UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ

FACULDADES LUIZ MENEGHEL

RAFAEL GUSTAVO PAIXÃO

CLASSIFICAÇÃO DE INFORMAÇÕES RELEVANTES CONTIDAS EM BASES TEXTUAIS DE WIKI USANDO MINERAÇÃO DE TEXTOS

Bandeirantes

2011

UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ

CAMPUS LUIZ MENEGHEL

RAFAEL GUSTAVO PAIXÃO

CLASSIFICAÇÃO DE INFORMAÇÕES RELEVANTES CONTIDAS EM BASES TEXTUAIS DE WIKI USANDO MINERAÇÃO DE TEXTOS

Bandeirantes

2011

RAFAEL GUSTAVO PAIXÃO

CLASSIFICAÇÃO DE INFORMAÇÕES RELEVANTES CONTIDAS EM BASES TEXTUAIS DE WIKI USANDO MINERAÇÃO DE TEXTOS

Trabalho de Conclusão de Curso submetido à Universidade Estadual do Norte do Paraná – *Campus* Luiz Meneghel - como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Msc. André Luis A. Menolli.

Bandeirantes

2011

RAFAEL GUSTAVO PAIXÃO

CLASSIFICAÇÃO DE INFORMAÇÕES RELEVANTES CONTIDAS EM BASES TEXTUAIS DE WIKI USANDO MINERAÇÃO DE TEXTOS

Trabalho de Conclusão de Curso submetido à Universidade Estadual do Norte do Paraná – *Campus* Luiz Meneghel - como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

COMISSÃO EXAMINADORA

_____
Prof. Msc. André Luis A. Menolli
UENP – *Campus* Luiz Meneghel

_____
Prof. Msc. Ederson Marcos Sgarbi
UENP – *Campus* Luiz Meneghel

_____
Prof. Msc. Glauco Carlos Silva
UENP – *Campus* Luiz Meneghel

Bandeirantes, __ de_____ de 2011

A meu pai, Moacir Paixão,

mesmo não estando aqui presente,

sinto muitas saudades e nesse momento deve estar

muito orgulhoso por essa conquista.

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus, por iluminar e abençoar para conquista de mais um sonho.

A minha mãe querida que sempre me apoio nesses quatros anos de faculdade, sempre me compreendendo e incentivando nos momentos difíceis e soube entender às vezes minha ausência devido aos estudos. Obrigado por acreditar em mim.

A minha namorada Carina, uma pessoa muito especial que a amo, que sempre acreditou em mim, me deu forças para continuar e teve muita paciência mesmo nas horas que não podíamos ficar juntos.

A minha irmã Simone e ao meu cunhado César que me apoiaram a lutar pelos estudos e passar suas experiências para seguir no caminho certo.

Ao meu irmão gêmeo Rodrigo, por se fazer presente, apesar da distância física, torcendo e acreditando em mim.

Ao meu orientador, professor André Menolli, o qual dedicou parte do seu tempo para ajudar no desenvolvimento do trabalho. Muito obrigado pela atenção, paciência, orientação e amizade.

Sou muito grato ao professor Luiz Fernando Nascimento do departamento de informática da UENP, pela sua paciência, orientação e atenção no desenvolvimento do estágio de extensão. Muito obrigado.

A amizade e companheirismo de todos da XIII turma de Sistemas de Informação, em especial meus grandes amigos durante o curso: Lélis, Wagner, Renan e Paulo.

Agradeço muito meus amigos da República “Thor óh Chinelo” (Jonatã “Caminhão”, Murilo “Mathias”, Rafael ”Pexe”, Renan “Pedreiro”, Felipe “Pato”, Ivan “Dylon”, Anderson “Califórnia”, Andrei “Museu”, João “Pagodero”, Lélis “Léli” e Alex “Ow rapaz”) pela amizade verdadeira que fiz durante os quatro anos de faculdade.

E por fim agradeço a todas as pessoas que direta ou indiretamente contribuíram para realização deste trabalho.

"A mente que se abre
a uma nova idéia
jamais voltará ao
seu tamanho original."

Albert Einstein

RESUMO

Nas organizações existem várias situações de incertezas e ambientes em constante mudança, com isso as empresas devem se preocupar em gerar novos conhecimentos. A aprendizagem organizacional tem uma grande importância para organizações, pois para obter sucesso e competitividade no mercado as organizações têm que aprender por meio dos indivíduos ou grupo de pessoas pela experiência adquiridas no dia-a-dia. Algumas tecnologias, como a denominada web 2.0 podem ser consideradas como facilitadoras para atingir uma organização de aprendizagem, pois auxiliam de alguma maneira no aprendizado. Com isso, a tecnologia *web* 2.0 escolhida neste trabalho é a wiki, uma ferramenta livre colaborativa que permite a criação de textos web para registros de informações. Dessa forma, o desenvolvimento do trabalho foi modelar um domínio por meio das ontologias, para construção de um ambiente semântico utilizando a ferramenta MediaWiki, com intuito de organizar o conteúdos relevantes através das propriedades semânticas contidas no ambiente, e aplicar a mineração de textos nas bases textuais da wiki com os algoritmos de classificação da ferramenta *Rapidminer*, para verificar o nível de precisão de eficiência dos algoritmos dos modelos proposto para classificação do conhecimento.

Palavras-chave: Aprendizagem Organizacional, Ontologia, Wiki, Mineração de Textos.

ABSTRACT

In organizations there are several situations of uncertainty and changing environments, with that companies should worry about generating new knowledge. The organizational learning has great importance for organizations, as for success and competitiveness in the market the organizations have to learn through individuals or persons group by the experience acquired daily. Some technologies, such as the so-called Web 2.0 can be considered as facilitators to achieve a learning organization, because they help in some way in learning. Thus, the web 2.0 technology chosen in this work is the wiki, a free collaborative tool that allows the creation of web texts for information records. Thus, the development work was to model a domain by means of ontologies for the construction of a semantic environment using the MediaWiki tool, aiming to organize relevant content through semantic properties contained in the environment, and apply text mining bases the wiki text classification algorithms with RapidMiner tool to verify the accuracy level of efficiency of algorithms in the models proposed for classification of knowledge.

Keywords: Organizational Learning, Wiki, Ontologies, Text Mining.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

GPL/GNU	Licença Pública Geral
HTML	*HyperText Markup Language*
OWL	*Web Ontology Language*
PHP	*Hipertext preprocessor*
TI	Tecnologia da Informação
W3C	*World Wide Web Consortium*
MySQL	*Structured Query Language*

SUMÁRIO

1 INTRODUÇÃO

Nas organizações existem várias situações de incertezas e ambientes em constante mudança, com isso as empresas devem se preocupar em aprender e envolver os seus funcionários para desenvolver e gerar novos conhecimentos.

A aprendizagem organizacional tem uma grande importância para organização dos conhecimentos nas organizações, pois para obter sucesso e competitividade no mercado, as organizações têm que aprender através dos indivíduos ou grupos de pessoas. O aprendizado ocorre pela convivência com outras pessoas e com as experiências adquiridas no dia-a-dia.

Peter Senge (1990), explica em seus textos que a aprendizagem organizacional está interligada com os aspectos do ser humano no ato de aprender, explorar e experimentar com a necessidade que o mundo exige.

Algumas tecnologias podem ser consideradas como facilitadores para atingir uma organização de aprendizagem, pois auxiliam de alguma maneira no aprendizado. As tecnologias da *Web* 2.0 facilitam a colaboração e a utilização das informações nas organizações, as principais tecnologias existentes utilizadas na aprendizagem organizacional são: *Web* Blogs, Wikis e *Folkosonomias* (RECH E RAS, 2008).

A tecnologia wiki escolhida para desenvolvimento desse trabalho, visa auxiliar na aprendizagem organizacional, focando a organização e compartilhamento do conhecimento. A tecnologia wiki, palavra em havaiano que significa “rápido”, trata-se de um ambiente colaborativo que permite a realização de textos *web* para registros de informações.

Segundo Schweitzer (2008), a ferramenta wiki é um site que pode ser editável por qualquer pessoa que tenha um computador com acesso a internet. As wikis utilizam em sua interface uma sintaxe simples e rápida, possibilitando aos usuários a criação, edição e *hiperlinks* em textos entre as páginas da wiki.

Gruber (1991) afirma que além das tecnologias da *Web* 2.0, existem também as ontologias que auxiliam na construção e conhecimento sobre um determinado domínio. A ontologia representa formalmente o conhecimento de um determinado domínio por meio de conceitualização.

Outra técnica utilizada nesse trabalho, à mineração de textos, pode ser definida como técnicas de análise textuais, extração do conhecimento, clusterização

e categorização que possibilita descobrir conhecimento inovador nos textos (REZENDE, 2005). Para organizar as informações contidas nas wikis, será aplicado o processo de mineração de textos que visa encontrar e organizar o conhecimento.

Portanto, o presente trabalho tem a finalidade de aplicar as técnicas de mineração de textos, com intuito de organizar os conhecimentos gerados nas bases textuais da wiki, estabelecendo assim qual algoritmo de classificação é o mais eficiente para estabelecer a classificação dos conhecimentos.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

O objetivo principal deste trabalho é modelar o conhecimento de um determinado domínio presente em uma wiki por meio de ontologias e aplicar a técnica de Mineração de Textos às informações geradas pelas wikis, com intuito de encontrar conteúdos relevantes e classificá-los para organizar o conhecimento.

1.1.2 Objetivos Específicos

Durante o desenvolvimento e ao término desta pesquisa, pretende-se:

- Escolha da ferramenta wiki;
- Gerar textos na wiki;
- Identificar e definir quais informações a serem armazenadas;
- Definir o domínio do problema;
- Modelar uma estrutura para organizar estas informações na forma de ontologias;
- Preparação dos dados para aplicação;
- Documentar as alterações realizadas na wiki;
- Escolha da ferramenta para mineração de textos;
- Escolha dos algoritmos a serem aplicados;
- Aplicação dos algoritmos nas bases de textos;
- Classificar o conhecimento gerado na wiki; e
- Análise dos resultados.

1.2 JUSTIFICATIVA

Esta pesquisa justifica-se pela necessidade de organizar o conhecimento que são gerados nas organizações e não são aproveitados de maneira correta, com isso são perdidos por não terem uma organização eficaz. Com o constante crescimento do mercado a competitividade torna-se um ponto de grande importância, entretanto o conhecimento é o principal diferencial em uma organização para o crescimento e a obtenção do sucesso (CASTILHO et.al., 2004).

Segundo Dias (2009), as organizações estão investindo cada vez mais nas evoluções da tecnologia da informação (TI), nas quais tornando as tecnologias um conjunto de facilidades para disseminações de informações e conhecimento. As organizações estão usando as tecnologias da *Web* 2.0 na expectativa de redução de custos e aumento da competitividade, os investimentos nas tecnologias *Web* 2.0, visam à ampliação da colaboração interna de uma organização.

Outra justificativa na realização do trabalho é aplicar a mineração de textos nas bases textuais da wiki, com intuito de obter informações com a precisão de índice de acertos com os melhores algoritmos de classificação para tornar o processo de descoberta de conhecimento mais eficaz.

No entanto, a tecnologia wiki ajudará com auxílio da mineração de textos na classificação e organização do conhecimento. A vantagem de utilizar a ferramenta wiki é a capacidade de interligar as informações de um grupo de pessoas na organização, gerando assim uma melhor qualidade e flexibilidade na disseminação do conhecimento.

1.3 MATERIAIS E MÉTODOS

O trabalho trata-se inicialmente de uma revisão bibliográfica, com utilização de livros, revistas, dissertações teses e artigos científicos para levantamento das informações referentes sobre o tema.

De acordo com os critérios de classificação, o trabalho apresenta a seguinte configuração metodológica (GONSALVES, 2001).

- Segundo aos objetivos: A pesquisa deste trabalho se define como exploratória, pois visa compreender sobre o assunto que está sendo estudado;

- Segundo o método de coleta: A pesquisa é um estudo de caso, pois será realizada uma análise específica do estudo e investigado profundamente o problema que será pesquisado;

- Segundo as fontes de informações: Foram obtidas através da pesquisa bibliográfica e documental, pois utiliza contribuições de vários autores sobre um determinado assunto;

- Segundo a natureza dos dados: De acordo com os dados é uma pesquisa qualitativa, por buscar profunda compreensão e interpretação do fenômeno estudado.

As ferramentas utilizadas no desenvolvimento do trabalho são: Protégé, MediaWiki , *Rapidminer* e *Cmap Tools*.

- Protégé: Ferramenta para construção das ontologias;
- MediaWiki: Utilização da MediaWiki para construção do ambiente semântico;
- *Rapidminer*: Ferramenta para aplicação da mineração de textos; e
- *Cmap Tools*: Para construção do mapa conceitual.

1.4 DIVISÃO DO TRABALHO

O restante do texto está dividido como segue. No capítulo 2, é apresentada a fundamentação teórica, sobre Aprendizagem Organizacional, teorias sobre Aprendizagem Organizacional e a diferença entre Organizações de Aprendizagem e Aprendizagem Organizacional.

Na capitulo 3 apresenta a tecnologia da ferramenta wiki, demonstrando o histórico, conceitos, características, vantagens e desvantagem e a ferramenta MediaWiki .

No capitulo 4 apresenta a *Web* Semântica, apresentando as ontologias, os componentes das ontologias, os tipos de ontologias, vantagens do uso da ontologia.

No capítulo 5 apresenta a Mineração de Textos, apresentando os processos, os tipos de abordagens de dados, preparação dos dados, análise dos dados, processamento dos dados e a pós-processamento.

No capitulo 6 apresenta o desenvolvimento do trabalho com a modelagem do mapa conceitual para organizar e representar de maneira clara o conhecimento e a modelagem de um determinado domínio por meio das ontologias na ferramenta Protégé.

No capitulo 7 apresenta o ambiente semântico MediaWiki, demonstrando as etapas criadas que são: propriedades, predefinições, categorias e formulários.

No capitulo 8 é aplicação da mineração de textos das bases textuais da wiki, realizando testes e validações com os algoritmos mais importantes de classificação da ferramenta *Rapidminer*.

No capitulo 9, 10, 11 são apresentados, respectivamente os resultados obtidos, conclusão e a referências bibliográficas.

2 APRENDIZAGEM ORGANIZACIONAL

As organizações encontram-se em um ambiente cada vez mais competitivo, com isso as organizações tem a necessidade de realizar um processo de mudança dando importância na aprendizagem.

Argyris (1978) define a aprendizagem organizacional como um processo que identifica os erros e tem por finalidade de fazer a correção. Já Kim (1993), entende que a aprendizagem organizacional é o aumento de valores em uma organização em executar ações eficientes. Outro conceito é descrito por Nevis (1995), o qual define que a aprendizagem organizacional é o desenvolvimento de processos e competência dentro da organização, com o objetivo de manter e melhorar a capacidade de desempenho na base de experiências.

De acordo com Probst e Buchel (1997), o processo de mudança da base e dos conhecimentos da organização, é o ato de possibilitar as resoluções de problemas que possam acontecer dentro da organização. Com isso, a aprendizagem organizacional é um elemento decisivo na resolução de problemas, tendo assim competitividade no mercado e possibilitando as mudanças de base tecnológica.

A aprendizagem organizacional é a características de uma organização em obter conhecimentos com sua experiência e com as experiências externas, com isso modificar sua forma de funcionar (ZANGISKI et al.,2009).

Silva et al.,(2009), define que aprendizagem organizacional é a admissão de pessoas e grupos para desenvolver e adquirir conhecimento e habilidades para agir de maneira eficaz na tomada de decisão dentro de uma organização.

2.1 TEORIA DA APRENDIZAGEM ORGANIZACIONAL

São encontradas muitas teorias de aprendizagem organizacional e modelos de conhecimento e aprendizagem. Esses modelos descrevem os processos de níveis individuais e organizacionais.

Bjørnson e Dingsøyr (2008) citam as quatro principais teorias de aprendizagem organizacional existente que são:

- O modelo de aprendizagem experiencial de Kolb;
- A teoria de aprendizagem de duplo circuito de Argyris e Schön;
- A teoria de comunidades de prática de Wenger; e

- A teoria de criação do conhecimento de Nonaka e Takeuchi.

O modelo de aprendizagem experiencial Kolb (1984), é chamado de “Aprendizagem Experiencial” por se tratar do papel que representa pela experiência no processo de aprendizagem.

Outra teoria é apresentada por Argyris e Schön (1978), para quais definem a duas maneiras de aprendizagem que são: aprendizagem de ciclo simples e aprendizagem de ciclo duplo.

- A aprendizagem de ciclo simples: Significa o ato de observar acontecimentos e efeitos, através destas observações possibilitando mudar e melhorar o processo;

- E a aprendizagem de ciclo duplo: Entendem não apenas os efeitos de um processo ou cadeia de eventos e sim como os fatores podem influenciar os efeitos.

Na teoria de comunidades de prática de Wenger (1998) é definida uma comunidade os quais são desenvolvidas práticas de aprendizagem rituais, rotinas, artefatos, símbolos, convenções e histórias.

A teoria de criação do conhecimento de Nonaka e Takeuchi é conhecida como modelo em espiral, esse modelo o aprendizado é gerado através da interação do conhecimento tácito e explícito. Existem quatros formas que compõem esse modelo que são: socialização, combinação, internalização e externalização como demonstrados na Figura 1.

Figura 1: Espiral do conhecimento (PEREIRA, 2010 et al., apud NONAKA; TAKEUCHI, 1997)

As definições desses quatro conceitos são:

1) Socialização: A socialização é adquirida pelas experiências compartilhadas dos indivíduos.

2) Externalização: Na externalização é caracterizada pelo processo, a qual existe a conversão do conhecimento tácito em explícito.

3) Internalização: O processo de internalização é o processo que existe a conversão do conhecimento explicito em tácito para outro indivíduo.

4) Combinação: O processo de combinação é a conversão de vários conjuntos de conhecimento explícitos em somente em um único conjunto.

2.2 ORGANIZAÇÕES DE APRENDIZAGEM X APRENDIZAGEM ORGANIZACIONAL

As definições dos conceitos de Aprendizagem Organizacional (*Organizational Learning*) e Organizações de Aprendizagem (*Learning Organizations*), geralmente as definições são confundidas e muitas vezes são considerados sinônimos. Nesta seção são demonstrados os conceitos e a diferenças desses dois termos.

Como visto anteriormente na literatura sobre aprendizagem organizacional é definida como indivíduos ou grupos aprendem em determinadas tipos de atividades dentro de uma organização. Já organização de aprendizagem significa os tipos de atividades que acontecem nas organizações.

De acordo com Garvin (1993), entende-se que uma organização de aprendizagem é aquela que aprende com a capacidade de gerar, transportar o conhecimento e refletir sobre o comportamento de novos conhecimentos.

Senge (1990) define organizações de aprendizagem como:

> *"(...) as pessoas expandem continuamente sua capacidade de criar os resultados que realmente desejam, onde surgem novos e elevados padrões de raciocínio, onde a aspiração coletiva é liberada e onde as pessoas aprendem continuamente a aprender em grupo."* (Senge, 1990, p. 11).

Teodoro e Ottoboni (2005), explica que existe diferença entre os dois conceitos, e que a aprendizagem organizacional é adquirida pela aprendizagem individual dentro da organização com um grupo de pessoas. Já as organizações de aprendizagem aprendem com as características obtidas dentro da organização.

3 TECNOLOGIA WIKI

3.1 HISTÓRICO

O termo wiki originou-se da palavra “WikiWiki”, que no seu significado em havaiano significa “rápido”. A primeira wiki foi criada em 1995 pelo norte americano Ward Cunningham e tornou-se popular com o aparecimento da enciclopédia livre (Wikipédia) (SCHONS E COUTO, 2007).

A primeira wiki era chamada Portland Pattern Repository, seu principal objetivo era publicar informações em um ambiente colaborativo.

3.2 CONCEITOS DE WIKI

Uma wiki é uma ferramenta colaborativa na *web*, permite que várias pessoas possam construir documentos em conjunto de forma colaborativa. Também possibilitam viabilizar os registros das informações e também podem corrigir informações desatualizadas e incorretas (SCHWEITZER, 2008).

Segundo Couto e Blattmann (2000), a ferramenta wiki é um software livre de aplicação *web*, que possibilita ao usuário a utilizar um ambiente ágil e simples para modificação e publicação de textos. Essa tecnologia é utilizada a partir de um *browser*.

Camara (2009) apud Valente e Mattar (2007) define wiki como:

> *“(...) O wiki é um software colaborativo que permite a edição coletiva de documentos de uma maneira simples. Em geral, não é necessário registro, e todos os usuários podem incluir, alterar ou até excluir textos sem que haja revisão antes de as modificações serem aceitas .”* Valente e Mattar, 2007, p. 102).

3.3 CARACTERÍSTICAS DA FERRAMENTA WIKI

Couto e Blattmann (2000) apud Faquetti e Alves (2006) descrevem as principais características da ferramenta wiki que são:

a) *Software* livre de fácil instalação e compatível com as plataformas *Linux* e *Windows*;

b) Permite discussão assíncrona;

c) Permite importação e exportação de textos e imagens facilitando a criação automática de hipertexto e *hiperlinks*;

d) Não existe qualquer mecanismo de revisão preliminar à publicação, portanto a responsabilidade pela qualidade das contribuições é de cada participante autorizado;

e) A autorização para contribuir no sistema pode ser programada pelo grupo gestor, podendo ser ampla e irrestrita ou possuir algumas restrições como, por exemplo, estar cadastrado.

3.4 VANTAGENS E DESVANTAGENS

Existem diversas ferramentas wikis para uso das organizações ou qualquer fins colaborativos, com isso alguns autores descrevem os benefícios e a qualidade que existem nas wikis em geral.

As vantagens apresentada por (BEAN e HOTT, 2005; WEL et al., 2005; O'LEARY, 2008) são:

- Os usuários têm uma grande facilidade de mexer com a ferramenta, podendo acrescentar textos e fazer sua edição com aparência do *Microsoft Word*;
- Existem várias versões wikis sendo livre e gratuito com seu código aberto;
- Pessoas em qualquer parte do mundo podem trabalhar simultaneamente;
- As wikis podem ser utilizadas como programas de treinamentos nas empresas;
- Muitas wikis permitem o mecanismo de 'discussão' e 'comentários' sobre um determinado assunto, os usuários podem colocar sua opinião e explicar a razão de suas posições;
- Pode ser encontrada gratuitamente uma grande diversidade de arquivos.

Já as desvantagens apresentada por (BEAN e HOTT, 2005; WEL et al., 2005; O'LEARY, 2008) são:

- Com a grande facilidade de edição de um documento na wiki, o usuário pode alterar e eliminar informações relevantes;

- Freqüentemente nas wikis não fornecem informações sobre os autores, com isso o usuário não tem a confiança, qualidade e a consistência do documento;

- Wikis estão sujeitas a vandalismo;

- Os usuários querem ter a segurança que o conteúdo obtido online seja de total confiança, ou seja, apoiado por alguma autoridade especializada;

- Nas wikis podem ocorrer riscos em relação da segurança da informação, nas quais usuários podem compartilhar dados indevidos que são de grande importância para uma organização;

Segundo Schweitzer (2008) apud Bordignon (2007) apresentam as vantagens e desvantagens de uma wiki:

> *Uma wiki em essência é um espaço de trabalho colaborativo, na qual um grupo de usuários trabalham em uma organização simples, constroem documentos virtuais de múltipla autoria, utilizando um formato simples. Das análises realizadas se observam os seguintes pontos fortes ou vantagens derivadas do uso desta ferramenta de edição aberta: mínimo treinamento e suporte técnico aos colaboradores; as atualizações de conteúdo são imediatas; possibilidade de acesso aberto; todos os conteúdos são revisados, corrigidos e expandidos pelos pares, é uma busca estratégica para preservar conhecimentos. As desvantagens são: a ausência de uma organização forte; problemas de vandalismo e que ainda não há uma única linguagem para definir a linguagem de edição.*

3.5 FERRAMENTA MEDIAWIKI

A ferramenta escolhida para o desenvolvimento do trabalho será MediaWiki[1], um programa gratuito com seu código aberto para edição. O MediaWiki é um programa feito na linguagem PHP e utiliza o sistema gerenciador de banco dados MySQL.

[1] Disponível em http://www.mediawiki.org/wiki/MediaWiki

Segundo Estivalete (2007), a MediaWiki é um software livre licenciado sob GNU/GPL[2], e as principais características são sua facilidade de uso e a criação e edição de páginas.

A Figura 2 representa a Tela Inicial do MediaWiki.

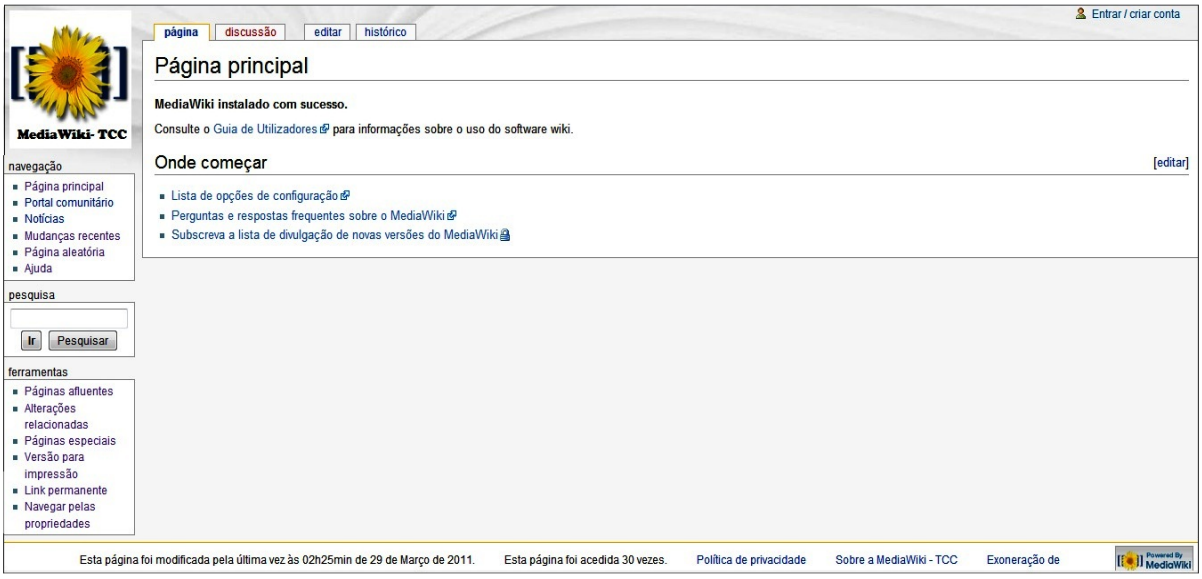


Figura 2: Tela Inicial do MediaWiki

[2] Disponível em http://www.gnu.org/licenses

4 WEB SEMÂNTICA

Uma das definições mais importantes apresentadas para a *Web* Semântica é do criador da linguagem HTML e líder da criação da W3C Tim Berners-Lee, ou seja, o idealizador da *web* semântica.

Para Berners-Lee et al. (2001) o conceito de *web* semântica é definida como:

> *"A Web Semântica é um extensão da web atual, onde a informação possui um significado claro bem definido, possibilitando uma melhor interação entre computadores e pessoas."*

Segundo Dziekaniak e Kirinus (2004), a *web* tradicional foi desenvolvida para entendimento para os usuários, já a *web* semântica está sendo desenvolvida para entendimento e compreensão das máquinas, com auxílio de agentes computacionais que são programas que operam na *web* em busca de informações eficientes, entendendo seu significado, com isso irão auxiliar os usuários em operações na *web*.

Já Moura (2002) define *Web* Semântica:

> *A Web Semântica é um dos objetivos de longo prazo da W3C. Deverá se desenvolver num ambiente de acesso inteligente à informação heterogênea e distribuída, através de agentes de softwares. Estes agentes irão mediar e realizar o brokering entre as necessidades de casa usuário e as fontes de informação disponíveis, permitindo pesquisas mais acuradas e eficientes.*

É possível através das ontologias representarem explicitamente a semântica dos dados, que possibilitam elaborar uma rede enorme de conhecimento humano e na melhora no processamento das máquinas e do nível de serviços na *web* (DZIEKANIAK E KIRINUS, 2004).

4.1 ONTOLOGIAS

O termo ontologia onto (ser) e "logia (discurso escrito ou falado) teve sua origem nos pensamentos filosóficos de Aristóteles. É um estudo introduzido na filosofia com objetivo de distinguir o estudo do ser e dos seres das ciências naturais. ALMEIDA (2003).

Segundo Borst (1997) define ontologia como: "Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada". Nessa

afirmação entende-se que “especificação formal” significa que é legível para compreensão dos computadores; “especificação explícita” pode entender os conceitos básicos de ontologias que são: conceitos, relações, instâncias, propriedades e axiomas; “conceitualização”, trata-se de algum modelo abstrato do fenômeno natural; “compartilhada” entende-se por conhecimento consensual.

Na construção de uma ontologia é necessário conhecer o tipo do domínio em questão, aos critérios de construção, ter uma metodologia para essa construção e a definição de uma linguagem e ambiente para ser trabalhado (ALMEIDA E BAX, 2003).

4.1.1 Componentes de Ontologias

Silva (2006) define os termos e as relações dos componentes básicos para construção das ontologias que são: Conceitos, classes, relações, funções, axiomas e instâncias. A seguir, estes itens são detalhados.

- Conceitos: A representação dos conceitos pode representar vários sentidos, o conceito pode ser abstrato ou concreto, real ou fictício;
- Classes: As classes constituem um grupo de indivíduos organizados em uma taxonomia;
- Relações: Esse componente apresenta uma associação entre os conceitos do domínio, por exemplo, os conceitos de Livro e Autor é um relacionamento escrito - por;
- Funções: A função é um componente especial de relações, tem a características de um conjunto de elementos apresentarem uma relação única com outro elemento;
- Axiomas: Servem para modelar sempre sentenças que são verdadeiras; e
- Instâncias: As instâncias apresentam são usadas para representar elementos da ontologia, por exemplo, Curitiba pode representar uma instância para o conceito capital.

4.1.2 Tipos de Ontologias

As ontologias nunca representam a mesma estrutura, mas em grandes partes delas é comum existir características e componentes básicos, a ontologia pode apresentar propriedades distintas e com tipos bem definidos (SILVA, 2006).

Alguns tipos de ontologias são apresentados a seguir:

- Ontologias de alto-nível: Descrevem conceitos muitos gerais, por exemplo, matéria, tempo, espaço, evento, objeto, etc.
- Ontologias de Tarefas: A ontologias de tarefa representa um vocabulário de tarefas genéricas, ou conceitos de especializações presentes nas ontologias de alto nível;
- Ontologias de Domínio: As ontologias de domínio representam o ramo de estudo de uma mesma área genérica de conhecimento abstrato;
- Ontologias de Aplicação: A característica dessa ontologia é procurar um problema específico de um domínio.

Guarino (1998) apresenta os tipos de ontologias e a classificação quanto à generalidade das ontologias, como pode ser visto na Figura 3.

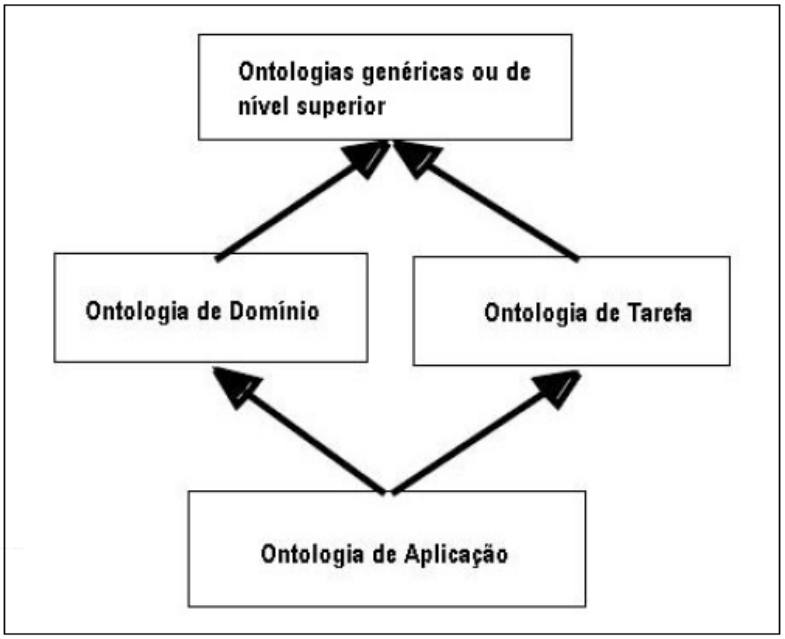


Figura 3: Tipos de ontologias (Adaptado de Guarino, 1998)

4.1.3 Vantagens dos usos de Ontologias

Neste tópico apresenta as vantagens de uso na utilização de uma ontologia. A seguir é apresentada uma lista das vantagens por Kern (2007).

- No termo de reuso das ontologias apresentam um vocabulário com grande consenso e o domínio apresentado de forma explícita, com isso possibilita um alto potencial de reuso;

- As ontologias apresentam um vocabulário para representação do conhecimento, que evita interpretações ambíguas;

- A comunicação nas ontologias são importantes e disponibilizadas em várias formas para ajudar as pessoas a se comunicarem sobre algum tipo de conhecimento, a comunicação possibilita que as pessoas entendam e raciocinem sobre o domínio do conhecimento;

- Ontologias permitem o compartilhamento dos conhecimentos. Por exemplo, caso exista alguma ontologia já modelada com algum tipo de domínio de conhecimento, assim pode ser compartilhada para quem deseja trabalhar com o mesmo domínio.

5 MINERAÇÃO DE TEXTOS (*TEXT DATA MINING*)

A definição para mineração de textos é caracterizada por um conjunto de técnicas e processos que possibilita descobrir o conhecimento inovador nos textos. (REZENDE, 2005).

Segundo Barion e Lago (2008) afirma que 80% das informações de uma organização e conteúdos que são disponibilizados on-line estão documentados em formatos de textos. Com isso, conclui-se que somente 20% das informações são utilizadas para manipulação com a finalidade de tomada de decisão.

O processo de extração em mineração de textos se preocupa em obter informações nos textos e tratá-los, com objetivo de gerar algum tipo de conhecimento que seja de grande utilidade e inovador para o usuário. Para gerar esse conhecimento eficiente para usuário torna-se a fase de pré-processamento a principal etapa do processo na mineração de textos.

A Figura 4 mostra as principais etapas do processo de mineração de texto que são: Pré-processamento, Extração do Conhecimento e o Pós-processamento.

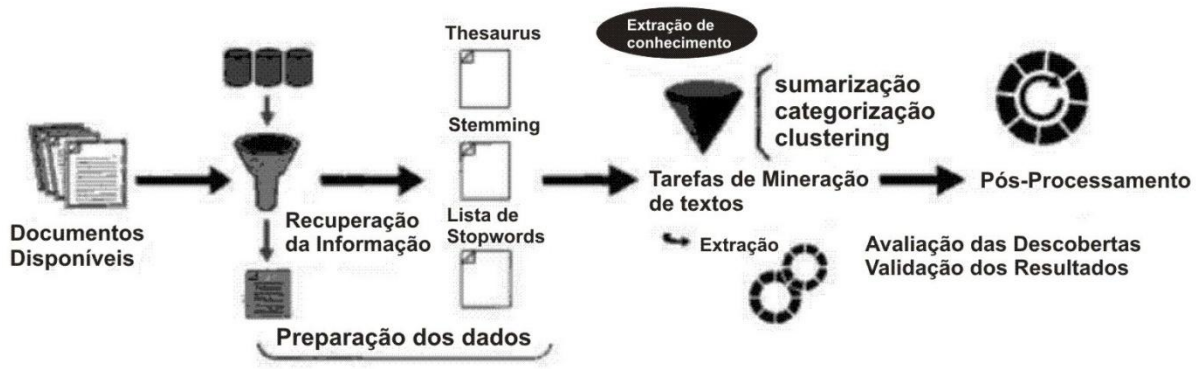


Figura 4: Etapas do processo de Mineração de Textos (REZENDE, 2005)

5.1 O PROCESSO DE MINERAÇÃO DE TEXTOS

O primeiro passo no processo de mineração de textos é identificar o tipo de dados que serão trabalhados para realização da extração do conhecimento. A identificação dos dados e o processo na mineração de textos são apresentados a seguir.

5.1.1 Tipos de Abordagem dos Dados

Segundo Rezende (2005) existe duas formas de abordagens de dados textuais na área de mineração de textos. A análise semântica e a análise estatística.

- Análise Semântica

A análise semântica caracteriza em empregar técnicas que possam fazer uma avaliação na seqüência dos termos no contexto da frase, para identificar a função correta em cada termo. Nessa análise são utilizadas técnicas baseadas no processamento de linguagem natural.

Na realização para o processamento da linguagem natural é preciso ter o entendimento dos vários tipos de linguagem do conhecimento morfológico, conhecimento sintático, conhecimento semântico, conhecimento programático, conhecimento do discurso e conhecimento do mundo. São definidos a seguir os tipos de conhecimentos:

a) Conhecimento Morfológico: São conhecimentos da forma e estrutura das inflexões das palavras;

b) Conhecimento Sintático: O conhecimento apresenta uma estrutura das listas de palavras e como a combinação das palavras podem produzir sentenças;

c) Conhecimento Semântico: Explica o conhecimento do significado das palavras no contexto independentes, e como é realizada a combinação das palavras nos significados mais complexos;

d) Conhecimento Programático: Conhecimento do uso da língua em diferentes contextos, e como é afetado o contexto pelo seu significado e interpretação;

e) Conhecimento do Discurso: E definida como as sentenças precedentes podem prejudicar na interpretação da próxima sentença;

f) Conhecimento do Mundo: Esse tipo de conhecimento significa a linguagem natural que se relaciona pela comunicação do domínio ou o mundo.

- Análise Estatística

A análise estatística entende-se na qual os termos relevantes que aparecem pelos números de vezes nos textos. Esse tipo de estratégia o aprendizado estatístico pode ser conduzido em qualquer idioma.

Serão abordados a seguir as etapas do aprendizado estatístico dos dados e os modelos de representação de documentos utilizados na análise estatística.

a) Codificação dos Dados: A codificação dos dados é escolhida por uma indicação de um especialista ou com algum critério das propriedades dos dados que reflitam aos objetivos da modelagem.

b) Estimativas de Dados: Nesse estágio tem a finalidade da procura de um modelo adequado a partir de um amplo conjunto de modelos possíveis. Para obter um modelo para dados é aplicado um algoritmo de aprendizado ou um método de estimativa.

c) Modelos de Representação de Documentos: Nos documentos de coleções de documentos, são chamados de recipientes de palavras. Essa abordagem é conhecida como *bag of words*, que significa ignorar a ordem das palavras de qualquer informação de pontuação ou estrutural. A codificação *bag of words* é vista como uma simplificação de uma grande quantidade de informações expressa por um documento, essa abordagem não oferece confiabilidade de seu conteúdo.

5.1.2 Preparação dos dados

Rezende (2005) afirma que a primeira etapa no processo de descoberta do conhecimento em textos é a preparação dos textos.

A principal etapa desse processo é a seleção dos dados, nesta fase irá se formar a base de textos para a realização da mineração de textos, no processo as informações que não forem relevantes poderá ser descartadas.

A Recuperação de Informação (RI) é uma área que é desenvolvida os modelos para representação de uma variedade de coleções de textos que possibilitam identificar tópicos específicos em documentos (REZENDE, 2005).

Os principais métodos de RI criados são: Modelo Booleano, Modelo Espaço Vetorial, Modelo Probabilístico, Modelo Busca Direta, Modelo aglomerado de (*Clusters*) e Modelo Contextual ou Conceitual.

1) Modelo Booleano: No modelo booleano é considerado como um conjuntos de palavras contidas nos documentos. São utilizados conectores de *boolean* (*and, or* e *not*) para descrever e manipular esses conjuntos.

2) Modelo Espaço Vetorial: Neste modelo é uma representação de vetores e cada vetores possui termos. Todos os termos nos vetores são associados com um valor (denominado peso) que indica importância de cada um.

3) Modelo Probabilístico: O modelo representa conceitos da área de estatística e probabilidade. Neste modelo a probabilidade busca saber se um documento pode ser relevante em uma consulta.

4) Modelo Busca Direta: O modelo é também conhecido como modelo de busca de padrões, esse modelo localiza documentos importantes através de métodos de busca de *strings*.

5) Modelo aglomerado de (*Clusters*): A características desse modelo é encontrar e identificar documentos similares em conteúdo e armazenar no mesmo grupo ou aglomerado que é chamado de '*clusters*'.

6) Modelo Contextual ou Conceitual: As consultas neste modelo levam-se em consideração o contexto dos documentos e da busca do usuário e não somente dos termos presentes na consulta.

5.1.3 Análise dos dados

O principal objetivo desta fase é identificar a similaridades dos significados entre as palavras.

Segundo Rezende (2005) para a realização da análise dos dados existem 3 processos que são: *Stopwords*, *Stemming* e Uso de Dicionário ou *Thesaurus*.

Stopwords são palavras que não tem relevância na análise de textos, são termos e palavras que não traduzem a essência dos textos. Isso pode ocorrer em função das palavras conectivas e auxiliares (pronomes, preposições, advérbios e outros auxiliares), pois elas não fornecem a informação na expressão do conteúdo dos textos.

No processo de *Stemming* é realizada a extração de cada palavra contida no texto, a característica nesse processo é considerar aquela palavra isolada e tentar reduzir a sua provável pela raiz.

Uso de Dicionário ou *Thesaurus* é um dicionário composto de um vocabulário de representação de sinônimos, hierarquias e relacionamentos associativos entre os termos que possibilitam aos usuários encontrar facilmente as informações que desejam.

5.1.4 Processamento dos Dados

Na etapa do processamento dos dados o objetivo do processo de mineração de textos, é realizar a definição das tarefas para sua execução, assim executando as tarefas mais relevantes do processo. No processo de extração do conhecimento existem várias tarefas que podem ser utilizadas, algumas tarefas usadas pela mineração de textos são descritas a seguir (REZENDE, 2005).

1) Indexação

A indexação consiste na procura eficiente em palavras chaves nos textos na busca de documentos que possam ser relevantes.

Existem dois tipos tradicionais de indexação no processo de mineração a indexação por *tags* e indexação semântica latente.

A indexação por *tags* são palavras-chaves que são extraídas com base nas *tags*. Já a indexação semântica latente são relacionamentos entre as palavras podendo ser deduzidos nos padrões da ocorrência em documentos.

2) Sumarização

A tarefa de sumarização caracteriza em selecionar as informações mais importantes de um texto, tornando compacta a sua descrição. Essa tarefa é uma das mais importantes e utilizadas em mineração de textos, com objetivo de reduzir e resumir nos documentos as palavras e frases mais importantes, mas mantendo seus significados-chave (BARION E LAGO, 2008).

Existem dois tipos específicos de sumarização são elas: sumarização por abstração e sumarização por extração.

Na sumarização por abstração tenta trabalhar da mesma maneira que um ser humano resume, enfatizando as principais idéias do texto original. E na sumarização por extração já é baseada na importância das palavras em um texto.

3) *Clustering*

É a tarefa que utiliza técnicas quando se deseja agrupar documentos similares. Os documentos similares têm a finalidade para facilitar na sua procura, tornando-se mais fácil para usuários finais.

Rezende (2005) apresenta na Figura 5 um esquema de ferramenta em *clustering* sobre um conjunto de documentos, é feita a separação em subgrupos de documentos chamados *clusters*.

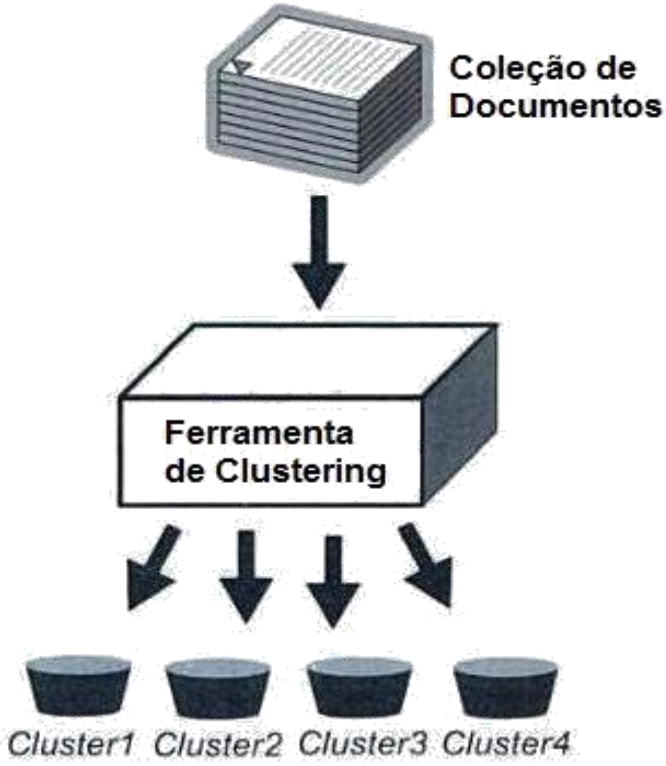


Figura 5: Processo básico de clustering (REZENDE, 2005)

4) Categorização

A categorização pode também ser chamada de classificação que é definida como um processo que realiza a identificação de principais tópicos em um documento, e suas associações é baseado em algoritmos pré-definido. O processo é construído a partir de um grupo de pessoas treinadas com experiência em um determinado assunto.

Segundo Barion e Lago (2008) o algoritmo analisa todos os possíveis exemplos de documentos, realiza o treinamento e aprende com suas regras, após essa etapa é feita o armazenamento em uma base de conhecimento. No

categorizador passam documentos que precisam ser classificados, baseados em regras inseridas previamente na base de conhecimento, com isso estabelecendo qual classe de cada documento, como demonstrada na Figura 6.

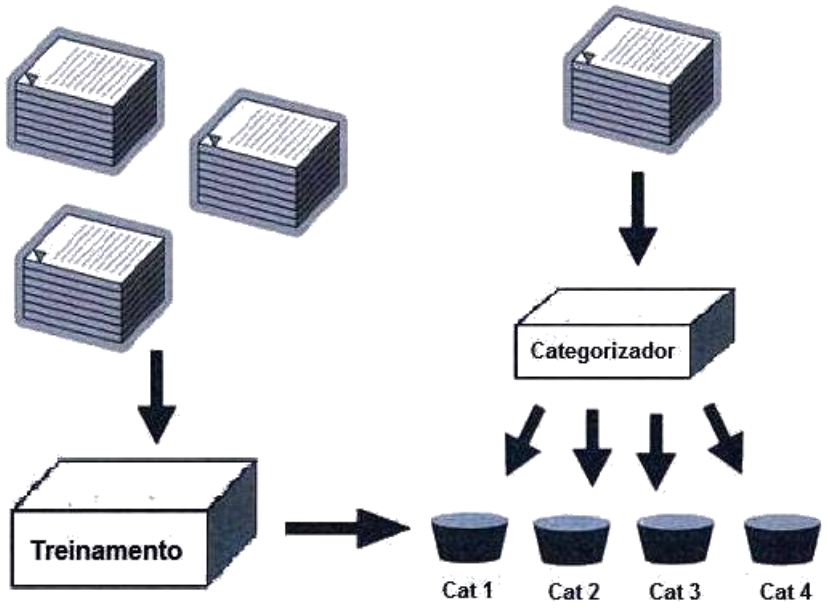


Figura 6: Processo de categorização (REZENDE, 2005)

5.1.5 Pós-Processamento dos Dados

A fase de Pós-Processamento consiste em validar as descobertas realizadas na etapa de processamento dos dados e a visualização dos resultados do processo de mineração de textos encontrados. São existentes métricas de avaliação de resultados que ajudam a consolidar o resultado (REZENDE, 2005).

As métricas de avaliação de resultados básicas que existem são: *recall* e *precision*.

- *Recall* (Abrangência ou revocação)

A *Recall* tem objetivo de fazer a medição da habilidade do sistema na recuperação dos documentos que são mais relevantes para o usuário, a fórmula *Recall* consiste em:

$$Recall = \frac{n\text{-}recuperados\text{-}relevantes}{n\text{-}possiveis\text{-}relevantes}$$, onde:

n-recuperados- relevantes = número de itens relevantes e recuperados

n-possiveis-relevantes = é o total de documentos relevantes.Essa informação só é obtida estatisticamente.

- *Precision* (Precisão)

Na precisão é feita a medição na capacidade do sistema de não retornar documentos não-relevantes nas consultas dos usuários, utiliza-se a seguinte fórmula:

$$precision = \frac{n\text{-}recuperados\text{-}relevantes}{n\text{-}total\text{-}recuperados}$$, onde:

n-recuperados-relevantes: número de documentos relevantes recuperados

n-total-recuperados: total de documentos que possui o sistema.

5.2 ALGORITMOS DE CLASSIFICAÇÃO

A tarefa mais comum em mineração é a classificação de textos ou dados através dos algoritmos, o objetivo da classificação é classificar textos ou dados em classes já determinadas. Existem vários algoritmos para esse tipo de tarefa de classificação, no desenvolvimento do trabalho será apresentados e analisados o grau de precisão de três importantes algoritmos que são: Árvore de Decisão, *K-NN* e *Naive Bayes*.

- Árvore de Decisão: O algoritmo eficiente do tipo aprendizagem de máquina para classificação dos documentos em uma representação baseado em árvores. Pode ser utilizadas para escrever praticamente qualquer tipo de função (COELHO, 2008);

- *K-NN*: Este algoritmo é conhecido como preguiçoso, é um método para classificar objetos baseados nos exemplos ou nos mais próximos (vizinhos) do treinamento. A vantagem de utilização do algoritmo é seu entendimento fácil, intuitivas, fácil de interpretar,

capacidade de generalizar problemas reais, robusto quando apresenta erros no conjunto de treinamento (PARVIN, 2008); e

- *Naive Bayes*: O algoritmo *Naive Bayes* é um dos mais simples classificadores probabilísticos. Usa um conjunto de probabilidades estimadas pela sua freqüência com as características das instâncias dos dados de treino. A vantagem de utilização desse algoritmo é sua estrutura simplificada (MAIA, 2005).

6 DESENVOLVIMENTO

No desenvolvimento inicial do trabalho, foi modelado um mapa conceitual na ferramenta *Cmap Tools*[3] para representação e organização do conhecimento para criação da ontologia na ferramenta Protégé[4], após a criação da ontologia foram construídas as páginas e o povoamento da ferramenta MediaWiki semântica com as classes devidamente criadas e na finalização do trabalho será aplicado os algoritmos de classificação da ferramenta *Rapidminer*[5] nas bases textuais da MediaWiki, possibilitando organizar o conhecimento gerado e classificá-los de acordo com o melhor algoritmo de classificação.

6.1 MODELAGEM MAPA CONCEITUAL

A modelagem do mapa conceitual realizada no trabalho tem por objetivo a organização e representação de maneira clara do conhecimento na criação da ontologia. No mapa conceitual os conceitos são representados por caixa de textos, as relações dos conceitos são interligadas por setas ou linhas ligadas por palavras de ligação, por exemplo, “tem”, “pode ser”, “pode ter” ou “é autor de” como pode ser visto na Figura 7 o mapa conceitual de biblioteca construído.

[3] Disponível em http://cmap.ihmc.us/

[4] Disponível em http://protege.stanford.edu

[5] Disponível em http://rapid-i.com/content/view/26/201/

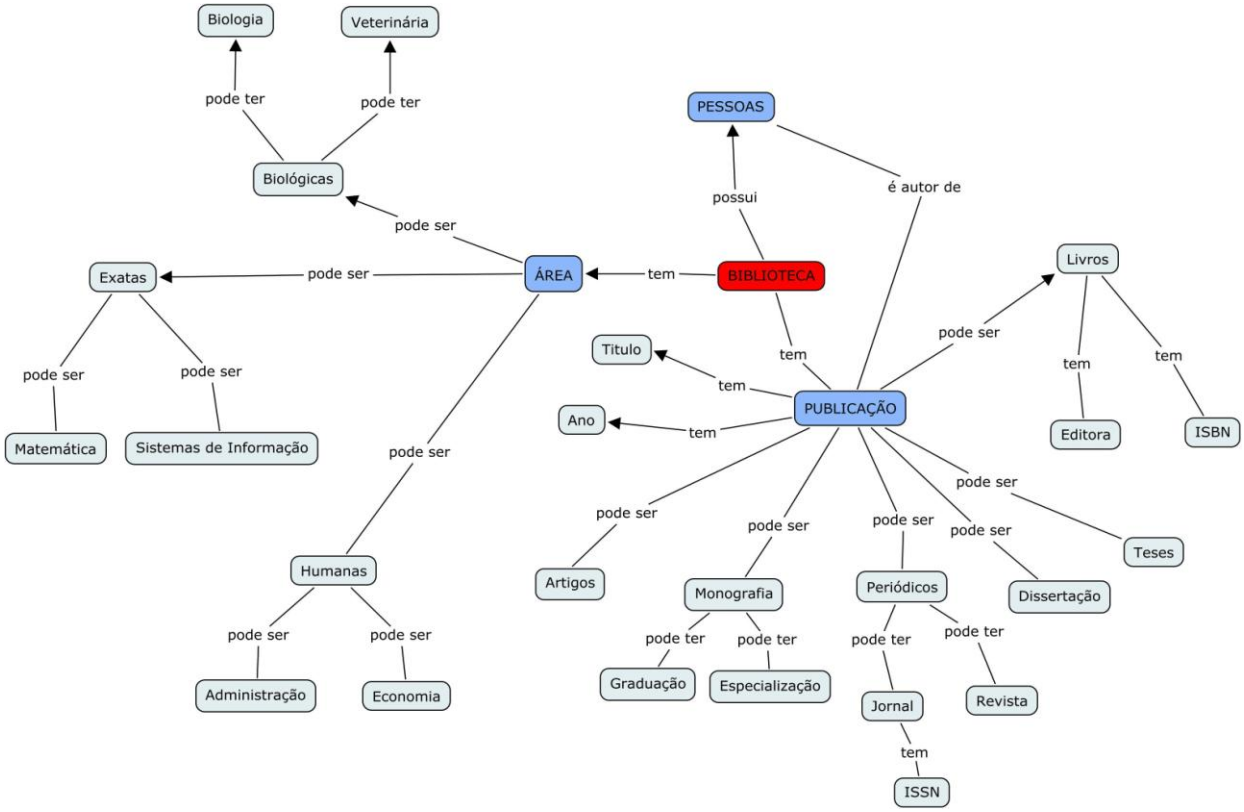


Figura 7: Mapa Conceitual de Biblioteca

6.2 CRIAÇÃO DA ONTOLOGIA

A ontologia criada foi baseada no mapa conceitual apresentado acima, a sua criação foi realizada na ferramenta Protégé que é uma ferramenta de editor de ontologias gratuita que utiliza a linguagem OWL (*Web Ontology Language*).

A criação da ontologia foi baseada em exemplos de ontologias encontradas no site Swoogle-semantic web search[6], a criação da ontologia não teve nenhum tipo de validação por algum especialista. No entanto, a ontologia criada chama-se biblioteca, que tem como objetivo de gerenciar todas as publicações e suas respectivas áreas existentes.

6.3 CONSTRUÇÃO DA ONTOLOGIA NA FERRAMENTA PROTÉGÉ

Para construção da ontologia na ferramenta Protégé primeiramente deve-se criar um novo Projeto (*New Project*) e escolher um determinado tipo de projeto, foi escolhido o tipo de projeto Arquivo OWL/RDF (*OWL/RDF Files*) que será um projeto

[6] Disponível em http://swoogle.umbc.edu/

padrão para a realização do trabalho após isso aparecerá uma janela para finalizar clique em OK.

A seguir é mostrada a tela inicial de criação de uma ontologia na ferramenta Protege que é apresentado na Figura 8.

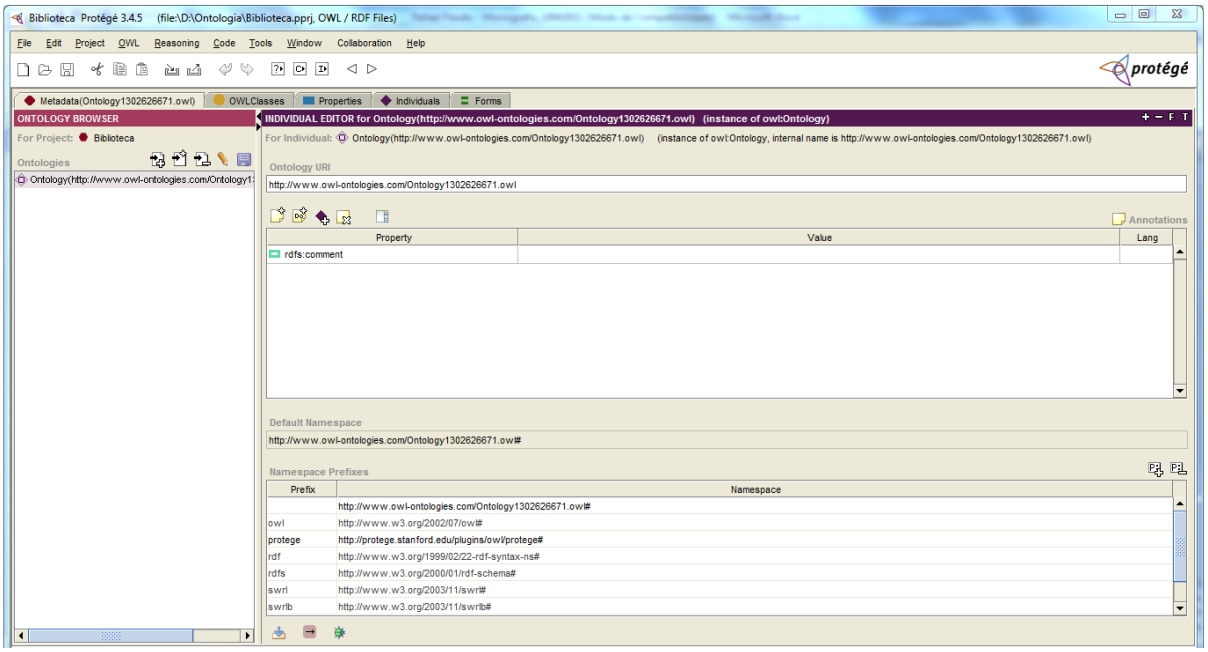


Figura 8: Tela inicial da ferramenta Protege

6.3.1 Criação das Classes

Uma das atividades mais importante da ferramenta ao iniciar um projeto é a criação das classes (*Create Class*), pois as classes são os objetos de um domínio. A criação das classes é apresentada em uma hierarquia contendo as classes e suas respectivas subclasses.

Como pode ser visto na Figura 9 são apresentadas as classes criadas na ontologia biblioteca que são: Pessoa, Publicações e Área. A classe selecionada na figura owl: Thing significa uma classe maior, pois todas as classes são subclasses de owl: Thing.

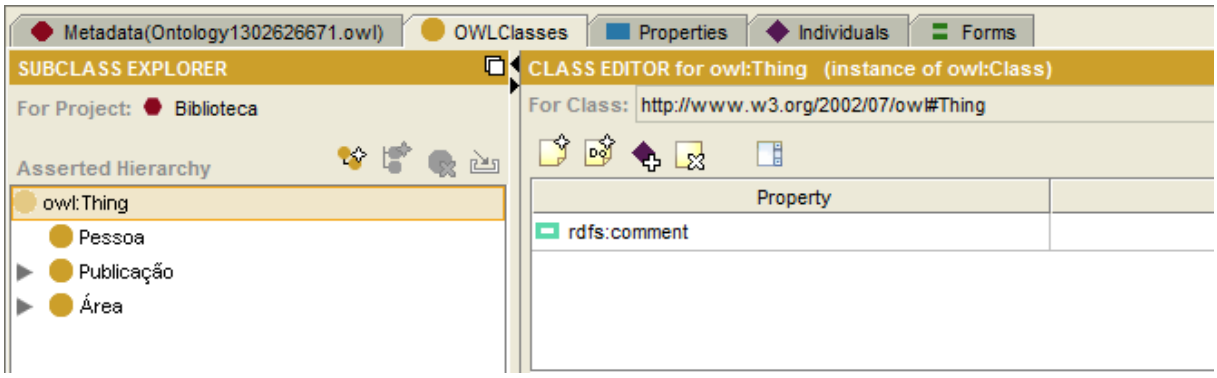


Figura 9: Criação das classes

6.3.2 Criação das Subclasses

Criadas as classes foi preciso criar subclasses (*Create Subclass*) da Ontologia Biblioteca, as classes que possuem as principais subclasses são Publicação e Área. A Figura 10 é apresentada as subclasses da ontologia.

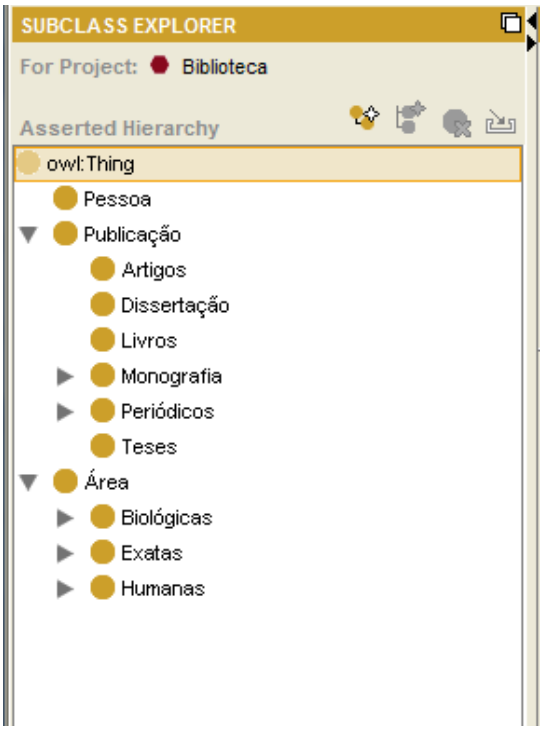


Figura 10: Criação das subclasses

A classe Publicação contém as subclasses Artigos, Dissertação, Livros, Monografia, Periódicos e Teses. E a classe Área possui suas subclasses de Biológicas, Exatas e Humanas.

Para as classes Monografia e Periódicos que são subclasses de Publicação foram criadas suas respectivas subclasses. Para classe Monografia foram criadas as

subclasses Graduação e Especialização, já a classe Periódicos apresenta as subclasses Jornal e Revista.

Na classe Área foram criadas as subclasses Biológicas, Exatas e Humanas com seus cursos, por exemplo, na classe Biológicas contendo subclasses dos cursos de Biologia e Veterinária, a classe Exatas contendo subclasses dos cursos Sistemas de Informação e Matemática e por fim a classe Humanas contendo subclasses Administração e Economia. Na Figura 11 são apresentadas todas as classes e subclasses criadas da ontologia de biblioteca.

Figura 11: Hierarquia das classes e subclasses completas da ontologia biblioteca

6.3.3 Classes Disjuntas

Após a criação de todas as classes da ontologia biblioteca, é necessário dizer que estas classes são disjuntas, pois cada classe não poderá ter indivíduos ou

objetos instanciados entres as outras classes. Para realizar quais classes serão disjuntas usa-se a forma gráfica Disjunção (*Disjoints*).

Para criar as classes disjuntas é necessário selecionar a classe que se deseja tornarem-se disjuntas e Clicar no botão Adicione todas as irmãs (*Add all siblings*) na caixa de dialogo de disjunção.

A Figura 12 demonstra um exemplo das classes disjuntas da ontologia de biblioteca, por exemplo, a classe selecionada Publicação é disjunta das classes Pessoa e Área e assim sucessivamente para as outras classes criadas.

Figura 12: Tela exemplo de disjunção

6.3.4 Criação das Propriedades

As propriedades da ontologia biblioteca foram criadas para existir ligações entre um indivíduo a outro indivíduo, as propriedades são criadas na aba Propriedades (*Properties*).

A Figura 13 é apresentada as propriedades criadas na ontologia biblioteca.

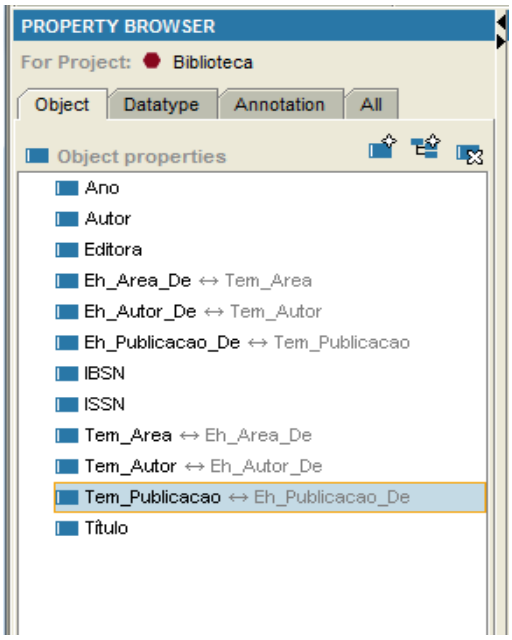


Figura 13: Criação das propriedades

As propriedades podem assumir quatro tipos de características que são: Propriedade Funcional, Inversa Funcional, Simétrica e Transitiva.

- Propriedade Funcional: Propriedades funcionais são aquelas que contêm um valor único;
- Propriedade Funcional Inversa: A propriedade funcional inversa significa que sua inversa é funcional, a propriedade inversa é automaticamente criada pela ferramenta quando se tem uma propriedade funcional;
- Propriedade Simétrica: A propriedade simétrica tem sua característica de relacionar um indivíduo A com um indivíduo B de forma que os dois indivíduos estejam relacionados entre si, ou seja, A relacionando com B e B relacionado com A.
- Propriedade Transitiva: A propriedade transitiva relaciona um indivíduo A com um B, e possibilitando que um indivíduo C relacione com os indivíduos B e A.

Uma propriedade contém um domínio de indivíduos que ligam aos indivíduos de um escopo. Para utilizar o domínio e um escopo em uma propriedade é

necessário selecionar uma propriedade e adicionar a classe que será o domínio e o escopo de uma propriedade. A Figura 14 é visto um exemplo de propriedade criada na ontologia Biblioteca utilizando domínio e um escopo.

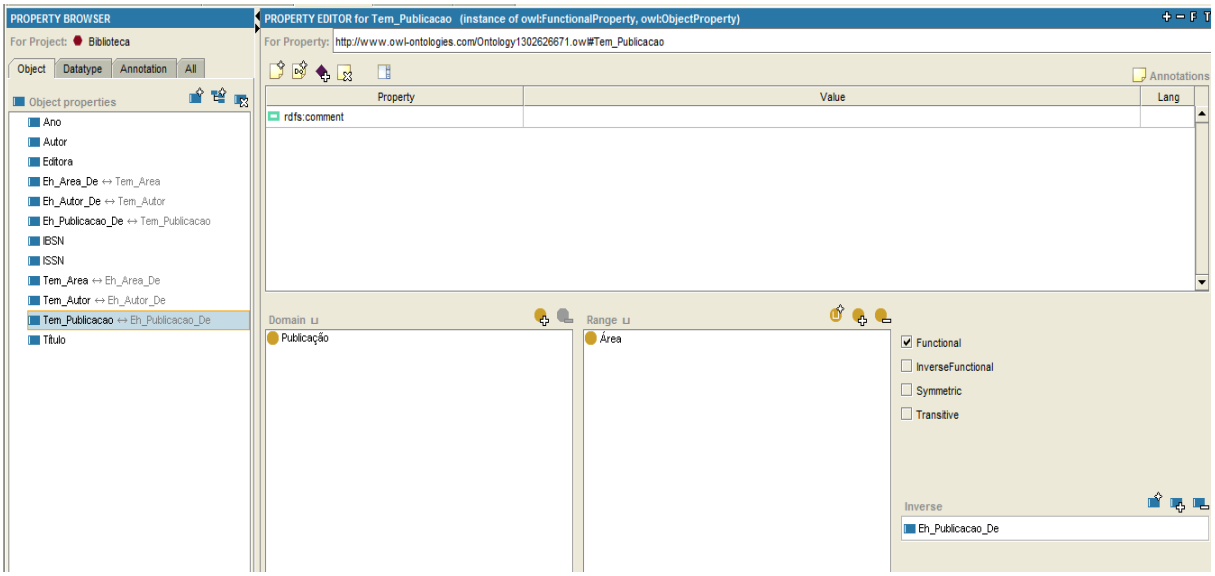


Figura 14: Característica, domínio e escopo de uma propriedade

A Figura 14 representa a característica de uma propriedade com seu domínio e escopo determinado, por exemplo, na ontologia biblioteca a propriedade *Tem_Publicacao* ligam os indivíduos pertencentes ao domínio Publicação com os indivíduos do escopo Área. A propriedade funcional foi determinada por se tratar que uma Publicação deve ser única para uma determinada Área.

6.3.5 Restrição das Propriedades

Depois de criadas as propriedades são necessárias criar algumas restrições, ou seja, as restrições são determinadas para restringir indivíduos de uma classe. Os tipos de restrições aplicadas na ontologia biblioteca foram a Restrição de Quantificador (*Quantifier Restrictions*) e a Restrição de Cardinalidade (*Cardinality Restrictions*).

A restrição de quantificador pode existir de dois tipos que são: Quantificador Existencial e Quantificador Universal.

- Quantificador Existencial: Pode ser lido “pelo menos um” ou “vários”;
- Quantificador Universal: Pode ser lido como “somente”.

Já a restrição de cardinalidade pode-se encontrar os valores de cardinalidade mínima, cardinalidade máxima e cardinalidade exata.

- Cardinalidade Mínima: Especifica que um indivíduo tem “pelo menos” uma quantidade mínima de relação para uma dada propriedade;
- Cardinalidade Máxima: Especifica que um indivíduo tem “no máximo” uma quantidade máxima de relação para uma dada propriedade;
- Cardinalidade Exata: Determina que um indivíduo tenha uma cardinalidade exata para uma propriedade.

Para criar as restrições selecione o botão criar restrição (*Create Restriction*) da caixa de interface de restrições, após isso abrirá uma caixa de Diálogo para selecionar o tipo de restrição que se deseja criar, na caixa de diálogo pode-se selecionar qual propriedade que será feita a restrição o tipo de quantificador e o tipo de condição (*Filler*) que deseja colocar na restrição.

A Figura 15 apresenta a primeira restrição criada na ontologia biblioteca de Quantificador Existencial da Classe Pessoa, que descreve um conjunto de indivíduos que tem pelo menos um Autor, e esse autor é um indivíduo da classe Pessoa.

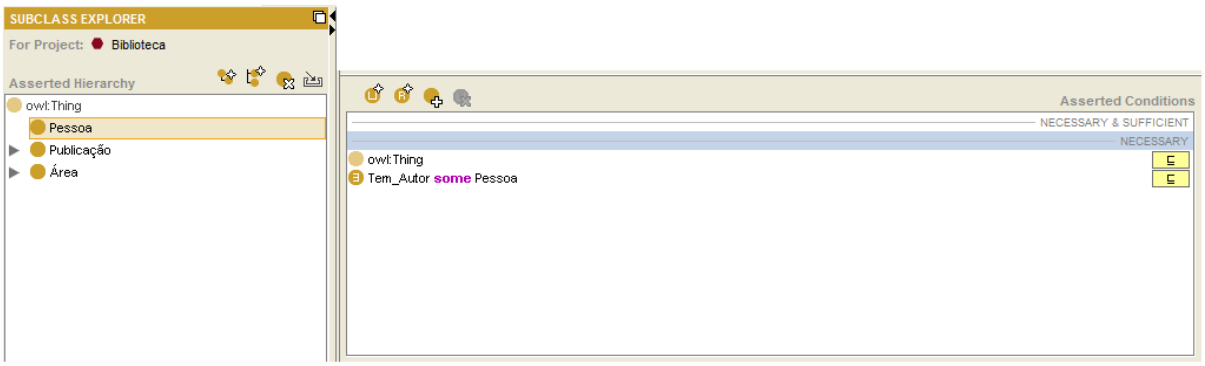


Figura 15: Restrição da classe pessoa

A segunda restrição criada envolve o quantificador existencial e as restrições de cardinalidade mínima e exata.

Na primeira restrição de quantificador existencial criada para a Classe Publicação, significa que a restrição descreve um conjunto de indivíduos que tem pelo menos um Autor, e esse autor é um indivíduo da classe Publicação. A segunda

representa um conjunto de indivíduos que tem pelo menos um Autor, e esse autor é um indivíduo da classe Área.

Já na restrição de cardinalidade apresenta dois tipos de restrição de cardinalidade mínima, que possibilita que uma classe Publicação deve existir no mínimo um Autor e um Ano. E por fim a cardinalidade exata determina que a classe Publicação deve constar exatamente um Título. A Figura 16 são demonstradas as restrições da classe Publicação criadas.

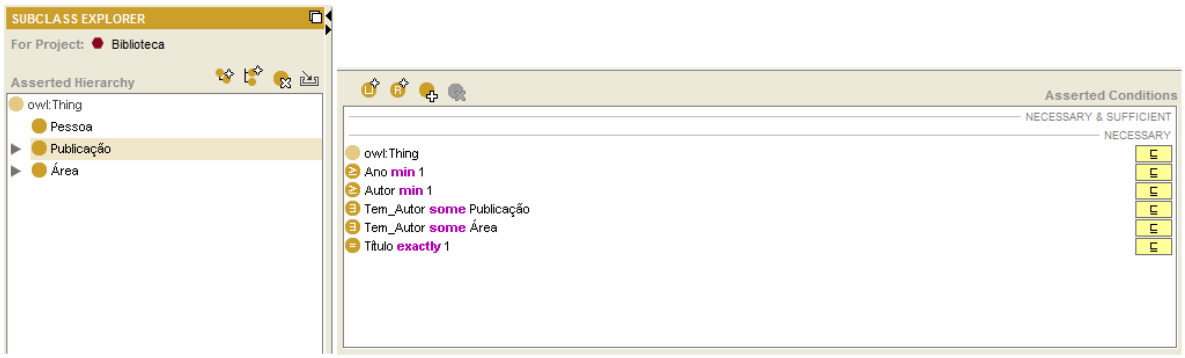


Figura 16: Restrição da classe publicação

A terceira e última restrição criada da Classe Área, descreve um conjunto de indivíduos que tem pelo menos uma Área, e essa Área contém um indivíduo da classe Pessoa. O mesmo para a outra restrição descreve um conjunto de indivíduos que tem pelo menos uma Área, e essa Área contém um indivíduo da classe Publicação. As restrições da Classe Área podem ser visto na Figura 17.

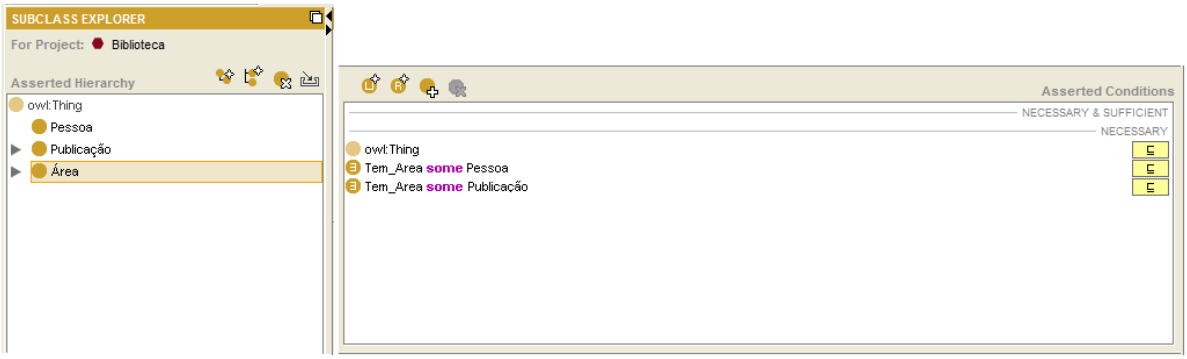


Figura 17: Restrição da classe área

7 AMBIENTE SEMÂNTICO MEDIAWIKI

Neste capítulo é apresentado a criação do ambiente semântico na MediaWiki, sendo demonstradas a criação das páginas e também como trabalhar com as propriedades semânticas.

A criação da MediaWiki foi baseada de acordo com a ontologia criada na ferramenta Protégé, que demonstra as classes devidamente criadas em hierarquia, com isso, as classes criadas na ontologia serão as páginas da MediaWiki.

Antes de criar as páginas da MediaWiki, existem vários procedimentos a serem feitos que são:

- Criação das propriedades;
- Criação de predefinições;
- Criação de categorias; e
- Criação de formulários.

Para criar uma propriedade, predefinição, formulário e uma categoria deve ser instalada uma extensão na MediaWiki chamada Formulários Semânticos[7] (Forms Semantics).

7.1 CRIAÇÃO DAS PROPRIEDADES

Para criação do ambiente semântico do domínio escolhido, devem-se criar as devidas propriedades semânticas para criação das páginas. As propriedades definidas são apresentadas abaixo no Quadro 1:

Quadro 1 Criação das propriedades

Número da Propriedade	Propriedades	Tipo
1	Foi escrito por	Page
2	Foi publicado no ano	Number
3	E do país	String
4	Tem ISBN	Number
5	Tem Editora	String
6	Tem ISSN	Number
7	Tem área	String

[7] Disponível em http://www.mediawiki.org/wiki/Extension:Semantic_Forms

Para criar uma propriedade no ambiente semântico basta clicar em páginas especiais, encontrada no início da ferramenta MediaWiki, e ir para formulários semântico, em seguida clicar em criar uma propriedade, como pode ser visto na Figura 18, a tela de criação de uma propriedade.

Figura 18: Tela de criação de propriedades

A Figura 18 é mostrado um exemplo de uma propriedade criada no ambiente semântico. Para criar uma propriedade basta inserir o nome da propriedade que se deseja trabalhar e seu tipo. No exemplo criado a propriedade possui o nome *Foi escrito por* com o tipo *Page*, no exemplo a propriedade possui o tipo *Page*, pois possibilita que cada autor tenha sua própria página.

7.2 CRIAÇÃO DE PREDEFINIÇÕES

Uma predefinição é um modelo de página, sendo assim para cada classe criada na ontologia terá uma predefinição. As predefinições possuem quatro campos, e para cada campo deve especificar os seus valores que são: o nome do campo, mostrar etiqueta para exibição em cada página, propriedade semântica e agregação que é opcional. A Figura 19 mostra um exemplo de predefinição criada no ambiente semântico.

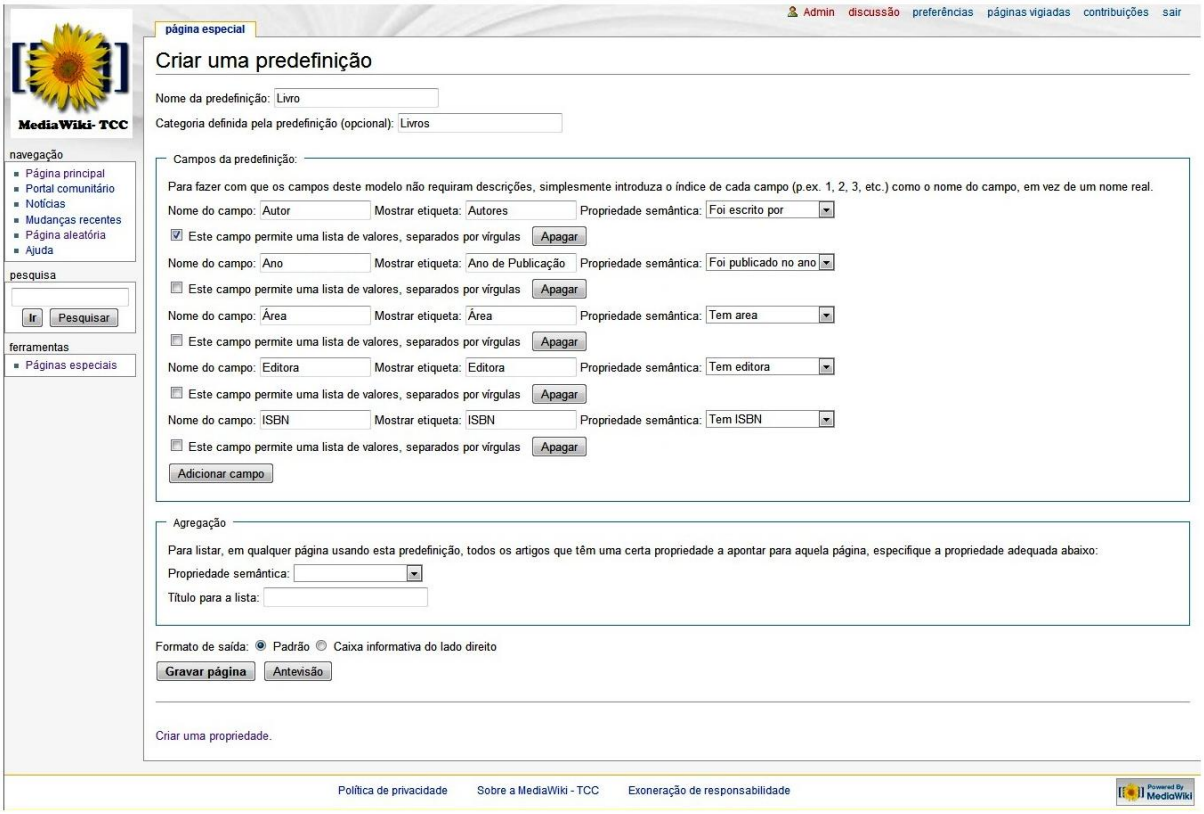


Figura 19: Exemplo de predefinição

O exemplo mostrado na Figura 19 é uma predefinição de livro que mostra todos os nomes do campo, etiqueta e a propriedade semântica preenchidos.

No primeiro campo da predefinição, foi definido o nome do campo como: autor, no mostrar etiqueta como: autores e na propriedade semântica como: *Foi escrito por*, esse campo tem objetivo de mostrar o autor da publicação de livro. No primeiro campo autor foi selecionado o campo que permite uma lista de valores separados por vírgulas, pois uma vez que o livro pode ter mais de um autor.

No segundo campo foi definido o nome do campo como: Ano de publicação, no mostrar etiqueta como: Ano de publicação e a propriedade semântica como: *Foi publicado no ano*, esse campo tem objetivo de mostrar o ano da publicação.

No terceiro campo foi definido o nome do campo como: área, no mostrar etiqueta como: área e a propriedade semântica como: *Tem área*, esse campo tem objetivo de mostrar a área da publicação.

No quarto campo foi definido o nome do campo como: editora, no mostrar etiqueta como: editora e a propriedade semântica como: *Tem editora*, esse campo tem objetivo de mostrar à editora que a publicação livro possui.

No quinto campo foi definido o nome do campo como: ISBN, no mostrar etiqueta como: ISBN e a propriedade semântica como: *Tem ISBN*, esse campo só existe para publicação livro, pois ISBN é um sistema internacional padronizado de numeração existente para identificação de livros.

Todas as predefinições criadas no ambiente semântico são mostrados no Quadro 2.

Quadro 2 Criação das predefinições

Número da Predefinição	**Nome do Campo**	**Etiqueta**	**Propriedade Semântica**
Nome da Predefinição: Livro			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
4	Editora	Editora	Tem Editora
5	ISBN	ISBN	Tem ISBN
Nome da Predefinição: Autores			
1	País	País de Origem	É do país
2	Área	Área	Tem área
Nome da Predefinição: Dissertação			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
Nome da Predefinição: Artigos			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
Nome da Predefinição: Monografia – Especialização			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
Nome da Predefinição: Monografia – Graduação			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
Nome da Predefinição: Periódicos Jornal			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
4	ISSN	ISSN	Tem ISSN

Nome da Predefinição: Periódicos Revista			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área
Nome da Predefinição: Teses			
1	Autor	Autores	Foi escrito por
2	Ano	Ano de Publicação	Foi publicado no Ano
3	Área	Área	Tem área

7.3 CRIAÇÃO DE CATEGORIAS

Nessa seção é realizada a construção das categorias, pois cada classe no ambiente semântico deve conter uma categoria para alocação das páginas que forem construídas. A Figura 20 demonstra a criação da categoria livros.

Figura 20: Criação da categoria

A Figura 20 mostra um exemplo da criação de uma categoria livro na MediaWiki. Na tela de criação da categoria basta definir o nome da categoria e o formulário padrão que deseja, no caso da categoria livro o formulário que deve ser escolhido é o formulário livro, pois cada categoria deve ser criada de acordo com o formulário construído para que não ocorram informações ou páginas erradas no ambiente semântico.

A Figura 21 mostra a categoria livros com seus respectivos livros criados.

Figura 21: Criação da categoria livro

7.4 CRIAÇÃO DE FORMULÁRIOS

Para criar as páginas na MediaWiki deve-se criar os formulários para todas as classes de acordo com a ontologia criada. A Figura 22 apresenta a tela de criação dos formulários.

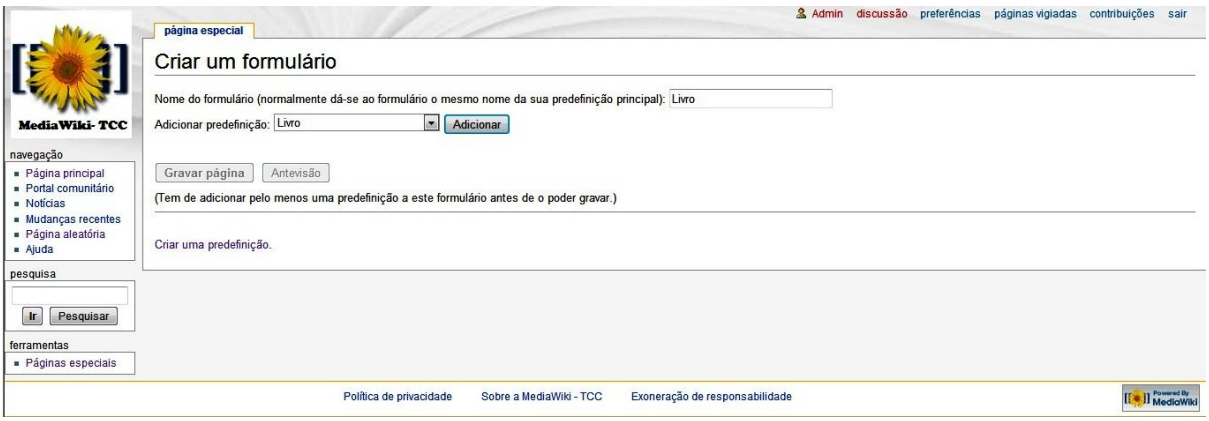


Figura 22: Tela de criação de formulário

A Figura 22 é um exemplo de criação de formulário livro. Para criar o formulário é indicado que o formulário tenha o mesmo nome que a predefinição. Após escolher o nome e selecionar qual predefinição que deseja basta clicar em adicionar, logo após irá abrir uma tela para colocar o título para criação da página, como pode ser visto na Figura 23.

Figura 23: Tela de criação de páginas

Depois dos passos realizados é mostrada a criação da página de publicação livro na Figura 24.

Figura 24: Criação da página livro

Na página criada na Figura 24 apresenta a página de uma publicação livro de genética contendo as predefinições definidas e um resumo de apresentação do livro.

A diferença das páginas na MediaWiki é a forma de organizar o conhecimento através da ontologia criada através das propriedades semânticas. As propriedades semânticas podem ser criadas através das predefinições ou manualmente no corpo do texto, como pode ser visto na Figura 25.

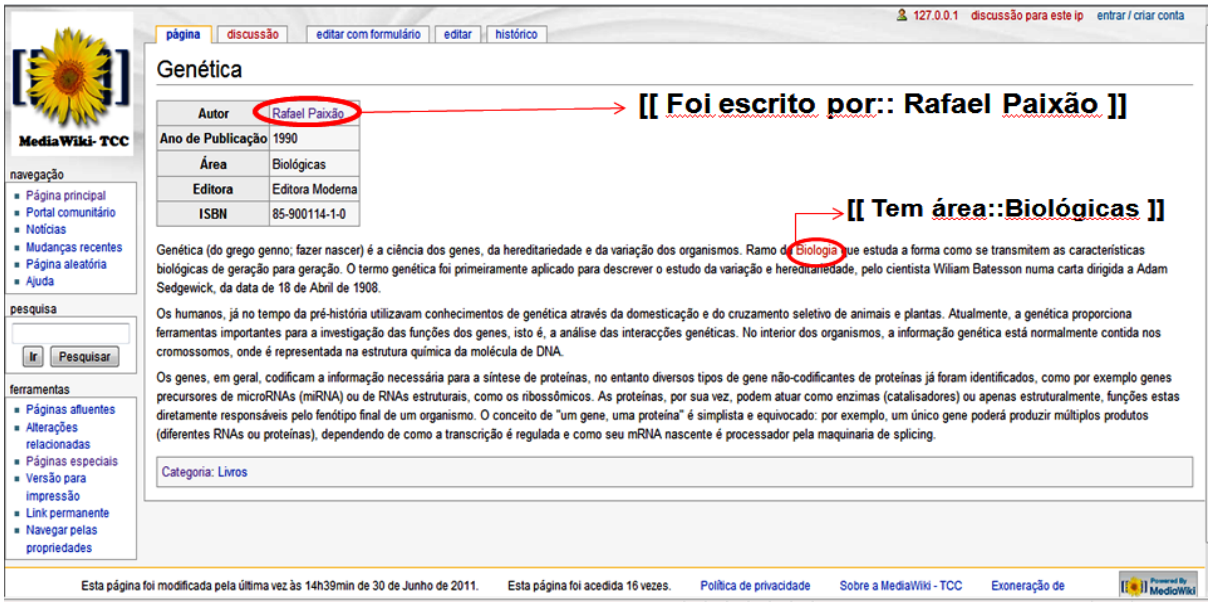


Figura 25 Exemplo de Propriedades Semânticas

A Figura 25 mostra os exemplos de criação das propriedades semânticas, as propriedades são representadas na forma de triplas (sujeito, predicado e objeto). No exemplo apresentado acima demonstra a propriedade para saber o autor da publicação livro, no caso o sujeito é o autor, o predicado é *Foi escrito por* e o objeto indicando o objeto que a relação semântica está apontando no caso Rafael Paixão, esse exemplo aplica-se também a outra propriedade apresentada.

8 APLICAÇÃO DA MINERAÇÃO DE TEXTOS

Para aplicação das técnicas de mineração de textos, foi realizada uma pesquisa para coletar as bases textuais para inserção dos textos na *wiki*. As bases textuais são compostas por resumos de Livros, Artigos, Monografias, Periódicos, Teses e Dissertações, e esses texto foram coletados do site Domínio Público[8], que é uma biblioteca digital desenvolvida em software livre, e também a partir da biblioteca digital Brasileira de Teses e Dissertações[9] (BDTD), do Ministério da Ciência e Tecnologia.

Foram criados dois modelos para mineração de textos, e aplicados algoritmos de mineração de dados. No modelo A, os dados foram organizados em áreas, que são: Biológicas, Exatas e Humanas, o qual cada área apresenta uma pasta heterogênea com 60 textos das devidas áreas. Já no modelo B, as bases textuais foram dividas em subáreas, cada área contém 60 textos divididos em dois cursos totalizando em 180 textos. Como pode ser visto na Figura 26.

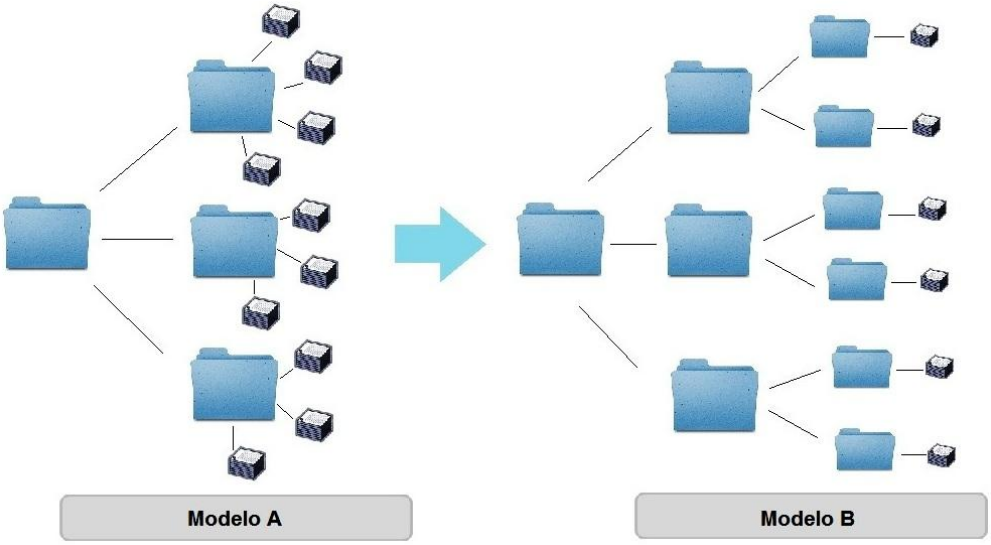


Figura 26: Modelo A e B

8.1 FERRAMENTA: RAPIDMINER

O *Rapidminer* na sua versão 5 é um ambiente para aprendizagem de máquina e mineração de dados desenvolvida em linguagem Java que permite

[8] Disponível em http://www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.jsp

[9] Disponível em http://bdtd.ibict.br/pt/inicio.html

utilizar diversos operadores para explorar os dados. A ferramenta é *open-source* e distribuída sob a licença GNU hospedada pelo *SourceForge* desde 2004.

A ferramenta *Rapidminer* foi escolhida para aplicação das técnicas de mineração de textos nas bases textuais da wiki. No desenvolvimento da mineração foram analisados e escolhidos três algoritmos de classificação, e aplicados aos dois modelos. Os algoritmos escolhidos para desenvolvimento foram: Árvore de decisão, *K-NN* e *Naive Bayes*.

A Figura 27 apresenta a interface da ferramenta *Rapidminer*, nas quais demonstra as funções e seus componentes bem distribuídos para criação dos processos.

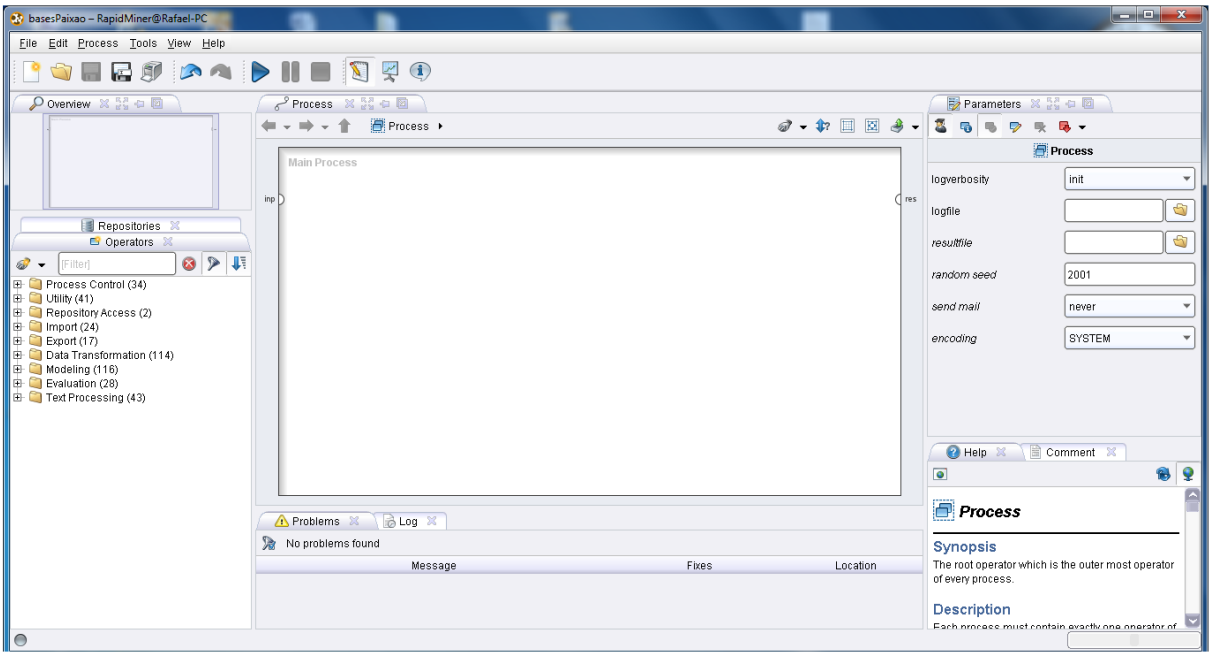


Figura 27: Interface Ferramenta Rapidminer

Para aplicação da mineração de textos foi usado os mesmos procedimentos para os dois modelos, a única diferença são os diretórios para realização de leitura das bases. Primeiramente deve-se realizar as etapas de Pré-Processamento que possibilita uma limpeza e seleção dos dados, já na etapa de Extração do Conhecimento foi utlizados os algorimtos de classificação, e por fim a etapa de Pós-processamento, sendo uma análise de precisão do melhor algoritmo que será apresentado os resultados em um capítulo separadamente.

8.2 PRÉ-PROCESSAMENTO

A pré-processamento é a primeira etapa do processo de mineração de textos, essa fase é importante para limpeza, redução de volume, seleção e na qualidade dos dados no processo de mineração, buscando colocar os dados em um formato adequado para extração de conhecimento.

A fase de pré-processamento passam por várias etapas fundamentais para preparação adequada dos dados. Os principais passos para etapa de seleção dos dados são:

- Leitura das bases textuais;
- Divisao do texto em termos;
- Padronização dos caracteres;
- Remoção dos *stopwords*; e
- Normalização morfológica.

A subções seguintes apresentam o processo na ferramenta Rapidminer.

8.2.1 Leitura da Base Textual

No processo de pré-processamento a primeira etapa a ser executada é a leitura da base textual, na ferramenta Rapidminer para fazer a leitura da base é utilizado o componente (*Process Documents from files).* O componente recebe de entrada as bases textuais coletadas através no diretório especificado, os diretórios dos modelos A e B são apresentados na Figura 28 e Figura 29, demonstrando assim a diferença dos diretórios nos dois modelos. No componente é definido o tipo de vetor de palavras, com isso foi escolhido no componente à freqüência inversa para desenvolvimento das análises do processo.

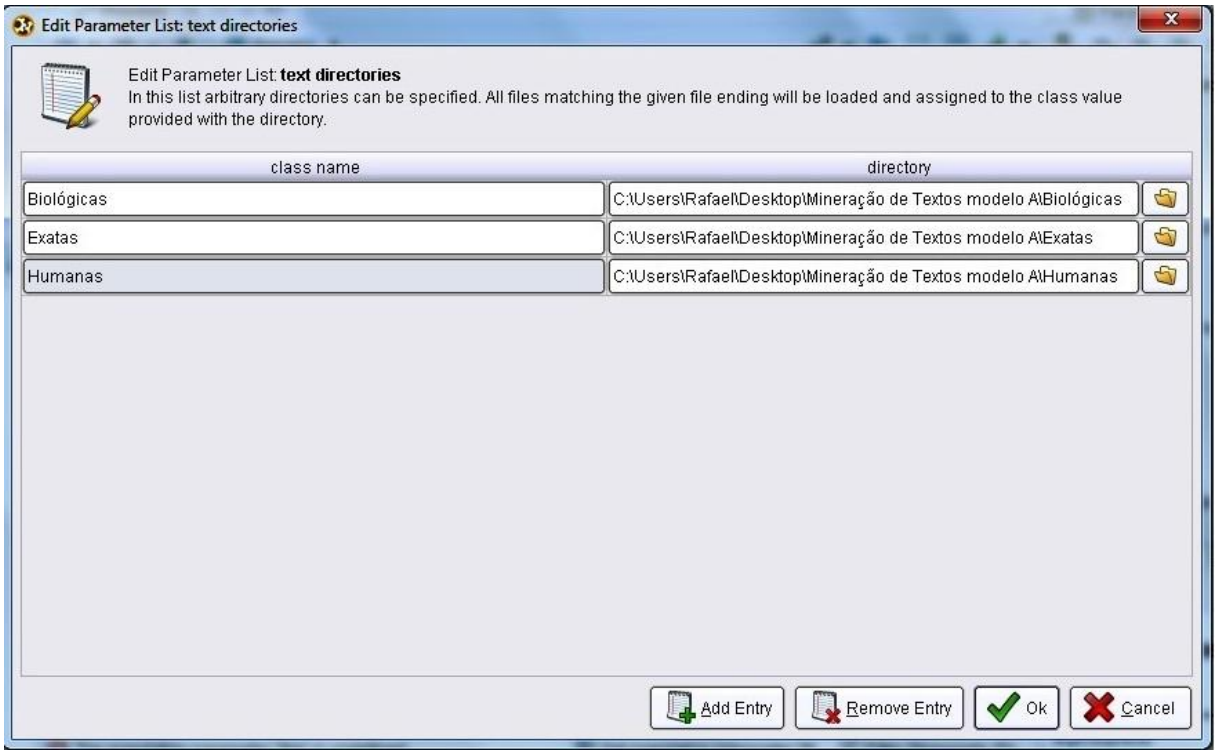


Figura 28: Diretório do Modelo A

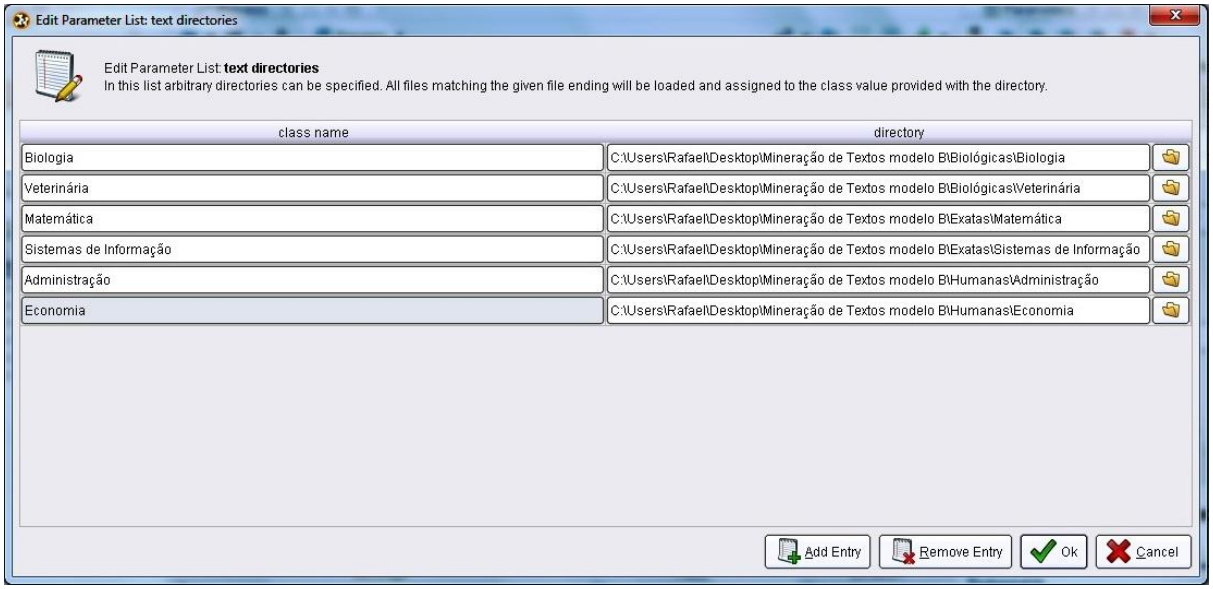


Figura 29: Diretório do Modelo B

A Figura 30 apresenta o componente de leitura da base textual.

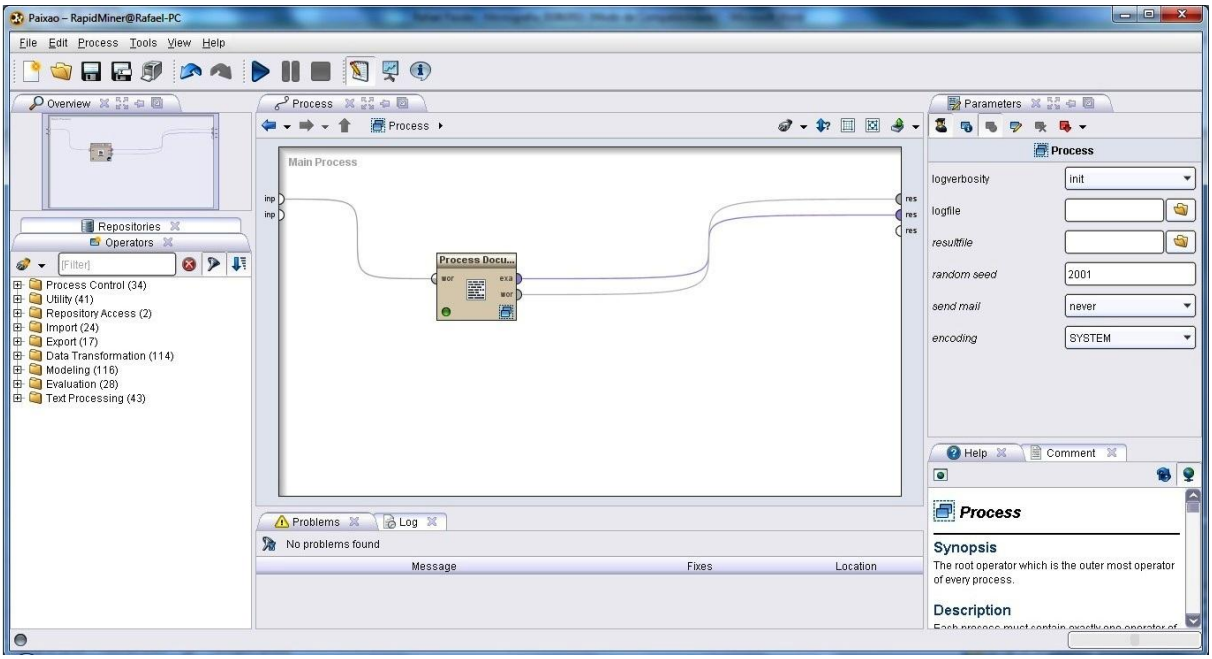


Figura 30: Leitura da Base Textual

As próximas etapas são realizadas no *Vetor Creation* interligadas com o componente (*Process Documents from files).*

8.2.2 Divisão do texto em termos

O próximo componente após a leitura da base para divisão do termo será o *Tokenize.* O componente tem o objetivo de separar as palavras para criação do vetor. No componente é marcada a opção não letra (*non-letters*) que realiza eliminar os espaços nos textos para gerar uma lista de palavras conhecidas como índice.

8.2.3 Padronização dos caracteres

Para padronização dos caracteres será usado o componente *Transform Cases* que tem a finalidade de padronizar os caracteres em maiúsculos ou minúsculos. No componente foi escolhido para transformar os caracteres todos em minúsculos, pois é importante para padronização dos caracteres para remoção dos termos do vetor.

8.2.4 Remoção de Stopwords

O objetivo do *stopwords* é remover palavras de pouca importância ou palavras em comum, como preposições, pronomes, artigos, pontuação e advérbios. O componente utilizado para remoção de *stopwords* é *Filter Stopwords Dictionary* o componente importa uma *stoplist* (conjunto de *stopwords*) que são padrão no

processo de remoção de palavras. Com isso foram eliminadas as palavras que não tem uma grande importância no significado dos textos baseado na comparação com os termos do índice criado anteriormente.

8.2.5 Normalização Morfológica

Esta etapa utiliza o componente *Stem (Snowball)*, no componente é selecionado o atributo na linguagem em português devido que a base textual é em português. O componente tem objetivo de eliminar prefixo e sufixo, características de gênero, número e grau, restando somente o radical de cada palavra.

A Figura 31 mostra todo o processo explicado anteriormente aplicado nos modelos.

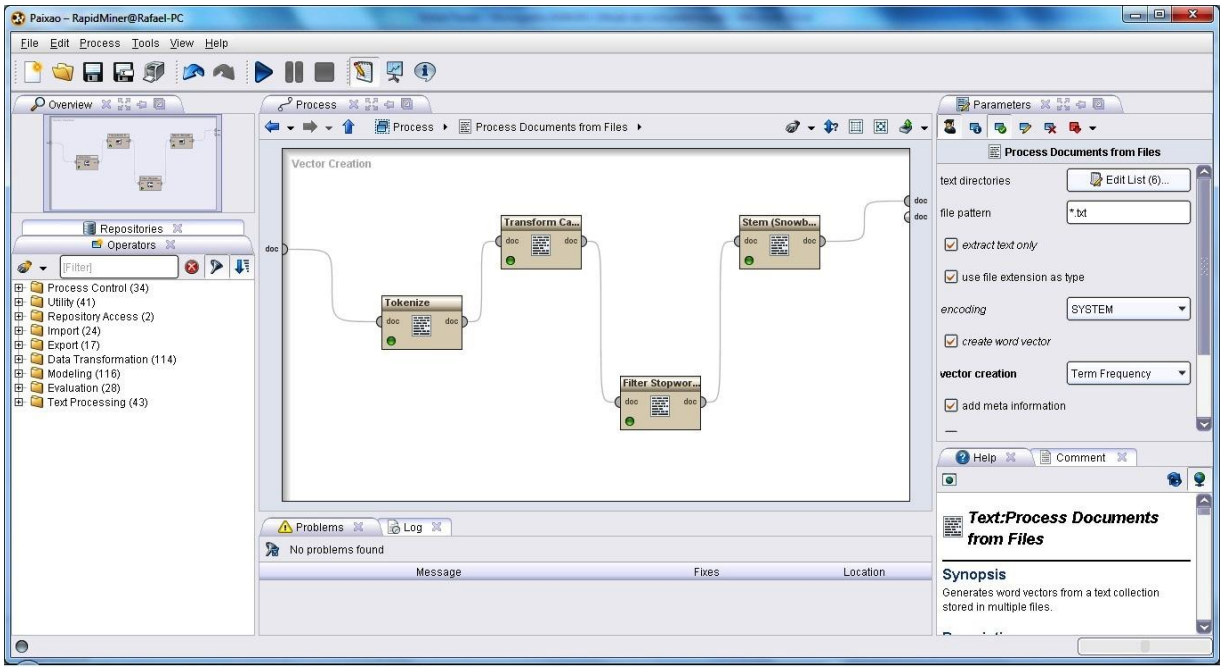


Figura 31: Etapas de limpeza da base

8.3 EXTRAÇÃO DO CONHECIMENTO E VALIDAÇÃO

Após a etapa de limpeza e seleção dos dados é gerado a *ExampleSet*, que mostra a base textual pré-classificada e a *WordList* uma lista das palavras que mais apareceram nas bases textuais e nos textos. A Figura 32 apresenta um exemplo da *ExampleSet*.

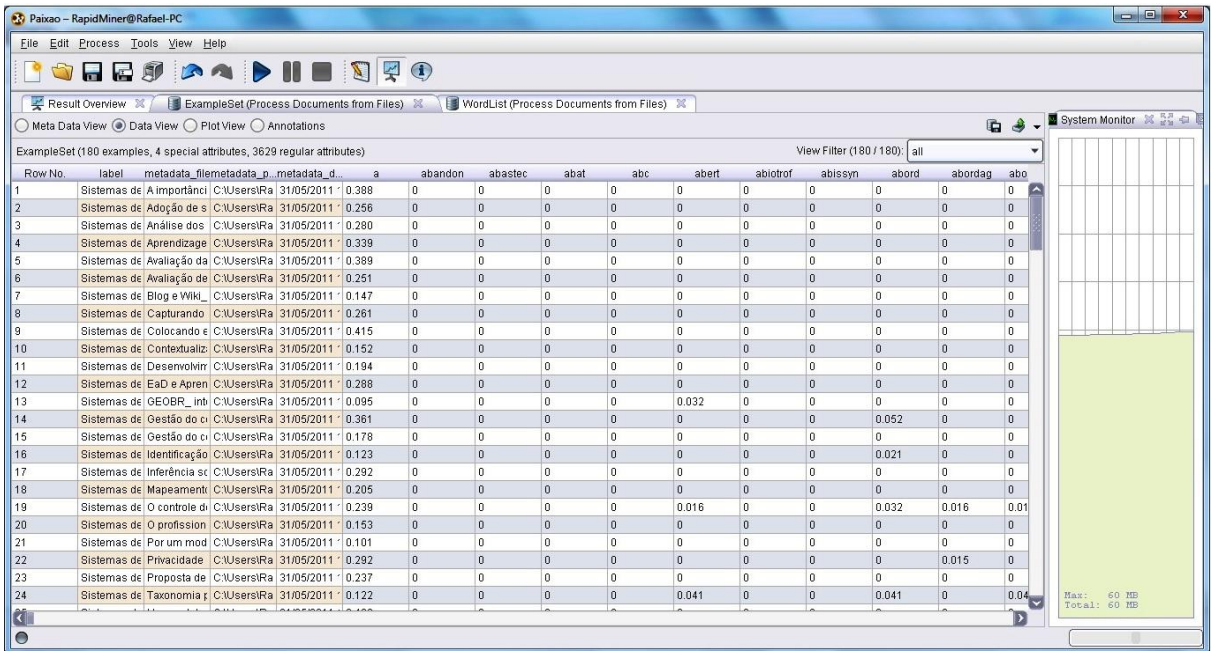
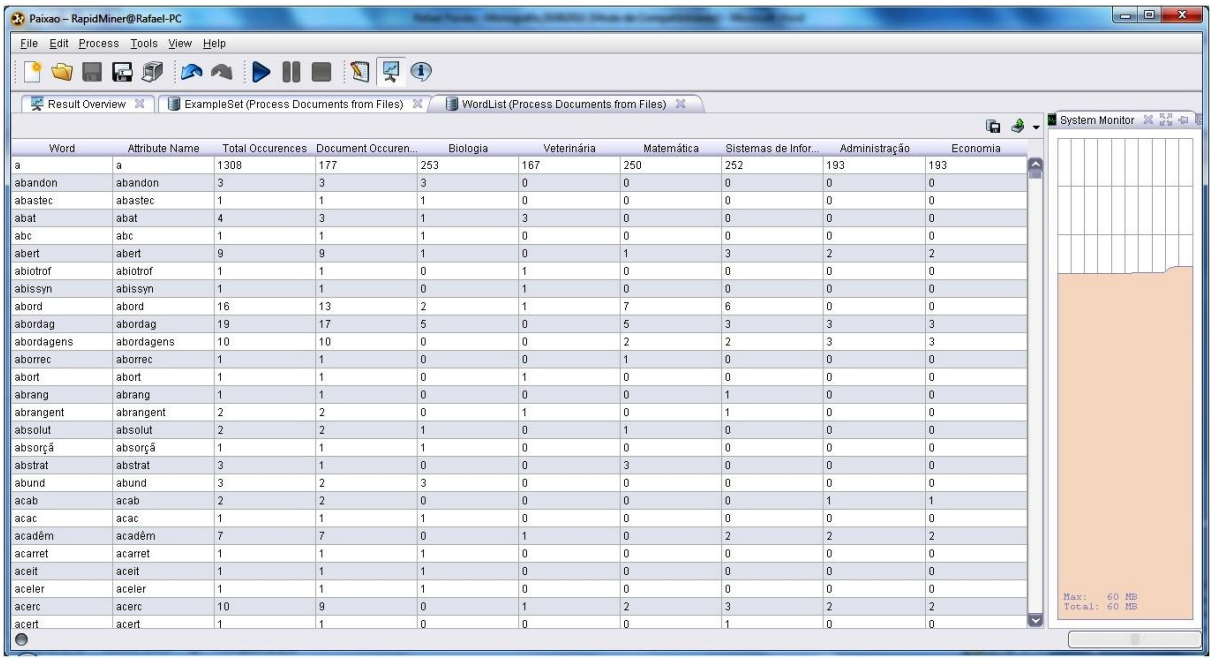


Figura 32: ExampleSet

A *WorList* foi substituída por uma lista *stoplist* (conjunto de *stopwords*) mostrada no Apêndice A. Com isso é gerada a *WordList* contendo as seguintes colunas que são:

- *Word*: São as palavras;
- *Total Occurrences*: Quantidade de vezes que uma determinada palavra foi encontrada na base;
- *Documents Ocurrences*: Quantas vezes as palavras apareceu em um determinado texto; e
- A coluna dos cursos de Biologia, Veterinária, Matemática, Sistemas de Informação, Administração e Economia: Refere-se à quantidade de vezes as palavras apareceram em um determinado curso.

A seguir é apresentado a Figura 33 um exemplo da *WordList.*



Word	Attribute Name	Total Occurences	Document Occuren...	Biologia	Veterinária	Matemática	Sistemas de Infor...	Administração	Economia
a	a	1308	177	253	167	250	252	193	193
abandon	abandon	3	3	3	0	0	0	0	0
abastec	abastec	1	1	1	0	0	0	0	0
abat	abat	4	3	1	3	0	0	0	0
abc	abc	1	1	1	0	0	0	0	0
abert	abert	9	9	1	0	1	3	2	2
abiotrof	abiotrof	1	1	0	1	0	0	0	0
abissyn	abissyn	1	1	0	1	0	0	0	0
abord	abord	16	13	2	1	7	6	0	0
abordag	abordag	19	17	5	0	5	3	3	3
abordagens	abordagens	10	10	0	0	2	2	3	3
aborrec	aborrec	1	1	0	0	1	0	0	0
abort	abort	1	1	0	1	0	0	0	0
abrang	abrang	1	1	0	0	0	1	0	0
abrangent	abrangent	2	2	0	1	0	1	0	0
absolut	absolut	2	2	1	0	1	0	0	0
absorçã	absorçã	1	1	1	0	0	0	0	0
abstrat	abstrat	3	1	0	0	3	0	0	0
abund	abund	3	2	3	0	0	0	0	0
acab	acab	2	2	0	0	0	0	1	1
acac	acac	1	1	1	0	0	0	0	0
acadêm	acadêm	7	7	0	1	0	2	2	2
acarret	acarret	1	1	1	0	0	0	0	0
aceit	aceit	1	1	1	0	0	0	0	0
aceler	aceler	1	1	1	0	0	0	0	0
acerc	acerc	10	9	0	1	2	3	2	2
acert	acert	1	1	0	0	0	1	0	0

Figura 33: WordList

8.3.1 Validação

Para validação das bases textuais será utilizada nos dois modelos a validação *BootStrap*, que é considerada uma das melhores validação para estimar desempenho de um conjunto de dados pequenos.

No processo de validação, *BootStrap* é um método de estimação que usa amostragem com reposição para formar o conjunto de treinamento, ou seja, a validação funciona retirando uma amostra aleatória de tamanho n de um conjunto de n exemplos com reposição, essa amostra é usada para treinamento, e por fim para realização dos testes são usados os exemplos dos dados originais que não estão contidos no conjunto de treino (LOPES, 2003).

No processo é utilizado o componente (*Retrieve)* sendo usado para receber um repositório de entrada no caso a *ExampleSet* gerada na fase de pré-processamento, que liga a outro componente de validação chamado (*Bootstrapping Validation),*o número de vezes de validação que cada algoritmo irá executar é de cinco validações. A Figura 34 apresenta o exemplo de um processo de validação *BootStrap*.

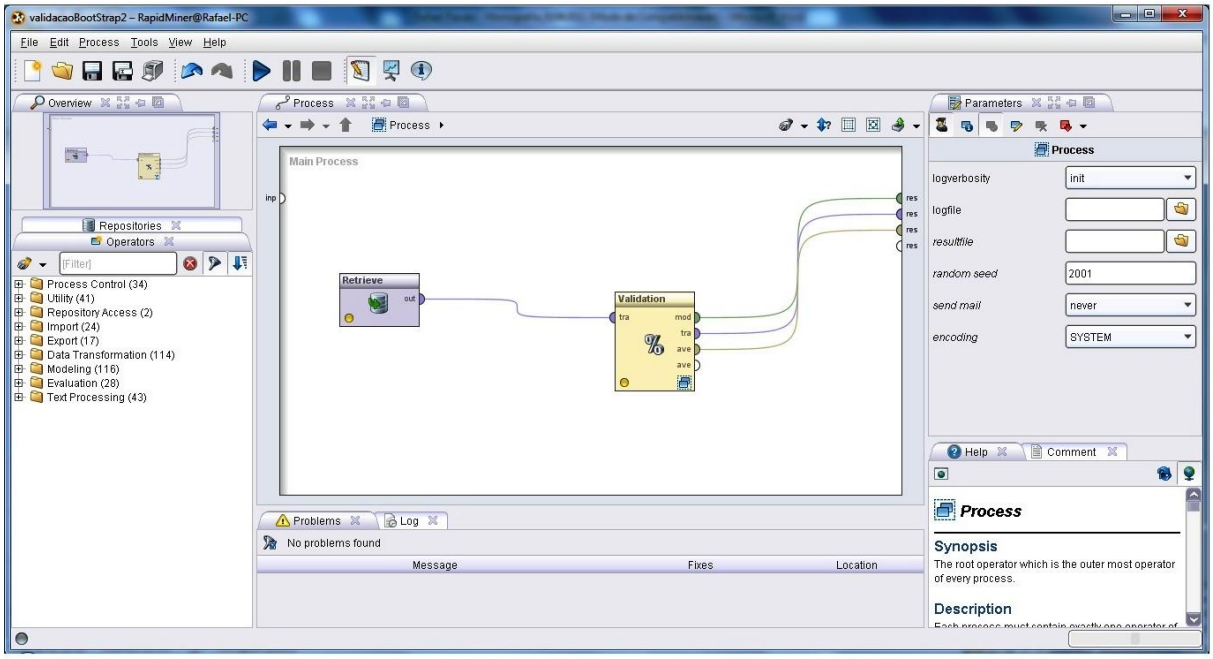


Figura 34: Validação BootStrap

No componente de validação são realizados os testes e o treinamento das bases com os seguintes algoritmos que são: Árvore de decisão, *Naive Bayes* e o *K-NN.*

A Figura 35 mostra um exemplo do processo de validação *BootStrap* com os algoritmos de classificação, sendo assim os processos são iguais para os demais algoritmos. O algoritmo usado para demonstrar o exemplo foi o *K-NN*, na parte de treinamento (*training)* é colocado o algoritmo e no Teste (*testing)* é colocado o componente (*Apply Model),* recebendo uma lista contendo o diretório da base, ligando com o componente de (*Performance),* com isso é finalizado o processo gerando o desempenho de cada algoritmo testado.

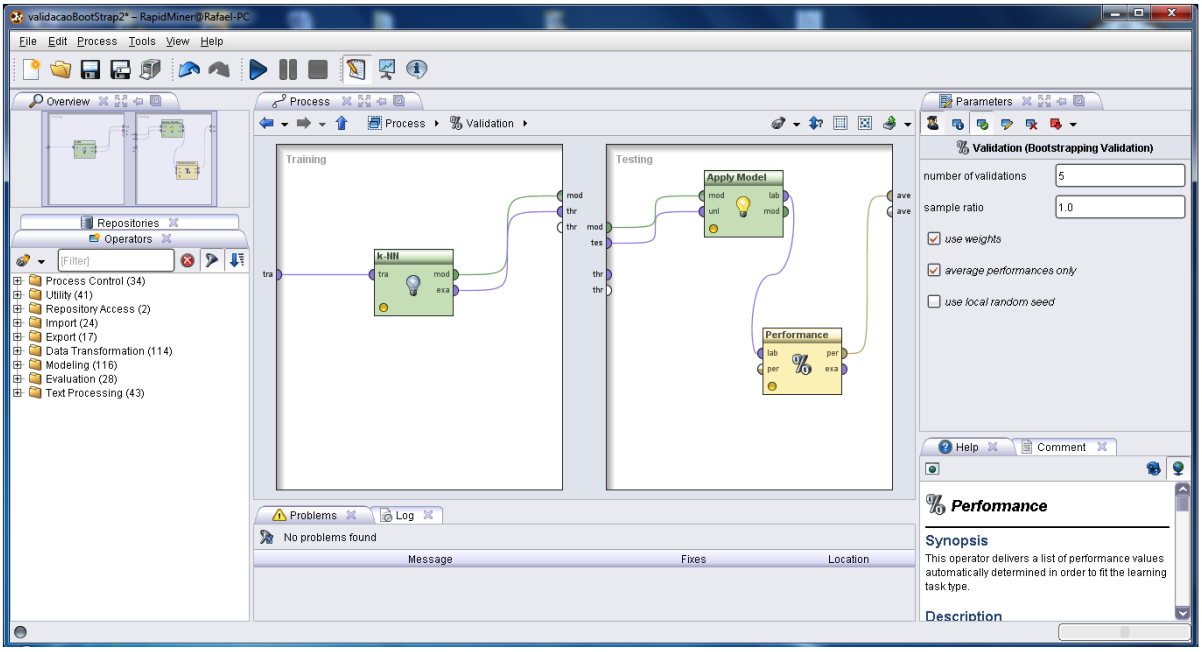


Figura 35: Validação BootStrap com algoritmo K-NN

9 RESULTADOS

Para obtenção de uma organização de conhecimento, as tecnologias *Web* 2.0, no caso a wiki, demonstrou ser uma ferramenta importante que auxilia no processo e ajuda acelerar e disseminar informações, gerando assim um nível de precisão mais eficaz, disponibilizando e compartilhando o conhecimento para toda uma organização, tornando-se assim importante em uma tomada de decisão em um ambiente corporativo.

Na aplicação da mineração de textos nas bases textuais da MediaWiki, os textos foram organizados em dois modelos A e B como demonstrados no capítulo 8 detalhadamente, a validação utilizada no processo foi a *BootStrap,* para realização dos resultados da eficiência de precisão dos algoritmos de classificação.

Para obtenção dos resultados da aplicação da mineração de textos foram realizadas as análises dos dois modelos A e B. A primeira análise foi realizar a eficiência da base com seu grau exatidão (*accuracy*), que determina uma porcentagem de desempenho total do algoritmo, a segunda realizando a porcentagem de precisão de acerto das classes, e por fim o tempo de execução de cada algoritmo. Todas as análises foram geradas pelo processo realizado na ferramenta *Rapidminer*.

A seguir as Figura 36 e Figura 37 apresenta graficamente os resultados de *accuracy* de cada algoritmo dos modelos A e B.

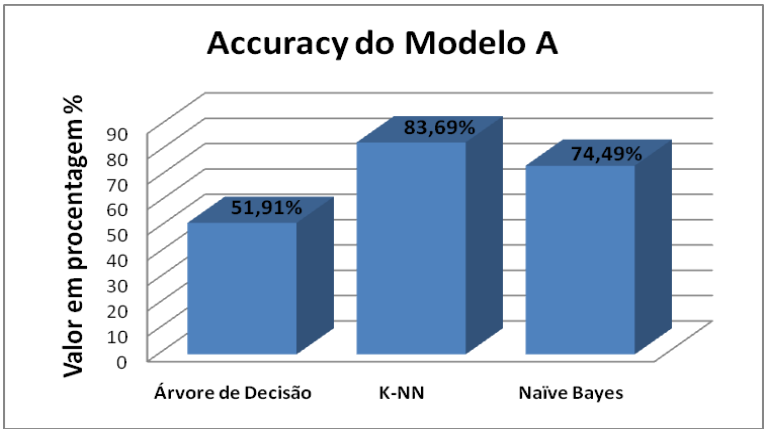


Figura 36: Accuracy do Modelo A

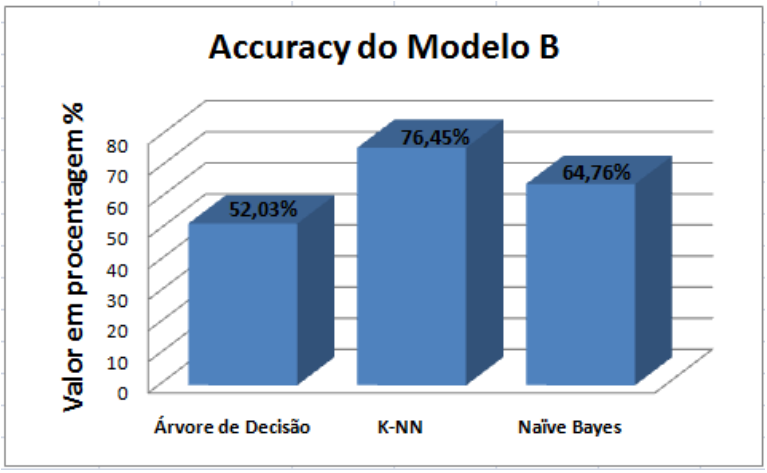


Figura 37: Accuracy do Modelo B

O melhor índice de precisão apresentado com a exatidão (*accuracy*) foi com o algoritmo *K-NN*. Tanto para o modelo A quanto para B. No entanto, pode-se observar que o algoritmo *Naive Bayes*, também apresentou uma eficiência satisfatória na classificação dos textos em áreas. Já o algoritmo de árvore de decisão nos dois modelos demonstrou-se insatisfatório, com o percentual de eficiência baixa, sendo um algoritmo não eficaz para este tipo de classificação.

A segunda análise realizada foi verificar a eficiência dos algoritmos na precisão de acertos das classes. A seguir é apresentado no Quadro 3 os resultados de precisão do modelo A.

Quadro 3: Precisão das classes - Modelo A

Precisão das Classes - Modelo A			
Decision Tree	Exatas	Biológicas	Humanas
	59,30%	**50,68%**	**47,62%**
K-NN	Exatas	Biológicas	Humanas
	74,10%	**91,26%**	**89,53%**
Naives Bayes	Exatas	Biológicas	Humanas
	67,52%	**78,29%**	**78,05%**

Na eficiência de precisão dos algoritmos de classificação por classes, mostram os melhores resultados, e novamente os algoritmos *K-NN* e *Naive Bayes* foram melhores, já o algoritmo árvore de decisão apresentou-se um baixo índice de acerto. Os algoritmos de classificação obtiveram resultados satisfatórios com um grande índice de acerto na classificação da área biológicas.

Já os resultados do modelo B são apresentados no Quadro 4.

Quadro 4: Precisão das classes - Modelo B

Precisão das Classes Modelo B						
Decision Tree	Biologia	Veterinária	Matemática	Sistemas de Informação	Administração	Economia
	38,33%	**57,14%**	**100%**	**50,94%**	**32,20%**	**38,89%**
K-NN	Biologia	Veterinária	Matemática	Sistemas de Informação	Administração	Economia
	81,63%	**95,24%**	**80,36%**	**60,61%**	**70,00%**	**68,75%**
Naives Bayes	Biologia	Veterinária	Matemática	Sistemas de Informação	Administração	Economia
	57,50%	**91,67%**	**91,89%**	**43,84%**	**48,28%**	**65,12%**

No modelo B, o índice de acerto por precisão das classes demonstra também o algoritmo *K-NN* como satisfatório para classificação por disciplinas. Já o algoritmo *Naive Bayes* teve médio índice de acertos e o algoritmo arvore decisão demonstrando não ideal para classificação por disciplinas. No modelo B os algoritmos de classificação obtiveram resultados satisfatórios com um grande índice de acerto na classificação na disciplina de matemática, por esse motivo o índice de acertos foi grande tendo em consideração com os termos específicos que existem na matemática.

A comparação entre os dois modelos mostra o melhor resultado para algoritmo *K-NN*, uma classificação mediana para o algoritmo *Naive Bayes* e por fim o algoritmo árvore de decisão apresentado um índice baixo de acerto.

A terceira análise foi verificar o tempo de execução de cada modelo A e B. O Quadro 5 apresenta o tempo de resposta de cada algoritmo para processamento das bases.

Quadro 5: Tempo de execução dos algoritmos

Tempo de Respostas dos Algoritmos – Modelo A	
Árvore de Decisão	1hr 5 min.
K-NN	2s
Naive Bayes	3s
Tempo de Respostas dos Algoritmos – Modelo B	
Árvore de Decisão	1hr 20 min.
K-NN	4s
Naive Bayes	6s

Nos dois modelos A e B, os algoritmos *K-NN* e o *Naive Bayes* apresentaram um processamento de segundos, o modelo B apresentou o dobro de segundos do modelo A. Já o algoritmo de Árvore de Decisão teve um tempo considerável alto para processamento dos dois modelos por área e cursos.

10 CONCLUSÕES

De acordo com os resultados apresentados no trabalho, o primeiro aspecto importante que podemos concluir foi à criação do ambiente semântico que possibilitou a organização dos conhecimentos e também a análise dos resultados da mineração de textos na classificação da precisão de acertos dos algoritmos.

O ambiente semântico apresentou-se viável para modelar um domínio através da ontologia e criá-lo na semântica MediaWiki, pois auxilia na organização dos conhecimentos através das propriedades semânticas de acordo com o modelo apresentado.

As análises realizadas com a validação *BootStrap* dos algoritmos de classificação, levando em consideração o índice de acerto de cada algoritmos nos modelos A e B, fica claro pelos resultados com que foi proposto o algoritmo com melhor índice de acertos em qualquer análise realizada foi o *K-NN*.

Já o algoritmo *Naive Bayes* obteve também um resultado favorável nos dois modelos, apresentando uma diferença baixa em relação o algoritmo *K-NN*, com isso sendo favorável para os resultados proposto.

O algoritmo Árvore de Decisão demonstrou quase os mesmo resultados nos modelos A e B. Demonstrou resultados não favoráveis e com um rendimento muito baixo de processamento.

Um resultado satisfatório após a realização da análise de mineração de textos, é a possibilidade de trabalhar para melhorar o algoritmo *K-NN* na integração das bases textuais com o domínio modelado inserido no ambiente semântico MediaWiki, gerando assim para o usuário uma eficiência na organização do conhecimento da MediaWiki de acordo com a classificação de desempenho do algoritmo.

No entanto, o melhor modelo apresentado com o melhor resultado de índice de precisão foi o modelo A apresentando um resultado melhor do que o modelo B, principalmente com o algoritmo de classificação *K-NN* com rápido tempo de processamento e o mais indicado para os dois modelos de bases textuais.

10.1 TRABALHOS FUTUROS

Nessa seção apresenta as possibilidades de trabalhos futuros. A primeira possibilidade seria melhorar o algoritmo *K-NN* na integração do ambiente semântico MediaWiki de acordo com as propriedades semânticas. Seria possível também estabelecer uma quantidade maior de propriedade semântica no ambiente para uma maior precisão na organização do conhecimento gerado na wiki.

REFERÊNCIAS

ARAGÃO, José L. Matos (2006). **Wikis e as Organizações: Usos e Aplicações**. Disponível em: http://wiki.softwarelivre.org/pub/Blogs/BlogPostVicenteAguiar20080911000357/ADIB1245.pdf
Acesso em: 02/09/2010.

ALMEIDA, Mauricio Barcellos, BAX, Marcelo P. "**Taxonomia para projetos de integração de fontes de dados baseados em ontologias**". *V Encontro Nacional de Pesquisa em Ciência da Informação*. Belo Horizonte, 2003.

_____.Uma visão geral sobre ontologias: pesquisas sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Revista Ci. Inf.,** Brasília, v. 32, n. 3, p. 7-20, set./dez. 2003.

ARGYRIS, C.; SCHON, D. **Organizational Learning*:* A theory of action perspective**. *Wesley*, Massachusetts, 1978.

BARION, Eliana Cristina Nogueira; LAGO, Decio. Mineração de Textos: Text Mining. **Revista de Ciências Exatas e Tecnologia**, Valinhos-SP, v. 3, n. 3, p.123-140, dez. 2008.

BEAN, Luan; HOTT, David. Wiki: A Speedy new tool to manage projects. **Journal of Corporate Accounting and Finance,** p .3-8, 2005.
Disponível em: http://kiwiwiki.co.nz/pmwiki/uploads/Technology/Software/Wiki%20a%20speedy%20tool%20to%20manage%20projects.pdf
Acesso em: 02/09/2010.

BERNERS-LEE, T.; HENDER, 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, New York, may.2001.

BJØRNSON, Finn; DINGSØYR, Torgeir. Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used. **Information and Software Technology**, v.50, n.11, p.1055-1068, fev./mar. 2008.

BORST, W. N. (1997). ***Construction of engineering ontologies***. Disponível em: http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf Acesso em: 02/04/2011.

CAMARA, Ferdinand da Costa. **A utilização de ambientes virtuais de aprendizagem no ensino presencial: Estudo de caso na disciplina de um programa de mestrado**. MACKENZIE-SP. São Paulo.2009.

CASTILHO, Noel Teodoro et.al. Aprendizagem organizacional e gestão do conhecimento. In: **XI SIMPEP.** Bauru - SP, 2004**.**

COELHO, F. L.. **Classificação semi-automática de monografias**. 2008. 71f. Trabalho de conclusão de curso (Graduação em Ciência da Computação) – Centro Universitário Feevale. Novo Hamburgo. 2008.

COUTO, Fabiano C. da Silva; BLATTMANN, Úrsula. Colaboração e interação na WEB 2.0 e Biblioteca 2.0. **Revista ABC: Biblioteconomia em Santa Catarina**, Florianópolis, v.12, n.2, p.191-215, jul./dez.2007.

DIAS, Gutenberg Marques. **Uso da Web 2.0 pelas organizações brasileiras: Quais são as contribuições dos novos recursos para alavancar a gestão do conhecimento?** UNIPEL- MG. Minas Gerais. 2009.

DUSYA, Vera; CROSSAN, Mary. **Organizational learning and knowledge management: toward an integrative framework**. In: EASTERBY-SMITH.

ESTIVALETE, Patricia Blini**. Documentação da arquitetura de sistemas e frameworks para processamento e análise de imagens: Uma abordagem baseada em visões da UML e padrões**. UFSM-RS. Rio Grande do Sul.2007.

GARVIN, D. Building a Learning Organization. **Harvard Business Review**, p. 45-49, july/aug.1993.

GONZAGA,M. V. **Modelagem: Uma aplicação com uma rede social**. UNIVERSIDADE DO RIO GRANDE DO SUL. Porto Alegre. 2009.

GRUBER, T. “Toward principles for the design of ontologies used for knowledge sharing?.” **International Journal of Human-Computer Studies 43**, n. 5-6, p.907-928, 1991.

GUARINO, N. (1998). **Proceesings of the fist International Conference on Formal Ontology in Information Sustems**.
Disponível em http://www.loa-cnr.it/Papers/FOIS98.pdf
Acesso em: 02/04/2011.

KERLI, Diego Cirino. **Inferência sobre ontologias**. Centro Universitário Feevale. Novo Hamburgo. 2007.

KIM, D.H. ***O elo entre aprendizagem individual e organizacional*** **em D. Klein. A gestão estratégica do capital intelectual**. Rio de Janeiro: Qualitymark, 1998.

KOLB, D. **Experiential Learning: Experience as the Source of Learning and Development**, Prentice Hall, Englewood Cliffs, NJ, USA, 1984.

LOPES, Fabrício Martins. **Um modelo perceptivo de limiarizacao de imagens digitais**. UFPR-PR. Curitiba.2003.

MAIA, Roberto Bomeny. **Intrusion detection ussing bayesian classifier**. UFRJ. Rio de Janeiro.2005.

MARJORIE, E.D. **Handbook of Organizational Learning and Knowledge Management. Malden: Blackwell**, p. 122-141, 2005.

MARTELETO, Regina Maria. Análise de redes sociais: Aplicação nos estudos de transferência da informação. **Revista Ciência da Informação**, Brasília, v. 30, n.1, p. 71-81, jan./abr. 2001.

MOURA, A.M.C. (2002). **A Web Semântica: fundamentos e tecnologias.**
Disponível em:
<http://www.des.ime.eb.br/%7Eanamoura/public/WebSemantica.pdf>
Acesso em: 04/04/2011.

NEVIS, E. C.; DI Bella, A, J.; Gould, J. M. Understanding organizations as learing systems. **Sloan Management Review**. v.36, n. 2 p. 73-85, 1995.

O´LEARY, Daniel: Wikis: From Each According to His Knowledge. **Published IEEE Computer Society,** p .34-41, 2008.
Disponível em:
https://msbfile03.usc.edu/digitalmeasures/oleary/intellcont/wikis-1.pdf
Acesso em: 02/09/2010.

PARVIN, Hamid (2008). **MKNN: Modified K-Nearest Neighbor**.
Disponível em:
http://www.iaeng.org/publication/WCECS2008/WCECS2008_pp831-834.pdf
Acesso em: 09/06/2011.

PEREIRA, Nivaldo Silva et al. A aprendizagem organizacional e a inovação: o caso da Concessionária de Energia do Sul do País. **Revista INGEPRO**, Santa Maria-RS, v.2, n.1, p.1-14, jan.2010.

PRIMO, Alex; RECUERO, Raquel. Hipertexto Cooperativo: Uma análise da escrita colaborativa a partir dos Blogs e da Wikipédia**. Revista da FAMECOS**, Porto Alegre, n.23, p.54-63, dez.2003.

PROBST, G.J.B; BÜCHEL, B.S.T. **Organizational Learning The competitive advantage of the future**. New York: Prentice Hall, 1997.

RECH, Jörg; RAS, Eric. **"The Future of Learning Software Organizations: Semantics – Collaboration - Aggregation New Technologies for Learning Software Organizations**" v.49, n.0, p.1-8, 2008.

RECUERO, Raquel (2004). **Teoria das Redes Sociais e Redes Socias na Internet**. Disponível em: http://galaxy.intercom.org.br:8180/dspace/bitstream/1904/17792/1/R0625-1.pdf Acesso em: 14/10/2010.

REZENDE, S. O. **Sistemas Inteligentes fundamentos e Aplicações**. 1ª. ed. Barueri-SP: Editora: Manole, v. I, 2005.

SCHONS, C. Henrique; COUTO, Fabiano. O uso de Wikis na gestão do conhecimento em organizações. **Revista de Biblioteconomia e Ciências da Informação**, v.8, n.27, p.1-10 mar. 2007.

SCHWEITZER, Fernanda. O serviço de referência da biblioteca central da UFSC e o programa de capacitação do usuário: desenvolvimento de uma ferramenta colaborativa com base na tecnologia Wiki. **Revista de Biblioteconomia e Documentação**, v.4, n.1, p.6-19, jan./jun. 2008.

SENGE, P. **A quinta disciplina**: **arte**, **teoria e prática da organização de aprendizagem**. São Paulo: Editora: Best Seller, 1990.

TEODORO, Camila Soares; OTTOBONI, Célia (2005). **Análise e conceituação de organizações que aprendem e aprendizagem organizacional – um estudo de caso**.
Disponível em: http://www.aedb.br/seget/artigos05/328_Analise%20organizacoes%20aprendem.pdf Acesso em: 14/09/2010.

SILVA, Janine P. da Silva et. al. (2009). **Gestão do conhecimento e aprendizagem organizacional: o caso da embasa.**
Disponível em:
http://www.simpoi.fgvsp.br/arquivo/2009/artigos/E2009_T00339_PCN51315.pdf
Acesso em: 15/09/2010.

SILVA, Eunice Palmeira. **Classifição de informação usando ontologias**. UFA. Maceió - PE. 2006.

WEL, Carolyn; MAUST, Brandon et.al. Wikis for Supporting Distributed Collaborative Writing. In: **Proceedings of the Society for Technical Communication 52nd Annual Conference**. Arlington-USA, 2005.
Disponível em:
http://depts.washington.edu/ibuxl/docs/STC_Wiki_2005_STC_Attribution.pdf
Acesso em: 02/09/2010

WENGER, E. **Communities of Practise: Learning, Meaning and Identity**, Cambridge University Press, Cambridge, UK, 1998.

ZANGISKI, Marlene; APARECIDA, S. Gonçalves et al.. Aprendizagem Organizacional e desenvolvimento de competências: uma síntese a partir da gestão do conhecimento. **Revista Produto & Produção**, v.10, n.1, p. 54-74, fev. 2009.

Apêndice A

WordList