DOI: 10.5433/1981-8920.2018v23n3p390

A UTILIZAÇÃO DA RECUPERAÇÃO DA INFORMAÇÃO NAS TESES DOUTORAIS DA BASE DO INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA – IBICT

THE USE OF INFORMATION RETRIEVAL IN DOCTORAL THESIS BASE OF THE BRAZILIAN INSTITUTE OF INFORMATION IN SCIENCE AND TECHNOLOGY- IBICT

Adilson Luiz Pinto[a]
Cleber da Silva André[b]
Ana Cristina de Albuquerque[c]

RESUMO
Introdução: Na atual conjuntura da sociedade a informação adquiriu grande valor, levando a necessidade de ferramentas que permitam a recuperação de informações relevantes a todo tipo de usuário. Considerando os Sistemas de Recuperação da Informação no meio acadêmico e sua importância na busca eficiente e eficaz de informações relevantes, questiona-se como averiguar a utilização nas teses doutorais? **Objetivo:** O artigo apresenta como Objetivo Geral a análise da incidência dos termos Recuperação de Informação nas teses encontradas na Base de Teses e Dissertações do IBICT. No decorrer da pesquisa conceituam-se alguns modelos e técnicas de Recuperação da Informação como forma demonstrar aplicações e teorias citadas nas teses que fazem parte do corpus da pesquisa. **Metodologia:** A pesquisa, quanto aos objetivos, apresenta caráter descritivo e exploratório, e documental quanto aos seus procedimentos. Também são abordados conceitos de estudos métricos da informação para verificar a incidência dos termos de Recuperação da Informação nos títulos, resumos e palavras-chave das teses para chegar aos objetivos propostos. **Resultados:** Com os dados levantados, apresenta-se a relação entre a citação

a Professor Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina. E-mail: adilson.pinto@ufsc.br.

b Bibliotecário na Biblioteca Pública Municipal Ary Cabral – Brusque-SC. Mestrando do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina. E-mail: cleber_csa@yahoo.com.br.

c Professora do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual de Londrina. E-mail: albuanati@uel.br.

do termo Recuperação nos títulos e palavras-chave e resumos. **Conclusões:** Com os resultados alcançados verifica-se a grande incidência de termos relacionados à Recuperação da Informação nas teses doutorais, reforçando sua aplicabilidade acadêmica e sua pluralidade teórica.

Palavras-chave: Recuperação da Informação. Análise de Teses. Análise de dados.

1 INTRODUÇÃO

O presente artigo visa à verificação, dentro da Base de Teses e Dissertações do Instituto Brasileiro de Informação Ciência e Tecnologia (BDTD do IBICT), da utilização dos sistemas de recuperação da informação nas teses de doutorado. Pretende-se fazer uma revisão bibliográfica para verificar, por meio dos termos utilizados nos resumos, títulos e palavras-chave, o que está sendo utilizado para recuperar as informações relevantes nas pesquisas dos doutorados.

Na atual conjuntura da sociedade, a informação adquiriu grande valor levando a necessidade de modelos que permitam a recuperação de informações relevantes a todo tipo de usuário. Com aumento da importância de métodos eficazes de recuperação da informação, de maneira automatizada, em decorrência do volume de dados não estruturados em todas as áreas, bem como o número crescente de fontes de documentos na Internet, implicou grande interesse no meio acadêmico/científico.

O uso de modelos de Recuperação da Informação mostra-se recorrente no meio acadêmico/científico, pois facilita e agiliza o acesso a informação desejada.

Uma forma de entender a importância da recuperação da informação é lançar mão de conceitos que corroborem para essa afirmação.

Manning, Raghavan e Schutze (2008) afirmam que o objetivo da recuperação de informação é encontrar materiais (geralmente documentos) de natureza não estruturada (geralmente texto), que satisfaça uma necessidade informacional de uma grande coleção de documentos (geralmente armazenados num computador).

Um processo de localização de itens de informação armazenados, a fim de permitir o acesso de usuários, por meio de solicitação é uma definição de recuperação de informação segundo Belkin e Croft (1987).

Ingwersen (1992) argumenta o interesse da recuperação da informação nos processos que envolvem a representação, armazenamento, pesquisa e descoberta da informação relevante às necessidades informacionais dos usuários.

Baeza-Yates e Ribeiro-Neto (1999) expõe que a recuperação da informação trata da representação, armazenamento, organização e acesso à informação.

Xie (2008) coloca a recuperação da informação como disciplina que lida com recuperação de dados não estruturados, de maneira especial os documentos textuais, em resposta a uma instrução de consulta ou tópico, que pode ser desestruturada, por exemplo, uma frase ou até mesmo outro documento, ou que possam ser estruturado, por exemplo, uma expressão booleana.

Robredo (2005) define a recuperação da informação como a finalidade do trabalho documentário que envolve os processos de seleção, aquisição, descrição bibliográfica, análise e indexação.

Lancaster (1978) confirma que recuperação da informação é um termo sinônimo de busca de literatura sendo, portanto, um processo para se buscar uma coleção de documentos.

A recuperação da informação, segundo Rijsberger (1995, p. 2), apresenta objetivos como:

> Estudar e entender os processos que acontecem na recuperação da informação para projetar, construir e testar sistemas de RI que facilitem a comunicação efetiva da informação desejada entre o gerador da informação e o usuário humano; A linguagem de consulta utilizada normalmente é a natural; A especificação de consulta é incompleta; O item desejado é relevante, devido muitas vezes à falta de informações; Resposta de erro pode não ser percebida devido à linguagem utilizada.

Sharma (2013) afirma que os principais problemas abordados pela recuperação da informação convergem na necessidade de métodos que recuperem a informação de maneira eficaz, de forma automatizada, devido ao grande volume de documentos não estruturados armazenados na internet, especialmente dando ênfase a documentos textuais, uma vez que, em uma coleção de documentos de

linguagem natural não estruturados, não há bem definida posição sintática em que um motor de busca poderia encontrar estes dados com uma determinada semântica.

Para Xie (2008) os estudos de RI estão divididos em duas abordagens: uma orientada para o usuário e outra para o sistema. Greengrass (2000) ainda reforça duas outras abordagens: a semântica e a estatística.

Souza (2006, p. 163) relaciona as atividades desempenhadas pelos Sistemas de Recuperação da informação:

> • Representação das informações contidas nos documentos, usualmente através dos processos de indexação e descrição dos documentos;
> • Armazenamento e gestão física e/ou lógica desses documentos e de suas representações;
> • Recuperação das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as necessidades de informação dos usuários. Para isso é necessário que haja uma interface na qual os usuários possam descrever suas necessidades e questões, e através da qual possam também examinar os documentos atinentes recuperados e/ou suas representações.

Os sistemas de recuperação apresentam uma série de modelos de busca da informação, para executar suas atividades, onde geralmente estão intrínsecos a softwares e ferramentas automatizadas, através de algoritmos de Recuperação da Informação.

Assim, questiona-se, como averiguar a utilização dos Sistemas de Recuperação de Informação nas teses de doutorado?

Para responder à questão, o artigo tem por **objetivo Geral** analisar, nas teses encontradas na BDTD do IBCTI, quais utilizaram o termo Recuperação da/de Informação, de maneira e mensurar sua incidência.

Com isso, busca-se alcançar os seguintes objetivos específicos: a) Verificar a incidência de teses que abordaram ou utilizaram a RI; b) Nomear as instituições onde as teses foram defendidas; c) elencar as áreas do conhecimento que utilizaram a RI; d) localizar a posição no texto do termo Recuperação de/da Informação; e) identificar os termos relacionados a RI mais recorrentes nas teses.

Como justificativa entende-se que a pesquisa aborda assunto utilizado no

meio acadêmico científico, apresentando grande variedade de ferramentas de auxílio a pesquisa e de fundamental importância na busca por informações relevantes.

A seguir, como referencial teórico, conceitos e modelos de recuperação da informação serão abordados.

2 MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO

A grande dificuldade atualmente, referente à busca da informação, principalmente na busca automatizada, está em definir quais os documentos recuperados são ou não relevantes as necessidades dos usuários. Como, por exemplo, em uma pesquisa utilizando algum motor de busca na web, se recuperam milhares de documentos, distribuído em *n* páginas. Como definir o que é ou não importante de maneira rápida e eficiente?

Os documentos precisam ser indexados, automática ou manualmente, de maneira a existir parâmetros para que as pesquisas possam ser delimitadas e os buscadores apresentarem respostas satisfatórias.

Dessa forma, algoritmos de recuperação são utilizados em modelos de Recuperação da Informação para facilitar as buscas, onde se dividem em modelos clássicos e modelos estruturados (BAEZA-YATES; RIBEIRO-NETO, 1999).

Devido ao grande número de modelos de RI existentes; a constatação da utilização dos modelos tradicionais serem mais recorrentes na literatura consultada no desenvolvimento da pesquisa; e o foco do artigo em explanar quantitativamente o seu uso e não os modelos abordados, no decorrer desta sessão serão conceituados os modelos clássicos e seus derivativos para ilustração das possibilidades de aplicação. Entender porque o assunto Recuperação da Informação é tão recorrente no meio acadêmico e, por consequência, nas teses doutorais, optou-se por apresentar uma série de conceitos e modelos que auxiliaram na compreensão da escolha do tema.

Antes de definirmos os modelos de recuperação da informação mais

conhecidos, resolveu-se selecionar alguns conceitos dos Sistemas de Recuperação da Informação mais abordados, como *query, adhoc, filtering* e *clustering.*

Segundo Souza (2006), *query* é uma questão levantada pelo usuário durante sua busca, uma necessidade de informação levada ao sistema. *Adhoc* é um modo de operação de recuperação da informação, quando há poucas alterações no acervo de documentos enquanto são submetidas novas *queries* ao sistema. Já o *filtering* (filtragem) é um modo de operação oposto, pois novos documentos são adicionados ao acervo, mesmo com os *queries* estáticos. Por fim, no método *Clustering*, grandes aglomerados são formados automaticamente de acordo com a semelhança de (índice) e termos que eles contêm. Estes são divididos em menores (e mais densos) *clusters*.

Crestani e Pasi (1999) descrevem que *Clustering* na Recuperação de Informação é um método para particionar um determinado conjunto de documentos *D* em grupos, usando uma medida de similaridade que é encontrado em todos os pares de documentos.

Souza (2006, p. 166) explica que “A filtragem acontece usualmente em processos de monitoração de fontes de informação, enquanto a recuperação *adhoc* representa as buscas usuais em SRIs.”

Estes são alguns conceitos básicos, que dão uma ideia de funcionamento e utilização dos SRIs, mostrando sua complexidade de implementação e aplicação.

Partindo dos conceitos anteriormente citados, a seguir estão descritos os modelos clássicos de RI, a fim de apresentar alguns que podem ter sido utilizados nas teses trabalhadas.

2.1 Modelos clássicos

Para ser considerado um modelo de Recuperação da Informação são necessários quatro elementos fundamentais: 1. Como será representado o documento; 2. Como serão feitas as consultas; 3. Como relacionar a relação das consultas e a representação dos documentos; 4. Como classificar (ranqueamento)

ou calcular a relevância dos textos selecionados.

Os modelos representados nesse tópico são o *booleano*, o *vetorial* e o *probabilístico*, conhecidos como modelos clássicos em Recuperação da Informação. Esses modelos consideram que cada documento é descrito por um conjunto de palavras-chave representativas chamadas *termos de índice*. Um termo de índice é simplesmente uma palavra (documento) cuja semântica ajuda a lembrar de principais temas do documento. Assim, os termos de índice são usados para indexar e resumir o conteúdo do documento.

Os conceitos de modelo booleano, vetorial e probabilístico, assim como os seus modelos derivados, descritos a seguir, são verificados em Baeza-Yates e Ribeiro-Neto (1999).

2.1.1 Modelo Booleano

No modelo booleano os termos representam os documentos, e a consulta utiliza termos e operadores booleanos, relaciona através da teoria matemática booleana. A vantagem é que se designa em modelos lógicos bem estabelecidos (binário - verdadeiro ou falso). Esse modelo booleano clássico apresenta estrutura formal e conhecida de como fazer a busca, utilizando os operadores de lógica booleana AND, OR, NOT, porém não há correspondência parcial, ou é relevante ou não é. Se o usuário não conhece os operadores pode aplicar formas que dificultem a recuperação desejada.

2.1.2 Modelo Vetorial

No modelo vetorial, documentos e consultas são representados como vetores num espaço *t-dimensional*. Assim, dizemos que o modelo é algébrico. O modelo de vetor reconhece que o uso de pesos binários é limitante e propõe um quadro no qual apresenta possibilidade de correspondência parcial. Isto é possível através da atribuição de pesos não binários para termos de índice em consultas e em

documentos. Ao classificar os documentos recuperados no fim deste grau de semelhança, o modelo vetorial leva em consideração os documentos que correspondem aos termos da consulta apenas parcialmente.

O principal efeito resultante é que o conjunto de respostas de documentos classificados é muito mais preciso (no sentido de que ele corresponda melhor as informações que o usuário procura) do que o conjunto de respostas obtido pelo modelo booleano. Apesar de sua simplicidade, o modelo vetor é uma estratégia de classificação flexível com coleções gerais. Ele produz respostas difíceis de refinar, sem expansão de consulta ou realimentação de relevância.

2.1.3 Modelo Probabilístico

No modelo probabilístico, conhecido como o modelo de recuperação de independência binário (BIR), as representações são baseadas na teoria de probabilidade. A indexação dos termos nos documentos e nas consultas não apresentam relevâncias definidas previamente, portanto os documentos são ordenados através de cálculo dinâmico de seus termos consultados.

Os modelos clássicos ainda possuem modelos derivados, que aprimoram sua usabilidade e desempenho: Teoria dos conjuntos (modelo booleano), teoria algébrica (vetores) e teoria probabilística (probabilidade).

2.1.4 Teoria dos conjuntos

Na teoria dos conjuntos encontramos o *modelo booleano estendido* e o modelo *fuzzy* (difuso).

O modelo booleano estendido tem por finalidade representar um documento como um conjunto difuso de termos, tornando assim a descrição do conteúdo de informações de documentos mais precisos. Para cada termo associado a um documento é atribuído um peso numérico que exprime o grau de pertinência da

informação contida no documento. Atribuir peso dos termos na indexação permite que o mecanismo de busca classifique os documentos por relevância, criando um ranking de pontuação, a recuperação de estado de valor.

Os modelos de lógica difusa (*fuzzy*), uma abordagem da teoria dos conjuntos difusos, onde trabalha com a diversificação e a parcialidade das informações recuperadas. Não apresenta fronteiras bem definidas, sendo bem flexível e permitindo que os membros de um conjunto possam, parcialmente, fazer parte de outros no sistema de recuperação da informação. Esse sistema vai além do modelo booleano clássico, pois sua pertinência não é binária, considerando intensidades de pertinência. O modelo difuso busca superar as limitações do modelo booleano clássico de maneira a generalizar a teoria de conjuntos tradicional.

2.1.5 Teoria Algébrica

Na teoria algébrica encontramos o modelo vetorial generalizado, redes neurais e as a indexação semântica latente.

No modelo vetorial generalizado a independência dos termos índice vista nos modelos clássicos é questionada, possibilitando que certos termos sejam relacionados, examinando co-ocorrências destes termos no texto de cada documento, além das relações semânticas estabelecidas por vocabulários controlados (SOUZA, 2006).

As redes neurais artificiais têm como propriedade importante a capacidade de aprender através de inferências de exemplos utilizados. Ele possui uma arquitetura similar a um sistema de neurônios biológicos que fazem interações, transmitindo sinais que variam de intensidade, que determinaram a relevância das informações transmitidas e recuperadas.

Segundo Ferneda (2006, p. 25),

> Redes neurais constituem um campo da ciência da computação ligado à inteligência artificial, buscando implementar modelos matemáticos que se assemelhem às estruturas neurais biológicas.

> Nesse sentido, apresentam capacidade de adaptar os seus parâmetros como resultado da interação com o meio externo, melhorando gradativamente o seu desempenho na solução de um determinado problema.

Existem duas maneiras do sistema neural aprender, o aprendizado supervisionado, onde um agente externo apresenta padrões de entrada e seus correspondentes padrões de saída, tendo conexões o valor resultante é de uma somatória que vai fazer a função de ativação do neurônio; e o não supervisionado, onde somente padrões de entrada estão disponibilizados para a rede neural, que processa e detecta suas dimensões e as classifica de maneira automática.

Vemos na indexação semântica latente o questionamento da significância das palavras-chave como candidatos a descritores, buscando estabelecer uma relação conceitual entre documentos e *queries*. Neste modelo busca-se mapear cada documento e cada *query* em um espaço menor, construído a partir dos conceitos relevantes que possuem os documentos no acervo (SOUZA, 2006).

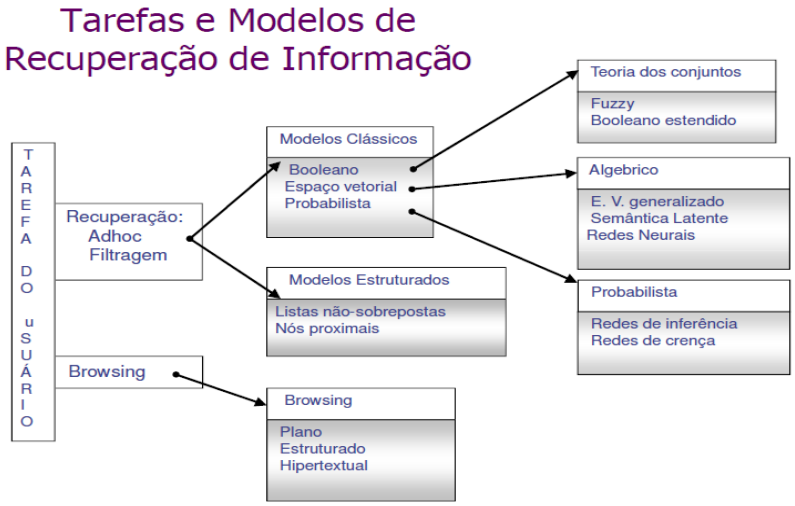
2.1.6 Teoria Probabilística

A teoria probabilística apresenta os modelos de rede de inferência e redes de crença.

> Redes de inferência: nesse modelo, associam-se variáveis aleatórias ao evento do atendimento de uma query específica por um documento específico. Essas variáveis podem ser alteradas de acordo com os eventos futuros, de forma a estabelecer relacionamentos baseados nos eventos observados.
> Redes de crença (belief networks): nesse modelo, similares às redes de inferência, documentos e queries são modelados como subconjuntos de um espaço de conceitos. A cada documento, associa-se a probabilidade de que o mesmo cubra os conceitos presentes no espaço de conceitos. Cada query é mapeada no espaço de conceitos, que por sua vez, está conectado ao espaço de documentos (SOUZA, 2006, p. 167-168).

Na **Figura 1** está representada a taxonomia dos modelos de Recuperação da informação traduzidos de Baeza-Yates e Ribeiro-Neto (1999, p. 21).

Figura 1- Modelos de Recuperação da Informação



Fonte: Baeza-Yates e Ribeiro-Neto (1999, p. 21).

Assim, vem crescendo a pesquisa de novas estratégias de uso dos modelos anteriores citados, pois a necessidades de encontrar formas de tornar as buscas por informação mais eficientes e eficazes é real, da mesma maneira que o aumento exponencial das informações disponibilizadas nos meios de comunicação e comunicação científica.

Ao verificar os conceitos de modelos de RI, tornando mais visível sua importância nas atividades acadêmicas e no desenvolvimento das pesquisas trabalhadas, outros conceitos são trabalhados no desenvolvimento deste artigo, os Estudos Métricos da Informação. Por meio dos Estudos Métricos da Informação é possível atingir os objetivos propostos.

3 CONCEITO DE ESTUDOS MÉTRICOS DA INFORMAÇÃO

Os estudos métricos da informação têm como objetivo avaliar a produção científica de forma a verificar seu impacto em cada área, a visibilidade das informações publicadas, e diversos fatores relacionados a produção, disseminação e recuperação de informação e informação científica.

Neste capítulo serão expostos, de maneira sucinta, alguns conceitos de estudos métricos da informação, apenas como forma de elucidação da importância da metodologia aplicada para evolução deste artigo.

Como se trata de um artigo científico, que tem como corpus de pesquisa a produção de teses de doutorados, optou-se por conceituar três das métricas mais relevantes para o mesmo, a *Bibliometria*, a *Cientometria* e a *Informetria*.

Marques (2010, p. 2), explica a "bibliometria como ferramenta que auxilia as pesquisas existentes entre a ciência da comunicação e ciência da informação utilizando a comunicação científica". Mugnaini (2003, p. 46) propõe bibliometria "[...] como ferramenta capaz de medir e facilitar a análise de informação armazenada." Urbizagástegui-Alvarado (1984, p. 91) coloca a utilização, na Biblioteconomia, da "Bibliometria para significar a aplicação de métodos matemáticos e estatísticos a livros e outros meios de comunicação escrita."

> A bibliometria se refere a uma variedade de regularidades tomadas de diferentes campos, exibindo uma variedade de formas. Embora as distribuições bibliométricas sejam muito diferentes em sua aparência, elas podem ser pensadas como versões de uma única regularidade, de modo que podemos falar em leis bibliométricas e suas manifestações. A Lei de dispersão de Bradford, a Lei de Zipf e a Lei de Lotka são as mais conhecidas, tratando de fenômenos importantes ou de "regularidades" encontradas na comunicação científica (WORMELL, 1998, p. 210).

Araújo (2006, p. 12) apresenta bibliometria "[...] consistindo na aplicação de técnicas estatísticas e matemáticas para descrever aspectos da literatura e de outros meios de comunicação (análise quantitativa da informação) [...]".

Atualmente os cientistas aceitam como correto o termo cientometria, mas durante um tempo havia outras denominações como Cienciometria. Algumas citações ainda apresentam termos já desconsiderados, porém seu conceito remete a Cientometria.

> Cienciometria é o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. A cienciometria é um segmento da sociologia da ciência, sendo aplicada no desenvolvimento de políticas científicas. Envolve estudos

> quantitativos das atividades científicas, incluindo a publicação e, portanto, sobrepondo-se à bibliometria (MACIAS-CHAPULA,1998, p. 134).

Assim como a Cientometria, ainda há discussão entre algumas nomenclaturas das métricas, como é o caso da Informetria, que também é conhecida como Infometria.

"Informetria é o estudo dos aspectos quantitativos da informação em qualquer formato [...], pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites tanto da bibliometria como da cienciometria" (MACIAS-CHAPULA,1998, p. 135).

Wormel (1998, p. 210) coloca que "A informetria é um subcampo emergente da ciência da informação, baseada na combinação de técnicas avançadas de recuperação da informação com estudos quantitativos dos fluxos da informação."

Conceituando essas três métricas da informação podemos entender, a grosso modo, que a Bibliometria é uma ferramenta que auxilia na quantificação e análise da informação bibliográfica armazenada, a Cientometria faz a análise quantitativa da ciência enquanto área de conhecimento ou disciplina e a Informetria trata quantitativamente da informação como um todo, sendo essas três métricas complementares entre si.

Também com os conceitos levantados, verifica-se o caráter relevante desta pesquisa, que tem como foco quantificar a incidência de termos relacionados ao uso de Recuperação da Informação nas Teses encontradas na BDTD do IBICT.

4 ASPECTOS METODOLÓGICOS

O artigo apresenta caráter exploratório e descritivo quanto aos objetivos estabelecidos, se valendo de análise sistemática de grupo específico. Também apresenta caráter documental, levando em consideração os procedimentos técnicos de análise de documentos analisados não processados. Ainda, quanto a abordagem, apresenta caráter misto, sendo quantitativa, pois foram coletados dados

numéricos do corpus pretendido, e foram apresentados resultados qualitativos que corroborando a uma análise qualitativa.

Com o intuito de verificar como a Recuperação da Informação aparece nas Teses da BDTD do IBICT, no dia 03 de abril de 2016, no campo de busca simples, foi utilizado o termo “recuperação da informação”, primeiramente sem aspas. Com o volume excessivo de teses que não correspondem com os objetivos almejados, utilizou o termo entre aspas. Para não correr o risco de perder algumas teses que também fizessem parte do objetivo estabelecido, foi também utilizado na busca o termo “recuperação de informação”.

Nos filtros disponibilizados pela Base, foram selecionadas apenas as teses publicadas no período de 2005 a 2015, mantendo assim um corpus fixo de dez anos de publicações, excluindo as dissertações e ainda restringindo a trabalhos em língua portuguesa.

Foram recuperadas 733 teses e dissertações, antes de aplicar a filtragem. Excluindo as dissertações, restaram 171 teses e, limitando o intervalo aos anos de 2005 a 2015, restaram 136 teses. Restringindo o idioma para apenas a língua portuguesa, finalizou-se o corpus em 106 teses a serem avaliadas.

A coleta das teses, assim como a filtragem, seleção do corpus final e a análise dos dados foram feitas de forma manual, não utilizando softwares de apoio para a extração. Também foi utilizado o método de mineração de dados com o auxílio dos softwares *Microsoft Word* e *Microsoft Ecxel*, assim como para tabulação e apresentação dos resultados, e a utilização de conceitos biblionométricos na análise dos dados tabulados.

Após as buscas e filtragem do corpus, optou-se por analisar a incidência dos termos de Recuperação da Informação nos títulos, resumos e palavras-chave das teses, como foco da pesquisa, a fim de determinar o que está efetivamente sendo aplicado de modelos e técnicas de Recuperação da Informação, por meio dos conceitos dos modelos e técnicas utilizadas na produção das teses.

5 RESULTADOS

Na área de busca da BDTD utilizou-se a expressão “Recuperação da Informação” para recuperação das teses, onde foram recuperadas 733 teses e dissertações, antes de aplicar a filtragem desejada. Ao excluírem-se as dissertações, restaram 171 teses e, limitando a pesquisa em teses de doutorado no intervalo dos anos de 2005 a 2015, restaram 136 teses. Restringindo o idioma para teses apresentadas apenas na língua portuguesa, finalizou-se o corpus em 106 teses a serem avaliadas.

Na **Tabela 1** está relacionado o tempo de publicação do corpus da pesquisa, onde se avaliou a partir do ano de 2005 até o ano de 2015, totalizando onze anos.

Tabela 1 – Tempo das publicações do corpus da pesquisa

N anos	Ano	Teses/Ano
1	2013	14
2	2011	13
3	2014	13
4	2015	11
5	2007	9
6	2009	9
7	2012	9
8	2005	8
9	2008	8
10	2010	7
11	2006	5
	Total	**106**

Fonte: Dados coletados e analisados na pesquisa

Pode constatar-se, por meio da **Tabela 1**, uma oscilação na defesa de teses que tratam da Recuperação da Informação, segundo o termo de busca utilizado. Aqui ainda não foram avaliadas quais teses efetivamente tratam do assunto. Apenas considera-se que o sistema de busca da BDTD do IBICT recuperou as teses do assunto solicitado.

A **Tabela 2** relaciona as instituições nas quais as teses foram defendidas. Num total de quinze instituições responsáveis pelas cento e seis Teses.

Tabela 2 – Instituições de defesa das Teses

N instituições	Instituição	Teses/Instituição
1	UFMG	35
2	UFSC	19
3	UNESP	11
4	USP	10
5	UNICAMP	7
6	UNB	5
7	UFRGS	4
8	ITA	3
9	PUCRS	3
10	FGV	2
11	UFRJ	2
12	UTFPR	2
13	FVG	1
14	UFSCar	1
15	UFUber	1
	Total	**106**

Fonte: Dados coletados e analisados na pesquisa

Com o levantamento dos dados verifica-se que a maioria das Teses, que apresentam o assunto foco da pesquisa, são produzidas em cinco instituições. A UFMG é a que apresenta maior número de teses sobre o tema Recuperação da Informação, seguida pela UFSC, UNESP, USP e UNICAMP. Salientando que esses dados são referentes a coleta de dados nos títulos, resumos e palavras-chave.

As áreas do conhecimento que mais trabalharam os temas de Recuperação da Informação no período definido, utilizando seus modelos, métodos, técnicas e conceitos são verificadas na **Tabela 3**.

Tabela 3 – Áreas do conhecimento das Teses

N áreas	Área	N teses
1	Ciência da Informação	40
2	Ciência da Computação	15
3	Engenharia e Gestão do Conhecimento	12
4	Engenharia Elétrica	6
5	Não determinada	6
6	Administração	4

7	Engenharia de Produção	3
8	Ciências	2
9	Comunicação e Informação	2
10	Engenharia da Computação	2
11	Engenharia Mecânica	2
12	Linguística	2
13	Cultura e Informação	1
14	Engenharia Biomédica	1
15	Engenharia Civil	1
16	Engenharia em Sistemas Eletrônicos	1
17	Física	1
18	Geografia	1
19	Informática Industrial	1
20	Informática na Educação	1
21	Letras	1
22	Psicobiologia	1
	Total	**106**

Fonte: Dados coletados e analisados na pesquisa

Como se pré-supunha, poucas áreas do conhecimento são responsáveis pela grande maioria das teses relacionadas a Recuperação da Informação. A Ciência da Informação, a Ciência da Computação e a Engenharia de Gestão do Conhecimentos são responsáveis, respectivamente, pelo maior número de Teses representadas no corpus da pesquisa.

Nas **Tabelas 4**, **5** e **6** são apresentados em que parte do documento foi encontrado o termo Recuperação da Informação, se no título, nas palavras-chave, e a relação de frequência entre eles, respectivamente.

Tabela 4 – Termos encontrados no título

Título	**Citadas**
Não	91
Sim	15
Total	**106**

Fonte: Dados coletados e analisados na pesquisa

Dos documentos recuperados, apenas quinze apresentavam o termo Recuperação da Informação no título, ficando evidente que a recuperação da informação não é o foco das Teses, ou pelo menos, não como área de pesquisa.

Tabela 5 – Termos encontrados nas palavras-chave

Palavras-Chave	**Citadas**
Não	69
Sim	37
Total	**106**

Fonte: Dados coletados e analisados na pesquisa

Em relação às palavras-chave, a constatação é a mesma verificada em relação ao título, porém, tem maior incidência, aparecendo em trinta e sete Teses.

Tabela 6 – Termos encontrados entre título e palavras-chave

No título/palavras-chave	**Relação**
Não é citado no título, nem nas palavras-chave	66
Não é citado no título, mas é nas palavras-chave	25
É citado no título e nas palavras-chave	12
É citado no título, mas não nas palavras-chave	3
Total	**106**

Fonte: Dados coletados e analisados na pesquisa

Cruzando os dados entre as teses que apresentam o termo sugerido, tanto nos títulos, quanto nas palavras-chave, verifica-se que a maior parte das Teses não apresenta o termo Recuperação da Informação tanto no título, quanto nas palavras-chave, sendo um total de sessenta e seis Teses. Logo em seguida aparecem as Teses que só apresentam citações nas palavras-chave, sendo vinte e cinco Teses. Em seguida temos as teses que apresentam citações tanto no título como nas palavras-chave, sendo doze citações. Por fim, aparecem três Teses que apresentam citações apenas no título.

O que realmente demonstra a aplicação da Recuperação da Informação nas Teses do corpus da pesquisa são os termos levantados nos resumos. Por meio dos resumos foram encontrados centenas de nomenclaturas relacionadas as técnicas, métodos, modelos e conceitos referentes aos Sistemas de Recuperação da Informação, os quais podem ser visualizados na **Tabela 7**.

Tabela 7 – Termos relacionados a Recuperação da Informação

N termos	**Termos citados**	**N teses**
1	Ontologias	54

2	Recuperação da informação	43
3	Não continha o assunto	25
4	Indexação	9
5	Web semântica	9
6	Processamento de linguagem natural	8
7	Metadados	7
8	Mineração de dados	6
9	Svm	4
10	Análise semântica	3
11	Inteligência artificial	3
12	Revocação	3
13	Similaridade	3
14	Folksonomia	2
15	Lógica fuzzy	2
16	Maquinas de busca web	2
17	Mineração de texto	2
18	Precisão	2
19	Pré-processamento de dados	2
20	Active learning	1
21	Agentes softwares	1
22	Busca por palavras-chave	1
23	Busca semântica interativa por palavras-chave	1
24	Cadeia de markov	1
25	Classificação facetada	1
26	Classificadores binários	1
27	Clusters	1
28	Data flouws	1
29	Dublin core	1
30	Estratégia de busca	1
31	Estratégia de índice invertido	1
32	Extração automática	1
33	Extração da informação	1
34	Frag-cubing	1
35	Fuzzy-bayernesiano	1
36	Hiperlinks	1
37	Indexação automática	1
38	Indexação facetada	1
39	Information seeking	1
40	Matriz kernel	1
41	Mecanismo de busca	1

42	Mecanismo de busca dinâmica facetada	1
43	Mecanismos de busca na web	1
44	Metabuscador	1
45	Método bayernesiano	1
46	Método k visinhos	1
47	Método ontomet	1
48	Método vetores	1
49	Mineração web	1
50	Modelagem de dados	1
51	Modelo espaço vetorial baseado em conjuntos	1
52	Modelo espaço vetorial padrão	1
53	Modelo fuzzy	1
54	Modelo modrsem	1
55	Modelo probabilístico	1
56	Modelo probabilístico coditional random fields (crf)	1
57	Modelo semântico estatístico	1
58	Modelo vetorial	1
59	Multimodal	1
60	Naivebayes	1
61	Navegação facetada	1
62	Ngram statistics package	1
63	Oai-pmh	1
64	Ogma	1
65	Opennlp estatístico	1
66	Optimum-path forest (opf)	1
67	Probabilidade	1
68	Processamento automatizado	1
69	Processo de busca	1
70	Query expansion	1
71	Recuperação de documentos	1
72	Rede de filas	1
73	Redes neurais artificiais	1
74	Representação interativa	1
75	Similaridade semântica	1
76	Simplekmeans	1
77	Sirilico	1
78	Sistemas de busca	1
79	Sistemas multiagentes	1
80	Smo (sequential minimal optimization)	1
81	Stemming	1

82	Stopwords	1
83	Taxonomia facetada	1
84	Taxonomias	1
85	Term extration	1
86	Vetor probabilidade	1
87	Web information mining	1
88	Weka	1
89	Wordnet	1
90	Wordpress	1
	Total	**260**

Fonte: Dados coletados e analisados na pesquisa

O primeiro fato visualizado ao fazer a tabulação dos dados foi que em vinte e cinco resumos não foram encontrados qualquer menção a recuperação da informação e seus conceitos.

Dois termos se destacam dos demais na frequência de aparecimento nas teses, o termo ontologias, que serve para diversas áreas do conhecimento, mas está diretamente ligada a Recuperação de Informação e o termo Recuperação da Informação propriamente dito. Lembrando que a frequência aqui citada não é o número de vezes que o termo aparece no resumo, sim o número de resumos que apresentam os termos.

Muitos outros termos foram encontrados, a **Tabela 7** apresenta a relação exatamente como se encontrava no resumo, porém muitos são equivalentes. Alguns desses conceitos são bem difundidos e não necessitam elucidações, porém existe a necessidade de conceituar outros. Deste modo é necessária a criação de um pequeno vocabulário (Quadro 1) para esclarecimentos de alguns dos termos encontrados.

Quadro 1- Vocabulário de termos de Recuperação da Informação

Termos	**Significado**
Ontologias	Define termos descritivos para representação de uma área do conhecimento, usadas em aplicações de compartilhamento de informações, tanto humanas como eletrônicas/digitais.

Indexação	É o ato de indexar - é, genericamente, apontar para algo, para alguma informação.
Web semântica	Web que apresenta ferramentas que permitem atribuir significado a seu conteúdo facilitando a interação com usuário.
Processamento de Linguagem Natural	Voltada a Inteligência artificial, aproveitando seus métodos para compreensão da linguagem natural por computadores estabelecendo comunicação homem/máquina.
Metadados	Tem importante função na descrição de um recurso informacional circulando em pontos de acesso sobre o mesmo, dado sobre dado.
Mineração de dados	Processo de descobrir dados importantes em documentos, geralmente digitalizados, tendo grande importância na busca de informações.
SVM	Algoritmo Support Vector Machine (Máquinas de vetores suporte) utilizado em aprendizado de máquinas para classificação automática de documentos de maneira eficiente.
Inteligência artificial	Técnicas de aprendizado de máquinas baseadas em redes neurais, que permitem que as máquinas aprendam de forma automática sem a ação humana.
Máquinas de busca web	Softwares utilizados para pesquisa de conteúdo na web.
Active learning	Técnica utilizada para treinamento de classificadores.
Agentes softwares	São sistemas capazes de perceber seu ambiente por meio de sensores e agir sobre esse ambiente por meio de atuadores. É (normalmente) um programa de software que apoia um utilizador na realização de alguma tarefa ou atividade.
Cadeia de markov	Estados anteriores são irrelevantes para predição dos estados seguintes desde que o estado atual seja conhecido.
Data flouws	Fluxo de dados.
Information seeking	Area de pesquisa preocupada em como obter informação, de origem humana ou computacional.
Matriz kernel	Produto escalar entre dois vetores.
Multimodal	Método de busca integrada (vários modelos)

Naivebayes	Algoritmos de agrupamento e classificação
Ngram statistics package	Software de extrassão de termos, utilizado em uma das teses para associação estatística.
Oai-pmh	É um protocolo baseado na arquitetura cliente-servidor, de iniciativa de arquivo aberto, utilizado para publicar e coletar metadados.
Ogma	Ferramenta para análise de texto.
Opennlp estatístico	Software baseado no modelo estatístico de PLN.
Optimum-path forest (opf)	Métodos de recuperação de imagens por conteúdo baseados em realimentação de relevância e no classificador por floresta de caminhos ótimos
Simplekmeans	Algoritmos de agrupamento e classificação
Sirílico	Sistema de Recuperação de Informação baseado em Teorias da Linguística Computacional e Ontologia
Sistemas multiagentes	Sistemas de agentes que cooperam entre si, se concentrando em um domínio, não em um problema específico.
Smo (sequential minimal optimization)	Otimização mínima sequencial, uma modificação do SVM,
Stemming	Processo de redução de grande número de termos que foram extraídos de um conjunto de documentos.
Stopwords	Palavras não significantes ao conteúdo
Web information mining	Prototipagem rápida em mineração
Weka	Agrega diversos algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas
Wordpress	Sistema de gestão de conteúdo

Fonte: Dados coletados e analisados na pesquisa

O vocabulário disponibilizado acima traz uma descrição dos termos recuperados nos resumos. Nem todos os termos citados foram disponibilizados, mas os recorrentes nas produções que tratam da disciplina de Recuperação da Informação.

6 CONSIDERAÇÕES FINAIS

Com a produção do artigo percebe-se que muitas técnicas e ferramentas da Recuperação da Informação são citadas e debatidas nas Teses analisadas, porém utiliza-se de seus conceitos para focar na forma como usuário os aplica, utilizando pouco nas metodologias como base para reforçar suas teses. É reconhecida a importância das ferramentas computacionais para recuperar informações, definir relevância, explicar as deficiências do usuário em criar estratégias de busca e desenvolver pesquisas de maneira mais eficiente e eficaz, mas o fator humano ainda é o mais discutido no corpus.

Constatou-se que, dentro da Base de Teses e Dissertações do IBICT, a área de Ciências da Informação é a que mais aborda o tema Recuperação da Informação, dentro dos critérios determinados na metodologia, porém, muitas outras áreas se utilizam dessa disciplina, podendo esta ser considerada multivalente.

Por estar diretamente relacionada com a área computacional, já que a maioria de seus modelos é aplicada em softwares de busca, mineração de documentos, reconhecimento semântico, inteligência artificial entre outros, a área de Ciências da Computação é outra grande colaboradora no desenvolvimento dos Sistemas de Recuperação da Informação.

Por fim, como a informação está em todo lugar e, atualmente, em quantidade ilimitada, é de vital importância, principalmente para o meio acadêmico, cada vez mais o desenvolvimento de pesquisa voltado para novas formas de se recuperar informações de maneira eficiente, eficaz, relevante, inteligente e rapidamente.

Também se faz necessário um esforço na produção de softwares e outras funcionalidades que sejam acessíveis, de baixo custo e exijam sistemas computacionais de menor poder de processamento, ou que os sistemas computacionais sejam barateados para que o acesso seja menos dispendioso.

REFERÊNCIAS

ARAÚJO, C. A. Á. Bibliometria: evolução histórica e questões atuais. **Em Questão**, Porto Alegre, v. 12, n. 1, p. 11-32, jan./jun. 2006. Disponível em: <http://revistas.univerciencia.org/index.php/revistaemquestao/article/viewFile/3707/3495>. Acesso em: 05 abr. 2016.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: Addison Wesley, 1999.

BELKIN, N. J.; CROFT, W. B. Retrieval tecniques. **Annual Review of Information Science and Tecnology**, [S.l], v. 22, p. 112-119, 1987.

CRESTANI, F.; PASI, G. Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks. In: **NeuroFuzzy Techniques for Intelligent Information Systems**. Publisher: Physica Verlag (Springer Verlag), 1999.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, v. 35, n. 1, p. 25-30, jan./abr. 2006.

GREENGRASS, E. **Information Retrieval**: A Survey, 2000.

INGWERSEN, P. The Traditional IR Research Approach. In: **Information Retrieval Interaction**. London: Taylor Graham, 1992. p. 72-78.

LANCASTER, F. W. **Information retrieval systems***:* charactistics, testing and evaluation. 2. ed. New York: Wiley, 1978.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, v. 27, n. 2, p. 134-140, maio/ago. 1998. Disponível em: <http://www.scielo.br/pdf/ci/v27n2/macias.pdf>. Acesso em: 05 abr. 2016.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **An introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

MARQUES, A. de A. A bibliometria: reflexões para comunicação científica na Ciência da Comunicação e Ciência da Informação. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 33., 2010, Caxias do Sul. **Anais eletrônicos...** Amazonas: UFAM, 2010. p. 1-10. Disponível em: <http://www.intercom.org.br/papers/nacionais/2010/resumos/R5-2437-1.pdf>. Acesso em: 05 abr. 2016.

MUGNAINI, R. A bibliometria na exploração de base de dados: a importância da Linguística. **Transinformação***,* Campinas, v. 15, n. 1, p. 45-52, jan./abr. 2003.

Disponível em: <http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/1475>. Acesso em: 05 abr. 2016.

RIJSBERGER, C. J. **Information retrieval**. University of Gasgow, 1995.

ROBREDO, J. **Documentação de hoje e de amanhã***:* uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. Brasília, 2005. 409 p.

SHARMA, A. Intelligent Information Retrieval System: A Survey. **Advance in Electronic and Electric Engineering**, v. 3, n. 1, p. 2231-1297, 2013.

SOUZA, R. R. Sistemas de Recuperação de Informações e Mecanismos de Busca na web: panorama atual e tendências. **Perspectiva em ciência da informação**, Belo Horizonte, v. 11 n. 2, p. 161 -173, maio/ago. 2006. Disponível em: <http://www.scielo.br/pdf/pci/v11n2/v11n2a02.pdf >. Acesso em: 22 maio 2016.

URBIZAGÁSTEGUI-ALVARADO, R. A bibliometria no Brasil. **Ciência da Informação**, Brasília, v. 13, n. 2, p. 91-105, jul./dez. 1984. Disponível em: <http://revista.ibict.br/ciinf/article/view/200>. Acesso em: 05 abr. 2016.

WORMELL, I. Informetria: explorando bases de dados como instrumentos de análise. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 210-216, maio/ago. 1998. Disponível em: <http://www.scielo.br/pdf/ci/v27n2/wormell.pdf>. Acesso em: 22 maio 2016.

XIE, I. **Interactive Information Retrieval in Digital Environments**. New York: IGI Publishing, 2008.

THE USE OF INFORMATION RETRIEVAL IN DOCTORAL THESIS BASE OF THE BRAZILIAN INSTITUTE OF INFORMATION IN SCIENCE AND TECHNOLOGY- IBICT

ABSTRACT:
Introduction: In the current conjuncture of the information society acquired great value, taking the need of tools that allow the information retrieval relevant to all types of users. Considering the Information Retrieval Systems in the academic world and its importance in the efficient and effective search of relevant information, it is questioned how to verify the use in the doctoral theses?. **Objective:** The article presents as General Objective the analysis of the incidence of the terms Information Retrieval in the theses found in the Thesis and

Dissertation Database of the IBICT. In the course of the research some models and techniques of Information Retrieval are conceptualized as a way to demonstrate applications and theories cited in the theses that are part of the corpus of the research. **Methodology:** A methodology presents with descriptive and exploratory character and documentary how many years its procedures. The concepts of metric information studies are also addressed to verify the incidence of the terms of Information Retrieval in titles, abstracts and theses keywords to reach the proposed objectives. **Results:** With the data collected, the relationship between the citation of the term Retrieval in titles and keywords and abstracts is presented. **Conclusions:** With the results achieved, there is a high incidence of terms related to Information Retrieval in doctoral theses, reinforcing its academic applicability and theoretical plurality.

Descriptors: Information Retrieval. Thesis Analysis. Data analysis.

LA UTILIZACIÓN DE LA RECUPERACIÓN DE LA INFORMACIÓN EN LAS TESIS DOCTORA DE LA BASE DEL INSTITUTO BRASILEÑO DE INFORMACIÓN EN CIENCIA Y TECNOLOGÍA - IBICT

RESUMEN:
Introducción: En la actual coyuntura de la sociedad la información adquirió gran valor, llevando la necesidad de herramientas que permitan la recuperación de informaciones relevantes a todo tipo de usuario. Considerando los Sistemas de Recuperación de la Información en el medio académico y su importancia en la búsqueda eficiente y eficaz de informaciones relevantes, se cuestiona cómo averiguar la utilización en las tesis doctorales? **Objetivo:** El artículo presenta como Objetivo General el análisis de la incidencia de los términos Recuperación de Información en las tesis encontradas en la Base de Tesis y Disertaciones del IBICT. En el curso de la investigación se conceptualizan algunos modelos y técnicas de Recuperación de la Información como forma de demostrar aplicaciones y teorías citadas en las tesis que forman parte del corpus de la investigación. **Metodología:** La metodología se presenta con carácter descriptivo y exploratorio, y documental en cuanto a sus procedimientos. También se abordan conceptos de estudios métricos de la información para verificar la incidencia de los términos de Recuperación de la Información en los títulos, resúmenes y palabras clave de las tesis para llegar a los objetivos propuestos. **Resultados:** Con los datos levantados, se presenta la relación entre la cita del término Recuperación en los títulos y palabras clave y resúmenes. **Conclusiones:** Con los resultados alcanzados se verifica la gran incidencia de términos relacionados a la Recuperación de la Información en las tesis doctorales, reforzando su aplicabilidad académica y su pluralidad teórica.

Descriptores: Recuperación de la información. Análisis de Tesis. Análisis de datos.