PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
MESTRADO EM CIÊNCIA DA INFORMAÇÃO
UNIVERSIDADE FEDERAL FLUMINENSE -INSTITUTO DE ARTE E COMUNICAÇÃO SOCIAL

EDUARDO BARÇANTE

PROPOSTAS E METODOLOGIAS DE PROCESSAMENTO AUTOMÁTICO DE DOCUMENTOS TEXTUAIS DIGITAIS: UMA ANÁLISE DA LITERATURA

Niterói
2011

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - UFF

MESTRADO

Área de Concentração:

Dimensões contemporâneas da informação e do conhecimento

Linha de Pesquisa:

Fluxos e Mediações Sociotécnicas da Informação

Projeto de Pesquisa de Dissertação

PROPOSTAS E METODOLOGIAS DE PROCESSAMENTO AUTOMÁTICO DE DOCUMENTOS TEXTUAIS DIGITAIS: UMA ANÁLISE DA LITERATURA

por

Eduardo Barçante

Projeto de Pesquisa de Dissertação apresentado para o Curso de Pós-Graduação em Ciência da Informação - UFF, como parte dos requisitos para aprovação no exame público de qualificação de Mestrado em Ciência da Informação.

Orientador: Professor Carlos Henrique Marcondes D. Sc.

Niterói

Maio de 2011

Barçante, Eduardo

Propostas e metodologias de processamento automático de documentos textuais digitais: uma análise da literatura / Eduardo Barçante. -- Niterói: UFF / 2011.

100 f.

Orientador: Carlos Henrique Marcondes

Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense / Programa de Pós-Graduação em Ciência da Informação, 2011.

Referências bibliográficas: f. 81- 88

1.Reuso. 2.Mineração de dados. 3.Texto digital. 4.Data-mining.

5. Ciência da Informação – Tese. I. Marcondes, Carlos Henrique (Orient.). II. Universidade Federal Fluminense, Ciência da Informação. III. Título.

EDUARDO BARÇANTE

PROPOSTAS E METODOLOGIAS DE PROCESSAMENTO AUTOMÁTICO DE DOCUMENTOS TEXTUAIS DIGITAIS: UMA ANÁLISE DA LITERATURA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação, UFF - Universidade Federal Fluminense, como parte dos requisitos necessários à obtenção do título de Mestre em Ciência da Informação.

_____

Carlos Henrique Marcondes, D.sc – UFF

_____

Maria Cristina Soares Guimarães, D.sc – FIOCRUZ

_____

Maria Luiza de Almeida Campos, D.sc – UFF

_____

Regina Cianconi, D.sc – UFRJ

Niterói
2011

Mas que universo é este? Que universo
entre muitos possíveis pelo tempo
acolhe o pensamento, o imaginário
a estruturar as linhas do poema?
Não percebo de todo, não entendo
o universo que jaz sob este canto.
Fernando Py, Antiuniverso, p. 53

RESUMO

BARÇANTE, Eduardo. **PROPOSTAS E METODOLOGIAS DE PROCESSAMENTO AUTOMÁTICO DE DOCUMENTOS TEXTUAIS DIGITAIS: UMA ANÁLISE DA LITERATURA**. Niterói, 2011. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense.

Em domínios interdisciplinares como Comunicação-Informação em Saúde, o surgimento da Web vem trazendo uma crescente oferta de documentos digitais diversos, como artigos científicos, notícias, legislação, manuais, normas, etc., de interesse.potencial. Dada a grande quantidade e a dispersão destes documentos por diferentes fontes, seu tratamento automático com vistas ao reuso e recontextualização segundo os interesses e semânticas de um domínio específico é de grande interesse. Esta pesquisa teve como objetivo investigar, a capacidade de identificar e analisar métodos de extrair automaticamente semânticas específicas a partir de textos digitais com objetivo de reutilizá-los para outros fins diferente dos quais estes foram inicialmente produzidos. Para tanto, foram levantados e classificados artigos científicos buscando responder as seguintes questões: Em que conjunto de dados textuais o método descrito no artigo foi aplicado? e como foi especificada a semântica a ser buscada no conjunto de dados textuais?. Após a análise, para cada texto identificado no levantamento emergiram as seguintes classes de métodos: Mineração de textos, Anotação Semântica, Análise Semântica, Análise em Linguagem Natural e Tratamento Estatístico de textos. Apresenta-se uma relação sistemática onde são descritas as características gerais de cada método, da classe formada por ele, e os artigos que compõe cada classe são discutidos e comentados. Espera-se que a pesquisa possa subsidiar propostas de sistemas de tratamento automático de textos publicados na Web com vistas a seu reuso e recontextualização.

Palavras-chave: Recuperação de informação; Mineração de dados; Interface de busca; Documento digital.

ABSTRACT

BARÇANTE, Eduardo. **PROPOSALS AND METHODOLOGIES OF AUTOMATIC PROCESSING OF DIGITAL TEXTUAL DOCUMENTS: A LITERATURE ANALYSIS**. Niterói, 2011. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense.

Inter-disciplinary fields such as Communication and Health Information, the emergence of the Web has brought an increasing availability of digital documents as diverse as scientific articles, news, legislation, manuals, standards, etc.., of potential interest. Given the large number and dispersion of these documents from different sources, their automatic treatment in order to reuse and recontextualization in the interests and semantics of a specific domain is of great interest. This study aimed to investigate the ability to identify and analyze methods for automatically extracting specific semantics from digital texts in order to reuse them for other purposes than that which they were first produced. It had been collected and classified papers seeking to answer the following questions: Which set of textual data with the method described in the article was applied? and semantics as specified was to be sought in the set of textual data?. After the analysis, for each text identified in the survey yielded the following classes of methods: Text mining, Semantic Annotation, Semantic Analysis, Natural Language Analysis and Statistical Treatment of texts. It presents a systematic relationship which describes the general characteristics of each method, the class formed by him, and items that compose each class are discussed and commented. It is hoped that the research will support proposals for systems of automatic processing of texts published on the Web with a view to their reuse and re-contextualization.

Keywords: Information Retrieval; Datamining; Search Interface; Digital document.

LISTA DE FIGURAS, TABELAS E QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT – Associação Brasileira de Normas Técnicas

ASCII – *American Standard Code for Information Interchange*

CI – Ciência da Informação

EI – Extração de Informação

HTML – *Hypertext Markup Language*

ISO – *International Standardization Organization*

KDD – *Knowledge Discovery in Databases*

KDT – *Knowledge Discovery in Texts*

OBO – *Open Biomedical Ontologies*

OWL – *Ontology Web Language*

PDF – *Portable Document Format*

RDF – *Resource Description Framework*

RDFS – *Resource Description Framework Schema*

SGML – *Standard Generalized Markup Language*

SQL – *Structured Query Language* ou Linguagem de Consulta Estruturada

SWRL – *Semantic Web Rule Language*

TIC – Tecnologias de Informação e Comunicação

URI – *Uniform Resource Identifier*

W3C – *World Wide Web Consortium*

WEB – *World Wide Web*

XML – *Extend Markup Language*

SUMÁRIO

1 INTRODUÇÃO

> "Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully" (BERNERS-LEE et al., 2001).

As ciências da vida, e mais especificamente as descobertas que redundaram na genômica, são apontadas como grandes orientadores de uma nova dinâmica socioeconômica das nações: a saúde é reconhecida como uma dimensão necessária, um requisito para o desenvolvimento, e não sua consequência. Os *Objetivos de Desenvolvimento do Milênio para 2015*, estratégia pactuada com a Organização das Nações Unidas (ONU), representam um imenso desafio para os países, que terão que organizar seus sistemas de pesquisa em saúde com uma orientação que procure assegurar a incorporação dos resultados às ações de saúde. Nesse sentido, devem ser buscados, cada vez mais, mecanismos e estratégias que possibilitem que os avanços na área de pesquisa biomédica possam ser mais rapidamente difundidos, testados, avaliados, utilizados e assimilados na base de conhecimento corrente, gerando inovações e consequentes melhorias na saúde.

Entretanto, os desafios que acompanham essa agenda política não são poucos e resultam de culturas e dinâmicas que nascem em diferentes áreas. A aceleração da dinâmica científica vem sendo acompanhada por um crescimento geométrico na produção de dados biomédicos e no volume e quantidade de literatura científica publicada. Não existe um número ou um cálculo oficial sobre esse quantitativo. Jinha (2010) afirma que qualquer resposta à pergunta "quantos artigos foram publicados até hoje?" é complicada de ser fornecida porque implica dizer o que se entende e como se define "periódico", "artigo", "publicado" e "até hoje", especialmente depois do advento da *web*. Segundo o referido autor, o *PubMed*[1] Central, um repositório de acesso livre, continha cerca de 1,7 milhão de artigos em 2009. O *Pubmed* teria cerca de 19 milhões de referências e cresceria à taxa de uma referência por minuto! *Scopus* e *ISI Web of Science* registrariam cerca de 40 milhões de artigos. Jinha (2010) finaliza propondo um total de 50 milhões de artigos publicados desde o lançamento do *Journal des Savants*, há 350 anos, sendo que as ciências biomédicas respondem por cerca de 40% desse total.

---

[1] Serviço da Biblioteca Nacional de Medicina dos Estados Unidos.

Para além do crescimento quantitativo, a pesquisa biomédica cresceu em especialidades e subespecialidades. Biólogos reconhecem, por exemplo, que ignoram grande parte da literatura-chave de outras subespecialidades da própria biologia. Não é trivial, portanto, o esforço de aproximar a “pesquisa de bancada” da “pesquisa no leito do paciente”. Todo conhecimento e informação produzidos que digam respeito aos genes, proteínas, doenças, drogas e seu papel nos processos biológicos são publicados na literatura. Se a abundância de ambos os dados biológicos e literários já é, por si só, um gargalo para interpretação e planejamento em larga escala de experimentos, o desafio é maior quando se sabe que esta inter e multidisciplinaridade implica diferentes tipologias de dados e informação, armazenados em diferentes fontes, com linguagens diferentes.

Esses desafios devem ainda ser considerados sob uma lógica de produção de novo conhecimento que é uma atividade intelectual essencialmente humana. Teorias e observações orientam novas hipóteses, que são testadas experimentalmente em função de um conhecimento prévio. Os resultados alcançados são publicados e comunicados aos pares, criticados e assimilados em novos experimentos e teorias. Esse ciclo entre teoria e prática forma uma rede socialmente complexa que se desenvolve e evolui em um tempo cronológico próprio. A avaliação por pares ainda é a principal estratégia e fonte para ter acesso a novo conhecimento, e o contínuo crescimento e diversificação da literatura vai demandar um grande esforço de sistematização de forma a tentar maximizar o uso da informação disponível.

Para todos os desafios colocados, as tecnologias de informação e comunicação – TICs, mais particularmente a *Web* e a Internet − despontam como uma alternativa e uma promessa para contornar, se não resolver, os problemas. Por um lado, a maioria dos periódicos da área da ciência, tecnologia e medicina (*STM* em inglês) possuem, na atualidade, versão digital, com um percentual significativo deles em acesso livre, o que em muito facilita o acesso e estimula o processo de disseminação do conhecimento e a interoperabilidade entre sistemas e repositórios (RENEAR; PALMER, 2009). Por outro lado, no cerne do processo de geração de hipótese e “descoberta” de novo conhecimento, as tecnologias e as metodologias ainda encontram-se em processo de maturação.

A Gartner Consult[2] registrou, em relatório de 2002, que pelo menos por mais uma década a principal via de relacionamento entre homem-máquina será realizada por meio de documentos em linguagem textual. Isso significa que homens e máquinas ainda operam de forma diferente. A *Web* atual é denominada por Breitman (2005) de *Web Sintática,* na qual os computadores fazem apenas a apresentação da informação enquanto o processo de interpretação fica a cargo dos seres humanos, já que isso exige um grande esforço para avaliar, classificar e selecionar informações e conhecimentos de interesse. Contrapondo essa *Web Sintática*, surge a *Web Semântica*, através da qual se buscam mecanismos que capturem o significado das páginas, criando um ambiente no qual os computadores possam processar e relacionar conteúdos provenientes de várias fontes. Para que isso se torne possível, é necessário embutir semântica na estrutura dos documentos disponíveis na *Web*. Enquanto a *Web Sintática* foi desenvolvida para ser entendida apenas pelos usuários, a *Web Semântica* está sendo projetada para ser compreendida pelas máquinas, na forma de "agentes" computacionais que serão capazes de operar eficientemente sobre as informações, podendo até entender (inferir) seus significados. Assim, esses agentes auxiliarão os usuários em suas diversas operações na *Web*.

Essa nova concepção de *Web* possibilita a substituição dos tradicionais sistemas de recuperação de informação baseados simplesmente em palavras-chave por avançados recursos capazes de manipular e realizar inferências, reconhecendo relações no conteúdo dos textos. É sob essa nova concepção de *Web* que surgem as abordagens de Extração de Informação (EI), as quais não apenas localizam informações para o usuário, mas também são capazes de compreender o conteúdo de textos em linguagem natural. Para um humano, essa tarefa é relativamente simples. No entanto, para que um sistema computacional seja capaz de simular EI, é necessária a construção de regras que orientem a relevância dos conteúdos/informação. Claro que não é tarefa simples implementar essa abordagem, então faz-se uso de ontologias, vocabulários, listas de palavras e outras formas para

[2] Empresa de consultoria que atua no segmento da Tecnologia e Informação. Disponível em: <http://www3.gartner.com/DisplayDocument?id = 379859>.

inúmeras relações e outras formas para enriquecer o número de vocábulos, acelerar, intensificar a leitura e finalmente extrair informação (RENEAR; PALMER, 2009).

Extração de Informação (EI) é uma forma de análise de linguagem natural e está se tornando uma tecnologia central para diminuir a distância entre um texto não estruturado e conhecimento formal expresso em ontologias. Os métodos de processamento de linguagem natural fornecem a fundamentação para as investigações no campo da mineração de textos biomédicos (ZWEIGENBAUM *et al*. 2007).

Na área biomédica, Cohen e Hunter (2008) fazem uso de técnicas baseadas em regras que reusam algum tipo de conhecimento ou classificadores para analisarem sentenças e documentos.

O MEDLINE, base de dados da literatura internacional da área médica e biomédica, é explorado por meio dos seus resumos, onde a informação é extraída com o uso de marcadores opcionais que rotulam classes gramaticais. (CHUN, H. et al. 2005; ZHOU, 2004).

Além do MEDLINE, há outras iniciativas e experiências relatadas em *workshops* e conferências em áreas do conhecimento envolvendo grupos de pesquisas em Estatística e Matemática (*Society for Industrial and Applied Mathematics* (SIAM), 2008), e também em grupos da ciência da computação na *Industrial Conference on Data Mining (ICDM)* (ICDM, 2010).

Assim, tem-se como apoio o desenvolvimento tecnológico, principalmente na área de *softwares* que agilizam o processamento de textos digitais com o objetivo de dar algum sentido ao seu conteúdo textual digital (GROSSMAN, R.; KARMATH, C.; KUMAR, 2001) permitindo seu processamento por programas. O objetivo de dar sentido ao conteúdo textual digital tem a finalidade de reutilizá-los para outro propósito diferente daquele para o qual esse texto foi originalmente criado.

Há várias definições para um sistema de mineração de textos. No sentido restrito, um sistema de mineração de texto deve ser capaz de identificar um conhecimento que não está explícito no texto. Em sentido mais amplo, diz respeito a qualquer sistema que faça a extração de informação ou englobe requisitos que o façam

(HEARST, 1999; ZWEIGENBAUM et al. 2007). A extração de fatos explícitos de textos científicos foi o principal objetivo dos primeiros desenvolvimentos de mineração orientados para os textos em Biologia Molecular (KRALLINGER; VALENCIA, 2005).

Entretanto, se o grande desafio é acelerar a produção de novo conhecimento, ou novas inferências e relações entre conceitos, a EI não avança muito porque as máquinas ainda não podem fazer isso; elas apenas identificam estruturas semânticas no texto, fazem a extração e relacionam fatos que já foram publicados/registrados. Cabe ressaltar a pesquisa inovadora de Don Swanson (1986; 1988) que, há mais de três décadas e por meio de métodos semi-automáticos, desenvolveu uma abordagem de descoberta de novos conhecimentos na literatura. O caminho é usar os fatos que foram extraídos de várias publicações (exemplo: A leva a B, e B leva a C) para inferir novas relações indiretas (A leva a C). Como a literatura é tão vasta que cada pesquisador só consegue ler um pequeno subconjunto, é provável que nenhum deles esteja ciente de todas as evidências que são necessárias para fazer essa inferência lógica. A metodologia de Swanson foi capaz de antecipar relações que somente anos mais tarde foram comprovadas em experimentos – óleo de peixe é benéfico para pacientes com Doença de Raynaud, e deficiência de magnésio exerce influência na enxaqueca. Ocorre, entretanto, que para Swanson essa inferência é uma comprovação de uma hipótese lançada pelo usuário que é testada nas buscas de informação. Pode-se assim argumentar que, nesse caso, o computador realmente não fez a “descoberta”, mas a comprovou.

Propostas como as de Swanson só corroboram a afirmação de que novas descobertas e, consequentemente, avanços da Ciência podem se dar identificando novas relações na literatura. No entanto, há um nível de complexidade muito maior para o problema; Attwood et al. (2009) afirmam:

> There is so much information available that we simply no longer know what we know, and finding what we want is hard – too hard. The knowledge we seek is often fragmentary and disconnected, spread thinly across thousands of databases and millions of articles in thousands of journals. The intellectual energy required to search this array of data-archives, and the time and money this wastes, has led several researchers to challenge the methods by which we traditionally commit newly acquired facts and knowledge to the scientific record. (ATTWOOD et al., 2009, p. 317).

Como já foi visto, atualmente, essa literatura científica já é publicada diretamente em formato digital e cresce enormemente. Tudo indica que essa quantidade de literatura só poderá ser tratada com o aporte das TIs.

Os sistemas de EI usam técnicas de processamento de linguagem natural, dentro de um domínio de conhecimento, em busca de padrões em documentos que levem a “fatos” que possibilitem fazer inferências. Esses “fatos” são identificados, extraídos e registrados em formatos processáveis por programas que podem permitir o seu reuso em outros contextos. Corney et al. (2004) observam que várias tentativas têm sido feitas para usar a EI em textos científicos, mas que ficaram restritas ao resumo dos artigos que, no geral, são bem estruturados, não têm subescritos, figuras, notas de rodapé, o que evita a interpretação de símbolos e letras gregas, por exemplo. Essa estruturação limita, por outro lado, a riqueza semântica e o detalhamento do texto completo. Zweigenbaum et al. (2007) ressaltam que um dos grandes desafios do futuro é a mineração de dados em textos completos.

Shatkay (2005) observa que o tratamento automatizado de texto é um campo de pesquisa multidisciplinar, abrangendo várias áreas do conhecimento e suas especialidades, como o processamento de linguagem natural (PLN) e o seu domínio mais específico de EI, aos quais se juntam a Ciência da Informação e sua preocupação com a organização e recuperação da informação. Saracevic (1996) descreve assim uma das forças que modelaram o nascimento da Ciência da Informação:

> Dentre os eventos históricos marcantes, o ímpeto de desenvolvimento e a própria origem da CI podem ser identificados com o artigo de VANNEVAR BUSH, respeitado cientista do MIT e chefe do esforço científico americano durante a Segunda Guerra Mundial (BUSH, 1945). Nesse importante artigo, BUSH fez duas coisas: (1) definiu sucintamente um problema crítico que estava por muito tempo na cabeça das pessoas, e (2) propôs uma solução que seria um ajuste tecnológico, em consonância com o espírito do tempo, além de estrategicamente atrativa. O problema era (e, basicamente, ainda é) "a tarefa massiva de tornar mais acessível, um acervo crescente de conhecimento"; BUSH identificou o problema da explosão informacional - o irreprimível crescimento exponencial da informação e de seus registros, particularmente em ciência e tecnologia. A solução por ele proposta era a de usar as incipientes tecnologias de informação para combater o problema. E foi mais longe, propôs uma máquina chamada MEMEX, incorporando (em suas palavras) a capacidade de associar idéias, que duplicaria "os

> processos mentais artificialmente". É bastante evidente a antecipação do nascimento da CI e, até mesmo, da inteligência artificial. (SARACEVIC, 1996, p. 42).

É também a partir da Ciência da Informação que Marcondes (2005) deu início a um programa de pesquisa que busca por um novo modelo de publicação eletrônica capaz de registrar os elementos semânticos que constituem o conteúdo de artigos científicos em formato tal que permita sua leitura por programas de computador. A proposta é criar um sistema de autopublicação de artigos científicos onde, por exemplo, o conteúdo dos mesmos fosse representado como instâncias de uma ontologia, o que permitiria a extração e codificação dos mesmos em formato inteligível por programas. Assim identificadas tais afirmações científicas constitutivas dos conteúdos dos artigos poderiam ser reutilizadas em outros contextos, abrindo a possibilidade de estimular a descoberta científica e a inovação.

A motivação para o presente trabalho foi a participação em projetos de pesquisa da Fundação Oswaldo Cruz (FIOCRUZ), mais precisamente no Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT); o grande material processado em ambientes de repositório digital como o ARCA; os registros em canais de comunicação como o sistema FALECONOSCO, um canal de comunicação que a FIOCRUZ disponibiliza para atender não somente a estudantes e pesquisadores, bem como a população que tem questionamentos para os mais diversos setores da saúde. Todos demonstram a grande massa de dados textuais que podem e devem ser reutilizados sob a forma de uma simples consulta, mineração de textos ou qualquer outro método que dê margem a novos conhecimentos ou possa transformá-los de modo que possam ser reutilizados em outros contextos.

Diante desse contexto, emerge o seguinte questionamento: quais metodologias existem hoje para o tratamento automático de textos digitais que viabilizem o reuso, quais são seus pontos fortes e fracos, e qual a sua aplicabilidade?

Sendo assim, o foco da presente dissertação está na discussão das propostas e metodologias para processamento automático do conteúdo de documentos textuais digitais, de modo a permitir seu reuso em outros contextos distintos daqueles para os quais esses documentos foram originalmente planejados. O objetivo é

compreender como estruturar esses conteúdos para que eles possam ser processados de formas mais flexíveis e semanticamente[3] mais consistentes, contribuindo para o grande ideal de unificação de todo o conhecimento.

A dissertação está estruturada da seguinte forma: no próximo capítulo é introduzido o potencial da estruturação de textos digitais e como essas iniciativas podem contribuir para a resolução de vários problemas e desafios em várias áreas do conhecimento. No terceiro capítulo são apresentadas as bases conceituais e teóricas que fundamentam os processos, o grande “guarda-chuva” que acolhe as técnicas, metodologia e abordagens da estruturação de textos digitais. Os objetivos da dissertação são apresentados no capítulo quatro, seguido pelo quinto capítulo em que se descreve como foi feita a escolha do material empírico a ser trabalhado. A discussão das metodologias identificadas é realizada no capítulo seis. O último capítulo encerra com algumas considerações e indicações de investigações futuras.

[3] Semântica é a ciência das significações, o estudo dos significados.

2 JUSTIFICATIVA

As Tecnologias de Informação e Comunicação (TICs), e mais especialmente a Internet, na medida em que possibilitam um aumento vertiginoso da produção, estoque e circulação da informação em formato digital, vêm desenhando, em anos recentes, um novo ambiente para o acesso, a integração e a produção de novo conhecimento.

Marcondes e Sayão (2001) afirmam que nesse novo ambiente de comunicação, principalmente na perspectiva das atividades de ciência e tecnologia, em que a informação é um insumo fundamental, a Internet promove a oferta de um novo conjunto de recursos para além das tradicionais formas de documento (artigos, teses dentre outros). Documentos multimídia, listas de discussão, fóruns eletrônicos, conferências em linha, imagens, modelos animados, bancos de *preprints* eletrônicos, os *e-prints* são alguns exemplos. Tanto como subsídios à pesquisa quanto como canais de comunicação e publicação de seus resultados, esses recursos vêm alterando profundamente os sistemas de recuperação de informação, antes voltados mais exclusivamente para identificação e recuperação de referências bibliográficas em bases de dados isoladas.

A lógica de operação desses sistemas não mudou: o usuário, partindo de uma necessidade de informação, formula uma questão que é então submetida à base de dados por meio de palavras-chave. A qualidade da resposta depende tanto da habilidade em expressar corretamente uma necessidade de informação, do controle terminológico[4] que a base de dados usa na representação do conteúdo dos documentos, e de como, efetivamente, esse conteúdo foi representado nas diversas possibilidades (manual ou automatizada) de representação quando da alimentação da base de dados. Como apontado por Marcondes e Sayão (2001), as dificuldades aumentam quando o processo de recuperação da informação volta-se para objetos digitais distribuídos e dispersos: textos completos, imagens fixas ou em movimento, som etc., estabelecendo como palavras de ordem a publicação na Internet e a

[4] Terminologia é o conjunto organizado de termos em um domínio especializado onde os significados foram explicados e definidos.

interoperabilidade entre fontes de informação heterogêneas e globalmente distribuídas.

Ao listar algumas fontes de informação disponíveis na Internet, os mecanismos de busca e os localizadores especializados, passando pelos portais temáticos e alcançando as bases de dados referencias, em geral pagas, os autores mencionados citam as seguintes dificuldades:

> [...] baixa qualidade da indexação, por ser feita automaticamente, que resulta em grande quantidade de informações recuperadas, a maioria sem relevância (em termos de recuperação de informação, oferecem alta revocação, mas baixa precisão); cobertura parcial da Internet; as ferramentas de busca não são especializadas; indexam páginas HTML isoladas, e não recursos; além disto, grande quantidade de informações disponíveis na Internet estão sob a forma de registros contidos em bases de dados, que ficam assim "escondidas"; estes registros são acessados somente por meio das interfaces destas bases de dados, o que pressupõe uma interação entre um usuário humano com a base de dados e, portanto, ficam inacessíveis aos programas robôs. (MARCONDES; SAYÃO, 2001, p. 26).

Em meio a tantas possibilidades, é muito provável que os resultados das buscas sejam pouco relevantes, e o processo, como um todo, cansativo. O esperado seria interagir com uma única interface de busca e receber como resultado registros e textos completos de diferentes fontes, de forma consolidada.

Atender a uma demanda em um universo tão diversificado apenas usando ferramentas de navegação torna-se tarefa exaustiva. Mecanismos de busca com o emprego de palavras-chave é uma saída mais comumente utilizada. Todos esses métodos resultam em algumas respostas que podem satisfazer ou não uma demanda, porém, exigem um enorme esforço para seu sucesso. Isso se dá em função da grande massa de dados a serem verificados e das limitações dos mecanismos de buscas baseados em processos simplesmente sintáticos, isso é, a leitura e comparação de vocábulo para vocábulo, palavra a palavra.

Todo o esforço para que a informação seja reutilizada envolve a compreensão e a concepção de processos, políticas e sistemas, tanto sociais quanto técnicas, para a identificação, seleção, compartilhamento e reutilização de recursos de informação (como o arquivamento, preservação, representação do conhecimento, resumo de texto da *Web Semântica*, metadados, classificação, arquitetura de informação e

memória organizacional). Isso representa uma parte do campo de trabalho para o reuso da informação.

Uma das mais significativas aplicações das propostas de processamento automático de textos digitais é justamente viabilizar o seu reuso. A informação biomédica clínica pode ter uso científico, assim como informação biomédica científica pode ter uso clínico. Os obstáculos são inúmeros como formatos diferentes, semânticas diferentes, informações fechadas em bancos de dados que não são interoperáveis, massa de dados produzidas em idiomas diferentes etc.

Experimentos são realizados em diferentes áreas que trabalham interpretação e integração de dados heterogêneos de modo que as ideias, técnicas e problemas têm sido partilhados e discutidos em um contexto mais amplo. O fazer e interagir com sucesso em um ambiente aberto e heterogêneo, e ser capaz de integrar de forma dinâmica e adaptável dados provenientes de outros sistemas, é algo fundamental. No entanto, os esforços caminham para que todo o processo, embora pouco preciso, encontre entendimento suficiente para permitir a interação com êxito. Com o advento da *Web*, há uma quantidade enorme de informação disponível online que pode ajudar nessa tarefa, mas essa informação é frequentemente organizada de forma caótica, armazenada em uma ampla variedade de formatos de dados e difícil de interpretar e tem que ser lida por seres humanos (LHD-11, 2011).

A mineração de textos tem auxiliado nesse trabalho procurando dar algum sentido ao conjunto de textos digitais coletados arbitrariamente ou de forma sistemática de outras bases de dados, na *Web*, em repositórios digitais etc.

Na assistência médica, os dados são quotidianamente gerados e armazenados como parte do processo de atendimento, sejam estes empregados para fins administrativos ou para a investigação científica (COIERA, 1997; PEÑA-REYES; SIPPER, 2000; SHORTLIFFE; BLOIS, 2001). Um único caso registrado na saúde ou na pesquisa pode produzir centenas de variáveis e gerar grandes quantidades de dados. Mesmo que esses dados sejam individuais, e dispostos de maneira diversificada, com alguns itens de pouco valor, informações valiosas podem estar contidas entre eles. Uma visão do conjunto processado forma subconjuntos diferentes que são, portanto, um grupo de subconjuntos. São dados recuperados e

reutilizados sob uma nova perspectiva de recuperação da informação, dando um sentido a um conjunto de dados com intuito de obter informação que não está aparente ou não está disposta para uso imediato, mas pode ser extraída e utilizada por meio da mineração de textos (KUO; CHANG; CHEN; LEE, 2001).

Essa disponibilidade de dados e de informações na saúde, juntamente com a necessidade de aumentar o conhecimento e entendimento das necessidades biológicas, bioquímicas, patológicas, psicossociais e ambientais − processos pelos quais a saúde e a doença são mediados −, significa que medicina e saúde são áreas adequadas para o emprego da mineração de dados (SHORTLIFFE; BARNETT, 2001; SHORTLIFFE; BLOIS, 2001).

Para além da sintática e mineração de textos, a *Web*, com sua proliferação e crescente volume de informação disponível, pede pela semântica – a *Web* semântica.

Em um ambiente como a Internet, há uma grande quantidade e variedade de dados interligados, há informações relevantes. No entanto, por ser desestruturado ou por haver a impossibilidade de ser de um único formato, padrão ou suporte faz com que se torne um ambiente desalinhado. A quantidade de informação a processar é de fato muito grande. Além disso, os sistemas que visam ao reuso devem reunir informações de fontes distribuídas e diversas, em que regimes de representação podem vir a ser muito heterogêneos, variáveis na sua qualidade, informação e confiabilidade em proveniência, algo difícil de se estabelecer. Contudo, a W*eb* Semântica deve ser baseada em padrões que podem operar neste mundo da informação heterogênea.

A *Web* Semântica é uma visão de uma rede de dados interligados, permitindo a integração de consulta e partilha de dados de fontes distribuídas em formatos heterogêneos, utilizando ontologias para dar uma interpretação associada e uma semântica explícita.

A *Web* Semântica diz respeito à adição de uma estrutura formal e semântica (metadados que descrevem seu conteúdo) para documentos na *Web* com o objetivo de uma gestão mais eficiente do conhecimento, especialmente por meio do acesso otimizado aos conteúdos. Assim, o mais importante seria a possibilidade de um

acesso mais direto ao conhecimento, ou seja, ir além da recuperação do documento e identificar e extrair, de forma mais rápida e direta, as declarações e assertivas de conhecimento registradas nos documentos.

A W*eb* Semântica exige, portanto, dois tipos de informação padrão para operar. Primeiro, exige formatos comuns para a integração de informações provenientes dessas fontes diversas. Segundo, precisa de uma linguagem para expressar o mapeamento entre os dados e objetos do mundo real, a fim de permitir uma compreensão perfeita de um conjunto de bases de dados distribuídas. Tais relações semânticas são muitas vezes óbvias para os seres humanos, mas não para os computadores. Um formalismo chave aqui é a ontologia, que define os conceitos e as relações as quais são usadas em aplicações específicas. Ontologias são centrais para a visão da *Web* Semântica, como o fornecimento de um dos principais meios pelos quais os termos usados em dados são compreendidos no resto do contexto (O'HARA; HALL, 2009).

Segundo Uren et al. (2005), as tecnologias que a *Web* Semântica permitem podem levar a uma possível geração de um tipo de documento "inteligente", difícil de ser pensado há dez anos: “[...] definimos um documento inteligente como um documento que "conhece" o seu próprio conteúdo, a fim de que processos automatizados possam "saber o que fazer" com ele.” (UREN et al., 2005, p. 1). A identificação do conhecimento inscrito nos documentos tem sido tradicionalmente realizada por meio do uso de metadados (por exemplo: registra-se em local específico o nome do autor, e, muitas vezes, pelo menos, parte do conteúdo, como as palavras-chave). A *Web* Semântica propõe anotar ou estruturar o conteúdo do documento utilizando informação semântica de ontologias de domínio. A principal contribuição da *Web* Semântica é, portanto, prover um padrão para uso global, o que abre a possibilidade de operar com recursos heterogêneos, criando uma ponte para uma sintaxe[5] e métodos comuns, entre outras perspectivas.

A estruturação de textos digitais traz, no mínimo, benefícios de dois tipos: melhoria no processo de recuperação de informação e aprimoramento da interoperabilidade. A recuperação da informação é melhorada com a capacidade para realizar buscas

[5] Sintaxe é a disposição material das palavras na frase.

que exploram ontologias para fazer inferências sobre dados oriundos de recursos heterogêneos. A interoperabilidade é particularmente importante para aquelas organizações que trabalham com sistemas legados, muitas vezes em diferentes formatos que não são facilmente integrados. Nessas circunstâncias, as anotações com base em uma ontologia comum podem fornecer um enquadramento para a integração de informações provenientes de fontes heterogêneas (UREN et al., 2005).

São vários os campos em que a anotação semântica vem sendo investigada: no domínio do conhecimento científico (FRIEDLAND et al., 2004), na genômica (RINALDI et al. 2004), na análise de notícias de rádio e TV (DOWMAN et al. 2005), na acessibilidade de páginas da *Web* para portadores de deficiência visual (PLESSERS et al. 2005), em compras *online* (SVAB et al. 2004), na descrição de artefatos culturais (HUNTER et al. 2004), entre outros.

O governo eletrônico (BRASIL, 2010) é outra área de aplicação potencialmente importante, em que o objeto informação é amplamente empregado, embora ainda bastante heterogêneo. Da maneira como o governo eletrônico trabalha a informação tornam a confidencialidade, a proveniência e a sua vida útil variável. Alguma informação vai ser boa por décadas ou mesmo séculos, enquanto outras informações podem estar fora de utilidade em questão de horas. Integrar essa informação de forma adequada é claramente um importante desafio.

Uma das aplicações mais importantes para a tecnologia da *Web* Semântica é a busca por dados. O aprimoramento dos meios de comunicação, o baixo custo de computadores e o desenvolvimento de *softwares* favorecem cada vez mais espaços onde a *Web* Semântica pode consolidar esses dados que estão distribuídos em diversos ambientes computacionais. Uma grande quantidade de dados é criada por meio de análises e experiências em disciplinas como física de partículas, meteorologia e ciências da vida. Além disso, em muitos contextos, as diferentes comunidades de cientistas estarão trabalhando de forma interdisciplinar, o que significa que os dados de diversas áreas (por exemplo, a genômica, os ensaios clínicos de drogas e epidemiologia) devem ser integrados. Muitos relatos de sistemas distintos e complexos (por exemplo, o corpo humano, o ambiente) são constituídos por dados trazidos de diferentes disciplinas, não só no vocabulário, mas

também na escala de descrição; a compreensão de tais sistemas e da maneira como os acontecimentos na microescala afetam a macroescala e vice-versa é claramente um imperativo importante. Muitas disciplinas científicas têm dedicado recursos para a criação de ontologias em larga escala para este e outros fins. O mais conhecido deles é o Gene Ontology[6] (GENE ONTOLOGY, 1999), um vocabulário controlado para descrever o produto do gene e seus atributos nos organismos e vocabulários relacionados, desenvolvido pela Open Biomedical Ontologies (OBO) (OBO, 2001).

Na área biomédica, há o projeto internacional *Semantic and Services-enabled Problem Solving Environment for Trypanossoma Cruzi*, para criar um ambiente *Web* integrado de acesso aos diferentes recursos e fontes de conhecimento sobre o *Trypanossa Cruzi* (SHETH, A., 2008). Esse projeto coloca em uma de suas linhas de ação a análise semântica de textos para extração de conhecimento da literatura biomédica. O *Pubmed* representa um recurso vasto e valioso para a investigação das ciências da vida (SHETH et al., 2003; SHETH, 2008).

> Semantic text analysis approaches for extraction of knowledge from biomedical literature - Biomedical literature, for example Pubmed, represents a vast and valuable resource for life sciences research. The ability to extract relevant knowledge from biomedical text and its representation in Semantic Web standard formats such as RDF is an important research issue that is being addressed in this project. (SHETH, 2008).

Também a *International Conference on Biomedical Ontology* (ICBO), conferência realizada na Universidade de Buffalo, Nova Iorque, em 2009, buscou abordar aspectos da ontologia biomédica com o objetivo de discutir o papel da ontologia no futuro da publicação científica.

> Ontologies are being used in a variety of ways by researchers in almost every life science discipline, and their use in annotation of both clinical and experimental data is now a common technique in integrative translational research. Principles-based ontologies are being developed for the description of biological and biomedical phenomena of almost every different type. To be maximally effective, such ontologies must work well together. But as ontologies become more commonly used, the problems involved in achieving

[6] Gene Ontology – iniciativa importante no campo da bioinformática com o objetivo de uniformizar a representação do gene, o produto do gene e seus atributos por meio das suas espécies e o registro em bancos de dados.

> coordination in ontology development become ever more urgent. (ICBO, 2011).

Pode-se dizer que, de acordo com a abordagem discutida nesta dissertação, torna-se importante investigar a capacidade de identificar e analisar métodos de extrair automaticamente conhecimento relevante a partir do texto biomédico digital, e sua representação na *Web* Semântica por meio de formatos padrões como RDF com objetivo de reutilizá-lo para outros fins.

3 OBJETIVOS

3.1 Objetivo geral

Levantar e sistematizar relatos na literatura de projetos, experiências, propostas e metodologias para processamento automático do conteúdo de documentos textuais digitais, de modo a permitir seu reuso em outros contextos distintos daqueles para os quais esses documentos foram originalmente planejados.

3.2 Objetivos específicos

1. Sistematizar os métodos identificados na literatura que empregam processamento automático de conteúdo de documentos digitais;
2. Organizar a literatura levantada de acordo com esta sistematização;
3. Avaliar as metodologias identificadas de acordo com sua aplicabilidade.

4 MARCO TEÓRICO

A revisão bibliográfica, apresentada a seguir, busca as principais teorias e suas abordagens, resultados e reflexões de pesquisadores para compor a base teórica da dissertação. O problema e as questões de pesquisa, apontados anteriormente, envolvem os seguintes conceitos, que serão apresentados e discutidos a seguir:

Linguagem natural

Textos

Documentos digitais

Web Semântica

Reuso e integração da Informação

Mineração de dados, mineração de textos

4.1 Linguagem natural

Chomsky (1955-56), linguista e filósofo, preocupou-se em elaborar uma teoria linguística que explicasse a natureza das línguas e da linguagem. Para o autor, essa teoria deveria responder às seguintes questões: (1) O que constitui o conhecimento linguístico de um falante? (2) Como esse conhecimento se desenvolve? (3) Como esse conhecimento é posto em uso?

As respostas implicam a concepção de um modelo teórico que trate amplamente dos mecanismos especificamente linguísticos e deem ao falante da língua a produção e compreensão das sentenças gramaticais. Os estudos realizados por Chomsky culminaram no início dos anos 1950 com a criação da gramática gerativa[7], sempre na tentativa de responder à primeira pergunta.

[7] Na teoria linguística, a gramática gerativa refere-se a uma abordagem particular para o estudo da sintaxe e tenta dar um conjunto de regras que irão prever corretamente quais combinações de

No entanto, a língua, no estruturalismo, como podemos verificar em Saussure (1989, p. 17), era entendida como um sistema de unidades e relações, com características que dependiam da sua natureza coletiva, a partir do consenso coletivo, do social.

Na mesma linha de pensamento, Saussure mais uma vez expõe essa concepção e questiona:

> Mas o que é a língua? Para nós, ela não se confunde com a linguagem; é somente uma parte determinada, o essencial dela, indubitavelmente. É, ao mesmo tempo, um produto social da faculdade de linguagem e um conjunto de convenções necessárias, adotadas pelo corpo social para permitir o exercício dessa faculdade nos indivíduos. (SAUSSURE, 1989, p. 17).

E reconhecendo a língua como um produto da sociedade, os estruturalistas abdicam os estudos da faculdade da linguagem para deslocar o foco de interesse, saindo do social para o mental. Isto é a gramática universal de Chomsky (1965), que postula a gramática compartilhada por todas as línguas e trata a língua como um tipo de conhecimento tácito, o qual todo ser humano é capaz de dominar desde os primeiros anos da infância.

Ao pensar na escrita como um sistema complexo de símbolos que podem ser processados em uma leitura e, se de alguma forma esse conjunto torna-se mais complexo, de modo que venha a ocultar numa palavra um nível ainda maior de compreensão ou até mesmo outro sistema que esta venha representar, então processar textos por meio de programas de computador requer sistemas eletrônicos complexos na sua essência. Esta complexidade advém de um número extenso de variáveis de interpretação, parâmetros tecnológicos e dos recursos empregados para a sua execução.

Se a linearidade para leitura e interpretação de textos é uma habilidade humana, a compreensão do conteúdo textual digital em grande escala por máquinas requer que esse conjunto de palavras do texto esteja estruturado ou com algum sentido para o que se deseja realizar.

---

palavras devem formar frases gramaticais e, na maioria das vezes, as regras também preveem a morfologia de uma frase.

Textos, inclusive os digitais, já têm *a priori* a estrutura gramatical investigada por Chomsky com sua gramática gerativa formalizada, que é a base para a construção dos programas *parsers*, os analisadores sintáticos. Além disso, a estrutura sintática de um texto fornece indicações importantes para sua estrutura semântica.

Coesão e estrutura aproximam-se da sintaxe, na Linguística, enquanto coerência e organização dão a ideia da semântica. "É a coerência sistêmica que dá o sentido às partes, constituindo o *substratum* de toda significação, logo da dimensão semântica." (SANTAELLA; VIEIRA, 2008, p. 37).

Syntactic Structures foi uma destilação do livro *Logical Structure of Linguistic Theory* (1955) no qual Chomsky apresenta sua ideia da gramática gerativa. Sua teoria postula que os enunciados ou frases das línguas naturais devem ser interpretados em dois tipos distintos de representação: as "estruturas superficiais", correspondendo à estrutura patente das frases, e as "estruturas profundas", uma representação abstrata das relações lógico-semânticas das mesmas (PINTO MOLINA, 1992, p. 50).

A gramática tradicional considera as funções sintáticas como elementos primitivos da gramática. No entanto, ao definir a função de *sujeito* (termo com o qual o verbo concorda), deve-se valer de noções mais básicas, como as de *verbo* e de *concordância*. Por causa disso, os linguistas gerativistas formularam uma teoria sintática incluindo apenas as classes gramaticais (Nome, Verbo, Adjetivo, Preposição, Advérbio) nas representações sintáticas (árvores) das sentenças.

A função do processamento sintático de acordo com Gonzalez e Lima (2003),

> Tradicionalmente ocupa posição de destaque, com a semântica sendo considerada uma interpretação da sintaxe. Mas, também, pode ser considerado em posição secundária, de acordo com os pesquisadores denominados semântico-gerativistas. Neste último caso, a sintaxe é uma projeção da semântica. Entretanto, qualquer que seja a visão adotada, o processamento sintático é uma etapa indispensável para viabilizar o processamento semântico. (GONZALEZ; LIMA, 2003, p. 352).

Nesse sentido, segundo Maia (2001), os *parsers* − programas de computador com capacidade de processar textos automaticamente e identificar sua estrutura sintática − surgem como analisadores para uma gramática que aceita como entrada

sentenças e cria para essas sentenças a sua árvore gramatical. A figura 1, a seguir, ilustra a visão estruturalista da linguagem, segundo Chomsky.

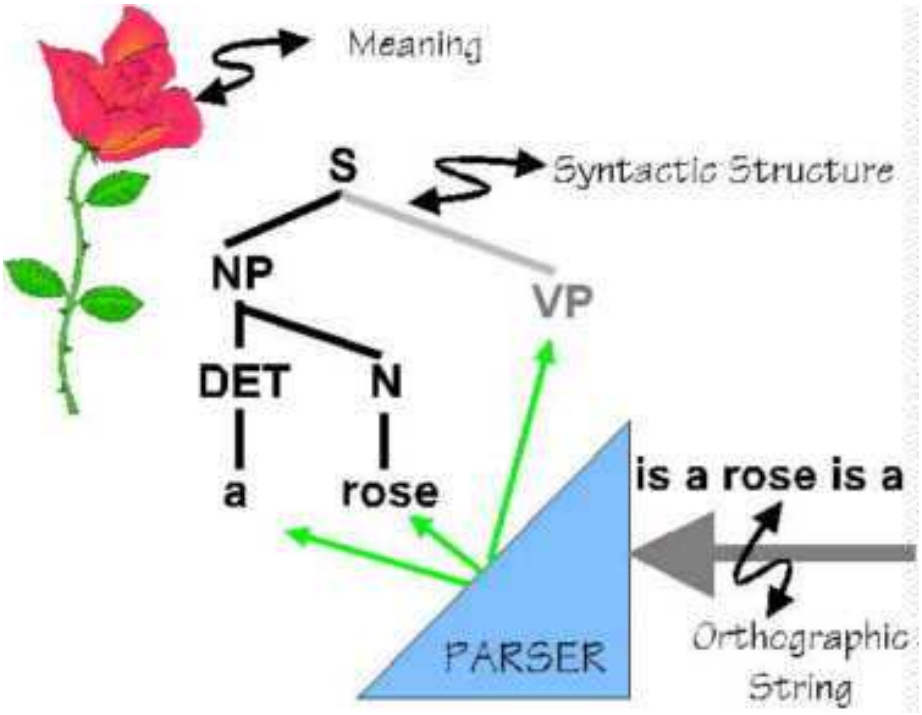

Figura 1 – Parser

Fonte: Natural Language Computing - Gramática Universal DOUGHERTY (1994) - A visão estruturalista da linguagem: Noam Chomsky

4.2 Textos

A palavra texto significa um conjunto de palavras de um autor que compõe uma carta, livro, folheto, documento; também é um trecho ou fragmento de uma obra, conteúdo de um telegrama etc. Etimologicamente, a palavra de origem latina vem de tecer, entrelaçar, compor de forma organizada o pensamento em obra escrita ou declamada. Diz-se também do material ilustrativo de uma obra que é impresso à parte, em papel especial e em folha(s), não numerado(s) ou numerados de forma autônoma entre cadernos de um livro. São citados como exemplo: mapas, fotos e desenhos. (HOUAISS, 2004, p. 2.713).

Platão e Fiorin (2006) sugerem que os significados das partes de um texto não podem ser obtidos isoladamente, há de se levar em conta correlações existentes de

forma a dar um sentido. O texto não é um conjunto de frases onde se pode obter significados autônomos.

Segundo Pierre Lévy (1977) quanto à leitura ou atualização do texto “o texto é um objeto virtual, abstrato, independente de um suporte específico. Essa entidade virtual atualiza-se em múltiplas versões, traduções, edições, exemplares e cópias.” (LÉVY, 1977, p. 19).

Artigos científicos são textos construídos segundo padrões e regras específicos. Para Azevedo (2001), o artigo científico é definido como um texto escrito e acrescenta uma finalidade, a comunicação. Sendo assim, “o artigo científico é um texto escrito para ser publicado num periódico especializado e tem o objetivo de comunicar os dados de uma pesquisa, seja ela experimental, quase experimental ou documental.” (AZEVEDO, 2001, p. 82).

Na literatura científica, os textos possuem características previamente definidas para sua aceitação, revisão e publicação: o artigo, como um texto escrito, é formalizado e provido de regras internacionais. No Brasil, a ABNT fornece os procedimentos de normatização que dão as definições e formas para artigo científico, cuja publicação ou parte dela, com autoria declarada, expõe e discute ideias, métodos, técnicas e apresenta resultados nas mais diversas áreas do conhecimento; o artigo de revisão tem a finalidade de resumir, analisar e discutir informações que foram disponibilizadas, isto é, já publicadas; o artigo original para todos os temas e abordagens originais. (NBR 6022, 2003, p. 2).

Segundo Van Dijk (1978, apud PINTO MOLINA, 1992), um texto apresenta três estruturas constituintes. A primeira, a microestrutura, situa-se no nível de sentenças e orações que vão satisfazer as condições de coerência e conexão. A segunda, a macroestrutura, refere-se à ideia geral do texto, a representação abstrata do seu significado, representando a estrutura semântica. E finalmente, a superestrutura que caracteriza o tipo de texto, sua forma, suas unidades, a organização sequencial de suas partes. Na figura 2 são apresentadas as Unidades Textuais de documento científico.

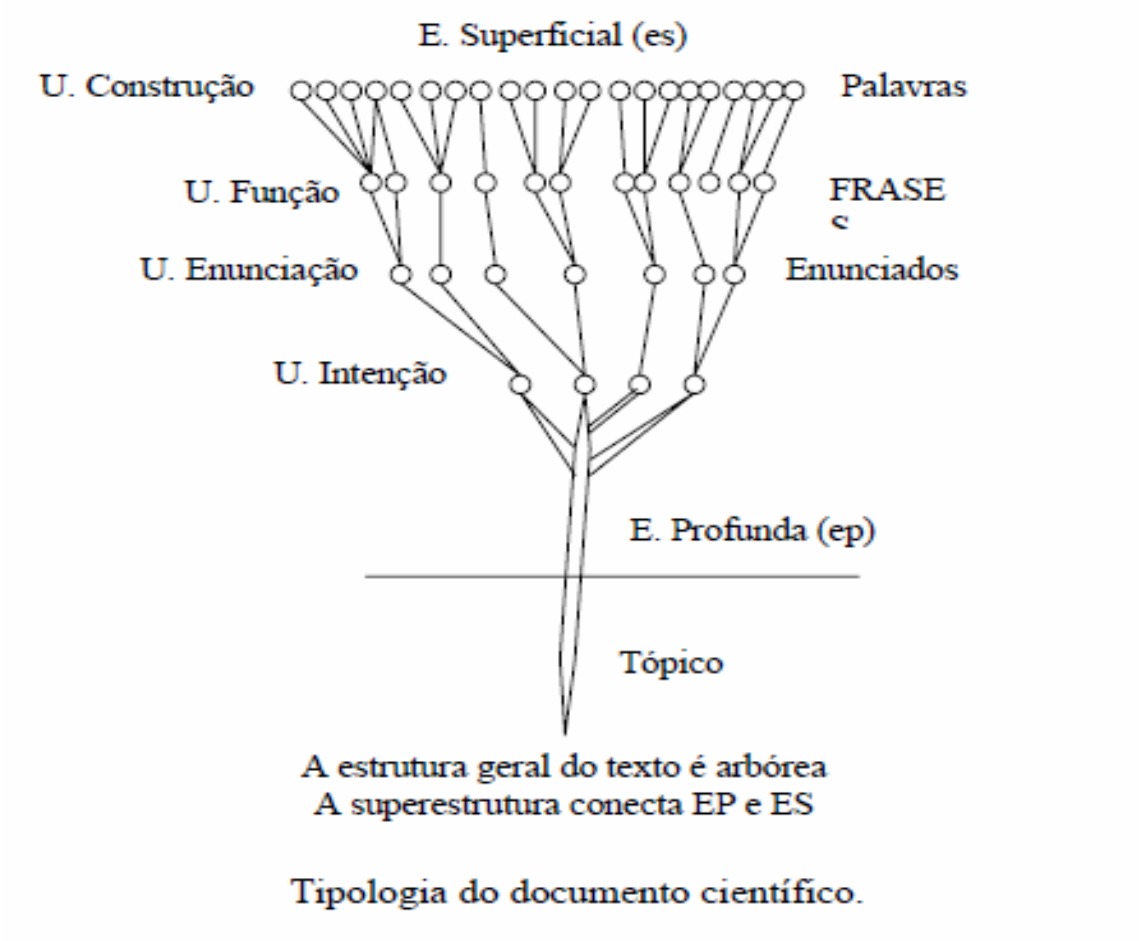

Figura 2: Unidades Textuais
Fonte: PINTO MOLINA (1992, p. 50)

Serão definidos alguns conceitos com objetivo de que estes não sejam interpretados de forma incorreta ou adquiram outro significado que não o desejado ao longo desta dissertação. O objetivo é auxiliar o leitor a identificar onde se deseja atuar, quais métodos poderão ser aplicados ou desenvolvidos após a revisão bibliográfica e os limites que serão estabelecidos diante dos desafios da pesquisa em questão.

Independente das inúmeras definições existentes para um texto, este é um documento textual e documentos textuais científicos na sua forma digital são objetos do escopo deste trabalho.

Pode-se dizer que o documento científico possui importância fundamental, uma vez que é considerado:

> A autêntica seiva da árvore da ciência. Estes documentos possuem uma personalidade características que nos permite distingui-los dos demais, não somente pelo seu conteúdo singular, evidentemente científico, mas sobretudo pela maneira de estruturar a apresentação desse conteúdo (PINTO MOLINA, 1992 apud FLAMINO et al., 2005, p. 10).

4.3 Documentos Digitais

A *Web* vem se tornando o maior repositório de documentos digitais. Por meio do modelo *Open Archives Initiative* (OAI) vários módulos de *softwares* e outras iniciativas possibilitaram o avanço dos repositórios e bibliotecas digitais. Dentre as inciativas, podemos observar em todo o mundo publicações de periódicos eletrônicos, teses e dissertações (KURAMOTO, 2006). Todo esse material sob a forma de um documento textual digital é disponibilizado em diversos suportes e formatos, ainda que entendamos, em uma definição mais simples, que um repositório ou biblioteca digital é um banco de dados com características definidas para armazenar e recuperar a informação nele contida. No contexto dessa dissertação, é necessário definir o que vem a ser um documento textual digital e a sua aplicabilidade citando alguns autores.

> An early development was to extend the notion of document beyond written texts, a usage to be found in major English and French dictionaries. (For historical background on "document" see also Sagredo Fernández & Izquierdo Arroyo (1982)). "Any expression of human thought" was a frequently used definition of "document" among documentalists. (BUCKLAND, 1997, p. 805).

Para Feitosa (2006),

> Um documento é um objeto que fornece um dado ou uma informação e pode ser diferenciado entre outros documentos, de acordo com suas características físicas ou intelectuais. As características físicas de um documento relacionam-se aos conceitos de material, natureza, tamanho, peso, forma de produção, suporte, entre outras. As características intelectuais relacionam-se aos conceitos de objetivo, conteúdo, assunto, tipo de autor, fonte, forma, forma de difusão, originalidade, entre outras. (FEITOSA, 2006, p. 17).

Por sua vez, Wives (2004) ressalta que:

> Inicialmente, documento foi um termo utilizado para denotar um registro textual qualquer (um texto). Isto porque os primeiros documentos eram, na verdade, informações elaboradas (descritas) na forma de um texto escrito em linguagem natural. Porém, existem outros objetos que podem conter e transmitir informações. (WIVES, 2004, p. 17)

Traçando um breve histórico do documento, Bodê (2006) relata que até o início do século XIX a produção documental estava registrada em papiros, pergaminhos e papel e, portanto, todo o conteúdo dos documentos de gênero textual não poderia ser removido sem que houvesse dano ao suporte que o conservava. Com a invenção da fotografia, em meados do século XIX, aumenta a diversidade desses acervos documentais e ao final do século XX tem-se a fotografia digital.

As possibilidades aumentam com os inventos para o registro de som, imagens e, ainda no final do século XX, surgem os CDs (Compact Disks), DVDs e o seu emprego com os computadores e suas tecnologias. A popularização dos computadores a partir dos anos 1980 representa um salto no âmbito da produção, registro, armazenamento e recuperação de documentos.

Ainda segundo Bodê (2006), quanto à caracterização do documento eletrônico comparando o conjunto de documentos produzidos até o final do século XIX com os documentos eletrônicos e digitais atuais, há evidências de peculiaridades:

> **A legibilidade por máquinas** por meio de *software*;
>
> **A independência entre suporte e conteúdo**, o que antes não poderia ter seu conteúdo removido sem a perda do documento devido ao dano causado ao suporte, porque estes são indissociáveis. O documento eletrônico é formado por um suporte e conteúdo; no entanto, um suporte físico ainda existirá e poderá ser facilmente substituído sem danos ao conteúdo;
>
> **Diversidade de conteúdos,** o conteúdo é diferente para cada tipo de documento: alguns são apenas textos, diferindo dos que possuem imagens fixas, fotografias e dos documentos que apresentam conteúdo sonoro ou vídeos. A diversidade apresentada se dá para textos, que

diferem dos que possuem imagens fixas, fotografias, dos documentos com conteúdo sonoro ou imagens etc.

Para os documentos eletrônicos, estes compõem um grupo único, porque a flexibilidade promovida pelos meios digitais se dá na codificação de decodificação numa linguagem digital única.

> A codificação. No caso de textos, os caracteres transmitidos correspondem aos diversos símbolos do alfabeto (maiúsculas e minúsculas), aos números decimais, bem como aos sinais relativos às operações aritméticas e lógicas e à pontuação, ou seja, um total de uma centena de caracteres. A codificação consiste em atribuir a cada um desses caracteres um número binário determinado, cujo número de bits depende do código adotado...O mais difundido hoje em dia é o American Standard Code for Information Interchange (ASCII[8])... (LE COADIC, 1996, p. 91-92).

Por sua vez Buckland (1997) considera o texto impresso:

> Documentation was a set of techniques developed to manage significant (or potentially significant) documents, meaning, in practice, printed texts. (BUCKLAND, 1997, p. 805).

O autor acrescenta a visão documentalista para a noção da evolução de documento entre Otlet, Briet, Schumeyer e outros que enfatizaram cada vez mais uma preocupação sobre o que funcionava como um documento em vez das tradicionais formas físicas dos documentos. A mudança advém da tecnologia digital e parece fazer essa distinção ainda mais importante.

Para Lévy (1993) e Berners-Lee et al*.* (2001), a comunidade científica tem feito pesquisas com o propósito de desenvolver metodologias para que computadores tornem-se uma extensão das capacidades cognitivas humanas. A transformação de documentos em papel para documentos textuais digitais está relacionada a uma mudança de qualidade, cujas consequências não estão totalmente claras. O

---

8 Código padrão que utiliza um conjunto de caracteres codificados composto de 7 bits (8-bits, incluindo verificação de paridade). O código é utilizado para o intercâmbio de informações entre os sistemas de processamento de dados, sistemas de comunicação de dados e equipamentos associados. O conjunto ASCII consiste de caracteres de controle e caracteres gráficos. IBM (2006b), ver Tabelas 1,2 e 3.

documento textual digital poderá vir a ser uma nova e poderosa ferramenta cognitiva, em especial no contexto do projeto *Web* Semântica.

O mundo digital traz novas mudanças para os profissionais que lidam com informação. Autores, bibliotecários, e porque não mencionar os tecnologistas com uma bagagem considerável de expertise na tradução, atestam o papel que a informática vem representando nesse contexto. Todos são envolvidos direta ou indiretamente nos processo de produzir, armazenar, tratar e recuperar documentos e informação.

O mundo digital altera radicalmente os processos que envolvem o aprendizado e o trabalho desses profissionais. Em função dessa modificação ocorrida, por meio da tecnologia digital, atualmente autores de textos, juntamente com profissionais dedicados às atividades de organizar esses conjuntos de informação, *clusters* de documentos, se preocupam em criar meios para recuperá-los ou, no mínimo, em dar visibilidade ao produto da sua criação.

Com essas novas instâncias postas em pauta, torna-se um desafio representar e identificar os vários tipos de conhecimentos, informações e instrumental tecnológico. Segundo Vickery (1986 apud ALVARENGA, 2003) diferentes tipos de dados demandam técnicas de representação distintas. Com isso ocorre a variedade de tipos de tratamento da informação em contextos como bibliotecas, museus, arquivos, e na própria *Web*, plataforma de exposição de objetos digitais.

Os textos digitais são codificados principalmente utilizando o código ASCII, no qual cada caractere é representado por um código de 8 bits, por exemplo, a letra “A” é representada pelo número decimal 65 ou pela sequência binária 01000001 (KOCHHAR, 2008). O código ASCII só é capaz de representar caracteres das línguas que utilizam o alfabeto latino. Atualmente o UNICODE[9] vem substituir o ASCII, podendo representar todas as letras de todos os alfabetos existentes no mundo.

---

[9] Código único para caracteres. Um padrão para representação no computador de caracteres de um texto em qualquer escrita existente.

É importante ressaltar que no processamento automático de textos digitais, a mineração de dados se constitui no campo da informação, principalmente para a informação médica.

Segundo Benoit (2002), minerar dados é parte de um processo pelo qual a informação pode ser extraída de dados ou bases de dados e a utilizada para auxiliar a tomada de decisão numa variedade de contextos.

Referindo-se à importância das estruturas, Adriaans e Zantinge (1996) destacam a relação de minerar dados com a descoberta do conhecimento em bases de dados. A mineração é a fase da extração de conhecimento no processo de descoberta do conhecimento que inclui, dentre outros métodos, as fases de selecionar, limpar, codificar e recodificar dados, o reuso da informação seguido pela apresentação e comunicação dos resultados às atividades de mineraração dos dados.

Portanto, documento digital nesta dissertação será todo aquele objeto que possuir conteúdo textual digital e do qual se possa extrair de forma automática ou por meio de software todo ou parte de seu conteúdo. Se necessário, serão contemplados também, outros tipos de documentos que não estão na sua forma digital, e sim registrados num suporte como papel, livro ou outro formato não digital; este poderá ter modificado o suporte de maneira que seu conteúdo possa ser extraído.

O documento textual digital, independente da sua estrutura e armazenamento, é o objeto principal desta dissertação, com suas estruturas textuais que servirão, após sua identificação, para organizar, recuperar, extrair e armazenar todo ou parte desse documento, sempre tendo em vista o seu reuso.

As metodologias e tecnologias trazidas pela *Web* Semântica podem trazer aportes significativos ao processamento automático de textos digitais e algumas dessas metodologias serão consideradas no próximo capítulo.

4.4 *Web* Semântica

Segundo Marcondes e Campos (2008), artigos científicos são bases de conhecimento registradas no modo texto e são processados por seres humanos. Com o advento da Internet e os diversos suportes eletrônicos oferecidos no

ambiente digital, uma massa de documentos é disponibilizada atualmente e, segundo os autores, atinge o mais alto grau com o surgimento da *Web* e da adoção das publicações veiculadas por meio de suportes eletrônicos. Devido ao fato destes documentos eletrônicos não estarem armazenados estruturadamente, isto é, de maneira tal que seja possível o processamento do conhecimento contido nesses documentos por meio do uso de computadores, a divisão do trabalho, desse processamento, se dá entre pessoas.

> São estas que têm que comparar, avaliar a coerência, relacionar, inferir e citar o conhecimento contido em textos. O formato textual não estruturado impede que este conhecimento seja processado por programas, como é a proposta da Web Semântica. (MARCONDES; CAMPOS, 2008, p. 110).

Para Berners-Lee et al. (2001) a *Web* Semântica é uma extensão da *Web*, e nesta é dada à informação um significado, e é por meio desta relação que surge a cooperação e o desenvolvimento do trabalho com uma melhor interação entre pessoas e computadores.

> A Ciência da Informação, ciência dos conteúdos registrados, das suas transferências com vistas à sua apropriação social, vem de uma longa tradição teórica, metodológica e prática que converge para as questões atuais colocadas pela proposta da Web Semântica e para a construção de ontologias. (MARCONDES; CAMPOS, 2008, p. 113).

Pode-se dizer que a organização e modelização de domínios de conhecimento são áreas de pesquisa recorrentes na Ciência da Informação (CAMPOS, 2004) em especial no ambiente *Web*.

Com base nos trabalhos realizados por Bush na década de 1940 e Ted Nelson na década de 1960, que envolviam o hipertexto, a *World Wide Web* (WWW) foi concebida por Tim Berners-Lee no período de 1989 até 1991. A WWW surgiu como uma proposta para que a informação pudesse adquirir um significado bem definido, com objetivo de viabilizar de forma rápida e ampla a comunicação e cooperação, contando para isso com recursos humanos e computacionais (FEITOSA, 2006; PICKLER, 2007; MARCONDES; SAYÃO, 2002).

Entretanto, segundo Pickler (2007), a *Web* voltou-se mais para a comunicação entre humanos e reconhecida como *Web* Sintática por alguns autores, como Breitman (2005), no qual os computadores desempenham o papel de apresentar a

informação, e a interpretação fica a cargo das faculdades intelectuais do homem, considerando o interesse por este nas formas de organizar, classificar e selecionar o que lhe é pertinente ou simplesmente conhecimentos do seu interesse.

Criada para facilitar o acesso, o projeto da *Web* traz consigo o intercâmbio e a recuperação da informação e a interoperabilidade (BRASIL, 2010; FONSECA et al., 2000), o que para outros autores, como Souza e Alvarenga (2004), é um modelo anárquico, desenhado de acordo com a liberdade de acesso, e as inúmeras formas de como a informação é disponibilizada nesse ambiente.

A falta ou a inexistência de estratégias e métodos que possam atender a um sistema de indexação dos documentos contidos na Web – que recuperem o conteúdo relevante nela contido, seja para sofisticar motores de busca que não somente empreguem palavras-chave porque essas são pouco eficientes e não se mostram eficazes por outras restrições como o acesso a bases de dados ou esbarram nos diversos tipos de estruturas, arquiteturas e suportes, isto é, onde e como a informação que se deseja recuperar foi armazenada – o que torna complexo o seu gerenciamento.

A estrutura e complexidade de um sistema são dadas pelo número de relações entre seus componentes. Por isso, a arquitetura da *Web* Semântica visa a estruturar conteúdos (documentos) de forma crescentemente mais complexa por meio de diferenciadas relações, como explicitado na proposta de Berners-Lee et al. (2001), como um nível sintático, com conteúdos em XML, que serve de base para níveis crescentemente estruturados, aumentando assim o potencial de processamento semântico: XML Schema, RDF, RDF Schema e ontologias em linguagem OWL.

Então, a estruturação de conteúdos seguindo os padrões da proposta da *Web* Semântica foi assim descrita:

> XML fornece uma sintaxe superficial para estruturar documentos, no entanto, não contém relações semânticas ou algo que atribua ou dê significado a este documento.

> XML Schema é uma linguagem para restringir a estrutura de documentos XML e se estende no padrão XML para definir tipos de dados.
>
> RDF é um modelo de dados para objetos “recursos” e relações entre eles, fornece uma semântica simples para o modelo de dados, e este modelo de dados pode ser representado numa sintaxe em XML.
>
> RDF Schema é um vocabulário utilizado para descrever propriedades e classes com uma semântica para descrever de forma genérica e hierárquica essas propriedades e classes.
>
> OWL adiciona mais vocabulário para descrever propriedades e classes: dentre outras, descreve também as relações entre as classes, cardinalidade, igualdade, características das propriedades e as classes enumeradas. (OWL Ontology *Web* Language Overview, 2004. p. 3).

Conforme a descrição observa-se que uma possível “semântica computacional” está intimamente ligada a uma crescente estruturação dos conteúdos disponibilizados na *Web*.

Assim, na figura 3, a seguir, buscou-se representar a complexidade de estruturas interrelacionadas da *Web* Semântica. Dentre outras características, encontra-se representado na figura 3, o Unicode[10] (IBM, 2006b), que é a representação de texto em todos os produtos de softwares. O Unicode substitui o que antes eram as normas definidas pelo padrão ISO 8859, e nada mais são do que conjuntos de caracteres ISO, isto é, uma extensão do *American Standard Code for Information Interchange*-ASCII. (FIPS 1968; IBM, 2006a).

---

[10] Os computadores lidam apenas com números. Eles armazenam letras e outros caracteres atribuindo um número para cada um. Antes do Unicode existiam centenas de diferentes sistemas de codificação para atribuir estes números. Nenhuma codificação única pode conter caracteres suficientes: por exemplo, só a União Europeia requer várias codificações diferentes para cobrir todas as suas línguas. UNICODE CONSORTIUM (1991).

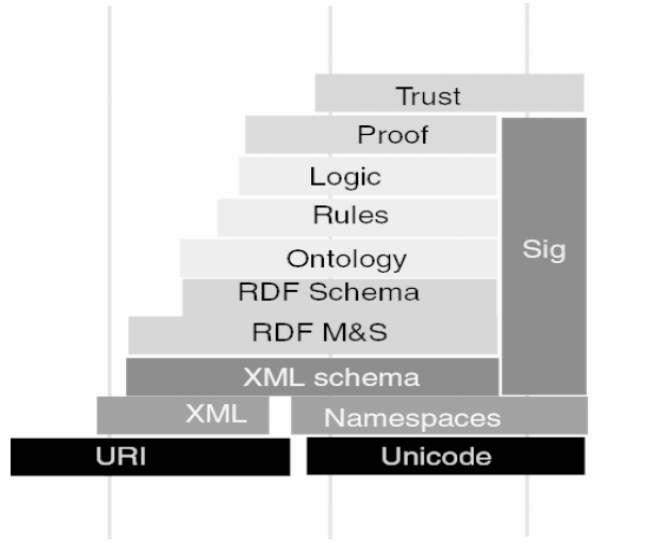

Figura 3: Estruturas interrelacionadas de tecnologias da WEB SEMÂNTICA

Fonte: www.w3.org/2001/ - W3C (2001)

O código ASCII é um conjunto de caracteres numéricos, alfabéticos e códigos especiais que foi amplamente utilizado em diversos países, com intuito de representar o texto em computadores, controlar periféricos e compor protocolos de comunicação. Atualmente o UNICODE incorpora o ASCII e o substitui, uma vez que representa caracteres e atende adequadamente às novas demandas que surgiram com a evolução tecnológica de *hardwares* e *softwares*.

Caso se deseje coletar dados neste rico e tão diversificado cenário, é necessário identificar as fontes e reconhecer quais estruturas são empregadas para que a coleta seja eficaz. Apesar dos avanços tecnológicos e dos formatos de apresentação como PDF, as publicações digitais ainda são quase que uma extensão do modelo impresso. Assim, determinar quais os melhores métodos de estruturar documentos digitais e quais as possíveis estruturas utilizadas são atividades relevantes para a solução de problemas dessa natureza, independente de estarem num ambiente automatizado ou manual.

Embora a massiva resposta obtida nas buscas realizadas por esses motores, sites indexados, arquivos e todo esse conjunto classificado, indexado e armazenado de acordo com uma nova base de dados empregada para esse propósito, é importante frisar que apenas uma pequena parcela do conteúdo existente na *Web* foi realmente checada. (CARDOSO, 2000; MARCONDES; SAYÃO, 2002; SOUZA, 2006)

Nesse sentido, para Pickler (2007) a *Web* Sintática vem melhorar a satisfação do usuário no momento da busca, com a finalidade de retornar as informações que atendam às suas necessidades e expectativas. De outra forma, há a *Web* Semântica que, por meio de outros mecanismos, espera-se capturar o significado das páginas, quando, ao criar um ambiente, permita que computadores possam processar e relacionar conteúdos provenientes de outras fontes. Segundo Breitman (2005), para que isso se torne possível, é preciso introduzir semântica na estrutura dos documentos disponíveis na *Web*.

> Se as palavras codificam um sentido de várias maneiras, podemos entender que a semântica é o estudo da função das palavras, função essa de transmitir um sentido e um significado relativos ao conteúdo. Sendo assim, percebemos que, se a intenção inicial da Web Semântica é justamente acrescentar semântica ao conteúdo Web, essa semântica servirá para determinar o sentido de um termo no contexto de determinado documento (PICKLER, 2007, p. 69).

Considerada por Tim Berners-Lee como uma extensão da *Web* atual, a *Web* Semântica tem como objeto apresentar uma estrutura que viabilize compreender e gerenciar conteúdos armazenados na *Web*, independente da tipologia empregada para que esses conteúdos sejam armazenados, isto é, da sua forma e suporte, como por exemplo, texto, imagem ou som (OLIVEIRA, 2002). Isso se dará por meio da valorização semântica e dos coletores de conteúdo que obtêm dados oriundos de outras fontes, e dependerá de esses coletores serem capazes de processar a informação, relacioná-la e disponibilizá-la oferecendo novos resultados a outros programas. (OLIVEIRA, 2002)

4.5 Reuso e Integração da Informação

Segundo Markus (2001), o termo “reutilização do conhecimento” é aplicado quando experiência e conhecimento são empregados, reaplicados ou adaptados em circunstâncias viáveis, independente do período e contexto que comportem o resultado dessas duas capacidades intelectuais adquiridas e armazenadas pelo homem, codificadas ou não, e transformadas em informação.

Segundo a autora, o ciclo do conhecimento pode ser descrito nas seguintes etapas: captura ou documentação do conhecimento; modularização e formatação; distribuição ou disseminação do conhecimento; e reutilização do conhecimento. Mas

é na última etapa, a de reutilização, que frequentemente encontra-se o interesse organizacional, uma vez que está relacionada à eficácia da organização.

Nesse sentido, a autora identifica três atores principais nesse processo de reutilização do conhecimento: o produtor de conhecimento, que origina e documenta o conhecimento, isto é, torna explícito o conhecimento tático; o intermediário, que organiza o conhecimento para utilização e facilita sua disseminação; e o reutilizador do conhecimento ou simplesmente usuário, que recupera conteúdos e os aplica.

Além disso, a autora ressalta que a utilização efetiva pelos usuários acontecerá apenas se os intermediários tiverem disponibilizado registros adequados às necessidades específicas de cada situação.

Ela concluiu que parte do papel dos intermediários pode vir a ser desempenhado pela tecnologia da informação, por meio da criação de categorizações automáticas, abstrações, filtros e disseminação de conteúdo.

Outra noção correlata ao reuso de recursos disponíveis na *Web* é o de *mashup*. Griffin (2008) afirma que um *mashup* é uma evolução da maneira como os aplicativos *Web* são desenvolvidos para que permita a um programador integrar produtos e serviços de empresas concorrentes como a Microsoft, Google, Amazon e Yahoo para criar novos produtos e serviços únicos.

As organizações são responsáveis por promoverem recursos que integrem outros produtos e serviços por meio da programação de aplicativos (Figura 4 e 5). São mais conhecidas como *Application Programming Interface* (API).

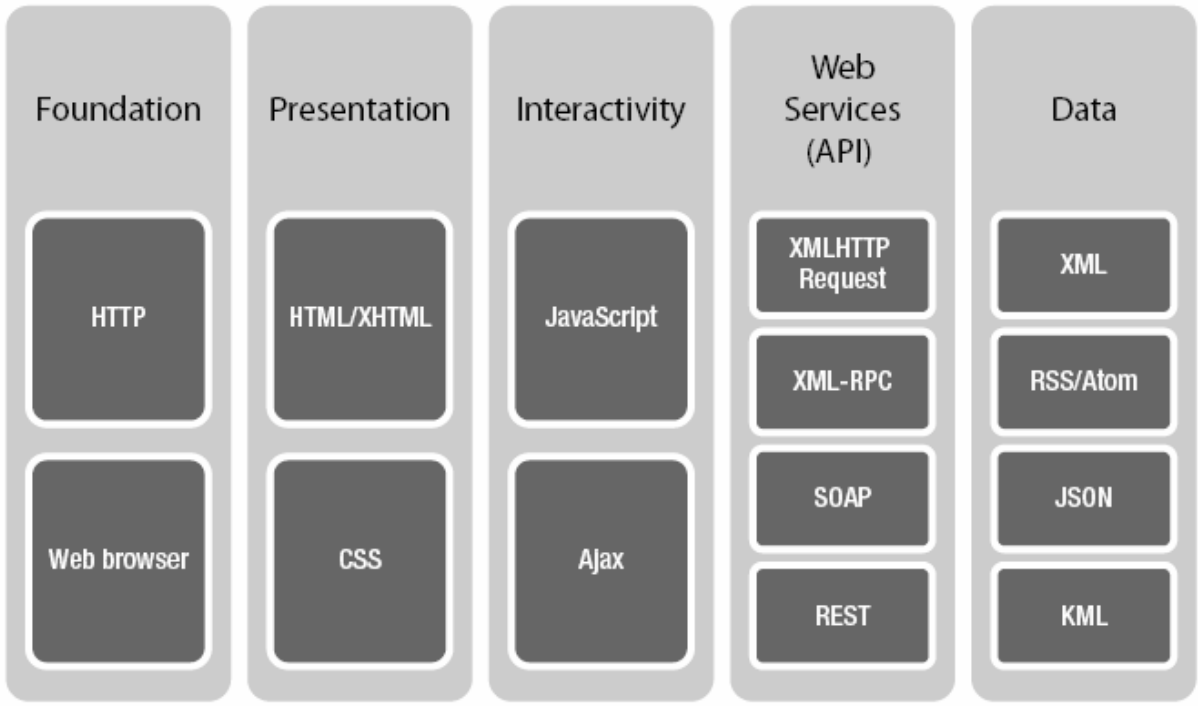

Figura 4: Tecnologias utilizadas por *mashups*
Fonte: GRIFFIN (2008, p. 5)

Para Rojas et al. (2007), o surgimento de novos paradigmas na gestão de recursos humanos, o desenvolvimento de outras áreas, o papel do conhecimento e a sua gestão em uma organização são variáveis cada vez mais relevantes no processo para tomada de decisão.

O registro desse conhecimento (JENNEX et al., 1998) é referência para toda uma organização que reside sob a forma de um repositório central, como uma biblioteca corporativa. Esses registros são: documentos em papel, relatórios, procedimentos e técnicas, normas descritas etc. Numa parte importante desta memória está a história cronológica de mudanças e revisões, pois refletem a evolução da cultura da organização e do processo de tomada de decisão.

Uma tipologia para situações de reutilização de conhecimento é apresentada por Markus (2001), na revisão de textos acadêmicos e práticos sobre a utilização dos documentos e repositórios de conhecimento, no qual a autora sugere que há pelo menos quatro diferentes tipos de situações em que o conhecimento é reutilizado, onde as características básicas para diferenciar esses tipos são o reutilizador do

conhecimento (em relação à fonte ou produtor de conhecimento) e os fins de reutilização.

Assim, os quatro tipos são: (1) reutilização dos produtores de conhecimento compartilhado, (2) reutilização compartilhada de profissionais do setor, (3) a reutilização por novatos em busca de conhecimentos, e a (3) reutilização secundária quanto àqueles que empregam a mineração para obter conhecimento.

Segundo Tim Berners-Lee (2004) e Oliveira (2002), na *Web* Atual, o reuso e integração da informação disponível na Internet, bancos de dados e em outros formatos e suportes torna-se possível quando é embutida semântica na estrutura desses documentos.

Apresentar as estruturas que viabilizem a compreensão e o gerenciamento de conteúdos digitais dará maior valorização semântica, o que permitirá uma resposta melhor e mais precisa no uso de coletores ou qualquer outra ferramenta capaz de relacionar termos, conceitos, paravras-chave etc. para a busca e recuperação da informação.

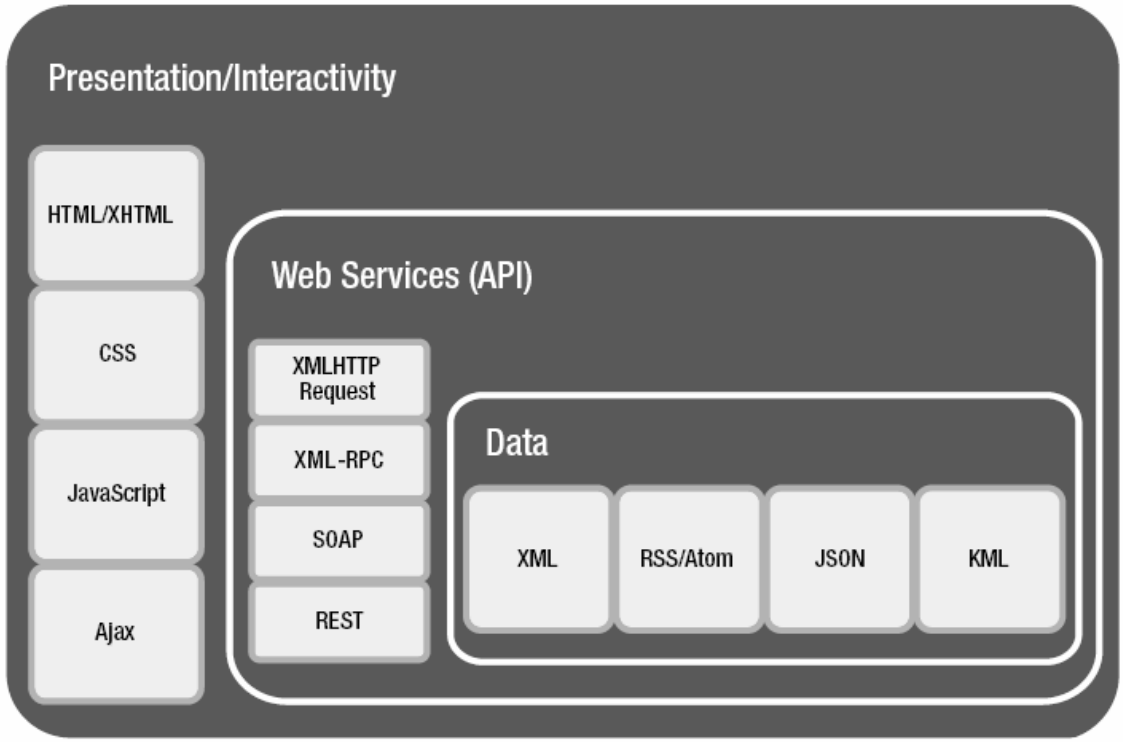

Figura 5: Arquitetura de camada *mashup*
Fonte: GRIFFIN (2008, p. 10)

Em 2007, o *Institute of Electrical and Electronics Engineers* (IEEE) − como uma associação profissional do mundo dedicada ao avanço da inovação tecnológica e excelência para o benefício da humanidade − promoveu a Conferência Internacional sobre Reuso e Integração da Informação (RII).

Neste contexto, segundo IEEE (2007, IRI-07), reutilizar e integrar informações é maximizar a disponibilidade de informações na criação de conhecimento. A finalidade do reuso dessas informações e conhecimentos é buscar respostas e evidências que apontem possíveis soluções para novos problemas. Assim, o RII desempenha um papel fundamental para capturar, manter, integrar, validar, extrapolar e aplicar informações e conhecimentos, aumentando a capacidade de decisão em domínios de aplicação.

Nesta dissertação, reuso é visto como o processamento de documentos textuais digitais por meio de um conjunto de métodos que envolvem pessoas e *softwares*, de modo que o produto desse processamento seja reutilizado em outros contextos diferentes dos quais esses documentos foram originalmente planejados e concebidos.

4.6 Mineração de dados e mineração de textos

Segundo Bath (2004, p. 331) “mineração de dados é parte de um processo através do qual informação é extraída de dados ou bancos de dados e usada na tomada de decisões numa variedade de contextos”. Nesta definição fica claro o contraste entre dados simplesmente e informação, ou seja, sentido.

Mineração de textos é uma aplicação, em dados textuais, da mineração de dados. Seu uso em domínios que produzem muitos textos como a biomedicina se torna obrigatório. Para Spasic et al. (2005), o aumento expressivo da literatura biomédica torna difícil localizar, recuperar e gerenciar informações sem utilizar a mineração de textos, que tem como objetivo “destilar informação, extrair fatos, descobrir relações implícitas e gerar hipóteses relevantes para as necessidades dos usuários” (SPASIC et al. ,2005, p. 241).

Mineração de dados e mineração de textos são métodos informáticos e pressupõem naturalmente que os dados a serem minerados estejam em formato digital.

5 METODOLOGIA

O objetivo do estudo permitiu uma opção metodológica que privilegiasse os aspectos qualitativos, de cunho exploratório (porque não se sabe o tamanho deste universo). A busca no *Scholar Google*[11] por "*semantic structure of digital texts*" que descreveria literalmente o interesse de pesquisa não encontrou nenhuma referência. Portanto, o critério de seleção estará baseado na identificação de artigos que, apesar de não falarem exatamente no tema dessa pesquisa, possam oferecer subsídios para equacionar a questão de pesquisa.

Foram pesquisadas propostas, experiências, projetos, entre outros, que identifiquem métodos de estruturação de textos digitais.

As fontes utilizadas foram artigos de periódicos nacionais e estrangeiros, trabalhos em eventos, *pré-prints* e documentos armazenados em repositórios das áreas de Ciência da Informação,[12] Ciência da Computação,[13] portal Capes,[14] Google Scholar,[15] Citeseer[16] etc.

Como a questão levantada era bastante específica, temas correlatos foram também levantados: formatos textuais digitais, *Web* Semântica, linguagens de marcação, metadados, mineração de textos, anotação de textos, anotação semântica, nos idiomas português e inglês.

O material levantado foi objeto de leitura, fichamento dos textos e análise dos dados, procurando identificar, em cada um, métodos de estruturação de textos digitais.

Com o resultado da análise efetuada, é proposto um esquema classificatório, sob a forma de um conjunto de critérios, apresentado e discutido no início do capítulo 7, e com o qual são classificados/agregados os diferentes projetos, experiências, propostas e metodologias para processamento automático do conteúdo de documentos textuais encontrados na literatura, na sequência do mesmo capítulo.

---

[11] Pesquisa avançada do Google Acadêmico
[12] http://dgz.org.br/
[13] http://www.portaldeperiodicos.sibi.ufrj.br/
[14] http://www.periodicos.capes.gov.br/portugues/index.jsp
[15] http://scholar.google.com.br/
[16] http://citeseerx.ist.psu.edu/

6 REVISÃO DA LITERATURA

O propósito deste capítulo é realizar uma revisão da literatura sobre o objeto da pesquisa, ou seja, focando na identificação dos diferentes projetos, experiências, propostas e metodologias para processamento automático do conteúdo de documentos textuais**.**

A literatura levantada foi também analisada visando a classificar e sistematizar os métodos de estruturação de documentos textuais digitais encontrados e apresentar a formulação de uma proposta de sistematização dos métodos de estruturação de textos digitais; estes resultados serão sistematizados no capítulo 7.

Teixeira (1974) descreve o método de recuperação de informação baseado numa linguagem de busca e programas de computador. É nesse ambiente que o usuário codifica dados do seu interesse, manipulando palavras-chave e descritores. O material a ser recuperado está em artigos publicados nos periódicos, trabalhos apresentados em congressos, monografias, relatórios etc. Esse material é indexado a partir de componentes básicos como: palavras-chave, termos auxiliares e nome de autores em que a linguagem faz uso de operações lógicas, do tipo OU, E e MAS NÃO. Qualquer conjunto de dados textuais poderá formar a base de dados textual permitindo ao usuário formar conjuntos ou subconjuntos sobre as várias características pertencentes aos materiais indexados e aos conjuntos de dados a serem recuperados.

Silva e Milidiú (1991) optam por um modelo que emprega funções de crença para indexar e recuperar documentos. Uma função de crença é reconhecida no ambiente de Inteligência Artificial como uma forma de representar e atualizar o conhecimento impreciso. Sendo assim, uma função de crença pode ser definida sobre um subconjunto dos descritores sem termos mais específicos, e os seus subconjuntos podem ser vistos como termos mais gerais ou como termos relacionados. A função aplicada sobre um subconjunto dos descritores sem termos mais específicos pode, então, segundo os autores, estimar o conteúdo semântico de um documento. O modelo é baseado em um vocabulário controlado e na frequência dos termos em cada documento. Um descritor desse vocabulário, um termo escolhido entre seus sinônimos, pode ter um subconjunto de descritores mais gerais, formando um

subconjunto de descritores específicos e outro de descritores relacionados. Os arranjos ou conjuntos formados por documentos textuais digitais, o seu conteúdo semântico e as consultas feitas às bases de documentos textuais digitais são representadas por funções de crença distintas.

Brito (1992) aborda a automatização de processos em linguagem natural, isto é, transpor automaticamente um texto, em linguagem natural, para uma metalinguagem de análise gramatical. Para tal, se faz uso do conteúdo descritivo e informacional de textos digitais, com a finalidade de automatizar a indexação e fornecer dados para sistemas de informação documentária, cujos elementos principais são obtidos pelo uso de sintagmas[17] nominais. O autor desenvolve uma ferramenta, a partir de uma linguagem de programação Starlet, para processos e análises morfossintáticos baseados no processamento (computacional) de linguagem natural (PLN). O conjunto de textos digitais foi obtido da *Agence France Presse* (AFP News Brieves) e forma um conjunto de testes para o analisador morfossintático. Todo o processo permitirá uma melhor interação com a formalização de fenômenos linguísticos e a sua aplicação com regras de análise fornecerá uma descrição mais sintática para programas adaptados ao meio computadorizado e às necessidades linguísticas.

O artigo de Pisanelli et al*.* (1998) faz um breve levantamento da análise ontológica e integração de várias terminologias. O trabalho realizado culminou na elaboração de uma metodologia denominada ONIONS. Esta metodologia tem o propósito de integrar termos e explorar uma biblioteca de teorias genéricas. No presente trabalho foram revistos: formalização de ontologias no domínio da Inteligência Artificial, filosofia, linguística e ciências cognitivas. Para as análises ontológicas, além dos dados obtidos numa biblioteca de teorias genéricas, terminologias médicas de alto nível também foram integradas, e algumas terminologias médicas também foram consideradas, como as de baixo nível, obtidas na *Unified Medical Language System* (UMLS), projeto realizado pela Biblioteca Nacional de Medicina dos EUA. O estudo visa a estabelecer relações entre as teorias genéricas, metatesauros, terminologias médicas, estejam elas sob a forma de listas ou agrupadas sistematicamente, fazendo uso da linguagem médica e, por fim, identificar as relações entre objetos,

[17] Sintagma é a combinação sucessiva de diversos elementos em um só discurso da cadeia fônica.

classes, linguagens de representação, consultas SQL a partir de um banco de dados, onde possam ser analisados padrões de tipos semânticos para relações de conceituação e classificação.

Wives (1999) apresenta seu estudo sobre métodos de agrupamento de objetos textuais digitais. Esses objetos são organizados automaticamente em grupos de acordo com a sua similaridade, isto é, faz uso de uma fórmula não convencional no cálculo de similaridade (lógica *fuzzy*) dos documentos textuais digitais. A metodologia de aplicação empregada no agrupamento deriva desse estudo e permite o seu reuso nas etapas seguintes sobre os mesmos dados com diferentes parâmetros.

Weeber M. (2000) em seu artigo reporta o desenvolvimento do sistema DAD, um conceito baseado em processamento de linguagem natural para citações do *PubMed*. O sistema DAD, a partir de uma doença, busca, em sites, relações e ações ligadas a essa doença (palavras-chave), de onde surgem novas hipóteses. As literaturas são examinadas de forma semelhante à Lindsay Gordon e Swanson. Os conceitos UMLS são empregados com a finalidade de obter conhecimento biométicо a partir das relações indicadas. A técnica não faz uso de uma lista excludente de vocábulos como preposições e advérbios, mas considera o uso de palavras compostas como, por exemplo, “pressão arterial”. Conceitos UMLS são utilizados e foram atribuídos a estes um ou mais tipos semânticos que, segundo os autores, formam um filtro semântico com objetivo de orientar o processamento, de forma que não se percam dados devido às inúmeras alternativas possíveis. Uma vez minimizadas essas alternativas, o usuário terá que fazer sua interpretação de acordo com o conhecimento atual e as suas metas. É papel do usuário fazer uma interpretação da lista de caminhos possíveis sugeridos pelo computador. O sistema emprega um modelo cliente-servidor em que o cliente é qualquer navegador *Web* padrão. Os recursos do sistema são bancos de dados, e as ferramentas da PNL e conectividade ao *PubMed* são as principais fontes de dados. Emprega-se o *software MetaMap* que, além do processamento de textos em linguagem natural, é utilizado para consultas e citações *PubMed*, onde conceitos UMLS são mapeados ou traduzidos. Identificam-se frases variantes, sistemas para avaliação linguisticamente métrica, onde são encontrados conceitos Metatesauro mais próximo do texto. O

MetaMap UMLS é baseado em dicionário de sinônimos e léxico, e compreende também o processo de geração de consultas *query terms.*

Kuramoto (2002) apresenta a proposta de desenvolvimento de um protótipo de interface de busca utilizando os sintagmas nominais como forma de acesso à informação. Seu conjunto de dados foi obtido de forma manual em 15 artigos da revista **Ciência da Informação**, dos quais foram extraídos 8.800 sintagmas nominais. Para isso, foi empregada uma abordagem lógico-semântica, em que o autor reconhece, extrai e indexa os sintagmas nominais. Duas alternativas são utilizadas para o experimento: implementar uma indexação automática baseada em palavras, apenas substituindo os índices contendo as palavras isoladas por índices contendo sintagmas nominais, modelos de classificação ou *ranking* como o modelo vetorial aplicado aos sintagmas nominais como unidade básica de acesso à informação; e a organização hierárquica em árvore dos sintagmas nominais. O arranjo de documentos textuais recuperados ganha sentido de acordo com a sua relação, que é estabelecida em qualquer uma das duas alternativas apresentadas.

Olsson et al. (2002) apresentam quatro noções para ajustar o resultado de técnicas métricas empregadas para medir o desempenho de marcadores de nomes de proteínas, considerando certas características do marcador sob avaliação. Os autores empregam sistemas como o *Yet another protein name extractor* (Yapex), desenvolvido pelo grupo, e o corpo textual para a realização do experimento é formado por um total de 101 (cento e um) resumos obtidos por meio de consulta do MEDLINE, e em todos os resumos, os nomes de proteínas foram marcados por especialistas no domínio. A metodologia segue com a análise do léxico para obter termos fundamentais que fazem parte do nome da proteína e de possíveis candidatos, baseados em regras pré-estabelecidas, aplicando filtros e bases de conhecimento com foco nas fórmulas químicas, nomes de substâncias químicas, referências bibliográficas etc. Os nomes de proteínas identificados formam um dicionário local e, para todos os documentos, esse dicionário é usado para detectar os nomes de proteínas que não foram percebidos no processo, o reuso de termos relevantes obtidos por meio da técnica, no próprio refinamento do processo de formação do dicionário e próprio do processo seletivo. A detecção de nomes de proteína em textos científicos servirá como apoio e reorganização de estratégias de

buscar, operacionalizar e navegar em bases de dados de alta qualidade como as de sequências de proteínas anotadas SWISS–PROT.

Tardelli et al. (2002) empregam técnicas apoiadas na Descoberta Baseada em Literatura de Swanson et al. (2006), em que são formuladas hipóteses científicas por meio das buscas e conexões entre estruturas de conhecimento publicamente disponíveis na medicina, o tipo MEDLINE, por exemplo. São empregados vocabulários controlados MeSH estruturados em 15 categorias hierárquicas, além de mais duas outras que compõem a versão traduzida do MeSH, o DeCS, utilizado pelas fontes de informação que compõem a Biblioteca Virtual em Saúde. O método está fundamentado na base de dados MeSH, na indexação humana e no registro de bibliografias. Assim, uma vez relacionado, um termo pode pertencer a mais de uma categoria hierárquica ou a outros ramos de uma mesma categoria hierárquica. A estrutura prevê e utiliza sistematicamente termos e termos seguidos de seus qualificadores de assunto. Os qualificadores de assunto representam o aspecto do assunto, o que permitirá um filtro restritivo com a finalidade de refinar o conjunto que se deseja recuperar, considerando também a nota de escopo ou anotação semântica e os elementos que vinculam o conjunto a outras literaturas complementares. Todas as relações visam a orientar o usuário que demanda informação quanto ao resultado e forma da seleção do conjunto buscado.

No trabalho de Oliveira et al. (2003) são formulados critérios sistemáticos para identificar preposições complexas ou locuções prepositivas, isto é, identificar uma combinação de duas ou mais palavras com uma função semântica específica para auxiliar na consulta a textos e na recuperação da informação. Para tal, constituem um léxico o conjunto de palavras, palavras compostas e frases que podem ser visualizados por meio de processador de linguagem natural e aplicados recursos de edição de textos, além de critérios estabelecidos para reconhecer preposições complexas ou locuções prepositivas: léxico *a priori*; substituição; valência do verbo anterior; inserção de um determinante. Os autores aplicam esse conjunto de recursos a um corpo textual em português brasileiro compilado pelo Núcleo Interinstitucional de Linguística Computacional (NILC). O método está baseado na própria estrutura sintática da língua portuguesa, e os resultados trazem evidências quanto ao reconhecimento de preposições complexas como uma classe aberta, isto

é, um conjunto finito ou uma lista finita de preposições que não contempla variações existentes na língua portuguesa, o que torna uma lista finita simples, ou universalmente conhecida, incompleta para se obterem resultados mais precisos em buscas textuais.

Celec (2004) apresenta o procedimento para Análise de Variância Rítmica (ANORVA) utilizado na detecção de variações cíclicas em séries temporais biológicas e a quantificação da sua probabilidade. É um método simples para a detecção de componentes periódicos de conjuntos de dados biológicos. O autor emprega o método em um conjunto de dados que representam trabalhos publicados por dia no *PubMed* em um determinado período. A partir dos conjuntos de dados completos e do uso de palavras-chave, o conjunto é submetido a métodos estatísticos para ajustes de entradas, interpolação; a cálculo da variância e a um conjunto de sistemas para tratamento estatístico, cujo objetivo é simplificar o uso de recursos matemáticos complexos. O resultado é comparado com um conjunto de dados ao acaso e apresenta evidências significativas do número de trabalhos publicados diariamente mostra um ritmo de publicações num período de sete dias "*circaseptan*".

Por meio de um método computacional, Wren et al. (2004) identificam as relações em relatórios científicos, de maneira que grandes conjuntos de relações, com itens como genes, doenças, fenótipos e produtos químicos extraídos de registros do MEDLINE sejam identificados, mensurados e estatisticamente classificados, estabelecendo, assim, a relevância do conjunto. O método consiste basicamente em associar esses itens que ocorrem concomitantemente a outros itens e verificar se há relação com outros de domínios diferentes. Assim, tem-se o uso de bases de dados como o MEDLINE, observados os títulos e resumos, o emprego da lógica fuzzy com objetivo de ponderar a importância de itens coocorrentes dentro de um mesmo registro na base de dados, ferramentas de desenvolvimento de *software* básico, banco de dados e consultas escritas em SQL. Todo esse conjunto das fontes de dados é armazenado, processado por meio de algoritmos e representado o seu resultado por meio de grafos.

Dieng-Kuntz et al. (2006) descrevem um método de reconstituição de uma ontologia médica por meio da tradução de um banco de dados médicos, com recursos

direcionados para a linguagem *Resource Description Framework* **(**RDF**).** O processo de interpretação dessa ontologia é feito por meio do processamento de linguagem natural de um corpus textual. A ferramenta *Virtual Staffa* foi criada para auxiliar no diagnóstico cooperativo, em que as várias hipóteses diagnósticas e terapêuticas podem ser representadas por um gráfico, usando o conceito de uma ontologia denominada Nautilus. As ontologias são usadas para representar conceitos compartilhados por membros de uma rede. A representação se dá por meio de recursos RDFs junto com as anotações oriundas do prontuário do paciente; motores de busca para consultar ontologias e anotações semânticas registradas sob a estrutura RDF; e finalmente intercâmbio de dados feito por XML. O que dá sentido ao relacionamento dos conjuntos acima é a possibilidade de estabelecer repositórios e responder a questões relativas à memória organizacional, apoio cooperativo e raciocínio baseado em casos.

No artigo de Spasic (2005), as ontologias são empregadas para a mineração de textos e aplicações na biomedicina. O método consiste em agrupar termos, estabelecer as relações desses termos e suas variações e eliminar ambiguidades. As ontologias fornecem descrições legíveis por máquina dos conceitos biomédicos e suas relações, ligando os termos específicos do domínio, isto é, representa textualmente esses conceitos, e as suas descrições oferecem uma plataforma para interpretação semântica. Uma camada semântica e o uso de ontologias permitem a extração de informações interpretáveis sobre conceitos biomédicos em oposição a simples correlações descobertas por mineração de dados textuais, utilizando informação estatística sobre coocorrências entre as classes específicas de termos biomédicos.

Araújo e Tarapanoff (2006) fazem uso do método de comparação entre a indexação manual e a mineração de textos, por meio de recursos estatísticos para análise do índice de precisão de resposta no processo de busca e recuperação da informação. O Centro de Referência e Informação em Habitação (Infohab) foi utilizado como fonte de dados, cuja base sobre habitação, saneamento e urbanização é indexada manualmente a partir de uma lista de palavras-chave previamente definida. Um protótipo foi desenvolvido para processar itens bibliográficos que correspondem às teses e dissertações contidas no Infohab, o que permitiu a aplicação do *software*

BR/Search, responsável pela mineração de textos. O método visa a auxiliar especialistas na identificação do conteúdo, isto é, dos assuntos contidos na base de dados.

O artigo de Swanson et al. (2006) é um dos pioneiros na antevisão das potencialidades da análise da literatura para identificar descobertas. Nesse artigo, os autores propõem um método que tem uma entrada de dados formada por dois conjuntos disjuntos de registros (A, C) da base de dados MEDLINE; o método irá produzir uma lista de palavras do título e frases (B) que são comuns para A e C, e apresentar um contexto onde títulos que ocorrem em termos de B ocorrem também em A e C. Assim, especialistas podem identificar relações A-B e B-C por meio dos termos de B, e sugerir relações indiretas C-A. O *Medical Subject Headings* (MeSH) é empregado nas entradas de registros por coesão textual, isto é, dar um contexto na medida que o número de relações aumenta. Dessa forma, o especialista pode avaliar se os títulos sugeridos de uma relação A-C são plausíveis baseado na densidade de termos do MeSH cujos registros A-B e B-C são comuns.

Diederich et al. (2007) analisam um conjunto de dados registrados eletronicamente cuja fonte são entrevistas na área da psiquiatria, dados, ou conversas gravadas em algum formato. Os candidatos verbalizam ou escrevem uma passagem ou história a partir de um tema e de um conjunto semântico previamente determinado. A transcrição desse conjunto de dados para ASCII permite a aplicação de métodos em dois experimentos. A partir dos atributos de valor dados a um registro eletrônico, a frequência com que a palavra ocorre na amostra da fala transcrita, no uso de uma lista de *stopword* e dos segmentos de texto de tamanho variável, extraídos aleatoriamente e convertidos para um formato de dados que traduz o aprendizado da máquina, chegou-se a uma decisão binária aprendida: esquizofrenia x controle. A extração de informações de diagnóstico de relatórios psiquiátricos consiste em gerar uma lista ordenada de “n” palavras que não estão em uma lista de palavras irrelevantes, classificá-las, aplicar e fazer análise de *cluster*, *stopwords*, excluir países e regiões, excluir também palavras de emoção com alta frequência como "feliz" e "triste". O modelo associado para determinar *clusters* abrange tanto os atributos numéricos e categóricos e constitui uma mistura de modelos gaussianos e

multinomial. O algoritmo sinaliza o surgimento de uma nova e inusitada lista de palavras e evidências de um distúrbio psiquiátrico.

Lima (2007) trabalha a construção e implementação de um Modelo Hipertextual para Organização de Documentos (MHTX) cujo objetivo é auxiliar a organização e representação do conhecimento humano em hipertextos. O trabalho de pesquisa de Lima é baseado em quatro referenciais: a Teoria da Análise Facetada, a Teoria dos Mapas Conceituais, a estrutura semântica de *links* hipertextuais e as normas técnicas da Associação Brasileira de Normas Técnicas (ABNT). O protótipo MHTX envolve uma estrutura semântica e um sumário expandido obtido por meio do sumário de uma tese de doutorado onde são agregados pontos de acesso (marcadores ou *tags*), que agregam mais informação (anotação semântica) para auxiliar o usuário na busca pela informação desejada. Os autores também fazem uso de *softwares*: para a representação gráfica; na construção e estruturação hierárquica do Mapa Conceitual, *Star Tree Studio;* e finalmente para a base de dados textuais *Greenstone Digital Library*.

Segundo Sirin (2007), SPARQL é uma linguagem para consultas da *Web* Semântica em arquivos RDF que disponibiliza recursos semânticos em aplicações, principalmente quando se trata de uma grande massa de dados que precisam ser integrados e relacionados. O agrupamento e organização dos conjuntos ocorrem na entrada sistematizada da descrição de serviços, diretório de informações e pedidos. Esse agrupamento direcionado pelos métodos definidos na linguagem garantem uma saída com resultados unificados. Os recursos RDF e OWL promovem representações gráficas, integração de modelos e interoperabilidade de diversas fontes. A técnica consiste em dividir um vocabulário em conjuntos disjuntos e associar a outros conjuntos que denotem: propriedades de um objeto; tipos de dados; anotação semântica ou propriedades descritivas que caracterizam indivíduos - *Uniform Resource Identifier* (URI); conjuntos de literais; recursos oferecidos pela *Web Ontology Language - Description Logics* (OWL-DL), que também contempla recursos oferecidos pela OWL, restrições de cardinalidade, retrições sobre classes e construtores de classes do tipo união, complemento e interseção.

A OWL-DL fornece várias construções para gerar expressões e classes complexas para uma ontologia que contém um número finito de classes, propriedades e axiomas individuais. Existem muitas formas sintáticas que se ajustam adequadamente em OWL-DL, mas a maioria são frases que podem ser expressas como uma combinação de subclasses, propriedades e sub-propriedades e outras formas que podem ser traduzidas e interpretadas e o resultado fornece uma nova forma também adequada ao modelo.

Wang et al. (2007) abordam descrições de genes em diferentes fontes de dados e buscam determinar as semelhanças funcionais por meio de informações e anotações de dados a partir de fontes de dados heterogêneos. O método semântico consiste em codificar um termo da GO (significados biológicos) em um valor numérico e agregar os dados do tempo que o precede, isto é, estabelcer uma hierarquia de significados que incluirá também o termo específico. Algoritmos são largamente empregados nas mais diversas funções, como: visualizar graficamente o encadeamento de GO; medir a similaridade semântica dos termos de GO; e, com base na similaridade semântica de termos utilizados na anotação do gene GO, medir a similaridade funcional dos genes. Os valores obtidos na similaridade formam o agrupamento de genes que, no caso do presente artigo, o *saccharomyces* do banco de dados do genoma (SGD) foi processado, identificadas as semelhanças funcionais de genes e os resultados do agrupamento foram coerentes com as perspectivas humanas desejadas.

Baumgartner et al. (2008) investigam a importância das avaliações estruturais trazendo evidências de que estas são importantes pré-requisito para avanços no campo da mineração de textos. A partir da combinação de *software* e dados aplicados para analisarem complexas comparações, uma estrutura flexível e extensa foi criada a partir de uma base de códigos que tem como objetivo facilitar e melhorar a performance dessas funções responsáveis pelo processamento, isto é, da linguagem de processamento e da linguagem em geral, e da biomédica em particular. O emprego da tecnologia é destinado a estabelecer as relações de proteínas e da identificação do gene, isto é, detectar onde os nomes de genes aparecem no texto. A metodologia segue com o uso de um gerenciador de dados para processamento, anotações (anotações semânticas), bem como com o

processamento de texto livre. Um *framework* para arquitetura de gerenciamento de informação não estruturada (UIMA) e uma camada denominada *middleware* facilita a interação de diferentes tipos de ferramentas, resolvendo a questão da interoperabilidade de recursos que não estão ou não foram concebidos para interagir com as existentes. Para comparar, o sistema usa uma variedade de técnicas métricas utilizadas por Olsson et al. (2002) e, para estabelecer um *score* das anotações e reutilizar outras comparações métricas, emprega *Common Analysis Structure* (CAS). O CAS é uma estrutura de dados flexível capaz de armazenar o texto do documento, as anotações de um texto e metadados.

Bodenreider (2008) fornece exemplos típicos de ontologias biomédicas em ação e enfatiza o papel desempenhado por essas ontologias na gestão do conhecimento, integração de dados e apoio à decisão. Destaca o uso das ontologias como fonte de vocabulário para anotação de dados ou documentos de indexação. Além dos exemplos prototípicos do MeSH, usados para indexação da literatura biomédica e da GO na notação funcional de produtos dos genes, a anotação semântica ganha outro papel: o de formar outro conjunto de termos, vocábulos e palavras que podem ser usados para indexar ou estabelecer relações num determinado domínio léxico. As relações e codificações estão fundamentadas de acordo como são empregadas determinadas fontes desses dados. Assim, para alguns países a Classificação Internacional de Doenças (CID) tem sido utilizada para a codificação de morbidade e mortalidade e, mais recentemente, como um sistema de codificação para efeitos de reembolso; o SNOMED CT adotado como terminologia padrão para registros de saúde eletrônicos também tem sido avaliado como uma fonte de vocabulário para pesquisa clínica, UMLS, tesauros, metatesauros. O reuso desses dados e as suas relações com o processo de identificar automaticamente as menções de entidades de interesse no texto trazem ao processamento de linguagem natural (NLP) técnicas e sistemas de reconhecimento, buscando na área biomédica, principalmente, explorar as ricas fontes de vocabulário fornecido por ontologias biomédicas. Selecionar e integrar dados, interagir, garantir a interoperabilidade de sistemas e de semântica, anotar, mapear ontologias biomédicas são evidências para transformação e reuso da informação.

Capuano (2009) demonstra o uso de um sistema de recuperação da informação composto por uma base de índices textuais de um conjunto de documentos textuais. No caso, os textos utilizados no experimento (ambiente simulado) constituem resumos das apresentações dos encontros *IA Summit* (nos EUA) de 2005 a 2008. É aplicado um *software* de rede neural artificial que faz uso de conceitos da Teoria da Ressonância Adaptativa, cuja função é processar, ordenar e apresentar os resultados realizados a partir de consultas demandadas por um usuário. A demonstração faz uso das técnicas de redes neurais, resolução semântica com índices sintagmáticos SiRILiCO, processamento de sintagmas nominais para associação semântica de termos, e contempla teorias da linguística computacional e ontologias. A combinação desses recursos contribui para a formação de *clusters,* automação do processo de aprendizado, e não supervisionado, cujo intuito é o de responder a consultas de usuários em redes de computadores.

Kiyavitskaya et al. (2009) propõem um quadro semiautomático de anotação semântica para documentos textuais baseado num modelo de domínio semântico específico. São empregadas técnicas e ferramentas destinadas a análises de código e marcação. Por meio da arquitetura Cerno, o método analisa o documento, reconhece fatos básicos, faz interpretação com relação a um modelo de domínio semântico e, finalmente, o mapeamento de informações identificadas registradas em um banco de dados externo. A linguagem TXL expressa as transformações estruturais a partir de uma fonte de dados. A fragmentação do texto em linguagem natural produzida pelo Cerno é dividida em: documento, frase, parágrafos etc. e podem ter seus itens escolhidos pelo usuário para anotação semântica. O sistema de marcação é empregado a partir do reconhecimento de instâncias de conceitos, ou seja, anota as unidades de texto que contêm informações relevantes baseadas numa lista de nomes, conceitos e vocabulários de domínio-dependente. O esquema de anotação é construído anteriormente usando métodos de aprendizagem ou manualmente. O resultado pode ser mapeado com recursos semelhantes ao XML e XML Schema em que frases e expressões são conjugadas e conceitos devidamente associados. A compilação por meio da metodologia apresentada no artigo faz com que um conjunto formado por documentos textuais, páginas *Web* etc., seja tranformado em um grupo de conceitos independentes de um domínio.

Lucca (2009) propõe um método para desambiguidade semântica de palavras num corpo textual. O programa proposto admite que o contexto é que determina o significado da palavra e faz uso de um corpo textual Corpus HispanoAmericano de Español (CHADES). O autor utiliza um algoritmo que tem a função de extrair do corpus candidatos a sentidos lexicais, baseados em uma estrutura capaz de armazenar palavras e suas ligações entre palavras que constituem determinado contexto, isto é, denominado pelos autores *HiperThesaurus* que tem capacidade de armazenar palavras e as relações entre elas, permitindo a recuperação de informações contextuais, não apenas por palavras-chave, mas também por um agrupamento de palavras.

A pesquisa de Maia e Souza (2010) faz uso de um grupo de documentos textuais com a finalidade de aplicar um método capaz de classificá-los automaticamente usando sintagmas nominais. O método faz uso de uma ferramenta própria denominada OGMA, que segue com a elaboração de um texto de léxico da língua portuguesa que serviu como um marcador de texto; a construção deste léxico se deu a partir do vocabulário BR/ISPELL, que também foi utilizado para verificação ortográfica. A verificação ortográfica é feita com o uso de uma tabela de nomes e adjetivos, uma tabela de verbos e de um processo manual de digitação com base na gramática de Tufano (1990); assim foram reunidas palavras de diversas classes gramaticais. O *software* WEKA foi aplicado para um conjunto de algoritmos da área de inteligência artificial, classificação e agrupamento; Naive Bayes e Simplekmeans, para resultados estatísticos. O conjunto de métodos interrelaciona textos, tabelas, vocabulários e *stopwords*. Assim tem-se a extração automática de descritores por meios estatísticos e pontuação. O reuso desses descritores comparativamente a matrizes de documentos e o agrupamento de sintagmas ou o sintagma nominal representa o conjunto de documentos textuais empregados.

A proposta de Rico et al. (2009) tem a finalidade de definir a semântica dos documentos utilizados em negócios eletrônicos trocados entre os parceiros comerciais em um relacionamento colaborativo. Um perfil de diagramas padronizados UMLs, linguagens de marcação e uso de protocolos traduzem a associação de um ato de fala com mensagens de negócios que representam a intenção que um parceiro comercial tem em relação com os Documentos Eletrônicos

de Negócios (*Electronic Business Documents* - EBDs) baseados em XML Schema. Por meio dessa estrutura padronizada, uma instância da ontologia é acordada para um EBD correspondente às ontologias dos departamentos da empresa e vice-versa. O protocolo de interação fica encarregado de estabelecer relações relevantes de ontologias internas e ontologias EBD, perfazendo um alinhamento e criando as regras de conversão (CR) para as instâncias, isto é, uma transformação da ontologia em instâncias correspondentes e expressas em uma outra ontologia (intercâmbio de dados entre diferentes domínios). Então, em tempo de execução, o protocolo de interação tem de executar um processo para a tradução dos dados e cumprir a definição de regras de conversão que permitem a interoperabilidade dos EBDs em um nível semântico.

Samwald (2009) elabora uma consulta restritiva que permite recuperar parte significativa e relevante do conjunto de registros no PUBMED, cujos parâmetros buscam artigos sobre emoção e cognição. O método sugere que as palavras abreviadas são retomadas na sua forma completa, utilizando-se o algoritmo Schwartz & Hearst. As *tags* ou marcadores são empregadas para extrair os dados, realizar anotações semânticas seguindo as normas e padrões da *Web* Semântica e ainda considerar os termos MeSH associados ao artigo. O método dá uma reorganização desses novos dados formatando-os de acordo com os padrões RDF e SIOC.

A partir de 100 conjuntos de dados compostos por registros do MEDLINE, Zhu et al. (2009) buscam estabelecer um método para integrar documentos textuais digitais, isto é, formar um agrupamento de documentos a partir da informação semântica do MeSH (*Medical Subject Headings* dicionário de sinônimos). Para tal, o método consiste em: aplicar um algoritmo para medir similaridade semântica entre dois documentos usando o dicionário de sinônimos do MeSH; combinar as semelhanças semânticas e de conteúdo para gerar a matriz de similaridade integrada entre os documentos; aplicar uma abordagem espectral para *clustering* de documentos por meio da integração matriz de similaridade, além de outros recursos como funções estatísticas e o uso de recursos do MeSH.

Chen, Donglin et al. (2010) mostram um modelo ontológico de catálogos eletrônicos com a finalidade de projetar um sistema de serviço *semantic personalized e-Catalog*

Service System (SPECSS), que consegue combinar *User Personalized Catalog Ontology* (UPCO) e o domínio *Domain e-Catalog Ontology* (DECO), com base na ontologia integrada e foco em quatro tecnologias fundamentais: catálogo de ontologias de usuário personalizado, catálogo eletrônico local, e de estabelecimento de uma correspondência semântica entre esses e o catálogo eletrônico do sistema de consulta semântica (*Query Semantic System*) baseadas em banco de dados de catálogos heterogêneos. O método propõe o uso de uma ontologia de domínio baseada em OWL; estruturas hierárquicas tipo pai-filho para classificação; anotação semântica, para enriquecer e possibilitar uma relação semântica entre propriedades e ontologias de produtos. As consultas são obtidas por meio do SPARQL. A aplicação da ontologia em catálogos eletrônicos e consultas são objetos do presente trabalho onde esforços foram concentrados na teoria do catálogo eletrônico de consulta semântica e de um serviço de catálogos personalizado, que pode expressar a preferência e intenção de potenciais usuários na busca por determinados produtos.

O método descrito em Dogan e Lu (2010) reconhece palavras essenciais para um documento na perspectiva do usuário. Tal procedimento se dá por meio do método de aprendizagem de máquina que aprende características únicas no clique-palavras onde cada palavra foi representada por um conjunto de características que incluem tipo semântico, *tags* de discurso, frequência inversa para termos do documento, peso e localização no resumo a partir das características mais importantes. São avaliadas no modelo com seis meses do *PubMed click-through logs,* que é o conjunto de artigos altamente acessados de um grande conjunto de *logs* de consulta. Os dados coletados são uma relação da consulta do usuário com o artigo clicado, que é independente da classificação do artigo na página. O uso de algoritmos se fez necessário para melhorar a seleção e os resultados de classificação, além de identificar características importantes para o modelo. São empregados marcadores para parte do discurso, tipo semântico, frequência de palavras, uso do MetaMap para mapeamento de palavras e conceitos semânticos onde se estabeleceu uma relação binária.

Janet e Reddy (2010) propõem um método de indexação de textos para viabilizar sua recuperação e mineração de dados. Este método de indexação é baseado num

léxico e no tratamento estatístico de termos do texto. O modelo é semelhante a um cubo de dados e o *Data Mining* dá uma visão múltipla das relações, ou seja, dimensões e fatos estabelecidos após o processamento. As dimensões são perspectivas ou entidades relacionadas ao que se deseja manter registrado, e os fatos são medidas numéricas calculadas na agregação de dados correspondentes a dimensão-valor que são definidos num determinado ponto como a frequência de termos, medidas de ponderação etc. O *Cube Index* permite modelar textos e visualizá-lo em múltiplas dimensões. As dimensões são palavras associadas nos documentos presentes e indicam uma palavra, a próxima palavra e o documento e frase em que ocorre. O processo hierárquico se dá por conceitos no cubo de dados que define uma sequência de mapeamentos de um conjunto de conceitos de baixo nível para um nível mais alto e conceitos gerais. Assim, tem-se a dimensão palavra que é a frase; frase, parágrafo ou bloco de texto. Toda essa organização é dada por meio do emprego de técnicas de *hash* para busca binária, sistemas de classificação, léxicos, *stopwords*, frequência de ocorrências do léxico da coleção etc. Todo o conjunto fornecerá ao usuário a flexibilidade para ver o texto a partir de diferentes perspectivas. O *Cube Text* ou *Cube Index* permite um novo olhar nas relações obtidas entre palavras, frases e documentos textuais a partir do uso de técnicas de programação e funções estatísticas.

Lendvai (2011) identifica relações semânticas entre conceitos em textos médicos e mineração de documentos estruturados. A partir de um sistema de atendimento holandês, as referências são manualmente anotadas, e os artigos são associados a marcadores semânticos que descrevem domínios predefinidos a partir de marcadores conceituais denominados no cabeçalho. Quando um desses artigos é recuperado e possui tal anotação semântica, as seções correspondentes são consideradas visando a manter o relacionamento taxonômico a partir das subclasses ou subtipos definidos.

7 PROPOSTAS LEVANTADAS NA LITERATURA – RESULTADOS

O capítulo apresenta sistematicamente as propostas vistas anteriormente que foram levantadas na literatura, propondo agrupá-las nos diferentes métodos identificados, segundo os seguintes critérios:

Para isto, cada artigo foi analisado procurando responder às seguintes questões:

1. Em que conjunto de dados textuais o método descrito no artigo foi aplicado?
2. Como foi especificada a semântica a ser buscada no conjunto de dados textuais?

Na análise, buscando responder essas questões, para cada texto identificado no levantamento emergiram as seguintes classes de métodos: Mineração de textos, Anotação Semântica, Análise Semântica, Análise em Linguagem Natural e Tratamento Estatístico de textos. Portanto, passou-se a descrever as características gerais de cada método, cada uma das classes e a discussão e comentários dos artigos que se enquadram em cada uma das classes.

A análise e classificação propostas permitirão a sistemas que propõem reutilizar documentos textuais, em contextos semelhantes à motivação discutida na Introdução, uma primeira aproximação dos métodos utilizados, permitindo assim viabilizar novas interpretações e assimilações do conhecimento ali registrado, ou seja, a reutilização com propósito diferente do qual o conjunto de textos digitais foi originalmente concebido. Dessa forma, aumentam as possibilidades de novas interpretações quando se viabilizam operações homem-máquina para reconhecer padrões, relacionar variáveis, e minimizar incertezas apoiadas em métodos quantitativos, estatísticos e matemáticos.

A análise de cada artigo considerará também todo ou parte do método, ou ainda, a parte do método que predomine ou venha estabelecer relevância diante da possível interpretação que o seu uso possa representar diante da metodologia adotada.

7.1 Anotação semântica

Anotação semântica é o processo por meio do qual um leitor assinala e vincula anotações e comentários a um determinado texto; existem métodos que utilizam as tecnologias da *Web* Semântica, como RDF(s) (RDF, 2004) para anotar textos e permitir sua recuperação semântica baseada nessas anotações. A anotação semântica é um processo que pode ser aplicado a qualquer corpo textual eletrônico, isto é, anotar com relevância trechos de tamanho variável, um complemento textual, que orienta o homem na compreensão do conteúdo ao qual a anotação está relacionada ou dá a um programa de computador complementos necessários para processar essas anotações com intuito de permitir refinar ou ampliar adequadamente os resultados desejados por um determinado algoritmo.

Na anotação semântica, os metadados são meios que viabilizam recursos para prover dados (LIMA, 2007), e onde se atribui a eles conceitos descritos numa ontologia, lista de palavras previamente definidas, ou conceitos no domínio de determinada língua (PISANELLI et al., 1998; OLIVEIRA et al., 2003). No caso, a anotação vem com o intuito de auxiliar nas relações e dar significado conciso e alternativo ao conjunto de palavras demarcado por meio de uma nova camada que descreve o seu conteúdo.

Este método é também presente em outros trabalhos, como: Brito (1992), que faz uso da anotação em complemento aos métodos de linguagem natural; Olsson et al. (2002), que o emprega como mais um recurso auxiliar para uma variedade de técnicas métricas; Tardelli et al. (2002), que faz uso de qualificadores de assunto com intuito de um filtro restritivo para refinar o conjunto que se deseja recuperar, considerando notas de escopo; Wang et al. (2007), que usam informações e anotações de dados para buscar e determinar as semelhanças funcionais nas descrições de genes em bases de dados e por fim Baumgartner et al. (2008) que, de certa forma, aplica um pouco de cada um dos usos citados para investigar a importância das avaliações estruturais.

Por outro lado, Diederich et al. (2007), Rico et al. (2009), Chen et al. (2010) exploram o conteúdo anotado, num processo semiautomático; Dieng-Kuntz et al. (2006), manualmente; Lendvai (2011) ou simplesmente selecionando por meio de

outros *softwares* de mineração; Kiyavitskaya et al. (2009) trazendo em seus artigos a importância da anotanção semântica e sua relavância em trabalhos acadêmicos que possam, de alguma forma, ser complementares e complementados com informações e características específicas de uma cultura, domínio, linguagem e aspectos técnicos completamente diferentes.

Sendo aplicável a qualquer tipo de texto digital, o processo de anotação semântica utilizado por Tardelli et al. (2002) acrescenta ao documento o marcador descritivo, ou seja, uma camada que descreve o seu conteúdo de acordo com o contexto e interpretação de um ser humano, no caso, um especialista responsável por completar com informação relevante o termo, palavra ou frase do documento, isto é, prover um conjunto de dados que sejam processados por *softwares* ou sistemas de computação que permitam uma melhor compreensão sobre o documento e suas partes, como também associá-lo a ontologias.

7.2 Mineração de textos

A mineração de textos geralmente é usada como parte de um processo que busca dar sentido a um conjunto de textos digitais dispostos nas mais diversas formas e suportes, por exemplo: frases sem verbo, bases de dados, páginas HTML, e-mails etc. Tem como função submeter esse conjunto a métodos para reorganizar e explorar os seus conteúdos digitais.

Para Teixeira (1974) sistemas de recuperação da informação, na sua interface com o usuário ou com outros sistemas, demandam por um método simples para buscar e recuperar informação, uma linguagem de busca que faça uso dos recursos lógicos e considere elementos comuns às referências bibliográficas como: autores, assuntos, ou ainda componentes agrupados por letras do alfabeto, dígitos, caracteres especiais etc. São elementos que, ao serem identificados no texto digital, reconhecidos numa leitura automática sequencial, contam com delimitadores representados por caracteres especiais ou com a posição de uma palavra no arquivo, e ainda na aplicação de operadores lógicos como: “E” ou “OU” poderão orientar sistemas de busca e recuperação da informação na identificação de documentos digitais.

Na prática, Samwald (2009) propõe uma aplicação para um método de consulta largamente utilizado para se obter dados relevantes de uma base de dados, ou seja, o emprego de uma consulta (*query*) estruturada em SQL. A aplicação desse método implica uma melhor definição para os parâmetros dessa consulta. Estes devem permitir individualizar ou isolar diferentes unidades textuais, isto é, uma palavra ou um termo previamente determinado como relevante. Estes parâmetros, no caso PARAM A e PARAM B, orientam a busca nos registros e compõem uma operação lógica usando o formato *query* que permite recortar um texto ou um conjunto de textos registrados eletronicamente em bancos de dados. No caso, o conjunto de textos resultante é aceito como válido e dá um novo sentido aos registros recuperados, a partir das palavras e termos empregados que podem ser submetidos a outros procedimentos estatísticos. No entanto, o novo sentido dado a esse novo conjunto recuperado já está realizado, uma vez que a precisão determinada pela forma como foi construída a consulta garante que o novo conjunto de dados recuperado seja significativamente satisfatório para atender a demanda desejada.

Os parâmetros utilizados nas consultas, isto é, os verbetes, termos, palavras-chaves empregados na linguagem utilizada pelo usuário ou nos sistemas de recuperação funcionarão no processo como uma espécie de filtro. Todas as repetições ocorrerão, caso não haja um refinamento dessa mesma linguagem, uma melhor seleção dos parâmetros desejados na orientação da consulta ou no refinamento da fonte que se deseja obter os documentos textuais desejados. A forma da linguagem facilita a recuperação, e toda a complexidade está na seleção dos parâmetros. O método é uma maneira de garimpar documentos textuais digitais, de modo que as formas como estes são armazenados sejam computadas sem maiores restrições quanto a suporte ou formato encontrado.

Sirin (2007) faz uso dos arquivos RDF e disponibiliza recursos semânticos em aplicações por meio da linguagem para consultas da *Web* Semântica, denominada SPARQL. O problema da interoperabilidade (UREN et al*.,* 2005; RENEAR; PALMER, 2009) é principalmente do tratamento de uma volumosa massa de dados (MARCONDES; SAYÃO, 2001) que precisa ser integrada e relacionada dentro de um processo que antecede o ato de minerar dados. Assim, a solução é dada pelo agrupamento e reorganização dos conjuntos, numa entrada sistematizada da

descrição de serviços, e por métodos definidos na linguagem e assim garantir uma saída com resultados unificados. Os recursos RDF e OWL promovem representações gráficas, integração de modelos e interoperabilidade das diversas fontes de dados. A linguagem SPARQL amplia a consulta em documentos no formato RDF/RDFS, representando virtualmente alguns conceitos e possibilitando que estes sejam identificados unicamente na *Web*. A vantagem apresentada no recurso RDF é a possibilidade de declararmos sentenças compostas por recursos, propriedades e sentenças, além de permitir a visualização dessas relações por meio de gráficos. Uma desvantagem ou limitação apresentada pelo recurso RDF é que não há uma representação completa de um domínio. O que se pode realizar sobre a sua composição é a construção de sentenças e não ser permissivo, numa forma mais ampla, a generalização sobre seus termos, isto é, agregar a uma sublasse que o represente ou o identifique num determinado grupo, problema proposto e com alternativas em RDFS. O SPARQL traduz uma consulta que compreende a estrutura semântica e sintática de documentos RDF/RDFS, podendo traduzir as consultas feitas em SPARQL para o modelo de uma linguagem natural.

Araújo e Tarapanoff (2006) e Capuano (2009) entendem a mineração como um processo de recuperação de itens. Embora os autores utilizem bases de índices textuais com objetivo de melhorar a precisão no processo de recuperação da informação, consideram também a participação do especialista, a figura humana, a sua contribuição significativa no processo de indexação, categorização com uso de palavras-chave para representação de um documento quanto ao seu uso na mineração de dados. A utilização de sistemas ocorre em ambos, no entanto, o método explorado por Capuano (2009) visa a simular o processo de recuperação fazendo uso de uma base de índices textuais, aliado a um sistema de inteligência artificial. Por outro lado, Araújo e Tarapanoff (2006) buscam comparar resultados em função da precisão no uso da mineração de textos, tendo em vista a aplicação de sistemas para esta ação. Independente dos objetivos dos autores, em relação à proposta de cada trabalho, o uso de índices textuais compostos automaticamente ou manualmente são conjuntos de palavras ou termos específicos, devidamente classificados e organizados de acordo com a especificidade de cada domínio, e prontos para auxiliar técnicas, sistemas, e outras formas e métodos para recuperar informação, dar uma resposta a um questionamento desejado, isto é, em qual grupo

textual está contido o termo e qual determinado termo ou conjunto de termos melhor representa um conjunto textual. A interpretação após o processamento de um conjunto de dados é trabalho do pesquisador, e este é responsável pelas diretrizes tanto antes quanto durante e depois do processamento de dados, respeitando as limitações que os sistemas venham a apresentar em função de como foram programados para auxiliar o trabalho e a qualidade dos dados que esses sistemas necessitam para o processamento. Assim, todo o conjunto trabalhará harmoniosamente e visando a atender ao máximo as expectativas criadas durante o estudo.

Lucca (2009) ressalta que o desenvolvimento de um modelo de ferramenta pode trazer soluções para problemas relacionados à lexicografia e ao agrupamento de contextos, ou seja, a formação de agrupamentos está baseada no contexto e é o contexto que dá significado à palavra. Então, o estudo baseado em *HyperThesaurus*, no caso, definido no artigo como uma estrutura capaz de armazenar palavras ou repositório de conhecimento e elos ou ligações entre palavras que constituem determinado contexto, conta com um número de verbetes oriundos do *Diccionario Salamanca de la Lengua Española* (DSLE).

Segundo o autor, o modelo baseado em corpus sem anotação vale para todas as palavras a desambiguar, e atende ao pressuposto de que qualquer palavra pode ganhar sentidos em vários contextos, portanto, toda e qualquer palavra pode se ligar a inúmeras fontes de informação, significados, e informações no campo semântico, sintático, contextual e estilístico.

Uma palavra ou conjunto de palavras constitui um nódulo que orienta a busca no corpo textual por frases ou orações que tenham uma pontuação forte - critério definido para frases que possuem pontos finais, ponto de exclamação, interrogação e reticências - e após esses sinais de pontuação, haja uma nova frase ou período. E nesse caso, especificamente, a frase foi definida como enunciado linguístico capaz de transmitir ideias. A adoção de tal método estabelece o peso e a relevância de um documento recuperado, uma vez que, não só a palavra ou palavra-chave responde a um processo de mineração, mas também considera o conteúdo do texto por meio de

perguntas e respostas que podem trazer significados relacionados ao título ou resumo de um documento, por exemplo.

Sendo o computador incapaz de compreender o sentido de uma palavra independente do contexto onde ela está impressa, tendo em vista que o computador foi concebido, ainda na sua atual estrutura, para processar sinais ou códigos numéricos, o ponto forte do artigo de Lucca (2009) é a utilização de um método ou parte do método que proporcione ao ser humano, especialista ou não em determinado domínio, associar letras e números, palavras e frases, e todas as variações possíveis dessa combinação textual, em diversos contextos. E ainda possibilitar o agrupamento de documentos de acordo com o entendimento do indivíduo que participa desse processo 'do fazer sentido' ao aplicar o resultado obtido por meio de *software*, bases de dados, dicionários, e o processamento de diversos algoritmos.

7.3 Análise Semântica

Resolver ambiguidades e analisar expressões em documentos textuais digitais são as principais características destacadas nesse grupo. O uso de um método ou o seu desenvolvimento − a partir de modelos que sigam uma lógica adequada em comparação ou auxílio a um modelo conceitual relevante ao usuário que o concebe, no sentido de que o seu resultado seja obtido empregando relações entre palavras, sintagmas, termos ou até mesmo *links* encontrados − deve dar algum sentido a um agrupamento de documentos textos digitais buscados ou recuperados.

A questão lógico-semântica abordada por Kuramoto (2002), a partir da indexação manual, reafirma o problema citado por Marcondes e Sayão (2001) sobre o uso de interfaces, mais o pressuposto da interação entre um usuário humano e bases de dados, e a indexação automática baseada em palavras. Ao recuperar e dar sentido aos arranjos de documentos textuais digitais, Kuramoto (2002) utiliza, numa de suas opções, o *ranking* que também é explorado por Silva e Milidiú (1991) e Swanson et al. (2006). O conjunto de dados ganha sentido a partir da extração de 8.800 sintagmas selecionados de 15 artigos da revista **Ciência da Informação**, em um processo criterioso de leitura, análise e substituição para indexar e avaliar métodos de busca e recuperação da informação, o reuso da informação a partir de métodos

cujas bases dependem da colaboração e avaliação, com alguma especificidade e técnicas inerentes, do ser humano, semelhante à proposta de Weber (2000)..

7.4 Análise em Linguagem Natural

Para que um sistema computacional interprete uma combinação de palavras em linguagem natural, uma frase ou palavra, é fundamental que esse sistema possa estabelecer e considerar no seu escopo informações de aspecto morfológico, sintático e semântico. Brito (1992), Weeber et al. (2000) e Araújo e Tarapanoff (2006) fazem uso de sistemas computacionais com intuito de organizar e agilizar os métodos empregados. No entanto, Dogan e Lu (2010) usam recursos desses sistemas computacionais para coletar palavras e identificar possíveis palavras-chave na perspectiva do usuário. É Um método simples em que o usuário navega ou simplesmente aciona com um clique do mouse palavras relevantes, à medida que avança na compreensão do texto digital ou por meio de um *link* desejado. A reorganização dessas palavras permite um *ranking* (KURAMOTO, 2002; SILVA; MILIDIÚ, 1991; SWANSON et al., 2006) de documentos textuais digitais com sua relevância definida sob o ponto de vista do homem que lê o texto digital e conduz a seleção a partir da sua compreensão, vontade e desejos. “[...] Nós, humanos, sabemos que em alguma parte de um texto existe uma significação, conhecimentos, que podemos facilmente extrair por meio de operações naturais como a leitura.” (BRITO, 1992, p. 224). Há, no método, a viabilidade da construção de palavras-chave indexadas por uma nova interpretação de documentos textuais digitais que serão selecionados sob o processo de leitura semiautomático, realizado por um especialista, pesquisador ou indivíduo interessado em agrupar a partir de uma temática ou, simplesmete, da leitura de uma sequência de documentos textuais digitais definida durante uma busca e recuperação da informação, isto é, o processo de aperfeiçoamento da indexação de documentos apontado em uma das conclusões de Araújo e Tarapanoff (2006).

A seleção de palavras não muda o conteúdo dos documentos. A seleção apenas reorganiza esses documentos e, no que tange ao processamento simbólico da linguagem natural, os parses são analisadores sintáticos baseados em regras rígidas. Brito (1992), sobre a análise de textos em linguagem natural, explora a

gramática de afixos baseada em uma estrutura lógica e formalizada por meio de grafos. A estreita relação é feita por meio da ligação da gramática do sintagma nominal até o modelo para a realização de um analisador morfossintático, seguindo a rigidez dos compiladores aproximando a interpretação do modelo gramatical ao das linguagens de programação. Isto corresponde a um tratamento para a unificação, restrição e controle do processo de análise sintática, assim, a intervenção humana, segundo o autor, será mais precisa na correção de um erro ou em um orientar ambiente de construção da linguagem, orientar para o que se deseja interpretar, na função de um parser, por exemplo, quando há percepção de novas interpretações ou ações desejadas.

O uso das técnicas para automatizar o processamento de linguagem natural, o aprender da máquina (CAPUANO, 2009) e o extrair informação (ZWEIGENBAUM et al. 2007) foram passos realizados na metodologia adotada por Diederich et al. (2007). Os autores associam dados obtidos por meio de entrevistas, obedecendo a um conjunto semântico previamente determinado, que foram convertidos em um formato textual digital. A transcrição desse conjunto de dados dá aos pesquisadores as possibilidades para relacionar o conjunto semântico a uma ontologia padrão (processo semiautomático, ou seja, com intervenção humana) ou ontologias de domínio específico (automático, por meio de sistemas de computação ou semiautomático). No caso, Diederich et al. (2007) obtêm uma lista de palavras.

7.5 Tratamento estatístico de textos

Documentos textuais digitais são analisados quantitativamente por meio de recursos estatísticos que contam ou estabelecem um *ranking;* ferramentas; *softwares* e sistemas envolvendo técnicos (planilhas eletrônicas, CAS, desenvolvidos em linguagem de programação, entre outros) e algoritmos complexos com vistas ao auxílio do pesquisador na exploração desses textos, sejam esses literários, científicos, anotações, e-mails ou qualquer outro formato onde seja possível processar caracteres digitais.

Os recursos estatísticos computacionais são normalmente disponibilizados para: contagem de palavras, léxico, vocabulário controlado, lista de palavras-chave, anotação semântica, e dentre outras possibilidades, a base de dados ou local de

onde informação ou dados registrados possam ser extraídos e manipulados por meio de funções matemáticas e/ou estatísticas.

Assim, analisar, interpretar dados textuais, reutilizar e dar novos sentidos é, na verdade, como um pesquisador, técnico, especialista ou qualquer indivíduo poderá interpretar as evidências apuradas nos experimentos envolvendo um conjunto de caracteres textuais aliados com teorias estatísticas e regras computacionais estabelecidas previamente por meio das funções matemáticas. São essas características dos métodos que serão aqui neste tópico classificadas:

Para Silva e Milidiú (1991), funções de crença operadas em conjunto com vocabulários, descritores e textos armazenados em bases de dados com ou sem relacionamentos formam a partir do conteúdo de textos digitais arranjos que podem ser definidos como funções de crenças distintas. Isto é uma metodologia que dá flexibilidade para reorganização de textos, agrupando-os de acordo com o conhecimento registrado e implícito no conteúdo. O reuso a partir de funções de crenças dão novas formas de arranjo e, consequentemente, uma nova abordagem diante da imprecisão de uma nova organização sintática ou do uso da anotação semântica. Como por exemplo, ao definir e empregar em parte do método um *ranking* (KURAMOTO, 2002), na identificação de termos concomitantes (WREN et al., 2004), no *score* de palavras em Baumgartner et al. (2008) ou em Swanson et al. (2006) onde o arranjo ganha sentido na medida em que as relações aumentam. São evidências que devem ser consideradas, dadas inúmeras possibilidades de novas leituras dos conteúdos textuais digitais submetidos ao modelo.

Contudo, Wives (1999) e Zhu et al. (2009) usam métodos baseados na similaridade que organizam automaticamente em grupos objetos textuais digitais. No caso, a estatística disponibilizada por meio de *softwares* não especifica, em alguns casos, a complexidade envolvida no processamento e, portanto, não está claramente exposto o quanto da estatística é aplicada com objetivo de dar um novo sentido aos conjuntos de textos. O uso de recursos estatísticos que está implícito nos sistemas de computação dá evidências sobre a real importância e o seu uso como uma poderosa ferramenta para auxiliar nas soluções e contemplar o trabalho de pesquisa com uma variedade de cálculos extremamente complexos. Isto pode ser observado em alguns trabalhos como: Celec (2004) sobre ANORVA; Wren et al. (2004) sobre

emprego da lógica *fuzzy*; Dieng-Kuntz et al. (2006) sobre *Virtual Staffa* e Janet e Reddy (2010) com o CUBE INDEX. Embora Janet e Reddy (2010) tenham desenvolvido seu trabalho voltado a problemas relacionados para *data mining*, o experimento emprega frequência de termos, frequência inversa e estabelece um termo peso. São essas as medidas que promovem as relações para a construção do cubo indexado, mais os pares de palavras, com sua respectiva identificação e identificação do documento. A complexidade está na forma tridimensional de como o cubo indexado é empregado nas relações de medidas para a recuperação da informação desejada. Desempenha aqui a estatística o seu papel na consolidação e representação das evidências apuradas durante o desenvolvimento metodológico adotado pelos autores.

Por sua vez Maia e Souza (2010) buscam a similaridade por meio de agrupamentos heterogêneos de documentos textuais digitais, cujo critério de similaridade de documentos, dentre outras medidas, baseia-se no peso dos termos e na maior frequência de termos comuns, técnica presente em Janet e Reddy (2010), a qual representa no método, um indicador de similaridade porque quanto mais termos em comum maior a tendência à similaridade.

A tabela 1 foi elaborada com o propósito de demonstrar a interseção de classes identificadas no presente estudo, alguns métodos são comuns e empregados para soluções diferentes. Atendem a problemas específicos para determinada classe, principalmente os estatísticos que são utilizados de forma complexa ou simplesmente para uma frequência de palavras; as anotações que complementam e auxiliam nas relações de termos, como por exemplo, na mineração de texto, contribuindo com um léxico parelelo ao conteúdo digitalizado; e a própria mineração de textos que reutiliza os métodos acima, na maioria das vezes implícitamente por meio dos *softwares* desenvolvidos e sistemas de computação.

TABELA 1 - REPRESENTAÇÃO DOS ARTIGOS POR CLASSES

No.	Autor(es)	Anotação Semântica	Mineração de textos	Tratamento Estatístico	Semântica	Linguagem Natural
1	TEIXEIRA (1974)		●			
2	SILVA e MILIDIÚ (1991)	●		●		
3	BRITO (1992)	●				●
4	PISANELLI et al. (1998)	●	●			
5	WIVES (1999)			●		
6	WEEBER M. (2000)	●		●		●
7	KURAMOTO (2002)				●	
8	OLSSON F. (2002)	●	●	●		
9	TARDELLI et al. (2002)	●	●			
10	OLIVEIRA et al. (2003)	●	●			
11	CELEC (2004)			●		
12	WREN et al. (2004)		●	●		
13	DIENG-KUNTZ et al. (2006)	●				
14	SPASIC et al. (2005)		●	●		
15	ARAÚJO JÚNIOR, TARAPANOFF (2006)		●			●
16	SWANSON et al. (2006)			●		
17	DIEDERICH et al. (2007)	●	●	●		
18	LIMA (2007)	●	●			
19	SIRIN (2007) SPARQL		●			
20	WANG et al. (2007)	●		●		
21	BAUMGARTNER, et al. (2008)	●	●	●		
22	BODENREIDER O. (2008)				●	
23	CAPUANO (2009)		●			
24	KIYAVITSKAYA et al. (2009)	●				
25	LUCCA (2009)		●			
26	RICO et al. (2009)	●				
27	SAMWALD (2009)	●	●			
28	ZHU et al. (2009)		●	●		
29	CHEN et al. (2010)	●		●		
30	DOGAN e LU (2010)					●
31	JANET e REDDY (2010)		●	●		
32	MAIA e SOUZA (2010)	●		●		
33	LENDVAI (2011)	●				

TABELA 2 - PERCENTUAL EM RELAÇÃO AO TOTAL DE ARTIGOS AVALIADOS

Classes	Participações	%
Anotação semântica	**18**	**55**
Mineração de textos	**17**	**51**
Análise Semântica	**2**	**6**
Análise Linguagem Natural	**4**	**12**
Tratamento Estatístico	**15**	**45**

A tabela 2 foi elaborada com intuito de apresentar percentuais das participações de cada classe nos artigos selecionados. O resultado foi obtido por meio do conjunto de artigos que compõem a amostra onde cada classe foi identificada e considerada relevante para contagem e posterior classificação. A tabela 2 sugere, dado o percentual obtido para cada classe, que a combinação de tratamento estatístico, mineração de textos e anotação semântica podem colaborar na obtenção de resultados mais significativos em trabalhos de pesquisa que façam uso de métodos similares aos identificados e agrupados no presente estudo.

8 DISCUSSÃO E CONSIDERAÇÕES FINAIS

Neste capítulo é apresentada a aplicabilidade de cada uma das propostas procurando identificar, respectivamente, seus pontos fortes e fracos. O objetivo é sempre tornar estes dados textuais mais ou menos uniformes, possibilitando se processamento segundo determinada uma intenção, visando ao seu posterior reuso.

Os algoritmos aplicados na mineração de textos digitais demonstram um melhor desempenho quando aliados a técnicas que depuram dados, ou seja, técnicas empregadas na fase de pré-processamento de textos eletrônicos e criação de conjuntos de dados.

A preparação de dados para mineração de textos deverá considerar o problema do espaço disponível para o corpo textual que será processado, o grau de complexidade exigido para o processamento em linguagem natural, além dos inúmeros recursos para leitura, extração, classificação e/ou agrupamento, e considerar também o tempo de processamento exigido de acordo com o nível e sentidos desejados do corpo textual.

As anotações semânticas são alternativas para expandir relações a partir de significados que fogem de um domínio de linguagem ou da aplicação de um determinado termo ou verbete, que se relaciona ou não com outro domínio diferente do contexto no qual o termo ou verbete está inserido. As anotações podem atribuir novos significados, inserir comentários, descrever sucintamente ou não dados complementares relacionados ao contexto em que a palavra ou verbete se aprensente.

As Consultas Semânticas tornam-se linguagens que permitem, no campo da Ciência da Informação, abrir janelas de conhecimento em bases de dados heterogêneas. São sequências de instruções previamente armazenadas ou descritas que podem ser acionadas por eventos de usuários cujas variáveis ou parâmetros promovem alterações significativas na formação de conjuntos de dados resultantes.

Embora não seja escopo do presente trabalho, o desenvolvimento de linguagens para dar sentido, buscar e reconhecer sentidos em um corpo textual contido em ou recuperado de recursos informacionais heterogênos, sejam estes bases de dados,

sites, *e-mails,* páginas em HTML etc. podem ter o seu conteúdo processado, desde que este seja descrito em um conjunto de caracteres legíveis por um programa de computador e, de certa forma, compatibilizado para leitura humana, considerando-se, também, os suportes em que esses textos digitais são disponiblizados para leitura, com seus inúmeros padrões e estruturas, tipo: MARC21,[18] Dublin core, METADADOS, PDFs, RDFs etc. Se esses permitem uma leitura sequencial do seu conteúdo textual, toda e qualquer ferramenta se torna livre para processar e produzir sentido a uma massa textual que se deseja processar.

Uma simples consulta feita por uma linguagem que utiliza comandos simples, que produza uma ordem simples, e especificações denotando variações extremamente complexas, oriundas de palavras que podem ser definidas de diferentes maneiras, em inúmeros artigos das mais diversas áreas do conhecimento, estes poderiam ser lidos por meio de uma simples instrução “LEIA SOBRE protozoário”. Se a *ordem* é dada a um motor de busca na Internet, o motor varreria *sites* sem nenhum propósito a não ser o de encontrar uma palavra denominada protozoário; no entanto, especificar critérios para essa busca, “LEIA SOBRE O protozoário Trypanossoma EM BIOLOGIA” implicaria diretamente na sofisticação de um sistema ou processo incluindo inúmeras teorias, práticas, experimentos, algoritmos, fórmulas matemáticas, funções estatísticas para que essa instrução, na qual foi adicionada um ou vários critérios, procure estabelecer a relevância cronológica ou as inúmeras variáveis abertas ao mundo da Biologia.

O que se pensa aqui não é sofisticar uma linguagem de busca com objetivo de atender a outras demandas. O que se deseja é simplesmente dizer que o sentido está no texto por meio do qual o autor expressa o que deseja e como esse texto pode ser reusado. Se entendermos que a forma que o autor usa para dar visibilidade ao seu conteúdo tem pouca relevância, e o seu instrumento de comunicação é um texto científico, a complexidade para dar um novo sentido a um corpo textual passa para *quem* ou *o que* é utilizado para processar o material disponível. O nível de complexidade no processamento muda, passa à construção de ferramentas ou mecanismos que façam uso do processo criativo do homem e de todas as teorias

[18] Formato Bibliográfico

que podem e estão disponíveis, porém dispostas e fragmentadas nos inúmeros domínios da ciência.

Por maior que seja um arranjo de documentos textuais digitais, um conjunto que represente um resultado de um processo de seleção, recuperação ou de um método que dê um novo sentido a esse conjunto, raramente esse conjunto representará toda a quantidade de documentos disponíveis. Esse arranjo representará uma amostra, um subconjunto de documentos extraído de um total de documentos disponíveis.

Na abordagem de *cluster*, os métodos estatítiscos são largamente empregados para agrupar documentos textuais digitais. Buscam grupos homogêneos e consideram que as palavras ou termos sejam indicadores de similaridade a partir das técnicas estatísticas para *ranking*, peso e frequência de palavras. Os marcadores e suas anotações semânticas viabilizados por modelos do W3C, os formatos e estruturas XML e RDF, os parsers, os compiladores e interpretadores para linguagem natural, os sistemas computacionais que incluem até a inteligência artificial são os meios encontrados para viabilizar estdos quantitativos que possam inferir junto à estatístisca meios mais precisos destinados à busca e recuperação da informação.

No entanto, particularidades reservadas às linguagens de consulta, como SQL para bases de dados ou a sua estrutura adaptada para identificar e extrair agrupamentos de documentos textuais digitais, se mostrou muito eficaz nos estudos de Samwald (2009), em que uma cadeia de parâmetros, criteriosamente definidos, dá sentido a um *cluster* de documentos.

Os esforços do W3C em desenvolver um ambiente que permita automatizar processos e tarefas, além de viabilizar os meios para que seres humanos possam interagir com recursos computacionais, *parsers,* por exemplo, são ampliados e cada vez mais utilizados. Os arquivos XML e RDF são uma parte do trabalho tecnológico que evoluiu para suprir e dar continuidade ao processo de reprensentar, organizar e dar tratamento semântico. São demandas que não foram supridas e oriundas das limitações e das questões que não foram respondidas com o uso das ontologias de domínio.

Não fez parte dessa dissertação o estudo de *softwares* ou de sistemas de computação desenvolvidos ou empregados para agilizar o processamento e/ou interpretar documentos textuais digitais. No entanto, diante da impossibilidade de processar uma grande massa de dados ou simplesmente separar e relacionar termos, palavras ou frases, os *softwares* desempenham um papel importante na realização desse trabalho porque diante da impossibilidade do homem em trabalhar em tempo hábil uma enorme quantidade de dados, eles dão velocidade no processamento e nas análises de resultados.

Não é possível afirmar com precisão quais são realmente os pontos fortes e fracos de cada uma dos métodos identificados e distribuídos nas classes. Cada um tem sua aplicabilidade diante da necessidade do pesquisador em resolver determinado problema. Uma fórmula matemática poderá contribuir muito bem para um termo que venha a ser usado como parâmetro numa consulta em SQL. O mesmo termo pode ser encontrado usando técnicas de mineração de texto ou simplesmente numa avaliação feita por um especialista na área de domínio desejada. E isto poderá levar a um resultado negativo, se a massa documental selecionada não possuir critérios rigorosos para a escolha das fontes de informação.

A análise efetuada, cujos resultados estão na Tabela 1, mostra que a maioria das soluções encontradas são uma combinação dos métodos identificados.

Há uma relação entre o que é produzido e registrado por meio dos artigos científicos e as respostas aos problemas que cientistas e pesquisadores buscam. O reuso desses documentos tem buscado outras relações, as implícitas, que não estão evidentes numa leitura feita por seres humanos e dos quais podem emergir soluções ou definir novos caminhos ou ainda dar indícios para a resolução de outros problemas.

A produção crescente de documentos textuais disponibilizados já diretamente na *Web*, de diversas naturezas e em diferentes domínios, em especial no domínio das ciências biomédicas, coloca a urgência da identificação e sistematização dos diferentes métodos e técnicas que viabilizem seu reuso em larga escala.

9. REFERÊNCIAS

ADRIAANS, P.; ZANTINGE, D. **Data mining**. Harlow, England. Addison Wesley, 1996.

ALVARENGA, L. Representação do conhecimento na perspectiva da Ciência da informação em tempo e espaços digitais. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, n. 15, 1. sem., 2003. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/viewFile/97/5233>.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 6023: informação e documentação - referências - elaboração. Rio de Janeiro: ABNT, 2002a.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 10520: Informação e documentação - citações em documento - apresentação. Rio de Janeiro: ABNT, 2002b

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14724: Informação e documentação - trabalhos acadêmicos - apresentação. Rio de Janeiro: ABNT, 2002c.

ATTWOOD, T. K.; KELL, D. B.; MCDERMOTT, P; MARSH, J.; PETTIFER S. R.; THORNE, D. Calling international rescue: Knowledge lost in literature and data landslide! **Biochem. J**., v. 424, p. 317–333, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805925/pdf/bj4240317.pdf>. Acesso em: 18 jan. 2011.

AZEVEDO, I. B. **O prazer da produção científica**: descubra como é fácil e agradável elaborar trabalhos acadêmicos. 10. ed. São Paulo: Hagnos, 2001.

BATH, P. A. Data mining in health and medical information. **Annual review of information science and technology**, v. 38, n. 1, p. 331-369, 2004.

BENOIT, G. Data mining. **Annual review of information science and technology**, v. 36, p. 265-310, 2002.

BERNERS-LEE, T; HENDLER, J.; LASSILA, O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American Magazine**. p. 34-43, May, 2001.

BODÊ, E. Formatos de arquivo e a preservação de documentos digitais. In: CONGRESSO NACIONAL DE ARQUIVOLOGIA, 2., 2006, Porto Alegre. **Artigo.** Disponível em: <http://www.cipedya.com/doc/101656>.

BRASIL, Ministério do Planejamento, Orçamento e Gestão. 2010. **O que é interoperabilidade**. Disponível em: <http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade/o-que-e-interoperabilidade>. Acesso em: 13 jul. 2010.

BRASIL, Ministério do Planejamento, Orçamento e Gestão. 2010. **Princípios e Diretrizes**. Disponível em: <https://www.governoeletronico.gov.br/o-gov.br/principios>. Acesso em: 13 jul. 2010.

BREITMAN, K. **Web Semântica**: a Internet do futuro. Rio de Janeiro: LTC, 2005.

BUCKLAND, M. What is a "document"? **Journal of the American society for information science,** v. 48, n. 9, p. 804-809, Sept. 1997.

CAMPOS, M. L. A. Modelização de domínios de conhecimento: uma investigação dos princípios fundamentais. **Ciência da informação**, Brasília, v. 33, n. 1, jan./abr., p. 22-32, 2004. Disponível em: <http://www.scielo.br/pdf/ci/v33n1/v33n1a03.pdf>.

CARDOSO, O. N. P. Recuperação de informação. Infocomp**, Revista de computação da UFLA** - Universidade Federal de Lavras, v. 2, n. 1, p.33-38, 2000. Disponível em: < http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf >. Acesso em: 20 jun. 2010.

CHOMSKY, N. **The logical structure of linguistic theory**. 1955-56. Tese (Doutorado) − Cambridge: Massachusetts Institute of Technology - MIT, 1955-56. Disponível em: <http://alpha-leonis.lids.mit.edu/chomsky/chomsky_thesis.pdf>. Acesso em: 13 jul. 2010.

CHOMSKY, N. **Aspectos da Teoria da Sintaxe**. Coimbra: A. Amado, 1965.

CHUN, H. et al. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, 13., October 2005. p. 4-15. Disponível em: <http://helix-web.stanford.edu/psb06/chun.pdf>. Acesso em: 31 de jan. 2011.

COHEN, K. B.; HUNTER, L. Getting started in text mining. **PLoS computational biology,** v. 4, n. 1, p. 1-3, 2008.

COIERA, E. **Guide to medical informatics, the Internet and telemedicine**. London: Chapman & Hall, 1997.

CORNEY, D. P.; BUXTON, B. F.; LANGDON, W. B.; JONES, D. T. A. BioRAT: Extracting biological information from full-length papers. **Bioinformatics**, v. 20, n. 17, p. 3206-3213, 2004.

DOUGHERTY, R. C. **Natural language computing:** an English generative grammar in Prolog. Hillsdale: Lawrence Erlbaum, 1994.

Dowman, M.; Tablan, V.; Cunningham, H.; Popov, B.;Web-assisted annotation, semantic indexing and search of television and radio news, in: Proceedings of the 14th International World Wide Web Conference (WWW2005), May 10–14, Chiba, Japan, 2005, pp. 225–234.

FEITOSA, A. **Organização da informação na web**: das tags à web semântica. Brasília: Thesaurus, 2006.

FIPS, Federal Information Processing Standards Publication. Code for information interchange,1968. Washington, DC (FIPS Pub) 1, National Bureau of Standards.

FLAMINO, A. N.; SANTOS, P. L. V. A. C.; FUJITA, M. S. L. Uma breve reflexão sobre documento, estruturas textuais e a xml nos repositórios institucionais digitais. In: SIMPÓSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, 3., 28 nov. - 2 dez. 2005, São Paulo. Disponível em: < http://bibliotecas-cruesp.usp.br/3sibd/docs/flamino194.pdf>. Acesso em: 10 mai. 2010.

FONSECA, F.; EGENHOFER, M.; BORGES, K. A. V. Ontologias e Interoperabilidade Semântica entre SIGs. **II Workshop Brasileiro em Geoinformática** - GeoInfo2000, São Paulo. Disponível em: <http://www.spatial.maine.edu/~max/GeoInfo2000.pdf>. Acesso em: 13 jul. 2010.

Friedland, N.S.; Allen, P.G. ; Matthews, G.; Witbrock, M.; Baxter, D.;Curtis, J.; Shepard, B.; Miraglia, P. ; Angele, J. ; Staab, S. ; Moench, E.;Oppermann, H.; Wenke, D. ; Israel, D. ; Chaudhri, V.; Porter, B. ; Barker, K.; Fan, J. ; Chaw, S.Y. ; Yeh, P. ; Tecuci, D.; Project halo: towards a digital Aristotle, AI Magazine, Winter 2004, 2004.

GENE ONTOLOGY. The Gene Ontology 1999. Disponível em: <http://www.geneontology.org/>. Acesso em: 11 dez. 2010.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento de linguagem natural. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais da III Jornada de Mini-Cursos de Inteligência Artificial**. Campinas: v. 3, p. 347-395, 2003.

GRIFFIN, E. **Foundations of popfly**: Rapid mashup development. Berkely, CA, USA: Apress, 2008.

GROSSMAN, R.; KARMATH, C.; KUMAR, V. **Data mining for scientific and engineering applications**, Tutorial at SC2001, November 12, 2001. Disponível em: <http://www-users.cs.umn.edu/~kumar/Presentation/M7-dm-chap5.pdf>. Acesso em: 31 jan. 2011.

HEARST, M. A. Untangling text data mining. In: **Annual meeting of the association for computational linguistics,** University of Maryland, 1999.

HOUAISS, A.; VILLAR, M. S.; FRANCO, M. M. **Dicionário Houaiss da Língua Portuguesa.** Rio de Janeiro: Objetiva, 2004.

IBM. Appendix B. ASCII character (2006a) Disponível em: <http://publib.boulder.ibm.com/infocenter/comphelp/v8v101/index.jsp?topic=/com.ibm.xlcpp8a.doc/compiler/ref/ruascii.htm>. Acesso em: 25 mai. 2010.

IBM. Glossary (2006b) Disponível em: <http://publib.boulder.ibm.com/infocenter/comphelp/v8v101/index.jsp?topic=/com.ibm.xlf101a.doc/xlfcr/glosslrm.htm>. Acesso em: 25 mai. 2010.

ICBO. International Conference on Biomedical Ontology. University at Buffalo, NY. July 2011. Disponível em: <http://icbo.buffalo.edu/>. Acesso em: 25 abr. 2011.

JENNEX, M.; OLFMAN, L.; PANTHAWI, P.; PARK, Y. An organizational memory information systems success model: An extension of DeLone and McLean's I/S success model. Proceedings of the Hawaii International Conference on Systems Sciences. 1998.

JINHA, A. E Article 50 million: an estimate of the number of scholarly articles in existence. **Learned publishing**, v. 23, n. 3, p. 258-263, July 2010.

KOCHHAR, R. **Basement computing**. 2008. University of Wisconsin. Disponível em: <http://www.neurophys.wisc.edu/comp/docs/ascii/>. Acesso em: 26 jan. 2011.

KRALLINGER, M.; VALENCIA, A. Text-mining and information-retrieval services for molecular biology. **Genome biology**, London, v. 6, p. 224, 2005.

KUO, W. J.; CHANG, R. F.; CHEN, D. R.; LEE, C. C. Data mining with decision trees for diagnosis of breast tumour in medical ultrasonic images. **Breast cancer research and treatment**, v. 66, p. 51-57, 2001.

KURAMOTO, H. Informação científica: proposta de um novo modelo para o Brasil. **Ciência da informação**, v. 35, n. 2, 2006. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/view/831/678>. Acesso em: 25 abr. 2011.

LE COADIC, Y-F. **A ciência da informação**. Brasília: Briquet de Lemos/Livros, 1996.

LÉVY, P. **As tecnologias da inteligência:** o futuro do pensamento na era da informática. Tradução: Carlos Irineu da Costa. Rio de Janeiro: Ed. 34, 1993.

LÉVY, P. **O que é o virtual?** Barcelona: Editorail Lúmen, 1977. Disponível em: <http://www.4shared.com/document/q8c3oeDd/Pierre_Levy_-_O_que__o_Virtual.html>. Acesso em: 15 mar. 2010.

LHD-11. Workshop on discovering meaning on the go in large heterogeneous data 2011. Held at the twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11). July 16, 2011, Barcelona, Spain. Disponível em: <http://dream.inf.ed.ac.uk/events/lhd-11/>. Acesso em: 2 fev. 2011.

MARCONDES, C. H. Metadados: descrição e recuperação de informações na web. In: MARCONDES, C. H.; KURAMOTO, H.; TOUTAIN, L. B.; SAYÃO; L. (Org.). **Bibliotecas digitais**: saberes e práticas. 2. ed. Salvador: EDUFBA; Brasília: IBICT, 2005.

MARCONDES, C. H.; SAYAO, L. F. Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ciência da informação**, Brasília, v. 31, n. 3, p. 42-54, 2002. Disponível em: < http://www.scielo.br/pdf/ci/v31n3/a05v31n3.pdf>.

MARCONDES, C. H.; SAYAO, L. F. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da informação**, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001. Disponível em: < http://www.scielo.br/pdf/ci/v30n3/7283.pdf>.

MARCONDES, C. H.; CAMPOS, M. L. A. Ontologia e web semântica: o espaço da pesquisa em ciência da informação. **PontodeAcesso,** Salvador, v. 2, n. 1, p. 107-136, jun./jul. 2008. Disponível em: <http://www.portalseer.ufba.br/index.php/revistaici/article/view/2669/1885>. Acesso em: 22 jul. 2010.

MARKUS, L. M. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. **Journal of management information systems**, v. 18, n. 1, p. 57-93, 2001.

OBO. The open biological and biomedical ontologies. Disponível em: <http://www.obofoundry.org/>. Acesso em: 21 jan. 2011.

O'HARA, K.; Hall, W. Semantic Web. In: Bates, M. J; Maack, M. N.; Drake, M. (Ed.). **Encyclopedia of library and information science**. 2. ed. Taylor & Francis. Disponível em: <http://eprints.ecs.soton.ac.uk/17126/>.

OLIVEIRA, R. M. V. B. Web Semântica: novo desafio para os profissionais da informação. In: **Seminário Nacional de Bibliotecas Universitárias**. 2002. Disponível em: <http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/124.a.pdf>. Acesso em: 1 mai. 2010.

PEÑA-REYES, C. A.; SIPPER, M. Evolutionary computation in medicine: An overview. **Artificial intelligence in medicine**, v. 19, n. 1, p. 1-23, May/2000.

PICKLER, M. E. V. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em ciência da informação**, Belo Horizonte, v. 12, n. 1, p. 65-83, jan./abr. 2007 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362007000100006&lng=en&nrm=iso>. Acesso em: 3 abr. 2010.

PINTO MOLINA**,** M. **El resumen documental**. Principios y métodos. Madrid: Fundación Germán Sánchez Ruipérez, 1992.

PLATÃO, F.; FIORIN, J. L. **Para entender o texto**: leitura e redação. São Paulo: Ática, 2006.

RDF, Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 2004. Disponível em: <http://www.w3.org/RDF/>. Acesso em: 22 abr. 2011.

RENEAR, A.; PALMER, C. Strategic reading, ontologies, and the future of scientific publishing**. Science**, v. 325, n. 5942, p. 828-832. 2009**.** Disponível em: <http://www.sciencemag.org/cgi/content/abstract/325/5942/828>.

ROJAS, T.; PÉREZ, M.; RIVAS, L.; Proposed knowledge reuse model for application in Venezuela. **Americas conference on information systems**, 2003. Disponível em: <http://www.lisi.usb.ve/publicaciones/04%20gestion%20del%20conocimiento/gestion_07.pdf>. Acesso em: 1 jan. 2011.

SANTAELLA, L; VIEIRA, J. A. **Metaciência**: como guia da pesquisa. São Paulo: Ed. Mérito, 2008.

SARACEVIC, T. Ciência da Informação: origens, evolução e relações. **Perspectivas em ciência da informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/235/22>.

SAUSSURE, Ferdinand. **Curso de lingüística general**. Ediciones Alkal, S.A. Madrid, España, 1989.

SHATKAY, H. Hairpins in bookstacks: Information retrieval from biomedical text. **Briefings in bioinformatics**, v. 6, n. 3, p. 222–238. September 2005. Disponível em: <http://research.cs.queensu.ca/~shatkay/papers/BrIB2005.pdf>.

SHETH, A. Semantics and services enabled problem solving environment for Tcruzi. 2008. Disponível em: <http://www.bioontology.org/PSE-talke/>. Acesso em: 25 abr. 2011.

SHETH, A. Semantics and services enabled problem solving environment for Tcruzi. 2008. Disponível em: <http://knoesis.wright.edu/research/semsci/application_domain/sem_life_sci/tcruzi_pse/ >. Acesso em: 10 jan. 2011.

SHETH, A; ARPINAR, I. B.; KHASYAP, V. Relationships at the heart of semantic web: Modeling, discovering and exploiting complex semantic relationships. In: **Technical report**, LSDIS Lab, Computer Science University of Georgia, Athens, GA, 2002.

SHORTLIFFE, E. H.; BARNETT, G. O. Medical data: Their acquisition, storage and use. In: SHORTLIFFE, E. H.; PERREAULT, L. E.; WIEDERHOLD, G.; FAGAN, L. M (Ed.). **Medical informatics computer applications in health care and biomedicine.** 2nd ed. New York: Springer, 2001.

SHORTLIFFE, E. H.; BLOIS, M. S. The computer meets biology and medicine: Emergence of a discipline. In: SHORTLIFFE, E. H.; PERREAULT, L. E.; WIEDERHOLD, G.; FAGAN, L. M. (Ed.). **Medical informatics computer applications in health care and biomedicine.** 2nd ed. New York: Springer, 2001.

SIAM. Society for industrial and applied Mathematics. 2008 SIAM CONFERENCE ON DATA MINING. Disponível em: < http://www.siam.org/meetings/sdm08/>. Acesso em: 31 jan. 2011.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em ciência da informação**, Belo Horizonte, v. 11, n. 2, p. 161-173, mai./ago. 2006. Disponível em:< http://www.scielo.br/pdf/pci/v11n2/v11n2a02.pdf>. Acesso em: 3 abr. 2010.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.

SPASIC, I; ANANIADOU, S; McNAUGHT, J.; KUMAR, A. Text mining and ontologies in biomedicine: making sense of raw text. **Briefings in bioinformatics**, v. 6, n. 3, p. 239-251, 2005. Disponível em: < http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/BIB.pdf>. Acesso em: 2 jan. 2011.

SWANSON, D. R. Fish-oil, Raynaud's syndrome and undiscovered public knowledge. **Perspectives in biology and medicine**, v. 30, n. 1, p. 7-18, 1986.

SWANSON, D. R. Two medical literatures that are logically but not bibliographically connected. **Journal of the American society for information science,** v. 38, n. 4, p. 228-333, jul. 1987.

SWANSON, D. R. Migraine and magnesium: eleven neglected connections. **Perspectives in biology and medicine**, v. 31, n. 4, p. 526–557, 1988.

UNICODE CONSORTIUM. 1991. What is Unicode? Disponível em: <http://unicode.org/http://www.unicode.org/standard/WhatIsUnicode.html>.

UREN, V.; CIMIANO, P.; IRIA, J.; HANDSCHUH, S.; VARGAS-VERA, M.; MOTTA, E. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. **Journal of websemantics,** 2005. Disponivel em: <http://siegfried-handschuh.net/pub/2006/semantic-annotation_websemantics2006.pdf>.

W3C (1999). HTML 4.01 Specification. W3C Recommendation, World Wide Web Consortium. Disponível em: <www.w3.org/TR/html401>. Acesso em: 15 dez. 2009.

W3C (2000a). Document Object Model (DOM) Level 1 Specification (Second Edition). W3C Working Draft, World Wide Web Consortium. Disponível em: <http://www.w3.org/TR/REC-DOM-Level-1>. Acesso em: 15 dez. 2009.`

W3C (2001). Web Architecture Specification. W3C Recommendation, World Wide Web Consortium. Disponível em: <www.w3.org/Tr/2001/>. Acesso em: 24 jul. 2010.

W3C (2004). OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 24 jul. 2010.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. 2004. 136 f.

Tese (Doutorado) - Universidade Federal do Rio Grande do Sul, Porto Alegre: Programa de Pós-Graduação em Computação. Disponível em: <http://www.lume.ufrgs.br/handle/10183/4576 >. Acesso em: 16 jan. 2009.

ZHOU, G.; Zhang, J.; Su, J.; Shen, D.; Tan, C. et al. Recognizing names in biomedical texts: a machine learning approach. **Bioinformatics**, v. 20, n. 7, p. 1178-1190, 2004.

ZWEIGENBAUM, P.; DEMNER-FUSHMAN, D.; HONG YU; COHEN, K. B. Frontiers of biomedical text mining: current progress. **Briefings in bioinformatics,** v. 8, n. 5, p. 358-375, 2007. Disponível em: <http://bib.oxfordjournals.org/content/8/5/358.full.pdf+html>.

RELAÇÃO BIBLIOGRÁFICA ANALISADA

ARAÚJO, J. R. H.; TARAPANOFF, K. 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, set./dez. 2006. Disponível em: <http://www.scielo.br/pdf/ci/v35n3/v35n3a23.pdf>. Acesso em: 7 mar. 2011.

BAUMGARTNER JR, W. A.; COHEN, K. B.; HUNTER, L. An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. **Journal of biomedical discovery and collaboration,** 2008. Disponível em: <http://www.biomedcentral.com/content/pdf/1747-5333-3-1.pdf>. Acesso em: 12 fev. 2011.

BODENREIDER, O. Biomedical ontologies in action: Role in knowledge management, data integration and decision support**. International medical informatics association (IMIA)**, p. 67-79, 2008.

BRITO, M. Sistemas de informação em linguagem natural: em busca de uma indexação automática. **Ciência da informação**, Brasília, v. 21, n. 3, p. 223-232, set./dez. 1992. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=9055>. Acesso em: 8 mar. 2011.

CAPUANO, E. A. O poder cognitivo das redes neurais artificiais modelo ART1 na recuperação da informação. **Ciência da informação**, Brasília, v. 38, n. 1, p. 9-30, jan./abr. 2009. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=6751>. Acesso em: 7 mar. 2011.

CELEC, P. Analysis of rhythmic variance - ANORVA. A new simple method for detecting rhythms in biological time series. **Biol. Res.**, Santiago, v. 37, n. 4, 2004. Disponível em: <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0716-97602004000500007&lng=es&nrm=iso>. Acesso em: 1 mar. 2011.

CHEN, D.; LI, X.; LIANG, Y.; ZHANG, J. A semantic query approach to personalized e-Catalogs service system. **Journal of theoretical and applied electronic commerce research**, Talca, v. 5, n. 3, dez. 2010. Disponível em: <http://www.scielo.cl/pdf/jtaer/v5n3/art05.pdf >. Acesso em: 1 mar. 2011.

DOGAN, R. I.; LU, Z. Click-words: Learning to predict document keywords from a user perspective. **Bioinformatics,** v. 26, n. 21, p. 2767-2775, 2010. Disponível em: <http://bioinformatics.oxfordjournals.org/content/26/21/2767.full.pdf+html>. Acesso em: 13 mar. 2011.

Diederich, J.; Al-Ajmi, A.; Yellowlees, P. Ex-ray: Data mining and mental health, Applied soft computing, v. 7 n. 3, p. 923-928, June, 2007 [doi>10.1016/j.asoc.2006.04.007]

DIENG-KUNTZ R.; MINIER D.; RUZICKA, M.; CORBY, F.; CORBY, O.; ALAMARGUY, L. Building and using a medical ontology for knowledge management

and cooperative work in a health care network. **Computers in biology and medicine,** v. 36*:* p. 871–892, 2006*.*

Hunter, J.; Schroeter, R.; Koopman, B.; Henderson, M. Using the semantic grid to build bridges between museums and indigenous communities, in: Proceedings of the GGF11—Semantic Grid Applications Workshop, Honolulu, June 10, 2004, 2004.

JANET, B.; REDDY, A. V. Cube Index: A text index model for retrieval and mining. **International journal of computer applications**, v. 1, n. 9, 2010. Disponível em: <http://www.ijcaonline.org/journal/number9/pxc387330.pdf>. Acesso em: 15 mar. 2011.

KIYAVITSKAYA, N.; ZENI, N.; CORDY, J. R.; MICH, L.; MYLOPOULOS, J.Cerno: Light-weight tool support for semantic annotation of textual documents. **Data & knowledge engineering**, v. 68, n. 12, p. 1470-1492, dez./2009. Disponível em: <http://research.cs.queensu.ca/~cordy/Papers/KZCMM_DKE_Cerno.pdf>. Acesso em: 4 jan. 2011.

KURAMOTO, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 1, fev. 2002. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=7479>. Acesso em: 8 mar. 2011.

LENDVAI, P. Conceptual taxonomy identification in medical documents. In: **Proceedings of the second international workshop on knowledge discovery and ontologies**, p. 31-38, 2005.

LIMA, G. A. B. O. Modelo hipertextual -MHTX: um modelo para organização hipertextual de documentos. **DataGramaZero - Revista de ciência da informação** Rio de Janeiro, v. 8, n. 4, ago. 2007. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=7550>. Acesso em: 8 mar. 2011.

LUCCA, J. L. De. Detecção e extração de candidatos a acepções baseadas em um tesaurus de colocados. **Informação & Informação**, Londrina, v. 14, n. esp., p. 125-144, 2009. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=10442>. Acesso em: 8 mar. 2011.

MAIA, L. C. G.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 154-172, 2010. Disponível em: < http://www.scielo.br/pdf/pci/v15n1/09.pdf >. Acesso em: 2 mar. 2011.

MAIA, M. A. R. Gramática e Parser. In: II Congresso Internacional da Abralin, 2001, Fortaleza, CE. **Anais do II Congresso Internacional da Abralin**, Boletim 26. Fortaleza. Imprensa Universitária UFC, 2001. v. I. p. 188-192.

OLIVEIRA, C.; GARRÃO, M.; AMARAL, L. Recognizing complex prepositions Prep+N+Prep as negative patterns in automatic term extraction from texts. In: **Proceedings of 1st workshop em tecnologia da informação e da linguagem humana (TIL2003).** São Carlos - SP. 2003. Disponível em: <http://www.nilc.icmc.usp.br/til2003/oral/oliveira_garrao_amaral25.pdf>.

OLSSON, F.; Eriksson, G.; Franzén, K.; Asker, L.; Lidén, P. **Notions of correctness when evaluating protein name taggers.** Proceedings of the 19th international conference on computational linguistics (COLING 2002) 2002:765-771. Disponível em: <http://www.aclweb.org/anthology/C/C02/C02-1110.pdf>. Acesso em: 31 mar. 2011.

Pisanelli, D. M.; Gangemi, A.; Steve, G. An ontological analysis of the UMLS metathesaurus. **Journal of American medical informatics association**, v. 5 (symposium supplement), 1998.

Plessers, P.; Casteleyn, S.; Yesilada, Y.; De Troyer, O.; Stevens, R.; Harper, S.; Goble, C.. Accessibility: a web engineering approach, in: Proceedings of the 14th International World Wide Web Conference (WWW2005), May 10–14, Chiba, Japan, 2005, pp. 353–362.

RICO, M.; TAVERNA, M. L.; CALIUSCO, M. L.; CHIOTTI, O.; GALLI, M. R. Adding semantics to electronic business documents exchanged in collaborative commerce relations. **Journal of theoretical and applied electronic commerce research**, v. 4, n. 1, abr./2009. Disponível em: < http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-18762009000100007&lng=es&nrm=iso >. Acesso em: 2 mar. 2011.

Rinaldi, F.; Schneider, G.; Kaljurand, K. ; Dowdall, J. ; Persidis, A.; Konstanti, O.. Mining relations in the GENIA corpus, Second European Workshop on Data Mining and Text Mining for Bioin-formatics, 24 September 2004, Pisa, Italy, 2004.

SAMWALD, M. Extracting conclusion sections from PubMed abstracts for rapid key assertion integration in biomedical research. **Nature precedings**, Set./2009. Disponível em: <http://precedings.nature.com/documents/3775/version/1>. Acesso em: 10 fev. 2011.

SILVA, W. T.; MILIDIÚ, R. L. Indexação e recuperação da informação com função de crença. **Ciência da informação**, Brasília, v. 20, n. 2, p. 155-164, jul./dez. 1991. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=9103>. Acesso em: 1 mar. 2011.

SIRIN, E.; PARSIA, B. SPARQL-DL: SPARQL Query for OWL-DL. **Third international workshop**, Innsbruck, 2007. Disponível em: <http://pellet.owldl.com/papers/sirin07sparqldl.pdf>. Acesso em: 1 mar. 2011.

SPASIC, I; ANANIADOU, S; McNAUGHT, J.; KUMAR, A. Text mining and ontologies in biomedicine: making sense of raw text. **Briefings in bioinformatics,** v. 6, n. 3, p. 239-251, 2005. Disponível em: < http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/BIB.pdf>. Acesso em: 2 jan. 2011.

Svab, O.; Labsky, M.; Svatek, V. RDF-based retrieval of information extracted from web product catalogues, in: Proceedings of the SIGIR'04 Semantic Web Workshop, Sheffield, 2004.

SWANSON, D. R.; SMALHEISER, N. R.; TORVIK, V. I. Ranking indirect connections in literature based discovery. The role of Medical Subject Headings. **Journal of the American society for information science and technology**, v. 57, n. 11, p. 1427–1439, 2006.

TARDELLI, A. O.; ANÇÃO, M. S.; PACKER, A. L.; SIGULEM, D. Descoberta baseada em literatura: um enfoque experimental para descoberta aberta em bases de dados do tipo MEDLINE. In: CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE – CBIS, 8., 2002. Disponível em: <http://ambienteaprendiz.bvs.br/pdf/aot_medline.pdf>. Acesso em: 2 jan. 2011.

TEIXEIRA, I. L. R. Uma linguagem de busca para sistemas de recuperação de informação. **Ciência da informação**, Brasília, v. 3, n. 1, p. 21-50, 1974. Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=9917>. Acesso em: 8 mar. 2011.

WANG, J. Z.; DU, Z.; PAYATTAKOO, R.; YU, P. S.; CHEN, C. F. A new method to measure the semantic similarity of GO terms. **Bioinformatics**, v. 23, n. 10, p. 1274–1281, 2007. Disponível em: <http://bioinformatics.oxfordjournals.org/content/23/10/1274.full.pdf+html>. Acesso em: 30 mar. 2011.

WEEBER, M.; KLEIN, H.; ARONSON, A. R.; MORK, J. G.; JONG, L. T. W.; VOS, R. Text-based discovery in biomedicine: The architecture of the DAD-system. **Proceedings of the American medical informatics association symposium**, Los Angeles, p. 903-907, 2000. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243779/pdf/procamiasymp00003-0938.pdf. Acesso em: 12 mar. 2011.

WIVES, L. K. **Estudo sobre agrupamento de documentos textuais em processamento de informação não estruturadas usando técnicas de clustering**. 1999. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999. Disponível em: <http://www.leandro.wives.nom.br/pt-br/publicacoes/dissertacao.pdf>. Acesso em: 7 mar. 2011.

WREN, J. D.; BEKEREDJIAN, R.; STEWART, J. A.; SHOHET, R. V.; GARNER, H. R. Knowledge discovery by automated identification and ranking of implicit relationships. **Bioinformatics,** v. 20, n. 3, p. 389–398, 2004.

ZHU, S.; ZENG, J.; MAMITSUKA, H. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. **Bioinformatics,** v. 25, n. 15, p. 1944-1951, 2009. Disponível em: <http://bioinformatics.oxfordjournals.org/content/early/2009/06/03/bioinformatics.btp338>. Acesso em: 10 fev. 2011.

ANEXOS

TABELA 3 - ASCII CARACTERES DE CONTROLE (CÓDIGOS DE CARACTERES 0-31)

DEC	OCT	HEX	BIN	Symbol	HTML Number	HTML Name	Description
0	000	00	00000000	NUL	&#000;		Null char
1	001	01	00000001	SOH	&#001;		Start of Heading
2	002	02	00000010	STX	&#002;		Start of Text
3	003	03	00000011	ETX	&#003;		End of Text
4	004	04	00000100	EOT	&#004;		End of Transmission
5	005	05	00000101	ENQ	&#005;		Enquiry
6	006	06	00000110	ACK	&#006;		Acknowledgment
7	007	07	00000111	BEL	&#007;		Bell
8	010	08	00001000	BS	&#008;		Back Space
9	011	09	00001001	HT			Horizontal Tab
10	012	0A	00001010	LF
	Line Feed
11	013	0B	00001011	VT	&#011;		Vertical Tab
12	014	0C	00001100	FF			Form Feed
13	015	0D	00001101	CR
	Carriage Return
14	016	0E	00001110	SO	&#014;		Shift Out / X-On
15	017	0F	00001111	SI	&#015;		Shift In / X-Off
16	020	10	00010000	DLE	&#016;		Data Line Escape
17	021	11	00010001	DC1	&#017;		Device Control 1 (oft. XON)
18	022	12	00010010	DC2	&#018;		Device Control 2
19	023	13	00010011	DC3	&#019;		Device Control 3 (oft. XOFF)
20	024	14	00010100	DC4	&#020;		Device Control 4
21	025	15	00010101	NAK	&#021;		Negative Acknowledgement
22	026	16	00010110	SYN	&#022;		Synchronous Idle
23	027	17	00010111	ETB	&#023;		End of Transmit Block
24	030	18	00011000	CAN	&#024;		Cancel
25	031	19	00011001	EM	&#025;		End of Medium
26	032	1A	00011010	SUB	&#026;		Substitute
27	033	1B	00011011	ESC	&#027;		Escape
28	034	1C	00011100	FS	&#028;		File Separator
29	035	1D	00011101	GS	&#029;		Group Separator
30	036	1E	00011110	RS	&#030;		Record Separator
31	037	1F	00011111	US	&#031;		Unit Separator

Os primeiros 32 caracteres na tabela ASCII são códigos de controle não imprimíveis e são usados para controlar periféricos, como impressoras.
Fonte: <http://www.ascii-code.com/>.

TABELA 4 - CARACTERES EM ASCII (CÓDIGO DE CARACTERE 32-127)

DEC	OCT	HEX	BIN	Symbol	HTML Number	HTML Name	Description
32	040	20	00100000		&#32;		Space
33	041	21	00100001	!	&#33;		Exclamation mark
34	042	22	00100010	"	&#34;	&quot;	Double quotes (or speech marks)
35	043	23	00100011	#	&#35;		Number
36	044	24	00100100	$	&#36;		Dollar
37	045	25	00100101	%	&#37;		Procenttecken
38	046	26	00100110	&	&#38;	&amp;	Ampersand
39	047	27	00100111	'	&#39;		Single quote
40	050	28	00101000	(	&#40;		Open parenthesis (or open bracket)
41	051	29	00101001	)	&#41;		Close parenthesis (or close bracket)
42	052	2A	00101010	*	&#42;		Asterisk
43	053	2B	00101011	+	&#43;		Plus
44	054	2C	00101100	,	&#44;		Comma
45	055	2D	00101101	-	&#45;		Hyphen
46	056	2E	00101110	.	&#46;		Period, dot or full stop
47	057	2F	00101111	/	&#47;		Slash or divide
48	060	30	00110000	0	&#48;		Zero
49	061	31	00110001	1	&#49;		One
50	062	32	00110010	2	&#50;		Two
51	063	33	00110011	3	&#51;		Three
52	064	34	00110100	4	&#52;		Four
53	065	35	00110101	5	&#53;		Five
54	066	36	00110110	6	&#54;		Six
55	067	37	00110111	7	&#55;		Seven
56	070	38	00111000	8	&#56;		Eight
57	071	39	00111001	9	&#57;		Nine
58	072	3A	00111010	:	&#58;		Colon
59	073	3B	00111011	;	&#59;		Semicolon
60	074	3C	00111100	<	&#60;	&lt;	Less than (or open angled bracket)
61	075	3D	00111101	=	&#61;		Equals
62	076	3E	00111110	>	&#62;	&gt;	Greater than (or close angled bracket)
63	077	3F	00111111	?	&#63;		Question mark
64	100	40	01000000	@	&#64;		At symbol
65	101	41	01000001	A	&#65;		Uppercase A
66	102	42	01000010	B	&#66;		Uppercase B

67	103	43	01000011	C	&#67;		Uppercase C
68	104	44	01000100	D	&#68;		Uppercase D
69	105	45	01000101	E	&#69;		Uppercase E
70	106	46	01000110	F	&#70;		Uppercase F
71	107	47	01000111	G	&#71;		Uppercase G
72	110	48	01001000	H	&#72;		Uppercase H
73	111	49	01001001	I	&#73;		Uppercase I
74	112	4A	01001010	J	&#74;		Uppercase J
75	113	4B	01001011	K	&#75;		Uppercase K
76	114	4C	01001100	L	&#76;		Uppercase L
77	115	4D	01001101	M	&#77;		Uppercase M
78	116	4E	01001110	N	&#78;		Uppercase N
79	117	4F	01001111	O	&#79;		Uppercase O
80	120	50	01010000	P	&#80;		Uppercase P
81	121	51	01010001	Q	&#81;		Uppercase Q
82	122	52	01010010	R	&#82;		Uppercase R
83	123	53	01010011	S	&#83;		Uppercase S
84	124	54	01010100	T	&#84;		Uppercase T
85	125	55	01010101	U	&#85;		Uppercase U
86	126	56	01010110	V	&#86;		Uppercase V
87	127	57	01010111	W	&#87;		Uppercase W
88	130	58	01011000	X	&#88;		Uppercase X
89	131	59	01011001	Y	&#89;		Uppercase Y
90	132	5A	01011010	Z	&#90;		Uppercase Z
91	133	5B	01011011	[	&#91;		Opening bracket
92	134	5C	01011100	\	&#92;		Backslash
93	135	5D	01011101	]	&#93;		Closing bracket
94	136	5E	01011110	^	&#94;		Caret - circumflex
95	137	5F	01011111	_	&#95;		Underscore
96	140	60	01100000	`	&#96;		Grave accent
97	141	61	01100001	a	&#97;		Lowercase a
98	142	62	01100010	b	&#98;		Lowercase b
99	143	63	01100011	c	&#99;		Lowercase c
100	144	64	01100100	d	&#100;		Lowercase d
101	145	65	01100101	e	&#101;		Lowercase e
102	146	66	01100110	f	&#102;		Lowercase f
103	147	67	01100111	g	&#103;		Lowercase g
104	150	68	01101000	h	&#104;		Lowercase h
105	151	69	01101001	i	&#105;		Lowercase i
106	152	6A	01101010	j	&#106;		Lowercase j
107	153	6B	01101011	k	&#107;		Lowercase k
108	154	6C	01101100	l	&#108;		Lowercase l
109	155	6D	01101101	m	&#109;		Lowercase m

110	156	6E	01101110	n	&#110;		Lowercase n
111	157	6F	01101111	o	&#111;		Lowercase o
112	160	70	01110000	p	&#112;		Lowercase p
113	161	71	01110001	q	&#113;		Lowercase q
114	162	72	01110010	r	&#114;		Lowercase r
115	163	73	01110011	s	&#115;		Lowercase s
116	164	74	01110100	t	&#116;		Lowercase t
117	165	75	01110101	u	&#117;		Lowercase u
118	166	76	01110110	v	&#118;		Lowercase v
119	167	77	01110111	w	&#119;		Lowercase w
120	170	78	01111000	x	&#120;		Lowercase x
121	171	79	01111001	y	&#121;		Lowercase y
122	172	7A	01111010	z	&#122;		Lowercase z
123	173	7B	01111011	{	&#123;		Opening brace
124	174	7C	01111100	|	&#124;		Vertical bar
125	175	7D	01111101	}	&#125;		Closing brace
126	176	7E	01111110	~	&#126;		Equivalency sign - tilde
127	177	7F	01111111		&#127;		Delete

Códigos de 32 até 127 são comuns para todas as diferentes variações da tabela ASCII. Eles são chamados de caracteres imprimíveis, utilizados para representar letras, algarismos, sinais de pontuação e alguns símbolos diversos. Você vai encontrar quase todos os caracteres em seu teclado. O número decimal 127 representa o comando DEL.
Fonte: http://www.ascii-code.com/

TABELA 5 - OS CÓDIGOS ESTENDIDOS ASCII (CÓDIGO DE CARACTERE 128-255)

DEC	OCT	HEX	BIN	Symbol	HTML Number	HTML Name	Description
128	200	80	10000000	€	&#128;	&euro;	Euro sign
129	201	81	10000001				
130	202	82	10000010	‚	&#130;	&sbquo;	Single low-9 quotation mark
131	203	83	10000011	ƒ	&#131;	&fnof;	Latin small letter f with hook
132	204	84	10000100	„	&#132;	&bdquo;	Double low-9 quotation mark
133	205	85	10000101	…	&#133;	&hellip;	Horizontal ellipsis
134	206	86	10000110	†	&#134;	&dagger;	Dagger
135	207	87	10000111	‡	&#135;	&Dagger;	Double dagger
136	210	88	10001000	ˆ	&#136;	&circ;	Modifier letter circumflex accent
137	211	89	10001001	‰	&#137;	&permil;	Per mille sign
138	212	8A	10001010	Š	&#138;	&Scaron;	Latin capital letter S with caron
139	213	8B	10001011	‹	&#139;	&lsaquo;	Single left-pointing angle quotation
140	214	8C	10001100	Œ	&#140;	&OElig;	Latin capital ligature OE
141	215	8D	10001101				
142	216	8E	10001110	Ž	&#142;		Latin captial letter Z with caron
143	217	8F	10001111				
144	220	90	10010000				
145	221	91	10010001	‘	&#145;	&lsquo;	Left single quotation mark
146	222	92	10010010	’	&#146;	&rsquo;	Right single quotation mark
147	223	93	10010011	“	&#147;	&ldquo;	Left double quotation mark
148	224	94	10010100	”	&#148;	&rdquo;	Right double quotation mark
149	225	95	10010101	•	&#149;	&bull;	Bullet
150	226	96	10010110	–	&#150;	&ndash;	En dash
151	227	97	10010111	—	&#151;	&mdash;	Em dash
152	230	98	10011000	˜	&#152;	&tilde;	Small tilde
153	231	99	10011001	™	&#153;	&trade;	Trade mark sign
154	232	9A	10011010	š	&#154;	&scaron;	Latin small letter S with caron
155	233	9B	10011011	›	&#155;	&rsaquo;	Single right-pointing angle quotation mark
156	234	9C	10011100	œ	&#156;	&oelig;	Latin small ligature oe
157	235	9D	10011101				
158	236	9E	10011110	ž	&#158;		Latin small letter z with caron
159	237	9F	10011111	Ÿ	&#159;	&yuml;	Latin capital letter Y with diaeresis
160	240	A0	10100000				Non-breaking space
161	241	A1	10100001	¡	&#161;	&iexcl;	Inverted exclamation mark
162	242	A2	10100010	¢	&#162;	&cent;	Cent sign
163	243	A3	10100011	£	&#163;	&pound;	Pound sign
164	244	A4	10100100	¤	&#164;	&curren;	Currency sign

165	245	A5	10100101	¥	&#165;	&yen;	Yen sign
166	246	A6	10100110	¦	&#166;	&brvbar;	Pipe, Broken vertical bar
167	247	A7	10100111	§	&#167;	&sect;	Section sign
168	250	A8	10101000	¨	&#168;	&uml;	Spacing diaeresis - umlaut
169	251	A9	10101001	©	&#169;	&copy;	Copyright sign
170	252	AA	10101010	ª	&#170;	&ordf;	Feminine ordinal indicator
171	253	AB	10101011	«	&#171;	&laquo;	Left double angle quotes
172	254	AC	10101100	¬	&#172;	&not;	Not sign
173	255	AD	10101101		&#173;	&shy;	Soft hyphen
174	256	AE	10101110	®	&#174;	&reg;	Registered trade mark sign
175	257	AF	10101111	¯	&#175;	&macr;	Spacing macron - overline
176	260	B0	10110000	°	&#176;	&deg;	Degree sign
177	261	B1	10110001	±	&#177;	&plusmn;	Plus-or-minus sign
178	262	B2	10110010	²	&#178;	&sup2;	Superscript two - squared
179	263	B3	10110011	³	&#179;	&sup3;	Superscript three - cubed
180	264	B4	10110100	´	&#180;	&acute;	Acute accent - spacing acute
181	265	B5	10110101	µ	&#181;	&micro;	Micro sign
182	266	B6	10110110	¶	&#182;	&para;	Pilcrow sign - paragraph sign
183	267	B7	10110111	·	&#183;	&middot;	Middle dot - Georgian comma
184	270	B8	10111000	¸	&#184;	&cedil;	Spacing cedilla
185	271	B9	10111001	¹	&#185;	&sup1;	Superscript one
186	272	BA	10111010	º	&#186;	&ordm;	Masculine ordinal indicator
187	273	BB	10111011	»	&#187;	&raquo;	Right double angle quotes
188	274	BC	10111100	¼	&#188;	&frac14;	Fraction one quarter
189	275	BD	10111101	½	&#189;	&frac12;	Fraction one half
190	276	BE	10111110	¾	&#190;	&frac34;	Fraction three quarters
191	277	BF	10111111	¿	&#191;	&iquest;	Inverted question mark
192	300	C0	11000000	À	&#192;	&Agrave;	Latin capital letter A with grave
193	301	C1	11000001	Á	&#193;	&Aacute;	Latin capital letter A with acute
194	302	C2	11000010	Â	&#194;	&Acirc;	Latin capital letter A with circumflex
195	303	C3	11000011	Ã	&#195;	&Atilde;	Latin capital letter A with tilde
196	304	C4	11000100	Ä	&#196;	&Auml;	Latin capital letter A with diaeresis
197	305	C5	11000101	Å	&#197;	&Aring;	Latin capital letter A with ring above
198	306	C6	11000110	Æ	&#198;	&AElig;	Latin capital letter AE
199	307	C7	11000111	Ç	&#199;	&Ccedil;	Latin capital letter C with cedilla
200	310	C8	11001000	È	&#200;	&Egrave;	Latin capital letter E with grave
201	311	C9	11001001	É	&#201;	&Eacute;	Latin capital letter E with acute
202	312	CA	11001010	Ê	&#202;	&Ecirc;	Latin capital letter E with circumflex
203	313	CB	11001011	Ë	&#203;	&Euml;	Latin capital letter E with diaeresis
204	314	CC	11001100	Ì	&#204;	&Igrave;	Latin capital letter I with grave
205	315	CD	11001101	Í	&#205;	&Iacute;	Latin capital letter I with acute
206	316	CE	11001110	Î	&#206;	&Icirc;	Latin capital letter I with circumflex
207	317	CF	11001111	Ï	&#207;	&Iuml;	Latin capital letter I with diaeresis

208	320	D0	11010000	Ð	Ð	Ð	Latin capital letter ETH
209	321	D1	11010001	Ñ	Ñ	Ñ	Latin capital letter N with tilde
210	322	D2	11010010	Ò	Ò	Ò	Latin capital letter O with grave
211	323	D3	11010011	Ó	Ó	Ó	Latin capital letter O with acute
212	324	D4	11010100	Ô	Ô	Ô	Latin capital letter O with circumflex
213	325	D5	11010101	Õ	Õ	Õ	Latin capital letter O with tilde
214	326	D6	11010110	Ö	Ö	Ö	Latin capital letter O with diaeresis
215	327	D7	11010111	×	×	×	Multiplication sign
216	330	D8	11011000	Ø	Ø	Ø	Latin capital letter O with slash
217	331	D9	11011001	Ù	Ù	Ù	Latin capital letter U with grave
218	332	DA	11011010	Ú	Ú	Ú	Latin capital letter U with acute
219	333	DB	11011011	Û	Û	Û	Latin capital letter U with circumflex
220	334	DC	11011100	Ü	Ü	Ü	Latin capital letter U with diaeresis
221	335	DD	11011101	Ý	Ý	Ý	Latin capital letter Y with acute
222	336	DE	11011110	Þ	Þ	Þ	Latin capital letter THORN
223	337	DF	11011111	ß	ß	ß	Latin small letter sharp s - ess-zed
224	340	E0	11100000	à	à	à	Latin small letter a with grave
225	341	E1	11100001	á	á	á	Latin small letter a with acute
226	342	E2	11100010	â	â	â	Latin small letter a with circumflex
227	343	E3	11100011	ã	ã	ã	Latin small letter a with tilde
228	344	E4	11100100	ä	ä	ä	Latin small letter a with diaeresis
229	345	E5	11100101	å	å	å	Latin small letter a with ring above
230	346	E6	11100110	æ	æ	æ	Latin small letter ae
231	347	E7	11100111	ç	ç	ç	Latin small letter c with cedilla
232	350	E8	11101000	è	è	è	Latin small letter e with grave
233	351	E9	11101001	é	é	é	Latin small letter e with acute
234	352	EA	11101010	ê	ê	ê	Latin small letter e with circumflex
235	353	EB	11101011	ë	ë	ë	Latin small letter e with diaeresis
236	354	EC	11101100	ì	ì	ì	Latin small letter i with grave
237	355	ED	11101101	í	í	í	Latin small letter i with acute
238	356	EE	11101110	î	î	î	Latin small letter i with circumflex
239	357	EF	11101111	ï	ï	ï	Latin small letter i with diaeresis
240	360	F0	11110000	ð	ð	ð	Latin small letter eth
241	361	F1	11110001	ñ	ñ	ñ	Latin small letter n with tilde
242	362	F2	11110010	ò	ò	ò	Latin small letter o with grave
243	363	F3	11110011	ó	ó	ó	Latin small letter o with acute
244	364	F4	11110100	ô	ô	ô	Latin small letter o with circumflex
245	365	F5	11110101	õ	õ	õ	Latin small letter o with tilde
246	366	F6	11110110	ö	ö	ö	Latin small letter o with diaeresis
247	367	F7	11110111	÷	÷	÷	Division sign
248	370	F8	11111000	ø	ø	ø	Latin small letter o with slash
249	371	F9	11111001	ù	ù	ù	Latin small letter u with grave
250	372	FA	11111010	ú	ú	ú	Latin small letter u with acute

251	373	FB	11111011	û	&#251;	&ucirc;	Latin small letter u with circumflex
252	374	FC	11111100	ü	&#252;	&uuml;	Latin small letter u with diaeresis
253	375	FD	11111101	ý	&#253;	&yacute;	Latin small letter y with acute
254	376	FE	11111110	þ	&#254;	&thorn;	Latin small letter thorn
255	377	FF	11111111	ÿ	&#255;	&yuml;	Latin small letter y with diaeresis

Existem diversas variações da tabela ASCII de 8 bits. A Tabela 5 é de acordo com ISO 8859-1, também chamado ISO Latin-1. Códigos 129-159 contêm o Microsoft ® Windows latino-1 caracteres estendidos.
Fonte: http://www.ascii-code.com/