UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS SOCIAIS E APLICADAS
FACULDADE DE BIBLIOTECONOMIA

PATRÍCIA TEIXEIRA DA SILVA

A INDEXAÇÃO AUTOMÁTICA NO PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO: técnicas de análise de assunto operadas pelos softwares de indexação SISA, MHTX e o buscador Google.

BELÉM
2018

PATRÍCIA TEIXEIRA DA SILVA

A INDEXAÇÃO AUTOMÁTICA NO PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO: técnicas de análise de assunto operadas pelos softwares de indexação SISA, MHTX e o buscador Google.

Trabalho de Conclusão de Curso apresentado à Faculdade de Biblioteconomia do Instituto de Ciências Sociais Aplicadas da Universidade Federal do Pará para obtenção de grau de Bacharel em Biblioteconomia.

Orientadora: Profª. Drª. Franciele Marques Redigolo.

BELÉM
2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S586i Silva, Patrícia Teixeira da

A INDEXAÇÃO AUTOMÁTICA NO PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO : técnicas de análise de assunto operadas pelos softwares de indexação SISA, MHTX e o buscador Google / Patrícia Teixeira da Silva. - 2018.

40 f. : il. color.

Trabalho de Conclusão de Curso (Graduação) - Faculdade de Biblioteconomia, Instituto de Ciências Sociais Aplicadas, Universidade Federal do Pará, Belém, 2018.

Orientação: Profa. Dra. Franciele Marques Redigolo

1. Indexação automática. 2. Análise documentária. 3. Softwares de indexação. I. Redigolo, Franciele Marques, *orient.* II. Título

---

CDD 025.47

PATRÍCIA TEIXEIRA DA SILVA

A INDEXAÇÃO AUTOMÁTICA NO PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO: técnicas de análise de assunto operadas pelos softwares de indexação SISA, MHTX e o buscador Google.

Trabalho de Conclusão de Curso apresentado à Faculdade de Biblioteconomia do Instituto de Ciências Sociais Aplicadas da Universidade Federal do Pará para obtenção de grau de Bacharel em Biblioteconomia.

Orientadora: Profª. Drª. Franciele Marques Redigolo.

Banca Examinadora:

_____ - Orientadora

Prof. Dra. Franciele Marques Redigolo

_____ - Membro

Profa. Nara Raimunda de Almeida Santos

_____- Membro

Prof. Ms. Willians Jorge Correa Pinheiro

Dedico este trabalho à minha orientadora Franciele Redigolo por todo apoio e incentivo prestado durante essa fase de estudo e às minhas amigas Patrícia Daniele, Roberta Suellen e Elane Patrícia pela amizade e força prestados em vários momentos da vida ao longo dos quatro anos de graduação.

AGRADECIMENTOS

Agradeço imensamente a minha orientadora Franciele Redigolo, por sua paciência nas orientações e apóio em momentos difíceis.

As minhas amigas Patrícia Daniele, Roberta Suelen e Elane Parícia, por todos esses quatro anos dividindo momentos bons e grande apoio em horas difíceis a as terei sempre em meus pensamentos.

Às amigas Amanda Garcia e Rosana Patrícia, por fazerem parte de momentos importantíssimos no início da vida acadêmica, conhecimentos compartilhados durante a confecção de artigo e pelos momentos de risos e comilanças nos intervalos.

A meu esposo Maurício por está compartilhando esse momento importante da minha vida.

Às amáveis colegas de faculdade e estágios Regina, Carolina Leão, Alcimar Gonzaga, Laís Lobo, Joane Aires, Fernanda por todos os momentos que passamos juntas, com muitas risadas, lanches e trocas de conhecimento.

Aos amigos Edvaldo Moura (Vevé) e Paulo Junior (Jr. Puk), pelo apóio e ajuda durante esses anos, afinal o que eu iria fazer sem computador funcionando para estudar e não fosse o Paulo Jr. para consertar.

As minhas amigas Brenda Lorena e Sheila Aguiar, pela amizade.

A bibliotecária Ivete Botelho, pelo apóio e ensinamentos durante o tempo de estágio no Ministério das Relações Exteriores.

A todos os colegas de faculdade que de forma direta ou indireta contribuíram durante essa trajetória.

Aos professores da Faculdade de Biblioteconomia, por partilharem experiências e ensinamentos nessa fase tão importante.

"...a informação e o conhecimento são criações humanas, grandes organizações só serão bem sucedidas se perceberem que o fluxo de informações dependem das pessoas, não de equipamentos"

Thomas H. Davenport

RESUMO

A pesquisa apresenta a temática no contexto de linguagens documentárias, na qual o problema identificado para o desenvolvimento desta está em descobrir como é desenvolvida a análise de assunto pelos softwares de indexação: SISA, MHTX e o buscador Google. Apresenta como objetivo geral em descobrir diferentes propostas da realização de análise de assunto na indexação automática; quanto aos objetivos específicos: a) apresentar o levantamento histórico da indexação automática e manual; b) identificar diferentes formas de Consistência da indexação automática; c) demonstrar quais são os critérios e instrumentos utilizados para representação nos softwares SISA, MHTX e o buscador Google. Para o desenvolvimento metodológico utilizou-se do método dedutivo e o caráter é exploratório e faz-se uso de levantamento bibliográfico para análise dos dados. Na análise dos dados constata-se diferentes formas para aplicação dos critérios e instrumentos para análise de assunto à indexação automática entre os softwares SISA (Sistema de Indización Semi-Automático) e MHTX e o buscador Google. Desta forma conclui-se que o uso de controle terminológico propicia melhor indexação tanto de forma manual quanto automática, tais controles dar-se por intermédio dos tesauros, vocabulários controlados e ontologias. Assim a aliança entre estudiosos de diferentes áreas do conhecimento empenhados neste propósito, proporcionará uma recuperação da informação com eficiência e eficácia.

Palavras Chaves: Indexação automática. Análise documentária. Software de indexação. Representação da informação.

ABSTRACT

The research presents the theme in the context of documentary languages, in which the problem identified for the development of this is to discover how the subject analysis is developed by indexing software: SISA, MHTX and the Google search engine. It presents as general objective in discovering different proposals of the accomplishment of analysis of subject in the automatic indexation; with specific objectives: a) present the historical survey of automatic and manual indexing; b) identify different forms of Consistency of automatic indexing; c) demonstrate the criteria and instruments used for representation in the software SISA, MHTX and the Google search engine. For the methodological development, the deductive method was used and the character is exploratory and a bibliographic survey is used to analyze the data. In the analysis of the data, different forms for application of the criteria and instruments for analysis of the subject to the automatic indexing between the software SISA (Semi-Automatic Indexing System) and MHTX and the Google search engine are verified. In this way, it is concluded that the use of terminological control provides better indexing both manually and automatically, such controls occur through thesauri, controlled vocabularies and ontologies. Thus the alliance between scholars from different areas of knowledge engaged in this purpose, will provide a recovery of information efficiently and effectively.

Keywords: Automatic indexing. Documentary analysis. Indexing software.

LISTA DE ILUSTRAÇÕES

LISTA DE SIGLAS

ENANCIB	ENCONTRO Nacional de pesquisa em Ciência da Informação
BTDECI	Programa de Pós Graduação da Escola de Ciências da Informação da UFMG
LDs	Linguagens documentárias
MC	Mapa conceitual
SE	Sumário expandido
TAF	Teoria da análise facetada
TIC’s	Tecnologia da Informação e Comunicação

SUMÁRIO

1 INTRODUÇÃO

O conhecimento precisa de métodos eficazes para que não haja perda de informação no infinito mundo do saber, para tanto para que tal objetivo seja alcançado muito vem sendo pesquisado a respeito de sua recuperação de modo eficaz e eficiente.

Guimarães e Dodebei (2012) foram responsáveis pela organização da obra "Desafios e perspectivas científicas para organização e representação do conhecimento na atualidade", nesta foram abordados temas inerentes a recuperação da informação, estudos abordando os rumos para representação do conhecimento de modo que este não seja perdido ou que sua recuperação tenha "ruídos" [1] e "silêncios" [2].

É nessa perspectiva de recuperação e representação do conhecimento em uma sociedade que produz um volume significativo de informação tanto técnico-científica quanto outros tipos de documentos no campo do (entretenimento, obras de artes, cartográficos, etc.), é necessário que a indexação seja um processo importante para alcançar um fim satisfatório para o usuário.

Deste modo, métodos e técnicas de indexação estão em constante estudo, sendo que com os avanços tecnológicos trouxeram a sociedade diferentes formas de comunicar-se e a internet principal meio desses avanços exerce função principal nesse processo de busca, divulgação e recuperação da informação.

Assim a indexação automática, surge como sendo alternativa para tornar esse processo mais ágil, pois sendo a indexação o ato de seleção de termos descritores significantes para posterior recuperação. Entretanto esse processo realizado manualmente torna-se encarecido e exige mais tempo dos indexadores.

Em Robredo (1982) apud Borges, Maculan e Lima (2008, p. 183) o processo de indexação automática é similar ao processo de leitura-memorização humano, sendo seu princípio geral baseado na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecida, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas.

[1] Dados irrelevantes obtidos na recuperação da informação por deficiência de programação, ou por tratamento inadequado da informação; informação parasita. (CUNHA; CAVALCANTE, 2008, p. 326).

[2] Ausência de documentos pertinentes, excluídos de lista fornecida por um sistema de informação, em decorrência de falha do próprio sistema. (CUNHA; CAVALCANTE, 2008, p. 338).

Em Robredo (1991, p.1) defende que a ideia de condensar um texto até reduzi-lo unicamente aos termos realmente significativos é, por outro lado, muito mais antiga, e todos nós a temos aplicado quando redigimos o texto de um telegrama ou de um telex.

Estudos no âmbito da indexação automática datam desde década de 50 com os trabalhos realizados por Lunh e Banxedele (1957 e 1958) respectivamente, nestes foram realizadas por diferentes formas um método que procurasse representar o conteúdo do documento através da indexação por extração, importa ressaltar que pode-se destacar a existência de dois tipos de indexação: a indexação por extração automática e a por atribuição, assim apontam Borges, Maculan e Lima (2008, p. 182).

A esses dois tipos, Lancaster (2004, p. 286; 289), as distingue da seguinte forma, a indexação por extração automática em que as palavras ou expressões do texto são extraídas e utilizadas para representar o texto como um todo, e a indexação por atribuição, esta segundo o mesmo autor defende sendo como mais difícil a ser feita por computador, sendo necessário desenvolver para cada termo a ser atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer com frequência nos documentos, o indexador humano atribuiria a esse termo, exemplo: chuva ácida incluiria expressões como chuva ácida, precipitação ácida, poluição atmosférica, dióxido de enxofre, etc.

Assim com intuito da redução tempo, custo e imparcialidade no ato de indexar é interessante o desenvolvimento de sistemas automáticos de indexação, a fim da contribuição com indexador humano, que em centros de informação onde o fluxo aumenta cada dia, torna-se humanamente inviável este processo ser realizado em tempo hábil e com qualidade.

Diante da importância do processo de indexação, o **problema** identificado para o desenvolvimento desta pesquisa está em descobrir como é desenvolvida a análise de assunto por softwares de indexação: SISA, MHTX e o buscador GOOGLE. Pois é neste processo que há escolha dos termos significativos do documento, e a seleção equivocada poderá gerar imprecisão na hora da recuperação da informação, assim a pesquisa propõe-se descrever a representação de conteúdo em softwares de indexação: SISA, MHTX e o buscador GOOGLE.

Como **objetivo geral** preocupa-se em descobrir diferentes propostas da realização de análise de conteúdo na indexação automática; e apresenta como **objetivos específicos**: a) apresentar o levantamento histórico da indexação automática e manual; b) identificar diferentes formas de consistência da indexação automática; e c) demonstrar quais são os critérios e instrumentos utilizados para representação nos *softwares* SISA, MHTX e o buscador Google.

Desse modo, este trabalho justifica-se pela necessidade de demonstrar dificuldades por parte de alguns usuários no ato da recuperação da informação, o gera desperdício de tempo, assim torna-se necessário levantar acerca do processo de tratamento temático efetuado por *software* de indexação automática.

Nesse contexto esta pesquisa adotou os procedimentos metodológicos o método dedutivo de caráter exploratório e fundamentou-se em levantamento bibliográfico.

Este trabalho encontra-se estruturado nas seguintes seções: esta sessão introdutória, seguida da sessão 2 – a qual irá apresentar o breve levantamento histórico da indexação automática; sessão 3 – irão ser apresentados os *softwares* de indexação automática; a sessão 4 irá ser descrito os procedimentos metodológicos; a sessão 5 apresenta análise dos dados levantados e conclui-se na sessão 6 com as considerações finais.

2 ESTUDOS EM INDEXAÇÃO AUTOMÁTICA: levantamento entre alguns estudiosos sobre a temática

Pode-se entender por indexação automatizada, como o ato de seleção de termos descritores de determinado documento, seja no formato impresso, som, imagem, etc. que por intermédio de linguagem computacional, irá posteriormente recuperar tal informação indexada.

Desde épocas remotas foram criados sistemas capazes de armazenar e recuperar informação e tais sistemas contam com importante processo do tratamento da informação, a indexação, esta passa por constate atualização referente a seus procedimentos de análise documental.

A indexação automática surge com a necessidade de tornar o processo de recuperação mais ágil, pois a produção de informação tanto no meio científico quanto de informações gerais cresce significativamente com o passar dos anos.

Alguns problemas ocorridos através da indexação manual também impulsionaram o estudo sobre indexação automática, dentre os quais pode-se destacar a morosidade manual desse processo, o autocusto, a ausência de subjetividade inerente à indexação (BORGES; LIMA; MACULAN, 2008, p. 182).

O estudo desta temática ganhou força com a chegada da internet, a qual caracteriza-se por aproximar seu usuário intercontinentalmente por ser uma rede formada por um número significativo de pessoas, o uso de internet móvel cresce vertiginosamente, serão cerca de 4,3 bilhões de pessoas conectadas até final de 2017, segundo relatório da ONU, assim Gil Leiva, 2007, p. 47 aborda que: “com a generalização da internet quase todos os ramos de atividades do ser humano foram florescendo na rede e consequentemente, a competência do mundo físico nos serviços, comércio ou cultura foi transladado ao âmbito digital”.

De certo o ambiente da web trouxe praticidade ao mundo, pois agora vive-se em uma sociedade em rede, as redes interativas de computadores estão crescendo exponencialmente, criando novas formas e canais de comunicação, moldando a vida e, ao mesmo tempo, sendo moldados por ela. (CASTELSS, 1996, p. 40).

Assim, o processo de indexação acompanhou a evolução tecnológica. Em Moura (2009, p. 60), Tim O’Reilly aponta a cerca do “beta eterno”, o que significa dizer que no

contexto da organização da informação os estudos tenderão a análise do fenômeno informacional através da netnografia[3] dos traços e vestígios da interação dos usuários na web.

Assim, o processo de indexação passa por imprescindível adaptação em meio ao crescente número de informação. Deste modo Evans et al. (1991, p. 108) apud Gil leiva (ano, p. 58) aborda:

A indexação automática baseada em processamento de linguagem natural (PLN) oferece alternativas atrativas para análise do documento, pois estas técnicas aliadas ao uso do tesauro para aperfeiçoar as estruturas lingüísticas, assim poderá provocar a indexação manual perfeição, consistência e precisão.

Deste modo constata-se a importância do estudo quanto às linguagens controladas com intuito de auxílio a sistemas operados por linguagem natural.

2.1 Indexação manual

Ao passar pelo processo evolutivo o homem passou por diversas transformações, na linguagem, pela forma de comunicar-se e este sentia a necessidade de deixar registrados acontecimentos corriqueiros ora presenciados.

Neste sentido Cintra, et al.(1994, p. 14), defende que "para que não haja perda do conhecimento este por sua vez fica registrado em formato de: livro imagem, foto, disco, etc., passando a se constituir um documento".

Ainda esta mesma autora (p. 15), a respeito da grande geração de conhecimento que gera o papel fundamental da área da documentação, a qual será responsável pela triagem, organização, conservação, da informação, e seu posterior acesso.

Diante destes eventos de guarda e recuperação da informação, a indexação encontra-se na segunda fase desse processo, pois essa faz o tratamento de conteúdo documental, através de Linguagens Documentárias (LDs).

A indexação é um dos principais processos na cadeia de desenvolvimento no serviço da mediação da informação, pois é através de uma indexação de qualidade que se atingirá umas das leis de Raganatah em "poupar o tempo do usuário".

[3]O termo *netnografia* é uma combinação das palavras *net* e *ethnographye* foi cunhado pelos pesquisadores norte-americanos Bishop, Star, Neumann, Ignácio, Sadunsky e Schatz em 1995 com objetivo de descrever o desafio metodológico de preservar os detalhes da observação em campo etnográfico usando o meio eletrônico para acompanhar os indivíduos.

O processo de indexação exige do indexador exaustividade e o devido conhecimento por parte deste de sua clientela, Hjorland (2001), apud Lancaster (2004, p. 10) “concorda que a indexação deve ser moldada para se ajustar às necessidades de determinada clientela”.

Bates (1998) apud Lancaster (2004, p. 10), aponta que o desafio do indexador é tentar antecipar quais os termos que as pessoas que possuem lacunas de informação de vários tipos, isto já possibilita meio caminho da satisfação do usuário.

De fato esta afirmativa se revela quando depara-se em centros de informação nos quais seus usuários detentores da “pergunta” almejando encontrar sua “resposta” de forma satisfatória, entretanto a perfeita combinação entre tais proposições será devidamente alcançada quando o indexador consegue usar os termos certos.

Entretanto, às vezes, para almejar tal objetivo é necessário o que Grogan chama de “negociação da pergunta”, pois em certos casos o cliente não consegue externar o assunto pretendido.

Todavia faz-se necessário um adequado uso dos sistemas, tanto por parte do indexador tanto por parte do usuário, quando este não “buscar” a informação pretendida por intermédio de um profissional da informação.

Segundo Salton (1983, p. 10), apud Fujita; Gil-Leiva (2004, p. 51):

> Os sistemas de recuperação de informação consistem em conjunto de itens de informação (documentos), um conjunto de petições de informação (perguntas) e algum mecanismo (comparação) para determinar quais documentos cumprem com as petições requeridas ao sistema.

Deste modo, nota-se o compromisso do indexador na escolha adequada do termo a ser indexado, “é função do indexador identificar no conteúdo dos documentos quais são os possíveis concretos e processos inerentes ao assunto que está sendo indexado, que irão servir à indexação”. (GUIMARÃES; SALES, 2016, p. 55).

Diante disso é correto afirmar que uma indexação adequada irá depender do alcance do conhecimento do indexador, para que isto ocorra é necessário por parte desse profissional o constante aperfeiçoamento, principalmente na instituição onde atua.

2.2 Recuperação da informação e a indexação

A recuperação vai estar completamente ligada ao tratamento da informação, ou seja, inclui a representação temática dos documentos. E deste modo, destaca-se que a

compatibilização de linguagens utilizadas por usuários entre as instituições de informação é fundamental para elaboração de estratégias de busca adequadas para a recuperação da informação.

A informação é um meio primordial para geração de conhecimento, entretanto o trajeto de sua disponibilização e acesso depende de tratamento temático adequado; em Cintra; et al. (1994, p. 14), apontam que "(...) a informação está diretamente ligada ao conhecimento e ao desenvolvimento de cada uma das áreas do saber, já que todo conhecimento começa por algum tipo de informação e se constitui em informação".

Assim pode-se dizer que a utilização de novas tecnologias, faz-se necessário para melhor atender a incessante busca de recuperação da informação, igualmente, esta recuperação com eficiência irá exigir substancial esforços no campo do tratamento temático.

No que tange as estratégias de busca da informação depende do alcance do conhecimento do assunto procurado, por parte do usuário, para tanto também, para que as estas funcionem é necessário que o indexador possua tal conhecimento.

Para acesso a informação, com qualidade, ou seja, sem que haja desperdício de tempo e que a recuperação da informação pretendida seja de qualidade, faz-se necessário o conhecimento e uso das estratégias de busca.

A recuperação da informação, como resume Saracevic (MOSTAFA, 2010, p. 162 apud BIOLCHINI; GIORDANO 2012, p. 127), "trata sobre o que pode ser feito para acessar, de maneira rápida e efetiva, a informação inserida em determinado repositório".

No ambiente informacional existem diferentes maneiras para recuperar informação, pois para cada sistema de informação existe uma estratégia de busca. Para cada instituição há um sistema e processos a serem seguidos.

2.2.3 Avaliação da indexação: implicância na geração de qualidade desse processo

A indexação é importante processo para posterior recuperação da informação, assim a Biblioteconomia estuda possibilidades de elaboração de linguagem controlada para melhor indexar o documento. (BORGES; LIMA; MACULAN, 2008, p. 187).

Nesse sentido quando não há indexação de qualidade não poderá ter-se recuperação da informação de maneira eficiente, assim e imprescindível fazer avaliação desse importante processo.

Assim em Fujita e Gil Leiva (2014, p. 51), abordam a questão da linguagem como instrumento para a indexação, pois cumpre a função do controle de vocabulário e realiza a mediação na recuperação por assuntos pelo usuário.

A linguagem controlada servirá tanto para o indexador no ato do tratamento, quanto para o usuário o auxiliando na busca.

Ainda os mesmos autores argumentam em relação ao processo de avaliação da indexação, na década de 1950 houve iniciativas para avaliação das linguagens documentárias e na década de 1960 teve a ocorrência para avaliação da indexação.

Assim estes abordam a questão da qualidade da indexação, antes que o documento seja disponibilizado em base de dados, esta avaliação dá-se como assim dispõem-se:

Quadro 1: Antes que o documento ingresse na base de dados

Métodos	**Avaliação**
Indexação realizada por especialistas	infração da política de indexação; especificidade relativa à exaustividade; falha no uso das linguagens de indexação (combinação incorreta de cabeçalho e subcabeçalho); incorreto uso dos termos, provocando erro de especificidade; designação inadequada de termos por erro ou por falta de conhecimento especializado sobre o tema tratado e omissão de um termo importante.
Indexação avaliada mediante simulação da realidade	Nesta os autores além da busca dos erros anteriores que é o que indica Lancaster (2004, p. 87), eles propõe uma metodologia mais rigorosa diante de uma simulação de uma avaliação real: esta baseia em seleção de documentos antes que chegue ao indexador; para cada documento, elaboração de três necessidades de informação para as quais o item seja uma resposta satisfatória; elaboração de estratégia de busca para cada uma das necessidades de informação por parte de profissionais especialistas em busca de informação; introdução dos documentos novamente ao fluxo de entrada para os indexadores fazerem seu trabalho de forma rotineira; comparar a indexação feita com as estratégias de busca para determinar se os termos designados aos documentos obteriam sua recuperação ou não de documentos relevantes do sistema.

Fonte: Fujita; Gil Leiva (2014, p. 51).

O quadro 2 demonstrará a avaliação após a entrada dos documentos na base de dados, estes procedimentos são indicados por Gil Leiva, sendo estas tão necessárias quanto a avaliação efetuada antes que o documento entre na base de dados.

Quadro 2: Após que o documento ingresse na base de dados

Avaliação	**Dados analisados**
Intrínseca qualitativa	Nesta são analisados elementos que proporcionam a qualidade de indexação, devem ser realizada por dois especialistas, número mínimo, os devem conhecer a política de indexação, a linguagem de indexação e as características dos usuários do sistema de informação, tais procedimentos ocorrem das seguintes formas: seleção ao acaso de um número significativo de registros do catálogo ou da base de dados; a reindexação dos documentos tomando como referência o texto completo do mesmo; comparação da indexação original com a dos especialistas.
Intrínseca quantitativa	É a indexação de um conjunto de documentos repetindo, na medida do possível, o entorno ao qual foi produzida a primeira indexação (indexação, política de indexação, linguagem de indexação, condições de trabalho, usuários potenciais, etc.). Esta avaliação é de grande utilidade para avaliação de periódica, na mesma unidade de informação através de ensaios intraconsistência, ou seja o indexador retorna ao documento indexado para comprovar se houve variação em relação a primeira indexação.
Extrínseca mediante a recuperação	Compara dois indexadores do mesmo sistema (intraconsistência) na indexação de duas unidades de indexação diferentes, isso se daria comparando a indexação manual de uma mesma base de dados ou biblioteca com a indexação automática. As adoções desses procedimentos se dão da seguinte forma: construção de base de dados com o mínimo de 100 registros; atribuição a cada um dos documentos incluídos sua relevância temática; seleção de um conjuntos de necessidades de informação reais que tenham relação com o conteúdo das bases de dados; construção de interrogações para cada petição de acordo com parâmetros próprios do sistema e iniciar as buscas; anotar o número de cada um dos documentos recuperados para cada uma das buscas executadas; encontrar os índices de exaustividade e precisão para cada uma das buscas; encontrar a média de exaustividade e de precisão para cada uma das bases de dados.

Fonte: Elaborado pela autora (2018).

Assim, pode-se dizer que tais avaliações para o processo de indexação, podem propiciar garantir a eficiência de análise documentária.

3 SOFTWARES DE INDEXAÇÃO AUTOMÁTICA

Com o cenário em ascensão da entrada tecnológica na sociedade, viu-se a necessidade de oferecer aos usuários de centros de informações e bibliotecas, o desenvolvimento de ferramentas que proporcionassem praticidade ao serviço informacional.

Neste âmbito de transformações tecnológicas Heeman (1994) apud Almeida e Cortê (2000, p. 11), aborda a questão da fluência de informação, que em uma sociedade informatizada, ocasionou o descontrole em acervos e catálogos, para integrar-se aos arquivos de computadores, assim formando a grande rede de troca de informação entre usuários, independente de proximidade.

Assim, identifica-se que os softwares desenvolvidos para computadores de grande porte exigia um amplo aparato computacional, como equipes especializadas, ambiente apropriado, além disso, tais programas não possuíam capacidade para alimentação em tempo real, desta forma colocava bibliotecários e usuários totalmente dependes da tecnologia gerando dessa forma pouca agilidade na prestação de serviços, ocasionando a sucumbência da informação à tecnologia. (ALMEIDA; CORTÊ, 2000, p. 12).

Almeida e Cortê (2000, p. 13), idenficaram aspectos a serem adotados antes de qualquer iniciativa de informatização tanto em bibliotecas quanto em centros de documentação: cultura, missão, objetivos e programas de trabalho da organização; características essenciais da biblioteca (abrangência temática, serviços e produtos oferecidos); interesse e necessidade de informação dos usuários; plataforma tecnológica existente na instituição (software e hardware), bem como sua capacidade de atualização e ampliação; recursos humanos disponíveis.

Levando em consideração a adoção de tais aspectos mencionados no parágrafo acima, estes tornam-se primordiais, pois através deles são estabelecidos os requisitos particulares às instituições, ou seja, a informatização ocorre perante as necessidades pré-estabelecidas ora identificadas. (ALMEIDA; CORTÊ, 2000, p. 13).

Assim, tratando-se de softwares de indexação, estes apresentam-se em constante processo de estudos, pois com intuito de automatizar o tratamento temático verificou-se a necessidade de ampliação neste âmbito de pesquisa. Tal interesse nesta área parte do princípio que estes possibilitarão disponibilização da informação com maior agilidade, qualidade e precisão.

Para tanto, a obtenção de sucesso no uso de softwares em sistemas de informação, estes devem operar por intermédio de diferentes linguagens de indexação, a escolha adequada

desta oferecerá alcance da consistência no processo de análise de termos. De acordo com Kobashi, Maimone e Mota (1988, p. 77), as linguagens organizadas como instrumentos de indexação e recuperação, recebem diferentes denominações de acordo com suas características, demonstrados no quadro 3.

Quadro 3: Instrumentos para indexação

LINGUAGENS/INSTRUMENTOS	CONCEITOS
Tesauros	uma linguagem especializada, normalizada, pós-coordenada, usadas com fins documentários, onde os elementos lingüísticos que o compõem, encontram-se relacionados entre si sintática e semanticamente
Vocabulários controlados	Dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem dos sistemas (linguagem de documentação, linguagem de informação) mais restrita
Lista de cabeçalhos de assunto	Palavra ou grupo de palavra que expressam o conteúdo de um documento
Taxonomias	Responsável pelo Controle terminológico
Ontologias	É uma especificação formal e explicita de uma conceitualização compartilhada

Fonte: Elaborado pela autora (2018).

No entanto para este estudo, além das denominações mencionadas por Kobashi, Maimone e Mota (1988, p. 77), faz-se necessário explanar das linguagens dos Mapas Conceituais e as Folkosomias, pois participam de forma direta ou indireta da construção do Sistema MHTX e o buscador Google, respectivamente.

Estes instrumentos são abordados por Moura (2009, p. 62), pertencentes às chamadas ferramentas ontológicas, estas são:

Utilizadas para designar estudos conceituais específicos que visam caracterizar dada área de conhecimento a partir de mapeamento de categorias mais gerais, através destes são desenvolvidos programas computacionais envolvendo inteligência artificial.

3.1 Uso das Tecnologias de Informação e Comunicação (TIC's) no processo de indexação

A sociedade é marcada por mudanças, desde a invenção da escrita até os dias atuais. Deste modo, Campos e Marcondes (2008, p. 108) abordam a respeito da relação do homem e sua necessidade de guarda e disseminação da informação pelo crescente volume de

informação, estes dedicaram esforços para criação de instituições especiais para manutenção e expansão do Conhecimento e da Cultura.

Nessa perspectiva de evolução, Arraes, et al. (2007, p. 5), aborda a questão do surgimento da documentação após a segunda guerra mundial, a qual foi demonstrada em estudos feitos a respeito da Ciência da Informação.

Em um momento pós-guerra, sentiu-se a necessidade de encontrar métodos organizacionais para volume de informação produzido, assim definiu-se o problema de gerenciamento com a proposta de uma máquina chamada Memex, esta era dotada com as mais modernas tecnologias de informação e comunicação, contudo o Memex não fora desenvolvido porem suas idéias serviram como bases inspiradoras em outros estudos. (ARRAES, et al., 2007, p. 5).

Nessa perspectiva, Castells (2000, p. 50) aborda a questão das revoluções tecnológicas atual, esta "originou-se e difundiu-se, em um período histórico da reestruturação global do capitalismo, para o qual foi uma ferramenta básica.

Isto demonstra que o uso das TIC's mudou a maneira de disseminar informação, a internet interfere na sociedade de forma significativa, já que em dias de desenvolvimento de aparelhos móveis com altas tecnologias, em que a informação está a um simples clique.

Tais mudanças advindas com uso das novas tecnologias de comunicação e informação, interferem no convívio social, o uso de informação via web tende crescer ao passar dos anos e gerar a cada dia transformações significativas o comportamento da busca por informação.

O acesso as TIC's com aporte da internet, proporcionaram a sociedade maior acessibilidade as informações, estas, disponibilizada com maior agilidade a seus usuários. É neste ambiente que surge a Web ou World Wild Web, surgida nos anos 90, com a intenção de implantar o hipertexto idealizado por Nelson & Douglas Engelbart em 1962, buscava oferecer interfaces mais amigáveis e intuitivas para organização e acesso ao crescente repositório de documentos que se tornava a Internet. (ALVARENGA; SOUZA, 2004, p. 182).

Neste âmbito percebe-se a necessidade de criação de métodos eficazes para recuperar a informação, já que esta já encontrava-se inserida no mundo caótico da rede mundial.

Foi nesse meio que as tecnologias aliadas com as áreas da Ciência da informação, tornaram acessíveis o conhecimento humano. (WERSIG, 1993 apud AUVARENGA; SOUZA 2004, p. 183).

Neste sentido vê-se que tais meios encontrados para tornar a intercomunicação mais eficiente trouxeram contribuição para disseminar e recuperar informação, isto deve-se ao surgimento de métodos capazes de proporcionar tal evento.

Assim a “explosão informacional” assim chamada por Marcondes (2001), deparou-se com a necessidade de formular um formato textual estruturado a fim de proporcionar reconhecimento por parte de programas, neste caso a web semântica. (CAMPOS E MARCONDES, 2008, p. 110).

Desta forma, as linguagens documentarias surgem para o auxiliar a organização informacional, através de suas técnicas propuseram estruturação e melhor sistematização para ambiente Web, através das ontologias.

Assim, Campos e Marcondes (2008, p. 110) abordam “o uso das ontologias como instrumentos de representação de conhecimento, surge no âmbito da Inteligência artificial na década de 90”.

Neste contexto serão demonstradas nas sessões seqüentes as técnicas de análise de assunto efetuada pelos *softwares* SISA e MHTX e o buscador Google.

3.1.2 Sistema de indizaciónsemi-automático (SISA)

O Sistema de Indización Semi-automático (SISA) é um sistema de indexação semi-automático, idealizado por Isidoro Gil Leiva da Universidade de Murcia na Espanha em 1997, inicialmente para indexar artigos de Biblioteconomia e Documentação.

A operacionalização deste ocorre por intermédio de vocabulário controlado, é composto por quase 3000 termos, dos quais 2200 são descritores e 800 não descritores. Ainda possui 273 palavras vazias, as quais não transmitem tema ou assunto com relevância, assim dispõe Gil Leiva (2003, p. 2).

Para tanto em Fujita, Gil Leiva e Narukawa0 (2009, p. 101), demonstram que este Gil Leiva em sua publicação (1999, p. 57; 2008, p. 320) identificou vinte expressões diferentes, as quais tais autores as delimitam em três conceitos: quanto a indexação assistida por computador – na qual os sistemas realizam o armazenamento dos termos de indexação identificados por um profissional; indexação semi-automática – nesta os programas realizam análise dos documentos de modo automático e quando necessário tais termos são validados por profissionais e a indexação automática,a objeto de estudos, em que os programas realizam análise documental sem ocorrência de validação por um profissional.

Ainda em Gil Leiva (2003), aborda quanto ao documento a ser indexado, este deve esta no formato txt[4] e com um conjunto de etiquetas que delimitam o início e o fim do título do artigo, o resumo e texto completo.

Da realização do processo da análise de assunto: o algoritmo SISA procura no artigo a ser indexado, por termos que serão possíveis de tornarem descritores, nestes haverá uma comparação com os descritores do vocabulário controlado, identifica a localização de onde ele foi retirado (título, resumo ou texto).

Nestes módulos ora apresentados é a divisão que o software realiza no seu processamento terminológico, no pré-processamento iram ocorrer às marcações do documento onde as partes deste são sinalizadas com marcadores #CTI# (começo do título), #FTI# (fim do título), #CR# (começo do resumo), #FR# (fim do resumo), #CTE# (começo do texto) e #FTE# (fim do texto).

No pré-processamento, ainda ocorre também a horizontalização das frases e orações localizadas entre sinais de pontuação. (FUJITA; GIL LEIVA; NARUKAUWA, 2009, p. 106).

Ainda os mesmos autores (2009, p.106), abordam os dois últimos módulos nos quais ocorrem a análise de conteúdo temático – segundo módulo -, como já descrita acima; e a validação dos termos no terceiro módulo, que consiste na aplicação de um critério de avaliação dos termos para que o sistema possa selecionar os termos de indexação que representarão o conteúdo do documento.

Em dissertação para título de mestre, Cristina Miyuki Narukawa intitulada "Estudo de vocabulário controlado na indexação automática: aplicação do processo de indexação do Sistema de Indización Semi-automatica (SISA)", apresentada em 2011, a autora irá apresentar resultados de alguns experimentos efetuados neste no qual foi feito comparações através da utilização do vocabulário ThesAgro empregado no SISA com a indexação manual da BINAGRI.

Nesta pesquisa a autora aponta que: "os problemas na indexação automática do sistema estão relacionados a fatores de natureza lingüística, i.e., de tratamento morfológico, sintático, semântico, assim como também a natureza metodológica do sistema e aplicação do vocabulário controlado por simples processo de cotejamento entre os termos do documento e do vocabulário".

[4]É uma espécie de ficheiro informático que é estruturado como uma sequência de linhas existe dentro de um computador do sistema de arquivos.

3.1.3 MHTX – Mapa Hipertextual

O *software* MHT baseia-se na adoção de mapas conceituais como instrumento no que concerne à linguagem de indexação. Estes baseiam-se na teoria de aprendizagem significativa, de David Ausubel na década de 60, foram introduzidos na educação por Joseph Novak na década de 70, com a intenção de apoiar a compreensão do processo de organização do conhecimento. (MOURA, 2009, p. 62).

Resultado de uma tese de doutorado de Gercina Ângela Borem Lima; é ampliação do protótipo mapa Hipertextual – MHTX, caracteriza-se como um modelo com objetivo de organização da informação hipertextual de documentos, teses e dissertações foram inseridas no formato de texto completo digitalizados, este programa encontra-se implantado na Biblioteca Digital do Programa de Pós-Graduação da Escola de Ciências da Informação da UFMG (BTDECI).

A pesquisa deste programa tem a pretensão de simplificar os processos de organização, acesso e recuperação da informação, ora contidas nas teses e dissertações.

Este estudo aborda mapa conceitual como uma ferramenta de organização do conhecimento, capaz de representar ideia ou conceitos na forma de um diagrama hierárquico escrito ou gráfico capaz de indicar as inter-relações entre os conceitos, procurando refletir a organização da estrutura cognitiva do indivíduo sobre um dado assunto. (LIMA, 2004, 92).

Para sua estruturação, adotou-se: identificação do documento básico de trabalho (tese); leitura da tese; Análise facetada do assunto da tese: a seleção de termos relevantes e categorias (facetas); o reconhecimento das subfacetas; a ordenação das facetas, subfacetas e focos a serem apresentados no mapa conceitual e finalmente a organização de todos os termos e suas relações; criação do mapa conceitual (MC) com seus links e suas relações; estruturação do sumário expandido (SE) e criação dos links do sumário para o texto.

Foram utilizadas duas técnicas da análise facetada para realização da análise de assunto, baseando-se no processo da análise e a classificação dos conceitos em categoria, tais procedimentos efetuam a identificação dos termos relevantes e representação das características. A autora ressalta que o processo referente à síntese, que ocorre quando cada conceito pertencente a essas categorias é combinado com outro para expressar um assunto composto, entretanto este procedimento não foi incluído na tese.

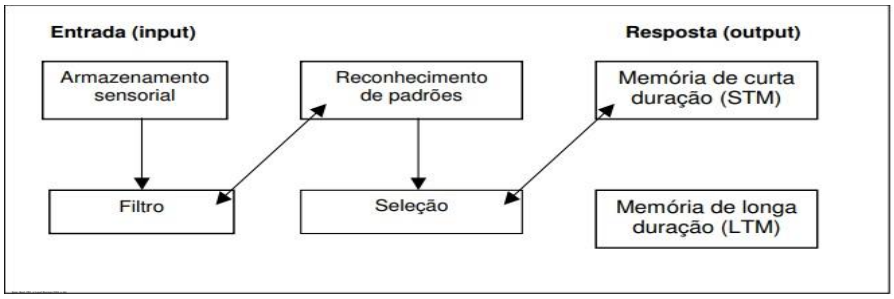
Na formulação das categorias foram utilizados os princípios normativos do plano das idéias constantes abordada no estudo “Modelo Simplificado para análise de facetas:

Ranghanathan 101" – SPITERI (1998); neste ocorre o processo de definição do assunto; seleção das características que constituem o assunto, seleção de um modelo para o mapeamento da informação dos conceitos, agrupamento e divisão dos conceitos conforme suas características comuns e diferentes e o arranjo de grupos e subgrupos.

Entende-se por Teoria da Análise facetada (TAF), um modelo com natureza metodológica dedutiva, onde considera-se primeiro o domínio ou contexto para posterior seleção dos termos representativos de uma área, possuindo mecanismos de representação para trabalhar-se com níveis conceituais na formação de categorias, nesta os conceitos são ordenados para formarem classes de conceitos.

Neste contexto Lima aborda quanto à relação da fase de indexação, pelo fato da indexação ser um processo de categorização, na qual poderia ocorrer harmonização entre as habilidades intelectuais envolvidas no processo com adoção da simulação dos processos cognitivos sensoriais.

Figura 1: Estágio do modelo de processamento da informação



Fonte: Reed (1992, p. 5) apud Lima (2014, p. 66)

Nesta identificação, Gercina Lima explica utilizando o estágio de processamento adotado por Reed no qual esta constatou que após as informações e os padrões de linguagem serem reconhecidos e filtrados, o processo de análise concentra-se na utilização do usuário na memória curta duração (STM) e a memória longa duração (LTM), no LTM será onde a informação ficará armazenada por um longo período, após esta passar pelo STM.

Haja vista o exposto demonstra-se no quadro 5, o processo da análise documental.

3.1.4 Google

Tratando-se deste buscador depara-se com um "mundo" de informação, este desde seu surgimento provocou grandes mudanças no ato de comunica-se, junto a grande "rede" internet, tem suas histórias entrelaçadas.

A busca por visibilidade na internet é o que empresas e pessoas buscam cada dia mais, cada um por interesses inerentes; empresas que conseguem aumentar seus rendimentos devido

a seus anúncios ora mostrados em diferentes sites; e pessoas em busca de amizades ou relacionamentos, estas através de redes sociais buscam demonstrar padrão de vida, muitas vezes diferente do real.

Deste modo assim como o manual deste buscador (2008), revela as seguintes informações: a evolução da internet, o Google evoluiu igualmente em relação à pesquisa na web, pois criam-se ferramentas constantemente com intuito de tornar a recuperação da informação mais ágil.Ferramentas para ajudar os webmasters[5], a maximizar a visibilidade de seu conteúdo, bem como controlar a forma como suas páginas da web são indexadas.

Assim como o processo de indexação o Google utiliza-se de um grupo de computadores, o *Googleboot*, estes rastreiam as páginas na web, a este processo designa-se de algorítmico, neste os programas de computadores determinam quais web sites devem ser rastreados, com que frequência e quantas páginas de cada web site devem ser analisadas, para realização desse processo a empresa não recebe pagamentos, e ainda matem a pesquisa independente do serviço de publicidade no Google *AdWords*.

Às vezes pode ocorrer dos *webmasters* que os web sites não aparecem nos resultados da pesquisa, este problema poderá está relacionado a capacidade de indexação.

[5] Profissional que gerencia tarefas tanto de webdesigner (elaboração de um projeto estético funcional de um *web site*) quanto de um web developer(parte da programação, como sistemas de login, cadastro, área administrativa).

4 METODOLOGIA

Para realização de pesquisa científica faz-se necessário adotar procedimentos inerentes para obtenção de resultados satisfatórios para conclusão do levantamento e reconhecimento no meio científico. Para Gil (2008, p. 8) defende que "a ciência tem como objetivo fundamental chegar à veracidade dos fatos. Neste sentido não se distingue de outras formas de conhecimento. O que torna, porém, o conhecimento científico distinto dos demais é que tem como característica fundamental a sua verificabilidade".

Assim, o corpus deste trabalho foram utilizados capítulos de livros, artigos científicos publicados em periódicos da área da Ciência da Informação, teses. Apresenta quanto aos procedimentos metodológicos o método é dedutivo de caráter exploratório, como procedimentos técnicos foram utilizado o levantamento bibliográfico e elaboração de uma análise simples para demonstração das técnicas análise de assunto nos softwares SISA, MHTX e o buscador Google.

A escolha do método dedutivo justifica-se pelo fato de esta pesquisa abordar uma temática na qual os estudos estão em constante atualização, ou seja, não apresentam estudos conclusivos já que tal método parte da observação de fatos ou fenômenos cujas causas desejam-se conhecer. (FREITAS; PRODANOVE, 2013, p. 29).

Apresenta caráter exploratório, pois pretende mostrar conhecimento acerca de estudos relacionados à análise de assunto na indexação automática, como procedimentos técnicos utilizaram-se o levantamento bibliográfico em livros, artigos e teses abordando assuntos da indexação manual e automática; tratamento da informação; recuperação da informação; linguagens documentárias; dos sistemas ora analisados – SISA, MHTX e o buscador Google.

Assim, destaca-se quanto à importância do uso de técnicas de pesquisa, Lakatos e Marconi (2003, p. 174) aborda sendo "um conjunto de preceitos ou processos de que se serve uma ciência ou arte, é a habilidade para usar esses preceitos ou normas, a parte prática. Toda ciência utiliza inúmeras técnicas na elaboração de seus propósitos".

Para análise dos dados foi utilizada a pesquisa bibliográfica, pois esta abrange toda a bibliografia já tornada pública em relação ao tema de estudo, desde publicações avulsas, boletins, jornais, revistas, livros, pesquisas, monografias, teses, material cartográfico, ect., até meios de comunicação orais Lakatos e Marconi (2003, p. 183).

Coleta de dados, inicialmente irá ser feita análise em trabalhos desenvolvidos sobre os sistemas SISA e o MHTX, os dados coletados nestes será a respeito da análise de assunto

neles efetuadas; e quanto ao buscador Google foi feita análise em publicações da própria empresa.

5 ANÁLISE E DISCUSSÃO DOS DADOS

Esta pesquisa propôs-se na identificação de diferentes propostas de análise de conteúdo efetuada na indexação automática, todavia adotou-se como referencial para apresentação desta identificação os *softwares* SISA, MHTX e o buscador Google.

Assim este estudo se fundamentou no levantamento bibliográfico, pois neste pode-se identificar na literatura científica acerca do assunto ora abordado, com devidas limitações, ampla abrangência de informações através de coleta de dados através de fichamento bibliográfico.

Na literatura consultada verificou-se amplos estudos para contribuição de diferentes técnicas, específicas a indexação automática no processo da representação de conteúdo temático, identifica-se duas formas de fazer análise documental, indexação automática e manual, nesse ambiente Lima (2008, p. 182) aborda que "a etapa da análise conceitual, determina do que trata um documento, isto é, qual seu assunto".

Nesse contexto, detectou-se a tendência em indexação por *softwares* que fazem uso de linguagens documentárias por meio automático.

Quadro 5 – Processos para automatização

	SISA	MHTX	GOOGLE
Processo de Automatização	Semi-automática	Automática	Automática
Critérios	Métodos estatísticos	Análise facetada	Ferramentas para webmasters
Instrumestos	Vocabulário controlado	Mapas conceituais	Googlebot

Fonte: Elaborada pela autora, 2018.

Desse modo quadro 5 demonstra diferentes critérios e instrumentos utilizados na representação de assunto proposta pelos softwares SISA E MHTX e o buscador Google.

Através do levantamento bibliográfico foi identificado em capítulos de livros, artigos científicos que o SISA elabora análise de assunto através de métodos estatísticos.

5.1 Critérios utilizados pelos softwares SISA, MHTX e o buscador Google

A indexação automática é um processo que depende de escolha de métodos pré-estabelecidos para extração do conteúdo do documento, assim como já mencionado no primeiro capítulo os precursores da indexação automática, Lunh (1957) e Baxendale (1958) utilizaram a freqüência de palavras para realizar tal procedimento.

Lancaster (2004, p. 286) defende a possibilidade de escrever programas simples, os quais efetuaram a contagem das palavras, estas já comparadas com uma lista de palavras vazias como: artigos, preposições, conjunções e outras iguais, e posteriormente colocá-las seguindo a freqüência de sua ocorrência.

Nesse contexto o Sistema de Indización Semi-Automático (SISA), emprega metodologia de comparação entre o documento e um vocabulário controlado utilizando para isto critérios estatísticos para contagem de palavras significativas, Gil Leiva faz um longo levantamento teórico para chegar a tal aplicação metodológica, investigações a cerca da validação de termos através dos títulos e resumos de artigos científicos, neste trabalho o autor identificou tanto teóricos em defesa quanto os não adeptos do uso de títulos e resumos para identificação de termos.

Diante destes fatos, pode-se constatar entre os softwares SISA (Sistema de Indización Semi-Automático) e MHTX e o buscador Google apresentam diferentes formas para aplicação de seus critérios para análise de assunto à indexação automática:

- O Sistema de Indización Semi-Automático (SISA), aplica seus critérios a partir da identificação de termos, os quais podem aparecer nas fontes – título, resumo e texto – assim em Fujita, Gil Leiva e Narukawa (2009, p. 106) é demonstrado que um termo autorizado pode ser indexado, quando aparecer no título e resumo, título e texto e resumo e texto, é dado também o conhecimento das etapas operacionalizadas pelo algoritmo as quais ocorrem em três fases (módulos), como já demonstrado no capítulo 3.2, no módulo 2 que é da análise do assunto onde ocorre a seleção dos termos preferidos também aparecem na linguagem documentária; os termos não preferidos que por serem sinônimos remetem para os termos preferidos e no módulo 3 ocorre aplicação de um critério de avaliação dos termos, pois com a ausência desta o sistema selecionaria todos os termos da linguagem documentária que coincidissem.

- O Modelo Hipertextual para Organização de Documentos (MHTX): para a seleção de termos utiliza os critérios da análise facetada, a qual apresenta-se em dois processos – o da análise no qual irá ocorrer a identificação dos termos relevantes e o da classificação dos

conceitos em categorias. Para a facetação do assunto foi feito o mapeamento cuidadoso do documento a fim de identificação correta de sua temática e determinação dos assuntos básicos e isolados na análise conceitual, foi utilizado o texto como o todo para indexação.

- O site de busca Google: apresenta facilitadores para editores da web, tais editores podem contar com a maximização de aparição na web cumprindo critérios pré-estabelecidos para os chamados webmasters, tais critérios estão ligados a adequada otimização dos motores de busca, o cumprimento destes irá propiciar melhor capacidade do Google de rastrear, indexar e classificar o conteúdo dos sites. Abaixo será demonstrado cinco dos critérios pertencentes ao otimizador SEO, os quais estritamente ligados a uma indexação de qualidade pelo Google.

1) Usar nas páginas o title tag, a fim da identificação do título, este deverá ser único para cada página, este serve para indicar qual tópico de uma determinada página aos utilizadores e aos motores de busca, deverá ser colocado entre o <head> tag no documento HTML, a figura 2 demonstra o título da página de cartões de baseball, onde aparece a indicação da empresa e as principais áreas em foco.

Figura 2: Demonstração da title tag

```
<html>
<head>
<title>Brandon's Baseball Cards - Buy Cards, Baseball News, Card Prices</title>
<meta name="description" content="Brandon's Baseball Cards provides a
large selection of vintage and modern baseball cards for sale. We also offer
daily baseball news and events in">
</head>
<body>
```

Fonte: Guia webnasters Google (2011, p. 4)

2) Utilizar metas tags descritivas, pois irá complementar a informação do title tag, a figura 3 demonstra que a meta tag descritiva oferece uma descrição geral do conteúdo.

Figura 3: Demonstração da meta tag

```
<html>
<head>
<title>Brandon's Baseball Cards - Buy Cards, Baseball News, Card Prices</title>
<meta name="description" content="Brandon's Baseball Cards provides a
large selection of vintage and modern baseball cards for sale. We also offer
daily baseball news and events in">
</head>
<body>
```

Fonte: Guia webnasters Google (2011, p. 6)

3) Estrutura dos URLs simplificadas irá propiciar a recuperação da informação com maior eficácia, evitando-se caracteres estranhos, pois estes podem dificultar o acesso de possíveis usuários, na figura 4 identifica-se quando é optado pela utilização de palavras, as quais podem informar tanto aos motores de busca quanto a seus usuários a respeito do conteúdo da página hiperligada.

Figura 4: Demonstração do uso de palavras em vez de caracteres nas URLs

Fonte: Guia webmasters Google (2011, p. 8)

4) Conteúdos e serviços de qualidade, antecipar o conhecimento do usuário sobre os conteúdos e tópicos, utilizando-se de palavras que provavelmente futuros utilizadores irão buscar.

5) Texto âncora de qualidade, localizado na zona URL permite a usuários e aos motores de busca a visibilidade do conteúdo da página

5.2 Dos instrumentos usados pelo SISA (Sistema de Indización Semi-Automático), MHTX (Modelo Hipertextual para Organização de Documentos)e o buscador Google.

Para minimizar equívocos na hora da indexação, profissionais da Ciência da Informação contam com importante acessório na hora de realizar análise do documento, estes identificados como instrumentos podem propiciar a geração do processo de indexação de qualidade, Bruzinga (ano, p.63) aborda a cerca destes como:

(...) chamados de linguagens de indexação ou vocabulário controlado, são criados através do processo de indexação e categorização da informação. Por organizarem a informação, podem ser considerados o coração de um sistema de informação.

Assim nesta análise demonstrará os instrumentos adotados pelos softwares SISA, MHTX e o buscador Google.

- O SISA (Sistema de Indización Semi-Automático): adota o uso de vocabulário controlado, desta forma irá identificar os termos pré-estabelecidos para futura comparação com a lista de termos preferidos, os quais serão prováveis a serem indexados;
- O Modelo Hipertextual para Organização de Documentos (MHTX): este protótipo adota dois instrumentos para organização e representação da informação, mapa conceitual (MC) apresenta como objetivo, assim aborda Lima (2004, p. 121) possibilitar visão geral da estrutura semântica do texto, facilitar a navegação semântica em contexto, por intermédio de seções e subseções, digitalizadas, inseridas na base hipertextual. Sua estrutura apresenta conceitos relevantes da tese tratada e organizada de acordo com a estrutura facetada, apresentado as relações semânticas, hierárquicas e associativas.
- O buscador Google: sua operacionalização conta *Googlebot* como instrumento para realizar o rastreamento dos conteúdos existentes na web, assim ele reúne as páginas em um índice, organizados pelas palavras encontradas e suas devidas localizações e as informações encontradas no conteúdo irão passar pelo processamento e postas nos principais atributos e *tags* de conteúdo, como *tags title* e atributo ALT.

6 CONSIDERAÇÕES FINAIS

O presente trabalho apresenta como objetivo em descobrir diferentes técnicas aplicadas a análise de assunto na indexação automática, assim este fundamentou-se na identificação de tais técnicas nos softwares SISA (Sistema de Indización Semi-Automático), Modelo Hipertextual para Organização de Documentos (MHTX) e o buscador Google.

Nesse sentido levantou-se os dados através de pesquisa bibliográfica com apresentação em estudos desenvolvidos a cerca da indexação automática, apresentando como objetivo geral diferentes propostas da análise de assunto neste tipo de processo e como objetivos específicos propôs-se apresentar a indexação automática e manual;

Quanto aos resultados encontrados verificou-se que o uso de controle terminológico propicia melhor indexação tanto de forma manual quanto a automática, tais controles podem dar-se por intermédio dos tesauros, vocabulários e ontologias. Por outro lado tem-se como aspecto negativo questões de união de estudos interdisciplinares, ou seja, estudiosos de diferentes áreas do conhecimento empenhado em um propósito, fornecer uma recuperação da informação com eficiência e eficácia.

Ainda nota-se também, que os mapas conceituais possuem importantes funcionalidades como: facilitação de navegação em hiperdocumentos e auxiliar no processo de recuperação do conteúdo semântico em textos completos em bibliotecas digitais.

Assim baseados nos levantamentos ora desenvolvidos, no qual buscou-se demonstrar técnicas do tratamento temático, através da análise dos critérios e instrumentos adotados pelos softwares e o Google foi possível detectar que o auxílio de metodologias de controle terminológico, as seguintes conclusões:

1 O SISA que apresenta o uso de critérios de frequência, para propor os termos de indexação aliado ao uso do vocabulário controlado como instrumento, isto fornece qualidade no funcionamento do deste.

2 O MHTX , um modelo hipertextual de documentos, dotado de estrutura hipertextual, adota o uso de mapas conceituais (MC) como instrumento os quais reproduzem a estrutura organizadora das páginas HTML que constituem o hipertexto do site, ainda funciona como um menu, assim cada elemento do mapa é um link que dá acesso ao documento, quanto ao critério adotado para representar o assunto do documento aplica a análise facetada, esta opera em dois processos – o da análise onde identifica os termos relevantes e o da classificação dos conceitos em categorias, fundamentada no cognitivismo já que para haver um processo de seleção de termos que irá propiciar posterior recuperação, há de se convir o envolvimento de

conhecimentos por parte do indexador, a este respeito uma das definições da categorização Lima (2004, p.) apoia-se na conceituação de Piedade (1983) a qual defende que este é um processo mental habitual do homem, já que vive-se de forma automática classificando ideias e coisas, a fim de conhecer e compreender.

- O Google: por ser um buscador de ampla abrangência, aprimora cada tempo que passa suas ferramentas e com isso oferecendo de forma eficaz, na medida do possível, a recuperação de informação. Trabalha com robôs de busca que são seus indexadores automáticos, pois rastreiam a web constantemente, tendo o Googlebot como instrumento para realizar tal processo, obedecendo a critérios apoiados no otimizador SEO, aqui destaca-se cinco desses critérios: uso title tag nas páginas; uso das meta tags descritivas; apresentar URLs simples; oferecer conteúdo de qualidade e apresentar um texto âncora de qualidade.

Por fim, independente de quais critérios e instrumentos adotados para efetuar o tratamento temático do documento, observou-se nos levantamentos apresentados que a pesquisa no âmbito desse importante processo há de partir de diferentes áreas do conhecimento, pois trata-se da recuperação da informação de maneira eficaz e eficiente.

REFERÊNCIAS

ALMEIDA, Iêda Muniz de; CORTÊ, Adelaide Ramos. Avaliação de softwares para bibliotecas. São Paulo: Polis, APB, 2000. 108 p. (Coleção Palavra-chave, 11)

ALVARENGA, Lídia; SOUZA, Renato Rocha. A Web Semântica e suas contribuições para ciência da informação. Ciência Informação, Brasília, v. 33, n. 1, p. 132-141, jan./abril, 2004. Disponível em: < http://www.scielo.br/pdf/ci/v33n1/v33n1a16.pdf>. Acesso em: 09/12/2017.

ARRAES, Bruno Henrique Rodrigues; CAMARGO, Liriane Soares de Araújo de. Tecnologias da Informação e Comunicação Como Recurso Interativo na Perspectiva da Ciência da Informação. Revista Eletrônica Informação e Cognição, v.6, n.1, p. 3-15, 2007. ISSN:1807-8281. Disponível em:< http://www.brapci.inf.br/_repositorio/2011/04/pdf_35426f3d80_0003767.pdf>. Acesso em: 04/12/2017.

BIOLCHINE, Jorge Calmon de Almeida; GIORDANO, Rafaela Boeira. Busca e recuperação da informação científica na web: comportamento informacional de profissionais da informação. InCID: R. Ci. Inf. e Doc, Ribeirão Preto, v. 3, n.1, p. 125-145, jan./jun., 2012. Disponível em: <>. Acesso em: 09/12/2017.

BORGES, GracianeBruzinga; LIMA, Gercina Ângela de. In: XVI Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB. Informação, memória e patrimônio: do documento as redes, 26 – 30 out./2015. Disponível em:< http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/viewFile/3172/1195 >. Acesso em:

CAVALCANTE, C. R. de O.; CUNHA, M. B. da; Dicionário de Biblioteconomia e Arquivologia. Briquet de Lemos Livros: Brasília, 2008.

CORRÊA, Maurício de Vargas; ROZADOS, Helen Beatriz Frota.A netnografia como método de pesquisa em Ciência da Informação. In: Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 22, n.49, p. 1-18, maio/ago., 2017. ISSN 1518-2924. Disponível em:<http://www.brapci.inf.br/index.php/article/view/0000023115/8fc17c52adf32b316ee46b4d0f6db34d>. Acesso em: 06/12/2017.

DODEBEI, Vera; GUIMARÃES, José Augusto Chaves (Org.). Desafios e perspectivas científicas para a organização e representação do conhecimento na atualidade. Marília: Sociedade Brasileira de Organização do Conhecimento (ISKO-Brasil), 2012. Disponível em: < https://www.marilia.unesp.br/Home/Extensao/CEDHUM/livro-isko-brasil-finalizado.pdf>. Acesso em: 04/12/2017.

FIGUEIREDO, Nice Menezes de. Paradigmas modernos da Ciência da Informação: em usuários, coleções, referências & informação. São Paulo: Polis, APB, 1999. 168 p. (Coleção Palavra-chave, 10)

FUJITA, Mariângela Spotti Lopes; GIL LEIVA, Isidoro; NARUKAWA, Cristina Miyuki.Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de odontologia. Inf. & Soc.:Est., João Pessoa, v.19, n.2, p. 99-118, maio/ago. 2009.Disponívelem:<https://repositorio.unesp.br/bitstream/handle/11449/10577/WOS000269446000009.pdf?sequence=2&isAllowed=y>. Acesso em: 31/05/2017.

FUJITA, Mariângela Spotti Lopes; GIL LEIVA, Isidoro. Avaliação da indexação por meio da recuperação da informação. . **Ci. Inf.**, Brasília, DF, v. 41 n. 1, p.50-66, jan./abr., 2014. Disponível em: < http://revista.ibict.br/ciinf/article/view/1418/1596>. Acesso em: 15/12/2017.
GIL, A. C. Como elaborar projetos de pesquisa. 5. ed. São Paulo: Atlas, 2010.

GIL LEIVA, I. **La automatización de La indización, propuesta teórico metodológica: aplicación al área de Biblioteconomía y Documentación**.1997. 268 p. Tese – Universidad de Murcia, Murcia, Espanha, 1997. Disponível em: < http://webs.um.es/isgil/resources/PhDissertation%20Gil-Leiva.pdf >. Acesso em: 25/06/2017.

GUIMARÃES, José Augusto Chaves; SALES, Rodrigo de Sales. A importância de Julius Kaiser para a Organização do Conhecimento: um estudo comparativo com as perspectivas de Cutter, Otlet e Ranganathan. **InCID: R. Ciência da Informação e Documentação**, Ribeirão Preto, v. 7, n. 1, p. 43-65, mar./ago. 2016. Disponível em: < https://www.revistas.usp.br/incid/article/view/110214/111647>. Acesso em: 15/12/2017

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. ver. atual. Brasília: Briquet de. Lemos, 2004.

MARCONI, M. A.; LAKATOS, E. M. **Metodologia do trabalho científico**: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos. 7. ed., 5. reimpr. São Paulo: Atlas, 2010.

MOURA, M. A. Informação, ferramentas ontológicas e redes sociais ad hoc: a interoperabilidade na construção de tesauros e ontologias. **Informação & Sociedade**: Estudos, v. 19, n. 1, p. 59-73, 2009. Disponível em: <http://www.brapci.inf.br/v/a/7570>. Acesso em: 27/11/ 2017.

ONUBR – ORGANIZAÇÃO DAS NAÇÕES UNIDAS BRASIL. Acesso em 04/12/2017. Disponível em:<https://nacoesunidas.org/brasil-ocupa-66o-lugar-em-ranking-da-onu-de-tecnologia-de-informacao-e-comunicacao/>

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, Brasília, v. 20, n. 2, p. 130-136, jul./dez. 1991. Disponível em: < http://www.brapci.inf.br/_repositorio/2010/04/pdf_2b09a726d5_0009108.pdf>. Acesso em: 10/03/2017.