organizadores
Thiago Henrique Bragato Barros
Rita do Carmo Ferreira Laipelt

Organização e Representação do Conhecimento em Múltiplas Abordagens

I São Paulo I 2022 I pimenta cultural

Direção editorial	Patricia Bieging e Raul Inácio Busarello
Editora executiva	Patricia Bieging
Coordenadora editorial	Landressa Rita Schiefelbein
Marketing digital	Lucas Andrius de Oliveira
Diretor de criação	Raul Inácio Busarello
Assistente de arte	Naiara Von Groll
Editoração eletrônica	Peter Valmorbida e Potira Manoela de Moraes
Imagens da capa	Bizkette1, Starline - Freepik.com
Tipografias	Swiss 721, Aileron, Libel Suit
Revisão	Maria Amália Cassol Lied
Organizadores	Thiago Henrique Bragato Barros Rita do Carmo Ferreira Laipelt

Dados Internacionais de Catalogação na Publicação (CIP)

O68

Organização e representação do conhecimento em múltiplas abordagens / Organizadores Thiago Henrique Bragato Barros, Rita do Carmo Ferreira Laipelt. – São Paulo: Pimenta Cultural, 2022.

Livro em PDF

ISBN 978-65-5939-561-3
DOI 10.31560/pimentacultural/2022.95613

1. Organização do conhecimento. 2. Metodologia. 3. Arquivologia. I. Barros, Thiago Henrique Bragato (Organizador). II. Laipelt, Rita do Carmo Ferreira (Organizadora). III. Título.

CDD 020

Índice para catálogo sistemático:

I. Organização do conhecimento

Janaina Ramos – Bibliotecária – CRB-8/9166

PIMENTA CULTURAL
São Paulo · SP
Telefone: +55 (11) 96766 2200
livro@pimentacultural.com
www.pimentacultural.com

pimenta cultural
2 0 2 2

6

Luciana Monteiro-Krebs

Rita do Carmo Ferreira Laipelt

Rafael Port da Rocha

Metodologia de análise de *logs* na ciência da informação: revisão de literatura e melhores práticas

Logs analysis as methodology for studies on information science: literature review and best practices

DOI: 10.31560/pimentacultural/2022.95613.6

Resumo:
Em um cenário de crescente acesso à informação através de plataformas digitais e sistemas de informação, pesquisas envolvendo a análise de *logs* tornam-se cada vez mais relevantes. Através dos *logs*, pesquisadores e profissionais da informação podem conhecer as demandas dos usuários com quem não mantêm contato direto, em função da prestação de serviço se dar remotamente, mediada via sistema de recuperação da informação. No entanto, estudos que utilizam a análise de *logs* como método no Brasil são ainda incipientes, e descrições sistemáticas dos procedimentos metodológicos na literatura nacional sobre o tema inexistem. Para preencher essa lacuna, o objetivo deste capítulo é investigar, sistematizar e descrever as etapas metodológicas e aspectos técnicos da análise de *logs* em CI, com enfoque em sistemas de recuperação da informação. Concretiza-se esse objetivo através da sistematização da seção metodológica dos 10 artigos mais citados disponíveis na *Web Of Science* que tenham feito uso de análise de *logs,* publicados no período entre 2006 e 2016. A proposta metodológica apresenta quatro etapas principais: (i) contextualização do *log*; (ii) seleção; (iii) coleta e preparação dos dados; e (iv) análise dos dados. A etapa de seleção (ii) é ainda dividida em três subetapas, nas quais se identificam a relevância e adequabilidade das informações do *log* aos objetivos de pesquisa, o recorte de dados e a disponibilização e uso dos dados para análise. Os procedimentos são apresentados com exemplos tanto dos textos do *corpus* de pesquisa quanto de estudos mais recentes baseados na experiência dos autores do presente capítulo.

Palavras-chave: Análise de *logs*; metodologia da pesquisa; sistemas de recuperação da informação; estudos de usuário.

Abstract:
In a scenario of increasing access to information through digital platforms and information systems, research involving log analysis becomes increasingly relevant. Through system logs, researchers and librarians can learn about the demands of users with whom they do not have direct contact, due to the fact that the informational service is provided remotely, usually mediated via an information retrieval system. However, studies that use log analysis as a method in Brazil are still incipient, and systematic descriptions of methodological procedures in the national literature on the subject do not exist. To fill this gap, the objective of this chapter is to investigate, systematize and describe the methodological steps and technical aspects of log analysis in Information Science with a focus on information retrieval systems. This objective is achieved through the systematization of the methodological section of the 10 most cited articles available on the Web Of Science that have made use of log analysis, published between 2006 and 2016. The methodological proposal has four main steps: (i) contextualization of the log; (ii) selection; (iii) data collection and preparation; and (iv) data analysis. The selection step (ii) is further divided into three sub-steps, being the relevance and suitability of the log information to the research objectives, the methodological cut and the availability of data for analysis. The procedures are presented with examples both from the texts of the research corpus and from more recent studies based on the experience of the authors of this chapter.

Keywords: *Log analysis; research methodology; information retrieval systems; user studies.*

1 INTRODUÇÃO

No Brasil, cerca de 152 milhões de indivíduos são usuários da internet, o equivalente a 81% da população com 10 anos ou mais (CETIC, 2021). Esse índice, crescente principalmente em centros urbanos, possibilita às unidades de informação no país ampliarem o conhecimento a respeito de seus usuários, expressos através de seu comportamento online. As interações do usuário com o sistema aplicativo são gravadas virtualmente em arquivos de *log*, registrando, para cada requisição, a data, o horário, o local de acesso (obtido via número IP do computador que realizou o acesso), assim como a ação determinada pela requisição e sua situação, entre outras informações. A análise de tais dados é importante para a melhoria dos serviços disponibilizados e o atendimento às necessidades informacionais dos usuários em um cenário cada vez menos presencial de acesso à informação. No que tange a estudos de usuário na Ciência da Informação (CI), a análise dos registros de uso em plataformas digitais (registros em *logs*) oferece a possibilidade de expandir tanto a quantidade de usuários incluídos nas amostras quanto a granularidade da informação disponível.

No entanto, a literatura da CI a respeito do uso dos registros de interações dos usuários em bases de dados, sites e/ou catálogos online ainda é incipiente no Brasil. Por exemplo, uma consulta[47] sobre o tema na base de dados BRAPCI[48] em junho de 2021 retornou apenas cinco publicações. Em sua maioria, os estudos sobre análise de *logs* na CI brasileira apresentam resultados de estudos empíricos, com pouco ou nenhum aprofundamento metodológico. O trabalho de Laipelt (2015) e Monteiro-Krebs, Rocha e Ribeiro (2017) são exemplos de pesquisas

47 A expressão de busca utilizada foi "análise de *logs*", no campo Resumo.

48 BRAPCI (http://www.brapci.ufpr.br/brapci/index.php/home) é uma base que indexa 57 periódicos científicos brasileiros na área de Ciência da Informação com cobertura temporal desde 1972.

realizadas no Brasil que analisam interações entre sistemas de recuperação da informação e usuários. Laipelt (2015) analisou expressões de busca dos usuários do Portal LexML (Senado Federal Brasileiro) para demonstrar o potencial dos *logs* como fonte de coleta de dados para a escolha de descritores para a representação da informação. Monteiro-Krebs, Rocha e Ribeiro (2017) analisaram o uso de um sistema de recomendação para catálogos on-line de bibliotecas universitárias. Já Aires (2003) descreveu a análise de *logs* e como ferramenta para incrementar a qualidade dos resultados das máquinas de busca, porém não realizou uma análise de *logs* propriamente dita, por não ter obtido acesso aos *logs* das máquinas de busca comerciais. O que de fato foi realizado foi um estudo das interações de usuários com motores de busca a partir de relatórios de consultas redigidos pelos próprios sujeitos da pesquisa. Efetivamente, esse trabalho não poderia ser considerado uma análise de *logs*, pois o *log* propriamente dito não foi utilizado como fonte de dados.

Observa-se, portanto, uma carência em publicações que tragam orientações metodológicas para pesquisas científicas focadas na análise de *logs*, em especial com relação aos procedimentos técnicos para a seleção, a coleta e a análise destes. O presente capítulo oferece subsídios metodológicos para pesquisadores e profissionais da informação interessados em implementar a análise de *logs* através de análise sistemática das principais publicações sobre o assunto e uma proposta metodológica com base nas lições aprendidas dessas publicações.

Tendo em vista o potencial do uso de *logs* em pesquisas no campo da CI, o objetivo deste capítulo é investigar, sistematizar e descrever as etapas metodológicas e aspectos técnicos da análise de *logs* em CI com enfoque em sistemas de recuperação da informação. O procedimento metodológico é proposto com base em uma revisão de literatura, utilizando como fonte pesquisas reconhecidas pela comunidade científica que utilizaram *logs*. Desta forma, colhemos as melhores práticas em estudos da CI com alto volume de citações que utilizam a análise de

logs como método. Ao encontrar diferentes descrições, buscamos identificar, analisar e discutir os procedimentos apresentados pelos autores na tentativa de sistematizá-los e apresentá-los de forma convergente. Essa contribuição visa oferecer subsídios aos profissionais e pesquisadores da área fornecendo, também, exemplos práticos de aplicação da análise de *logs* em benefício das unidades de informação no país, além de identificar um panorama de como a questão dos procedimentos técnicos relacionados à análise de *logs* é abordada nas pesquisas.

Este capítulo está organizado da seguinte forma. A seção 2 caracteriza a análise de *logs* no contexto da recuperação da informação em sistemas de informação, apresentando a definição de análise de *logs* e introduzindo diferentes tipos de análise. A seção 3 apresenta a metodologia, identificando as pesquisas na área da CI que utilizam análise de *logs* e que apresentam os procedimentos técnicos utilizados com relação aos *logs*, tendo como fonte a plataforma *Web of Science*. A seção 4 apresenta etapas que caracterizam procedimentos metodológicos na análise de *logs*, delineadas a partir da análise das pesquisas identificadas na seção 3. O capítulo é encerrado com uma breve conclusão na seção 5.

2 ANÁLISE DE *LOGS* NA RECUPERAÇÃO DA INFORMAÇÃO

Trabalhos empíricos utilizando *logs* começaram sendo chamados de *web searching studies* (JANSEN; POOCH, 2000; DAVIS, 2004), *search engine transaction log studies*, *log analysis*, entre outros. Ao longo do tempo, alguns termos foram se consolidando na literatura da CI, como é o caso da análise do *log* de transações (*transaction log analysis - TLA*). "A análise do *log* de transações é o uso de dados coletados em um *log* de transações para investigar uma questão de pesquisa

específica relacionada ao usuário, ao sistema ou ao conteúdo" (SPINK; JANSEN, 2005, p. 36). As transações são todas as interações entre o usuário e o sistema, que não precisam, necessariamente, dar-se no ambiente web. Para estas últimas, Jansen, Taksa e Spink (2009) usam o termo *Web log analysis*.

> A análise do log de transações é uma ampla categorização de métodos que abrange várias subcategorizações, incluindo análise de logs da Web (ou seja, análise de logs do sistema da Web), análise de blog e análise de logs de pesquisa (análise de logs de mecanismos de busca) (JANSEN; TAKSA; SPINK, 2009, p. 2).

Segundo Jansen (2006, p. 408), "[...] um *log* de transações é um registro eletrônico de interações que ocorreram durante um episódio de pesquisa entre um mecanismo de pesquisa da Web e usuários pesquisando informações nesse mecanismo de pesquisa da Web.". Um *log* de transações, segundo Jansen (2006), registra a comunicação entre os usuários e um sistema em um arquivo. Os *logs* "[...] representam os usuários, são pegadas de informação digital" (NICHOLAS; HUNTINGTON; WATKINSON, 2005, p. 250).

Na literatura, no entanto, não há convergência para essa classificação. Diferentes termos designam técnicas operacionalmente distintas entre si, mas, em linhas gerais, contempladas pela TLA. Assim como Jansen, Taksa e Spink (2009) por vezes utilizam *log analysis* para se referir à TLA, Nicholas, Huntington e Watkinson (2005) também utilizam *log analysis* quando se referem a *deep log analysis* (DLA - análise profunda de *logs*), que é diferente de TLA, embora, ao analisar os métodos descritos em seus artigos e artigos citados por eles, percebe-se que a *deep log analysis* é, na prática, um aprimoramento da TLA. A DLA agrega, aos dados de pesquisa/navegação obtidos via *log,* informações contextuais relativas aos usuários, que podem ser demográficas ou pessoais extraídas a partir de outras técnicas de coletas de dados (questionários, entrevistas, dados de *login*/identificação etc.), com isso, tem-se a caracterização como análise "profunda" de *logs*.

Um exemplo desse tipo de estudo é o realizado por Nicholas *et al.* (2006a) que investigam o comportamento de busca de informações de quase três milhões de usuários à medida que exploram o site[49], o número de visitas realizadas, bem como o tipo de itens e o conteúdo visualizado em duas bibliotecas digitais de periódicos, a EmeraldInsight e a Blackwell Synergy. O que caracteriza esse estudo como DLA é a combinação desses dados com a identificação de perfis de usuários por profissão, local de trabalho, tipo de assinatura do periódico, localização geográfica, tipo de universidade, número de itens visualizados em uma sessão etc. (NICHOLAS *et al.,* 2006a, p. 1345). Uma excelente revisão dos estudos de pesquisa na Web (*Web searching studies*) é resumida por Jansen e Pooch (2001). Para acessar um histórico mais antigo de TLA, recomendamos consultar Peters (1993).

Na busca de informação, o usuário pode utilizar filtros e categorias que determinem o conjunto de resultados que ele quer obter, para além da expressão de busca. Após realizar a busca, o usuário navega pelos resultados, selecionando as obras que decidir ler ou acessar o registro. É a partir dos registros gerados por essas interações entre os usuários e o sistema que a análise de *logs* se faz possível. As expressões de busca, em que itens dos resultados o usuário clicou, quais documentos foram visualizados e/ou baixados e todos os demais dados podem ser usados como variáveis para pesquisas no campo da Ciência da Informação.

Através de estudos com análise de *logs*, pode-se identificar quais métodos de recuperação da informação são preferidos pelos usuários de determinadas áreas do conhecimento. No ambiente de informação eletrônica, estudos de usuários de periódicos on-line apontam a preferência geral dos usuários para pesquisar por assunto em bancos de dados ao invés de navegar pelos periódicos (CHEN, 2010;

49 O número de itens ou páginas visualizadas em uma sessão é chamado de "site penetration", que traduzimos como "penetração no site" (ver seção 4.1) (NICHOLAS *et al.*, 2006a; NICHOLAS *et al.*, 2007).

MENG-XING; CHUN-XIAO, YONG, 2010). O estudo de Vakkari e Talja (2006) identifica essa tendência, principalmente para ciências naturais e medicina, em relação a outras disciplinas. Além da busca e da navegação, outros métodos também se mostram relevantes para a RI, (Recuperacao da Informação) como o encadeamento de citações, que se mostrou "[...] um método de pesquisa significativamente mais relevante em economia e engenharia em comparação com humanidades e medicina" (VAKKARI; TALJA, 2006, documento eletrônico).

Apesar de haver distinção conceitual entre busca de informação e navegação, para a análise de *logs*, ambas são interessantes e podem auxiliar nos estudos que visam à melhoria dos serviços informacionais oferecidos pela instituição. Na busca de informação (ou pesquisa), podem-se analisar os termos que compõem a expressão de busca – no intuito de melhorar o vocabulário controlado e rede de remissivas, assim como identificar tendências temáticas de pesquisa. Na navegação, por outro lado, pode-se analisar a jornada do usuário, que itens ele visualizou e em quais ele clicou etc. – no intuito de melhorar a estrutura do site, formatos de conteúdo, caminhos internos.

Recentemente, outros tipos de análises que incluem informações oriundas de *logs* começaram a ganhar popularidade entre gestores de sistemas e bibliotecas, chamando a atenção de pesquisadores também. Com o advento de ferramentas de análise estatística (como AWStat, Google Analytics, entre outras), os gestores passam a ter acesso a gráficos e dashboards com resumos de tráfego para os sites que gerenciam. Essas ferramentas leem os *logs* do sistema e mostram os resultados em formas de tabelas e até gráficos, o que facilita a tomada de decisão no que tange a potenciais melhorias do sistema. No entanto, tais ferramentas são diferentes da análise de *logs* como concebida em nosso trabalho (DLA e TLA) por três motivos:

Primeiro, o pesquisador não possui acesso ao *log* em si, apenas aos resultados por meio da análise estatística da ferramenta. Isso

limita as informações que a ferramenta apresenta, ou seja, nem todas as variáveis são computadas nem são passíveis de cruzamentos. A decisão do que é ou não apresentado é feita anteriormente à consulta do pesquisador, ainda na fase de *design to dashboard*, e não há a possibilidade de o pesquisador escolher quais variáveis deseja cruzar para atender seus objetivos de pesquisa. Por exemplo, se a ferramenta determina que o IP dos acessos será apagado dos registros para preservar a privacidade dos usuários, o pesquisador não poderia se valer dessa informação (IP) para analisar a jornada de um usuário específico. Apenas dados agregados estariam disponíveis, permitindo somente estudos de conjuntos de usuários, o que é mais interessante para estudos quantitativos e menos valioso para análises qualitativas.

Segundo, em geral, essas ferramentas também não documentam exaustivamente de onde vêm os valores apresentados nas estatísticas. Assume-se que as informações provêm dos *logs*, porém não é possível encontrar nenhuma documentação que indique qual campo do *log*/http é utilizado em cada análise estatística. Nas análises TLA e DLA, esses dados são conhecidos pelos pesquisadores, até por uma questão de transparência metodológica e replicabilidade.

Terceiro, os relatórios estatísticos também reúnem dados de outras fontes, como, por exemplo, o site visitado pelo usuário antes de chegar na *home* do sistema analisado. Através de APIs, data e hora de consulta são incorporadas à análise e, em alguns casos, até expressões inseridas nos motores de busca são oferecidas nos relatórios. Todas essas características impedem o pesquisador de isolar e cruzar variáveis, além de estabelecer livremente os recortes de dados de acordo com os seus objetivos de pesquisa. As ferramentas de análise estatística, como Google Analytics e AWStat, estão limitadas ao que os relatórios que essas empresas definem previamente e, portanto, não se encaixam na definição de análise de *logs* conforme a metodologia aqui descrita.

3 PESQUISAS EM ANÁLISE DE *LOGS* NA CIÊNCIA DA INFORMAÇÃO

Várias pesquisas mencionam que seus estudos foram feitos a partir da análise de *logs*. Entretanto, essas pesquisas pouco relatam os procedimentos utilizados com relação aos *logs*. Esta seção identifica e caracteriza pesquisas que usam *logs*.

Realizou-se uma busca na bibliografia especializada em que a análise de *logs* consta como assunto principal da publicação ou foi mencionada junto aos métodos em pesquisas na Ciência da Informação. Optou-se pela busca na *Web of Science* pela abrangência da base, pelos recursos de recuperação e filtragem dos conteúdos e pela facilidade em gerar relatórios preliminares para rápida visualização de características do *corpus*. A coleta foi realizada em 20 de agosto de 2017. Foi realizada busca na plataforma *Web of Science* por trabalhos sobre análise de *logs* ou que usam análise de *logs* como metodologia, na área da Ciência da Informação[50]. A maior parte dos artigos encontrados são estudos aplicados.

A triagem dos textos selecionados obedeceu a dois critérios: impacto (número de citações) e adequação ao escopo desse artigo. Após a busca na base, ordenamos a lista de resultados (141 artigos) por número de citações (começando pelo mais citado). O artigo mais citado da lista contém 297 citações e o menos citado, uma citação.

A partir desse extrato, utilizamos como segundo filtro os objetivos do presente capítulo. Nosso intuito é criar um protocolo metodológico para auxiliar profissionais e pesquisadores da Ciência da

50 Expressão de busca utilizada na Web of Science: Você pesquisou por: **Tópico:** ("log analysis" OR "análise de logs" OR TLA OR "transaction log analysis") *OR* **Título:** ("log analysis" OR "análise de logs" OR TLA OR "transaction log analysis") **Refinado por: Categorias do Web of Science:** (INFORMATION SCIENCE LIBRARY SCIENCE) **Tempo estipulado:** 2006-2016. Índices: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI. **Resultados: 141** (de Principal Coleção do Web of Science).

Informação a realizar estudos utilizando análise de *logs*. Assim, faz sentido nos guiarmos pelos materiais que trouxessem embasamento teórico sobre os métodos e, por isso, artigos exclusivamente de relatos de pesquisa foram desconsiderados. Como a maioria está relatando estudos aplicados a partir do uso da análise de *log*, selecionamos os que mais traziam definições de *log*, das variáveis que o compõem e as etapas do processo de pesquisa. O quadro abaixo mostra os artigos que serviram de fundamento para a sistematização apresentada neste capítulo. Ele fornece os dados dos artigos analisados e como eles foram utilizados nessa revisão bibliográfica.

Quadro 1 – Artigos mais citados entre os que utilizam análise de *logs*

Título	Autor(es)	Periódico	Data	# citações	Utilizado na seção
How are we searching the World Wide Web? A comparison of nine search engine transaction logs	Jansen, BJ; Spink, A.	INFORMATION PROCESSING & MANAGEMENT. v. 42, n. 1, p. 248-263	JAN 2006	297	4.3
Search log analysis: What it is, what's been done, how to do it	Jansen, BJ.	LIBRARY & INFORMATION SCIENCE RESEARCH, v. 28, n. 3, p. 407-432	2006	81	4.1 4.2 4.4
Defining a session on web search engines	Jansen, BJ.; Spink, A.; Blakely, C.; Koshman, S.	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, v. 58, n. 6, p. 862-871	APR 2007	63	4.1
The information seeking behaviour of the users of digital scholarly journals	Nicholas, D.; Huntington, P.; Jamali, HR.; Watkinson, A.	INFORMATION PROCESSING & MANAGEMENT, v. 42, n. 5, p. 1345-1365	2006a	40	4.3

E-textbook use, information seeking behaviour and its impact: Case study business and management	Nicholas, D.; Rowlands, I.; Jamali, HR.	JOURNAL OF INFORMATION SCIENCE, v. 36, n. 2, p. 263-280	APR 2010	37	4.2.1 4.4
What deep log analysis tells us about the impact of big deals: case study OhioLINK	Nicholas, D.; Huntington, P.; Jamali, HR.; Tenopir, C.	JOURNAL OF DOCUMENTATION, v. 62, n. 4, p. 482-508	2006b	32	4.4
Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the 'bouncer'	Nicholas, D.; Huntington, P.; Jamali, HR.; Dobrowolski, T.	INFORMATION PROCESSING & MANAGEMENT, v. 43, n. 4, p. 1085-1102	2007	22	4.1
Online use and information seeking behaviour: institutional and subject comparisons of UK researchers	Nicholas, D.; Clark, D.; Rowlands, I.; Jamali HR.	JOURNAL OF INFORMATION SCIENCE, v. 35, n. 6, p. 660-676	2009	18	4.3
Empirical observations on the session timeout threshold	Huynh, T.; Miller, J.	INFORMATION PROCESSING & MANAGEMENT, v. 45, n. 5, p. 513-528	2009	9	4.1
Library and information resources and users of digital resources in the humanities	Warwick, C.; Terras, M.; Galina, I.; Huntington, P.; Pappa, N.	PROGRAM-ELECTRONIC LIBRARY AND INFORMATION SYSTEMS, v. 42, n. 1, p. 5-27	2008	9	4.3

Fonte: elaborado pelos autores.

A partir dos artigos selecionados, realizamos a sistematização dos conteúdos com foco nas principais contribuições potenciais para a CI. Analisando os estudos, observou-se a relevância de estágios anteriores à coleta de dados para a análise de *logs*, conforme destacado pelos trabalhos de Nicholas *et al.* (2005), Jansen (2006), Nicholas *et al.* (2007), Jansen *et al.* (2007) e Huynh e Miller (2009). Por exemplo, Nicholas *et al.* (2005) explicam como a análise de *logs* se diferencia de outros métodos de pesquisa, ao refletir

ações dos usuários sem a interferência do pesquisador. Os demais autores apresentam definições de diversos conceitos relevantes, como, sessão, consulta, penetração no site e itens visualizados. Ao sistematizar os procedimentos metodológicos, essas informações foram incorporadas na etapa de contextualização do *log* (seção 4.1).

Também se observou como os autores explicam a etapa de seleção dos *logs* (seção 4.2), o que é destacado por Jansen (2006) e Nicholas, Rowlands e Jamali (2010). Especificamente, três subetapas da seleção foram identificadas. Primeiro, como o pesquisador pode compreender a relevância e adequabilidade das informações do *log* (seção 4.2.1). Nessa fase, a sistematização traz exemplos de linhas de *log* e informações fornecidas pelo registro. Segundo, como o recorte de dados pode ser feito dependendo dos objetivos da pesquisa (seção 4.2.2). E terceiro, a disponibilização e uso dos dados (seção 4.2.3), quando ocorre a negociação com os responsáveis pela guarda do *log* para acesso a estes. Cuidados com a privacidade dos usuários são necessários e explicados nesta seção.

Diversos trabalhos entre os mais citados apontam ainda como a etapa de coleta e preparação de dados (seção 4.3) se dá empiricamente. A sistematização traz exemplos de possibilidades de análise de expressões de busca (JANSEN; SPINK, 2006) e análises de diversas outras variáveis, especialmente aplicáveis para DLA, conforme destacado por Nicholas *et al.* (2006a; 2009) e Warwick *et al.* (2008).

Finalmente, a análise de dados (seção 4.4) é explorada. Apresenta-se como os estudos de *logs* podem contribuir para melhoria de sistemas de recuperação da informação, com exemplos de aplicabilidade baseados em Jansen (2006). Evidencia-se também a possibilidade de cruzamento de variáveis do *log* com dados de diferentes fontes, como questionário e grupo focal. Exemplos desses cruzamentos (típico de DLA) e os resultados esperados são demonstrados com citações de Nicholas, Rowlands e Jamali (2010) e Nicholas *et al.*, (2006b). As limitações dos estudos com *logs* (JANSEN, 2006) são apresentadas na conclusão.

4 SISTEMATIZAÇÃO DOS PROCEDIMENTOS TÉCNICOS DE ANÁLISE DE LOGS NA CIÊNCIA DA INFORMAÇÃO

A análise de *logs* não é um método simples de ser empregado, por dois motivos principais. O primeiro é a dificuldade de acesso à matéria-prima, haja vista que diversos cuidados devem ser empregados para garantir a segurança da informação e o respeito à privacidade dos usuários cujas interações estão registradas nos *logs*. A anonimização de IP é um exemplo. O segundo motivo é o conhecimento técnico que se demanda para manusear os *logs*. É necessário entender minimamente de programação para entender a linguagem e a estrutura dos arquivos que se vai receber.

Porém, informações extraídas dessa fonte são valiosas por, pelo menos, duas razões. Em primeiro lugar, a confiabilidade dos *logs.* Usar *logs* como fonte de dados para conhecer o comportamento informacional elimina qualquer eventual interferência do pesquisador sobre o usuário no momento da consulta e acesso à informação. As informações dos *logs* são fiéis aos fatos, pois provém um registro direto e imediato do que as pessoas de fato fizeram durante a interação com o sistema, "[...] não o que dizem que poderiam ou gostariam de fazer; nem o que eles foram levados a dizer, tampouco o que eles pensam que fizeram" (NICHOLAS *et al.*, 2005, p. 1445). Em segundo lugar, as pessoas têm dedicado cada vez menos tempo para participar de pesquisas e preencher questionários (LAIPELT, 2015). Portanto, através dos *logs*, é possível coletar grandes volumes de dados com riqueza de detalhes de forma quase instantânea, sem as limitações de tempo e espaço impostas pela disponibilidade dos participantes da pesquisa para questionários ou entrevistas, por exemplo.

Para facilitar essa tarefa, nessa seção apresenta-se uma sistematização dos procedimentos e aspectos técnicos utilizados em análise de *logs*, a partir do exame dos trabalhos apresentados no Quadro 1. Essa sistematização inicia pela compreensão das interações entre usuário e sistema aplicativo e o registro dessas interações em arquivos de *log*, prosseguindo por aspectos e procedimentos que envolvem seleção, coleta, preparação e análise.

Segundo Jansen (2006), a análise de *logs* envolve três etapas principais: (i) coleta: processo de recolha dos dados de interação de um determinado período em um *log* de transações; (ii) preparação: processo de limpeza e organização dos dados de *log* de transações; e (iii) análise: processo de exame dos dados com vistas a alcançar os resultados da pesquisa. Nesse estudo, adiciona-se às etapas de Jansen a contextualização do *log*, na qual os *logs* são primeiramente compreendidos pelo pesquisador no contexto das interações entre usuários e Sistemas de Recuperação da Informação (SRI), e a seleção, na qual são feitas análises prévias do escopo e das viabilidades do uso de um *log* em uma pesquisa. Essas duas questões são fundamentais para a tomada da decisão sobre o uso ou não de um *log* e para minimizar riscos que possam vir a ocorrer em passos subsequentes da pesquisa. Elas foram propostas por este estudo à medida que se observou nos artigos analisados preocupações em demonstrar *logs* e seu funcionamento, assim como as viabilidades e as possibilidades do uso desses *logs* nos estudos.

Partindo dessas etapas, a seguir apresenta-se a proposta metodológica baseada em uma sistematização da literatura sobre a análise de *logs*, considerando, além das etapas previstas por Jansen, outras questões que se consideram relevantes do ponto de vista de diversos autores que são citados ao longo do texto.

4.1 CONTEXTUALIZAÇÃO DO LOG

A contextualização compreende a percepção, por parte do pesquisador que utilizará *logs* em suas análises, do contexto em que ocorrem as interações entre o sistema aplicativo, os usuários e o registro dessas interações no *log*. No contexto dos sistemas aplicativos que operam na plataforma Word Wide Web, um *log* representa o registro das interações entre usuários e sistemas que rodam em Servidores Web, que, no caso desse capítulo, são SRI. Jansen (2006, p. 419) define interação como "[...] qualquer troca específica entre o pesquisador e o sistema (ou seja, enviar uma consulta, clicar em um hiperlink etc.)".

Ao realizar suas buscas e ao navegar pelos resultados dessas buscas, um usuário de um SRI realiza várias interações, determinadas através de requisições ao servidor. Por exemplo, ao preencher e enviar um formulário que contém uma expressão de busca, o usuário está realizando uma requisição, que terá como resposta o resultado da consulta expressa no formulário. Ao clicar em um item do resultado dessa consulta, o usuário está realizando uma outra requisição. Essa requisição terá como resposta a visualização das informações bibliográficas do item.

As interações ocorrem através do protocolo HTTP, que é um protocolo de requisição-resposta entre o cliente (usuário) e o servidor (sistema). Os *logs* são arquivos situados no servidor em que ficam registradas informações referentes às requisições. *Common Log Format* (CFL) é o formato básico para representação de arquivos de *logs*. *Logs* representados em CFL registram: o endereço **IP** do computador do usuário que realizou a requisição; a identificação do **usuário**; a **data** e o **horário** da requisição; a **requisição**; um código que identifica o **estado** da requisição (como sucesso ou erro); e o tamanho em **bytes da resposta** à requisição.

Em uma requisição, informações são passadas ao servidor. Essas informações são específicas de cada sistema aplicativo e basicamente indicam a ação solicitada ao sistema (como a de realizar uma consulta) e os parâmetros necessários para a execução dessa ação (como a expressão de uma consulta).

No protocolo HTTP, as requisições não possuem estado, isto é, para o servidor, cada requisição é independente da outra, não havendo o conhecimento de que duas ou mais requisições participam de um mesmo conjunto de interações e compartilham uma mesma situação ou estado. Entretanto, para permitir a realização de interações com estado, sessões podem ser implementadas pelos sistemas aplicativos. Segundo Jansen (2006), uma **sessão** compreende uma série de interações que ocorrem em um determinado período para atender a uma ou mais necessidades de informação.

O sistema aplicativo determina quando uma sessão é iniciada ou encerrada. Por exemplo, uma sessão pode ser iniciada por um determinado sistema aplicativo quando o usuário acessa o site ou quando este executa uma determinada função da aplicação. Já o término de uma sessão, por exemplo, pode ser determinado quando o usuário executa uma determinada função ou por situações como tempo limite ou desativação do navegador. As sessões são identificadas por um código interno, e todas as requisições que são realizadas em uma mesma sessão carregam o código dessa sessão, que também fica registrado no *log*, junto ao registro de cada requisição. Na análise de *logs*, esse código de sessão é usado para recuperar todas as requisições que ocorreram em uma mesma sessão, e o início e o fim de uma sessão são conhecidos, respectivamente, pelas datas da requisição mais antiga e da última requisição registrada na sessão.

A delimitação das sessões faz parte da estratégia que cada sistema estabelece para realizar suas interações com o usuário. Jansen (2006), considerando os *logs* analisados em sua pesquisa, caracteriza

a sessão como "episódio de pesquisa", e afirma que "A duração da sessão é o tempo total que o usuário passou interagindo com o mecanismo de pesquisa, incluindo o tempo gasto visualizando o primeiro e os documentos subsequentes da Web, exceto o documento final." (JANSEN, 2006, p. 419)[51]. No sistema da biblioteca digital utilizado por Nicholas *et al.* (2007), a sessão é identificada por um número único no *log*. Para cálculo de tempo da sessão, eles utilizam etiqueta (*tag*) de início de sessão e uma *tag* de finalização de sessão. "Se isso não estiver disponível, os pesquisadores escolhem um prazo máximo no qual um usuário pode estar inativo e considere isso como o final dessa sessão" (NICHOLAS *et al.*, 2007). Huynh e Miller (2009) trazem diversos exemplos de estudos com diferentes métricas para definir o tempo de sessão. Para contribuir com os estudos no campo, os autores apresentam um modelo matemático para estimar o tempo de sessão baseado em observações empíricas. Jansen *et al.* (2007) compararam três métodos que contemplam diferentes variáveis para estabelecer qual a melhor maneira de determinar o que é uma sessão de pesquisa. Para os autores, o método que apresenta a melhor identificação contextual, com extensão e duração da sessão, é o que utiliza as variáveis endereço IP, *cookie*, e alterações de conteúdo das consultas.

Na análise de *logs* de sistemas aplicativos de recuperação de informação, normalmente os seguintes conceitos são relevantes:

- **Consulta**: "Uma consulta é definida como uma sequência de zero ou mais termos submetidos a um mecanismo de pesquisa" (JANSEN, 2006, p. 418, tradução nossa). A consulta é registrada no *log* como informação adicionada a uma requisição que determina a execução de uma consulta. Cada sistema aplicativo possui uma forma específica para representar uma requisição que solicita a execução de consulta.

51 A duração da sessão é tipicamente curta, com pesquisadores da Web usando entre cinco e 120 minutos (JANSEN, 2006).

- **Penetração no site**: Nicholas *et al.* (2006a) e Nicholas *et al.* (2007) utilizam esse termo (*site penetration*) para nomear a quantidade de itens ou páginas visualizadas em uma sessão. Isso pode ser obtido a partir da identificação das requisições registradas no *log* que determinam a visualização de itens.

- **Itens visualizados**: Esse conceito varia de sistema para sistema. Para Nicholas *et al.* (2007, 1088, tradução nossa), considera-se um "[…] item 'completo' retornado pelo servidor ao cliente em resposta a uma ação do usuário". Segundo o autor, arquivos de *log* tradicionais diferem grandemente dos registros de biblioteca digital. Os primeiros registram imagens e documentos de texto separadamente. Já nos registros de biblioteca digital, um item completo pode ser todas as páginas, gráficos etc. de um artigo, e isso é registrado como um único item, incluindo um resumo, um artigo ou um sumário. (NICHOLAS *et al.*, 2007).

4.2 SELEÇÃO

A seleção compreende a definição do escopo e a identificação das viabilidades do uso dos *logs* na pesquisa. Envolve verificar se o sistema produz informações relevantes e adequadas para a realização de uma pesquisa científica, definir o recorte temporal, negociar a disponibilidade e o uso dos dados, incluindo questões com relação à privacidade dos dados.

4.2.1 Relevância e adequabilidade das informações do *log*

As questões de pesquisa vão nortear quais dados devem ser coletados e em que período, porém deve-se observar que cada sistema possui características específicas criadas durante o seu

desenvolvimento, o que implica padrões predefinidos pelos aplicativos de software (JANSEN, 2006, p. 412; LAIPELT, 2015). Outro motivo de variação é a "[...] técnica empregada no momento da coleta dos mesmos" (LAIPELT, 2015, p. 166). Nicholas *et al.* (2005, p. 1446) evidenciam diferentes formas de medir o uso dos documentos: o registro pelo servidor de um único item "[...] pode incluir um resumo, um artigo ou um índice", ou pode-se considerar como um item completo todas as páginas, gráficos etc. de um artigo, se isso for assim registrado pelo servidor. Tradicionalmente, no entanto, os arquivos de *log* gravam imagens e documentos de texto separadamente.

Fica claro, portanto, que a variedade de possibilidades de informação a serem obtidas nos *logs* depende do que o servidor apresenta, e cabe ao pesquisador definir as métricas utilizadas para análise. Assim, o pesquisador precisa adequar-se ao que oferece o *log* em termos de dados para articular suas questões de pesquisa tendo em mente o que a fonte é capaz de fornecer. Isso evidencia a importância de uma etapa de planejamento antes da coleta dos dados nos *logs*.

A seleção envolve verificar se o sistema em questão contém informações relevantes para a análise e se o *log* é adequado e suficiente para a realização dessa análise. Isso pode ser realizado primeiramente através da experimentação (operação) do ambiente, pela sua exploração via interface de usuário, com a anotação das **interações relevantes ao estudo** e das informações associadas a cada interação. Com base nesse levantamento de interações, pode-se identificar **variáveis** e **indicadores** potenciais para o estudo.

O pesquisador pode navegar no sistema, realizar consultas e anotar os resultados, procurando entender que tipo de retorno o sistema oferece via pesquisa. Por exemplo, uma consulta a um catálogo de biblioteca geralmente envolve um campo de busca e, às vezes, alguns parâmetros (como os campos onde a busca será realizada). Em outros casos, esses parâmetros aparecem somente na busca

avançada. Após a realização da busca, é esperada uma lista de resultados (ordenados por algum critério, que também pode ser observado) que apresenta algumas informações básicas de cada documento, como título, autor, data etc. Este é um exemplo de interação com o sistema, do qual se pode apreender as possibilidades que o sistema oferece, o que pode ser solicitado e o que ele devolve da requisição.

Por exemplo, Monteiro-Krebs (2013) analisou o uso de sistemas de recomendação, selecionando as variáveis de uso ou não da recomendação feita pelo sistema, quantidade de documentos visualizados, tipos de recomendação usados, tempo de sessão, data, IP do usuário. Observa-se que, nesse caso, as expressões de busca não foram coletadas pois não eram relevantes para o objetivo da pesquisa. Se, por outro lado, o objetivo da pesquisa fosse examinar as tendências temáticas de um domínio, seria recomendado considerar a análise dos termos de busca, estabelecendo o ponto de corte a partir da saturação dos dados (LAIPELT; MONTEIRO-KREBS, 2021).

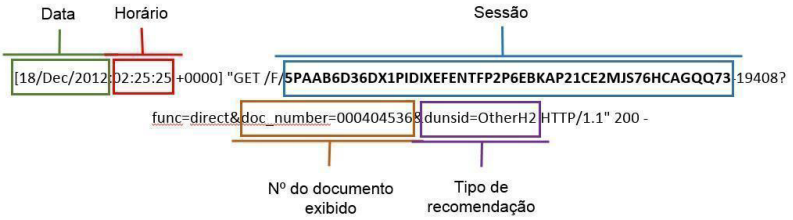
Foi o caso da pesquisa de Laipelt (2015), que coletou *logs* do Portal LexML referentes a um período de 15 dias. A partir da aplicação de um filtro, a pesquisadora selecionou os parâmetros relevantes para a sua pesquisa, que foram IP, data, horário, expressão de busca e tipo de documento (Legislação, Doutrina e Jurisprudência). Além disso, usou para seleção de seus *corpora* de estudo os critérios de frequência e funcionalidade (LAIPELT; MONTEIRO-KREBS, 2021). Como observa-se nesse caso, se o intuito da pesquisa é estudar a linguagem, as expressões de busca dos usuários são mais relevantes do que o tempo da sessão.

Após identificar as interações relevantes ao estudo, o passo seguinte é verificar como essas interações são representadas no *log*, a partir da análise de uma pequena amostra de registros. O *log* de interações de sistemas de recuperação da informação tipicamente oferece informações como quem fez a consulta (IP), o que foi procurado (expressão de busca

e filtros utilizados – se aplicável), o que obteve como resposta (lista de resultados), qual(is) documento(s) escolheu(ram)-se para ver em detalhes (item(ns) clicado(s)/visualizado(s)) e que ações realizou (apenas abriu o registro, fez *download* do documento, realizou uma reserva etc.).

A interpretação do *log* envolve a identificação do que cada registro de *log* informa, porque cada sistema possui uma estrutura específica de informações, que pode variar bastante. Cada registro de *log* é composto por informações que representam uma interação do usuário com o sistema, na forma de requisição. Os elementos que compõem uma requisição são chamados de parâmetros, e cada parâmetro pode ser utilizado como uma variável na pesquisa. Na Figura 1, apresentamos a decodificação de um registro de requisição do trabalho de Monteiro-Krebs (2013):

Figura 1 – Exemplo de registro do *log* decodificado



Fonte: Monteiro-Krebs (2013, p. 65) sistematizado pela autora.

A partir da decodificação representada na Figura 1, pode-se identificar a ação do usuário: ocorre no dia 18 de dezembro de 2012, às 02h25min25seg, o usuário (IP oculto), dentro da sessão de número 5PAAB6[...]QQ73, acessou o registro do documento 0004045368, que foi originado de uma recomendação do tipo OtherH2 (que significa ligação por classificação, ou seja, o sistema está recomendando outras obras sobre o mesmo assunto).

Já na Figura 2, o *log* analisado por Nicholas, Rowlands e Jamali (2009) registra informações diferentes do exemplo anterior. Os dois primeiros campos mostram data e hora de quando o arquivo foi enviado pelo servidor ao computador do usuário. O próximo campo fornece a ação (*download*) e descreve o tipo de página; aqui /browse/open.asp refere-se à página inicial. O campo a seguir é o número do IP. "Mozilla / 5.0 + (Windows; + U; + Windows + NT + 6.0; + fr; + rv: 1.8.1.11) + Gecko/20071127 + Firefox/2.0.0.11" identifica o tipo de navegador da máquina cliente. O campo http://www.sussex.ac.uk/library/resources/e-books.php é o campo de referência e fornece os detalhes do site e da página da página anterior visualizada pelo usuário. Nesse caso, esta era uma página de recursos de e-book ou links na Universidade de Sussex (NICHOLAS; ROWLANDS; JAMALI, 2010).

Figura 2 – Exemplo de registro do *log* decodificado

2007-12-01 04:33:38 GET /browse/open.asp - 139.184.30.131 HTTP/1.0
Mozilla/5.0+(Windows;+U;+Windows+NT+6.0;+fr;+rv:1.8.1.11)+Gecko/20071127
+Firefox/2.0.0.11 http://www.sussex.ac.uk/library/resources/e-books.php
www.myilibrary.com

Fonte: Nicholas, Rowlands e Jamali, 2010, p. 4.

Uma questão importante a ser analisada é que *logs* podem não registrar informações que permitam a identificação do encadeamento entre as interações, impossibilitando análises "relacionais". Por exemplo, pode ser de interesse do pesquisador investigar a quantidade de itens resultantes de uma consulta que são realmente visitados pelo usuário. A análise dessa relação envolve dois tipos de requisições, que estão inter-relacionadas: a requisição da consulta e as requisições das visualizações dos itens da consulta. Essa análise somente é possível se, no registro da requisição de visualização de item, houver uma informação que permita a identificação da requisição que gerou a respectiva consulta. Por exemplo, o registro de *log* da Figura 1 contém a informação que diz que a requisição de visualização de um item foi originada de uma interação prévia de recomendação, mas não contém informação que permita

identificar a recomendação de origem. O pesquisador, portanto, precisa ter em mente o que cada *log* pode oferecer (e o que não pode) em termos informacionais, para que não incorra em conclusões imprecisas.

4.2.2 Recorte dos Dados

Na etapa de seleção, são definidos os **períodos de interações** relevantes para a análise, isto é, os recortes temporais das interações que serão coletadas. Como os *logs* registram todas as interações, não é necessário se pensar em estratificar uma amostra para fins de representatividade, ou seja, o pesquisador poderá ter acesso a todos os registros de todos os usuários. Por outro lado, uma característica do uso de *logs* é o grande volume de dados aos quais se tem acesso. Esse fator pode dificultar a pesquisa, já que pode sobrecarregar o pesquisador com a quantidade de dados a ser manipulada. Assim, é recomendável que se considere a aplicação de algum filtro de seleção dos dados de acordo com os objetivos da pesquisa – escolhendo quais interações se deseja observar.

O volume de dados disponível normalmente é muito grande, e os dados exigem limpeza e pré-processamento antes da análise. Por isso, é desejável que se observe a representatividade dos dados levando em conta a otimização dos recursos disponíveis para análise destes. Uma forma possível de recortar uma amostra representativa de um sistema é, por exemplo, determinar um **recorte estatístico** a partir do universo total de acessos ao sistema. Isso pode ser feito através do acompanhamento de acessos mensais por usuários únicos. Através de consulta prévia diretamente no arquivo de *log*, é possível recuperar a quantidade de IPs diferentes que acessaram o sistema no mês. Acompanhando-se esse indicador, pode-se calcular a média mensal de usuários dos serviços e, assim, determinar a amostra suficientemente representativa de acessos que devem compor o *corpus*.

4.2.3 Disponibilização e uso dos dados

Tendo conhecimento das possibilidades do estudo a partir da experimentação do ambiente e interpretação do *log*, deve-se verificar a viabilidade do uso deste. Nesse momento, o responsável pela guarda do *log* deve ser contatado para identificar se o *log* registra as interações anotadas no levantamento de forma suficiente e adequada para a análise. Muitos ambientes podem optar por não registrar todas as interações ou guardar o registro por um período limitado.

Outra questão relevante na negociação é a presença no *log* de **informações sensíveis**. Em geral, dados pessoais dos usuários – aqueles que permitem sua identificação – são considerados sensíveis, como, por exemplo, o IP de acesso ao sistema. Pela natureza dessas informações, é recomendado que projetos de pesquisa que utilizem dados sensíveis em *logs* passem pela avaliação de um Comitê de Ética, preferencialmente indicando a anonimização dos participantes da pesquisa.

Por questão de privacidade dos usuários, é recomendável que seja feito um tratamento no endereço de IP do computador originário das interações. Os três primeiros algarismos do IP identificam o país de origem do acesso. Se for relevante para a pesquisa regionalizar os usuários, essa informação pode ser mantida, mas os demais algarismos devem ser embaralhados (com o mesmo padrão, para que se identifique quais interações são de cada usuário) ou substituídos por letras. Se a regionalização não for relevante ou mesmo a identificação de usuários únicos, o IP pode ser simplesmente omitido. Outra opção é fazer esse tratamento após a etapa de análise e antes da divulgação dos resultados da pesquisa, de acordo com o protocolo do projeto aprovado previamente pelo Comitê de Ética (LAIPELT; MONTEIRO-KREBS, 2021).

Em sistemas que exigem *login*, é possível obter ainda mais informações, além das presentes nos *logs*, como nome, endereço, idade, gênero, profissão etc. No entanto, cada instituição estabelece

critérios próprios para definição do que entende como informação sensível. Assim, é necessário consultar o responsável pela guarda do *log* para entender quais dados poderão ser disponibilizados obedecendo a política de privacidade de dados da instituição. Outro ponto de atenção é a adequação à Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), que entrou em vigor no Brasil em 2020. Ao lidar com dados pessoais dos usuários, pesquisadores precisam atentar para o fato de que, com o marco legal, a coleta e o uso de dados devem ser feitos sempre com consentimento explícito dos usuários. Assim, a LGPD (Lei 13.709, de 2018) garante maior controle dos cidadãos sobre suas informações pessoais.

Caso existam informações sensíveis a serem retiradas do *log* antes de sua disponibilização, deve-se analisar os impactos dessa retirada ao estudo e possíveis alternativas para contornar essas limitações. Se os objetivos da pesquisa demandarem que seja possível agrupar todas as interações de um mesmo usuário (como estudos de análise da jornada do usuário), não se pode abrir mão do uso do IP (em casos de sistemas sem login mandatório). Nesse caso, o embaralhamento dos IPs de acesso ao sistema é recomendado, que pode ser feito pela própria instituição antes de sua disponibilização. Dessa forma, garante-se a privacidade dos usuários sem comprometer a pesquisa.

Em resumo, nessa fase o pesquisador determina quais os parâmetros ou variáveis que deseja analisar, e esse planejamento é crucial para evitar retrabalho e garantir a viabilidade da pesquisa de forma legal e ética (LAIPELT, MONTEIRO-KREBS, 2021). Os parâmetros precisam dar conta do objetivo do estudo em questão, tendo em vista que cada sistema de informação possui critérios próprios para organização dos *logs*, para determinação do(s) período(s) de guarda dos arquivos e políticas de compartilhamento (ou não) desses dados.

4.3 COLETA E PREPARAÇÃO DOS DADOS

Após a verificação das viabilidades do uso dos dados e a identificação dos dados a serem colhidos, a próxima etapa é a recolha dos dados de interação de um determinado período em um *log* de transações, isto é, a coleta dos dados.

Na coleta, as informações são preparadas pela equipe que gerencia o sistema aplicativo para serem entregues ao pesquisador, de acordo com o que foi previamente especificado, com a remoção de informações sensíveis e de informações não pertinentes à pesquisa, através de comando no servidor. Então o pesquisador passa a preparar esses dados.

> Depois que os dados são coletados, passa-se para o estágio de preparação de dados do processo TLA. Para a preparação de dados, o foco é importar os dados do registro de transações para um banco de dados relacional (ou outro *software* de análise), atribuindo a cada registro uma chave primária, limpando os dados (ou seja, verificando dados incorretos em cada campo) e calculando métricas de interação padrão que servirá de base para uma análise mais aprofundada (JANSEN, 2006, p. 414, tradução nossa).

Na etapa de preparação dos dados, o pesquisador realiza a limpeza e organização (ou tratamento) dos dados obtidos na fase de coleta. Procede-se, então com a normalização, isto é, a adaptação do formato de registro de *log* (normalmente em formato texto) para uma estrutura adequada para a análise (que pode ser texto, no caso de análise exclusiva de expressões de busca, e/ou tabela no caso de análises que envolvam outras variáveis).

Geralmente, os *logs* são recebidos pelo pesquisador no formato texto (.txt) e, para adaptá-los a um formato adequado à pesquisa, é necessário que estes sejam exportados para uma planilha (.csv ou

.xls). Assim, recomenda-se usar um *software* que realize as funções de acordo com o que o pesquisador deseja investigar. São exemplos: eliminar as URLs repetidas do *log*; ordenar cronologicamente as URLs; contar o intervalo de tempo entre o primeiro e o último registro de uma sessão (em horas, minutos ou segundos); exportar os dados para uma tabela (formato .csv ou .xls) organizando-os em colunas por tipo de informação; entre outras ações. É possível encomendar o desenvolvimento de um *software* a um profissional da área de tecnologia da informação, conforme a parametrização da base de dados que registra o *log*, como feito por Monteiro-Krebs (2013) e Laipelt (2015)52. Outra opção é utilizar *softwares* de mercado disponíveis para análise de *logs*, como o Graylog, Logstash, Loggly e Splunk.

A limpeza consiste em eliminar registros corrompidos, que são passíveis de identificação através da ordenação sequencial dos campos, assim, registros fora do formato padrão dos dados de cada campo ficarão agrupados (JANSEN, 2006). Desse modo, fica mais rápida a eliminação dos erros em grandes volumes de dados, em que a identificação visual é inviável. "As funções padrão de banco de dados de dados para somar e agrupar campos-chave, como horário e endereço IP, geralmente identificam quaisquer erros adicionais" (JANSEN, 2006, p. 415, tradução nossa). Além disso, na limpeza, podem ser eliminados registros de agentes não humanos (robôs) se o objetivo da pesquisa for analisar exclusivamente interações entre humanos e o sistema. Infelizmente, não há como determinar com absoluta certeza quais interações são realizadas por humanos ou robôs (SILVERSTEIN *et al.*, 1999; JANSEN, 2006). Assim, é necessário pensar em uma técnica para estabelecer um ponto de corte no *corpus*. Jansen (2006), buscando especificamente as interações de humanos com o sistema, determinou arbitrariamente como ponto de corte a quantidade de 101 ou mais consultas. Pode-se, alternativamente, construir um histograma,

52 Agradecemos a Vicente Grassi-Filho pelo desenvolvimento do *software* extrator de *logs* para o formato texto, já utilizado em diferentes projetos do grupo de pesquisa Orcalab (UFRGS).

com base no tempo de sessão, e eliminar os "*outliers*" (tanto sessões muito grandes quanto aquelas muito pequenas ou sem consulta), que podem distorcer as métricas de análise.

Laipelt e Monteiro-Krebs (2021, p. 133) explicam que "Com a decodificação das interações a partir dos parâmetros é possível separar os registros de acordo com o tipo de interação do usuário." Assim, os arquivos extraídos podem ser importados para tabelas de um banco de dados como o Microsoft Access ou MySQL. A partir de então, podem ser realizadas consultas ao banco de dados, através da linguagem SQL. As consultas resultarão em tabelas de planilhas eletrônicas (Por exemplo: Microsoft Excel) em que as linhas normalmente representam requisições e as colunas representam os parâmetros/variáveis. Essas tabelas podem, então, ser analisadas diretamente e/ou serem transformadas em gráficos, servindo de base para a análise de dados.

4.4 ANÁLISE DOS DADOS

Na etapa de análise, as tabelas e gráficos gerados são comparados e interpretados à luz dos conhecimentos teóricos concernentes à pesquisa. Jansen (2006) traz diversos exemplos de *queries* SQL para análise de *logs*. A partir dessas consultas, pode-se obter muitos resultados estatísticos que servirão de indicadores para o gestor do sistema, como média de consultas realizadas, média de páginas de resultados visualizadas, frequência e coocorrência de termos, grau de associação de pares de termos, e muitos outros.

Essas descobertas servem tanto para melhorar a arquitetura de informação do sistema quanto para planejar capacitações aos usuários, com o intuito de contribuir para que os sistemas possam ser planejados e avaliados a partir das necessidades reais destes. Apresentamos a seguir, com base nos estudos recuperados para a

redação desse capítulo, algumas alternativas de análise das interações dos usuários nos sistemas de recuperação da informação a partir dos dados registrados em *log*.

Segundo Jansen (2006, p. 410, tradução nossa), a análise de *logs* vem sendo utilizada como método "[...] para avaliar sistemas de bibliotecas, sistemas de recuperação de informação (IR) e, mais recentemente, sistemas da Web.". Para tanto, o autor aponta três possibilidades de análise: análise no nível do termo, análise no nível da consulta e análise no nível da sessão (JANSEN, 2006, p. 418). Gouveia (2013) afirma que, assim como o *page tagging*, o método é utilizado nos estudos de métricas de acesso na Web, também designados por webmetria (*Webmetrics* ou *Web Metrics*, em inglês). A análise de *logs* pode ser utilizada para diversos fins no âmbito da Ciência da Informação. O método mostra-se útil para validar sistemas, compreender como os usuários se comportam e o que de fato esperam dos recursos oferecidos pela instituição etc.

Através de um estudo de *logs*, Foust *et al.* (2007) descobriram que 90% das pesquisas realizadas em um sistema de busca por livros eletrônicos (*e-books*) configuraram-se como palavras-chave ou frases. Apenas 10% utilizaram recursos avançados de pesquisa, com construções complexas, operadores booleanos, frases de busca estruturadas e tentativas de pesquisa de títulos de livros ou autores. Os resultados desse estudo sugerem que a maioria dos usuários não esperava recursos avançados de pesquisa e compreendia com facilidade o tipo de pesquisa no estilo "Google" no sistema de pesquisa de livros eletrônicos analisado. Ou seja, os usuários aparentemente não queriam ter que articular recursos complicados de filtragem, concatenação de operadores lógicos etc., mas esperam, cada vez mais, que o sistema faça isso por eles.

A seguir, elencamos algumas possibilidades de descobertas com a análise de *logs* no campo da CI, a partir de parâmetros que comumente estão presentes nos registros de sistemas de informação:

a. **Análises dos acessos ao sistema**: através do IP, pode-se identificar a região de onde o sistema é acessado, possibilitando a geração de relatórios regionais (conforme mencionado na seção 4.1.4); também é possível verificar os dias de maior ou menor acesso (no mês ou na semana) através da análise das variáveis data e horário no *log*; assim como verificar quantos pesquisadores acessaram o mecanismo de pesquisa durante determinado período. Para responder a essa questão, pode-se fazer uma consulta SQL que provê "uma lista de usuários únicos e o número de consultas que eles enviaram durante o período" (JANSEN, 2006, p. 420, tradução nossa)[53].

b. **Duração e comprimento das sessões**: a variável horário também possibilita identificar a duração da sessão e quanto tempo em média os usuários levam para concluir uma ação (seja visualizar um documento, fazer uma reserva ou renovação, realizar o *download* de um documento etc.). Isso pode ser medido ao se subtrair o horário final (última interação) do horário inicial (primeira interação). Assim, é possível identificar se alguma ação específica influencia o tempo das sessões (se a sessão será, em média, mais longa ou curta do que sem a realização dessa ação). Monteiro-Krebs (2013) analisou a duração de sessões com e sem o uso de recomendação de obras no catálogo de uma biblioteca universitária, concluindo que o uso de recomendação implica acesso otimizado (visualização de mais registros em menos tempo); adicionalmente, pode-se

53 Jansen (2006, p. 426) exemplifica essa consulta através da SQL Query 03 (qry_03_list_of_unique_ips):
"SELECT tbl_searching_episodes.uid, Count(tbl_searching_episodes.search_url)
AS CountOfsearch_url
FROM tbl_searching_episodes
GROUP BY tbl_searching_episodes.uid
ORDER BY Count(tbl_searching_episodes.search_url) DESC".

identificar o comprimento das sessões, realizando uma consulta pela quantidade de pesquisas realizadas por cada usuário[54].

c. **Visualização, leitura ou *download* de documentos**: esse parâmetro pode medir sozinho ou cruzado com outras variáveis a quantidade de acessos aos registros de documentos e, em caso de obras disponíveis para acesso ao texto integral, quantas leituras do texto ou mesmo *downloads* foram feitos pelo usuário. Um exemplo desse tipo de estudo é o de Nicholas, Rowlands e Jamali (2010), que empregaram três métricas a fim de obter estimativas mais precisas de "uso" de *e-books*: número de páginas visualizadas, número de sessões conduzidas e quantidade de tempo gasto on-line. Esse tipo de estudo é caracterizado como DLA, pois, além dos *logs*, utiliza questionário, dados de circulação da biblioteca e grupo focal.

d. **Dados demográficos e de pesquisa**: em um sistema que pede *login*, no qual se tem acesso aos dados demográficos do usuário, essas informações podem ser cruzadas com todas as demais variáveis identificadas no *log*. Conforme explicado na introdução desse capítulo, estudos que cruzam os dados de *login* do usuário com dados obtidos no *log* (variáveis acima listadas), chamam-se *deep log analysis* (DLA) (NICHOLAS *et al.*, 2006a; WARWICK *et al.*, 2008; NICHOLAS *et al.*, 2009; NICHOLAS *et al.*, 2006b). Uma possibilidade de uso da DLA é cruzar estratégias de busca (no *log*) com dados demográficos para identificar padrões por perfil de usuário (por área de conhecimento, nível de especialização, tipo de vínculo com a instituição etc.).

54 Jansen (2006, p. 426) exemplifica essa ação com a SQL Query 05 (qry_05_session_length):
"SELECT qry_03_list_of_unique_ips.CountOfsearch_url, Count(qry_03_list_of_unique_ips.CountOfsearch_url) AS CountOfCountOfsearch_url
FROM qry_03_list_of_unique_ips
GROUP BY qry_03_list_of_unique_ips.CountOfsearch_url
ORDER BY Count(qry_03_list_of_unique_ips.CountOfsearch_url) DESC".

e. **Comportamento do usuário**: através do IP, pode-se agrupar todas as interações de um usuário com o sistema. Pelas variáveis data e horário, é possível ordenar cronologicamente todas as ações de um único usuário. Quando não se tem acesso ao IP, a sessão pode exercer esse papel, porém, tem limitações. A primeira limitação do uso de sessão no lugar do IP é que não é possível afirmar que uma sessão seja de um único usuário ou um acesso de vários usuários em um mesmo computador (por exemplo, quando o acesso é realizado a partir de locais públicos como bibliotecas e *lan houses*). A segunda é que, apenas com o dado de sessão, não se pode determinar quais sessões foram realizadas por um mesmo usuário (podem ser várias) e, por isso, não é possível agrupar todas as sessões de um mesmo usuário e visualizar seu comportamento informacional na íntegra. Pode-se também realizar uma análise de exibição de página, também conhecida como análise de cliques, que, de acordo com Jansen (2006, p. 420), serve para medir o comportamento de visualizações de página dos usuários da Web. Para tanto, mensura-se a duração da visualização do documento desde o momento em que o usuário clica em uma URL em uma página de resultados até o momento em que ele retorna ao mecanismo de pesquisa. Se o pesquisador tiver interesse em conhecer a linguagem do usuário, pode fazê-lo através da análise das expressões de busca. Estudos dessa natureza visam melhorar a indexação (remissivas, variantes denominativas, variantes morfológicas – numeral, gênero –, alterações gramaticais etc.).

f. **Estratégias de busca**: visa perceber padrões de navegação (usuários mais ou menos avançados); jornada do usuário (caminhos percorridos no site, ou as diferentes tentativas de recuperação da informação através da análise das expressões de busca); ver se o usuário utiliza todos os recursos disponíveis no sistema de recuperação da informação (SRI), como filtros,

operadores booleanos, comprimento da expressão de busca; identificar termos simples e compostos. Jansen e Spink (2006) fazem uma análise comparativa de nove estudos do tipo TLA para descobrir características e mudanças nos padrões de busca na Web (através dos *logs* de buscadores). A análise realizada por Laipelt (2015, p. 160-161), por sua vez, permitiu inferir características do perfil do usuário através de sua jornada: agrupando todas as interações com o sistema, pode-se verificar o nível de especialização do usuário a partir do seu domínio sobre a terminologia da área do direito.

g. **Uso do acervo**: pode-se analisar as expressões de busca para encontrar os títulos mais buscados. Um exemplo de estudo para identificar obras buscadas com maior ou menor frequência é o de Warwick *et al.* (2008), no qual os pesquisadores cruzaram essa informação com um questionário de percepção sobre uso de recursos digitais (usando método DLA). Além disso, é possível cruzar uma obra mais buscada (no *log*) com os períodos em que esta estava emprestada (no catálogo), verificando assim obras não retiradas ou reservadas por indisponibilidade (se a busca pela obra poderia ter se efetivada em retirada, caso o item estivesse disponível e sem fila de reserva).

h. **Tendências temáticas**: a partir da identificação dos assuntos mais procurados, é possível identificar se o acervo está atendendo ou não as necessidades dos usuários, podendo suprir eventual demanda reprimida. A demanda reprimida também poderia ser identificada através da lista de reservas no catálogo, porém, esse indicador somente informa a quantidade insuficiente de itens já existentes no acervo, enquanto no *log* se amplia a visão, possibilitando identificar novos itens para aquisição que contemplem essas necessidades.

i. **Penetração no site**: a partir da quantidade de itens ou páginas visualizadas, é possível identificar o comportamento de busca de informações dos usuários. Identificando quantos cliques o usuário precisa dar para acessar a informação desejada e os caminhos percorridos por ele (páginas visitadas), é possível avaliar criticamente a usabilidade do site, e implementar melhorias para poupar o tempo do leitor. A penetração do site pode ser combinada com o número de visitas, tempo de permanência no site, tipo de itens e conteúdo visualizados para indicar características do perfil do usuário e suas demandas.

5 CONSIDERAÇÕES FINAIS

Com a disponibilidade cada vez maior de informação através de dispositivos diversos, os sistemas de informação têm evoluído de forma gradual à medida que os usuários exigem maior precisão e personalização em seus resultados de busca. Com o objetivo de munir pesquisadores e profissionais da informação com recursos de investigação atuais e muitas vezes disponíveis dentro da própria unidade de informação, nesse capítulo discorreu-se sobre a análise de *logs* como uma alternativa para identificar comportamentos de usuários e antecipar suas demandas.

As etapas metodológicas foram apresentadas e exemplificadas, também foram expostas diversas possibilidades de uso dos parâmetros do *log* tanto separadamente quanto de forma conjunta (cruzando as variáveis). Buscou-se, ainda, apresentar a análise de *logs* para explorar as potencialidades das expressões de busca para identificação de tendências temáticas e diferenças de perfis entre os usuários.

Apesar de a obtenção de um *log* para análise não ser uma tarefa das mais simples, por envolver uma série de questões técnicas e até

políticas comentadas nesse trabalho, acredita-se que o esforço é recompensado com a riqueza de *insights* que se pode obter com esse recurso. Isso foi possível demonstrar nesse capítulo, considerando as práticas consolidadas na área da Ciência da Informação e a expertise dos autores na realização de pesquisas envolvendo o uso e a análise de *logs*.

6 REFERÊNCIAS

AIRES, R. V. N. X.; ALUÍSIO, S. M. Como incrementar a qualidade dos resultados das máquinas de busca: da análise de *logs* à interação em português. **Ciência da Informação**, v. 32, n. 1, p. 5-16, 2003. Disponível em: http://www.brapci.ufpr.br/brapci/v/a/689. Acesso em: 27 Jul 2017.

BRASIL. **Lei nº 13.709**, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Brasília, DF: Presidência da República; 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 22 nov. 2021.

CETIC. **Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros - TIC Domicílios 2020**. *S. l.*: Cetic.br, 25 nov. 2021. Disponível em:https://cetic.br/media/docs/publicacoes/2/2011124201505/resumo_executivo_tic_domicicilios_2020_.pdf. Acesso em: 10 dez. 2021.

CHEN, LS. Applying swarm intelligence to a library system. **Library Collections, Acquisitions, and Technical Services**, v. 34, n. 1, p. 1-10, 2010. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/14649055.2010.10766254. Acesso em: 20 ago. 2018.

DAVIS, P. M. Information-seeking behavior of chemists: A transaction log analysis of referral URLs. **JASIST**, v. 55(4), 326 –332, 2004. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.10384. Acesso em: 21 dez. 2018.

FOUST, J. E. *et al.* Improving e-book access via a library-developed full-text search tool. **Journal of the Medical Library Association**, Bethesda, v. 95, n. 1, p. 40-45, 2007. Disponível em: https://www-ncbi-nlm-nih-gov.ez45.periodicos.capes.gov.br/pmc/articles/PMC1773047/. Acesso em: 02 jul. 2018.

GOUVEIA, F. C.. Altmetria: métricas de produção científica para além das citações. **Liinc em Revista**, Rio de Janeiro, v. 9, n. 1, p. 214-227, maio

2013 – Disponível em: http://revista.ibict.br/liinc/article/view/3434/3004. Acesso em: 27 jul. 2017.

HUYNH, T.; MILLER, J. Empirical observations on the session timeout threshold. **Information Processing & Management**, v. 45, n. 5, p. 513-528. 2009.

JANSEN, B. J. Search log analysis: what it is, what's been done, how to do it. **Library & Information Science Research**, Pennsylvania, v. 28, p. 407-432, 2006. Disponível em: http://lincs.hum.iit.edu/sites/default/files/JansenSearchLog.pdf. Acesso em: 03 maio 2013.

JANSEN, B. J., & POOCH, U. A review of Web searching studies and a framework for future research. **Journal of the American Society for Information Science and Technology** (JASIST), v. 52, p. 235–246, 2001.

JANSEN, B. J.; TAKSA, I.; SPINK, A. Research and Methodological Foundations of Transaction Log Analysis. *In*: JANSEN, Bernard J.; SPINK, Amanda; TAKSA, Isak. **Handbook of web log analysis**. Hershey, PA: Information Science Reference, 2009. p. 1-16.

JANSEN, B.J.; SPINK, A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. **Information Processing & Management**, v. 42, n. 1, p. 248-263, jan. 2006.

JANSEN, BJ.; SPINK, A.; Blakely, C.; Koshman, S. Defining a session on web search engines. **Journal of the American Society for Information Science and Technology** (JASIST), v. 58, n. 6, p. 862-871, Apr. 2007.

LAIPELT, R. C. F. A análise de logs como estratégia para a realização da garantia do usuário. **Em Questão**, Porto Alegre, v. 21, n. 3, p. 150-170, set/dez. 2015. Disponível em: http://www.redalyc.org/html/4656/465645968009/. Acesso 16 jul. 2017.

LAIPELT, R. C. F.; MONTEIRO-KREBS, L. **Termos sob a superfície**: elementos teóricos, metodológicos e terminológicos para a Representação do Conhecimento. Rio de Janeiro: Editora Interciência, 2021.

MENG-XING, H.; CHUN-XIAO, X.; YONG, Z.. Supply chain management model for digital libraries. **The Electronic Library**, v. 28, n. 1, p. 29-37, 2010. Disponível em: https://www.emeraldinsight.com/doi/abs/10.1108/02640471011023351. Acesso em: 20 ago. 2018.

MONTEIRO-KREBS, L. **Sistema de recomendação para bibliotecas universitárias.** 2013. 95f. Trabalho de Conclusão de Curso (Graduação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: http://www.lume.ufrgs.br/handle/10183/78367. Acesso em: 26 jul. 2017.

MONTEIRO-KREBS, L.; ROCHA, R. P. da; RIBEIRO, C. Quem leu este também leu...: sistema de recomendação na biblioteca universitária. **Perspectivas em Ciência da Informação**, v. 22, n. 1, p. 151-169, mar. 2017. DOI: 10.1590/1981-5344/2496. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2496. Acesso em: 17 abr. 2019.

NICHOLAS, D. *et al*. Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the 'bouncer'. Information Processing & Management, v. 43, n. 4, p. 1085-1102, 2007. Disponível em: https://doi.org/10.1002/asi.20564. Acesso em: 22 dez. 2021.

NICHOLAS, D. *et al.* The information seeking behaviour of the users of digital scholarly journals. **Information Processing & Management**, v. 42, n. 5, p. 1345-1365, September 2006a. Disponível em: https://doi.org/10.1016/j.ipm.2006.02.001. Acesso em: 22 dez. 2021.

Nicholas, D. *et al.*. Online use and information seeking behaviour: institutional and subject comparisons of UK researchers. **Journal of Information Science**, v. 35, n. 6, p. 660-676, 2009.

NICHOLAS, D.; HUNTINGTON, P.; WATKINSON; A. Scholarly journal usage: the results of deep log analysis. **Journal of Documentation**, London, v. 61, n. 2, p. 248-280, 2005. Disponível em: http://www.emeraldinsight.com/doi/pdfplus/10.1108/00220410510585214. Acesso em: 26 jul. 2017.

NICHOLAS, D.; HUTTINGTON, P; WATKINSON, A. Revisiting 'obsolescence' and journal article 'decay' through usage data: an analysis of digital journal use by year of publication. **Information Processing & Management**, v. 41, n. 6, p. 1441-1461, dez. 2005. Disponível em: https://doi.org/10.1016/j.ipm.2005.03.014. Acesso em: 20 ago. 2018.

NICHOLAS, D.; ROWLANDS, I.; JAMALI, H.R.. E-textbook use, information seeking behaviour and its impact: Case study business and management. **Journal Of Information Science**, v. 36, n. 2, p. 263-280, abr. 2010.

NICHOLAS, D.; ROWLANDS, I.; JAMALI, H.R.. E-textbook use, information seeking behaviour and its impact: Case study business and management. **Journal of Information Science**, v. 36, n. 2, p. 263-280, abr. 2010.

NICHOLAS, David *et al*. What deep log analysis tells us about the impact of big deals: case study OhioLINK. **Journal of Documentation**, v. 62, n. 4, p. 482-508, 2006b. Disponível em: https://doi.org/10.1108/00220410610673864. Acesso em: 22 dez. 2021.

PETERS, T.A.. The history and development of transaction log analysis. **Library Hi Tech**, v. 11, n. 2, p. 41-58, 1993.

SILVERSTEIN, C. *et al.*. Analysis of a very largeWeb search engine query log. **SIGIR Forum**, v. 33, n. 1, p. 6-12, 1999.

SPINK, A., JANSEN, B. J.. **Web search**: public searching of the web. New York: Kluwer, 2005. 198 p.

VAKKARI, P.; TALJA, S.. Searching for electronic journal articles to support academic tasks. A case study of the use of the Finnish National Electronic Library (FinELib). **Information Research**, v. 12, n. 1, 2006. Disponível em: https://www.emeraldinsight.com/doi/abs/10.1108/02640471011023351. Acesso em 20 ago. 2018.

WARWICK, C. *et al.*. Library and information resources and users of digital resources in the humanities. **Program-electronic Library And Information Systems**, v. 42, n. 1, p. 5-27, 2008.