XIX encontro nacional de pesquisa em ciência da informação
ENANCIB
// SUJEITO INFORMACIONAL E AS PERSPECTIVAS ATUAIS EM CIÊNCIA DA INFORMAÇÃO. //
22-26 OUTUBRO 2018
LONDRINA/PR
Universidade Estadual de Londrina

XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2018

GT- 7 - Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação

ESTUDO MÉTRICO SOBRE BIBLIOTECA DIGITAL: USO DO SOFTWARE IRAMUTEQ

Márcio Henrique Wanderley Ferreira (Universidade Federal de Pernambuco)

Renato Fernandes Côrrea (Universidade Federal de Pernambuco)

METRIC STUDY ON DIGITAL LIBRARY: USE OF IRAMUTEQ SOFTWARE

Modalidade da Apresentação: Comunicação Oral

Resumo: Apresenta os principais assuntos abordados dentro da temática “Biblioteca digital” na literatura científica brasileira em Ciência da Informação, analisando os termos presentes nos campos do título, resumo e das palavras-chave, dos trabalhos publicados em artigos de periódicos e indexados na BRAPCI contendo o termo “Biblioteca digital” nas palavras-chaves. Realiza um estudo de caso que envolve visualização dos termos por meio de uma proposta metodológica para estudo métrico temático utilizando o software IRAMUTEQ. Possui natureza quali-quantitativa, quanto aos objetivos, classifica-se como exploratória. Quanto aos procedimentos, caracteriza-se num estudo de caso e se utiliza da análise estatística de termos e mineração de texto. A proposta metodológica envolve o levantamento e visualização dos termos em forma de gráficos gerados a partir da análise de frequência (lei de Zipf) e co-ocorrência das palavras-chave identificadas. Como resultado da aplicação, o método permite a visualização dos assuntos que envolvem o tema “Biblioteca digital” na Ciência da Informação brasileira no período de 2001 até 2017. Foram analisados 82 artigos, eles apontam a existência de 12440 ocorrências de termos e 2087 termos únicos, assinalam o termo composto “Biblioteca digital” com 203 ocorrências e seu grau de proximidade com outras temáticas como informação, tecnologia, artigo, digital, biblioteca, acesso, serviço, teses e dissertações.

Palavras–chave: Biblioteca digital. Ciência da Informação. Estudo temático. Bibliometria. Mineração de Texto.

Abstract: This work presents the main topics covered within the theme "Digital Library" in the Brazilian scientific literature in the Information Science area, through analyses of the terms present in the title, abstract and keywords fields, of papers published in periodical articles and indexed in BRAPCI with the term "digital library" in keywords. It carries out a case study that involves visualization of the terms through a methodological proposal for thematic metrical study using IRAMUTEQ software. It has a qualitative-quantitative nature, in terms of objectives it is an exploratory research. As for the procedures, it consists of a case study and uses the statistical analysis of terms and text mining. The

methodological proposal involves the survey and visualization of the terms in the form of graphs generated through the frequency analysis (Zipf's law) and co-occurrence of the identified terms. The method application generates visualizations of the subjects that involve the theme of "Digital Library" in the Brazilian Information Science from 2001 to 2017. The analyse of the 82 articles points to the existence of 12,440 occurrences of terms and 2,087 unique terms, the term "digital library" with 203 occurrences and its degree of proximity to other themes like information, technology, article, digital, library, access, service, thesis and dissertations.

Keywords: Digital library. Information Science. Thematic study. Bibliometrics. Text Mining.

1 INTRODUÇÃO

A sociedade da informação evolui de acordo com suas necessidades sociais e culturais. Com o advento das tecnologias da informação e comunicação, o ser humano passou a ser capaz de ampliar suas possibilidades de comunicação e de acesso à informação. Ambientes digitais, como a *world wide web*, vem se desenvolvendo com o objetivo de fornecer informações de maneira mais eficiente e de prover acesso aos conteúdos armazenados.

Neste sentido, serviços de armazenamento e busca de informação digital tornam-se cada vez mais comuns, já que a sociedade da informação necessita e demanda cada vez mais acesso a textos, imagens, sons e vídeos em formato digital. Portanto, surgem as Bibliotecas digitais, com a capacidade de fornecer acesso aos conteúdos digitais de forma irrestrita no tempo e no espaço.

O conceito de biblioteca digital no entanto se encontra em constante evolução, sendo influenciadora e ao mesmo tempo influenciada pela sociedade da informação (TAMMARO; SALARELLI, 2006). De acordo com Borgman (1999) a Biblioteca digital (BD) representa uma reunião de recursos eletrônicos, que reunidos possibilitam a criação, pesquisa e uso de informações.

Outra importante definição é trazida por Ciotti e Roncaglia (2002, apud TAMMARO e SALARELLI, 2006, p. 122), na qual afirmam que a “Biblioteca digital é uma coleção de documentos digitais estruturados, produzidos mediante digitalização de materiais existentes ou preparados de modo digital na origem […]”. Enquanto que para Cunha (1997), a Biblioteca digital é também conhecida como biblioteca eletrônica, biblioteca virtual, biblioteca sem paredes e biblioteca cibernética.

Para Fox et al (2002), o termo *Digital Library* é utilizado de diversas maneiras dependendo da linha temática na qual o pesquisador esteja trabalhando. Isso ocorre, pois, o

tema envolve diversas áreas, dentre elas a preservação digital, a catalogação, classificação, recuperação, revocação, arquitetura da informação, estudo dos usuários, design, além de computação.

Já para a *Digital Library Federation* (DLF):

> Bibliotecas digitais são organizações que disponibilizam os recursos, incluindo pessoal especializado, para selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir, preservar a integridade e assegurar a persistência ao longo do tempo de coleções de trabalhos digitais, de forma que eles estejam pronta e economicamente disponíveis para uso de uma comunidade definida ou um conjunto de comunidades. (DLF, 1998, apud SAYÃO, 2009, p.15).

Estudos sobre a evolução da temática sobre BD na literatura científica de Ciência da Informação (CI), mostram-se importantes para a apropriação e consolidação do conceito por parte dos pesquisadores (OHIRA; PRADO, 2002).

Entre os estudos que realizaram uma análise diacrônica de artigos na temática BD na CI brasileira, destaca-se o trabalho de Ohira e Prado (2002). As autoras realizaram um estudo de produção científica de 1995 aos anos 2000 através da análise de 33 artigos sobre BD. Foram elucidados quais assuntos estavam sendo discutidos e as diversas definições iniciais sobre os conceitos em volta da Biblioteca digital. Este estudo inicial sobre os conceitos relacionados à um determinado tema, remete ao trabalho de Santos (2015), no qual ela propõe estabelecer uma discussão sobre o que seriam estudos métricos que envolvessem a análise de temas e assuntos. Neste caso, ela considerou que a padronização terminológica no uso de grandes áreas temáticas, seja pela reindexação dos dados coletados ou pela atribuição de vocabulário controlado, poderiam contribuir nos procedimentos bibliométricos.

Nesse contexto, o objetivo do trabalho aqui proposto é analisar os assuntos desenvolvidos dentro da temática “Biblioteca digital”, entre termos presentes em artigos de periódicos de Ciência da Informação indexados na BRAPCI, para isso, se utilizou de método estatístico de análise de frequência de termos e mineração de texto, para identificar e refletir sobre os assuntos verificados no decorrer de 17 anos de estudos desenvolvidos (2001 até 2017).

Dessa forma, o trabalho busca responder ao problema existente da ausência de controle terminológico das bases de dados existentes, que comprometem a compreensão e relação temática sobre determinados temas. Para isto propõe um método de análise que

consiste de um pré-processamento para estruturar os textos de maneira a retirar ambiguidades, eliminar palavras com classificações gramaticais que não servissem ao propósito da pesquisa e padronizar termos que estivessem no plural ou escrito de forma equivocada. Para realização do pré-processamento e posterior análise estatística dos termos, faz uso da ferramenta estatística IRAMUTEQ (*Interface de R pour analyses Multidimensionnelles de Textes et de Questioneeires*).

O método proposto corrobora com a necessidade de buscar os temas mais relevantes que estão sendo desenvolvidos na CI sobre esta temática.

2 TRABALHOS QUE UTILIZARAM O IRAMUTEQ

Nesta seção, serão descritas algumas pesquisas que foram desenvolvidas com o uso do software IRAMUTEQ. Ao ser realizada uma pesquisa simples na ferramenta de busca da Biblioteca de teses e dissertações do IBICT, foi possível identificar que existem 111 teses e dissertações que utilizam o IRAMUTEQ como ferramenta ou descrevem as suas principais funcionalidades. Dentre as instituições que tiveram maior frequência de trabalhos relacionados, essas representam as 13 principais: UFPB(16), UFRN(12), UNB(11), UCB(8), UFSC(8), UNIFOR(7), UFC(6), UFPE(6), UTP(6), UFS(5), UFFS(4), USP(4), UFBA(3). Ressalta-se que os trabalhos foram defendidos num período de 2014 a 2018, demonstrando serem pesquisas recentes.

Dentre as pesquisas desenvolvidas, destaca-se a de Lima (2017), dissertação da área de matemática, mas que se aproximou da proposta aqui apresentada. Ela trouxe um estudo investigativo sobre a utilização das tecnologias digitais de informação e comunicação pelos professores de matemática da rede estadual de educação de Goiás. Nesse estudo, ela utilizou o IRAMUTEQ para construção de um gráfico de “Árvore de Similitude”, a partir das respostas dos questionários, feitas pelos professores, e construiu uma nuvem de tags dos termos mais utilizados numa pergunta feita aos entrevistados. Dessa forma, obteve dados de conexidade entre as palavras, que foram utilizadas nas repostas, e pode dissertar sobre as razões que levaram ao quadro obtido no gráfico. De acordo com Marchand e Ratinaud (2012), a análise de similitude se baseia na teoria dos grafos e permite a identificação das coocorrências entre as palavras resultando em indicações de conexidade entre os termos. Essa análise auxilia na identificação da estrutura de um corpus textual distinguindo entre as partes comuns e as especificidades, ela foi uma das escolhidas nesta pesquisa.

Outra pesquisa interessante foi a tese de Ferreira Júnior (2017), da área de Ciência da Informação, na qual ele conduziu um estudo de redes de relacionamento conceituais. O objetivo do trabalho foi buscar a caracterização de conceitos para a análise de sistemas de informação em ambientes digitais. Para isso, ele utilizou de técnicas e medidas dos estudos de redes para identificar e caracterizar esses conceitos. Por meio do IRAMUTEQ ele conseguiu formatar categorias temáticas, com frequências simples e relativa, na qual ele pudesse exportar os dados por meio de gráficos de rede e identificar características das conexões, medidas de conexão e densidade do gráfico. Os resultados alcançados permitiram identificar as relações de proximidade dos termos, explicando as relações de conectividade, e tornando possível a análise das relações e atribuição de padrões de comportamento nas respostas obtidas.

Neste sentido, percebe-se de maneira geral, que o software IRAMUTEQ serve de apoio à construção de análises estatísticas de termos realizadas nas pesquisas, e serve como ferramenta para construção de gráficos. Nesse ínterim, cabe uma reflexão sobre o uso da ferramenta como meio de auxiliar nos processos de pré-processamento, análise e visualização de resultados. Nesta pesquisa são utilizados apenas alguns métodos do software, que são apontados na próxima seção.

3 PROCEDIMENTOS METODOLÓGICOS

Diante do exposto, a pesquisa aqui desenvolvida possui natureza quali-quantitativa e quanto aos objetivos classifica-se como exploratória, pois busca proporcionar uma maior explanação sobre o problema de correlação temática. Quanto aos meios, caracteriza-se num estudo de caso e se utiliza da análise de conteúdo como procedimento para aprofundamento das análises. (GIL, 2009).

A necessidade do estudo parte da busca por compreender os níveis de correlação entre os temas identificados e tentar abranger como ocorre a construção do conhecimento na área.

Para isso, utiliza-se de elementos teóricos do processamento de linguagem natural. Dentro da temática da “Biblioteca digital”, realiza tarefas de mineração de texto, tais como: agrupamento de termos, frequência de termos e correlação temática. Segundo Rezende (2003), o processo de mineração de textos pode ser dividido em cinco etapas iniciais: 1-identificação do problema; 2-pré-processamento textual; 3-extração de padrões; 4-pós-

processamento; 5-utilização do conhecimento, que no caso seria a visualização e o resultado do processamento.

Dessa maneira, o estudo utilizou do método estatístico de análise textual. Essa análise, leva em consideração que a dispersão consequente da variação dos assuntos abordados no conjunto de artigos é grande ao serem avaliados termos únicos como as palavras-chave. Neste cenário, ocorrem inconsistências, uso de sinônimos, variações linguísticas, assim como critérios diversos para nivelar o tratamento dos termos oferecidos pelos autores, o que torna relevante a análise dos procedimentos adotados. (SANTOS, 2015).

Inicialmente, buscou na BRAPCI, o termo composto “Biblioteca digital” no campo das palavras-chave. Ressalta-se que, para efeitos dessa pesquisa, os resultados da busca pelo termo composto no singular englobam aqueles que se apresentam no plural. Ao realizar a busca que compreende o período de 1972 a 2017, realizada no mês de novembro de 2017, foram encontrados 100 trabalhos que possuem a palavra-chave “Biblioteca digital” indexada. Contudo, apesar da recuperação de 100 trabalhos indexados, apenas 82 trabalhos foram utilizados na pesquisa, por possuírem “Biblioteca digital” no campo das palavras-chave e por serem identificados como artigos de periódicos. Observou-se também que a busca duplicou 3 trabalhos, recuperando 2 vezes, por terem sido descritos em inglês e português.

Posteriormente, utilizou-se o software IRAMUTEQ, que possui os instrumentos necessários para as análises. Conforme Camargo e Justo (2013), o IRAMUTEQ é um software gratuito desenvolvido originalmente no idioma francês, pelo cientista Pierre Ratinaud em 2009, como ferramenta de organização de dados. Ele ancora-se no ambiente estatístico do software R (www.r-project.org) e na linguagem python. Permite a viabilização de diferentes tipos de análise de dados textuais, como lexicografia básica, lematização, cálculo de frequência de palavras, análises multivariadas, análise pós-fatorial e análise de similitude. Dessa forma, tal ferramenta possui o rigor científico, no qual é diretriz para o desenvolvimento desta pesquisa.

Para atingir os objetivos da pesquisa, foram realizadas as seguintes etapas:

1- Realizou-se a pesquisa da palavra-chave “biblioteca-digital” na BRAPCI (Base de dados Referencial de Artigos de Periódicos em Ciência da Informação), marcando apenas o campo “palavras-chave”. Após essa pesquisa foram encontrados 100 artigos que possuíam “Biblioteca digital” como palavra-chave indexada na base de dados. Em seguida foi realizado

o refinamento dos arquivos, identificando 82 artigos científicos passíveis de análise. Os outros documentos não tinham BD como palavra-chave no campo, não eram artigos de periódicos, ou estavam duplicados.

2- No passo seguinte foi feita uma coleta manual das palavras-chave, dos títulos e dos resumos dos campos dos 82 artigos para uma base de dados no software bloco de notas. Foram coletadas 430 palavras-chave, 82 resumos e 82 títulos.

3- Posteriormente foi realizada uma formatação do arquivo de extensão (.txt) correspondente ao corpus da pesquisa, no formato padrão exigido pelo IRAMUTEQ. Nesta etapa, o conteúdo dos campos de metadados de cada artigo foi numerado com um padrão de caracteres (**** *Artigo_X , onde X corresponde a um inteiro único para cada artigo) que permitiu seu processamento. O arquivo foi salvo com codificação UTF-8 no bloco de notas.

4- No quinto passo, buscou-se realizar o pré-processamento textual, introduzindo a padronização dos termos e retirada (limpeza) daqueles que não tivessem relevância nos resultados. O software IRAMUTEQ não consegue analisar termos compostos, por esse motivo, exigiu que fosse inserido o símbolo "_" entre as palavras dos termos compostos, para que o software considerasse como um único termo para análise. Essa etapa exigiu a geração de uma tabela de termos compostos que foram ordenados no Microsoft Excel, para que em seguida fosse possível a substituição no *corpus*. Essa tabela foi criada a partir da coleta de todos os termos compostos identificados no *corpus* pela frequência em que apareciam, e foram ordenadas de acordo com a frequência. Por exemplo, o termo "Ciência da Informação" foi substituído por "Ciência_da_Informação".

5- Na última etapa, implementou-se a construção de gráficos de análise textual e de frequência de palavras utilizando o IRAMUTEQ. Para a referida análise textual, utilizou-se das seguintes ferramentas do software: aplicação da Lei de Zipf; cálculo de frequência de palavras; análise de similitude e nuvem de palavras. Essas análises foram escolhidas pois se adequariam aos dados coletados e forneceriam informações suficientes sobre os temas trabalhados.

4 RESULTADOS

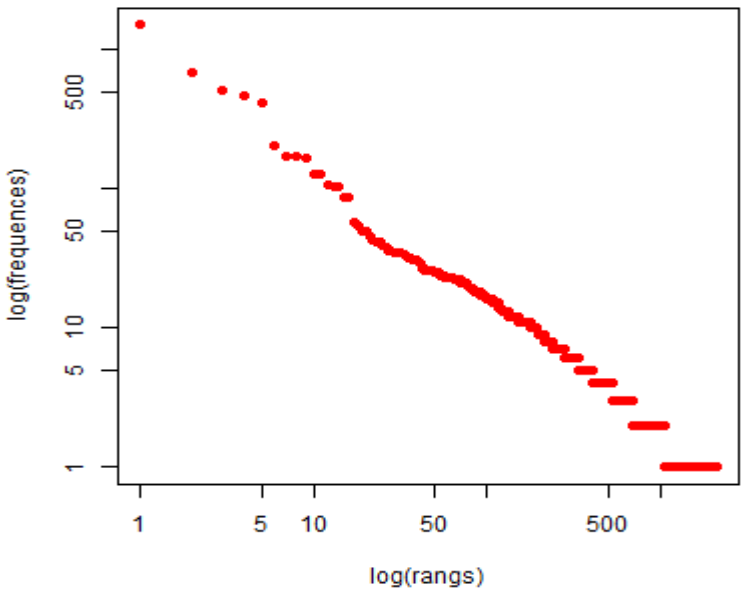
Primeiramente, optou-se por realizar uma análise que demonstra os índices de dispersão dos termos analisados na pesquisa (Figura 1). Nesta análise foi possível identificar que nos 82 artigos analisados existiam 12440 ocorrências de termos e 2087 termos únicos, onde cada vocábulo corresponde a um tipo de termo, e a quantidade de vezes em que aparece

corresponde à ocorrência. Ressalta-se que cada conjunto de termos dos campos estudados possuíam uma média de 152 ocorrências por texto.

Segundo Molinos et al (2016), o gráfico de Zipf se baseia na lei de mesmo nome, a lei de Zipf. Segundo Guedes (2012), Zipf observou que num texto longo existia uma relação entre a frequência que uma dada palavra ocorria e sua posição na lista de palavras ordenadas segundo sua frequência de ocorrências. Esse produto da ordem de série r de uma palavra pela sua frequência de ocorrência f era aproximadamente c, enunciou-se o que ficou conhecido como primeira lei de Zipf (r * f = c). Essa lei é uma técnica que identifica a frequência de ocorrência de palavras dentro de um texto longo, dessa forma, consegue-se perceber uma correlação entre a frequência em que um determinado termo aparece e sua posição na lista de palavras. Isso leva a conclusão de que existe uma regularidade na seleção e uso de palavras. Portanto o produto de uma ordem de série R de uma palavra pela sua frequência de ocorrência F é aproximadamente uma constante C (R x F = C).

Dessa maneira na figura 1 é possível visualizar a aplicação da Lei de Zipf no *corpus* analisado. A leitura do grafo é feita da seguinte maneira: na amostra de 2087 tipos de termos, cerca de 10 termos que podem ser identificados no eixo horizontal, apresentam frequência mais alta, e supõe-se que cerca de 30% das ocorrências do texto são representadas por esses 10 termos.

Figura 1: Aplicação da Lei de zipf no *Corpus.*



Fonte: Dados da pesquisa – 2018.

Essa suposição é feita ao serem comparados os eixos vertical e horizontal, percebe-se que no eixo horizontal, cerca de 10 termos representam os 30% mais frequentes, de acordo com uma escala possível de percentual. Além disso, pode-se chegar a essa conclusão pois percebesse que existem os pontos mais altos no eixo vertical, e esses pontos vão descendo na medida em que a ocorrência vai diminuindo no texto. Da mesma maneira, o eixo da vertical, que vai de 1 a 500, corresponde as faixas de frequências dos 2087 termos. Assim, conclui-se que poucos termos representam o conteúdo semântico da amostra. O que pode ser ratificado ao voltar-se aos dados apresentados na tabela 1, na próxima análise.

Em seguida foi realizada uma análise de conteúdo, com o objetivo de identificar as classes gramaticais mais frequentes dentro do *corpus* observado. Esses resultados são apresentados na tabela 1, com o objetivo de ratificar a quantidade de termos que poderiam aparecer na análise se não houvesse uma "filtragem" das classes. Dessa maneira, se não houvesse a eliminação das classes gramaticais que não possuem representatividade semântica, os termos mais frequentes seriam as preposições, artigos e conjunções. Observa-se que, dos 10 primeiros termos identificados, apenas 2 possuem grau de importância semântica para o corpus, que são: "Biblioteca digital", com 203 ocorrências, e "informação", com 129 ocorrências. Dos 20 termos mais frequentes, sem a utilização do filtro, apenas 5 possuem grau de relevância para análise, completam: "biblioteca", com 57 ocorrências, "pesquisa", com 54 ocorrências e "digital", com 49 ocorrências.

Tabela 1: 20 termos mais frequentes e classes gramaticais (sem filtro).

TERMOS	FREQUÊNCIA	CLASSES	TERMOS	FREQ.	CLASSES
De	1504	Preposição	**Se**	128	Pronome
A	685	Artigo/preposição	**Por**	106	Preposição
E	510	Conjunção	**Como**	105	Advérbio
Em	461	Preposição	**Uma**	105	Artigo indef.
O	413	Artigo definido	**Um**	88	Artigo indef.
Biblioteca_digital	203	Substantivo composto	**Com**	87	Preposição
Que	173	Pronome relativo	**Biblioteca**	57	Substantivo
Para	171	Preposição	**Pesquisa**	54	Substantivo
Ser	167	Verbo	**Digital**	49	Adjetivo
Informação	129	Substantivo	**Sobre**	49	Preposição

Fonte: Dados da pesquisa – 2018.

Entretanto, o objetivo da pesquisa é o de apresentar os termos mais representativos, no que diz respeito a relevância, para isso foi construída a tabela 2, onde é possível identificar os termos mais frequentes presentes de apenas 2 classes gramaticais, substantivos e adjetivos. Essa tabela demonstra como os dados podem ser filtrados para se demostrar

apenas aquilo que for mais relevante para o estudo. O software utilizado permite, previamente, que se excluam classes gramaticais que não serviriam para a análise. Dessa maneira, os resultados podem ser "filtrados" para que apenas sejam demonstrados os dados desejados. Neste sentido, percebe-se que as classes gramaticais mais frequentes que representam os termos mais significativos são os substantivos. Isso ocorre pois dentro da construção dos conceitos, os termos mais frequentes são aqueles que pertencem a esse tipo de classe gramatical, dentre eles destacam-se: biblioteca digital, informação e biblioteca.

Tabela 2: 20 termos mais frequentes e classes gramaticais (com filtro).

TERMOS	FREQUÊNCIA	CLASSES	TERMOS	FREQ.	CLASSES
Biblioteca_digital	203	Substantivo composto	**Projeto**	35	Substantivo
Informação	129	Substantivo	**Serviço**	35	Substantivo
Biblioteca	57	Substantivo	**Universidade**	34	Substantivo
Pesquisa	54	Substantivo	**Usuário**	33	Substantivo
Digital	49	Adjetivo	**Desenvolvimento**	32	Substantivo
Estudo	43	Substantivo	**Novo**	32	Adjetivo
Artigo	42	Substantivo	**Sistema**	31	Substantivo
Conhecimento	41	Substantivo	**Área**	31	Substantivo
Trabalho	38	Substantivo	**Acesso**	30	Substantivo
Tecnologia	36	Substantivo	**Processo**	27	Substantivo

Fonte: Dados da pesquisa – 2018.

Ressalta-se que, durante a geração dos resultados, identificaram-se termos compostos com frequência um pouco abaixo desses 20 termos mais frequentes, sendo eles com os respectivos valores de frequência absoluta: Biblioteca digital de teses e dissertações (26), teses e dissertações (24), preservação digital (24), acesso aberto (23) e ciência da informação (23).

Outro importante resultado, é a análise de similitude com as principais informações encontradas nos artigos. Este gráfico é gerado para facilitar a visualização da informação, que de acordo com Vieira e Côrrea (2011), está relacionada com a transformação de dados abstratos em gráficos ou imagens. Dessa maneira, as visualizações auxiliam na compreensão de determinado assunto e minimizam o esforço cognitivo nesta compreensão.

A análise de similitude se baseia na teoria dos grafos e possibilita identificar as coocorrências entre as palavras. Esse resultado traz indicações da conexidade entre as palavras, auxiliando na identificação da estrutura de um corpus textual, distinguindo as partes comuns e as especificidades das variáveis descritas. (MARCHAND; RATINAUD, 2012).

Para realizar essa análise foram excluídas as seguintes classes gramaticais: advérbios, advérbios suplementares, conjunções, preposições, verbos e verbos suplementares. Outro ponto a ser ressaltado, é que foram utilizados para essa amostra, os 70 vocábulos com maior

índice de frequência e relevância temática foram selecionados, para os quais a frequência variou de 03 até 203. Esse método foi adotado pois permitiu uma melhor análise visual. Dentre os 70 termos, que a partir deste momento chamaremos de vértices, 2 foram escolhidos, pois ao ser feita uma análise visual, percebeu-se que eles possuíam valor semântico, pela quantidade de agrupamentos de vértices que englobavam, e pela quantidade de ligações(arestas) que se conectam com outros vértices. As configurações gráficas utilizaram os seguintes parâmetros: escore -> coocorrência; apresentação -> *fruchterman reingold*; tipo de gráfico -> estatístico; utilizou comunidades e *halo*. Os principais vértices identificados foram: biblioteca digital e informação. (Figura 2)

Ao observar o gráfico da Figura 2, pode-se identificar 2 polos centrais de termos em volta de duas temáticas mais frequentes.

Figura 2 : Agrupamentos de termos em nível de proximidade e similitude.

Fonte: Os autores – 2018.

Na análise de frequência foi identificado que os termos “biblioteca digital” e “informação” são os 2 termos mais frequentes respectivamente, apresentando 203 e 129 ocorrências cada um. Observa-se que os termos se ligam a outros por meio de fios cinza-claro. A espessura desses “fios” representa o grau de conexão entre os termos observados. Dessa maneira, observa-se que “biblioteca digital” se conecta com termos como, “informação”, “biblioteca”, “repositório”, “preservação”, “digitalização”, “digital”, “artigo”, “acesso”, “metadados”, “Biblioteca digital de teses e dissertações” e “teses e dissertações eletrônicas” (vide Figura 2). Esses termos aparecem em agrupamentos externos ao agrupamento central em verde claro.

Em relação à “Biblioteca digital”, percebe-se que está mais intensamente conectado à Informação (apresentando 46 ligações), enquanto que o termo “Informação”, está mais conectado à “Tecnologia” (apresentando 23 ligações). Isso significa que os termos ocorrem X vezes de forma conjunta no corpus. Tal visualização só é possível graças a verificação da espessura das linhas que conectam os termos e os números em azul, que representam a quantidade de vezes que eles ocorrem de forma simultânea.

Na figura 2, por meio da visualização das espessuras das arestas, na cor cinza claro, e por meio dos índices, na cor azul, é possível identificar os 11 principais vértices com grau de proximidade e similitude mais próximos de “Biblioteca digital”, verificam-se: “informação” (46), “tecnologia” (23), “artigo” (21), “digital” (21), “biblioteca” (18), “acesso” (16), “serviço” (16), “teses e dissertações” (16), “organização” (12), “recurso” (12) e “usuário” (12). Os índices de conexidade, apresentados entre parênteses, representam o grau de proximidade dos termos com o vértice “Biblioteca digital”.

Outra análise interessante, apresenta o agrupamento do termo “informação”, em laranja, conectado mais intensamente com “tecnologia”, “conhecimento”, “comunicação”, “recuperação” e “ciência da informação”, com graus de conexidade que variam de 8 à 23. E demonstra estar menos conectados com termos como, ”Biblioteca digital educativa”, “pós graduação” e “biblioteca escolar”. Essas visualizações podem ser justificadas no âmbito das pesquisas que envolvem a informação na CI e mais especificamente no ambiente das Bibliotecas digitais. Muitos trabalhos do *corpus* envolvem estudos das seguintes temáticas: sistemas de informação, tecnologia da informação, informação e comunicação, representação da informação, profissional da informação, recuperação da informação, busca de informação,

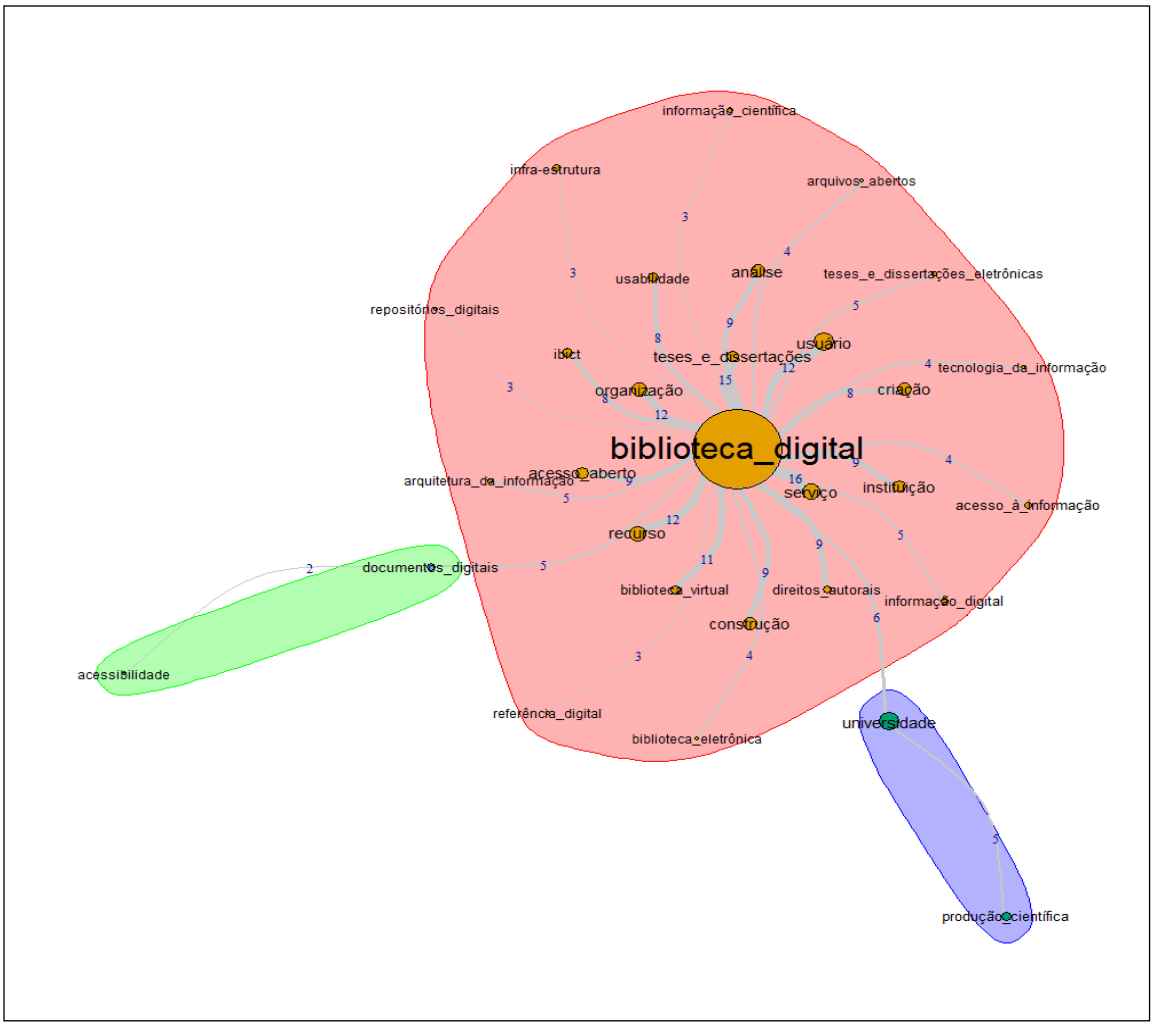
necessidade de informação, disseminação da informação, produção de informação, rede de informação e ciência da informação. Por esse ângulo, justifica-se o grau de conexidade dos termos em volta da palavra “informação”, esses gráficos de similitude apresentam indícios das correlações temáticas existentes entre os vocábulos.

Observa-se que os termos mais periféricos em agrupamentos em outras cores, se conectam ao conjunto principal, mesmo apresentando números inferiores de ligação, como o vértice “biblioteconomia”, por exemplo, apresentado no agrupamento em azul marinho, na parte inferior esquerda do grafo, com grau 2.

Em seguida optou-se por construir um gráfico de similitude que aprofundasse o processo de compreensão das ligações existentes no agrupamento em verde claro. Para o gráfico apresentado na figura 3, foram utilizados os mesmos parâmetros da figura 2. Ressalta-se que para esta análise utilizaram-se os 30 termos centrais inseridos no grupo mais próximo ao vértice “Biblioteca digital”.

O gráfico presente na figura 3, foi importante para elucidar termos que estavam embaralhados na figura 2, como por exemplo: “referência digital” (3), “acessibilidade” (2), “direitos autorais” (9), “instituição” (9) e “usuário” com grau de conexidade (12). A sobreposição de termos e arestas é algo comum na geração de gráficos de similitude pelo IRAMUTEQ, necessitando de alternativas que facilitem o processo de análise. Outro ponto importante desta figura, é que são observados os índices de frequência dos termos no *corpus*, por meio do tamanho da área das circunferências em amarelo e azul. Quanto maior a área da circunferência, maior é a frequência do termo. Assim, observa-se que apesar de alguns termos apresentarem um baixo índice de frequência, como “bibliotecas eletrônicas” (5 ocorrências) e “arquivos abertos” (4 ocorrências), eles apresentam relativo grau de conexidade com o vértice “Biblioteca digital”. (Figura 3)

Figura 3 : Agrupamentos dos 30 termos centrais em nível de proximidade, similitude e frequência.



Fonte: Os autores – 2018.

Outro resultado apresentado na figura 3, indica que os 2 agrupamentos menores, nas cores azul e verde, apresentam 4 vértices: (em azul; universidade e produção científica), (em verde; acessibilidade e documentos digitais). Esses vértices, inicialmente, faziam parte do mesmo agrupamento em verde na figura 2. Essa característica indica que o software, ao calcular o grau de conexidade relativo à proximidade e similitude, dividiu os 4 termos em comunidades distintas. Dentre os resultados importantes que são apresentados na figura 3, podem-se destacar os principais termos com grau de conexidade superior ou igual a 10, que são: serviço (16), teses e dissertações (15), organização (12), recurso (12), usuário (12) e biblioteca virtual (11). Os vértices/termos mencionados apontam quais temáticas estão mais próximas de BD, no contexto apresentado na figura.

Supõe-se que tais observações iniciais são visualizadas pois tais termos possuem algum grau de similitude dentre os estudos praticados na CI. Como a pesquisa aqui desenvolvida

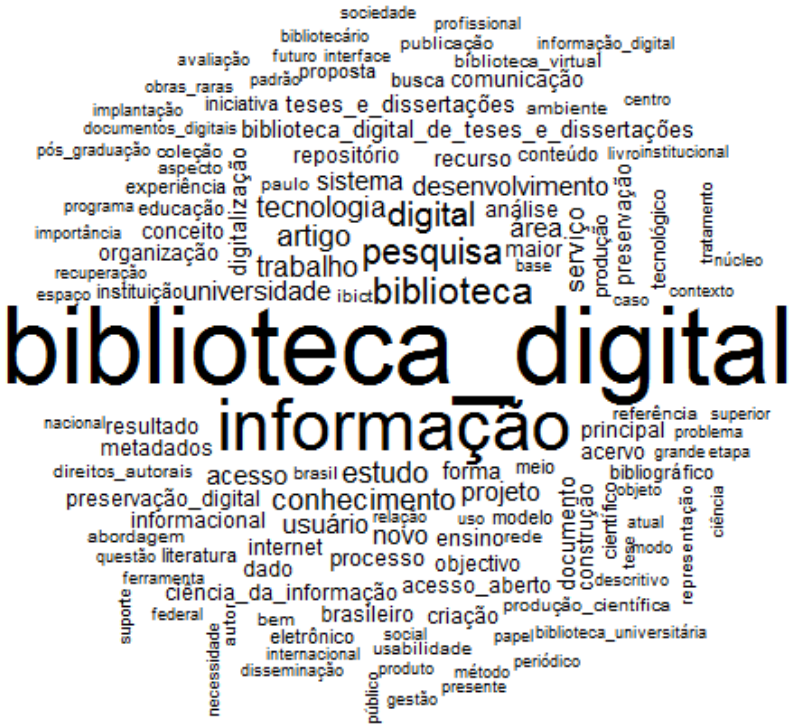
volta-se para a identificação de termos que envolvem estudos sobre a BD na CI, naturalmente tais visualizações correspondem a temáticas mais desenvolvidas dentro desse campo.

No instante em que “biblioteca digital” está conectado com “organização”, percebe-se que os conceitos que envolvem os dois termos possuem relação, pois a BD é um ambiente que necessita de um nível adequado de organização na sua estrutura de informação. No momento em que “informação” está conectado com “tecnologia”, subtende-se que tal ligação ocorreu devido a aplicação das tecnologias da informação no contexto das bibliotecas digitais.

Neste sentido, os grafos de conexidade servem para facilitar a identificação dos termos relacionados e quais são as possíveis tendências de relação conceitual entre os termos. Esclarece-se que os estudos de redes de conexidade são amplamente utilizados em áreas como as ciências sociais aplicadas, justamente por facilitarem a compreensão visual dos termos de uma determinada amostra.

Por fim, optou-se pelo gráfico de nuvem de palavras para facilitar a visualização dos termos mais utilizados no *corpus*. Nesse gráfico foram excluídos os termos das seguintes classes gramaticais: artigos, verbos, advérbios, conjunções, onomatopeias e preposições. Ressalta-se que a título de melhor visualização, a amostra foi delimitada aos 133 termos mais frequentes dentro do corpus, de uma variação de frequência entre 10 e 203. (Figura 4)

Figura 4 – Nuvem de palavras da amostra



Fonte: Iramuteq – 2018.

Ao analisar a figura 4 percebe-se que os termos correspondentes na tabela 1 aparecem na imagem, assim como evidencia-se os termos mais frequentes com o tamanho de fonte maior. Quanto maior a fonte do termo na figura, maior sua frequência na amostra, assim como menor o tamanho da fonte, menor sua frequência. Assim, pode-se perceber que os 10 primeiros termos de maior frequência aparecem na seguinte ordem decrescente: Biblioteca digital, informação, biblioteca, pesquisa, digital, estudo, artigo, conhecimento, trabalho e tecnologia. Enquanto que os 10 menores em termos de frequência, que aparecem com fonte menor, são eles: uso, suporte, sociedade, Pós graduação, institucional, informação digital, implantação, avaliação, espaço e centro. Nesse sentido, tal análise é importante para elucidar, num conjunto textual, quais são os temas mais trabalhados e permitir que o pesquisador compreenda quais são os principais assuntos de determinado conjunto de dados textual.

5 CONSIDERAÇÕES FINAIS

Os estudos que envolvem mecanismos de processamento textual, com o propósito de obter resultados numéricos e gráficos, sobre os temas desenvolvidos, surgem como alternativa de pesquisa no âmbito da CI. Eles podem fornecer resultados diferenciados no momento em que utilizam teorias e metodologias das áreas da estatística e computação para realizar análise textual e obter representações e visualizações que facilitam a compreensão de um determinado tema.

Tais ferramentas, tornam-se instrumento de apoio às etapas de análise, nas quais, o pesquisador está submetido. As figuras resultantes das análises podem elucidar questões que não conseguiriam ser identificadas e servem para ratificar os argumentos propostos nas pesquisas.

Ao serem analisados os campos do título, resumo e palavras-chave dos 82 artigos indexados na BRAPCI por meio do software IRAMUTEQ, foi possível esclarecer sobre a relação temática existente nos artigos e o tema “Biblioteca digital”, solucionando parcialmente o problema do descontrole terminológico da base de dados que contém os artigos. Percebeu-se até o momento, que os artigos se voltam para análise de bibliotecas digitais no formato de bibliotecas de teses e dissertações. Essa situação pode ser enumerada na figura 2, no momento em que surgem os termos compostos, “Biblioteca digital de teses e dissertações”, “teses e dissertações”, “teses e dissertações eletrônicas”, com os 3 termos compostos, em

nível de proximidade semântica muito próximos, e em grau de conexidade, ao termo "Biblioteca digital". Quanto aos índices conexidade tais termos apresentam respectivamente os seguintes valores como índice de proximidade e similitude: 9; 12; 5.

Nesse ínterim, a opção metodológica de escolher a palavra-chave "Biblioteca digital" como prerrogativa de seleção dos artigos, trouxe maior confiabilidade sobre a recuperação de assuntos relacionados ao tema, pois a preferência de determinada palavra-chave é feita pelo autor no momento em que está indexando o artigo científico.

A recente aplicação do IRAMUTEQ nos estudos desenvolvidos nas ciências humanas e sociais legitimam o grau de importância ao qual vem agregando nas áreas, sobretudo nas áreas da saúde como a Enfermagem. Neste sentido, a CI pode desfrutar de um método seguro que vem sendo testado e ratificado pelos pares de diversas áreas. Essa segurança científica é importante pois promove uma maior confiabilidade nas análises e credibilidade aos argumentos propostos.

As opções gráficas que foram utilizadas no texto trouxeram maior segurança às análises estatísticas obtidas. Portanto as análises aqui propostas apontaram como o tema da "Biblioteca digital" vem se conectando com outras temáticas.

Dessa maneira, fica evidente os benefícios das análises aqui propostas e pretende-se, com este trabalho, promover a disseminação do uso de ferramentas que possam trazer melhorias na realização de estudos métricos temáticos da área da CI. O IRAMUTEQ se apresenta como mais um instrumento em proveito da mineração de informações textuais e pode servir para outros desafios que outros pesquisadores se propuserem a enfrentar.

Como possíveis limitações, foi identificado a dificuldade de analisar grupos de termos que possuíssem representação semântica. Percebeu-se que o software é excelente para palavras isoladas porém possui limitação na análise de palavras compostas. Além disso, foi preciso realizar um exaustivo trabalho na construção do corpus de análise exigido pelo software. Outro ponto limitante foi a dificuldade em ter a possibilidade de customizar os gráficos gerados. Desta maneira, como sugestão de estudos futuros, pretende-se aprofundar nas análises estatísticas do método e realizar outros tipos de análise no âmbito da CI, além disso, sugere-se que sejam utilizadas outras ferramentas, com fins de comparação entre os resultados.

REFERÊNCIAS

BORGMAN, C. L. What are digital libraries? Competing visions. **Information Processing and Management**, Los Angeles, v.35, n.3, p. 227-243, 1999. Disponível em: < https://pdfs.semanticscholar.org/d0e6/90b74b3b9513d9d1f97cf366e31a3920a4bf.pdf >. Acesso em 15 jul. 2018.

CAMARGO, B. V.; JUSTO, A. M. IRAMUTEQ: Um software gratuito para análise de dados textuais. **Temas em psicologia**, Ribeirão preto, v.21, n.2, p. 513-518, 2013. Disponível em: <http://pepsic.bvsalud.org/pdf/tp/v21n2/v21n2a16.pdf >. Acesso em 10 nov. 2017.

CUNHA, M. B. Biblioteca Digital: bibliografia internacional anotada. **Ciência da Informação**, Brasília, v.26, n.2, 1997. Disponível em: < http://www.scielo.br/pdf/ci/v26n2/v26n2-12.pdf >. Acesso em 11 abr. 2018.

FOX, E. et al. The Networked Digital Library of Thesis and Dissertations: changes in the university community. **Springer**, Virginia, v.13, n.2, p.102-124. 2002. Disponível em: < https://link.springer.com/article/10.1007/BF02940968 >. Acesso em: 10 jun. 2018.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2009.

GUEDES, V. N. L. S. A bibliometria e a gestão da informação e do conhecimento científico e tecnológico: uma revisão da literatura. **Ponto de Acesso**, Salvador, v.6, n.2, p.74-109, 2012. Disponível em: < https://portalseer.ufba.br/index.php/revistaici/article/view/5695/4591 >. Acesso em 4 maio 2018.

FERREIRA JÚNIOR, A. A. **Princípios para análise do uso de sistemas de informação**. 2017. 221 f. Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação da Universidade de São Paulo, Universidade de São Paulo (USP), São Paulo, 2017. Disponível em: < http://www.teses.usp.br/teses/disponiveis/27/27151/tde-07072017-113025/pt-br.php >. Acesso em 24 jul. 2018.

LIMA, T. V. **Professores de Matemática da Rede Estadual em Goiânia:** TDCI em perspectiva. 2017. 203f. Dissertação (Mestrado) – Programa de Mestrado em Educação em Ciências e Matemática, Mestrado em Educação em Ciências e Matemática, Universidade Federal de Goiás (UFG), Goiás, 2017. Disponível em: < https://repositorio.bc.ufg.br/tede/handle/tede/7925 >. Acesso em: 20 nov. 2017.

MARCHAND, P.; RATINAUD, P. L'analyse de similitude appliquq´q aux corpux textueles: les primaires socialistes pour l'election présidentielle française. **JADT,** Liège, v.1, n.1 p.687-699, abril. 2012. Disponível em: < http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Marchand,%20Pascal%20et%20al.%20-%20L%27analyse%20de%20similitude%20appliquee%20aux%20corpus%20textuels.pdf >. Acesso em: 5 nov. 2017.

MOLINOS, D. N. et al. ZipfTool: Uma ferramenta bibliométrica para auxílio na pesquisa teórica. **Revista Informática e Teórica Aplicada**, Porto Alegre, v.23, n.1 , p.293-317, maio. 2016.

Disponível em: < http://seer.ufrgs.br/index.php/rita/article/view/RITA-VOL23-NR1-293 >. Acesso em 5 jan. 2018.

OHIRA, M. L. B.; PRADO, N. S. Bibliotecas virtuais e digitais: análise de artigos de periódicos brasileiros (1995/2000). **Ciência da Informação,** Brasília, v.31, n.1, p.61-74, jan./abr. 2002. Disponível em: < http://www.scielo.br/pdf/ci/v31n1/a07v31n1.pdf >. Acesso em: 13 jun. 2018.

REZENDE, S. O. **Sistemas Inteligentes: fundamentos e aplicações**. Editora Manole: 2003.

SAYÃO, L. F. Afinal, o que é biblioteca digital? **Revista USP**, São Paulo, v.80, n.1, p.6-17, fev. 2009. Disponível em: < http://www.revistas.usp.br/revusp/article/view/13709/15527 > . Acesso em: 2 nov. 2017.

SANTOS, C. A. C. M. Organização e representação do conhecimento: bibliometria temática em artigos de periódicos brasileiros. **Revista Brasileira de Biblioteconomia e Documentação**, Bela Vista São Paulo, v.11, n.especial, p.640-653, 2015. Disponível em: < https://rbbd.febab.org.br/rbbd/article/view/494/458 >. Acesso em: 26 jul. 2017.

TAMMARO, A. M.; SALARELLI, A. **A Biblioteca Digital**. Brasília: Briquet de Lemos, 2006.

VIEIRA, J. M. L.; CÔRREA, R. F. Visualização da Informação na Construção de Interfaces Amigáveis para Sistemas de Recuperação de Informação. **Encontros Bibli**, Florianópolis, v.16, n.32, p.73-93, set. 2011. Disponível em: < https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2011v16n32p73 >. Acesso em: 3 jul. 2018.