em questão
E-ISSN 1808-5245

A análise de logs como estratégia para a realização da garantia do usuário

Rita do Carmo Ferreira Laipelt
Doutora; Universidade Federal do Rio Grande do Sul (UFGRS), Porto Alegre, RS, Brasil;
ritacarmo@yahoo.com.br

Resumo: Tem como objetivo discutir o potencial da coleta de logs de pesquisa para realizar a garantia do usuário na elaboração de tesauros e/ou para gestão de catálogos de autoridades. Demonstra como a metodologia de análise de logs possibilita identificar a linguagem dos usuários a partir do próprio sistema de recuperação da informação. Utiliza como objeto de estudo as lexias contidas nos logs de pesquisa dos usuários do Portal LexMl, do Senado Federal Brasileiro, nas áreas do Direito do Trabalho e do Direito Previdenciário. Apresenta os resultados da análise de lexias com aspectos conceituais, que indicam o nível de conhecimento do usuário em relação ao assunto pesquisado. Conclui que os logs são relevantes para a identificação de características, demandas e necessidades dos usuários. Sugere o estabelecimento de políticas para sua coleta, visto que, no Brasil, poucas pesquisas utilizam os logs como objeto de estudo na área da Ciência da Informação.

Palavras-chave*:* Representação do conhecimento. Recuperação da informação. Análise de logs. Garantia do usuário.

1 Introdução

Com os avanços tecnológicos decorrentes da Segunda Guerra Mundial, o tratamento da informação passou a ser fundamental para o controle da literatura e acompanhamento de pesquisas em desenvolvimento. Em função dos problemas gerados pela explosão da informação, a ciência avançou rapidamente e o valor da informação passou a ser percebido em função de interesses políticos, científicos e tecnológicos que se tornaram evidentes. Em 1945, Vannevar Bush, preocupado com a questão da recuperação da informação, falava sobre a associação de conceitos ou palavras para a organização da informação, argumentando que esse é o padrão utilizado pelo cérebro humano para transformar informação em conhecimento. Desde então, a Ciência da Informação tem dedicado maior atenção a esse tema e, com isso, tem

desenvolvido pesquisas visando o aprimoramento de técnicas e instrumentos que auxiliem na realização das atividades de representação e organização da informação, tais como os Sistemas de Organização do Conhecimento (SOC).

No contexto atual, com a utilização dos catálogos *on-line* nas bibliotecas e dos repositórios digitais, que permitem aos usuários fazerem suas pesquisas remotamente, os problemas relacionados à recuperação da informação se tornaram ainda mais preocupantes. Hoje, o usuário não precisa ir até biblioteca para fazer suas pesquisas e, consequentemente, não interage, ou interage muito pouco, com os bibliotecários. O resultado da pouca interação entre usuários e bibliotecários é uma das principais causas das dificuldades dos usuários para a recuperação da informação. Visto que, muitas vezes, os descritores atribuídos por bibliotecários aos documentos, durante o processo de representação do conteúdo dos mesmos (indexação), são diferentes dos termos utilizados pelos usuários durante o processo de busca de informação. Isso dificulta, e pode até impossibilitar a recuperação de documentos pertinentes aos interesses/necessidades dos usuários.

Nesse cenário, a interface de pesquisa do Sistema de Recuperação da Informação (SRI), com suas diferentes possibilidades de busca, exerce um papel intermediário entre o usuário e o acervo documental de uma instituição. Desse modo, o papel tradicional de mediador da informação, fortemente exercido pelos bibliotecários quando os catálogos eram manuais, está desaparecendo em função da utilização e disponibilização dos catálogos eletrônicos, que podem ser acessados a distância através da internet. Isso não quer dizer que os bibliotecários vão deixar de exercer seu papel de mediador. Porém, certamente terão de exercer essa função de outra maneira; terão de encontrar uma forma alternativa para auxiliar os diferentes usuários dos SRI a realizarem suas pesquisas com autonomia e sucesso.

Acreditamos que a melhor maneira de auxiliar os usuários em suas atividades de busca de informação, no cenário atual, é investir no aperfeiçoamento de metodologias e ferramentas de representação e recuperação da informação. Porém, para que esse aperfeiçoamento seja possível, é imprescindível realizar a garantia do usuário, ou seja, identificar e acompanhar a

linguagem dos usuários que utilizam os sistemas de recuperação da informação. Em vista disto, apresentamos neste artigo uma metodologia que nos possibilita conhecer/identificar a linguagem dos usuários a partir do próprio sistema de recuperação da informação. Nosso objetivo é demonstrar a partir da discussão dos resultados, o potencial dos logs como fonte de coleta de dados, sobretudo, para a identificação do léxico dos usuários para a elaboração de tesauros e/ou para gestão de catálogos de autoridades, ou seja, para a escolha de descritores para a representação da informação. Para isso, escolhemos como objeto de estudo as lexias de buscas contidas nos "logs" de pesquisa dos usuários do Portal LexMl do Senado Federal Brasileiro nas áreas do Direito do Trabalho e do Direito Previdenciário, visto que, toda pesquisa realizada no Portal fica registrada em um arquivo log armazenado no servidor Web da instituição e, através de sua análise, é possível verificar a linguagem utilizada pelos usuários para a recuperação de informação. Ressaltamos que o LexML é um portal especializado em informação jurídica e legislativa, que reúne leis, decretos, acórdãos, súmulas, projetos de leis, entre outros documentos das esferas federal, estadual e municipal dos Poderes Executivo, Legislativo e Judiciário de todo o Brasil. De responsabilidade do Senado Federal Brasileiro, seu acervo é o reflexo da cooperação existente entre quatorze bibliotecas jurídicas. O Portal encontra-se disponível para consulta gratuita na internet no seguinte endereço: <www.lexml.gov.br>.

2 A organização e representação da informação e do conhecimento e os sistemas de recuperação da informação

A área de Organização e Representação do Conhecimento, de acordo com Pinho (2009), investiga os fundamentos científicos, as habilidades e os instrumentos que auxiliam o profissional nas atividades de extração, descrição, nomeação e rotulagem do conhecimento que será objeto de sistemas de recuperação da informação. Seu principal objetivo, de acordo com Lima e Alvares (2012, p. 39), "[...] é a recuperação de objetos

informacionais, que são as informações registradas nos diferentes suportes existentes (texto, imagem, registro sonoro, mapas, páginas da web, etc.).” Por isso, “[...] as atividades de organização e representação do conhecimento são o cerne da atuação do profissional da informação.” (PINHO, 2009, p. 18).

Dahlberg (2006) considera diferentes o conceito e a aplicação dos termos, organização do conhecimento e organização da informação. Para a autora, a organização do conhecimento se dá quando elaboramos sistemas conceituais, como os tesauros; já a organização da informação ocorre quando utilizamos esses sistemas conceituais para descrever o conteúdo dos documentos, ou seja, ocorre quando realizamos a indexação das obras de um acervo. Deste modo, a organização da informação está relacionada aos meios de recuperar a própria informação e, portanto, também ao arranjo de acervos, tradicionais ou eletrônicos, através da descrição do assunto dos documentos (LIMA; ALVARES, 2012). Bräscher e Café (2010, p. 92) definem organização da informação como “[...] um processo que envolve a descrição física e de conteúdo dos objetos informacionais.”. Para as autoras, a representação da informação é o produto dessa descrição que deve ser “[...] entendida como um conjunto de elementos descritivos que representam os atributos de um objeto informacional específico.” (BRÄSCHER; CAFÉ, 2010, p. 92). Ou seja, a organização da informação refere-se à organização de suportes físicos nos quais estão contidas as informações. A organização do conhecimento, por outro lado, refere-se aos conceitos contidos nos documentos, “[...] visa à construção de modelos de mundo que se constituem em abstrações da realidade.” (BRÄSCHER; CAFÉ, 2010, p. 93). As autoras descrevem a organização do conhecimento como

> [...] o processo de modelagem do conhecimento que visa a construção de representações do conhecimento. Esse processo tem por base a análise do conceito e de suas características, para o estabelecimento da posição que cada conceito ocupa num determinado domínio, bem como das suas relações com os demais conceitos que compõem esse sistema nocional. (BRÄSCHER; CAFÉ, 2010, p. 95).

Por isso, para as autoras, a representação do conhecimento é uma estrutura conceitual

> [...] feita por meio de diferentes tipos de sistemas de organização do conhecimento (SOC), que são sistemas conceituais que representam determinado domínio por meio da sistematização dos conceitos e das relações semânticas que se estabelecem entre eles. (BRÄSCHER; CAFÉ, 2010, p. 96).

Logo, entendemos que, no âmbito da Ciência da Informação, a representação, a organização e a recuperação da informação fazem parte de um fluxo contínuo em que cada etapa depende das demais para o seu sucesso. Representamos a informação contida nos documentos com o objetivo de organizá-la e, posteriormente, recuperá-la. Por esse motivo, não se pode pensar nessas três atividades isoladamente, já que são concomitantes, no sentido de que a decisão tomada durante a representação vai determinar a organização e, consequentemente, interferir na recuperação da informação. Em um SRI, a qualidade da recuperação da informação depende dos procedimentos utilizados durante o processo de organização da informação. Por isso, os padrões de organização devem ser definidos desde a concepção do sistema (LIMA, ALVARES, 2012).

Consequentemente, quando planejamos um sistema ou um instrumento como os tesauros, cuja função é auxiliar na representação da informação (indexação), é preciso ter em mente que as decisões tomadas (como categorias, subcategorias, relações de equivalência e associação) irão, inevitavelmente, afetar a forma de organização da informação, sua estrutura de apresentação, e a maneira de recuperá-la. Ou seja, ao determinar a forma de entrada dos dados em um SRI, também estamos determinando sua organização e a forma de recuperação. Esse processo depende e é influenciado por diferentes variáveis, tais como: público-alvo do SRI, contexto socioeconômico, contexto cultural, recursos tecnológicos, recursos humanos, físicos e financeiros disponíveis, entre outros.

Para muitos profissionais e pesquisadores do campo da Biblioteconomia e da Ciência da Informação, a indexação, além de ser muito relevante, é uma das principais atividades que exercem. Guedes e Dias (2010, p. 42) definem indexação como "[...] um conjunto de procedimentos com objetivo de expressar/representar o conteúdo temático de documentos através de linguagens de indexação ou documentárias visando à recuperação posterior." Lancaster

(2004) explica que a indexação é realizada principalmente em duas etapas; são elas: a análise conceitual e a tradução. Durante a análise conceitual, ocorre a identificação do assunto do documento. Já na etapa de tradução, o assunto identificado a partir da análise conceitual é transformado em termos de indexação. A escolha dos termos de indexação, por sua vez, pode ocorrer a partir da extração de palavras que ocorrem no documento ou através da atribuição de termos extraídos de outra fonte. Para Lancaster (2004), o mais frequente é que os termos sejam atribuídos a partir do uso de algum vocabulário controlado, mas também pode acontecer de os mesmos serem selecionados arbitrariamente, conforme a percepção do indexador, o que Rowley (2002) denomina linguagem de indexação livre.

Convencionalmente, a indexação pode ser realizada com o suporte de três tipos de linguagens, descritas por Rowley (2002) da seguinte forma:

a) linguagem de indexação controlada: definida como um conjunto de termos autorizados (descritores) para uso na indexação do assunto de documentos. É subdividida em dois tipos: as linguagens alfabéticas de indexação, como os tesauros e listas de cabeçalhos de assunto; e os sistemas de classificação, representados por código ou notação;
b) linguagem de indexação natural: refere-se a quaisquer expressões que ocorram em alguma parte do documento. Todos os termos no corpo do documento são candidatos a serem termos de indexação (descritores);
c) linguagem de indexação livre: para esta linguagem não existem limitações quanto aos termos a serem utilizados no processo de indexação. Sua diferença em relação a linguagem natural é que os termos utilizados para a indexação são selecionados de maneira arbitrária e não ocorrem no texto do documento.

O tipo de linguagem de indexação a ser adotada por uma unidade de informação pode variar de acordo com a área de conhecimento ou o tipo de instituição, especialmente pelo fato de que em algumas áreas do conhecimento, por exemplo, não existem tesauros ou cabeçalhos de assunto publicados. A vantagem de usar linguagens controladas é que elas conferem maior qualidade à indexação e possibilitam a manutenção da consistência da representação da

informação. A desvantagem de utilizá-las, no entanto, é que nem sempre, ou nunca, o usuário emprega esses descritores em sua busca e, com isso, caso o SRI não tenha um sistema de remissivas integrado, ou seja, um dispositivo que liga os termos utilizados na indexação com as variantes dos mesmos (sinônimos e quase-sinônimos), o usuário provavelmente ficará sem resposta para a sua busca. Em relação às linguagens naturais, bastante utilizadas em unidades de informação de um modo geral, como os termos são extraídos direto dos textos dos especialistas, há um grande risco de "poluir" o SRI e torná-lo inconsistente, em função dos diferentes tipos de variações existentes (variações denominativas e conceituais). Pela mesma razão, as linguagens de indexação livres também apresentam um grande risco à consistência do SRI, uma vez que sequer a linguagem dos especialistas da área de conhecimento do documento a ser indexado é consultada.

O processo de indexação envolve tomadas de decisões em diferentes níveis, tais como:

a) nível de exaustividade - quantidade de termos utilizados para a representação dos assuntos de um documento - pode ser determinada conforme o tipo de documento;
b) nível de especificidade - refere-se a intensão das características de um termo, de modo que, sua especificidade possibilita, ao mesmo tempo, maior precisão e menor revocação à recuperação da informação;
c) tipo de linguagem a ser adotada - ou seja, se livre, natural ou controlada (lembrando que o uso de uma linguagem livre, sem padronização, exige mais tempo e trabalho para a busca e recuperação da informação pelo usuário);
d) recursos disponíveis para a gestão do acervo - capacidade de revocação e precisão, características igualmente importantes, pois expressam a capacidade de filtragem do sistema em deixar passar o que é solicitado e impedir o que não é solicitado (CARNEIRO, 1985).

No mesmo sentido, Chaumier (1988) afirma que a indexação é essencial para que os documentos possam ser recuperados de um acervo e, ao mesmo tempo, possam oferecer uma resposta adequada, a toda solicitação dos usuários,

sem que haja "ruídos" (ou seja, sem oferecer como resposta documentos que não correspondem à questão de pesquisa do usuário), nem "silêncios" (quando o sistema não recupera um documento que existe no acervo). A ausência de precisão, conforme Moreiro González (2004), tem sido comum em qualquer sistema de recuperação baseado em palavras, justamente em função da ambiguidade presente nelas. Para o autor, o problema é que a grande maioria dos sistemas realiza uma busca léxica, pela simples ocorrência das palavras no registro bibliográfico, e não semântica, ou seja, através do conceito representado pelo termo de um domínio, fator que torna ainda mais complexa a indexação e a recuperação da informação.

Para Guedes e Dias (2010), a indexação também pode ser analisada a partir da perspectiva do agente executor do processo. Rafferty e Hidderley[1] (apud GUEDES; DIAS, 2010, p. 45) distinguem três grupos de candidatos a atores no processo de indexação, são eles:

a) indexação orientada por especialistas: baseia-se no tratamento da informação através da intervenção de intermediários (bibliotecários, indexadores, editores voluntários), é a indexação feita por especialistas, sendo mais dispendiosa e cara;
b) indexação orientada pelo autor: esta abordagem pressupõe que o autor irá utilizar termos que são comumente compreendidos e geralmente aceitos. Um problema que essa abordagem enfrenta é o fato de o autor não ser necessariamente um gestor de informação, com os conhecimentos profissionais de um especialista;
c) indexação orientada pelo usuário: esse tipo de indexação possibilita um elevado nível de interação com a comunidade que, provavelmente, não seria possível se tivesse decisões a serem tomadas sobre códigos, convenções e regras que regem qualquer taxonomia controlada.

A indexação orientada por especialistas é utilizada principalmente em SRI de instituições. Contudo, destacamos que a indexação orientada pelo autor e pelos usuários são métodos bastante utilizados na internet, por exemplo, em repositórios digitais e em ambientes caracterizados como pertencentes a *Web 2.0*, embora esse tipo de indexação seja problemática, do ponto de vista da

consistência e qualidade da representação das informações contidas nos documentos. Brown *et al*. (1996)[2] apud Lancaster (2004), chamam a atenção para a necessidade de um tratamento democrático da indexação, em que os usuários acrescentariam aos registros termos de sua própria escolha, quando isso fosse necessário e apropriado. Acreditamos que, diante da riqueza vocabular existente no contexto discursivo de autores e usuários de sistemas de informação, seria uma grande evolução se os SRIs pudessem realizar uma indexação colaborativa entre bibliotecários, autores e usuários. Visto que, termos e conceitos de uma área de conhecimento constituem-se como tais a partir da interação e negociação de sentidos entre os sujeitos. Nesse sentido Gaudin explica que "[...] não há a palavra certa em si. Há somente palavras apropriadas a interações definidas." (2005, p. 86, tradução nossa).

A partir dessa breve apresentação, constatamos que, para atender às necessidades de busca de informação dos usuários – considerando os diferentes tipos de variações existentes, em consonância com padrões de controle terminológico, a fim de manter sua consistência e eficácia– é necessário que os SRIs sejam aperfeiçoados. É importante salientar que, em termos de desenvolvimento tecnológico, já existem *softwares*, como o ALEPH, por exemplo, com ferramentas capazes de melhorar consideravelmente a recuperação da informação nos SRIs, tais como as redes de remissivas que ligam os descritores aos termos equivalentes (variantes denominativas) facilitando a recuperação da informação pelo usuário. Contudo, infelizmente, do ponto de vista prático, o uso dessas ferramentas ainda é limitado e muitas instituições, ainda que disponham deste recurso, não o utilizam. A subutilização de *softwares* que permitem a elaboração de redes de remissivas reforça a necessidade de maior domínio e compreensão de um conjunto de princípios teóricos que podem contribuir para o aperfeiçoamento dos sistemas de recuperação da informação, tais como as técnicas de elaboração de tesauros e o estudo do fenômeno da variação no âmbito da Terminologia, a título de exemplificação.

3 Análise de logs

Para realizar a garantia do usuário, ou seja, identificar a terminologia utilizada pelos usuários na recuperação da informação, optamos pela análise das lexias de buscas contidas no arquivo log das pesquisas realizadas pelos usuários do Portal LexMl. Essa escolha se justifica pelo fato de encontrarmos nos logs o registro da interação dos usuários com o sistema de recuperação da informação. Nicholas, Huntington e Watkinson explicam que os logs "[...] representam os usuários, são pegadas de informação digital." (2005, p. 250). De acordo com os autores, uma das grandes vantagens de trabalharmos com os logs, é que eles fornecem o registro imediato das ações das pessoas, ou seja, são o retrato da realidade dos usuários, pois "Os dados não são filtrados e falam por si [...]" (NICHOLAS, HUNTINGTON, WATKINSON, 2005, p. 250). Por essa razão, acreditamos que a análise de logs de pesquisa representa uma nova alternativa de fonte de coleta de termos e descritores tanto para tesauros como para a gestão de catálogos de autoridades, pois, desse modo, é possível verificar de forma direta as expressões utilizadas pelos usuários para a recuperação da informação sem a interferência do bibliotecário e/ou do pesquisador. Além disto, também são uma ótima fonte para a identificação de demandas e necessidades dos usuários de um sistema de informação.

No entanto, é importante mencionar que os dados registrados em um arquivo log podem variar de acordo com a configuração utilizada pelo servidor do sistema de recuperação da informação e, também, pela técnica empregada no momento da coleta dos mesmos. No caso do Portal LexML, cada pesquisa realizada pelos usuários gera um registro (log) que informa o número de IP da máquina utilizada para a realização da pesquisa, bem como a data, o horário, o tipo de fonte (Legislação, Jurisprudência, Doutrina, Proposições Legislativas, Publicação Oficial e Outras Manifestações) e as lexias utilizadas pelos usuários em suas buscas.

É importante ressaltar que, além das lexias dos usuários (expressões de busca), informações como número de IP da máquina utilizada para a realização da pesquisa, a data, e o horário, são elementos fundamentais para a realização de pesquisas que visam a análise e identificação de características dos usuários.

Visto que, sem essas informações não é possível observar onde começa e onde termina a interação de cada indivíduo. Com tudo, esclarecemos que embora seja necessário identificar o número de endereço IP das máquinas utilizadas, a privacidade dos usuários é preservada, pois a coleta realizada não nos permite identificar dados pessoais dos usuários. Nos logs, visualizamos apenas os dados referentes à interação dos usuários com o sistema de buscas, no caso específico desta pesquisa, a interação dos usuários do Portal LexMl. Ainda assim, no intuito de garantir a privacidade dos usuários, não divulgaremos os números de endereço IP das máquinas identificadas nos logs. Razão pela qual, também optamos por substituir os primeiros números do endereço IP por letras, na figura 1, que apresentamos a seguir, para demonstrar a interação de um usuário com o sistema bem como o tipo de informações que podemos obter com a análise de logs.

Figura 1 – Log de pesquisa de um usuário

IP	Ano	Mês	Dia	Hora	Minuto	Segundo	Pais	Lexia de Busca	Tipo de Fonte
xxx.y.241.251	2012	8	11	17	55	4	Brasil	"execução contra herdeiro"	Jurisprudência
xxx.y.241.251	2012	8	11	17	55	15	Brasil	"execução contra herdeiro"	""
xxx.y.241.251	2012	8	11	17	55	19	Brasil	"execução"	""
xxx.y.241.251	2012	8	11	18	25	20	Brasil	"execução trabalhista contra herdeiro"	""
xxx.y.241.251	2012	8	11	18	25	50	Brasil	"execução trabalhista contra executado falecido"	""
xxx.y.241.251	2012	8	11	18	27	55	Brasil	"execução trabalhista de cujus"	""
xxx.y.241.251	2012	8	11	18	29	35	Brasil	"execução trabalhista contra empregador falecido"	""
xxx.y.241.251	2012	8	11	21	18	37	Brasil	"execução trabalhista contra herdeiros"	""
xxx.y.241.251	2012	8	11	21	19	33	Brasil	"execução trabalhista empregador falecido"	""

Fonte: Elaborado pela autora.

A partir da análise da interação exemplificada na figura 1, percebemos que o usuário reformulou sua questão de pesquisa diversas vezes utilizando termos diferentes, mas com o mesmo sentido (variantes denominativas), como executado falecido e de cujos. Outro recurso de busca também utilizado pelo usuário, que podemos observar na figura 1, é a busca por campo semântico, nesse caso especificamente por oposição, como podemos identificar nas expressões: “execução trabalhista contra herdeiros” e “execução trabalhista contra empregador falecido”. Nesse exemplo, fica evidente o esforço do usuário para recuperar a informação que deseja, tanto pelo uso de variantes como pelo

uso de termos que embora sejam opostos preservam o sentido da sua questão de busca. Outro aspecto importante a destacar, também na figura 1, são as características do usuário. A estratégia de pesquisa empregada oferece fortes indícios sobre o seu perfil, e nos permite inferir que se trata de um usuário especializado, que conhece muito bem a área e o assunto que está buscando, visto que, demonstra bastante domínio sobre a terminologia da área do direito do trabalho e a emprega de forma estratégica para a obtenção de melhores resultados de busca.

Diante do exposto, apresentamos a seguir as etapas metodológicas empregadas para a realização desta pesquisa.

Quanto à constituição do *corpus* de estudo, o mesmo é formado, portanto pelas lexias de buscas utilizadas pelos usuários do Portal LexML, identificadas nos logs de pesquisa armazenados no servidor do Senado Federal Brasileiro. Tendo em vista o grande volume de dados obtidos (276.129 expressões de busca em 32 dias de coleta), optamos por realizar a análise referente a apenas 15 dias de coleta: de 26 de julho a 11 de agosto de 2012. Nesse *corpus*, encontramos informações referentes a 65.536 lexias de buscas de usuários de todas as áreas do Direito. Desse montante, identificamos 2.617 lexias de busca (com ocorrência de lexias repetidas) relacionadas ao Direito do Trabalho e ao Direito Previdenciário.

O exame preliminar do *corpus* de estudo nos levou à identificação de três categorias de análise de lexias, são elas:

a) aspectos verbais;
b) recursos não verbais;
c) aspectos conceituais.

A categoria aspectos verbais é constituída por variantes denominativas que contemplam expressões verbais das lexias de buscas empregadas pelos usuários do Portal LexMl. A categoria recursos não verbais, por sua vez, é formada por lexias de buscas constituídas pela combinação de números e palavras ou por números puramente, razão pela qual preferimos chamá-las de "lexias alfanuméricas" e "lexias numéricas", respectivamente. A categoria aspectos conceituais, por outro lado, é constituída por lexias que apresentam

como característica as partes da definição de um termo ou uma pergunta. Neste artigo, vamos abordar apenas a categoria aspectos conceituais, que é constituída por 32 lexias. Essas lexias, com aspectos conceituais, foram identificadas a partir da aplicação dos critérios de funcionalidade e ocorrência terminológica, ou seja, foram selecionadas lexias que desempenham o papel de termo no âmbito do Direito. Suas características apresentam indícios de uma busca realizada por um usuário menos especializado, que sabe o que quer mas não domina a terminologia da área. No caso das perguntas, destaca-se o fato das mesmas serem muito parecidas com as perguntas que os usuários fazem/faziam para o bibliotecário (ver entrevista de referência) quando iam/vão até uma biblioteca para buscar informação.

Para organização dos dados, realizamos os seguintes procedimentos:

a) extração dos logs de pesquisa para planilhas do *software* Microsoft Excel, 2013. A extração foi realizada com o auxílio de um extrator de logs elaborado por um profissional da área de informática, especificamente para a realização desta pesquisa. O *software* em questão foi programado para dispor o conjunto de dados relativos à pesquisa de cada usuário do Portal em linhas e colunas. Desta forma, para cada pesquisa é possível identificar o número de IP, a data, o horário, e a lexia de busca utilizada pelo usuário para recuperação da informação;
b) após a extração, os logs foram agrupados pelo número de IP, para que pudéssemos fazer uma análise global de todas as lexias de buscas utilizadas pelos usuários no decorrer de sua interação com o Portal;
c) foi realizada a listagem das lexias de busca relacionadas às áreas do Direito do Trabalho e do Direito Previdenciário as quais foram identificadas com o auxílio da Classificação Decimal de Direito e da fonte especializada (jurisprudência);
d) verificação da ocorrência ou não das lexias de busca utilizadas pelos usuários do Portal LexML, no corpus textual especializado (constituído pela CLT e pelas leis da Previdência) e na fonte especializada (base de jurisprudência disponível no Portal LexML);

e) comparação das lexias de buscas utilizadas pelos usuários do Portal LexML com os descritores existentes no Vocabulário Controlado Básico (VCB).

4 Resultados: aspectos conceituais

A categoria aspectos conceituais surgiu a partir da identificação, em alguns casos, do conhecimento por parte dos usuários das definições dos conceitos que buscam, em detrimento do termo normalmente utilizado no âmbito do Direito do Trabalho e do Direito Previdenciário para designá-los; bem como da não identificação de uma categoria em que se pudesse classificar essa tipologia de variantes na Classificação Formal de Variantes Denominativas, de Freixa (2002, 2014) que utilizamos como parâmetro para a identificação e classificação das variantes denominativas do corpora de pesquisa. A identificação desta tipologia de variação mostra a importância da dimensão conceitual das pesquisas dos usuários e a necessidade de também considerarmos esses aspectos no momento da representação dos assuntos identificados em uma obra (indexação).

A quantidade de lexias identificadas nessa categoria não foi muito expressiva no *corpus* analisado – apenas 32 (listadas no APÊNDICE A), no entanto, devido à relevância da dimensão conceitual na recuperação da informação, optamos por apresentá-las em uma categoria própria. A busca por uma parte da definição de um conceito, ou a formulação de uma pergunta como expressão de busca é também um indício sobre o nível de conhecimento do usuário em relação ao assunto pesquisado. Sua utilização nos permite inferir se o usuário que a emprega, como parte da sua estratégia de busca em um sistema de informação, é ou não um especialista das áreas do Direito pesquisadas. A busca por partes da definição de um termo é uma característica de pesquisa importante de ser destacada, pois nos permite inferir que o usuário que a emprega não é um especialista da área pesquisada, se trata, possivelmente, de um usuário semi-leigo no assunto. Nesses casos, o usuário sabe o que quer, mas não sabe expressar de forma objetiva a sua necessidade de informação, em outras palavras, sua forma de expressão ainda não é reconhecida como um descritor ou

termo equivalente na linguagem utilizada pelo SRI. Um exemplo dessa situação é a lexia "auxílio-alimentação" que, embora seja um tipo de "salário utilidade", não se encontra como descritor na linguagem documentária utilizada pelo Portal LexMl (VCB- Vocabulário Básico e Autoridades).

Outro exemplo que se enquadra nessa categoria são as lexias de busca "estabilidade gestante" e "estabilidade gravidez". Em verificação realizada no VCB e no *corpus* especializado constituído pelas leis do Trabalho e da Previdência, não identificamos a ocorrência dessas lexias. O termo semanticamente mais próximo que encontramos, nessas duas fontes, foi "estabilidade provisória". Porém, tanto a lexia "estabilidade gestante", como "estabilidade gravidez", ocorrem no *corpus* especializado de jurisprudência. Ao buscarmos a definição dessas lexias no dicionário de Direito, nosso *corpus* de referência, identificamos a ocorrência de variações apenas da lexia "estabilidade gestante" no corpo do texto da definição do termo "estabilidade provisória". Efetivamente, as lexias utilizadas pelos usuários do Portal LexMl não ocorrem como um termo, com entrada específica no Dicionário de Direito Processual do Trabalho. No entanto, ocorre a utilização de variações da lexia "estabilidade gestante" no discurso do especialista, conforme podemos observar em alguns trechos da definição do termo "estabilidade provisória", que transcrevemos a seguir:

> Estabilidade Provisória: [...] A estabilidade provisória consiste na restrição transitória e temporária, decorrente de um fato ou evento específico, ao direito *potestativo* do empregador de resilir o contrato de trabalho, do que resulta que o empregado que detém estabilidade provisória não pode ser despedido sem justa causa. [...] A estabilidade provisória encontra-se expressa em lei ou em acordos ou convenções coletivas de trabalho. Modalidades legais de estabilidades temporárias: *Gestante*. Para a empregada, é vedada a dispensa sem justa causa no período gestacional, desde a confirmação da gravidez, estendendo-se a sua estabilidade até cinco meses após o parto. A empregada doméstica faz juz à *estabilidade gestacional*, nos termos da Lei n. 11.324/2006. [...] Atualmente, encontram-se consolidados no âmbito do Tribunal Superior do Trabalho os seguintes entendimentos sobre a *estabilidade da gestante*: (SCHWARZ, 2012, p. 431-432, grifo nosso).

Diante dos resultados, fica evidente a relevância da análise das variações dessas duas lexias como termo e/ou descritor de um tesauro já que constam no discurso do especialista, tendo sido identificadas tanto em textos de

jurisprudência (especificamente na Base de Dados de Jurisprudência do Portal LexMl), como no texto de um dicionário especializado nas áreas analisadas. Soma-se a isso o fato de percebermos, através da definição de "estabilidade provisória", que o termo é genérico e comporta diferentes modalidades de estabilidades temporárias: um claro indício da relação hierárquica gênero/espécie. Identificamos a mesma situação com o termo "auxílio-alimentação", pois, ao verificarmos o VCB e o *corpus* especializado constituído pelas leis do Trabalho e da Previdência, não identificamos a ocorrência dessa lexia. O termo semanticamente mais próximo que encontramos nessas fontes foi "salário utilidade". Porém, a lexia "auxílio-alimentação" ocorre tanto no *corpus* especializado de jurisprudência, como no corpo do texto da definição do termo "salário utilidade" do Dicionário de Direito Processual do Trabalho. E, assim como o termo "estabilidade provisória", o termo "salário utilidade" também é genérico e abrange outros tipos de utilidade, tais como: auxílio transporte, auxílio moradia, etc.

No caso das perguntas, identificamos expressões como: "qual a forma de cálculo do FGTS"; "projeto de lei que quer aumentar jornada dos pilotos"; "nova lei para motoboys". Nesse tipo de busca, além das características do usuário é possível identificar também a manifestação de uma necessidade dos mesmos. Quanto às características do usuário, é possível inferir que se trata de um leigo no assunto, visto que não conhece a terminologia da área que está pesquisando. Quanto à necessidade, as expressões de busca desses usuários estão demonstrando que os catálogos tradicionais precisam incluir novas funcionalidades aos recursos de buscas existentes, como a possibilidade de resposta direta à uma pergunta além da oferta dos documentos que podem responde-la. Para atender à essa demanda, é necessário pensar no uso e implantação de ontologias, bem como reformular as técnicas tradicionais de representação da informação e do conhecimento tradicionalmente empregada pelos bibliotecários.

Entendemos que as lexias identificadas nessa categoria reforçam nosso argumento de que os sistemas de busca precisam estar o mais próximo possível

da linguagem dos usuários, visto que a pergunta que antes era feita aos bibliotecários, agora é realizada por meio dos sistemas de busca.

5 Conclusão

No contexto atual, é preciso contar com diferentes estratégias para acompanhar as necessidades informacionais e as características específicas dos usuários de informação. Entre as estratégias possíveis, destacamos o estudo de logs, em função das diferentes possibilidades de pesquisa que eles viabilizam. No âmbito da CI, há um grande potencial para a utilização dos logs como objeto de estudo e fonte de coleta de dados, visto que ainda há poucos estudos que os utilizem.

Em levantamento realizado em 2012, encontramos apenas um trabalho no Brasil que menciona a análise de logs. Pontes (2006), em sua monografia de especialização em Gestão de Unidades de Informação pela Universidade Federal da Paraíba, fala sobre a contribuição do estudo de logs em pesquisas sobre gestão da informação, destacando a importância desses dados como ferramentas de apoio para a tomada de decisões e aperfeiçoamento dos Catálogos Online de Acesso Público (OPAC). Atualmente no Brasil, além do trabalho mencionado, encontramos uma tese (LAIPELT, 2015) e quatro trabalhos de conclusão de curso (TCC) realizados por alunos do curso de Biblioteconomia da Universidade Federal do Rio Grande do Sul que utilizam a análise de logs como fonte para a coleta de dados (FERRO, 2014; KREBS, 2013; RECH, 2013; SILVA, 2013).

Isto é resultado, também, da baixa frequência com que as unidades de informação coletam os logs referentes às pesquisas dos usuários. Geralmente, os *softwares* de gestão da informação das instituições estão configurados para apagar esses dados automaticamente dos sistemas, para que não ocupem espaço em seus servidores. Entendemos que, em função do espaço ocupado nos servidores, não seria necessário guardar todos os logs de acesso dos usuários ao sistema. No entanto, é possível configurar os *softwares* para coletarem e guardarem os logs apenas por alguns meses, de modo que o espaço ocupado nos servidores seja mínimo. Para isso, basta determinar uma política para a coleta e

armazenagem desses dados e configurar os *softwares* de gestão de acervos, realizando, então, essa tarefa automaticamente.

Os resultados obtidos neste estudo confirmam, portanto, a relevância dos logs de pesquisa como fonte para a coleta de dados referentes ao léxico empregado pelos usuários para a recuperação da informação disponibilizada em sistemas de recuperação da informação, tais como: catálogos de bibliotecas, repositórios, bases de dados, etc. A observação da interação dos usuários com o SRI, durante o processo de recuperação da informação, sobretudo do léxico empregado, possibilita o reconhecimento de características importantes desses usuários; tanto para a tomada de decisão em relação à determinação de descritores, como para a constituição e ampliação da rede de remissivas e a identificação de demandas e necessidades informacionais dos usuários.

Referências

BRASCHER, M.; CAFÉ, L. Organização da Informação ou Organização do Conhecimento?. In: LARA, M. L. G.; SMIT, J. W. (Org.). **Temas de pesquisa em Ciência da Informação no Brasil**. São Paulo: USP, 2010. p. 85-102

CARNEIRO, M. V. Diretrizes para uma política de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v.14. n.2, p. 221-241, set. 1985.

CHAUMIER, J. Indexação: conceito, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v.21, n.1/2, p. 63-79, jan./jun. 1988.

DAHLBERG, I. Knowledge organization: a new science? **Knowledge Organization**, Würzburg: Ergon-Verlag, v. 33, n.1, p. 11-19, 2006.

FERRO, V. **A indexação e o usuário**: análise de expressões de busca do direito penal no portal LEXML. 2014. 124f. Trabalho de Conclusão de Curso (Graduação)-Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

FREIXA, J. **La variació terminològica:** anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient. 2002. 569f. Tese (Doutorado) - Universitat Pompeu Fabra. Barcelona, 2002.

FREIXA, J. La variación denominativa en terminología: tipos y causas. In: ISQUERDO, A. N. D. C.; MANTOVANI, G. O. **As ciências do léxico**: lexicologia, lexicografia e terminologia. Campo Grande: UFMS, 2014. vol. 6.

GAUDIN, F. La socioterminologie. **Langages**, Paris, v. 39, n. 157, p. 81-93, 2005. Disponível em: <http://www.persee.fr/web/revues/home/prescript/article/lgge_0458-726x_2005_num_39_157_976>. Acesso em: 1 out. 2014.

GUEDES, R. M.; DIAS, E. J. W. Indexação social: abordagem conceitual. **Revista ACB**, Florianópolis, v. 15, n. 1, p. 39-53, 2010. Disponível em : <http://www.brapci.inf.br/_repositorio/2010/06/pdf_fcb17df2cd_0010808.pdf >. Acesso em : 12 nov. 2012.

KREBS, L. M. **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.

LAIPELT, R. C. F. **Metodologia para seleção de termos equivalentes e descritores de tesauros**: um estudo no âmbito do Direito do Trabalho e do Direito Previdenciário. 2015. Tese (Doutorado) - Programa de Pós-Graduação em Linguística Aplicada, Escola da Indústria Criativa : comunicação, design e linguagens, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2015.

LANCASTER, F.W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos Livros, 2004.

LIMA, J. O.; ALVARES, L. Organização e representação da informação e do conhecimento. In: ALVARES, L. (Org.) **Organização da informação e do conhecimento**: conceitos, subsídios interdisciplinares e aplicações. São Paulo: B4 Editores, 2012. 248p.

MOREIRO GONZÁLEZ, J. A. **El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural**. Gijón: Trea, 2004.

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.

PINHO, F. A. **Fundamentos da organização e representação da informação e do conhecimento**. Recife : UFPE, 2009.

PONTES, A. M. de. **OPAC como recurso para a gestão da informação no contexto da biblioteca central da UFPB.** 2006. Trabalho de Conclusão de Curso (Especialização em Gestão de Unidades de Informação) - Departamento

de Biblioteconomia e Documentação, Centro de Ciências Sociais Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2006.

RECH, C. A. **Expressões de busca dos usuários do portal LexML do Senado Federal Brasileiro:** uma análise comparativa com a linguagem documentária utilizada pelo portal. 2013. 53 f. Trabalho de Conclusão de Curso (Graduação) - Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

ROWLEY, J. **A biblioteca eletrônica**. 2. ed. Brasília: Briquet de Lemos, 2002.

SCHWARZ, R. G. (Org.) **Dicionário de direito do trabalho, de direito processual do trabalho e de direito previdenciário aplicado ao direito do trabalho**. São Paulo: LTR, 2012.

SILVA. C. G. M. da. **Avaliação das expressões de busca para a recuperação da informação utilizadas pelos usuários do ramo do Direito de Família do Portal LexML do Senado Federal.** 2013. 65f. Trabalho de Conclusão de Curso (Graduação)-Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

Log analysis as a strategy for user warranty

Abstract: The aim of this paper is to discuss the potential for collection of research logs for user warranty in the construction of thesauri and/or for the management of authorities catalogs. It demonstrates how log analysis allows identifying users' language from the information retrieval system itself. The lexical units in the fields of labor law and social security law found in the Brazilian Senate's web portal (LexMl) research logs are used as the object of study. It presents the results for the analysis of lexical units with conceptual aspects, which indicate the user's level of knowledge about the researched topic. The paper concludes that logs are important for the identification of users' characteristics, demands, and needs. It is suggested that collection policies be formulated, given that, in Brazil, few research studies use logs as their object of study in the field of information science.

Keywords*:* Knowledge representation. Information retrieval. Log analysis. User warranty.

Recebido em 03/11/2015

Aceito em 22/12/2015

[1] RAFFERTY, Pauline; HIDDERLEY, Rob. Flickr and democratic indexing: dialogic approaches to indexing. **Aslib Proceedings**, Bingley, v. 59, n. 4/5, 2007. p. 397-410. Disponível em: <http://www.emeraldinsight.com/Insight/viewPDF.jsp?Filename=html/Output/Published/EmeraldFullTextArticle/Pdf/2760590407.pdf>. Acesso em: 1 maio 2008.

[2] BROWN, Pauline et al. The democratic indexing of images. **New Review of Hypermedia and Multimedia**, Abingdon, v. 2, n. 1, p. 107-120, 1996.

APÊNDICE A – Lexias aspectos conceituais

1.	Falta de deposito de fundo de garantia
2.	Rescisão indireta
3.	Licença maternidade estabilidade
4.	Estabilidade provisória
5.	Legislação motoboy
6.	Nova lei para motoboys
7.	Lei do estágio
8.	Lei do idoso
9.	Lei do trabalho
10.	Qual a forma de cálculo dos depósitos em atraso do FGTS
11.	Qual a forma de cálculo do FGTS
12.	Lei que exeta imposto de pessoas portadoras de doenças grave
13.	Projeto de lei que quer aumentar jornada dos aeronautas
14.	Projeto de lei que quer aumentar jornada dos pilotos
15.	Pec da obrigatoriedade do curso de jornalista
16.	Pec dos jornalistas;
17.	Pec da exigencia de curso de jornalista
18.	regulamentação taxistas
19.	regulamenta a profissão de físico
20.	lei que fala sobre aposentadoria por invalides
21.	lei sobre doenças ocupacionais
22.	lei sobre medidas preventivas acidentes de trabalho
23.	lei trabalhista rural
24.	Lei das domesticas
25.	Lei Estéticista
26.	Lei Barbeiro, Cabelereiro, Estéticista, Manicuri
27.	Lei Massagem
28.	Lei Massoterapia
29.	auxílio-alimentação
30.	estabilidade gestante
31.	estabilidade gravidez
32.	Falta de deposito de fundo de garantia

Fonte: Elaborado pela autora a partir dos dados da pesquisa.