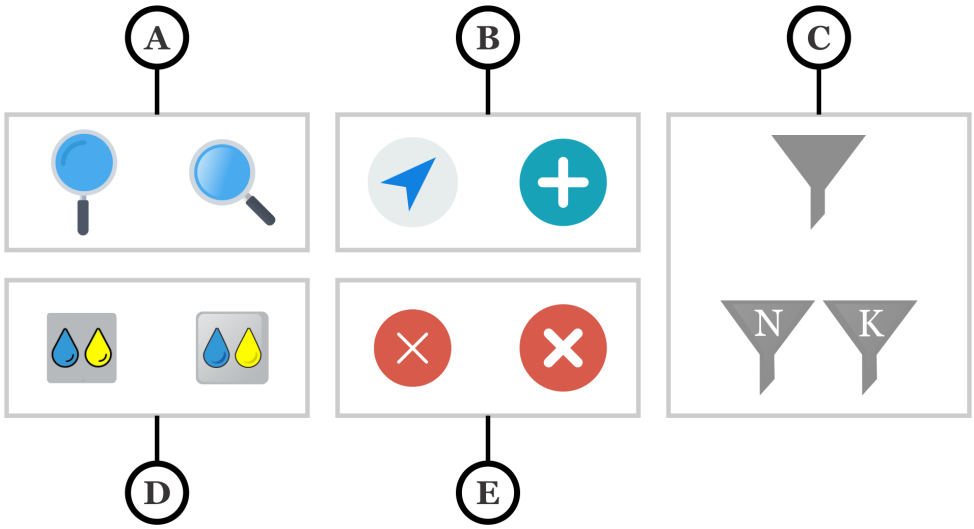


Fonte: Elaborada pelo autor.

Outro aspecto levantado por U4, e também por U3, U5 e U6 durante os estudos foi a adequação dos ícones utilizados no TRIVIR 1.0. Para apresentar as motivações por trás das modificações realizadas, observe a Figura 30, sendo A-E as principais modificações realizadas. À

esquerda das caixas {A,B,D,E} e acima na caixa C estão os ícones do TRIVIR 1.0, enquanto que à direita em {A,B,D,E} e abaixo em C estão os novos ícones do TRIVIR 2.0. Foi relatado que as versões antigas de A e B remetiam à localização, como no Google Mapas, embora devessem remeter a uma lupa e à adição de documentos como relevantes. Em C foi relatado que o ícone causava confusão de acordo com sua funcionalidade, uma vez que o mesmo ícone era aplicado em dois filtros com finalidades distintas. Para contornar esta situação, foi adicionado um K denotando o filtro kNN e um N denotando o filtro por n-gramas. Em D e E foi relatado que os tons das cores e legibilidade do conteúdo deveriam estar mais definido visualmente. Para D foi introduzida uma iluminação, aprimorando o aspecto de "botão" clicável, enquanto em E o 'X' que indica fechamento foi realçado.

Figura 30 – Novos ícones adicionados na interface do sistema.

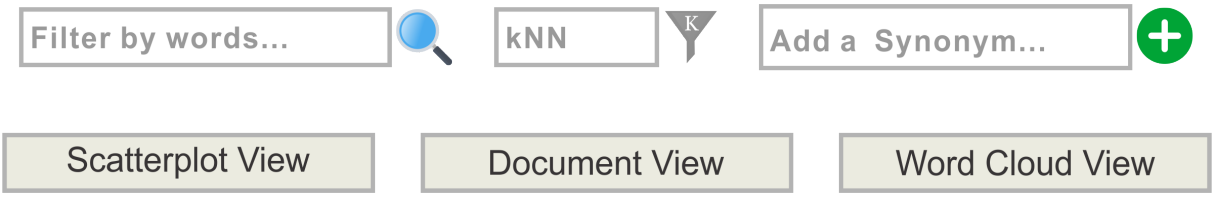


Fonte: Elaborada pelo autor.

Um questionamento recorrente por parte dos usuários relacionava-se à identificação das funcionalidades na interface e sua finalidade. Embora a apresentação do sistema contemplasse todos estes aspectos, recordar todos os detalhes não é uma tarefa fácil. Logo, foi solicitado que informações relevantes fossem adicionadas à interface, principalmente quanto ao significado das categorias de documentos. Para atender esta demanda, foi adicionado o filtro por classe (4) (Figura 27), o qual identifica e define cada classe, bem como descrições em caixas textuais e cabeçalhos identificando as principais abas do sistema (Figura 31). Os usuários em geral relataram que estas modificações tornaram ágil o processo de introdução às funcionalidades.

Em Nielsen (1994) erro é definido como uma ação que não leva ao resultado esperado. Neste sentido, o sistema deve apresentar uma baixa taxa de erros, ou seja, o usuário não pode cometer muitos erros durante o seu uso, não pode perder seu trabalho e deve perceber que errou, quando for o caso. Em outras palavras, um dos princípios da interatividade de sistemas é manter o usuário informado acerca do *status* de execução e das informações da sessão. Assim, a funcionalidade *Session Data* foi desenvolvida para apresentar ao usuário o seu progresso

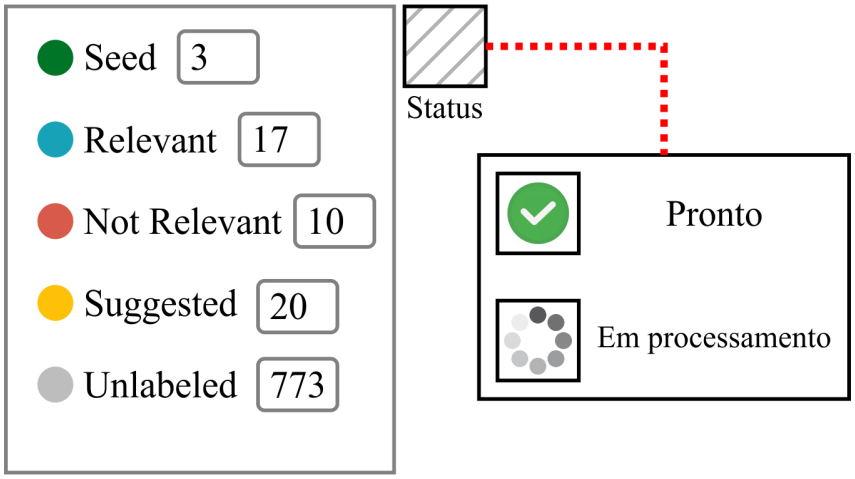
Figura 31 – Adição de descrições em caixas textuais e cabeçalhos nas principais funcionalidades.



Fonte: Elaborada pelo autor.

durante a sessão, a partir da quantidade de documentos classificados e os respectivos resultados em cada categoria, bem como informar em tempo real o *status* de processamento do sistema. Ambas as informações estão diretamente associadas com o *workflow* do protocolo CAL, uma vez que a finalização da sessão depende de o usuário considerar suficiente a quantidade de documentos rotulados. As informações agora apresentadas facilitam essa decisão. Além disso, a continuidade do processo de rotulação só é possível ao terminar o processamento da rotulação anterior, uma vez que são tratados como requisições sequenciais ao *back-end* do sistema. Isso acarretava em frustração dos usuários, que não tinham *feedback* do sistema sobre o *status* do processamento e continuavam a rotulação sem sucesso, o que inclusive causou travamentos do sistema por múltiplas requisições. Com a indicação do *status* do sistema em tempo real, os usuários reportaram que se sentem mais confiantes ao realizar o processo de rotulação.

Figura 32 – A funcionalidade *Session Data* apresenta a quantidade de documentos por classe e apresenta o *status* de execução do sistema.



Fonte: Elaborada pelo autor.

Outro aspecto importante em um sistema é permitir realizar todas as operações disponíveis a partir da sua interface. No TRIVIR 1.0 o carregamento de um corpus e a indicação de um documento de consulta inicial eram realizados diretamente pelo código fonte do servidor, o que demandava conhecimentos de programação por parte dos usuários finais. Como o sistema foi avaliado em ambientes controlados, sendo configurado com antecedência pelo pesquisador responsável, tais funcionalidades não demandaram esforço dos usuários. No entanto,

isso inviabiliza o uso do sistema em larga escala, principalmente por usuários não especialistas. Portanto, alguns botões foram desenvolvidos para possibilitar esta comunicação da interface do sistema com o servidor, veja a Figura 33. O botão *Load* carrega o corpus e a seguir o botão *Seed* carrega o documento de consulta (semente). O usuário pode utilizar o botão *Train* para retreinar o classificador. Vale ressaltar que o *Train* é o único botão dos apresentados na figura que estava presente no TRIVIR 1.0, no entanto vários usuários criticaram o seu posicionamento como de difícil acesso, visto que ficava dentro da aba *Suggestion*. O botão *Export* foi incluído para permitir o *download* dos resultados obtidos durante a utilização do sistema de forma legível aos usuários, que originalmente estavam limitados a arquivos de configuração em formato ***.json*** no servidor. Embora a implementação atual gere um arquivo textual (***.txt***) contendo todos os documentos rotulados e recomendados, é possível estender a implementação em uma versão futura para exportar diferentes formatos, como ***.csv*** ou ***.bib***, permitindo compatibilidade com outros sistemas, como o Parsifal, ou o Mendeley.

Figura 33 – Novos botões adicionados na interface do sistema.

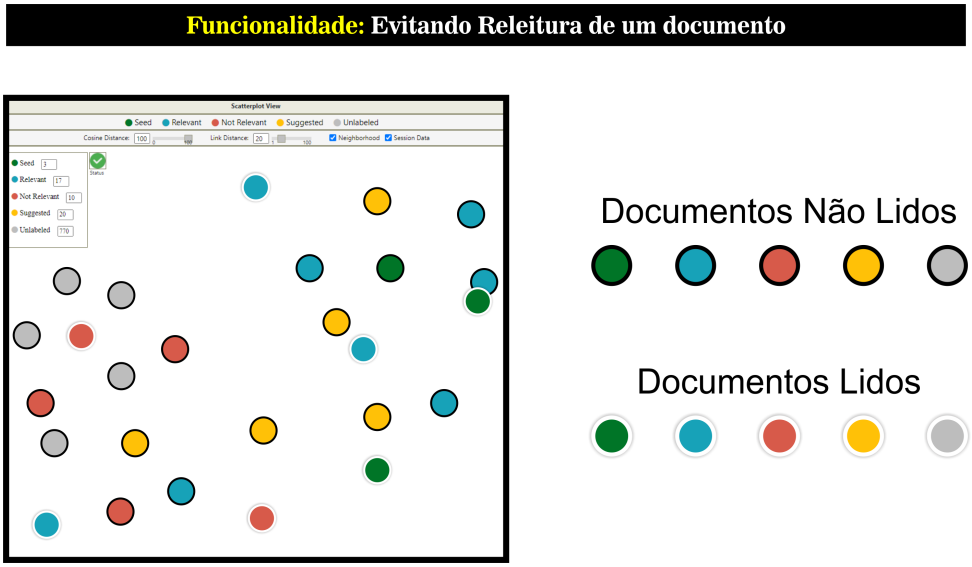
Fonte: Elaborada pelo autor.

Visto que os corpora investigados podem conter centenas, ou até mesmo milhares de documentos, facilmente o usuário poderá repetir a exploração de documentos de maneira não intencional, realizando um retrabalho. Para evitar este processo de releitura, assumimos que um documento classificado pelo usuário foi lido e então foi decidido espontaneamente rotulá-lo, ou seja, tal documento pode ser definido como já explorado. Assim, a interface modifica a cor da borda do círculo para diferenciá-lo. Círculos com borda preta por padrão representam documentos ainda não rotulados (lidos) pelo usuário, enquanto documentos com borda branca já o foram, observe o exemplo na Figura 34.

Alguns usuários relataram dificuldades na legibilidade do conteúdo dos documentos e também no desgaste causado pelo alto tempo investido na leitura do conteúdo. A Figura 35 ilustra as soluções propostas. No *Document View* (esquerda) o conteúdo foi justificado e os metadados foram ressaltados, como o Título em maiúsculo e negrito e as definições dos campos seguintes, como "Autores", "Ano" e "Resumo", indicando visualmente as subdivisões do conteúdo. Enquanto na *Wordcloud View* (direita) as palavras mais frequentes são apresentadas para agilizar a identificação do tema do documento.

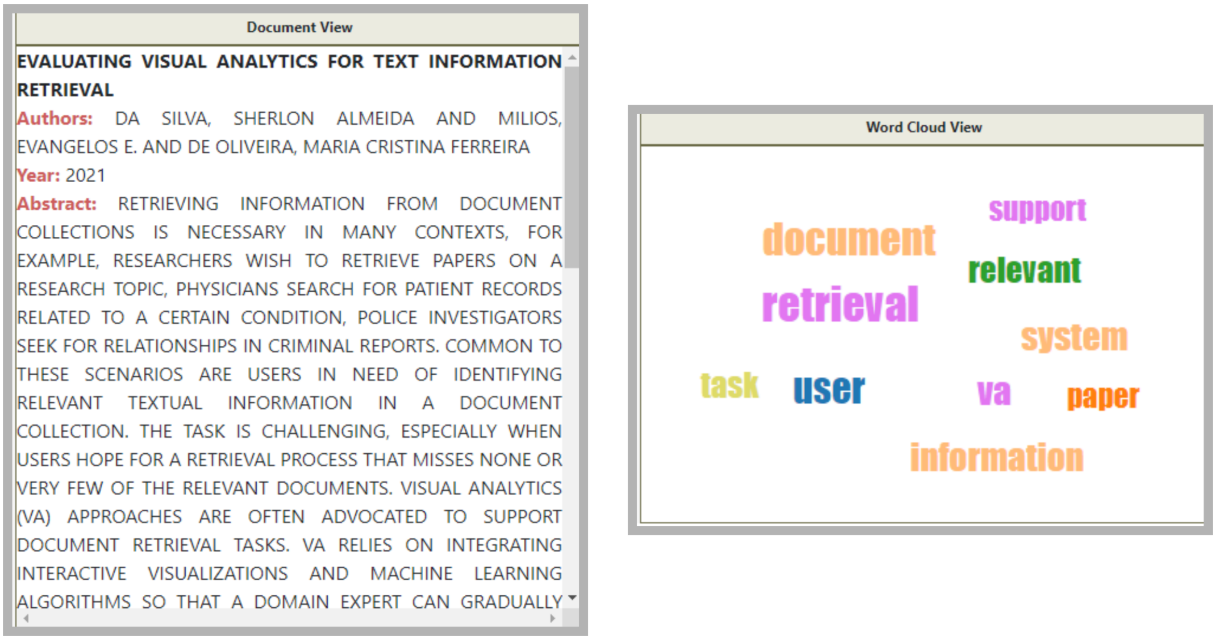
A utilização de palavras chave reconhecidas como relevantes permite identificar documentos relacionados mais facilmente. O filtro por palavras (11) no TRIVIR 1.0 permitia identificar os documentos que continham certos termos, no entanto uma limitação ocorria ao identificar o contexto deste termo dentro do conteúdo do documento. Para agilizar este processo de identificação os termos agora são destacados diretamente no conteúdo do documento sele-

Figura 34 – Esta funcionalidade indica documentos já lidos pelo usuário, evitando o processo de releitura.



Fonte: Elaborada pelo autor.

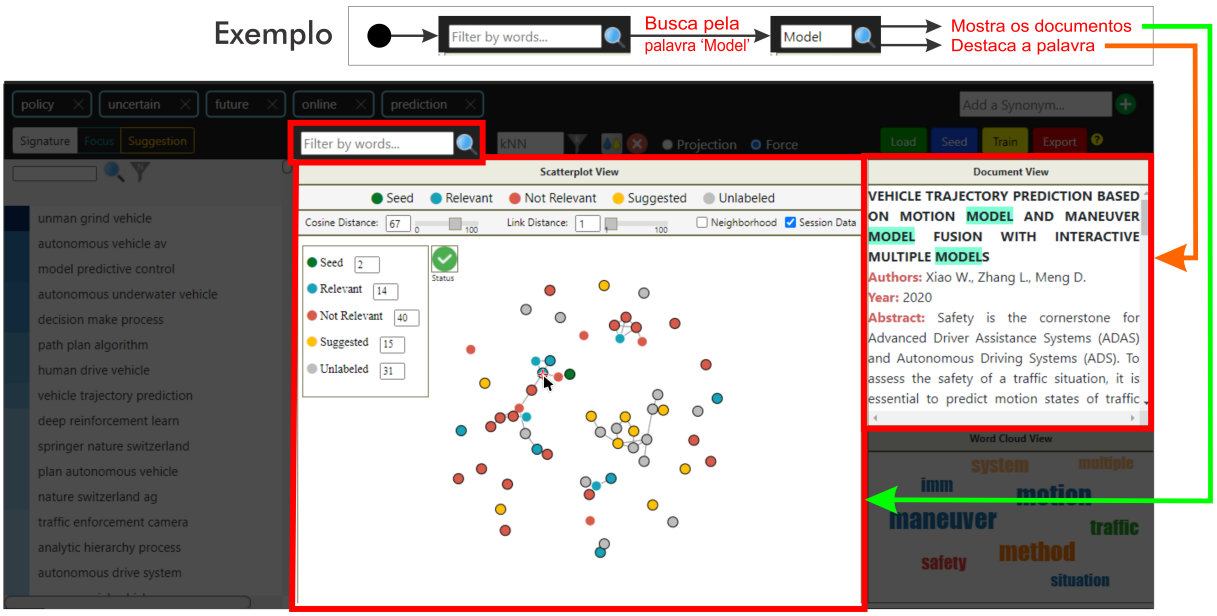
Figura 35 – *Document View* (esquerda) apresenta o conteúdo do documento selecionado, enquanto a *Wordcloud View* (direita) apresenta as suas palavras mais frequentes.



Fonte: Elaborada pelo autor.

cionado. Veja um exemplo de utilização na Figura 36, em que a palavra "Model" é buscada e destacada em um documento selecionado pelo usuário.

Figura 36 – A funcionalidade de busca por termos permite identificar documentos que contém uma palavra desejada, destacando-a no *Document View*.

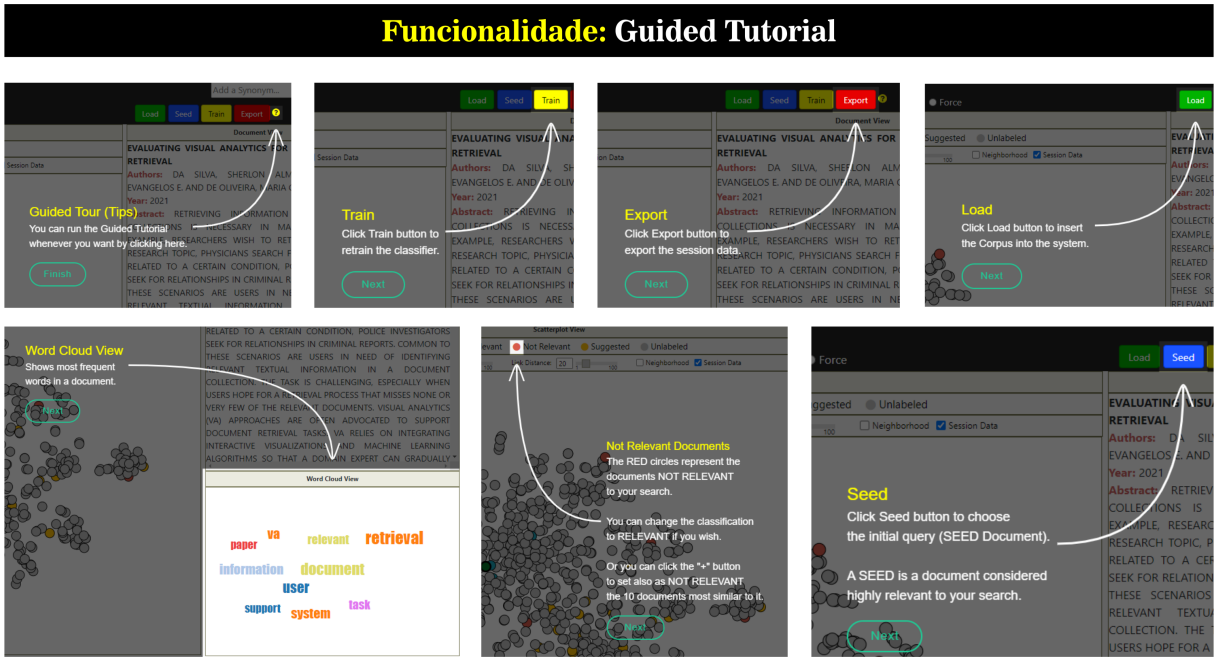


Fonte: Elaborada pelo autor.

De acordo com Nielsen (1994), todo sistema precisa ser fácil de aprender de forma que o usuário possa rapidamente começar a interagir. Segundo ele, é o mais importante atributo de usabilidade, por ser a primeira experiência de qualquer usuário com o sistema. Por isso, foi inserida uma sumarização de todas as funcionalidades do sistema, disponibilizada como um tutorial interativo, que o usuário pode acessar a qualquer o momento a partir do botão representado pelo símbolo de interrogação '**?**' na Figura 33. Exemplos de uso de cada funcionalidade são apresentados no tutorial, bem como a finalidade de cada funcionalidade. Alguns exemplos de telas do tutorial são apresentados na Figura 37.

Ressalta-se que todas as propostas de melhorias foram baseadas nas limitações encontradas durante o estudo S1 a partir dos relatos e sugestões dos usuários. Limitações mais críticas, bem como melhorias de contribuição significativa ao propósito do sistema foram priorizadas. A partir dos questionários foram criados dois grupos de modificações: o primeiro diz respeito àquelas contribuições para pequenas correções de problemas identificados (Melhorias e correção de *bugs*) e o segundo a contribuições com alto potencial de melhoria (Desenvolvimento de novas funcionalidades). As respostas e sugestões dos usuários foram categorizadas e agrupadas por funcionalidades do TRIVIR 1.0, e na sequência os problemas identificados foram ranqueados prioritariamente. Este trabalho não teve por objetivo propor soluções para todos os problemas encontrados no TRIVIR, mas sim aplicar conceitos de VA e IHC para as principais limitações identificadas, a fim de obter a perspectiva dos usuários em tarefas reais de RI.

Figura 37 – A funcionalidade *Guided Tutorial* permite ao usuário explorar interativamente todas as funcionalidades do sistema e obter informações de suas finalidades.



Fonte: Elaborada pelo autor.

4.2.3 Avaliando o sistema TRIVIR 2.0

A segunda rodada de estudos observacionais foi realizada com 11 participantes usando o sistema TRIVIR 2.0 [3], 8 dos quais (U1-U8) já haviam participado de S1, veja a Tabela 2. O participante U9 de S1 não participou de S2, enquanto U10-U12 são três novos voluntários. A Parte 2 do questionário foi expandida para contemplar as novas funcionalidades, conforme descrito na Tabela 5. Foram preservadas todas as perguntas de S1, embora as perguntas tenham sido reformuladas e numeradas conforme necessário para consistência de apresentação no novo ambiente.

Primeiramente, ao analisar as pontuações das 10 funcionalidades já avaliadas em S1, representadas nas Figuras 38a e 38b, observa-se um padrão semelhante ao observado em S1, de classificações mais baixas para alguns recursos por alguns usuários, novamente com os usuários mais experientes sendo mais críticos, de modo geral. Suas avaliações do filtro kNN no mapa de similaridade (Q11) e a *Terms View* (Q13) permanecem baixas. Em relação à *Terms View*, os participantes reconheceram que as cinco palavras iniciais são interessantes, mas os sinônimos mostrados não são. Também consistentes são as pontuações mais baixas atribuídas às três visualizações de lista (Q14, Q15, Q16).

As Figuras 38c e 38d mostram as avaliações em resposta às novas questões introduzidas em S2. Aqui, as pontuações mais baixas aparecem principalmente no controle de distância do cosseno (Q4), no *Wordcloud View* (Q9) e no *Scatterplot View* (Q20.a). Novamente, os

[3] <https://github.com/SherlonAlmeida/TrivirV2.0>

Tabela 5 – Questionário Parte 2 (S2) que visa avaliar a utilidade percebida das funcionalidades do sistema. Respostas na faixa [0-10], de menos útil (0) a mais útil (10), exceto em Q19 (verifique na lista as funcionalidades a serem melhoradas) e Q21 (a-e) (as respostas estão na Escala Likert).

Questionário Parte 2: S2
How do you rate the ... :
1) system as a whole? 2) Scatterplot View (Similarity Map of Documents)? 3) filtering of the Scatterplot view by document label? 4) Cosine Distance control for connecting the similar documents in the Scatterplot view? 5) Link Distance control to bring similar documents closer or push them apart in the Scatterplot view? 6) Neighborhood control to display only the closest neighbors to a selected document in the Scatterplot view? 7) Session Data control for showing the numeric distribution of documents per label and system status? 8) Document View for displaying a document´s content? 9) Word Cloud view of the most frequent words in a document? 10) functionality to filter the Scatterplot view to display only documents containing specific terms? 11) functionality to filter the Scatterplot view to display only the K documents most similar to the query document? 12) functionality to filter the Scatterplot view by document label (filter by seed, relevant or suggested documents) ? 13) Terms view to help identifying synonyms to important terms in the Seed documents? 14) Signature List View, which displays the most frequent tri-grams in the corpus? 15) Focus List view, which shows the documents currently labeled as Relevant? 16) Suggestion List view, which shows the documents currently suggested as Relevant by the ML algorithm? 17) functionalities to load the Corpus (Load), define the initial query document (Seed), retrain the classifier (Train) and export the session result (Export), respectively? 18) Guided Tour (Help) of the system's functionalities?
19) In your opinion, which functionalities need further improvement?
20.a) Considering the scatterplot views generated by the Projection (A) and Force Layout (B), how do you rate their integration to identify patterns in the data?
20.b) How do you prefer to use the Scatterplot View?
Comparing the first version of the system (A) and the current version (B):
21.a) the system support to the document retrieval task has improved. 21.b) the system interface has improved regarding the placement of the windows. 21.c) the system interface has improved regarding the use of available screen space. 21.d) the interface has improved regarding the suitability of icons and use of colors.
21.e) The coloring of the documents borders as white (read) and black (unread) facilitated the process of reading the documents, since it mitigates the re-reading process.

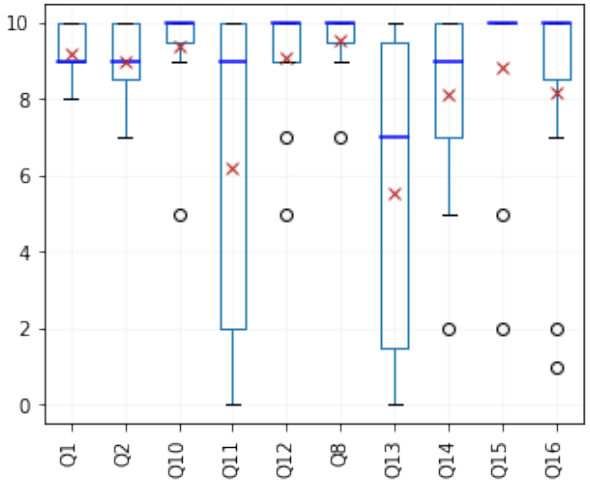
participantes experientes U3, U4 e U12 expressam as opiniões mais críticas. A maioria dos participantes considerou o *Wordcloud View* (Q9) útil como um complemento para auxiliar na interpretação de agrupamentos observados no *Scatterplot View*, mas alguns sugeriram que poderia haver melhor uso das cores, visto que as cores não têm significado particular na *Wordcloud View*.

A Questão Q20 aborda a exibição do *Scatterplot View* que descreve o mapa de similaridade de documentos. Q20a solicita ao participante para avaliar sua utilidade. Como agora é possível alternar entre as visualizações de projeção e força, em Q20.b pedimos sua preferência por uma delas. Sete dos 11 participantes declararam preferir o mapa do *Force Layout*, enquanto 4 declararam alternar entre os dois tipos de mapas em situações diferentes e os consideraram complementares. Curiosamente, o *Scatterplot View* não recebeu pontuações baixas na rodada S1. A complexidade adicional de ter duas visões alternativas, além de vários controles adicionais, pode ser responsável pelas classificações mais baixas.

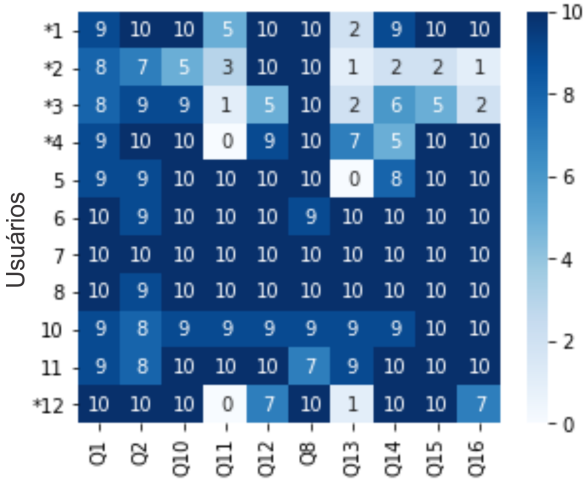
Na segunda rodada (S2) também foi solicitado aos participantes uma avaliação geral da interface, quanto à sua capacidade de apoiar a tarefa de recuperação de documentos e a disposição e qualidade de seus componentes visuais. Foram incluídas as questões Q21 (a-e), perguntando se consideram que as modificações no TRIVIR 2.0 contribuíram para uma melhoria em sua percepção geral sobre o sistema. As pontuações atribuídas pelos 8 participantes que usaram o TRIVIR 1.0 em S1, mais U12, que não participou de S1, mas estava familiarizado com o TRIVIR 1.0, são apresentadas nas Figuras 38e e 38f. O *feedback* geral é positivo, mas novamente os participantes com alguma experiência (U3, U4 e U12) expressam uma opinião mais crítica sobre certos aspectos. O participante U4 discordou veementemente do posicionamento das janelas (Q21.b) e do uso do espaço da tela (Q21.c), ele os considerou tão bons quanto antes, mas criticou as cores usadas nos botões e no *Wordcloud*, que são semelhantes aos usados para codificar as categorias de cada documento. Os participantes também apontaram o baixo contraste da borda branca nos círculos, modificação introduzida para indicar os documentos já lidos. No entanto, os três participantes U3, U4 e U12 concordaram fortemente que as modificações contribuíram para melhorar a tarefa de recuperação de documentos (Q21a).

Por fim, em ambas as rodadas de estudos S1 e S2 foi solicitado aos participantes que identificassem quaisquer funcionalidades que eles consideravam que poderiam ser melhoradas, os resultados são apresentados em Tabela 6. No geral, a tabela reflete as críticas já discutidas, por exemplo, a *Terms View* recebeu muitas menções em S1 e S2. Porém, outros recursos, como as listas *Signature*, *Focus* e *Suggestion*, receberam menos menções no S2, embora não tenham sofrido alterações específicas no TRIVIR 2.0. Talvez a apresentação mais cuidadosa dos recursos da interface e a familiaridade prévia dos participantes com o sistema tenham favorecido um melhor entendimento de seu papel, principalmente a *Signature List View*, pois foi observado que ela foi acessada por vários participantes para identificar trigramas relevantes para a sua busca.

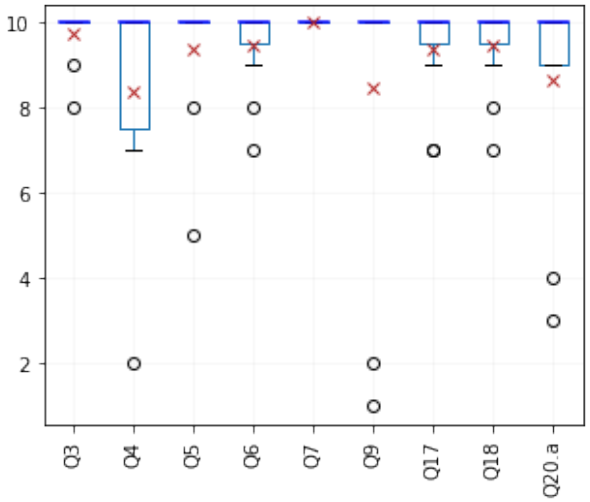
Figura 38 – Boxplots e heatmaps para as avaliações no estudo S2.



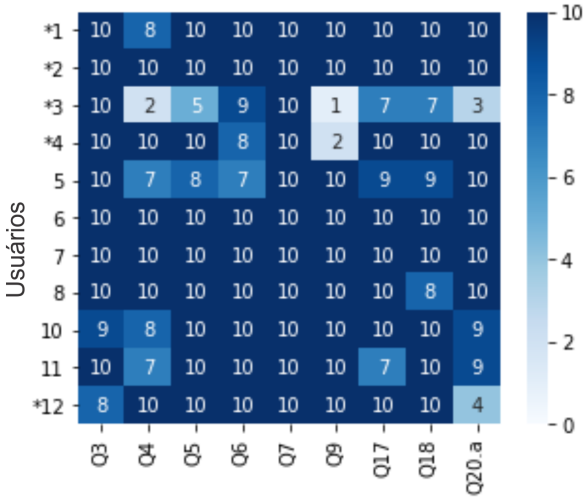
(a) Distribuição das avaliações.



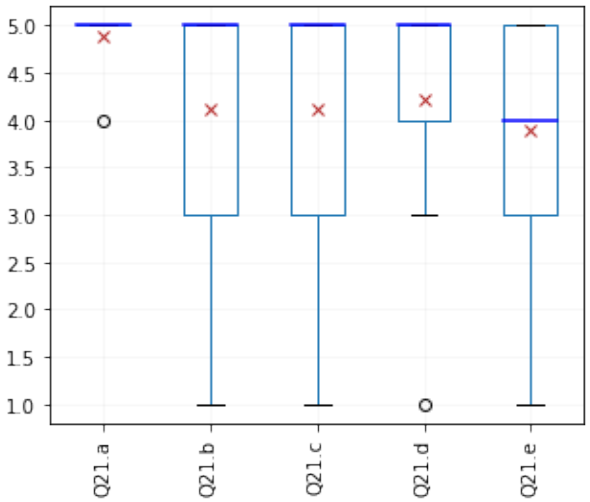
(b) Avaliações por participante.



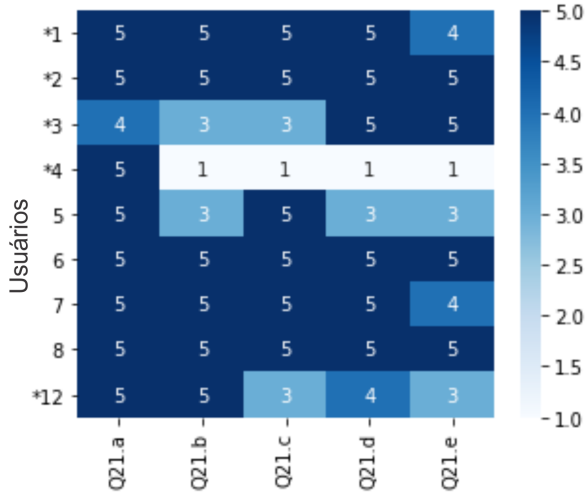
(c) Distribuição das avaliações.



(d) Avaliações por participante.



(e) Distribuição das pontuações Q21(a-e).



(f) Pontuações por participante Q21(a-e).

Fonte: Elaborada pelo autor.

Tabela 6 – TRIVIR: *checkbox* nos estudos S1 e S2. O prefixo SV indica os parâmetros relacionados ao Scatterplot View.

S1 - Questão 11			
ID	**Funcionalidade**	**Votos**	**Usuários**
15 (a)	Signature List View	6	66.7%
14	Terms View	5	55.6%
15 (c)	Suggested List View	5	55.6%
15 (b)	Focus List View	4	44.4%
1	Similarity Map of Documents	3	33.3%
9	Document View	2	22.2%
12	Filter by K most similar	2	22.2%
11	Filter by Terms	2	22.2%
13	Point clutter reduction	2	22.2%
S2 - Questão 19			
ID	**Funcionalidade**	**Votos**	**Usuários**
14	Terms View	7	63.6%
10	Word Cloud View	5	45.5%
1	Similarity Map of Documents	3	27.3%
9	Document View	3	27.3%
12	Filter by K most similar	3	27.3%
15 (b)	Focus List View	3	27.3%
17 (a-d)	Buttons	3	27.3%
5	SV: Cosine Distance Filter	2	18.2%
7	SV: Neighborhood Filter	2	18.2%
11	Filter by Terms	2	18.2%
15 (a)	Signature List View	2	18.2%
15 (c)	Suggested List View	2	18.2%
17 (e)	Guided Tour	2	18.2%
2	SV: Projection	1	9.1%
4	SV: Filter by labels	1	9.1%
6	SV: Link Distance Filter	1	9.1%
13	Point clutter reduction	1	9.1%

4.2.4 SUS: analisando usabilidade

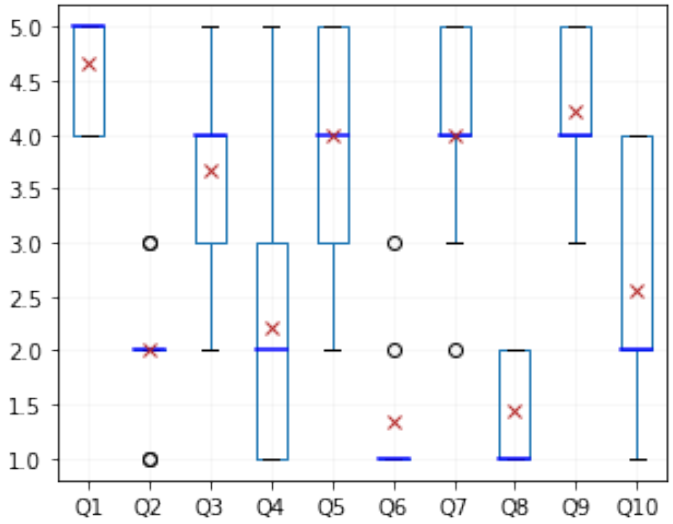
A Parte 1 do questionário respondido pelos participantes corresponde às questões do SUS, listadas na Tabela 7. A Figura 39 mostra *boxplots* e *heatmaps* das pontuações atribuídas por cada participante em S1. Os *heatmaps* que representam as respostas foram divididos em dois: o da esquerda com as questões ímpares, para os quais os números mais altos indicam melhores pontuações e o da direita com os números pares, para os quais os valores mais baixos são melhores. Tons mais escuros de azul indicam os melhores resultados. A Figura 40 mostra os *boxplots* correspondentes e as pontuações dos participantes da rodada S2.

Observando as respostas de forma individual, nota-se que os participantes em geral demonstraram de forma positiva sua disposição em utilizar o sistema com frequência (Q1) e as respostas são consistentes em S1 e S2. As questões Q2, Q3 e Q8 captam a opinião do participante

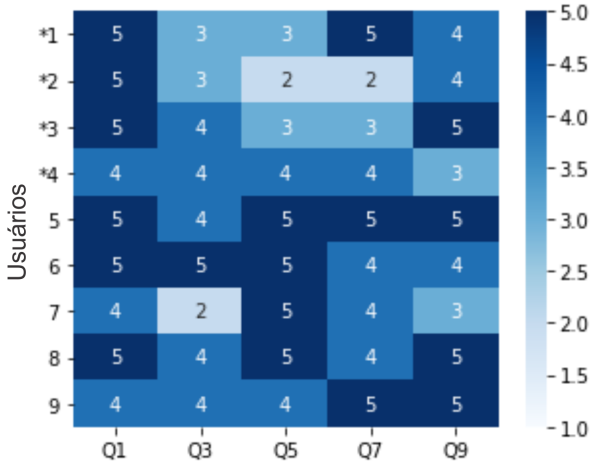
Tabela 7 – Questionário Parte 1, *System Usability Scale* (SUS). Respostas na escala Likert (1: Discordo totalmente, 2: Discordo parcialmente, 3: Indiferente, 4: Concordo parcialmente, 5: Concordo totalmente).

Questões SUS
1) I think that I would like to use this system frequently.
2) I found the system unnecessarily complex.
3) I thought the system was easy to use.
4) I think that I would need the support of a technical person to be able to use this system.
5) I found the various functions in this system were well integrated.
6) I thought there was too much inconsistency in this system.
7) I would imagine that most people would learn to use this system very quickly.
8) I found the system very cumbersome to use.
9) I felt very confident using the system.
10) I needed to learn a lot of things before I could get going with this system.

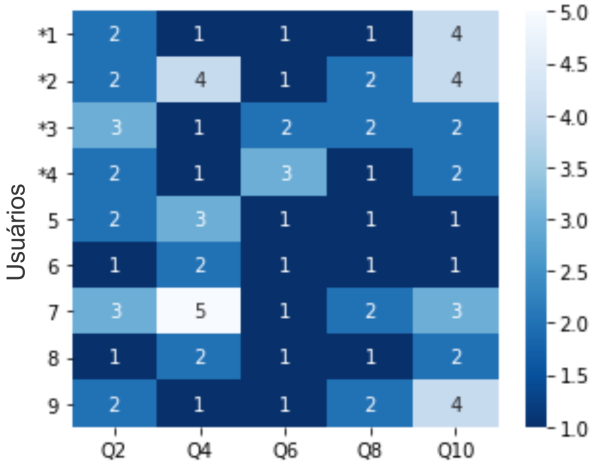
Figura 39 – Pontuações do SUS no estudo S1.



(a) Distribuição das pontuações do SUS.



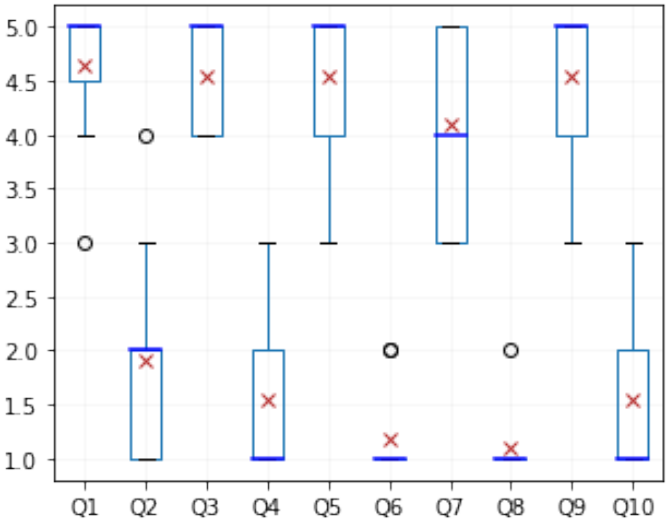
(b) Pontuações dos participantes (melhores ≈ 5)



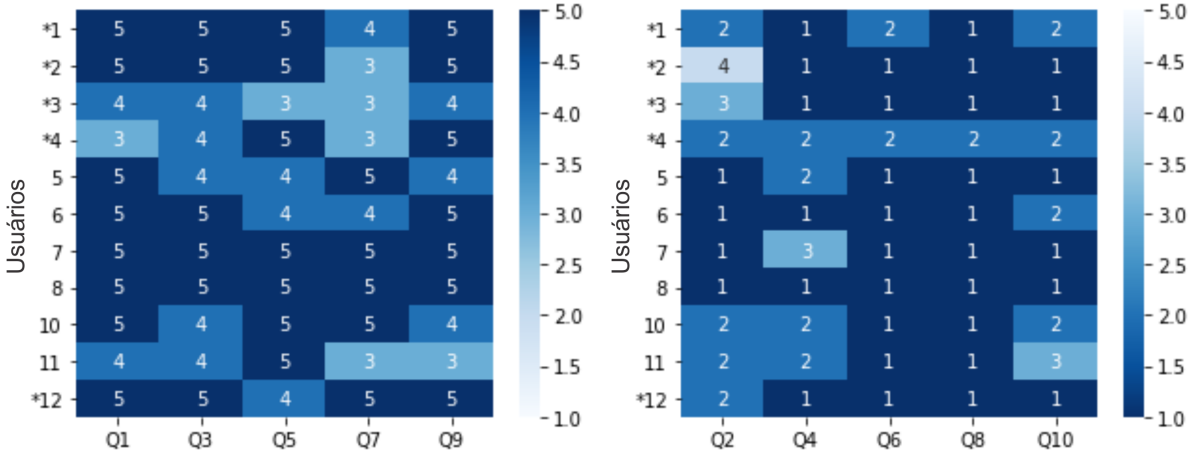
(c) Pontuações dos participantes (melhores ≈ 1)

Fonte: Elaborada pelo autor.

Figura 40 – Pontuações do SUS no estudo S2.



(a) Distribuição das pontuações do SUS.



(b) Pontuações dos participantes (melhores ≈ 5)

(c) Pontuações dos participantes (melhores ≈ 1)

Fonte: Elaborada pelo autor.

sobre a complexidade do sistema (Q2), facilidade e conforto de uso (Q3 e Q8). No S2 observa-se uma melhora nas pontuações destas questões, em relação ao S1. Os participantes mais familiarizados com as interfaces visuais permaneceram críticos no S2, atribuindo pontuações mais altas (piores) ao Q2 em ambos os estudos, embora tenham expressado opiniões muito positivas sobre a facilidade de uso (Q3) e não terem considerado o sistema complicado (Q8). Observando as avaliações atribuídas às questões Q5 e Q6, sobre a percepção geral quanto à integração das funcionalidades do sistema e consistência, também percebe-se uma melhora em relação aos resultados S1.

As perguntas Q4 e Q9 consideram o conforto e a confiança do usuário ao usar o sistema. A Questão Q4 pergunta se os usuários acreditam que a assistência de um especialista é necessária para usar o sistema. Em S1, 2 dos 9 usuários (U2, U7) sentiram mais fortemente que precisariam

de ajuda e um deles (U2) possui familiaridade com visualização e/ou IHC. A percepção destes usuários mudou em S2, embora o U7 ainda esteja moderadamente confiante. Em S2 também observa-se uma pequena melhora nas avaliações da confiança do usuário no uso do sistema (Q9).

As questões Q7 e Q10 receberam pontuações mais baixas e com maior dispersão; elas questionam se os participantes acham que precisariam investir muito tempo para aprender a usar o sistema de forma eficaz. Curiosamente, embora em S1 a maioria dos participantes acreditasse que a maioria das pessoas conseguiria aprender a usar o sistema rapidamente (Q7), essa percepção diminuiu em S2, e os mais experientes estão menos confiantes de que isso seja verdade do que os menos experientes. Além disso, vários participantes sentiram que precisavam aprender muito antes de começar (Q10), o que é compreensível, pois o uso do sistema exigiu assistência e treinamento. No entanto, também se observa que as pontuações de ambas as questões melhoraram em S2.

No geral, as pontuações de usabilidade melhoraram e os participantes acharam o TRIVIR 2.0 mais fácil de usar do que o TRIVIR 1.0 e se sentiram mais confiantes em usá-lo sem o suporte de um especialista. As pontuações mais altas nas questões de usabilidade em S2 podem refletir uma percepção melhorada resultante tanto das modificações introduzidas quanto do aumento da familiaridade com o sistema, já que a maioria dos participantes estava executando sua segunda sessão. No entanto, os usuários mais experientes parecem evidentemente cientes dos desafios introduzidos pela complexidade inerente de um sistema VA que implementa um modelo conceitual sofisticado.

CAPÍTULO

5

CONCLUSÕES

Recuperar informações em bases de dados textuais pode ser uma tarefa bastante complexa, uma vez que usuários alternam continuamente entre tarefas de recuperação e exploração a fim de obter novas perspectivas quanto às consultas a serem realizadas, a medida em que aprendem novas informações interessantes sobre o corpus. Sistemas de *Visual Analytics* associam visualizações abstratas a algoritmos de mineração de texto para aprimorar a intuição do usuário em identificar padrões, mas tais técnicas são baseadas em conceitos que podem ser difíceis de entender. Compreender como e em que medida eles realmente facilitam a recuperação é um problema de pesquisa relevante em visualização e interação humano-computador. No entanto, os sistemas são frequentemente avaliados em cenários simulados ou simplistas, e raramente por um grupo representativo de usuários reais. Neste trabalho, descrevemos os resultados da avaliação de um sistema VA específico para recuperação de informações textuais, o qual havia sido previamente submetido a uma validação preliminar com usuários em potencial, mas não havia passado por uma avaliação formal. Conduzimos estudos observacionais em um ambiente realista com uma categoria de usuários potenciais: doze pesquisadores utilizaram o sistema para realizar uma tarefa de recuperação de informação relevante a partir da exploração e identificação de material interessante em um corpus sobre um tema de seu interesse. Os participantes utilizaram o sistema com o apoio do pesquisador responsável e foram solicitados a opinar sobre aspectos de usabilidade e avaliar as funcionalidades do sistema, com base na percepção de sua utilidade.

De maneira geral, os participantes expressaram visões positivas sobre o potencial do sistema para facilitar suas tarefas de recuperação de informação e declararam-se dispostos a adotá-lo. No entanto, observamos que os participantes com formação em áreas diferentes de Ciência da Computação, e pouco familiarizados com VA, fizeram as melhores avaliações de quase todas as funcionalidades, enquanto aqueles que declararam familiaridade anterior com VA ou IHC, ou possuem formação em Ciência da Computação, ofereceram avaliações críticas para várias funcionalidades específicas do sistema. A avaliação do usuário em algumas funcionalidades variou consideravelmente, e aparentemente foi afetada por sua formação. Isso

pode sugerir que as configurações de avaliação devem considerar diferentes perfis de usuários. Em relação às visualizações específicas no TRIVIR, a interação com o mapa de similaridade do documento fornecido no *Scatterplot View* rendeu as observações mais críticas, e a maioria dos participantes preferiu interagir com o mapa de similaridade dos documentos no *Force-Layout* do que com sua contraparte criada com t-SNE.

Como conduzimos duas rodadas de sessões com praticamente o mesmo grupo de participantes, percebemos o papel fundamental desempenhado por uma introdução cuidadosa às funções de interface para transmitir os modelos subjacentes do sistema e suas limitações. Na verdade, os usuários menos familiarizados com as abstrações subjacentes podem enfrentar dificuldades em usar o sistema de forma autônoma e efetiva, ainda que possam não estar cientes disso.

Medir a efetividade da tarefa de recuperação em relação a um resultado *ground truth* exigiria uma configuração experimental diferente, mas seria uma iniciativa relevante. Neste trabalho apenas solicitamos aos participantes uma apreciação informal do resultado obtido em sua interação com o sistema: vários comentaram que tinham identificado artigos interessantes e dois pediram para ter o sistema instalado em suas próprias máquinas. Outras questões surgiram nos estudos, no entanto, não relacionadas diretamente com a interface de visualização e não foram investigadas neste trabalho. Em particular, os participantes muitas vezes desaprovaram as sugestões do classificador. Na verdade, o TRIVIR foi apresentado como uma prova de conceito de adoção do protocolo CAL para recuperação de documentos em geral. As escolhas ideais para a representação de texto subjacente e o algoritmo de classificação são possivelmente específicas do domínio e exigiriam uma investigação mais aprofundada, pois provavelmente afetariam a qualidade de visualização e recuperação e, indiretamente, a usabilidade do sistema. Um trabalho posterior também pode verificar se nossas observações são válidas em tarefas de recuperação em domínios de aplicação diferentes da revisão da literatura.

REFERÊNCIAS

ABDELGHAFAR, S.; DARWISH, A.; HASSANIEN, A. E. Intelligent health monitoring systems for space missions based on data mining techniques. In: **Machine learning and data mining in aerospace technology**. [S.l.]: Springer, 2020. p. 65–78. Citado na página 23.

AÏT-SAHALIA, Y.; XIU, D. Principal component analysis of high-frequency data. **Journal of the American Statistical Association**, Taylor & Francis, v. 114, n. 525, p. 287–303, 2019. Citado na página 34.

AL-ANZI, F. S.; ABUZEINA, D. Beyond vector space model for hierarchical arabic text classification: A markov chain approach. **Information Processing & Management**, Elsevier, v. 54, n. 1, p. 105–115, 2018. Citado na página 31.

ALENCAR, A. B.; OLIVEIRA, M. C. F. de; PAULOVICH, F. V. Seeing beyond reading: a survey on visual text analytics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 2, n. 6, p. 476–492, 2012. Citado nas páginas 23, 37 e 48.

AMERSHI, S.; CAKMAK, M.; KNOX, W. B.; KULESZA, T. Power to the people: The role of humans in interactive machine learning. **Ai Magazine**, v. 35, n. 4, p. 105–120, 2014. Citado na página 60.

AMIRI, I.; NIKOUKAR, A. Secured binary codes generation for computer network communication. **Cultural Studies**, v. 8, p. 2, 2017. Citado na página 30.

AN, X.; HUANG, J. X.; WANG, Y. Diversity and novelty in biomedical information retrieval. In: **Biomedical Information Technology**. [S.l.]: Elsevier, 2020. p. 369–396. Citado na página 27.

ARMENI, I.; SENER, O.; ZAMIR, A. R.; FISCHER, M.; SAVARESE, S. **Systems and methods for performing three-dimensional semantic parsing of indoor spaces**. [S.l.]: Google Patents, 2019. US Patent 10,424,065. Citado na página 34.

BAE, J.; HELLDIN, T.; RIVEIRO, M.; NOWACZYK, S.; BOUGUELIA, M.-R.; FALKMAN, G. Interactive clustering: A comprehensive review. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 1, fev. 2020. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3340960>. Citado na página 50.

BASCUR, J. P.; ECK, N. J. van; WALTMAN, L. An interactive visual tool for scientific literature search: Proposal and algorithmic specification. In: **BIR@ ECIR**. [S.l.: s.n.], 2019. p. 76–87. Citado na página 110.

BECK, F.; KOCH, S.; WEISKOPF, D. Visual analysis and dissemination of scientific literature collections with survis. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 180–189, 2016. Citado nas páginas 48, 49 e 110.

BELKEBIR, R.; GUESSOUM, A. A supervised approach to arabic text summarization using adaboost. In: **New contributions in information systems and technologies**. [S.l.]: Springer, 2015. p. 227–236. Citado na página 37.

BIAS, R. Interface-walkthroughs: efficient collaborative testing. **IEEE Software**, IEEE, v. 8, n. 5, p. 94–95, 1991. Citado na página 43.

BLUMENSTEIN, M.; VERMA, B.; BASLI, H. A novel feature extraction technique for the recognition of segmented handwritten characters. In: IEEE. **Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.** [S.l.], 2003. p. 137–141. Citado na página 30.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **arXiv preprint arXiv:1607.04606**, 2016. Citado na página 33.

____. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 5, p. 135–146, 2017. Citado na página 56.

BONIFÁCIO, B.; VIANA, D.; VIEIRA, S.; ARAÚJO, C.; CONTE, T. Aplicando técnicas de inspeção de usabilidade para avaliar aplicações móveis. In: **Proceedings of the IX Symposium on Human Factors in Computing Systems**. [S.l.: s.n.], 2010. p. 189–192. Citado na página 43.

BOUCINHA, R. M.; TAROUCO, L. M. R. Avaliaçao de ambiente virtual de aprendizagem com o uso do sus - system usability scale. **RENOTE**, v. 11, n. 3, 2013. Citado na página 41.

CABRAL, E. M. **Interactive keyterm-based document clustering and visualization via neural language models**. Dissertação (Mestrado) — Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, 2020. Disponível em: <https://doi.org/10.11606/D.55.2020.tde-20082020-093906>. Citado nas páginas 35 e 36.

CANTINI, L.; ZAKERI, P.; HERNANDEZ, C.; NALDI, A.; THIEFFRY, D.; REMY, E.; BAUDOT, A. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. **bioRxiv**, Cold Spring Harbor Laboratory, 2020. Citado na página 33.

CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. In: **Proceedings of the 23rd International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2006. (ICML '06), p. 161–168. ISBN 1595933832. Disponível em: <https://doi.org/10.1145/1143844.1143865>. Citado na página 37.

CHEN, C.; SONG, M. **Visualizing a Field of Research: A Methodology of Systematic Scientometric Reviews**. 2019. Citado na página 110.

CHEN, C.-H. Reducing the dimensionality of time-series data with deep learning techniques. 2019. Citado na página 37.

CHEN, K.; ZHANG, Z.; LONG, J.; ZHANG, H. Turning from tf-idf to tf-igm for term weighting in text classification. **Expert Systems with Applications**, Elsevier, v. 66, p. 245–260, 2016. Citado na página 31.

CHEONG, S.-H.; SI, Y.-W. Force-directed algorithms for schematic drawings and placement: A survey. **Information Visualization**, SAGE Publications Sage UK: London, England, v. 19, n. 1, p. 65–91, 2020. Citado nas páginas 36 e 72.

CHERAPANUKORN, V.; CHAROENKWAN, P. Word cloud of online hotel reviews in thailand for customers' satisfaction analysis. 2018. Citado na página 38.

CHI, E. H.; ROSIEN, A.; SUPATTANASIRI, G.; WILLIAMS, A.; ROYER, C.; CHOW, C.; ROBLES, E.; DALAL, B.; CHEN, J.; COUSINS, S. The bloodhound project: automating discovery of web usability issues using the infoscentπ simulator. In: **Proceedings of the SIGCHI conference on Human factors in computing systems**. [S.l.: s.n.], 2003. p. 505–512. Citado na página 43.

CHOO, J.; KIM, H.; CLARKSON, E.; LIU, Z.; LEE, C.; LI, F.; LEE, H.; KANNAN, R.; STOLPER, C. D.; STASKO, J.; PARK, H. Visirr: A visual analytics system for information retrieval and recommendation for large-scale document data. In: . New York, NY, USA: ACM, 2018. v. 12, n. 1, p. 8:1–8:20. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/3070616>. Citado nas páginas 54 e 110.

CHUANG, J.; MANNING, C. D.; HEER, J. Termite: Visualization techniques for assessing textual topic models. In: **Proceedings of the international working conference on advanced visual interfaces**. [S.l.: s.n.], 2012. p. 74–77. Citado na página 28.

CORMACK, G. V.; GROSSMAN, M. R. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2014. (SIGIR '14), p. 153–162. ISBN 9781450322577. Disponível em: <https://doi.org/10.1145/2600428.2609601>. Citado na página 29.

____. Autonomy and reliability of continuous active learning for technology-assisted review. **arXiv preprint arXiv:1504.06868**, 2015. Citado nas páginas 37 e 55.

____. Multi-faceted recall of continuous active learning for technology-assisted review. In: **Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2015. p. 763–766. Citado na página 37.

____. Scalability of continuous active learning for reliable high-recall text classification. In: **Proceedings of the 25th ACM international on conference on information and knowledge management**. [S.l.: s.n.], 2016. p. 1039–1048. Citado na página 55.

CUI, Z.; LI, F.; ZHANG, W. Bat algorithm with principal component analysis. **International Journal of Machine Learning and Cybernetics**, Springer, v. 10, n. 3, p. 603–622, 2019. Citado na página 34.

CUTTING, D. R.; KARGER, D. R.; PEDERSEN, J. O.; TUKEY, J. W. Scatter/gather: A cluster-based approach to browsing large document collections. In: ACM NEW YORK, NY, USA. **ACM SIGIR Forum**. [S.l.], 2017. v. 51, n. 2, p. 148–159. Citado na página 50.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Citado na página 33.

DIAS, A. G. **TRIVIR: A Visualization System to Support Document Retrieval with High Recall**. Dissertação (Mestrado) — Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, 2019. Disponível em: <https://doi.org/10.11606/D.55.2019.tde-11092019-090930>. Citado nas páginas 35, 58, 63, 66 e 68.

DIAS, A. G.; MILIOS, E. E.; OLIVEIRA, M. C. F. de. Trivir: A visualization system to support document retrieval with high recall. In: **Proceedings of the ACM Symposium on Document Engineering 2019**. New York, NY, USA: ACM, 2019. (DocEng '19), p. 10:1–10:10. ISBN 978-1-4503-6887-2. Disponível em: <http://doi.acm.org/10.1145/3342558.3345401>. Citado nas páginas 24, 29, 31, 38, 55, 56, 58, 69 e 110.

DRAGAN, D.; PETROVIĆ, V. B.; GAJIĆ, D. B.; ŽIVANOV, Ž.; IVETIĆ, D. An empirical study of data visualization techniques in pacs design. **Computer Science and Information Systems**, v. 16, n. 1, p. 247–271, 2019. Citado nas páginas 33 e 38.

EL-KHAIR, I. A. Effects of stop words elimination for arabic information retrieval: a comparative study. **arXiv preprint arXiv:1702.01925**, 2017. Citado na página 30.

ENDERT, A.; HOSSAIN, M. S.; RAMAKRISHNAN, N.; NORTH, C.; FIAUX, P.; ANDREWS, C. The human is the loop: new directions for visual analytics. **Journal of intelligent information systems**, Springer, v. 43, n. 3, p. 411–435, 2014. Citado na página 60.

FERRAZ, R. **Acessibilidade na web**. [S.l.]: Senac, 2017. Citado na página 40.

FRANÇA, R.; TEDESCO, P. Pensamento computacional sob a perspectiva de licenciandos em computação. In: **Anais do Workshop de Informática na Escola**. [S.l.: s.n.], 2017. v. 23, n. 1, p. 795–804. Citado na página 44.

FRICK, A.; LUDWIG, A.; MEHLDAU, H. A fast adaptive layout algorithm for undirected graphs. In: SPRINGER. **International Symposium on Graph Drawing**. [S.l.], 1994. p. 388–403. Citado na página 36.

FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. **Software: Practice and experience**, Wiley Online Library, v. 21, n. 11, p. 1129–1164, 1991. Citado nas páginas 36 e 72.

GALLAGHER, B.; REVER, M.; LOVELAND, D.; MUNDHENK, T. N.; BEAUCHAMP, B.; ROBERTSON, E.; JAMAN, G. G.; HISZPANSKI, A. M.; HAN, T. Y.-J. Predicting compressive strength of consolidated molecular solids using computer vision and deep learning. **Materials & Design**, Elsevier, p. 108541, 2020. Citado na página 33.

GAMALLO, P.; GARCIA, M. **FreeLing e TreeTagger: um estudo comparativo no âmbito do Português**. [S.l.], 2013. Citado na página 30.

Gomez-Nieto, E.; Roman, F. S.; Pagliosa, P.; Casaca, W.; Helou, E. S.; de Oliveira, M. C. F.; Nonato, L. G. Similarity preserving snippet-based visualization of web search results. **IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 3, p. 457–470, March 2014. ISSN 2160-9306. Citado nas páginas 51 e 52.

GONÇALVES, V. P.; NERIS, V. P.; MORANDINI, M.; NAKAGAWA, E. Y.; UEYAMA, J. Uma revisão sistemática sobre métodos de avaliação de usabilidade aplicados em software de telefones celulares. In: **Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction**. [S.l.: s.n.], 2011. p. 197–201. Citado na página 43.

GÖRTLER, J.; SCHULZ, C.; WEISKOPF, D.; DEUSSEN, O. Bubble treemaps for uncertainty visualization. **IEEE transactions on visualization and computer graphics**, IEEE, v. 24, n. 1, p. 719–728, 2017. Citado na página 49.

GROSSMAN, M. R.; CORMACK, G. Continuous active learning for tar. **The Journal**, v. 4, n. 3, p. 1–7, 2016. Citado na página 55.

GROSSMAN, M. R.; CORMACK, G. V. Continuous active learning for tar. practical law the journal: Litigation. p. 32–37, 2016. Citado na página 29.

____. A tour of technology-assisted review. **Perspectives on Predictive Coding and Other Advanced Search and Review Technologies for the Legal Practitioner (ABA 2016)**, 2016. Citado na página 29.

GUO, Y.; KORHONEN, A.; POIBEAU, T. A weakly-supervised approach to argumentative zoning of scientific documents. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the conference on empirical methods in natural language processing**. [S.l.], 2011. p. 273–283. Citado na página 37.

GURU, D.; SWARNALATHA, K.; KUMAR, N. V.; ANAMI, B. S. Effective technique to reduce the dimension of text data. **International Journal of Computer Vision and Image Processing (IJCVIP)**, IGI Global, v. 10, n. 1, p. 67–85, 2020. Citado na página 33.

GVILY, Y. **Snippet selection**. [S.l.]: Google Patents, 2006. US Patent 7,085,994. Citado na página 51.

HASHIMOTO, K.; KONTONATSIOS, G.; MIWA, M.; ANANIADOU, S. Topic detection using paragraph vectors to support active learning in systematic reviews. **Journal of biomedical informatics**, Elsevier, v. 62, p. 59–65, 2016. Citado na página 31.

HAVRE, S.; HETZLER, B.; NOWELL, L. Themeriver: Visualizing theme changes over time. In: IEEE. **IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings**. [S.l.], 2000. p. 115–123. Citado na página 49.

HE, J.; PING, Q.; LOU, W.; CHEN, C. Paperpoles: Facilitating adaptive visual exploration of scientific publications by citation links. **JASIST**, v. 70, p. 843–857, 2019. Citado nas páginas 52, 53 e 110.

HEINRICH, J.; WEISKOPF, D. State of the art of parallel coordinates. In: **Eurographics (State of the Art Reports)**. [S.l.: s.n.], 2013. p. 95–116. Citado na página 50.

HOULDING, S. **3D geoscience modeling: computer techniques for geological characterization**. [S.l.]: Springer Science & Business Media, 2012. Citado na página 37.

HU, Y. Efficient, high-quality force-directed graph drawing. **Mathematica journal**, Redwood City, Ca.: Advanced Book Program, Addison-Wesley Pub. Co., c1990-, v. 10, n. 1, p. 37–71, 2005. Citado na página 36.

HUANG, Y.; WANG, Y.; YE, F. A study of the application of word cloud visualization in college english teaching. **International Journal of Information and Education Technology**, v. 9, n. 2, 2019. Citado na página 38.

HURTER, C. Image-based information visualization techniques. In: . [S.l.: s.n.], 2018. Citado na página 37.

INSELBERG, A. The plane with parallel coordinates. **The visual computer**, Springer, v. 1, n. 2, p. 69–91, 1985. Citado na página 50.

INSELBERG, A.; DIMSDALE, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: IEEE. **Proceedings of the First IEEE Conference on Visualization: Visualization90**. [S.l.], 1990. p. 361–378. Citado na página 50.

ISENBERG, T.; ISENBERG, P.; CHEN, J.; SEDLMAIR, M.; MÖLLER, T. A systematic review on the practice of evaluating visualization. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 19, n. 12, p. 2818–2827, 2013. Citado na página 58.

JADHAV, D. V.; HOLAMBE, R. S. Radon and discrete cosine transforms based feature extraction and dimensionality reduction approach for face recognition. **Signal Processing**, Elsevier, v. 88, n. 10, p. 2604–2609, 2008. Citado na página 33.

JANAKIRAMAIAH, B.; KALYANI, G.; NARAYANA, S.; KRISHNA, T. B. M. Reducing dimensionality of data using autoencoders. In: **Smart Intelligent Computing and Applications**. [S.l.]: Springer, 2020. p. 51–58. Citado na página 33.

JAPIASSÚ, H.; MARCONDES, D. Dicionário de filosofia. **Rio de**, 1996. Citado na página 45.

JARDINE, N.; RIJSBERGEN, C. J. van. The use of hierarchic clustering in information retrieval. **Information storage and retrieval**, Elsevier, v. 7, n. 5, p. 217–240, 1971. Citado na página 49.

JARROUSH, J.; KHELL, B.; SEGEV, I.; HAKIM, Y. S. **System and method for automatically acquiring two-dimensional images and three-dimensional point cloud data of a field to be surveyed**. [S.l.]: Google Patents, 2019. US Patent App. 16/346,600. Citado na página 33.

JAYASHANKAR, S.; SRIDARAN, R. Superlative model using word cloud for short answers evaluation in elearning. **Education and Information Technologies**, Springer, v. 22, n. 5, p. 2383–2402, 2017. Citado na página 38.

JEMISON, J. M.; WELCOMER, S.; KERSBERGEN, R.; MAJEWSKI, C. Word cloud analysis of early adopter no-till farmer interviews. **Journal of Extension**, v. 56, n. 3, p. 11, 2018. Citado na página 38.

JENTNER, W.; KEIM, D. A. Visualization and visual analytic techniques for patterns. In: **High-Utility Pattern Mining**. [S.l.]: Springer, 2019. p. 303–337. Citado na página 33.

JI, S.; KRISHNAPURAM, B.; CARIN, L. Variational bayes for continuous hidden markov models and its application to active learning. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 28, n. 4, p. 522–532, 2006. Citado na página 55.

JOLLIFFE, I. **Principal Component Analysis**. [S.l.]: Springer Verlag, 1986. Citado na página 34.

____. **Principal Component Analysis**. 2nd. ed. New York: Springer-Verlag, 2002. ISSN 0172-7397. ISBN 0-387-95442-2. Disponível em: <http://link.springer.com/10.1007/b98835>. Citado na página 34.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, v. 1, n. 82, p. 1–10, 1972. Citado na página 56.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; MIKOLOV, T. Bag of tricks for efficient text classification. **arXiv preprint arXiv:1607.01759**, v. 1, p. 1–5, 2016. Citado nas páginas 33 e 56.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; DOUZE, M.; JÉGOU, H.; MIKOLOV, T. Fasttext.zip: Compressing text classification models. **arXiv preprint arXiv:1612.03651**, 2016. Citado nas páginas 33 e 56.

KASER, O.; LEMIRE, D. Tag-cloud drawing: Algorithms for cloud visualization. **arXiv preprint cs/0703109**, 2007. Citado na página 49.

KAUR, J.; BUTTAR, P. K. A systematic review on stopword removal algorithms. **Int. J. Future Revolut. Comput. Sci. Commun. Eng**, v. 4, n. 4, 2018. Citado na página 30.

KEIM, D. A. Information visualization and visual data mining. **IEEE transactions on Visualization and Computer Graphics**, IEEE, v. 8, n. 1, p. 1–8, 2002. Citado na página 23.

KEIM, D. A.; MANSMANN, F.; SCHNEIDEWIND, J.; THOMAS, J.; ZIEGLER, H. Visual analytics: Scope and challenges. In: **Visual data mining**. [S.l.]: Springer, 2008. p. 76–90. Citado na página 23.

KOBAK, D.; LINDERMAN, G. C. Umap does not preserve global structure any better than t-sne when using the same initialization. **bioRxiv**, Cold Spring Harbor Laboratory, 2019. Citado na página 34.

KOSS, A. R.; CANAGARATNA, M. R.; ZAYTSEV, A.; KRECHMER, J. E.; BREITENLECHNER, M.; NIHILL, K. J.; LIM, C. Y.; ROWE, J. C.; ROSCIOLI, J. R.; KEUTSCH, F. N. *et al.* Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments. **Atmospheric Chemistry and Physics**, Copernicus GmbH, v. 20, n. 2, p. 1021–1041, 2020. Citado na página 33.

KUCHER, K.; KERREN, A. Text visualization browser: A visual survey of text visualization techniques. **Poster Abstracts of IEEE VIS**, v. 2014, 2014. Citado na página 37.

LAM, H.; BERTINI, E.; ISENBERG, P.; PLAISANT, C.; CARPENDALE, S. Empirical studies in information visualization: Seven scenarios. **IEEE transactions on visualization and computer graphics**, IEEE, v. 18, n. 9, p. 1520–1536, 2011. Citado na página 58.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **International conference on machine learning**. [S.l.: s.n.], 2014. p. 1188–1196. Citado na página 33.

LEHMANN, D. J.; THEISEL, H. General projective maps for multidimensional data projection. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2016. v. 35, n. 2, p. 443–453. Citado na página 33.

LEITÃO, C. F.; PRATES, R. O. A aplicação de métodos qualitativos em computação. **Jornadas de Atualização em Informática**, v. 2017, p. 43–90, 2017. Citado nas páginas 44 e 45.

LINDERMAN, G. C.; RACHH, M.; HOSKINS, J. G.; STEINERBERGER, S.; KLUGER, Y. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. **Nature methods**, Nature Publishing Group, v. 16, n. 3, p. 243–245, 2019. Citado na página 34.

LINDERMAN, G. C.; STEINERBERGER, S. Clustering with t-sne, provably. **SIAM Journal on Mathematics of Data Science**, SIAM, v. 1, n. 2, p. 313–332, 2019. Citado na página 34.

LOPES, M. A. D. S.; NETO, A. D. D.; MARTINS, A. D. M. Parallel t-sne applied to data visualization in smart cities. **IEEE Access**, IEEE, v. 8, p. 11482–11490, 2020. Citado na página 34.

LUZ, L. P. da; CONEGLIAN, C. S.; SEGUNDO, J. E. S. Tecnologias da web semântica para a recuperação da informação no wikidata. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, v. 17, p. e019003–e019003, 2019. Citado na página 27.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. Nov, p. 2579–2605, 2008. Disponível em: <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>. Citado nas páginas 34, 35 e 71.

MACHADO, J. T.; LOPES, A. M. Multidimensional scaling and visualization of patterns in prime numbers. **Communications in Nonlinear Science and Numerical Simulation**, Elsevier, v. 83, p. 105128, 2020. Citado na página 35.

MACIEL, C.; NOGUEIRA, J. L. T.; CIUFFO, L. N.; GARCIA, A. C. B. Avaliação heurística de sítios na web. **VII ESCOLA DE INFORMÁTICA DO SBC-CENTROOESTE**, 2004. Citado na página 43.

MACK, R. L.; NIELSEN, J. Usability inspection methods: Executive summary. In: **Readings in Human–Computer Interaction**. [S.l.]: Elsevier, 1995. p. 170–181. Citado na página 41.

MACKAY, K.; KUSALIK, A. Stohi-c: Using t-distributed stochastic neighbor embedding (t-sne) to predict 3d genome structure from hi-c data. **bioRxiv**, Cold Spring Harbor Laboratory, 2020. Citado na página 34.

MADAAN, R.; BHATIA, K. K. Prevalence of visualization techniques in data mining. In: **Data Visualization and Knowledge Engineering**. [S.l.]: Springer, 2020. p. 273–298. Citado na página 23.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. [S.l.]: Cambridge university press, 2008. Citado na página 31.

MATERA, M.; RIZZO, F.; CARUGHI, G. T. Web usability: Principles and evaluation methods. In: **Web engineering**. [S.l.]: Springer, 2006. p. 143–180. Citado na página 43.

MAZZA, R. **Introduction to information visualization**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 41.

MCGOWAN, B.; CHANEY, U. Creatively facilitating reflection and learning using wordclouds and social media. In: **38th Annual International Nursing & Midwifery Research and Education Conference 2019: The Future of Nursing and Midwifery Practice, Education and Research'.** [S.l.: s.n.], 2019. Citado na página 38.

MELO, F.; MARTINS, B. Automated geocoding of textual documents: A survey of current approaches. **Transactions in GIS**, Wiley Online Library, v. 21, n. 1, p. 3–38, 2017. Citado na página 30.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citado nas páginas 32 e 33.

MIKOLOV, T.; YIH, S. W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human**

Language Technologies (NAACL-HLT-2013). Association for Computational Linguistics, 2013. Disponível em: <https://www.microsoft.com/en-us/research/publication/linguistic-regularities-in-continuous-space-word-representations/>. Citado na página 33.

MITRA, B.; NALISNICK, E.; CRASWELL, N.; CARUANA, R. A dual embedding space model for document ranking. **arXiv preprint arXiv:1602.01137**, 2016. Citado na página 31.

MOGGRIDGE, B.; ATKINSON, B. **Designing interactions**. [S.l.]: MIT press Cambridge, MA, 2007. v. 17. Citado na página 40.

MOUMANE, K.; IDRI, A.; ABRAN, A. Usability evaluation of mobile applications using iso 9241 and iso 25062 standards. **SpringerPlus**, Springer, v. 5, n. 1, p. 1–15, 2016. Citado na página 40.

MUNZNER, T. A nested model for visualization design and validation. **IEEE transactions on visualization and computer graphics**, IEEE, v. 15, n. 6, p. 921–928, 2009. Citado nas páginas 58 e 59.

____. **Visualization analysis and design**. [S.l.]: CRC press, 2014. Citado na página 59.

NIELSEN, J. **Usability engineering**. [S.l.]: Morgan Kaufmann, 1994. Citado nas páginas 64, 78 e 82.

____. **Multimídia e hipertexto: A Internet e além**. [S.l.]: Morgan Kaufmann, 1995. Citado nas páginas 40 e 41.

NIELSEN, J.; MOLICH, R. Heuristic evaluation of user interfaces. In: **Proceedings of the SIGCHI conference on Human factors in computing systems**. [S.l.: s.n.], 1990. p. 249–256. Citado nas páginas 41 e 42.

NILASHI, M.; IBRAHIM, O.; BAGHERIFARD, K. A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. **Expert Systems with Applications**, Elsevier, v. 92, p. 507–520, 2018. Citado na página 33.

NOACK, A. Modularity clustering is force-directed layout. **Physical Review E**, APS, v. 79, n. 2, p. 026102, 2009. Citado na página 50.

NONATO, L. G.; AUPETIT, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 25, n. 8, p. 2650–2673, 2018. Citado na página 33.

NOURASHRAFEDDIN, S.; SHERKAT, E.; MINGHIM, R.; MILIOS, E. E. A visual approach for interactive keyterm-based clustering. **ACM Trans. Interact. Intell. Syst.**, ACM, New York, NY, USA, v. 8, n. 1, p. 6:1–6:35, fev. 2018. ISSN 2160-6455. Disponível em: <http://doi.acm.org/10.1145/3181669>. Citado na página 50.

NYMAN, D. **A Study of Isomap Extensions and Topological Data Analysis with Applications to Financial Data**. 2019. Citado na página 33.

OLIVEIRA, M. C. Ferreira de; LEVKOWITZ, H. From visual data exploration to visual data mining: A survey. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, USA, v. 9, n. 3, p. 378–394, jul. 2003. ISSN 1077-2626. Disponível em: <https://doi.org/10.1109/TVCG.2003.1207445>. Citado na página 23.

OLIVEIRA, M. R. de; SILVA, C. G. da. Adapting heuristic evaluation to information visualization-a method for defining a heuristic set by heuristic grouping. In: SCITEPRESS. **International Conference on Information Visualization Theory and Applications**. [S.l.], 2017. v. 4, p. 225–232. Citado na página 41.

PAREJO, U. T.; CAMPAÑA, J. R.; VILA, M. A.; DELGADO, M. A survey of tag clouds as tools for information retrieval and content representation. **Information Visualization**, SAGE Publications Sage UK: London, England, v. 20, n. 1, p. 83–97, 2021. Citado na página 39.

PATHAK, S.; PATHAK, S. Data visualization techniques, model and taxonomy. In: **Data Visualization and Knowledge Engineering**. [S.l.]: Springer, 2020. p. 249–271. Citado na página 23.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 14, n. 3, p. 564–575, 2008. Citado nas páginas 34, 36 e 71.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Empirical Methods in Natural Language Processing (EMNLP)**. [s.n.], 2014. p. 1532–1543. Disponível em: <http://www.aclweb.org/anthology/D14-1162>. Citado na página 33.

PETERSON, L. E. K-nearest neighbor. **Scholarpedia**, v. 4, n. 2, p. 1883, 2009. Citado na página 55.

PHAM, P. G. **Interactive visual data query & exploration: techniques for visual data analytics through visual query modelling and multidimensional data interaction**. Tese (Doutorado) — University of Technology Sydney, 2018. Disponível em: <http://hdl.handle.net/10453/123228>. Citado na página 48.

PHILIP, R. K. Word cloud analysis and single word summarisation as a new paediatric educational tool: Results of a neonatal application. **Journal of Paediatrics and Child Health**, Wiley Online Library, 2020. Citado na página 38.

PIENTA, R.; HOHMAN, F.; ENDERT, A.; TAMERSOY, A.; ROUNDY, K.; GATES, C.; NAVATHE, S.; CHAU, D. H. Vigor: interactive visual exploration of graph query results. **IEEE transactions on visualization and computer graphics**, IEEE, v. 24, n. 1, p. 215–225, 2017. Citado na página 48.

PLEUSS, A.; RABISER, R.; BOTTERWECK, G. Visualization techniques for application in interactive product configuration. In: **Proceedings of the 15th International Software Product Line Conference, Volume 2**. [S.l.: s.n.], 2011. p. 1–8. Citado na página 28.

RAMSAY, N.; WAMPLER, T. **Visualization and interaction with financial data using sunburst visualization**. [S.l.]: Google Patents, 2015. US Patent 9,021,397. Citado na página 49.

RASTOGI, N.; VERMA, P.; KUMAR, P. Ontological design of information retrieval model for real estate documents. In: **Microservices in Big Data Analytics**. [S.l.]: Springer, 2020. p. 73–85. Citado na página 27.

RAVICHANDRAN, T.; MOHANTA, K.; NALINI, C. An efficient approach for data preprocessing by using improved stop word removal algorithm. **International Journal of Pure and Applied Mathematics**, v. 119, n. 16, p. 231–237, 2018. Citado na página 30.

REINBOLD, C.; KUMPF, A.; WESTERMANN, R. Visualizing the stability of 2d point sets from dimensionality reduction techniques. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2019. Citado na página 33.

RIEHMANN, P.; HANFLER, M.; FROEHLICH, B. Interactive sankey diagrams. In: IEEE. **IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.** [S.l.], 2005. p. 233–240. Citado na página 38.

RITZ, T.; BAUES, J.; KRENKEL, O.; SCHIRMACHER, P.; LONGERICH, T. Multispectral imaging to define morpho-molecular classes of human hcc. **Zeitschrift für Gastroenterologie**, Georg Thieme Verlag KG, v. 58, n. 01, p. 4–44, 2020. Citado na página 34.

ROCHA, H. V. da; BARANAUSKAS, M. C. C. Design e avaliação de interfaces humano-computador. **Campinas: Unicamp**, 2003. Citado nas páginas 39, 40 e 42.

RODRIGUES, K. R.; CANAL, M. C.; XAVIER, R. A.; ALENCAR, T. S.; NERIS, V. P. Avaliando aspectos de privacidade no facebook pelas lentes de usabilidade, acessibilidade e fatores emocionais. In: **Companion Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems**. [S.l.: s.n.], 2012. p. 75–76. Citado na página 41.

ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2323–2326, 2000. Citado na página 34.

RUBENS, N.; ELAHI, M.; SUGIYAMA, M.; KAPLAN, D. Active learning in recommender systems. In: ____. **Recommender Systems Handbook**. Boston, MA: Springer US, 2015. p. 809–846. ISBN 978-1-4899-7637-6. Disponível em: <https://doi.org/10.1007/978-1-4899-7637-6_24>. Citado na página 28.

SAMMON, J. W. A nonlinear mapping for data structure analysis. **IEEE Transactions on computers**, Ieee, v. 100, n. 5, p. 401–409, 1969. Citado na página 34.

SAUL, L. K.; ROWEIS, S. T. An introduction to locally linear embedding. **unpublished. Available at: http://www. cs. toronto. edu/~ roweis/lle/publications. html**, 2000. Citado na página 34.

SCHIRRMEISTER, R. T.; GEMEIN, L.; EGGENSPERGER, K.; HUTTER, F.; BALL, T. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. **arXiv preprint arXiv:1708.08012**, 2017. Citado na página 37.

SCHÖNFELD, M.; PFEFFER, J. Fruchterman/reingold (1991): Graph drawing by force-directed placement. In: **Schlüsselwerke der Netzwerkforschung**. [S.l.]: Springer, 2019. p. 217–220. Citado na página 36.

SETTLES, B. **Active learning literature survey**. [S.l.], 2009. Citado na página 28.

SHAHABI, H.; SHIRZADI, A.; GHADERI, K.; OMIDVAR, E.; AL-ANSARI, N.; CLAGUE, J. J.; GEERTSEMA, M.; KHOSRAVI, K.; AMINI, A.; BAHRAMI, S. *et al.* Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach:

Hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 12, n. 2, p. 266, 2020. Citado na página 55.

SHAO, H.; JIANG, H.; LI, X.; LIANG, T. Rolling bearing fault detection using continuous deep belief network with locally linear embedding. **Computers in Industry**, Elsevier, v. 96, p. 27–39, 2018. Citado na página 33.

SHERKAT, E.; MILIOS, E. E.; MINGHIM, R. A visual analytics approach for interactive document clustering. **ACM Transactions on Interactive Intelligent Systems (TiiS)**, ACM, New York, NY, USA, v. 10, n. 1, p. 1–33, ago. 2019. ISSN 2160-6455. Disponível em: <http://doi.acm.org/10.1145/3241380>. Citado na página 50.

SHERKAT, E.; NOURASHRAFEDDIN, S.; MILIOS, E. E.; MINGHIM, R. Interactive document clustering revisited: A visual analytics approach. In: **23rd International Conference on Intelligent User Interfaces**. New York, NY, USA: ACM, 2018. (IUI '18), p. 281–292. ISBN 978-1-4503-4945-1. Disponível em: <http://doi.acm.org/10.1145/3172944.3172964>. Citado nas páginas 38, 50, 51, 72 e 110.

SI, H.; SUN, C.; QIAO, H.; LI, Y. Application of improved multidimensional spatial data mining algorithm in agricultural informationization. **Journal of Intelligent & Fuzzy Systems**, IOS Press, v. 38, n. 2, p. 1359–1369, 2020. Citado na página 27.

SILVA, C.; RIBEIRO, B. The importance of stop word removal on recall values in text categorization. In: IEEE. **Proceedings of the International Joint Conference on Neural Networks, 2003.** [S.l.], 2003. v. 3, p. 1661–1666. Citado na página 30.

SILVA, D. E.; SOBRINHO, M. C.; VALENTIM, N. M. Educação 4.0: um estudo de caso com atividades de computação desplugada na amazônia brasileira. **Anais do Computer on the Beach**, v. 11, n. 1, p. 141–147, 2020. Citado na página 44.

SINGH, J.; GUPTA, V. Text stemming: Approaches, applications, and challenges. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 49, n. 3, p. 1–46, 2016. Citado na página 30.

SONI, R.; SHARMA, S.; FAGNA, H.; MITTAL, S. *et al.* News analysis using word cloud. In: **Advances in Signal Processing and Communication**. [S.l.]: Springer, 2019. p. 55–64. Citado na página 38.

SOUZA, D. d. M. *et al.* Um estudo qualitativo sobre escolha inicial de carreira de concluintes em um curso de ciência da computação. Universidade Federal de Campina Grande, 2019. Citado na página 44.

SPERRLE, F.; EL-ASSADY, M.; GUO, G.; BORGO, R.; CHAU, D. H.; ENDERT, A.; KEIM, D. A survey of human-centered evaluations in human-centered machine learning. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2021. v. 40, n. 3, p. 543–567. Citado na página 60.

TEER, J. V. A lematização de expressões idiomáticas em dicionários para aprendizes. **Domínios de Lingu@ gem**, v. 12, n. 4, p. 2363–2383, 2018. Citado na página 30.

TENENBAUM, J. B. Mapping a manifold of perceptual observations. In: **Advances in neural information processing systems**. [S.l.: s.n.], 1998. p. 682–688. Citado na página 34.

TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. Citado na página 34.

THOMAS, J.; COOK, K. **Illuminating the Path: Research and Development Agenda for Visual Analytics.** [S.l.]: IEEE National Visualization and Analytics Center, 2005. Citado na página 23.

____. A visual analytics agenda. **IEEE Computer Graphics and Applications**, v. 26, n. 1, p. 10–13, 2006. Citado na página 23.

TOMINSKI, C.; GLADISCH, S.; KISTER, U.; DACHSELT, R.; SCHUMANN, H. Interactive lenses for visualization: An extended survey. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2017. v. 36, n. 6, p. 173–200. Citado na página 48.

TORGERSON, W. S. Multidimensional scaling: I. theory and method. **Psychometrika**, Springer, v. 17, n. 4, p. 401–419, 1952. Citado na página 35.

TRAVIS, H. Estimating the economic impact of mass digitization projects on copyright holders: Evidence from the google book search litigation. **J. Copyright Soc'y USA**, HeinOnline, v. 57, p. 907, 2009. Citado na página 51.

TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. **Journal of artificial intelligence research**, v. 37, p. 141–188, 2010. Citado na página 32.

UZNAŃSKI, P. Approximating text-to-pattern distance via dimensionality reduction. **arXiv preprint arXiv:2002.03459**, 2020. Citado na página 33.

VASCONCELOS, V.; ANDRADE, E. Análise da evasão de alunos na licenciatura em computação. In: SBC. **Anais do XXVI Workshop sobre Educação em Computação**. [S.l.], 2018. Citado na página 44.

VERMEULEN, A. F. Supervised learning: Using labeled data for insights. In: ____. **Industrial Machine Learning: Using Artificial Intelligence as a Transformational Disruptor**. Berkeley, CA: Apress, 2020. p. 63–136. ISBN 978-1-4842-5316-8. Disponível em: <https://doi.org/10.1007/978-1-4842-5316-8_4>. Citado na página 28.

WALLACH, J. D.; GONSALVES, G. S.; ROSS, J. S. Research, regulatory, and clinical decision-making: the importance of scientific integrity. **Journal of clinical epidemiology**, Elsevier, v. 93, p. 88–93, 2018. Citado na página 48.

WANG, C. J.; FANG, H.; WANG, H. Esammon: A computationaly enhanced sammon mapping based on data density. In: IEEE. **2016 International Conference on Computing, Networking and Communications (ICNC)**. [S.l.], 2016. p. 1–5. Citado na página 33.

WANG, F.; SUN, J. Survey on distance metric learning and dimensionality reduction in data mining. **Data mining and knowledge discovery**, Springer, v. 29, n. 2, p. 534–564, 2015. Citado na página 37.

WANG, Z. J.; WANG, V. Y.; GAMAGE, T. P. B.; RAJAGOPAL, V.; CAO, J. J.; NIELSEN, P. M.; BRADLEY, C. P.; YOUNG, A. A.; NASH, M. P. Efficient estimation of load-free left ventricular geometry and passive myocardial properties using principal component analysis. **International Journal for Numerical Methods in Biomedical Engineering**, Wiley Online Library, p. e3313, 2020. Citado na página 34.

WANICHAVORAPONG, N.; YUSOF, A. F. *et al.* Extracting factors of physical activity tracking technology using wordcloud and relevancy ranking. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2019. v. 1196, n. 1, p. 012015. Citado na página 38.

WU, Y.; CAO, N.; GOTZ, D.; TAN, Y.-P.; KEIM, D. A. A survey on visual analytics of social media data. **IEEE Transactions on Multimedia**, IEEE, v. 18, n. 11, p. 2135–2148, 2016. Citado na página 48.

XIAO, R.; HAO, Q.; WANG, C.; CAI, R.; ZHANG, L. **Snippet extraction and ranking**. [S.l.]: Google Patents, 2015. US Patent 8,954,425. Citado na página 51.

XU, K.; ZHANG, L.; PÉREZ, D.; NGUYEN, P. H.; OGILVIE-SMITH, A. Evaluating interactive visualization of multidimensional data projection with feature transformation. **Multimodal Technologies and Interaction**, Multidisciplinary Digital Publishing Institute, v. 1, n. 3, p. 13, 2017. Citado na página 33.

XU, T.; YANG, J.; GOU, G. A force-directed algorithm for drawing directed graphs symmetrically. **Mathematical Problems in Engineering**, Hindawi, v. 2018, 2018. Citado na página 36.

YANG, Y.; YAO, Q.; QU, H. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. **Visual Informatics**, v. 1, n. 1, p. 40 – 47, 2017. ISSN 2468-502X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2468502X17300074>. Citado nas páginas 48, 50 e 110.

ZHANG, Y.; WANG, X.; WEI, H.; GE, G. Em códigos de matriz de recuperação de informações privadas. **Transações IEEE sobre Teoria da Informação**, v. 65, n. 9, p. 5565–5573. Citado na página 27.

ZHOU, B.; JIN, W. Visualization of single cell rna-seq data using t-sne in r. In: **Stem Cell Transcriptional Networks**. [S.l.]: Springer, 2020. p. 159–167. Citado na página 34.

APÊNDICE

A

ESTADO DA ARTE EM SISTEMAS DE RI E VA

Para a exploração dos trabalhos relacionados foram definidos os seguintes pontos a serem investigados:

1. **Entrada:** o trabalho foi validado utilizando documentos textuais para:
   1.1. Propósito geral (e.g Tweets, Notícias, etc)
   1.2. Revisão da literatura (e.g Artigos)
2. **Consulta:** O método de consulta utilizado é (são):
   2.1. Palavras chave ou Tópicos.
   2.2. Documentos exemplo.
3. **Representação:** qual a representação dos dados?
   3.1. TF-IDF - VSM
   3.2. TF-IDF - *Latent Dirichlet Allocation* (LDA)
   3.3. TF-IDF - *Non-Negative Matrix Factorization* (NMF)
   3.4. TF-IDF - *Hierarchical Latent Tree Model* (HLTM)
   3.5. *Word Embeddings* - FastText
   3.6. *Citation Chaining*
4. **Agrupamentos:** há utilização de algoritmos de *Clustering*?
   4.1. K-means
   4.2. iK-means

4.3. Louvain

4.4. LDC

4.5. Leiden

4.6. NMF

4.7. LDA

4.8. *Information Bottleneck* (IB)

4.9. *Scatter/Gather*

5. **Redução da Dimensionalidade:** há utilização de técnicas de projeção multidimensional?

5.1. PCA

5.2. t-SNE

5.3. MDS

5.4. LSP

5.5. *Extended* t-SNE (t-SNE + *Force Layout*)

5.6. *Linear Discriminant Analysis* (LDA)

6. **Visualização:** Qual (is) técnica (s) de visualização utilizada (s)?

6.1. *List-Based* (*Ranking* de relevância)

6.2. *Scatterplot*

6.3. *Bubble-Based*

6.4. *Force-Layout*

6.5. *Tree*

6.6. *Sunburst Diagram*

6.7. *Theme River*

6.8. *Parallel Coordinates*

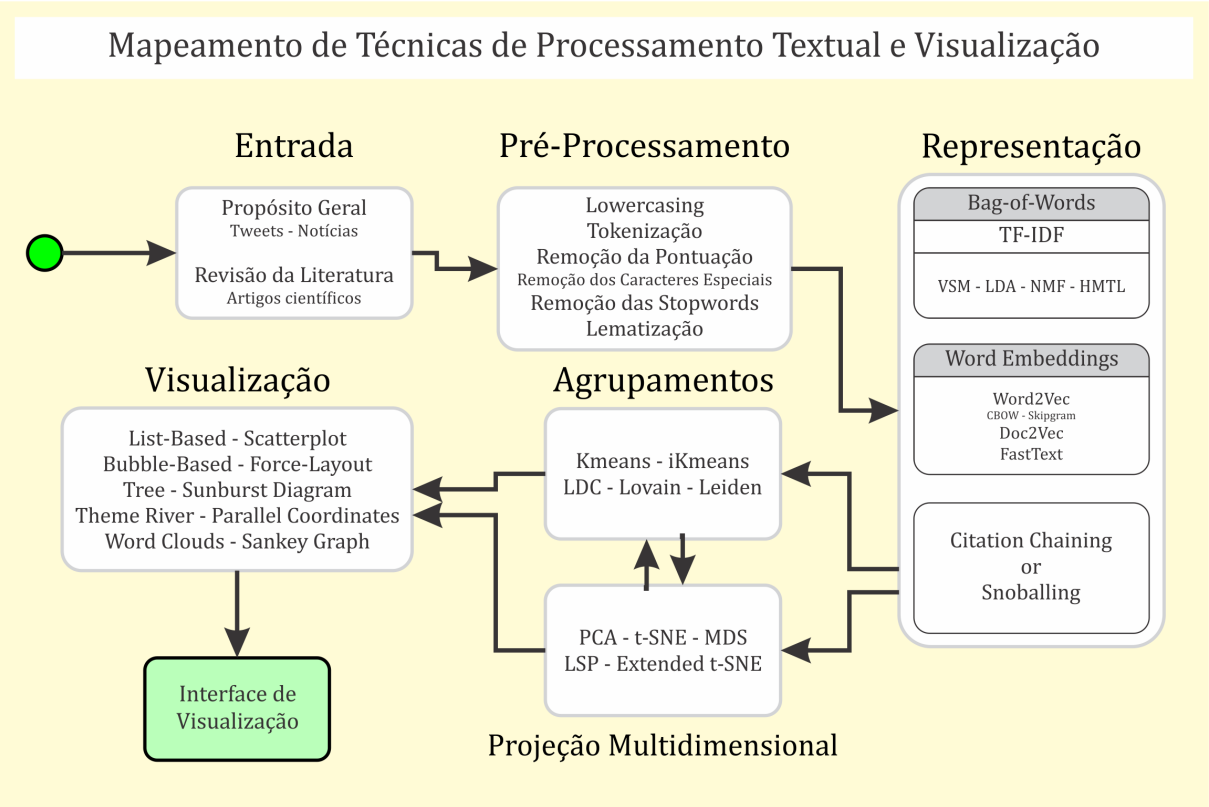
6.9. *Word Clouds*

6.10. *Timeline*

6.11. *Network*

Todas as técnicas identificadas nos trabalhos relacionados foram mapeadas no *Workflow* apresentado na Figura 41, bem como foram destacadas para cada trabalho individualmente na Tabela 8.

Figura 41 – Mapeamento de técnicas de processamento textual e visualização de sistemas de RI e VA.



Fonte: Elaborada pelo autor.

Tabela 8 – Aspectos analisados nos artigos

Artigo	Entrada	Consulta	Representação	Agrupamentos	Projeção	Visualização
(BECK; KOCH; WEISKOPF, 2016)	[1.2.]	[2.1.], [2.2.]	[3.6.]	[4.1.]	-	[6.1.], [6.9.], [6.10.]
(YANG; YAO; QU, 2017)	[1.2.]	[2.1.]	[3.4.]	-	-	[6.3.], [6.4.], [6.6.], [6.7.]
(CHOO *et al.*, 2018)	[1.2.]	[2.1.]	[3.1.], [3.2.], [3.3.]	[4.1.], [4.6.], [4.7.], [4.8.]	[5.6.]	[6.1.], [6.2.], [6.3.]
(SHERKAT *et al.*, 2018)	[1.1.], [1.2.]	[2.1.]	[3.1.]	[4.1.], [4.2.], [4.4.]	[5.5.]	[6.4.], [6.8.], [6.9.]
(CHEN; SONG, 2019)	[1.2.]	[2.1.], [2.2.]	[3.6.]	[4.9.]	-	[6.1.], [6.5.], [6.11.]
(BASCUR; ECK; WALTMAN, 2019)	[1.2.]	-	[3.6.]	[4.5.], [4.9.]	-	[6.3.]
(DIAS; MILIOS; OLIVEIRA, 2019)	[1.2.]	[2.1.], [2.2.]	[3.1.], [3.5.]	-	[5.2.], [5.4.]	[6.1.], [6.2.]
(HE *et al.*, 2019)	[1.2.]	[2.2.]	[3.6.]	[4.3.]	-	[6.1.], [6.3.], [6.4.]

ANEXO

A

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

A seguir encontra-se o Termo de Consentimento Livre esclarecido apresentado à cada participante dos estudos conduzidos, logo antes de cada início de sessão.

UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO E MATEMÁTICA COMPUTACIONAL

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

(Resolução 466/2012 do CNS)

Projeto de Pesquisa:
Visualização no Apoio à Recuperação de Informação em Coleções de Documentos

Eu, **Sherlon Almeida da Silva**, aluno do Programa de Pós Graduação em Ciência da Computação e Matemática Computacional da Universidade de São Paulo – USP o (a) convido a participar da pesquisa "**Estudos em Visualização no Apoio à Recuperação de Informação em Coleções de Documentos**", conduzida sob a orientação da Profª Drª Maria Cristina Ferreira de Oliveira, docente no Instituto de Ciências Matemáticas e Computação (ICMC) da Universidade de São Paulo (USP), em São Carlos/SP.

A necessidade de recuperar informações em repositórios de documentos textuais tem se mostrado essencial em diversas áreas do conhecimento. Por exemplo, pesquisadores desejam identificar documentos relevantes para seu tema de pesquisa, médicos ou advogados buscam relatos relacionados a um caso específico, entre outros cenários. A recuperação de informação relevante em diversos cenários têm em comum a necessidade de explorar e investigar conteúdos em coleções de documentos textuais que satisfaçam as necessidades de informação de um usuário. Trata-se de um problema desafiador, principalmente em situações em que o usuário precisa garantir alta cobertura, isto é, em que o processo de recuperação deve garantir que nenhum (ou muito poucos) documentos relevantes sejam ignorados. Técnicas de *Visual Analytics* (VA), baseadas em representações visuais interativas, podem auxiliar usuários nestas tarefas. Entretanto, o emprego dessas técnicas no contexto de sistemas de Recuperação de Informação (RI) precisa ser investigado quanto à sua adequação às reais necessidades e objetivos do usuário.

Este projeto de pesquisa propõe a realização de um estudo sobre a perspectiva de usuários quanto ao uso de técnicas de VA atualmente empregadas em um sistema de RI em coleções de documentos (TRIVIR). Para isso, pretende-se conduzir sessões de múltiplos usuários com um sistema existente.

Rubrica	
Pesquisador Responsável	***Participante da Pesquisa***

Você está sendo convidado (a) a participar desta pesquisa como um potencial usuário de um sistema dessa natureza. Ao aceitar participar desta pesquisa, o estudo será conduzido da seguinte forma. Primeiramente, você será convidado a realizar uma sessão de interação com o sistema TRIVIR utilizando uma coleção de textos do seu interesse, na expectativa de simular um cenário de busca por documentos relevantes para a sua necessidade de informação. O estudo será conduzido pelo pesquisador responsável, que irá apresentar-lhe este termo de consentimento, introduzirá o sistema TRIVIR e suas funcionalidades, bem como as tarefas a serem realizadas durante a sessão. O pesquisador responsável permanecerá disponível para esclarecer dúvidas. Ao final da sessão você será convidado a responder um questionário sobre sua experiência com o sistema.

O estudo será realizado individualmente e com duração máxima de 3 horas, a sessão será realizada remotamente (online), através do sistema Team Viewer, em respeito ao distanciamento social adotado em atual situação de pandemia, e a data/horário serão combinados conforme a sua conveniência. As perguntas não serão invasivas à sua intimidade, entretanto, pode ser que a participação resulte em algum cansaço físico dada a duração do estudo e a necessidade de expressar opiniões pessoais acerca da sua experiência durante a utilização do sistema. Sinta-se livre para solicitar pausas sempre que necessário, bem como para não responder as perguntas caso considere-as inadequadas. Caso sinta-se desconfortável poderá interromper o estudo a qualquer momento, alertando o pesquisador responsável o qual o (a) encaminhará aos profissionais da saúde adequados e oferecerá todo o suporte necessário durante seu tratamento. Caso a sua participação na pesquisa resulte em algum dano à sua integridade física ou mental, você terá direito à indenização e receberá assistência integral e imediata de forma gratuita pelo tempo que se fizer necessário durante ou após a pesquisa.

Sua participação é voluntária e será organizada de modo a não incorrer em nenhum custo financeiro para você. Custos referentes ao transporte e alimentação, mas não limitados a estes, serão ressarcidos no dia da sua participação. Porém, não haverá compensação financeira ou de outra natureza pela sua participação. Você pode, a qualquer momento, desistir de participar e retirar seu consentimento, sem que isso lhe traga nenhum prejuízo pessoal ou profissional, seja em sua relação ao pesquisador, à Instituição em que trabalha ou à Universidade de São Paulo. Todas as informações obtidas serão confidenciais, sendo assegurado o sigilo sobre sua participação em todas as etapas do estudo. Os nomes dos participantes não serão identificados, com garantia de anonimato nos resultados e publicações.

Esperamos, a partir da análise dos resultados desses estudos, identificar limitações existentes no sistema atual e as características desejadas de uma solução mais adequada às necessidades de usuários. Sua participação nessa pesquisa auxiliará no levantamento de informações a serem utilizadas exclusivamente para os fins científicos expressos neste termo.

Você e o pesquisador responsável assinarão duas vias deste termo. Você receberá uma das vias enquanto o pesquisador responsável receberá a outra. Todas as páginas deste termo

Rubrica	
Pesquisador Responsável	***Participante da Pesquisa***

serão rubricadas por você e pelo pesquisador responsável. Se você tiver qualquer problema ou dúvida durante a sua participação na pesquisa poderá comunicar-se com o pesquisador responsável pelo telefone (055) 99916-3043 ou dirigir-se ao prédio 3, sala 3-253 do Instituto de Ciências Matemáticas e de Computação de 2ª. à 6ª. das 14:00 às 18:00h.

Esta pesquisa obteve aprovação do Comitê de Ética em Pesquisa com Seres Humanos, assegurando à sua participação todos aspectos éticos necessários descritos pela Resolução 466/12 do Conselho Nacional da Saúde (CNS) para que ambas as partes, você e o pesquisador responsável, estejam amparados durante a realização do estudo.

Declaro que entendi os objetivos, riscos e benefícios de minha participação na pesquisa e concordo em participar. O pesquisador me informou que o projeto foi aprovado em 27 de Maio de 2020, sob o parecer 4.052.532, pelo Comitê de Ética em Pesquisa com Seres Humanos da Escola de Educação Física e Esporte de Ribeirão Preto (EEFERP) na Universidade de São Paulo, localizada na Avenida Bandeirantes, 3900, Vila Monte Alegre, CEP: 14040-907, Ribeirão Preto/SP. Fone: (16) 3315-0494. Email: cep90@usp.br

ENDEREÇO PARA CONTATO:

Pesquisador Responsável:
Sherlon Almeida da Silva

Endereço Profissional:
Av. Trab. São Carlense, 400 - Parque Arnold Schimidt, São Carlos - SP, 13566-590
Instituto de Ciências Matemáticas e de Computação - ICMC/USP
Laboratório de Visualização, Imagens e Computação Gráfica -VICG
Prédio 3 - Sala 3-253

Contato telefônico: (55) 9 9916-3043 **E-mail:** sherlon@usp.br

_____ **de** _____ **de 2021, São Carlos/SP**

Sherlon Almeida da Silva	Maria Cristina Oliveira
Sherlon Almeida da Silva	**Maria Cristina Ferreira de Oliveira**
Pesquisador Responsável	Orientadora do Projeto de Pesquisa
Mestrando no PPGCCMC - ICMC/USP	Docente e Diretora do ICMC/USP

Participante da Pesquisa

Rubrica	
Pesquisador Responsável	***Participante da Pesquisa***


ICMC
USP
SÃO CARLOS