VIVIANE NEVES DOS SANTOS

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTUAIS: INICIATIVAS DOS GRUPOS DE PESQUISA DE UNIVERSIDADES PÚBLICAS BRASILEIRAS

São Paulo
2009

VIVIANE NEVES DOS SANTOS

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTUAIS: INICIATIVAS DOS GRUPOS DE PESQUISA DE UNIVERSIDADES PÚBLICAS BRASILEIRAS

Trabalho de conclusão de curso apresentado ao Departamento de Biblioteconomia e Documentação da Escola de Comunicações e Artes da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia.

Orientadora: Profª Drª Nair Yumiko Kobashi

São Paulo

2009

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Contatos

E-mail 1: vivianeneves81@gmail.com

E-mail 2: vivivns@yahoo.com.br

Catalogação na publicação elaborada pelo próprio autor

SANTOS, Viviane Neves dos

Indexação automática de documentos textuais: iniciativas dos grupos de pesquisa de universidades públicas brasileiras / Viviane Neves dos Santos; Nair Yumiko Kobashi (Orientadora). São Paulo, 2009.

72 p.

Trabalho de Conclusão de Curso (Bacharelado em Biblioteconomia) - Departamento de Biblioteconomia e Documentação. Escola de Comunicações e Artes. Universidade de São Paulo.

1. Indexação automática. 2. Modelos de indexação automática. 3. Grupos de Pesquisa em Indexação automática. I. Autor. II. Título. III. Orientadora.

FOLHA DE APROVAÇÃO

Nome: SANTOS, Viviane Neves dos

Título: Indexação automática de documentos textuais: iniciativas dos grupos de pesquisa de universidades públicas brasileiras

Trabalho de conclusão de curso apresentado ao Departamento de Biblioteconomia e Documentação da Escola de Comunicações e Artes da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia.

Banca Examinadora

Presidente da Banca: Profª Drª Nair Yumiko Kobashi

Profª Drª. Vânia Mara Alves Lima Instituição: Universidade de São Paulo

Prof. Dr. Marcelo dos Santos Instituição: Universidade de São Paulo

Aprovada em:

____/____/_____

AGRADECIMENTOS

A Deus que me permitiu chegar até aqui. Meu abrigo, amparo e conforto nos momentos de calmaria e de turbulência.

À minha mãe, exemplo de caráter e força, pelo seu amor incondicional, amor só possível vindo de uma mãe, a quem devo tudo que fui, que sou e que serei, aquela que me inspira a continuar andando.

A Antônio (Magrinho) companheiro de minha mãe, que me considera sua sexta filha, mesmo não sendo, que vibrou comigo quando entrei na faculdade e que sempre tem uma palavra de apoio e um bom conselho a me dar.

Às mães que tive durante a vida, tia Tereza, tia Lourdes e minha prima Marilene (Ziza). Aos pais do meu amigo Carlos, Aldeniza e Marcos, que adotei como meus pais.

Aos amigos Carlos, Elisangela, Renata, Heloísa Kodama, Ricardo, Sarah, Larissa Raci, Virgínia, Larissa Neves (priminha), Andrea Laila, Vanessa Madeleine, à tríade (que não é mais de três) Luciana, Patrícia, Maria Irene, Geslaine (*in memorian*). Às amigas Carol e Juju e aos amigos Agamenon, Alex, Gledson, Thiago Gaudêncio e Tiago Murakami. Agradeço a vocês pelas conversas, pela ajuda, por poder compartilhar os bons e os difíceis momentos dessa vida.

A Demetrios, meu namorado, amigo e companheiro da vida, pela força, carinho e bom humor sempre.

À Michely Vogel pelo incentivo, pela leitura do trabalho, sugestões e correções.

A todos os amigos que mesmo não citados estão sempre em meu coração.

À equipe do Instituto Fernando Henrique Cardoso e da Grifo por proporcionarem meu primeiro estágio. Agradecimento especial à bibliotecária Francisca Evrard, mestra e amiga, que me guiou nos primeiros passos da Biblioteconomia. À equipe da BIREME, meu segundo estágio, em especial a Luciano Soares Duarte, Selma Palombo, Sueli, Maria Anália e Ernesto Spinak. À Ana Belluzzo e equipe do Projeto "Arte no Brasil". Agradeço a todos pelas contribuições em minha formação profissional e aprendizado.

A todos os professores da Escola Estadual Nossa Senhora Aparecida, meus primeiros mestres no Ensino Fundamental e Médio, que um dia, lá atrás, disseram que eu poderia conseguir.

Aos professores do CBD que me mostraram a importância de ser bibliotecária, além de contribuirem para minha formação.

Agradeço ao Prof. Marcos Mucheroni pelo incentivo e solução de dúvidas, contribuindo para este trabalho.

À professora Nair Kobashi, pela orientação, paciência, pela disposição em ajudar, bem como pelas correções e *insights* que contribuíram muito neste trabalho. Sou muito grata, também, por acreditar em mim, mesmo quando nem eu mesma acreditava.

À Biblioteconomia, que me possibilita uma vida melhor por saber que há sentido no que faço e confirmar, a cada dia, que decidi pelo caminho certo, pois faço por amor e sem amor eu nada seria.

SANTOS, Viviane Neves dos. **Indexação automática de documentos textuais:** iniciativas dos grupos de pesquisa de universidades públicas brasileiras. 2009. 72 p**.** Trabalho de Conclusão de Curso (Bacharelado em Biblioteconomia) – Departamento de Biblioteconomia e Documentação, Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2009.

RESUMO

Estudo sobre os modelos de indexação automática e seu uso no tratamento e recuperação de documentos na *Web* e em processos de indexação em bancos de dados bibliográficos. Apresenta-se um breve histórico da indexação automática, seus principais conceitos, as áreas relacionadas e a classificação dos métodos de indexação automática. São também identificados e caracterizados os grupos de pesquisa brasileiros que se dedicam ao tema. Conclui-se que os grupos desenvolvem pesquisas sobre o Processamento de Linguagem Natural (PLN), Sistemas Inteligentes, bem como Sistemas Inteligentes combinados com PLN. As propostas de indexação automática tendem à integração de diferentes perspectivas, de modo a pemitir o uso da linguagem natural como linguagem de intercâmbio entre usuário e sistema. Confirma-se a característica interdisciplinar da indexação automática, sendo sugerida a parceria entre os grupos para compartilhamento de recursos que conduzam ao avanço das pesquisas sobre a indexação automática.

PALAVRAS-CHAVE: Indexação automática; Modelos de indexação automática; Grupos de Pesquisa em Indexação automática

LISTA DE ABREVIATURAS E SIGLAS

CDD - Classificação Decimal de Dewey

CDU - Classificação Decimal Universal

CID - Classificação Internacional de Doenças

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

DeCS - Descritores em Ciências da Saúde

GED -Gestão Eletrônica de Documentos

HTML - HyperText Markup Language

IBICT - Instituto Brasileiro de Informação em Ciência e Tecnologia

IDF - Inverse Document Frequency

IHC - Interação Humano-Computador

KWAC - Keyword alongside context

KWIC - Keyword in context

KWOC - Keyword out of context

LaLiC - Laboratório de Linguística Computacional

LCC - Library of Congress Classification

LCSH - Library of Congress Subject Headings

LD - Linguagem Documentária

MHTX - Modelo Hipertextual para Organização de Documentos

NILC - Núcleo Interinstitucional de Linguística Computacional

PLN - Processamento de Linguagem Natural

URL - Uniform Resource Locator

XML - eXtensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

O advento da Internet promoveu a aceleração dos processos de comunicação, provocando grandes mudanças nas formas de fazer ciência, desenvolver tecnologias, como também em diversos aspectos da vida em sociedade.

A área da Biblioteconomia e o fazer bibliotecário também vêm sendo afetados pela informatização. Inicialmente, os computadores foram utilizados para otimizar os processos de tratamento da informação e, logo depois, para disponibilizar informação, substituindo os catálogos manuais por catálogos *online.* Estes últimos superam os catálogos manuais porque oferecem maior número de pontos de acesso para pesquisa, permitem fazer buscas simultâneas em campos diferentes e, desse modo, promovem rapidez na recuperação. Pode-se afirmar, ainda, que representar descritiva e tematicamente um documento tem sido a solução mais adequada para que o usuário recupere informação no imenso universo de documentos produzidos e disponibilizados.

Com a invenção da *World Wild Web*, na década de 90, foram criados novos tipos de documentos, as páginas HTML (*HyperText Markup Language*) e a possibilidade de navegação por *links*. Essa rede, mais conhecida como *Web*, pode ser definida como um sistema de documentos (hipertextos, sons, figuras) interligados e executados na Internet.

Para dar alguma ordem ao atual universo de informação e comunicação, o *World Wide Web Consortium*, fundado em 1994 por Tim Berners Lee, passa a elaborar padrões e recomendações para o desenvolvimento de recursos para a *Web*. Com relação aos documentos textuais, emerge como padrão de metadados a linguagem de marcação XML (*eXtensible Markup Language*).

Em 2001, surge a *Web 2.0* que, dentre outras características, destaca-se por ser uma rede colaborativa, com forte participação do usuário, tendo a própria *Web* como plataforma de trabalho. Assim, cresce a tendência de utilização de aplicativos diretos na *Web*.

Atualmente, a grande dificuldade enfrentada pelo usuário é encontrar informação pertinente dentro da imensa massa de documentos que circula na Rede. É neste cenário que surge a Web semântica que objetiva criar uma rede de dados que possam ser utilizados e combinados entre os aplicativos, empresas e comunidade em geral. Pretende-se, na Web semântica, atribuir sentido aos dados para que possam ser mais facilmente compartilhados e manipulados.

Neste contexto, podem ser destacadas duas questões: a informação não está mais presa a formatos fixos e a tendência é de que os documentos tenham estruturas descritas com a utilização de padrões abertos, como o XML. Isso permite que uma mesma página *Web* possa ser exibida em navegadores diferentes, não sendo necessário recorrer a softwares específicos para ter acesso à informação. Além disso, os documentos e páginas podem ser exibidos de maneiras diferentes, com estilos e cores variados, uma vez que há autonomia entre a apresentação da informação, a estrutura do documento e seu conteúdo. O outro aspecto refere-se ao usuário, que passa a ser também proudutor de conteúdos.

Um fenômeno marcante do período atual é o surgimento de novas formas de interação: há uma explosão de Blogs, Repositórios, Comunidades Virtuais e Foruns constituídos de pessoas que se reúnem por afinidades e interesses comuns. Nesses espaços, o usuário não só publica, como também descreve e indexa as informações. A alteração produzida pela Internet é de tal monta que praticamente toda informação produzida está na Internet, desde a comunicação científica até as informações voltadas para entretenimento ou lazer. Considerando a difusão desse meio e sua larga utilização, há quem diga que o que não está na Internet, não existe. Todavia, pode-se afirmar que a informação que está na Internet e não é recuperada, portanto perdida na Rede, também não existe. Isso pode ser aplicado também às instituições, principalmente as de caráter científico, que hoje têm na Internet uma das principais ferramentas de divulgação de resultados de pesquisas e de comunicação entre

pesquisadores. Todavia, as instituições enfrentam o desafio de adaptação ao novo contexto informacional e tecnológico, sendo necessário introduzir modificações radicais nas formas de processar e apresentar informação.

Padrões de descrição e importação de dados, como os metadados utilizados pelo SciELO e políticas de catalogação colaborativa, como a desenvolvida para a catalogação de materiais para a base LILACS, iniciativas da BIREME, são alguns dos métodos utilizados para otimizar os processos de tratamento dos documentos para rápida disponibilização.

Por outro lado, devido ao volume crescente de documentos e à velocidade de sua produção, há iniciativas que visam a representar conteúdos de documentos automaticamente (ou de maneira semi-automática), recorrendo às técnicas da área de Apredizado de Máquina, um ramo da Inteligência Artifical, associadas às técnicas de Processamento de Linguagem Natural (PLN). São os chamados métodos de indexação automática, cujas origens remontam a técnicas mais simples, baseadas em frequência/ocorrência de palavras, utilizados desde os anos 1960.

Observou-se, durante o levantamento bibliográfico feito para a presente pesquisa, que há atualmente um considerável número de trabalhos sobre indexação automática, que utiliza variados tipos de técnicas e algoritmos aplicados principalmente à indexação de documentos na Internet. Esses trabalhos nomeiam a atividade ora como Classificação automática, ora como Categorização automática.

Este cenário justifica os objetivos deste Trabalho de Conclusão de Curso, de estudar os modelos atuais de indexação automática, em particular os projetos de grupos de pesquisas de universidades públicas brasileiras e identificar os produtos/softwares desenvolvidos para a indexação de textos em língua portuguesa.

Foram definidos, a partir daí, os objetivos da pesquisa, apresentados a seguir.

- Objetivo Geral:

Realizar um estudo exploratório sobre os métodos ou modelos de indexação automática de documentos textuais.

- Objetivos Específicos:

Verificar, com base na literatura publicada na área da Ciência da Informação, os atuais modelos de indexação automática.

Verificar seu uso no tratamento e na recuperação de documentos na *Web* e em processos de indexação em bancos de dados bibliográficos já constituídos.

Levantar as inciativas de desenvolvimento de sistemas ou pesquisa de métodos de indexação automática em grupos de pesquisas de universidades públicas brasileiras.

- Procedimentos Metodológicos:

Para o desenvolvimento da pesquisa, foi feito inicialmente um levantamento da bibliografia relacionada ao assunto estudado neste projeto. Foram feitas buscas nas fontes Library and Information Science Abstracts (LISA), Web of Science, Annual Review of Information Science and Technology (ARIST), Portal do CRUESP, que reúne os catálogos e recursos da USP, Unesp e Unicamp, Pesquisa Brasileira em Ciência da Informação e Biblioteconomia, outros periódicos na área de Biblioteconomia e Ciência da Informação e pesquisas na Internet. Foram consideradas também obras referenciadas nos materiais já lidos para a redação do projeto.

Para levantamento da literatura foram utilizados, em português e inglês, os assuntos “Indexação automática”, “Indexação semi-automática”, “Categorização automática”, “Classificação automática”, combinando-os com “Sistemas” e “Métodos”, os assuntos “Sistemas de Recuperação de Informação” e “Mineração de texto” também foram considerados.

Em cada fonte foi feita busca com os assuntos selecionados, nos campos “Título”, “Palavras-chave”, “Resumo” e no próprio texto do item quando disponível o texto completo.

Cabe dizer que, uma vez que a literatura sobre o assunto publicada até 1998 foi revista na obra de Isidoro Gil Leiva (1999), as buscas se concentraram em obras publicadas do ano de 1998 em diante.

Com base nos subsídios teóricos fornecidos pelas obras lidas, foi elaborada uma grade de análise para classificar as tendências em indexação automática dos grupos de pesquisa brasileiros.

Os resultados da pesquisa estão organizados nas seguintes seções:

O Capítulo 1 – Introdução, apresenta os objetivos da pesquisa, justificativas e procedimentos metodológicos para levantamento da literatura utilizada para a parte teórica do trabalho.

O Capítulo 2 – Sobre a Indexação, contém o conceito de indexação, bem como localiza a indexação dentro do fluxo do Sistema de Recuperação de Informação e levanta resumidamente os fatores que interferem no produto final da indexação.

O Capítulo 3 – Indexação automática: conceitos, versa sobre a automatização da indexação e conceitos relacionados (indexação automática, semi-automática e assistida por computador).

A história da indexação automática é vista no Capitulo 4 – História da indexação automática.

No Capítulo 5 – Razões para uma indexação automática, faz-se uma breve revisão dos fatores favoráveis e contrários à indexação automática.

No Capítulo 6 – A interdisciplinaridade da indexação automática, caracteriza-se a indexação automática como área interdisciplinar e são relacionas as disciplinas que contribuem para seu desenvolvimento.

O Capítulo 7 – Indexação nos dias atuais, indexação automática e indexação na Internet, trata dos dias atuais da indexação automática, incluindo a indexação na Internet.

No Capítulo 8 – Modelos de indexação automática, é feita uma classificação dos modelos indexação automática de acordo com a literatura.

O Capítulo 9 – Grupos de pesquisa no Brasil na área de indexação automática, trata da metodologia para levantamento dos grupos de pesquisa, incluindo a descrição de cada um e sua classificação de acordo com quadro de modelos de indexação automática elaborado a partir dos modelos identificados na literatura descritos no Capítulo 8.

Nas "Considerações Finais" são apresentados e discutidos os resultados alcançados, bem como feitas indicações sobre trabalhos futuros.

2 SOBRE A INDEXAÇÃO

Uma das missões da Biblioteconomia e da Documentação é tratar e organizar informação para sua difusão. Para cumprir tal missão, o bibliotecário ou profissional da informação desenvolve atividades que envolvem a seleção de documentos e seu tratamento, tendo em vista as necessidades dos usuários. Para atender aos usuários, é necessário também promover a adaptação contínua dos sistemas de informação.

Dentre as atividades bibliotecárias típicas podem ser citadas a representação descritiva e a análise documentária, que tratam, respectivamente, da descrição dos aspectos formais e de conteúdo dos itens de informação. O processo de indexação é uma das principais atividades realizadas pela Biblioteconomia e Documentação e constitui uma das formas de representação do conteúdo de documentos**.**

Para melhor compreensão do funcionamento de um sistema automático ou semi-automático de indexação, será feita, a seguir, uma caracterização resumida do conceito de indexação, seus objetivos, estapas e instrumentos, bem como sua relação com a recuperação de informação.

Os sistemas de recuperação de informação, de modo geral, apresentam uma entrada onde é recebido um documento selecionado de acordo com a política do serviço de informação. Este passa pelas tarefas de representação descritiva e representação temática. Os produtos da descrição são armazenados em uma base de dados para acesso. Neste processo, um usuário com uma determinada necessidade informacional, fará perguntas ao sistema por meio de estratégias de busca (LANCASTER, 2004, p.2).

Apesar das diferentes correntes teóricas que conceituam a indexação, aceita-se aqui, que ela é uma operação de representação documetária com a finalidade de recuperar informação, localizando-se a Indexação dentro da área de “Análise Documentária” (CINTRA et al., 2002, p.33).

Por “Análise Documentária” compreende-se, no sentido apresentado por Silva e Fujita (2004) como:

> "[...] área teórica e metodológica com o objetivo de tratamento temático de documentos, que abrange as atividades de Indexação, Classificação e elaboração de resumos, considerando as diferentes finalidades de recuperação da informação." (p. 138)

Com base no modelo de sistema de recuperação de informação apresentado por Lancaster, verifica-se que, na entrada no sistema, é feita a representação do documento para armazenamento. As etapas, segundo o autor, são constituídas da Análise Conceitual e da Tradução. Na saída do sistema, há a pergunta do usuário, formulada por meio de estratégias de busca, feita com o intuito de recuperar documentos. Ambos os processos são mediados pelo “Vocabulário do Sistema”, necessitando também, na saída do sistema, que seja feita a Análise Conceitual e a Tradução da pergunta do usuário para a linguagem do sistema.

Apesar de ser possível descrever de diversas formas as etapas do processo de indexação, pode-se dizer, com base no exposto até agora, que ele ocorre na entrada dos sistema de recuperação de informação, quando é feita a leitura, análise e representação do conteúdo do documento, com base em um vocabulário ou linguagem documentária do sistema; e ocorre também na saída do sistema, quando é feita a “leitura” da pergunta do usuário e a elaboração de uma expressão de busca, de acordo com o vocabulário ou linguagem documentária do sistema.

Gil Leiva (1999, p.19-20), com relação ao conceito de indexação, afirma que a maioria dos conceitos são incompletos por se referirem, muitas vezes, apenas aos documentos como fontes de análise, ignorando a pergunta do usuário.

Para o autor, a indexação ocorre em dois momentos: a *indexação do documento*, para armazenamento; e a *indexação da pergunta* do usuário, cujo objetivo é obter o que o autor chamou de “resposta documental”, ou seja, para recuperar documentos que atendam à necessidade do usuário, materializada na expressão de busca.

Gil Leiva divide a *indexação dos documentos* em duas etapas. A primeira refere-se à *leitura do documento*, que por sua vez se divide em uma "leitura horizontal", em que são analisados e selecionados os conceitos presentes no documento; e em uma "leitura vertical", onde são identificados e atribuídos termos referentes aos conceitos implícitos no documento. Na segunda etapa, os conceitos em linguagem natural podem ser armazenados em linguagem natural ou convertidos para os termos de uma linguagem documentária.

Um sistema automático apenas seria capaz de identificar conceitos implícitos se ele fosse capaz de reconhecer um padrão e inferir que uma expressão refere-se a um conceito; esta é uma tarefa difícil pois a identificação requer lidar com um universo simbólico, aspecto ainda não suficientemente descrito para ser incorporado aos sistemas de indexação automática.

A possibilidade de a máquina interpretar um texto tal como um ser humano o faria, é tarefa subjetiva ainda não realizável pelos sistemas de indexação, dado que a maioria deles, apesar de fazerem algum processamento baseado em referenciais linguísticos, e/ou com uma ajuda de uma linguagem documentária, baseiam-se predominantemente em métodos estatísticos. Com relação aos conceitos explícitos, estes já seriam de fácil identificação dado que o sistema trabalharia com o que está presente "materialmente" no texto (os símbolos), ou seja, seria feito um reconhecimento dos significantes presentes no documento.

Na saída do sistema, tanto Cleveland e Cleveland (1990, p.20) quanto Gil Leiva (1999, p.21), afirmam que a indexação da pergunta passa pelo mesmo processo realizado sobre o documento na entrada do sistema. Todavia, enquanto neste processo, a indexação do documento é orientada às possíveis perguntas dos usuários, naquele, a indexação da pergunta é orientada para o documento, ou seja, tenta-se elaborar uma expressão de busca com os termos que possam constar como termos de indexação de determinado documento.

São utilizados diferentes nomes para designar o processo de indexação. Ora se emprega o termo indexação de assuntos, ora classificação, categorização e, ainda, catalogação de assuntos. Por vezes, classificar e indexar são vistos como processos distintos, pois o primeiro utiliza-se de um sistema de classificação e o segundo pode utilizar palavras ou expressões do próprio texto para a representação o conteúdo. Porém, o ato em si consiste em atribuir uma representação a um documento, com o intuito de armazená-lo e recuperá-lo do ponto de vista de seu conteúdo. Portanto, em essência, classificar e indexar significam praticamente a mesma coisa. Adota-se aqui, pela frequência na literatura, o termo indexação para todos os processos, com base na afirmação de Lancaster, de que:

> "O processo que consiste em decidir do que trata um item e de atribuir-lhe um rótulo que represente esta decisão é conceitualmente o mesmo, quer o rótulo atribuído seja extraído de um esquema de classificação, de um tesauro ou de uma lista de cabeçalhos de assuntos, quer o item seja uma entidade bibliográfica completa ou parte dela, quer o rótulo seja subseqüentemente arquivado em ordem alfabética ou em outra seqüência (ou, com efeito, não arquivado de modo algum), quer o objeto do exercício seja organizar documentos em estantes ou registros em catálogos, índices impressos ou bases de dados eletrônicas." (LANCASTER, 2004, p. 21)

Há outro elemento importante na indexação: a linguagem. A Indexação, enquanto atividade que cria representações de conteúdos explícitos e implícitos dos documentos, utiliza-se de uma linguagem constituída de termos que podem ser armazenados ou usados para busca em linguagem natural ou convertidos para o vocabulário do sistema, ou seja, uma Linguagem Documentária (LD).

Segundo Cintra et al. (2002, p. 33) as linguagens documentárias têm sua origem associada à necessidade de resolução das dificuldades de armazenamento e recuperação de informações, surgidas nas décadas de 50 e 60, dado o crescimento do conhecimento científico e tecnológico.

Essas linguagens podem ser definidas como linguagens construídas e constituídas de símbolos para representação do conteúdo dos documentos, para armazenamento e

recuperação. De maneira geral, operam no sentido de evitar ambiguidades na representação da informação, bem como no agrupamento daqueles que possam ser representados por termos sinônimos, além de tentar garantir a univocidade dos termos, evitando a polissemia. Esta característica, que torna a linguagem natural rica, por outro lado, dificulta a recuperação da informação. Além disso as LDs deixam explícitas as normas ou regras de como devem ser utlizados os termos no ato da indexação e as relações entre os termos da linguagem (sinonímicas, hierárquicas ou associativas) (CINTRA et al., 2002; GIL LEIVA, 1999).

Os aspectos anteriormente descritos evidenciam que, além de exercer a função de instrumento de representação, as linguagens documentárias também têm uma função comunicativa, ou seja, "a normalização das representações documentárias como meio de viabilizar sua comunicação" (LARA, 1993, p.223), portanto, envolve também a questão da significação, tendo como problema a representação de um documento de modo a não alterar o seu significado. Além de outros fatores, as características da linguagem utilizada na indexação influenciam a comunicação que será estabelecida entre o sistema de informação e o usuário, no ato da busca; assim, uma indexação incoerente ou a limitação da linguagem de indexação podem provocar a comunicação incorreta do conteúdo do documento. De maneira geral, as linguagens documentárias, enquanto linguagens de tratamento e recuperação da informação, podem ser classificadas de acordo com três pontos de vista (GIL URDICIAIN, 1996, p. 22 citado por GIL LEIVA, 1999, p. 49):

a) Tipo de controle de vocabulário – nesse sentido, podem ser livres (listas de descritores livres) ou controladas (classificações, listas de cabeçalhos de assuntos e tesauros).
b) Pela coordenação – pré-coordenadas (classificações e listas de cabeçalhos de assunto) e pós-coordenadas (lista de descritores livres, listas de palavras-chave e tesauros)

c) Pela sua estrutura – podem apresentar-se como hierárquicas (classificações), alfabéticas (listas de cabeçalhos de assunto), ou ambas (tesauros).

Exemplo de lista de cabeçalhos de assunto é a Lista de Cabeçalho de Assunto da Library of Congress (Library of Congress Subject Headings – LCSH). Exemplos de classificações são a Classificação Decimal de Dewey (CDD), Classificação Decimal Universal (CDU), Library of Congress Classification (LCC) e a Classificação de Ranganathan (Colon Classification). Estas são denominadas classificações enciclopédicas, posto que procuram abarcam todo conhecimento humano. Todavia, há classificações especializadas, como a Classificação Internacional de Doenças (CID), utilizada na área da Saúde para "indexação" das doenças, em prontuários médicos ou atestados de óbito, possibilitando a análise estatística de doenças, causas de mortes etc (WORLD HEALTH ORGANIZATION, [2009?]). Como exemplo de tesauro, pode ser citado o Tesauro da Unesco (UNESCO Thesaurus).

Incluem-se aqui também as Taxonomias, que vêm sendo utilizadas para a recuperação de informação em portais e bibliotecas digitais. As taxonomias permitem acesso por meio de navegação baseada em estruturação lógica de termos, organizados em classes e sub-classes, com quantidade de subdivisões definida de acordo com a necessidade. As Ontologias também são consideradas neste grupo que, tal como as taxonomias, desempenham papel importante na Web Semântica.

Campos e Gomes (2008) resumidamente ilustram a importância do papel das ontologias e taxonomias para a Web Semântica:

> "Para que a Web semântica venha a funcionar de forma efetiva, computadores têm que ter acesso às coleções estruturadas de informações e a conjuntos de regras de inferência que se consolidam através de mecanismos como as ontologias. Estas são meios poderosos de inter-relacionar sistemas e neste contexto elas possuem papel de destaque, como podemos observar através dos componentes que integram uma ontologia, ou seja: Termos e Definições; Classes e subclasses - que podem estar organizadas em uma taxonomia; Relações (também chamadas de propriedades), que devem representar os tipos de interação entre as classes de um domínio; Axiomas

> que são regras para determinar a verdade das sentenças; e Instâncias que são utilizadas para representar elementos específicos, ou seja, os próprios dados."

Enquanto instrumentos utilizados para a representação do conteúdo de documentos, possibilitando armazenamento e recuperação de informação na Internet, consideram-se aqui as ontologias e taxonomias como tipos de linguagens documentárias.

No estudo e no exercício da atividade de indexação, há que se considerar outros fatores que influenciam o produto final da indexação e, consequentemente, a recuperação do documento, sendo alguns deles relacionados à política de indexação da instituição. São exemplos, as partes do documento utilizadas para a indexação (se títulos, resumos ou texto completo), o tempo dedicado à indexação, a exaustividade, a especificidade e o grau de pré-coordenação da linguagem documentária ou vocabulário do sistema.

Outras características mais alinhadas com a qualidade do produto da indexação são a correção e a coerência.

A indexação correta é caracterizada pela ausência de erros. Os erros podem ser causados pela omissão de um descritor necessário ou pela atribuição de um descritor incorreto. Esse fator afeta diretamente a qualidade da recuperação de informação.

A coerência pode ser definida como o grau de concordância entre as indexações feitas por diferentes indexadores, bem como o grau de concordância entre as indexações de um mesmo indexador. (LANCASTER, 2004, p.68; GIL LEIVA, 1999, p.26).

A coerência pode ser medida pela razão entre os termos coincidentes atribuídos a um documento pelos sistemas ou indexadores avaliados, e a soma dos termos atribuídos por ambos, subtraindo-se os termos coincidentes. (GIL LEIVA, 1999, p.31).

Essa proposta de Salton e McGill (citados por GIL LEIVA, 1999) foi inicialmente pensada para avaliar a consistência entre indexação manual e indexação automática. Ela pode ser empregada para avaliação de sistemas de indexação automática, não no sentido de

oposição entre a indexação manual e automática, mas no sentido de verificação da consistência para posterior melhora ou correção dos parâmetros do sistema de indexação, em um trabalho conjunto da indexação manual e automática.

Outra forma de avaliar a indexação é por meio da recuperação de documentos pelos índices de precisão e revocação do sistema. São conceitos de grande importância para elaborar indicadores de desempenho de bases de dados ou sistemas de recuperação de informação.

Lancaster (2004, p. 4), com relação ao uso de revocação e precisão, diz que, apesar de existirem outras abordagens, elas são medidas a serem utilizadas para expressar os resultados de qualquer busca que simplesmeste divida uma base de dados em recuperados e não recuperados. Emprega revocação (*recall*) como a capacidade de um sistema de informação de recuperar documentos úteis; e precisão, a capacidade evitar documentos inúteis.

O coeficiente de revocação é constituído pela razão entre os documentos relevantes recuperados em uma busca e o total de documentos relevantes do sistema (CLEVELAND e CLEVELAND, 1990, p.149). A princípio já se pode afirmar que quanto mais tendente a um o coeficiente, maior é a capacidade do sistema de recuperar itens relevantes para uma determinada busca.

O coeficiente de precisão leva em conta a razão entre documentos relevantes recuperados e o total de documentos recuperados em uma busca (relevantes e irrelevantes). Pode-se inferir, então, que quanto mais tendente a um, mais precisa será a busca, pois maior será a quantidade de itens relevantes recuperados efetivamente.

Outra consideração é a de que precisão e revocação são inversamente proporcionais, ou seja, quanto maior a precisão de um sistema, menor será sua revocação (LANCASTER, 2004, p. 4; CLEVELAND e CLEVELAND, 1990, p. 150).

Se o objetivo de um sistema de informação é recuperar informação, então precisão e revocação são pontos a serem considerados e medidas que podem fornecer, de certa forma, parâmetros para a avaliação de um sistema de indexação, seja ele automático ou não.

Cabe ressaltar que para grandes volumes de informação é recomendável que se tenha maior índice de precisão e não de revocação. Lancaster recomenda isso ao afirmar que "quanto maior for a base de dados, menos aceitável será uma baixa precisão." (2004, p. 4). Como explicação, o autor alega que o usuário pode ter disposição para examinar 57 itens com o fim de encontrar 6 que lhe sejam satisfatórios, mas não examinaria 570 itens com o fim de selecionar 60. Tal afirmação é igualmente aplicável ao resultado de uma busca na Internet, em que o usuário se dispõe a verificar os *links* da primeira página de resultado, todavia não há garantia de que ele verifique as demais.

Pode-se dizer, então, que em grandes bancos de dados e na *Web*, além de recuperar informação, as inciativas devem objetivar, principalmente, a precisão dos resultados das buscas, pois um aumento na revocação geraria como resultado muitos registros a serem examinados, sob risco de poucos deles serem relevantes para a busca efetuada.

Para uma melhora na precisão de sistemas de recuperação de informação, o emprego da indexação é necessário. Dada a característica descentralizada da Internet e o volume de informação crescente, as iniciativas que visem à automatização do processo são bem-vindas, posto ser de difícil realização a indexação manual de todos os documentos disponíveis na Rede.

Considerando a indexação tradicional, são muitos os fatores que influenciam sua qualidade, mesmo havendo uma política delimitada, uma linguagem bem estruturada e pessoas bem treinadas para o tratamento e a difusão da informação. A indexação exige um esforço intelectual e requer padrões e métodos para contornar a subjetivade da compreensão

de mensagens presentes em textos. As possibilidades de várias interpretações de um texto, uma característica inerente a eles, pode causar incoerência nas indexações.

As iniciativas de automatização são propostas, portanto, para facilitar o trabalho do indexador, conferindo padronização à indexação e para tentar resolver o problema de tratamento da crescente massa documental com a qual os serviços de informação precisam lidar na atualidade.

Sistemas automáticos que abarquem todas as etapas do processo de tratamento, armazenamento e recuperação da informação e os agentes envolvidos (usuários, profissionais da indexação, autores dos documentos e instituições que abrigam os serviços de informação) podem ser de grande valia e se deve considerá-los como alternativas para a melhoria dos resultados de busca e dos produtos da indexação. É necessário observar, no entanto, que estes sistemas ainda estão por vir.

3 INDEXAÇÃO AUTOMÁTICA: CONCEITOS

Sendo a indexação a representação de um documento ou das perguntas feitas pelos usuários, no ato de busca, por meio de linguagem natural ou uma linguagem documentária, a indexação automática seria a execução deste processo por meio de programas ou algoritmos de computador que "varrem" o documento (ou registros de documentos) e realizam a representação do conteúdo sem a intervenção do documentalista.

Em revisão de literatura feita por Gil Leiva (1999, p.57-58), foi identificada uma grande variedade de termos utilizados para denominar a automatização da indexação, sendo o termo "Automatic indexing" (Indexação Automática) a forma mais utilizada.

Todos os termos levantados pelo autor referiam-se à automatização da indexação, representando três conceitos diferentes:

- Indexação assistida por computador durante o armazenamento: sistemas que auxiliam o processo de armazenamento dos termos de indexação extraídos pelo indexador na etapa de análise conceitual. São facilitadores do processo de indexação uma vez que proporcionam, por meio de janelas de ajuda, notas explicativas sobre os termos e seus relacionados e, às vezes, acesso a documentos já indexados, para solução de dúvidas.
- Indexação semi-automática: sistemas que indexam automaticamente o documento e, se necessário, dão a possibilidade de edição e validação dos termos pelo documentalista.
- Indexação automática: sistemas sem nenhuma validação por parte do documentalista; os termos de indexação são armazenados diretamente como descritores do documento.

Anderson e Perez-Carballo (2001b, p.256) definem indexação automática como a "análise do texto por meio de algoritmos de computador".

Na mesma linha, Hjørland (2008) define indexação automática como "a indexação feita por procedimentos algorítmicos". O algoritmo pode trabalhar em uma base contendo representações dos documentos, e/ou texto completo, registros bibliográficos ou partes do texto, bem como pode ser efetuada em bases de materiais não-textuais, como imagens ou música.

Ainda de acordo com o autor acima, algumas técnicas podem ser totalmente automáticas, enquanto outras, semi-automáticas. Cita como processamento semi-automático a técnica "*Machine-Aided indexing*". Exemplos dessa abordagem são os sistemas NewsIndexer (REDMOND-NEAL, 2003) e o M.A.I. (*Machine Aided Indexer*) um aplicativo do sistema Data Harmony da Access Innovation, Inc.(HLAVA, 2003), que utilizam um vocabulário controlado e realizam a operação de comparar as expressões extraídas do documento com as expressões de uma linguagem documentária. Como processamentos totalmente automáticos há aqueles que utilizam técnicas de "Categorização de Texto" (*Text Categorization*) e agrupamento (*clustering*).

GOLUB[*](2005, p.52-53), em pesquisa sobre indexação automática para páginas *Web* utilizando vocabulários controlados, diferencia as três "técnicas" acima citadas de acordo com sua área predominante.

A Categorização de Textos, de acordo com a autora, é uma abordagem da área de Apredizado de Máquina (*Machine-Learning*), na qual os métodos da área recuperação da informação são também aplicados. Envolve, basicamente, a construção de indexadores automáticos (classificadores automáticos) que são capazes de aprender e classificar documentos apoiando-se em um conjunto de categorias pré-definidas e uma "instância de

[*]A Dra. Koraljka Golub é pesquisadora do Grupo UKOLN da Universidade de Bath (Reino Unido) e faz parte do conselho editorial do periódico International Journal of Digital Library Systems.

treino" de documentos já pré-classificados manualmente, que servem para que o sistema aprenda as características dos documentos e possa classificar um novo documento incorporado ao conjunto (GOLUB, 2005, p. 52). Esta é considerada uma abordagem de "*aprendizado supervisionado*"(SEBASTIANI[*], 2002, p.8).

O *clustering* ou agrupamento (*document clustering*) é uma abordagem de recuperação de informação (da área da Ciência da Informação) e, diferente da técnica anterior, não envolve uso de categorias pré-definidas ou uma "instância de treino" de documentos já classificados manualmente, o que o caracteriza como *não-supervisionado*. Os agrupamentos (*clusters*) e as relações entre eles derivam automaticamente dos documentos a serem agrupados e, posteriormente, os documentos são inseridos nos *clusters*.

GOLUB (2005) denominou, também, de Classificação de Documentos (*Document classification*) a técnica ligada à Ciência da Informação que envolve um vocabulário controlado (uma linguagem documentária) intelectualmente criada e utilizada por um sistema semi-automático que sugere termos de indexação (*Machine-Aided Indexing*). Ainda ressalta uma abordagem mista, na qual as duas primeiras técnicas são combinadas com a terceira, ou seja, o uso de vocabulários controlados em categorização de textos e em *clustering* (GOLUB, 2005, p.19).

Com relação à "Categorização Automática", Farmer (2006) afirma que se trata de uma nova tecnologia feita para lidar com o grande volume de conteúdos digitais não-estruturados, não indexados e "desorganizados". É utilizada conjuntamente com taxonomias e metadados para melhorar o desempenho das ferramentas de busca.

De acordo com a autora (FARMER, 2006, p.93) essas ferramentas desempenham três funções:

[*]Fabrizio Sebastiani dedica-se ao estudo do Aprendizado de Máquina aplicado à Categorização Automática de Textos e é pesquisador do Conselho Nacional de Pesquisa da Itália.

1) Categorização de conteúdos digitais de acordo com uma taxonomia pré-estabelecida.
2) Extração de conceitos e entidades dos documentos para desenvolvimento de uma taxonomia.
3) Extração de metadados dos conteúdos ou extração do conteúdo de *tags* de acordo com um esquema de metadados pré-definido.

Ainda segundo a mesma autora (FARMER, 2006, p.94-95) são três as técnicas de processamento de texto para atribuir um documento a uma categoria:

1) Abordagem baseada em regra – as regras são expressas por especialistas no formato "SE... ENTÃO", como nos sistemas *Machine-Aided Indexing*.
2) Análise estatística – para verificar frequência de palavras, usando também algoritmos de co-ocorrência de termos. Esta abordagem inclui a Categorização de Texto, citada por outros autores anteriormente, que necessita de um conjunto de documentos pré-classificados para que o classificador "aprenda" as regras de inferência.
3) Agrupamento (*Clustering*) linguístico e semântico – esse tipo de tecnologia, considerado pela autora como o mais sofisticado, possibilita a criação de taxonomias e não necessita de documentos pré-classificados. Baseia-se no sentido das palavras para agrupá-las, utilizando instrumentos como tesauros, dicionários, analisadores morfossintáticos, lematizadores, gramáticas etc.

Observa-se, portanto, que há classificadores que se baseiam em técnicas de Processamento de Linguagem Natural para indexação de documentos digitais. Além disso, vê-se uma variedade de técnicas, umas apoiadas no PLN e outras em modelos matemáticos (estatísticos ou probabilísticos), conjuntamente com técnicas de Aprendizado de Máquina.

Em essência, essas técnicas podem ser consideradas dentro do âmbito da automatização da indexação, haja vista a utilização de algoritmos que realizam funções de representação automática do conteúdo de um documento, com o objetivo de armazenamento e/ou recuperação de informação, seja em bases de dados ou na Internet. Além disso, a atividade de indexação pode utilizar uma linguagem documentária ou basear-se em termos em linguagem natural extraídos dos próprios documentos. A Categorização de Textos também se utiliza de um conjunto de categorias definido previamente ou uma taxonomia, assemelhando-se aos processos já praticados pela Biblioteconomia e Documentação.

Com relação à terminologia, verificou-se na literatura a utilização do termo indexação automatizada, sem a menção aos sistemas semi-automáticos (MÉNDEZ RODRÍGUEZ e MOREIRO GONZÁLEZ, 1999), às vezes referindo-se apenas à indexação que requer validação do documentalista (semi-automática) (SILVA e FUJITA, 2004, p.145), bem como seguindo a mesma linha apresentada por Gil Leiva (RODRIGUEZ PEROJO e RONDA LEON, 2006). Por outro lado, ao se referir às abordagens de indexação automática, Hjørland (2008) inclui a indexação semi-automática.

Reconhece-se que a validação dos termos propostos por um sistema semi-automático implica uma outra indexação por parte do documentalista, exigindo o mesmo esforço intelectual necessário na indexação manual ou assistida por computador. No entanto, o sistema em si executa a tarefa de indexação como um sistema automático, com a diferença de que há um processo de verificação ou validação do produto final.

Em seu sentido denotativo, de acordo com o Dicionário Houaiss da Língua Portuguesa, automatizar é “prover de máquinas ou de dispositivos mecânicos ou eletrônicos, para agilização e otimização da produção, dos serviços etc”. Logo, a automatização da indexação é o emprego de dispositivos que agilizam e otimizam o processo de indexação e adota-se aqui a classificação proposta por Gil Leiva, em que a indexação assistida por

computador, a indexação semi-automática e a indexação automática enquadram-se no âmbido da indexação automatizada.

Alguns sistemas baseados em regras têm na correção da indexação por humanos subsídios para a melhoria dos processos. De modo geral, o *feedback* dos indexadores fornece dados que permitem a correção das regras do sistema, bem como fornecem parâmetros de avaliação para melhorar a precisão da indexação. Exemplos desse tipo de sistema são os já citados anteriormente, NewsIndexer e M.A.I. (*Machine Aided Indexer*).

4 HISTÓRIA DA INDEXAÇÃO AUTOMÁTICA

A história da indexação automática foi consistentemente revista por Gil Leiva (1999), portanto decidiu-se basear o histórico do tema em sua obra, recorrendo-se, quando necessário, a outros autores.

Os primeiros passos dados em direção à indexação automática são atribuídos a Hans Peter Luhn, que por volta do final dos anos 1950, durante suas atividades na IBM, propôs que a frequência das palavras em um documento ou cojunto de documentos estaria relacionada com sua utilidade para a indexação.(GIL LEIVA, 1999, p. 64; HJØRLAND, 2008).

Luhn baseou-se nos estudos desenvolvidos por Zipf. Este observou que havia um "princípio do mínimo esforço" na comunicação escrita ou falada, relativo à tendência de repetição de certas palavras ao invés da utilização de palavras diferentes na comunicação oral ou escrita. Analisando a frequência de aparição das palavras, verificou que o produto da frequência pela posição (classificação) da palavra no *ranking,* resultava em uma constante.

Com base no exposto acima, Luhn propôs o primeiro método de indexação automática, que considerava a frequência das palavras dos títulos dos documentos, compondo um índice permutado, chamado KWIC (*Keyword in Context*).

A ideia de um índice KWIC é atribuída por Borko e Bernier a William Frederick Poole com a publicação de "Poole's Index" em 1882 (1978, p.8 citados por SILVA e FUJITA, 2004, p.146). Todavia, sua aplicação em processos automáticos de indexação dá-se a partir das iniciativas de Luhn.

KWIC e suas variantes KWOC (*Keyword out of context*) e KWAC (*Keyword alongside context*) são as inciativas mais simples de indexação automática que baseavam-se em extração de palavras, geralmente dos títulos, e cálculo de sua frequência. (HJØRLAND, 2008; ANDERSON e PÉREZ-CARBALLO, 2001b, p.258). Luhn propunha que as melhores

palavras para indexação seriam as de frequência média e já previa a retirada de palavras vazias como artigos, preposições etc.

Seguindo essa linha estatística da indexação automática, Spärk Jones propôs, em 1972, um método de ponderação de termos, o IDF (*Inverse Document Frequency*), "que mede a escassez de aparição de um termo em uma coleção". Essa forma de ponderação é utilizada atualmente em combinação com a frequência de aparição do termo em um documento (*Term frequency-Inverse Term Frequency* – TF-IDF), em experiências de indexação automática e em recuperação da informação (GIL LEIVA, 1999, p.65; HJØRLAND, 2005).

Outro método de ponderação, também da década de 1970, é o valor de discriminação de termos, proposto por um grupo de investigadores liderados por Gerald Salton. Basicamente, a técnica classificava vocábulos de um texto segundo sua capacidade para diferenciar um documento de outro em uma dada coleção. Segundo este método, são atribuídos pesos aos termos que, quanto mais altos, significam que se referem a termos que causam a máxima separação possível entre os documentos, sendo estes os melhores termos para indexação. Ainda consideravam que, se havia mais de três termos identificando um documento, poderia-se recorrer ao vetor espacial para representar uma coleção (GIL LEIVA e RODRÍGUEZ MUÑOZ, 1996, p.276).

Técnicas não linguísticas, baseadas não só na frequência das palavras, mas se apoiando na probabilidade e relevância de termos, são iniciativas que também surgiram até a década de 1980. Experiências baseadas em referenciais probabilísticos, que consideravam uma base com documentos pré-classificados por indexadores humanos como "exemplos" para o indexador automático inferir regras já tiveram as primeiras iniciativas testadas na década de 80 (GIL LEIVA e RODRÍGUEZ MUÑOZ, 1996).

O emprego de métodos estatísticos contribuiu para o desenvolvimento inicial da indexação automática. No entanto, estavam sujeitos a limitações que influenciam os

resultados da ponderação dos termos. Como limitações, consta que esses sistemas não possibilitavam reconhecer relações semânticas (como o sinônimo de uma palavra); não reconheciam termos compostos, pois não trabalhavam com sintagmas e requeriam a normalização das palavras, pois computavam, por exemplo, a forma singular e plural de um termo como ocorrências distintas.

É nos anos 60 que se inicia a aplicação de técnicas da área de Processamento de Linguagem Natural (PLN) na indexação automática. Segundo Gil Leiva (1999, p.69) as técnicas de PLN são organizadas de acordo com diferentes análises, chegando a enumerar um processamento morfológico, um sintático e um semântico.

Em seu estudo, Gil Leiva (1999, p.77) evidencia que, exceto as primeiras propostas dos anos 60, que eram totalmente baseadas em métodos estatísticos, as propostas posteriores poderiam ser híbridas, considerando:

1) Sistemas estatísticos e PLN;
2) Sistemas estatísticos com a utilização de vocabulário controlado;
3) Sistemas fundamentados em PLN com a utilização de vocabulários controlados;
4) Sistemas que consideramvam as três abordagens anteriores.

Cabe ressaltar aqui a interdisciplinaridade na construção desses sistemas, que une profissionais de PLN, estatísticos e bibliotecários em trabalhos conjuntos.

No Brasil, a aplicação da indexação automática tem seu início no final dos anos 60, com a utilização do programa KWIC para elaborar os índices das bibliografias especializadas publicados pelo Instituto Brasileiro de Bibliografia e Documentação (IBBD), atual Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Nos anos de 1970 surgem os primeiros estudos com a aplicação de leis bibliométricas na indexação automática, com a utilização das Leis de Zipf e de Bradford, assim como a utilização da transição de

Goffman. Posteriormente, foram desenvolvidos estudos já considerando a co-ocorrência de palavras, bem como indexação baseada em raízes vazias e raízes significativas das palavras, como alternativa para otimização do processo de indexação e recuperação, proposta por Jaime Robredo. (VIEIRA, 1988, p.52-53).

Na década de 80 surgem também estudos já baseados em referenciais linguísticos, conjuntamente com uma abordagem estatística, como por exemplo o estudo de Andreewski e Ruas (1983) que trata da adaptação do sistema francês SPIRIT para documentos em língua portuguesa.

Verifica-se, atualmente, o uso de referenciais linguísticos ou de PLN, mais exatamente de critérios sintático-semânticos, tal como a proposta de uso de sintagmas nominais como unidades de análise, presentes nos trabalhos de alguns autores brasileiros (KURAMOTO, 1996; SOUZA, 2006; BORGES, MACULAN e LIMA, 2008).

5 RAZÕES PARA UMA INDEXAÇÃO AUTOMÁTICA

Méndez Rodríguez e Moreiro González (1999, p.4-8) identificam 4 fatores que levaram às experiências de indexação automatizada:

- O alto custo da indexação humana, em termos de tempo, suscitou a ideia de explorar de maneira eficaz, a um custo e tempo reduzidos, o volume constantemente crescente de informação. Essa questão motivou estudos que para comparar a indexação humana e a indexação automática.
- Aumento exponencial da informação eletrônica e a proliferação de textos completos.
- A Gestão Eletrônica de Documentos (GED) e a informatização dos processo documentais.
- A automatização de processos cognitivos e a pesquisa crescente e os avanços em Processamento de Linguagem Natural (PLN). A automatização de processos cognitivos permite o surgimento de sistemas inteligentes, que somados ao PLN, podem lidar com a atividade de indexação. Porém, os autores alertam para complexidade da linguagem e afirmam que um sistema não pode lidar globalmente com ela, sendo capaz apenas de reconhecer cadeias de caracteres.

Os autores chegam a citar a digitalização de documentos e seu manejo por meio de sistemas GED como um dos fatores que fortaleceram os estudos de indexação automática. O aumento da capacidade de memória dos computadores, o avanço nas interfaces gráficas, o advento da Internet e depois da *Web*, dentre outros motivos, também criaram um conjunto de condições que podem facilitar o trabalho dos profissionais de informação. Se por um lado se

presencia um crescimento da produção e disponibilização de informação, por outro, há também o desenvolvimento de tecnologias e aplicativos para seu tratamento.

A subjetividade inerente à indexação humana é outro forte fator levantado em favor da automatização da indexação (GIL LEIVA, 1999, p.61; MÉNDEZ RODRÍGUEZ e MOREIRO GONZÁLEZ, 1999, p.6; BORGES, MACULAN e LIMA, 2008, p.183). A indexação pode variar de um indexador para outro, bem como pode variar a indexação de um mesmo indexador em momentos diferentes. Logo, outro forte argumento em favor dos sistemas automáticos é que eles são mais objetivos, posto que aplicam sempre os mesmos parâmetros para a indexação dos documentos, enquanto o indexador humano está sujeito à variação de humor, além de sua indexação refletir, até inconscientemente, sua visão de mundo, preconceitos e valores.

A riqueza, traduzida pela exaustividade da indexação, é outra característica favorável, embora a indexação humana pareça ser mais precisa. (GIL LEIVA, 1999, p.62; ANDERSON e PEREZ-CARBALLO, 2001a, p.234). Porém, a exaustividade também pode significar alta revocação, fato que interfere na precisão dos resultado de buscas, efeito nem sempre desejável.

Ainda segundo os autores (ANDERSON e PEREZ-CARBALLO, 2001a) a indexação automática parece funcionar tão bem como a indexação humana, mas de maneira diferente. E endossam o baixo custo (com relação ao tempo) desse tipo de indexação, além de sua facilidade de aplicação a grandes conjuntos de documentos (como na Internet), onde o volume de informação cresce constantemente, dificultando a indexação humana.

Importante enfatizar que a indexação automática pode ser vista como um instrumento facilitador da atividade de indexação, não sendo oposta à indexação humana. Nesse sentido, os autores Anderson e Perez-Carballo (2001b, p.270-271) sugerem que a indexação humana poderia ser concentrada nos documentos mais importantes, ressaltando,

por exemplo, as abordagens metodológicas, os pontos de vista ou os valores qualitativos, aspectos que não são facilmente identificáveis por procedimentos automáticos.

Concorda-se aqui com esta abordagem para serviços de informação já constituídos, como as bibliotecas, posto que a indexação automática é uma realidade e pode ser considerada uma solução com relação a grandes volumes de informação. Porém, os sistemas automáticos ainda não lidam satisfatoriamente com a linguagem humana ao ponto de indexar documentos textuais com alto grau de qualidade.

O indexador também pode contribuir nas atividades de construção e avaliação dos sistemas automáticos, postura defendida por Gil Leiva. O autor ainda afirma que, uma vez que este tipo de tecnologia venha a ser aplicada à area de Ciência da Informação, o profissional da informação poderá dedicar-se mais às atividades fins, consequentemente, à difusão da informação, tarefa que constitui sua principal missão (1999, p. 60).

Ainda sobre a indexação automática, Farmer (2006, p. 99-100), quando se refere às ferramentas de categorização automática de documentos digitais, confirma a necessidade da parceria Homem-Máquina para somar à capacidade de processamento de textos dessas ferramentas, a inteligência, julgamentos e experiência humanas. Essa parceria produz melhoria na efetividade das taxonomias e no desempenho dos sistemas. De acordo com a autora, as habilidades humanas podem ser aplicadas à configuração das ferramentas, ao controle de qualidade da indexação (avaliação), à criação das taxonomias, em testes e treinamento dos sistemas e à criação de regras de classificação.

As atividades elencadas pela autora já são práticas correntes de bibliotecários, já tendo portanto, este profissional da informação, instrumentos metodológicos e ferramentas para a execução dessas atividades.

6 A INTERDISCIPLINARIDADE DA INDEXAÇÃO AUTOMÁTICA

A interdisciplinaridade é uma característica inerente à indexação automática. Muitos autores reconhecem não só a interdisciplinaridade como recomendam a criação de grupos interdisciplinares para que se avance nas pesquisas da área (GIL LEIVA, 1999, p.82-83).

Em relação aos sistemas de indexação automática, Lamarca Lapuente (2007) afirma que os mesmos, hoje, norteiam-se pela equação "Linguística + Estatística + Informática + Base de conhecimento", utilizando cada elemento da equação em graus diferentes.

De acordo com Gil Leiva, as áreas que contribuem com a indexação automática de documentos são:

Linguística – Como a indexação lida com a linguagem para a representação dos conceitos, falar em indexação de documentos textuais é falar também no uso de componentes da Linguística que ajudam os sistemas automáticos, por exemplo, a padronizar palavras para contagem (morfologia), desambiguação gramatical (sintaxe) e determinação do sentido de uma palavra (semântica).

Terminologia – A Terminologia tem como principal contribuição o fornecimento de bases para a construção de linguagens documentárias. Relação herdada da Ciência da Informação, pode-se constatar a contribuição da área nos sistemas de indexação automática que utilizam essas linguagens para representação dos documentos.

Informática – Área que permite, desde os anos 50, o tratamento automático da informação e seu armazenamento. Permite não só a indexação automática, como também o armazenamento dos termos de indexação selecionados manualmente.

Linguística Computacional – Trabalha a compreensão da língua e de técnicas apropriadas à sua interpretação, escrita ou falada, tentando imitar a capacidade humana de

compreender textos. Essa área intersdisciplinar, que fica entre a Linguística e a Informática, utiliza elementos de sintaxe, semântica, fonética e fonologia, pragmática e análise do discurso, e pode ser dividida em Linguística de Corpus e Processamento da Língua Natural (PLN). O PLN tem relação direta com a indexação automática, pois se preocupa com o estudo da linguagem para a construção de *softwares* de tradução automática, reconhecedores automáticos de voz, geradores automáticos de resumos, *parsers*, entre outros. É da área da PLN que surgem as tecnologias que permitem à indexação automática a realização de processamentos sintáticos, morfológicos, semânticos e pragmáticos. Hoje ela contribui com a área de Inteligência Artificial na construção de Sistemas Inteligentes.(GIL LEIVA, 1999, p.88; BORGES, MACULAN e LIMA, 2008, p.187).

Estatísticta – A estatística geralmente é aplicada a processos automáticos de indexação com o intuito de calcular a capacidade informativa das palavras, determinada, geralmente, por frequência de aparição nos documentos. Posteriormente, com a PLN, foi possível obter melhores resultados, uma vez que o processamento linguístico contribui para a normalização dos termos e maior correção em sua contagem.

Inteligência Artificial – A área contribui com os "Sistemas Inteligentes", ou seja, sistemas baseados em conhecimento, operando com uma base de conhecimento, que dota o sistema da capacidade de realização de inferências para a resolução de problemas. Uma das formas mais comuns de expressão desse conhecimento é por meio de regras. Atualmente, a área tem contribuído com várias áreas por meio do fornecimento de algoritmos e técnicas de Aprendizado de Máquina, incluindo a indexação automática de documentos.

Como exemplo desse tipo de abordagem é o WADCS (*Web-based automatic document classification system*) criado por Pong *et al* (2007), testado no ambiente de biblioteca, com dois algoritmos da área, o *k-nearest neighbours* (KNN) e *Naïve Bayes,* e usando categorias da Library of Congress Classification (LCC). Exemplo de estudo para

aplicação na tarefa de classificação de páginas *Web* foi relatado por Indra Devi, Rajaram e Selvakuberan (2008).

Outras técnicas e disciplinas que também contribuem para o tratamento e recuperação de informação são:

Mineração de Texto (*Text mining*) – tendo como base a Mineração de Dados, a Mineração de Texto dedica-se à extração de informação de dados não estruturados ou semi-estruturados, ou seja, textos em linguagem natural. Assim como na Mineração de Dados, a área também trabalha com classificação automática de textos e agrupamento (*clustering*), utilizando algoritmos de Aprendizado de Máquina para a construção de seus sistemas. Na classficação de textos, geralmente, o aprendizado é supervisionado e no *clustering* é não-supervisionado.

Um estudo sobre a aplicação de Mineração de Texto aos processos de busca e recuperação de informação de materiais textuais, em língua portuguesa, foi feito por Araújo Júnior e Tarapanoff (2006). Os autores chegaram à conclusão de que o processo poderia ser aplicado como auxiliar da atividade de indexação manual, na melhoria da precisão da indexação.

No Brasil, um exemplo de busca realizada com base em *clusters* é o IAHx, sistema de pesquisa integrado desenvolvido pela BIREME. Esse sistema objetiva, de maneira geral, aperfeiçoar a apresentação dos resultados de buscas da Biblioteca Virtual em Saúde e de sua coleção de fontes de informação, possibilitando a visualização de forma integrada, individualizada e ordenada por diferentes critérios e *clusters*. (BIREME – CENTRO LATINO-AMERICANO E DO CARIBE DE INFORMAÇÃO EM CIÊNCIAS DA SAÚDE, 2008).

A aplicação de técnicas de Aprendizado de Máquina na Classificação Automática de Textos (*Text categorization*) foi bastante estudada por Sebastiani (2002). O autor afirma

que a Categorização de Textos data dos anos 60, mas foi popularizada nos anos 90. Até os anos 80, a abordagem mais popular era baseada na Engenharia do Conhecimento, que consistia em elaborar uma série de regras, sobre como classificar um documento sob uma determinada categoria, alimentadas manualmente por especialistas (base do *Machine-Aided Indexing*). Nos anos 90, cresce a adoção do paradigma do Aprendizado de Máquina que se constitui na construção de um classificador automático, capaz de inferir regras, de acordo com uma base de documentos já pré-classificada. Isto faz com que a Categorização de Textos seja uma disciplina que compartilha elementos do Aprendizado de Máquina e da Recuperação de Informação, além de contribuir a execução de tarefas como extração de conhecimento/informação e mineração de texto.

Das aplicações da Categorização de Textos destacam-se a Indexação Automática para Sistemas de Recuperação de Informação, mecanismos de filtragem de texto (por exemplo para disseminação seletiva de informação), desambiguação do sentido das palavras (*Word sense disambiguation*), e a categorização hierárquica de páginas *Web* (indexação de páginas *Web*).

Alguns métodos para a construção de classificadores automáticos são:

- Probabilísticos (Exemplo: *Naïve Bayes*)
- Árvores de Decisão (não-numéricos, ou seja, "simbólicos")
- Regras de Decisão
- Métodos de Regressão
- Métodos *On-line* (classificadores lineares e Método Rocchio)
- Redes Neurais
- Classificadores baseados em exemplos (Exemplo: "*k-nearest neighbours*")
- *Support Vector Machines*
- *Classifier Commitees* (quando mais de um classificador é utilizado)

Moens* (2000, p.132) também estuda as abordagens baseadas em Aprendizado de Máquina para indexação automática e ressalta a importância do uso de termos de linguagens controladas, pois o conhecimento sobre as palavras e expressões é necessário, exigindo que o conceito esteja presente, seja em um tesauro ou uma base de conhecimentos (base dos Sistemas Inteligentes).

Há que se considerar aqui a complexidade dos indexadores automáticos de Categorização de Textos, que podem ser construídos de acordo com uma grande variedade de métodos e algoritmos, podendo ser automáticos ou semi-automáticos. Um estudo mais qualitativo dos algoritmos faz-se necessário para identificação daqueles de melhor performance em documentos textuais para utilização em indexação automática de documentos em língua portuguesa.

Rodriguez Perojo e Ronda León (2006) propõem a participação de outra disciplina não só na Ciência da Informação no geral, como também na criação de sistemas de indexação automática. Trata-se da Interação Humano-Computador (IHC). Esta disciplina, nascida no contexto da explosão tecnológica da década de 1970, tem como palavra de ordem a interação, procurando assim desenhar, avaliar e implementar sistemas interativos para o uso dos seres humanos.

A IHC pode ser analisada em fução do ***estilo***, ou seja, a forma como o usuário introduz e recebe informação; ***estrutura***, que se refere à forma de organizar os componentes (distribuição dos comandos em janelas ou campos em um formulário); e ***conteúdo***, relativo aos significados semânticos e pragmáticos produzidos no diálogo homem/máquina.

Nesse sentido, a área está ligada à Inteligência Artificial, mais exatamente ao desenvolvimento de Sistemas Inteligentes que simulam o comportamento humano para resolução de problemas. Somadas ao Processamento de Linguagem Natural, as áreas

*Marie-Francine Moens é professora do Departamento de Ciência da Computação da Katholieke Universiteit Leuven (Bélgica), dedicando-se ao estudo de métodos automáticos de indexação e resumo.

contribuem para o desenvolvimento de sistemas de indexação automática, que atualmente possuem componentes dos sistemas inteligentes, como segue:

Base de conhecimentos – que compreende o conhecimento factual (dados) e inferencial (regras) que são introduzidos no sistema com a ajuda de especialistas da área à qual o sistema é voltado.

Ferramenta de inferência – composta de regras e princípios aplicados de forma consistente para garantir a estabilidade e previsibilidade do sistema.

Interface com o usuário – que precisa estar preparada para receber dados e acrescentá-los à memória de trabalho durante a sessão. Acredita-se que aqui a IHC pode ter grande contribuição.

De maneira geral, foi feito aqui um resumo sobre as áreas e disciplinas relacionadas à indexação automática. Cabe enfatizar que, muitas vezes, é difícil tratar algumas disciplinas separadamente, uma vez que muitas colaboram entre si, tal como a Inteligência Artificial, que se baseia em preceitos de PLN para construção de Sistemas Inteligentes, bem como a Mineração de Textos, a Categorização de Textos e o *clustering*.

A interdisciplinaridade dessas áreas também se reflete na composição dos grupos que realizam pesquisas, requerendo cada vez mais profissionais de diferentes áreas para desenvolvimento de projetos, inclusive na área de Indexação Automática que, como visto, vale-se de referenciais de variadas áreas para seu desenvolvimento.

7 INDEXAÇÃO NOS DIAS ATUAIS, INDEXAÇÃO AUTOMÁTICA E INDEXAÇÃO NA INTERNET

O fluxo de recuperação de informação de Lancaster (2004, p.2), descrito no Capítulo 2 – Sobre a Indexação, é pensado com relação a sistemas como as bases de dados tradicionais. Contudo, o autor afirma que o esquema tradicional de um Sistema de Recuperação de Informações pode ser aplicado aos documentos da Internet, embora este não apresente as mesmas características de uma base de dados. Isto se deve ao fato de que, na Internet, qualquer pessoa pode criar uma página, inexistindo um processo de seleção e estruturação de documentos para recuperação (LANCASTER, 2004, p. 5).

Embora alguns sites da *Web* possam incluir algum tipo de dado sobre seu conteúdo (metadados), nem todos o fazem. Os metadados são a maneira encontrada para a estruturação de dados nos recursos da *Web*, uma vez que estes não se encontram centralizados em uma base de dados. Essa "estrutura descritiva" faz parte do próprio documento e possibilita que seus dados sejam manipulados e consultados.

O autor ainda prevê que a "indexação e resumos na Rede, provavelmente serão feitos por processo automáticos, por meio de várias etapas de processamento informatizado." (LANCASTER, 2004, p. 5). O que se torna bastante nítido é que a cada dia há um volume maior e de tendência crescente de informações disponíveis. O tratamento de todo este volume de informação necessita de mecanismos que otimizem sua execução.

Outro fator a ser levantado é que a Internet, que abriga repositórios de informações produzidas de maneira descentralizada, não conta com profissionais de referência (como os que existem em serviços de informação ou biblioteca). Esse papel é desempenhado ou pela ferramenta busca ou pelo próprio usuário, quando clica em um *link* de uma página *Web* que ele supõe atender às suas necessidades, quando navega pelas categorias de um site ou quando elabora uma expressão de busca.

Em algumas bases de dados e serviços de informação *on-line* é possível que o usuário busque pelo termo de uma linguagem documentária, como no caso da BIREME, que tem o DeCS como linguagem que integra todas as fontes de informação disponíveis em seu portal e permite que as buscas sejam feitas por termos DeCS.

Com relação à indexação na Internet, Gil Leiva menciona o que ele chamou de "Universo da Indexação *Web*", confirmando a presença e a necessidade da indexação de documentos nesse meio. Esse Universo foi alcançado por uma "extensão progressiva, tanto de conhecimentos e práticas próprios dos indexadores como dos profissionais da informação e da documentação em geral, com vistas à popularização da Internet" (2007, p.47-48).

Esta extensão dos conhecimentos e práticas é facilmente identificável. Quem desenvolve uma página *Web* ou procura informações na Internet, tanto pessoas como instituições, estão, de fato, exercendo funções ou lidando com conceitos da área da Biblioteconomia e Documentação. Os metadados são exemplos disso. Eles são formas de representação descritiva e temática do documento (ou descrição da forma e do conteúdo). Quem define as *tags* de uma página está, portanto, realizando a atividade de um documentalista. Generalizando, o usuário que insere tags de assunto nos seus conteúdos na *Web* está indexando, bem como os padrões de descrição de dados se assemelham a campos de bancos de dados.

De maneira geral, as *tags* são marcações no próprio texto que "qualificam o objeto do texto" (autor, título, descritores) permitindo que essa informação seja tratável por computador. A separação entre conteúdo, estrutura e estilo permite que o documento seja portável e as linguagens de marcação permitem a estruturação desses documentos (descrição de sua forma e conteúdo). (BAX, 2001; ALMEIDA, 2002).

Nesse sentido, o metadado sempre existiu. Antes ele era estruturado e centralizado em um banco de dados, constituindo os campos do banco. Hoje, ele está em um meio

descentralizado, a Internet, e faz parte da estrutura do próprio documento. Este pode ser apresentado de várias formas, bem como podem ser feitas buscas em seu conteúdo, uma vez que a informação apresenta uma estrutura inteligível a navegadores e sistemas de busca.

Para Gil Leiva (2007), esse Universo de Indexação *Web* ou "Ambiente de Indexação *Web*" está "impregnado" pela indexação e é formado por quatro elementos interrelacionados, a saber:

- Metadados: ordenam e descrevem a informação no documento, do ponto de vista formal e de conteúdo, facilitando seu acesso na Internet.
- O posicionamento *Web*: também chamado de *Search Engine Optimization*, refere-se ao conjunto de técnicas utilizadas pelas ferramentas de busca para o ranqueamento das pesquisas. Cada ferramenta de busca tem um critério para ranqueamento dos resultados. Todavia a utilização das tags ou de palavras significativas na URL (*Uniform Resource Locator*), palavras-chave, títulos, *links* externos, dentre outros, são fatores que contribuem para um bom posicionamento e visibilidade na *Web*. Essa "catalogação" da página fica a cargo da entidade que a produz.
- Buscadores: podem ser um diretório ou uma ferramenta de busca. Os diretórios são organizados manualmente e apresentam uma estrutura de categorias navegável. As ferramentas de busca operam com algoritmos que classificam as páginas do resultado de busca por relevância, de acordo com critérios estabelecidos pelas instituições que produzem a ferramenta. Estes nem sempre são divulgados, mas o que se pode perceber é que além dos *links* internos de uma página, estes buscadores podem verificar também a frequência de uma palavra no texto ou sua posição no documento.

- Usuários: O usuário aqui é visto pelo autor como um "paradocumentalista", pois recorre constantemente à Internet para localização de informações e já está familiarizado com conceitos da área da documentação.

Com todos esses agentes, a tarefa de "organizar" os documentos na *Web* não é fácil, dado o nível de subjetividade no momento de descrever o conteúdo (mesmo havendo uma estrutura de metadados) e dada a diversidade das informações presentes na Internet (estruturadas ou não). Cada entidade "catalogará" suas informações de acordo com seu ponto de vista e este não será necessariamente o ponto de vista do usuário. Quem busca, buscará informações de acordo com sua *praxis* e fica para o buscador a tarefa de mediação entre as duas pontas do sistema.

A subjetividade da indexação vê-se intensificada na Internet. Moens (2000, p.21) alerta que a inserção de marcações em documentos eletrônicos quando considera atributos relativos ao conteúdo (por exemplo, a atribuição de uma *tag* de descritores), pode ser considerada uma indexação manual e pode ser custosa, subjetiva e inconsistente.

Como a Internet é descentralizada, uma boa solução seria aumentar o número de inciativas automáticas. Porém, basear-se apenas no documento não é considerar todos os agentes que contribuem para a indexação, pois ignora o ponto de vista do usuário. Portanto, tecnologias que indexam conteúdos, tanto na Internet como em serviços de informação constituídos formalmente, precisam de alguma avaliação ou validação de seus produtos.

Pensando na Internet, os mecanimsos de busca voltam-se para os documentos, mas o registro da busca feita pelo usuário pode contribuir bastante para a melhoria das buscas e ordenação de seus resultados. O mesmo pode ser considerado para serviços de informação tradicionais, onde a possibilidade de acesso aos *logs* de buscas efetuadas pode fornecer subsídios para avaliação do vocabulário do sistema e da indexação.

Em serviços de informação que utilizam sistemas automáticos de indexação, o problema é que por mais que seja avançado um sistema, este não entende ou interpreta um texto como o ser humano. Assim, parece ser imprescindível a avaliação constante do produto da indexação para verificar se o documento está sendo representado coerentemente, se a indexação está permitindo a recuperação do item ou se a linguagem utilizada precisa de atualizações ou adaptações. Em caso de sistemas que indexam e simultaneamente constróem a linguagem documentária, a supervisão é igualmente necessária.

A fase atual dos sistemas de indexação automática é marcada pela união de referenciais teóricos de PLN e dos "Sistemas Inteligentes", sistemas de indexação apoiados em referenciais da Inteligência Artificial. Méndez Rodríguez e Moreiro González (1999, p.17) dão um panorama da "nova geração de sistemas de indexação automática". Essa nova geração seria caracterizada pelo acesso direto aos documentos por meio de processamento linguístico automático e pela utilização da linguagem natural, combinando técnicas de análise estatística ou ponderação de termos.

Os autores acima afimam que aqui são integrados todos os modelos anteriores (matemáticos e linguísticos) com o intuito de fornecer competências linguísticas e cognitivas às máquinas, baseadas tanto na Linguística como nas bases de conhecimento.

Há a possiblidade de se contar também com interfaces inteligentes que viabilizam a utilização da linguagem natural como linguagem de intercâmbio de "conhecimento" entre o documentalista, o usuário e o sistema.

Com relação às bases de conhecimento, estas podem ser consideradas um tesauro enriquecido com informação morfológica, sintática e semântica, cujo vocabulário é extraído dos documentos de uma área específica do conhecimento.

Como já citado anteriormente, os Sistemas Inteligentes possuem três componentes fundamentais, de acordo com Rodriguez Perojo e Ronda León (2006): a base de

conhecimento, a ferramenta de inferência e a interface com o usuário. As competências necessárias a este sistema, ou seja, o conhecimento da área, são retirados diretamente dos documentos, "do conhecimento que os especialistas colocam neles", um conhecimento pragmático, uma vez que vem da realidade (semântica de mundo), o que contribui também para que a linguagem do sistema esteja atualizada. (MÉNDEZ RODRÍGUEZ E MOREIRO GONZÁLEZ, 1999, p.18-19; LAMARCA LAPUENTE, 2007).

Documentos *Web* que são marcados com alguma linguagem de marcação (como XML) podem servir como uma base de dados, ou seja, pode ser gerida a partir de sua estrutura e com o uso de um programa. Como nem todos documentos possuem esta estrutura marcada, há o desenvolvimento de ferramentas que manipulam esses tipos de dados. Lamarca Lapuente cita sistemas comerciais que indexam de forma automática, mas admite que essas ferramentas não realizam somente as funções de indexação, elas também processam, armazenam e recuperam documentos.

Méndez Rodríguez e Moreiro González (1999, p.14-16), bem como Lamarca Lapuente (2007), resumem quatro processamentos (ou *parsers* linguísticos) sucessivos no PLN:

O primeiro é o **processamento morfológico-léxico,** que tem como principal função obter um léxico que serve como base para as análises posteriores (sintática e semântica), além de fornecer dados coerentes e semanticamente únívocos para uma análise estatística de frequências.

Neste processamento há a segmentação do conjunto de textos em pequenas unidades, realizando uma verticalização das orações e atribuindo-lhes identificadores que serão utilizados como referência nas análises posteriores, marcando-se, assim, não só as palavras, mas os sintagmas, as locuções, siglas, etc. São utilizados como auxiliares dois dicionários, um contendo todas as entradas da língua e outro as locuções e expressões

idiomáticas. Neste processo também pode ocorrer a lematização para a conversão das palavras em sua forma canônica (por exemplo a transformação de verbos conjugados em seu infinitivo, ou substantivos no plural para o singular).

O segundo é o **processamento sintático,** aqui são utilizados dicionários e gramáticas para a descrição da estrutura das orações e separação das unidades linguísticas, bem como desambiguação das categorias gramaticias atribuídas no processamento anterior e realimentação dos dicionários de aplicação. Utilizam-se "analisadores sintáticos" que podem determinar as funções das palavras no texto (sujeito, verbo, etc). As etapas morfológica e sintática podem, também, ser realizadas de uma única vez, com um analisador morfossintático.

O **processamento semântico** é a análise que permite agrupar e hierarquizar o conteúdo do texto por meio de um novo reconhecimento morfológico, que tenha em conta os significados, por meio de reconhecimento de sinônimos e termos genéricos. Pode-se realizar uma análise semântica que estude as relações do termo no contexto da frase ou no documento completo. Posteriormente, pode-se sistematizar os termos (em árvores) que mostrem as relações dos termos dentro do esquema. Nesta etapa, são utilizados tesauros especializados.

O **processamento pragmático** é considerado pelos autores como o mais complexo por não se basear somente no conhecimento linguístico, mas no conhecimento do mundo real (semântica de mundo). Este processamento analisa as relações contextuais, valendo-se de algoritmos que permitem compreender o contexto do discurso.

Uma área mais avançada dessa corrente baseia-se na "Análise Cognitiva do Discurso", com o fim de extrair o que se denomina estrutura fundamental do significado. Para isso, são utilizadas outras técnicas, como a de construção de Redes Semânticas. Este tipo de processamento já tem características de Sistemas Inteligentes.

Como já visto, os sistemas automáticos utilizados para documentos digitais, incluindo páginas da Internet, utilizam-se de algoritmos de aprendizado de máquina, inclusive já com base em PLN, constituindo, de acordo com Farmer (2006, p. 96) a técnica mais sofisticada de ferramentas de categorização automática que já conta com analisadores morfossintáticos, dicionários e tesauros.

Pode ser percebido, de acordo com o exposto até o momento, que as técnicas podem ser utilizadas conjuntamente (não sendo excludentes), permitindo pensar em uma "evolução" dos sistemas de indexação automática que antes eram baseados em abordagens estatíscas mais simples (frequência e ocorrência de palavras). Hoje são caracterizadas por algoritmos complexos e teorias de PLN, que permitem a utilização de linguagem natural no processo de recuperação de informação, em uma união de modelos matemáticos (não linguísticos) de indexação automática e modelos linguísticos.

8 MODELOS DE INDEXAÇÃO AUTOMÁTICA

Méndez Rodríguez e Moreiro González (1999), ao falarem sobre a classificação dos modelos de indexação automática, afirmam que o mais comum é o critério evolutivo, mas que apesar das classificações, os modelos não são excludentes e não tendem a se suplantarem, mas a conviverem e se unirem com um propósito comum que é a obtenção de uma indexação totalmente automática.

Das formas de classificação identificadas, podem ser destacadas, com base em Méndez Rodríguez e Moreiro González (1999) e Lamarca Lapuente (2007):

1) Segundo o métodos de extração terminológica:

Com relação ao método de extração terminológica, este se subdivide em: métodos linguísticos e métodos não linguísticos.

Os métodos lingüísticos envolvem análise do léxico, sintática, semântica e conceitual, com a utilização de ferramentas automáticas. São os processamentos morfológico-léxico, sintático, semântico e pragmático citados anteriormente (Capítulo 7).

Os métodos não linguísticos são aqueles de características quantitativas, baseados em:

- Extração estatística dos termos – por exemplo o método KWIC de Luhn.
- Extração probabilística dos termos – baseada na frequência média de aparecimento dos termos.
- Extração bibliométrica dos termos – baseada na análise quantitativa de determinados termos presentes nos documentos da bibliografia empregada em um campo concreto.
- Extração infométrica dos termos – baseada no tratamento informático dos termos e na engenharia do conhecimento. É o denominado "*data mining*" ou

mineração de dados. Toma-se a liberdade de inserir também a mineração de texto, uma vez que é derivada da mineração de dados.

2) Segundo as partes do documento que indexam: Os sistemas automáticos de indexação são divididos naqueles que indexam apenas as partes principais do documento (título, resumo etc) e os que indexam texto completo.

3) Segundo o controle de vocabulário: Os sistemas são divididos de acordo com a linguagem utilizada pelos sistemas, se liguagens controladas (taxonomias, ontologias, tesauros, listas de cabeçalho de assunto etc) ou linguagens livres (lista de termos livres).

4) Segundo a evolução dos sistemas de indexação automática:

Essa abordagem foi utilizada por Gil Leiva e Rodríguez Muñoz (1996). Méndez Rodríguez e Moreiro González (1999) citam as gerações de sistemas de indexação automatizada para propor uma classificação de acordo com o papel da linguagem natural em cada um deles.

- *1ª geração* – Palavras como objeto: aqui encontram-se os primeiros estudos baseados nos métodos estatísticos e probabilísticos, onde as palavras são entendidas como objetos, e o processamento da linguagem ainda se dá em nível morfológico.
- *2ª geração* – Análise linguística para a desambiguação das palavras: nesta geração já se aplicam as técnicas de Processamento de Linguagem Natural na desambiguação das palavras. Abrange os processamentos morfológico-léxico, sintático, semântico e pragmático, com o intuito de compreender o "significado dos documentos".
- *3ª geração* – "Indexação inteligente" – Sistemas que se apoiam em Sistemas Inteligentes em combinação com os modelos anteriores (modelos estatísticos, probabilísticos etc). Possibilitariam o acesso direto aos documentos por meio

de linguagem natural e a utilização de bases de conhecimento para dotar os sistemas de competência linguística e cognitiva.

Cabe aqui citar também a sugestão de Hjørland (2008) que propõe uma classificação voltada para o ponto de vista epistemológico. O trabalho de Moreiro González (2002) também fornece subsídios para classificação dos modelos de acordo com a teoria matemática da informação. Estes últimos foram apenas citados, sendo importante considerá-los em estudos futuros sobre a classificação dos modelos de indexação automática.

Em síntese, foi percebido, de acordo com as leituras feitas, que os sistemas de categorização automática aplicados em documentos digitais (como páginas da Internet), como exposto, podem basear-se em referenciais de Aprendizado de Máquina, utilizando-se de uma base de documentos já pré-classificada, e por vezes não contam com todos os instrumentos linguísticos dos sistemas baseados em Processamento de Linguagem Natural. Portanto, são considerados neste estudo como Sistemas Inteligentes apenas, com exceção dos sistemas de agrupamento (*clustering*) linguístico e semântico citados por Farmer (2006) que já têm características da terceira geração de sistemas.

Pode ser dito, também, que há diversos modelos de indexação automática atualmente, tendo sido percebido que as mais recentes são as assentadas em Sistemas Inteligentes ou na combinação de Sistemas Inteligentes com o Processamento de Linguagem Natural.

Nesse sentido, considerando uma abordagem evolutiva e as características dessas novas ferramentas, será apresentado, na próxima seção, um quadro geral que possa identificar os métodos que os grupos de pesquisa de universidades públicas brasileiras vêm desenvolvendo.

9 GRUPOS DE PESQUISA NO BRASIL NA ÁREA DE INDEXAÇÃO AUTOMÁTICA

Neste item serão analisados os trabalhos dos grupos de pesquisa brasileiros que se dedicam à indexação automática. Os critérios de análise estão expostos no quadro abaixo.

Quadro 1: Critérios para Classificação dos Modelos de Indexação Automática

Modelo de Indexação Automática	**Descrição**
Sistemas não linguísticos	Inclui as linhas que seguem modelos estatísticos, probabilísticos, bibliométricos e infométricos.
Sistemas linguísticos (PLN)	São as linhas que já consideram um processamento de linguagem natural nos níveis morfológico, sintático e semântico. Por exemplo com a utilização de vocabulários controlados ou o uso dos sintagmas nominais para representação; e sistemas baseados em regras (*Machine Aided-Indexing*).
Sistemas Inteligentes	Sistemas de indexação automática que se baseiam em algoritmos de Aprendizado de Máquina, permitindo a inferência automática das regras para a classificação dos documentos, podendo incluir o uso de um conjunto de documentos pré-classificados manualmente.
PLN + Sistemas Inteligentes	Trata-se da última geração de sistemas de indexação que une todos os modelos existentes, com a utilização de técnicas e instrumentos de Processamento de Linguagem Natural (incluindo os intrumentos de processamento morfológico, sintático, semântico, pragmático para a composição de uma base de conhecimentos).

Sabe-se que algumas das técnicas descritas no Quadro 1 podem estar incluídas em outras. Um exemplo é a categoria chamada aqui “Sistemas Inteligentes” que comporta

sistemas de categorização automática que podem utilizar algoritmos probabilísticos. Um classificador automático também pode contar com recursos linguísticos de PLN (dicionários, lematizadores, analisadores morfossintáticos), podendo se enquadrar na categoria (PLN + Sistemas Inteligentes).

O levantamento dos grupos de pesquisa de universidades brasileiras que se dedicam ao estudo da indexação automática foi feito por consulta à Base Corrente do Diretório dos Grupos de Pesquisa no Brasil do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Para a busca foram utilizadas as expressões "indexação automática", "classificação automática" e "categorização automática". A princípio, considerou-se apenas os grupos de Ciência da Informação, todavia, foram incluídos grupos de Ciências da Computação que realizam estudos e desenvolvem tecnologias na área de Processamento de Linguagem Natural (PLN), Mineração de Texto, Aprendizado de Máquina, voltados para o tratamento de documentos textuais.

Para a definição dos métodos de indexação automática estudados pelos grupos de pesquisa, quando este não estava explícito na descrição do grupo ou em alguma linha pesquisa dele, recorreu-se ao Currículo Lattes do líder para verificação de sua área de atuação, bem como dos trabalhos que tenha publicado recentemente.

Os grupos identificados foram:

- Laboratório de Linguística Computacional – LaLiC
- Modelagem Conceitual para Organização Hipertextual de Documentos – MHTX
- Núcleo Interinstitucional de Linguística Computacional – NILC
- Recuperação Inteligente da Informação
- Representação do Conhecimento, Ontologias e Linguagem

São detalhadas, a seguir, as características de cada grupo.

9.1 LABORATÓRIO DE LINGUÍSTICA COMPUTACIONAL (LALIC)

O LaLiC (Laboratório de Linguística Computacional) é um grupo de pesquisa do Departamento de Computação, Centro de Ciências Exatas e de Tecnologia, Universidade Federal de São Carlos (UFSCar), cuja área principal é "Ciências da Computação".

Formado em 2006, o grupo conta com a lider Lucia Helena Machado Rino, e se dedica a técnicas de PLN na sumarização automática e tradução automática.

O grupo trabalha com uma equipe multidisciplinar de linguistas e cientistas da computação, envolvendo pesquisadores colaboradores, bolsistas de projetos e estudantes dos departamentos de Computação e Letras da UFSCar.

Importante ressaltar que o grupo colabora com outro que será descrito adiante, o Núcleo Interinstitucional de Linguística Computacional (NILC) em projetos de sumarização automática.

Linhas de Pesquisa:

- Construção de Recursos Linguísticos e Computacionais para o PLN
- Geração Automática de Textos
- Mineração de textos
- Modelagem do discurso para o projeto e desenvolvimento de sistemas de PLN
- Sumarização Automática
- Tradução Automática

Observando as linhas de pesquisa do grupo, pesquisadores e estudantes de Ciência da Informação poderiam ser parceiros deste grupo com o intuito de troca de informações para a pesquisa e desenvolvimento de tecnologias para tratamento e recuperação de informações. Sumarização automática, técnicas de PLN aplicadas a essa área, Mineração de Texto e

Tradução automática são exemplos de linhas que seriam muito importantes e úteis aos grupos de pesquisa em organização da informação.

Classificou-se o grupo como de tendência a um modelo de PLN com Sistemas Inteligentes, dada a sua participação no grupo NILC, no desenvolvimento de sumarizadores automáticos.

9.2 Modelagem Conceitual Para Organização Hipertextual De Documentos (MHTX)

Grupo do Departamento de Organização e Tratamento da Informação da Escola de Ciência da Informação, Universidade Federal de Minas Gerais (UFMG) formado em 2004.

O MHTX tem como área predominante a Ciência da Informação, tendo como líder a professora Profª Drª Gercina Ângela Borém de Oliveira Lima, da área de Biblioteconomia.

O grupo faz pesquisas sobre o MHTX (Modelo Hipertextual para Organização de Documentos), sistema proposto em 2004 na tese da líder do grupo, hoje estudado para melhoria dos processos de tratamento e organização de informação.

O sistema é um Mapa semântico Conceitual e Sumário Expandido, ao qual são acrescidos pontos de acesso. Foi instalado em uma base de dados digital de teses e dissertações em texto completo, pertencente à Biblioteca de Teses e Dissertações do Programa de Pós-Graduação em Ciência da Informação da UFMG.

Apesar de não haver nenhuma menção à pesquisa em indexação automática na descrição do grupo, o mesmo foi selecionado para este estudo porque em publicação recente (BORGES, MACULAN e LIMA, 2008) os participantes relataram as bases teóricas para o desenvolvimento de um sistema de indexação automática para fazer parte do protótipo do MHTX. Essa publicação revela uma tendência do grupo a seguir a linha de Processamento de Linguagem Natural (PLN), todavia sem vinculação com Sistemas Inteligentes, pois a pesquisa

do grupo, até onde se pôde verificar, não chega a contar com recursos de Aprendizado de Máquina, como o uso de uma base de conhecimentos construída com base em referenciais de PLN.

Linha de Pesquisa:

- Organização e Uso da Informação (OIU)

9.3 NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL (NILC)

O Núcleo Interinstitucional de Linguística Computacional (NILC) é um grupo do Departamento de Ciências da Computação e Estatística, do Instituto de Ciências Matemáticas e de Computação São Carlos da Universidade de São Paulo (USP), em atuação desde 1993, sob a liderança da Prof.ª Dr.ª Maria das Graças Volpe Nunes.

Tem como área predominante a Ciência da Computação e, apesar de não ser diretamente ligado ao desenvolvimento de pesquisas em Indexação Automática, foi selecionado por trabalhar com PLN e referenciais de Aprendizado de Máquina.

O NILC conta uma equipe multidisciplinar, de linguistas e cientistas da computação para o Processamento de Linguagem Natural (PLN) em português. Além de pesquisadores da USP de São Carlos, o grupo trabalha em parceria com pesquisadores da Universidade Federal de São Carlos (UFSCar) e Universidade Estadual Paulista (Unesp).

Em mais de quinze anos de atuação, o grupo já desenvolveu tecnologias para o processamento de textos em língua portuguesa, como analisadores de discurso, lematizadores, sumarizadores, dicionários etc, que estão disponíveis na página do NILC (http://www.nilc.icmc.usp.br/nilc/).

Algumas iniciativas são destacadas aqui devido à possibilidade de seu uso em pesquisas em tratamento e organização de informação:

- Stemmer –programa que converte as palavras em língua portuguesa para sua raiz, retirando as terminações (flexões de número etc).

- Unitex-PB – projeto que visou à construção de recursos linguístico-computacionais para um sistema de processamento de *corpus* em língua portuguesa. Dentre os recursos desenvolvidos estão um dicionário e uma gramática para resolução de ambiguidades.

- CURUPIRA – *parser* desenvolvido para procesamento morfossintático de texto em língua portuguesa.

- DiZer-PBr – analisador automático de discurso. Inclui *corpus* em língua portuguesa já anotado (RHETALHO) e um segmentador de textos em sentenças (SENTER).

- GistSumm – programa que prodruz resumos automáticos por meio da identificação automática das principais ideias do texto para a construção do resumo.

- Lácio-Web – projeto desenvolvido pelo NILC em parceria com a Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) da USP, cujo objetivo é divulgar e disponibilizar na Internet: *corpus* do português brasileiro escrito contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados em um padrão que possibilite fácil intercâmbio, navegação e análise; e *ferramentas linguístico-computacionais*, tais como contadores de frequência, concordanciadores e etiquetadores morfossintáticos.

O Núcleo desenvolveu também outro projeto: "Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (PLN-BR)"[*] que contava, além das atuais instituições participantes do NILC, com pesquisadores da PUC-RS, Unisinos e Mackenzie e possuía sub-grupos nas linhas "Categorização de Textos" (mais alinhada com Aprendizado de Máquina) e "Sumarização Automática e Recuperação da Informação Textual".

[*] http://www.nilc.icmc.usp.br/plnbr/index.htm

O grupo LaLiC (Laboratório de Linguística Computacional) da UFSCar, também descrito neste estudo, contribui com o NILC nas pesquisas relativas à sumarização automática.

Linhas de Pesquisa:

- Aprendizado de Máquina e PLN
- Construção de Recursos Linguísticos e Computacionais para PLN
- Extração de Informação
- Ferramentas de Auxílio à Escrita
- Ferramentas de Avaliação da Proficiência em Línguas Não Nativas
- Geração de Textos e Sumarização Automática
- Linguística de Corpus
- Redes Complexas e PLN
- Revisão Gramatical Automática do Português do Brasil
- Simplificação Textual
- Terminótica
- Text Mining
- Textos Paralelos e Bilingues
- Tradução Automática

Percebeu-se que muito foi desenvolvido pelo grupo, principalmente na área de desenvolvimento de Corpora e tecnologias para Processamento de Linguagem Natural. Caracterizando-se o grupo em uma tendência de PLN e Sistemas Inteligentes.

Algumas linhas de pesquisa como Aprendizado de Máquina e PLN, Extração de Informação, Geração de Textos e Sumarização Automática, *Text mining* e Tradução Automática poderiam contar com profissionais da Ciência da Informação para o desenvolvimento de tecnologias para tratamento e recuperação de informação.

A união das tecnologias já produzidas pelo grupo, somada à contribuição da Ciência da Informação na área de tratamento e recuperação de informação, poderia resultar em uma base de conhecimentos para sistemas de indexação automática de documentos em língua portuguesa.

9.4 RECUPERAÇÃO INTELIGENTE DA INFORMAÇÃO

Recuperação Inteligente da Informação é um grupo formado em 2004, tendo como área predominante a Ciência da Informação.

Formado por pesquisadores e estudantes da área de Ciências da Informação e Ciências da Computação do Departamento de Ciência da Informação do Centro de Ciências Jurídicas e Econômicas da Universidade Federal do Espírito Santo (UFES), o grupo realiza pesquisas em classificação automática de documentos, baseada em técnicas de Inteligência Artificial, o que pode ser percebido pelas publicações e formação do líder do grupo o Prof. Dr. Elias Silva de Oliveira, bem como pelo perfil das linhas de pesquisa do grupo, pois sistemas inteligentes de classificação automática apoiam-se, geralmemte, em algoritmos de Aprendizado de Máquina.

Linhas de Pesquisa:

- Bibliotecas Digitais
- Classificação Automática de Documentos
- Ferramentas para Apoio ao Ensino
- Visualização da informação

Pela descrição do grupo e perfil das publicações do líder, o que se pôde constatar foi a ausência de pesquisas relativas ao Processamento de Linguagem Natural. Todavia, o grupo já se utiliza de técnicas de Aprendizado de Máquina, podendo ser classificado em uma abordagem de Sistemas Inteligentes.

9.5 Representação do Conhecimento, Ontologias e Linguagem

Grupo do Departamento de Organização e Tratamento da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais (UFMG), formado em 2004, cujos líderes são os Prof. Dr. Renato Rocha Souza e o Prof. Dr. Maurício Barcellos Almeida.

Sua área predominante é a Ciência da Informação e o grupo tem uma orientação para a pesquisa, dentre outros tópicos, para a indexação automática associada ao Processamento de Linguagem Natural.

Linhas de Pesquisa:

- Gestão de Conteúdo e Portais Semânticos
- Instrumentos de Representação do Conhecimento e Ontologias
- Organização e Uso da Informação
- Processamento de Linguagem Natural e Indexação Automática

O grupo já apresenta uma aproximação da indexação automática com o PLN, mas pela descrição do grupo e produção do líder, não foi identificada relação com pesquisas em Sistemas Inteligentes, sendo enquadrado o grupo dentro da categoria de PLN.

9.6 Considerações Gerais Sobre os Grupos de Pesquisa

De maneira geral, os grupos analisados concentram as pesquisas na área de Processamento de Linguagem Natural ou já procuram uma abordagem de Sistemas Inteligentes ou Processamento de Linguagem Natural associado a Sistemas Inteligentes, indicando que há grupos que acompanham as tendências atuais em indexação automática, como pode ser visto no quadro resumo a seguir:

Quadro 2: Grupos de Pesquisa do Brasil e Modelos de Indexação Automática

GRUPO	INSTITUIÇÃO	ÁREA	MÉTODO
Laboratório de Linguística Computacional (LaLiC)	UFScar	Ciência da Computação	PLN+Sistemas Inteligentes
Modelagem Conceitual Para Organização Hipertextual De Documentos (MHTX)	UFMG	Ciência da Informação	PLN
Núcleo Interinstitucional de Linguística Computacional (NILC)	USP/São Carlos	Ciência da Computação	PLN+ Sistemas Inteligentes
Recuperação Inteligente da Informação	UFES	Ciência da Informação	Sistemas Inteligentes
Representação do Conhecimento, Ontologias e Linguagem	UFMG	Ciência da Informação	PLN

Grupos de pesquisa que se baseiam predominantemente em modelos de sistemas não linguísticos não foram encontrados.

Cabe ressaltar que iniciativas que já têm algum software disponível são aquelas voltadas para o processamento de textos em português dos grupos de Ciência da Computação, mais destinadas ao estudo da língua portuguesa do que à recuperação de informação.

Percebeu-se, pelas linhas de pesquisa dos grupos de Ciência da Informação descritas, a preocupação com o tratamento da informação também da Internet, já pensando a questão das bibliotecas digitais, os hipertextos, a classificação automática, ontologias, taxonomias etc.

Como visto na literatura, a interdisciplinaridade é uma característica inerente à Indexação Automática. O desenvolvimento de pesquisas e de *softwares* de indexação automática de documentos textuais em língua portuguesa poderia ser realizado por meio de parcerias entre os grupos estudados.

Os grupos, de uma maneira geral, são constituídos de pesquisadores da Ciência da Computação e da Ciência da Informação ou da Linguística, todavia uma maior multidisciplinaridade das equipes, considerando também profissionais da Matemática, mais

profissionais da Linguística, profissionais da área de Ciência da Informação nos grupos de Ciência da Computação, bem como a manutenção de uma "porta" sempre aberta a novas contribuições, podem levar a experiências mais enriquecedoras.

10 CONSIDERAÇÕES FINAIS

A difusão da informação a um determinado público pode ser considerada a principal missão da Ciência da Informação. No esforço de cumprí-la, os profissionais valem-se de técnicas e intrumentos para o tratamento e organização da informação. Dentre os instrumentos e técnicas estão aqueles relacionados à representação do conteúdo de documentos por meios automáticos: os métodos de indexação automática.

A indexação automática, atualmente, tanto na Internet como em serviços de informação tradicionais, conta com os mais variados modelos. Apesar de não ter sido possível distinguir os métodos especificamente aplicados em bases de dados daqueles aplicados somente na Internet, questão que merece estudos complementares futuros, pôde-se perceber uma tendência ao desenvolvimento de sistemas que combinam técnicas de Processamento de Linguagem Natural (PLN) com Sistemas Inteligentes, resultando em ferramentas dotadas de "conhecimento" que permitem busca em linguagem natural.

Outro fator importante é a interdisciplinaridade da área. Para o desenvolvimento de tecnologias de indexação automática parece ser necessário que a Ciência da Informação busque apoio em outras áreas, compondo grupos de pesquisas interdisciplinares para a realização de projetos conjuntos.

O bibliotecário pode participar de projetos de indexação automática principalmente nas áreas de desenvolvimento, gerenciamento e avaliação dos sistemas, bem como na construção de linguagens documentárias para sistemas que utilizem essa abordagem.

Para verificação do modelo de indexação automática seguido pelos grupos analisados, tentou-se seguir uma organização "evolutiva" baseada nas características dos sistemas atuais identificados na literatura, elaborando-se um quadro geral de modelos. Conclui-se que a maioria dos grupos analisados concentram-se ou em pesquisas na área de Processamento de Linguagem Natural (PLN) ou já procuram uma abordagem de Sistemas

Inteligentes ou PLN com Sistemas Inteligentes. Isso indica que já há grupos que acompanham as tendências atuais em indexação automática, não sendo identificados grupos que se baseiam somente em modelos não linguísticos.

Apenas as inciativas voltadas para o processamento de textos em português nos grupos de Ciências da Computação já possuem *softwares* disponíveis, mas verifica-se a possibilidade de um trabalho conjunto para a exploração desses sistemas para tratamento e recuperação de informação.

A formação de parcerias entre os grupos estudados para o desenvolvimento de pesquisas e de *softwares* de indexação automática de documentos textuais em língua portuguesa poderia ser uma boa oportunidade para troca de experiências e união de recursos e forças para o avanço das pesquisas na área.

REFERÊNCIAS

ALMEIDA, Maurício Barcellos. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. **Ciência da Informação**, Brasília, v. 31, n. 2, p.5-13, 2002. Disponível em: < http://revista.ibict.br/ciinf/index.php/ciinf/article/view/140/120 >. Acesso em: 15 out. 2008.

ANDERSON, J. D.; PEREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval: Part 1: Research, and the nature of human indexing. **Information Processing and Management**, v. 37, n. 2, p.231-254, Mar. 2001a.

_____. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part 2: Machine indexing, and the allocation of human versus machine effort. **Information Processing and Management**, v. 37, n. 2, p.255-277, Mar. 2001b.

ANDREEWSKI, Alexandre; RUAS, Vitoriano. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ciência da Informação**, Brasília, v. 12, n. 1, p. 61-73, 1983. Disponível em: < http://revista.ibict.br/index.php/ciinf/article/view/1550/1167 >.

ARAÚJO JÚNIOR, Rogério Henrique de; TARAPANOFF, Kira. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. **Ciência da Informação**, Brasília, v. 35, n. 3, p.236-247, 2006. Disponível em: < http://revista.ibict.br/index.php/ciinf/article/view/786/643 >.

AUTOMATIZAR. In: DICIONÁRIO Houaiss da Língua Portuguesa. [s.l]:[Instituto Antonio Houaiss], [2009?]. Edição eletrônica para assinantes Uol. Disponível em: < http://houaiss.uol.com.br/busca.jhtm?verbete=automatizar&stype=k >. Acesso em: 12 dez. 2008.

BAX, Marcello Peixoto. Introdução às linguagens de marcas. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 32-38, jan./abr. 2001. Disponível em: < http://revista.ibict.br/ciinf/index.php/ciinf/article/view/221/196 >. Acesso em: 15 out. 2008.

BIREME – CENTRO LATINO-AMERICANO E DO CARIBE DE INFORMAÇÃO EM CIÊNCIAS DA SAÚDE. **IAHx:** recuperação de informação baseada em clusters. [São Paulo], 2008. Disponível em: < http://wiki.reddes.bvsalud.org/index.php/IAHx_-_Recupera%C3%A7%C3%A3o_de_informa%C3%A7%C3%A3o_baseada_em_clusters >. Acesso em: 10 dez. 2008.

BORGES, Graciane Silva Bruzinga; MACULAN, Benildes Coura Moreira dos Santos; LIMA, Gercina Angela Borem de Oliveira. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade**: Estudos, v. 18, n. 2, p.181-193, 2008. Disponível em: < http://www.ies.ufpb.br/ojs2/index.php/ies/article/download/1759/2129 >. Acesso em: 15 fev. 2009.

CAMPOS, Maria Luiza de Almeida; GOMES, Hagar Espanha. Taxonomia e Classificação: o princípio de categorização. **DataGramaZero**: Revista de Ciência da Informação, v. 9, n. 4, ago. 2008. Disponível em: < http://dgz.org.br/ago08/F_I_art.htm >. Acesso em: 2 mar. 2009.

CINTRA, Anna Maria Marques et al. **Para entender as linguagens documentárias**. 2. ed. rev. ampl., 1. reimp. São Paulo: Polis, 2005. 92 p. (Coleção Palavra-Chave, 4).

CLEVELAND, Donald B.; CLEVELAND, Ana D. **Introduction to indexing and abstracting.** 2nd. ed. Englewood:Libraries Unlimited, 1990. 329 p.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. **Diretório dos Grupos de Pesquisa no Brasil** [base de dados]. [Brasília], 2009. Disponível em: < http://dgp.cnpq.br/buscaoperacional/ >. Acesso em: 30 jun. 2009.

FARMER, Linda. Automatic categorization: what's it all about?. **Serials Librarian**, v. 51, n. 2, p.91-101, 2006.

GIL LEIVA, I.; RODRÍGUEZ MUÑOZ, J. V. Tendencias en los sistemas de indización automática: estudio evolutivo. **Revista Española de Documentación Científica**, 1996, v. 19, n. 3, p 273-291.

GIL LEIVA, Isidoro. A indexação na Internet. **Brazilian Journal of Information Science**, v.1, n.2, p.47-68, jul./dez. 2007. Disponível em: < http://www.bjis.unesp.br/pt/include/getdoc.php?id=72&article=21&mode=pdf >. Acesso em: 20 out. 2008.

GIL LEIVA, Isidoro. **La automatización de la indización de documentos**. Gijón (Astúrias): Eciciones Trea, 1999. 220 p.

GOLUB, Koraljka. **Automated subject classification of textual Web pages, for browsing**. Lund: Lund University, Department of Information Technology, 2005. 139 p. Disponível em: < http://www.it.lth.se/koraljka/Lund/publ/LicE.pdf >. Acesso em: 15 jan. 2009.

HJØRLAND, Birger. Automatic Indexing. In: _____. **Lifeboat for Knowledge Organization**. [s.l.]:[s.n.], 2008. Disponível em: < http://www.db.dk/bh/lifeboat_ko/CONCEPTS/automatic_indexing.htm >. Acesso em: 5 dez. 2008.

_____. **Core Concepts in Library and Information Science (LIS)**. [s.l.]:[s.n.], 2005. Disponível em:< http://www.db.dk/bh/Core%20Concepts%20in%20LIS/home.htm >. Acesso em: 5 dez. 2008.

HLAVA, Marjorie M. K. NewsIndexer: machine-aided indexing customized for the news industry. In: SCHROEDER, Sandi (Ed.). **Software for indexing**. Medford, NJ: American Society of Indexers, 2003. p.253-261.

INDRA DEVI, M.; RAJARAM, R.; SELVAKUBERAN, K. Generating best features for Web page classification. **Webology**, v. 5 n. 1, article 52, mar. 2008. Disponível em: < http://www.webology.ir/2008/v5n1/a52.html >. Acesso em: 15 out. 2008.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, p.1-18, 1996. Disponível em: < http://revista.ibict.br/index.php/ciinf/article/view/435/393 >. Acesso em: 15 abr. 2009.

LAMARCA LAPUENTE, María Jesús. Indización automática. In:_____. **Hipertexto**: El nuevo concepto de documento en la cultura de la imagen. Tesis doctoral - Universidad Complutense de Madrid, 2007. Disponível em: < http://www.hipertexto.info/documentos/indiz_automat.htm >. Acesso em: 27 out. 2007.

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

LARA, Marilda Lopes Ginez de. Algumas contribuições da semiologia e da semiótica para análise das linguagens documentárias. **Ciência da Informação**, Brasília, v. 22, n. 3, p. 223-226, set./dez. 1993. Disponível em: < http://revista.ibict.br/ciinf/index.php/ciinf/article/download/1129/778 >. Acesso em: 26 set. 2008.

MÉNDEZ RODRÍGUEZ, Eva M.; MOREIRO GONZÁLEZ, José A. Lenguaje natural e indización automatizada. **Ciencias de la Información**, v. 30 , n. 3, p.1-23, sept. 1999. Disponível em: < http://www.bib.uc3m.es/~mendez/publicaciones/articulos/indizacion99.pdf >. Acesso em: 15 maio 2008.

MOENS, Marie-Francine. **Automatic indexing and abstracting of document texts**. Boston : Kluwer Academic Publishers, c2000. 265 p. (The Kluwer international series on information retrieval, 6). Disponível em: < http://site.ebrary.com/lib/usp/Doc?id=10046957 >. Acesso em: 2 mar. 2009.

MOREIRO GONZÁLEZ, José Antonio. Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información. **Anales de documentación**, n. 5, p.273-286, 2002. Disponível em: < http://revistas.um.es/analesdoc/article/viewFile/2101/2091 >. Acesso em: 15 maio 2009.

PONG, Joanna Yi-Hang et al. A comparative study of two automatic document classification methods in a library setting. **Journal of Information Science**, v. 34, n. 2, p. 213-230.

REDMOND-NEAL, Alice. NewsIndexer: machine-aided indexing customized for the news industry. SCHROEDER, Sandi (Ed.). **Software for indexing**. Medford, NJ: American Society of Indexers, 2003. p.247-251.

RODRIGUEZ PEROJO, K.; RONDA LEON, R. Organización y recuperación de la información: un enfoque desde la perspectiva de la automatización. **ACIMED**, Habana, v. 14, n. 1, ene./feb., 2006. Disponível em: < http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352006000100004&lng=es&nrm=iso&tlng=es >. Acesso em: 15 maio 2008.

SEBASTIANI, F. Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, v. 34, n. 1, p.1-47, 2002. Disponível em: < http://citeseer.ist.psu.edu/article/sebastiani99machine.html >. Acesso em: 15 maio 2008.

SILVA, M. R. da; FUJITA, M. S. L. A prática de indexação: análise da evolução e tendências teóricas e metodológica. **TransInformação**, Campinas, v. 16, n. 2, p.133-161, 2004. Disponível em: < http://revistas.puc-campinas.edu.br/transinfo/include/getdoc.php?id=196&article=65&mode=pdf&OJSSID=3bcd6d818e45ebfecdcb30215f9b0c5b >. Acesso em: 15 jan. 2009.

SOUZA, Renato Rocha . Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, n. esp, p.42-59, 1. sem. 2006. Disponível em: < http://www.periodicos.ufsc.br/index.php/eb/article/view/329/385 >. Acesso em: 15 abr. 2009.

VIEIRA, Simone Bastos. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, Brasília, v. 17, n. 1, p.43-57, jan./jun. 1988. Disponível em: < http://revista.ibict.br/index.php/ciinf/article/viewPDFInterstitial/1391/1017 >. Acesso em: 20 abr. 2009.

WORLD HEALTH ORGANIZATION. **International Classification of Diseases (ICD)**. [s.l.], [2009?]. Disponível em: < http://www.who.int/classifications/icd/en/ >. Acesso em: 25 maio 2009.