EXTRAÇÃO DE CONHECIMENTO EM PROCESSOS DE CERTIFICAÇÃO PROFISSIONAL

Sérgio Assis Rodrigues

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Jano Moreira de Souza, Ph.D.

Prof. Geraldo Bonorino Xexéo, D.Sc.

Drª. Raquel Borba Balceiro, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2007

RODRIGUES, SÉRGIO ASSIS

Extração de Conhecimento em Processos de Certificação Profissional [Rio de Janeiro] 2007

XIII, 146 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2007)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Recuperação de Informações Textuais

2. Certificação Profissional

I. COPPE/UFRJ II. Título (série)

AGRADECIMENTOS

Agradeço a Deus por todas as maravilhas que Ele me proporcionou até hoje.

Agradeço aos meus pais, Antônio José e Layse, pelo apoio irrestrito nas minhas ações e por nortear meus caminhos.

Agradeço aos meus irmãos Fabiano, Maurício, Cristina e André, meus melhores amigos.

A minha linda Jaqueline que está ao meu lado em todos os momentos.

Ao Professor Jano, meu orientador, pela oportunidade de compartilhar suas brilhantes idéias.

A Jonice, minha co-orientadora, que me ajudou imensamente no decorrer deste trabalho.

Aos professores Geraldo Xexéo e Raquel Balceiro, por aceitarem fazer parte da banca.

Ao Marco pelas oportunidades e incentivos no trabalho.

Ao Maurício, por me propiciar grandes idéias neste estudo de caso.

Aos amigos de trabalho Ricardo, Rodrigo, Marcos e Carlos pela convivência e ajuda espontânea praticada.

Aos amigos de longa data, em especial, Fernando, Nicoliti e Sávio.

A Patrícia e a Carol pelas constantes conversas na sala de espera.

Gostaria ainda de agradecer a todas as pessoas que de alguma forma contribuíram para realização deste trabalho e que não foram citadas aqui.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

EXTRAÇÃO DE CONHECIMENTO EM PROCESSOS DE CERTIFICAÇÃO PROFISSIONAL

Sérgio Assis Rodrigues

Março / 2007

Orientador: Jano Moreira de Souza

Programa: Engenharia de Sistemas e Computação

Este trabalho apresenta uma forma de extração de conhecimento a partir de processos de certificação profissional. Neste contexto, muitos insumos foram considerados para uma análise qualitativa, entre eles: livros, apostilas, documentos de referência, documentos de opinião (críticas e sugestões), questões de provas, simulados e dicionários técnicos. Além disso, um sistema computacional foi concebido baseado na tecnologia de mineração de textos e nas particularidades encontradas nos processos de certificação que embasam o estudo de caso desta dissertação. O *software* em questão possibilita buscas por palavras chave, seja por freqüência ou proximidade das palavras e ainda indica a probabilidade de uma questão estar ou não em um conjunto de materiais didáticos. Abordou-se inicialmente o conceito de Mineração de Textos bem como Gestão do Conhecimento e Gestão por Competências, que embasaram a construção da ferramenta desenvolvida. Após a introdução destes conceitos, introduziu-se uma reflexão sobre como as técnicas de Mineração de Textos podem auxiliar na extração de conhecimento a partir dos materiais didáticos dos processos de certificação e qualificação profissional. Em seguida, a ferramenta proposta foi apresentada, elucidando as principais funcionalidades e a aplicabilidade de cada uma delas. E finalmente, como forma de aferir a ferramenta, um conjunto de testes foi realizado a partir de dados reais de uma grande empresa brasileira.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

KNOWLEDGE EXTRACTION IN PROFESSIONAL CERTIFICATION

Sérgio Assis Rodrigues

March / 2007

Advisor: Jano Moreira de Souza

Department: Computing and Systems Engineering

This work presents a kind of knowledge extraction in professional certification processes. In this context, a lot of materials was considered for a quality analysis, as example: books, notes, reference documents, opinions, suggestions, questions of tests and technical dictionaries. In addition, a computational system was conceived based on text mining technology and particularities found in the certification processes which base this dissertation. The software affords searches with key words, using frequency or proximity among words to compute the documents relevance, and still calculates the probability of matchmaking questions in a book. This work introduced concepts about Text Mining, Knowledge Management and Competences, which embassy the developed tool. After that, it is discussed the best way of using the text mining techniques to find knowledge from the certification documents and materials. The text presented also this environment, explaining its functionalities and the usability of each one. Finally, the case study showed the tests accomplished in a large Brazilian company.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

- **ASCII** *American Standard Code for Information Interchange*
- **CNE** Conselho Nacional de Educação
- **CNPq** Conselho Nacional de Desenvolvimento Científico e Tecnológico
- **DTS** *Data Transformation Service*
- **GCC** Sistema de Gestão de Conhecimento Científico
- **IDC** *Institute Data Corporation* Brasil
- **HTML** *HyperText Markup Language* (Linguagem de Formatação de Hipertexto)
- **KDD** *Knowledge Discovery from Data* (Descoberta de Conhecimento em Dados)
- **KDT** *Knowledge Discovery from Texts* (Descoberta de Conhecimento em Textos)
- **LDB** Lei de Diretrizes e Bases
- **MEC** Ministério da Educação
- **OLAP** On Line Analytic Processing
- **PCE** Pesquisa Contextual Estruturada
- **RC2D** Representação do Conhecimento Contextualizado Dinamicamente
- **SRI** Sistema de Recuperação de Informações
- **TI** Tecnologia de Informação
- **VSM** *Vector Space Model*

I - INTRODUÇÃO

O interesse pela extração de conhecimento cresce como auxílio ao processo de tomada de decisão, dado que, em geral, as principais fontes de armazenamento de informações se encontram em formato não estruturado (TAN, 1999).

Os processos de manipulação de informações vêm sendo gradativamente incorporados no cotidiano das pessoas. Com o desenvolvimento de tecnologias de comunicação cada vez mais eficientes e, principalmente, com o advento da Internet, o volume de informações que uma pessoa tem acesso cresce diariamente. Entretanto, esta massa de informações torna mais difícil a sua absorção.

Nestas situações, onde uma pessoa tem facilidade de acesso a uma grande quantidade de informações, faz-se necessária a utilização de mecanismos que indiquem quais caminhos devem ser tomados para aumentar as chances de obter a informação de que se necessita, ou seja, aquilo que é realmente relevante.

I.1 - CONTEXTO

O contexto analisado neste trabalho corresponde ao modelo de qualificação e certificação profissional utilizado em uma grande indústria brasileira. A finalidade deste modelo é a capacitação da mão-de-obra interna por meio de cursos e provas de certificação. Neste âmbito, há um grande número de pessoas envolvidas produzindo diferentes documentos diariamente.

Lavin (1992) esclarece que, em ambientes onde existe diversidade de produtores de informação, há também uma variedade de materiais que se apresentam em diferentes formas. Alguns desses materiais aparecem em uma forma comum, como um livro ou jornal, enquanto outros são peculiares ao negócio praticado, apresentando características especiais, como banco de dados, planilhas eletrônicas e apresentações, por exemplo. Desta forma, a ferramenta de extração de informações proposta neste trabalho irá considerar o contexto organizacional apresentado.

A empresa abordada é dividida em vários departamentos e unidades de negócio espalhadas pelo Brasil. Para garantir a confidencialidade das informações disponibilizadas, serão utilizados neste texto os nomes EmpresaXPTO e DepartamentoA cujas denominações ocultam os nomes reais da empresa e do departamento abordados. Para um melhor entendimento, será chamado de Operador o

funcionário que trabalha na área de operação das unidades de negócios (indústrias) da EmpresaXPTO.

I.2 - OBJETIVOS

Conforme descrito, a EmpresaXPTO possui uma grande diversidade de documentos. A área do DepartamentoA realiza há alguns anos o Programa de Qualificação e Certificação Profissional da sua Força de Trabalho (Operadores). Muitos dados foram gerados neste período e a busca de informações qualitativas referentes a estes processos é dificultada pela ausência de padrões nos documentos.

O objetivo principal deste trabalho é apresentar como mecanismos de recuperação e manipulação de informações textuais podem ser utilizados em ambientes de Certificação Profissional, proporcionando uma forma de acesso facilitada às informações relevantes em meio aos insumos envolvidos, como por exemplo: livros, questões de prova, dicionários técnicos, documentos de críticas e sugestões.

As técnicas e métodos utilizados são provenientes da área de recuperação de informações textuais. Deste modo, é possível identificar dentre uma grande coleção de documentos, aqueles com maior relevância segundo tipo de busca proposto pelo usuário.

Para realizar a análise dos métodos e técnicas, foi desenvolvido um *software* de manipulação de informações textuais (mineração de textos) que permite a aplicação prática dos métodos analisados. Desta forma, foram desenvolvidas ferramentas de busca de informações a partir de palavras chave, bem como mecanismos que possam analisar estas consultas de forma contextual (considerando a proximidade das palavras procuradas).

As análises poderão ser efetuadas sobre um conjunto de documentos textuais gerados a partir dos processos de Certificação Profissional, ou seja, livros, questões de prova, críticas e sugestões, dicionários técnicos, e documentos de referência pedagógica.

A partir da ferramenta de manipulação destes insumos, os envolvidos nos processos de Certificação poderão encontrar materiais didáticos com maior rapidez, assim como verificar possíveis divergências entre questões e livros, por exemplo.

I.3 - ORGANIZAÇÃO DO TRABALHO

Os capítulos 2 e 3 correspondem à revisão bibliográfica, com o objetivo de demonstrar a abrangência do tema. No 2º capítulo, há uma introdução sobre Gerência do Conhecimento, que abordará Gestão de Conhecimento e Competências além de introduzir o tema Certificação Profissional. No capítulo 3, constam as principais etapas dos processos de Mineração de Textos, assim como ferramentas e trabalhos desenvolvidos nesta área.

No capítulo 4 é apresentado o ambiente organizacional e a metodologia utilizada nos processos de Qualificação e Certificação Profissional da Empresa analisada.

A implementação computacional e os procedimentos de mineração de textos utilizados em função do cenário são descritos no capítulo 5, enquanto as interfaces utilizadas na ferramenta proposta podem ser vistas no capítulo 6.

No capítulo 7 são apresentados os testes usados para validar a metodologia adotada, utilizando a ferramenta desenvolvida e mecanismos de visualização de análise de resultados em múltiplas dimensões.

E, finalmente, o texto segue com as conclusões obtidas no decorrer deste estudo, além de destacar os possíveis trabalhos futuros e outras observações do autor.

II - GERÊNCIA DO CONHECIMENTO

O capital intelectual vem sendo cada vez mais valorizado nas organizações. Não se trata de um fenômeno restrito ao mundo empresarial, pois educação, capacitação profissional, conhecimento, competências e habilidades, de modo geral, ganham cada vez maior visibilidade na contemporânea sociedade do conhecimento (DRUCKER, 1996).

Segundo Drucker (1996), este novo paradigma, fruto das exigências do mundo globalizado, requer pessoas não só alfabetizadas, sabendo ler, escrever, fazer contas e obedecer a normas. Exige que as pessoas dominem, por exemplo, habilidades básicas de computador e compreendam e questionem o sistema histórico-político-social no qual estão inseridas. O novo mundo do trabalho passa a demandar trabalhadores do conhecimento, pessoas capacitadas e motivadas a criar inovações que possam trazer benefícios aos processos produtivos da empresa, às relações interpessoais e, obviamente, maior lucratividade.

Este tipo de trabalhador, raro em décadas passadas, ficou conhecido no mundo organizacional como *intrapreneur* (PINCHOT, 1989), uma espécie de intra-empreendedor da empresa, que ousa inovar dentro da organização, buscando, na maioria das vezes, simplesmente desenvolver novas formas de melhorar o seu dia-a-dia (e dos demais também) no ambiente de trabalho. É um funcionário automotivado e autônomo que se sente muito à vontade com novos negócios e atividades (JACOBSEN, 1992). De acordo com Jacobsen (1992), um *intrapreneur* legítimo tem como características: buscar chances de desenvolver suas potencialidades e automotivar-se pelo simples prazer de ver seus inventos funcionarem; sentir-se útil por criar inovações (bens e serviços) que facilitem a sua vida e dos demais; trabalhar em cooperação com os demais, em tarefas construtivas e partilhadas.

Jacobsen (1992) complementa dizendo que é importante que as empresas desenvolvam, motivem e reconheçam seus *intrapreneurs*, em benefício da própria organização, mesmo que uma inovação tecnológica, organizacional ou estratégica possa parecer insegura e longínqua quando comparada ao cotidiano seguro das práticas tradicionais conhecidas.

Assim, o novo mundo do trabalho demanda não só autonomia e criatividade dos trabalhadores, como passa a valorizar um conhecimento particular e subjetivo de cada trabalhador: o chamado conhecimento tácito (NONAKA & TAKEUCHI, 1997). Este tipo de conhecimento pode ser externalizado e, desta forma, trabalhado em ferramentas computacionais.

O objetivo desta seção é discutir como vem sendo desenvolvido no mundo das organizações o gerenciamento destes tipos de conhecimento, difíceis de serem mensurados, formulados e comunicados, pela subjetividade do conhecimento tácito e pela dificuldade de externalizar as crenças, valores e experiências pessoais dos trabalhadores.

II.1 - Gestão do Conhecimento

Inicialmente, é importante destacar uma pequena diferenciação entre dado, informação e conhecimento, pois a confusão no entendimento do significado de cada um deles pode gerar enormes dispêndios para a organização. Consideram-se dados como sendo uma seqüência de números, palavras, sob nenhum contexto específico. Quando organizamos estes dados e apresentamos o contexto onde eles se situam, pode-se considerar que eles passam a ser uma informação. Por sua vez, o conhecimento é a informação organizada, com o entendimento de seu significado. Entre estes 3 elementos, os dados são aqueles que são menos importantes, pois não sofrem qualquer agregação de valor. Normalmente, os dados precisam ser manipulados e tratados para conterem algum valor e, então, se transformarem em informação (NONAKA & TAKEUCHI, 1997).

Os dados, no contexto organizacional, são descritos como registros de transações, pois apenas descrevem parte daquilo que aconteceu, não fornecendo julgamento, nem interpretação e nem qualquer base sustentável para a tomada de ação ou decisão. Segundo Peter Drucker (1996), informações são "dados dotados de relevância e propósito". Desta forma, a informação, pode, em algumas situações, ter grande valor. É possível citar como exemplo, um grande furo jornalístico que pode valer milhões, naquele exato instante, mas que a partir de então, com o passar do tempo, passará a ter seu valor depreciado.

Segundo Davenport & Prusak (1998), dados só se tornam informação a partir dos seguintes métodos:

- Contextualização: saber a finalidade dos dados coletados;
- Categorização: saber as unidades de análise;
- Cálculo: os dados que podem ser analisados matematicamente;
- Correção: os erros são eliminados dos dados;
- Condensação: os dados podem ser sumarizados.

Com relação ao processo de formação do conhecimento, Rodriguez y Rodriguez (2001), afirma que este se inicia através de eventos que ocorrem e, por sua vez, geram fatos e dados. Estes dados quando devidamente tratados, manipulados e interpretados, geram informações. Estas informações quando testadas, validadas e codificadas, transformam-se em conhecimento.

Daft (2002) define conhecimento coletivo como a combinação de informações pelos cérebros coletivos que se baseia em conhecimento anterior. Antonelli e Quéré (2004) acrescentam que o conhecimento externo do indivíduo é uma contribuição importante no processo de produção de novos conhecimentos.

Para Davenport & Prusak (1998), o conhecimento é uma mistura fluida de experiências, valores, informação contextual e *insight*, a qual possibilita a existência de uma estrutura que permite a avaliação e incorporação de novas experiências e informações. O conhecimento tem origem na cabeça das pessoas e, nas organizações, está presente não apenas em documentos, mas também em rotinas, processos e práticas.

O conhecimento deriva da informação assim como esta se origina de dados. Segundo Davenport & Prusak (1998), a transformação da informação em conhecimento é possível a partir da:

- Comparação: como as informações relativas a um determinado assunto podem ter alguma relação ou aplicação em outras situações;
- Conseqüências: implicação que determinada informação pode trazer para tomada de alguma decisão e/ou ação;
- Conexões: relação entre o conhecimento adquirido e o existente;
- Conversação: o que as pessoas pensam sobre aquela informação.

Neste contexto, há o pressuposto de que o conhecimento humano é criado e expandindo através da interação social entre o conhecimento tácito e o explícito, que se

chama "conversão do conhecimento". O trabalho de Nonaka & Takeuchi (1997) considera que as empresas "criadoras de conhecimento" são aquelas que criam novos conhecimentos, disseminam esses conhecimentos pela organização inteira e os incorporam em seus produtos e serviços, sendo que o processo de conversão de conhecimento ocorre conforme o modelo da Figura 1.

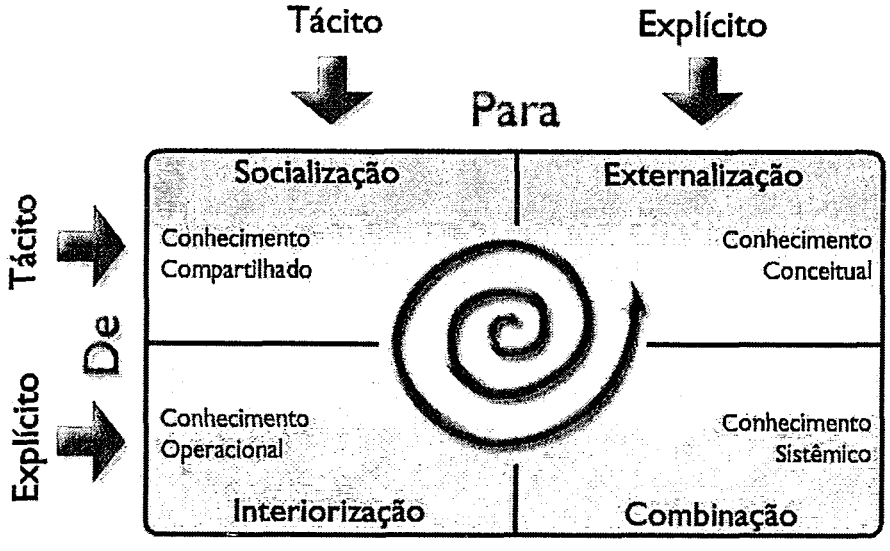


Figura 1 - Espiral do Conhecimento (Modelo SECI – NONAKA & TAKEUCHI, 1997)

O conhecimento explícito é aquele que está materializado nos livros, manuais, documentos, periódicos, base de dados, repositórios etc. Por ser um produto concreto, o conhecimento explícito normalmente é captado pelas organizações. O conhecimento tácito é aquele gerado e utilizado no processo de produção do conhecimento explícito, constituindo-se de idéias, fatos, suposições, decisões, questões, conjecturas, experiências e pontos de vista (NONAKA & TAKEUCHI, 1997).

Os autores resumem este modelo espiral da seguinte forma: inicialmente a **socialização** desenvolve um campo de interação que permite o compartilhamento das experiências dos indivíduos. A partir da **externalização**, é gerado o diálogo ou reflexão coletiva com uso de metáforas ou analogias, o que gera conceito. O modo de **combinação** possibilita a junção do conhecimento recém criado com o conhecimento existente, o que resulta em um novo processo, sistema ou modo de fazer. A **interiorização** (ou internalização) ocorre a partir do "aprender fazendo".

Para Drucker (1996), o conhecimento é a capacidade de aplicar a informação a um trabalho específico, e só vem com um ser humano e sua capacidade intelectual e

habilidade. Segundo o autor, a gestão do conhecimento contempla o processo de obter, gerenciar e compartilhar a experiência e especialização da força de trabalho, com objetivo de se ter acesso à melhor informação no tempo certo, utilizando-se de tecnologias de forma corporativa (DRUCKER, 1996).

Para Wah (2000), a gestão do conhecimento agrega valor às informações, filtrando, resumindo e sintetizando estas, e dessa forma, desenvolvendo um perfil de utilização pessoal que ajuda a levá-las à ação.

Ainda segundo o autor (WAH, 2000), o tema central da gestão do conhecimento é aproveitar os recursos existentes na empresa para que as pessoas procurem, encontrem e empreguem as melhores práticas em vez de tentar reinventar a roda. Para que isto ocorra, é importante que haja uma cultura de identificação, captação, organização, disseminação e proteção das experiências individuais relevantes e soluções criativas. Além disso, deve existir um ambiente de aprendizado interativo, no qual as pessoas transfiram prontamente o conhecimento, internalizem-no e apliquem-no para criar novos conhecimentos.

Para Nonaka & Takeuchi (1997), a grande competitividade do mundo moderno obriga as empresas a terem uma grande capacidade de aprender. Sendo assim, o aprendizado adquirido durante a execução de uma nova atividade deve ser proliferado para toda a empresa e não ficar restrito às pessoas ou somente ao grupo participante. A função da organização é incentivar o processo de criação de conhecimento e fornecer o contexto apropriado para facilitar as atividades em grupo, criando e acumulando o conhecimento em nível organizacional.

Segundo Davenport & Prusak (1998), também é necessário combater a chamada "Cultura Nacional de Individualismo Possessivo". Além disso, em um ambiente de alta competitividade, o ciclo de inovação tende a ser cada vez menor. Para isto, é necessário ter pessoas criativas nesse processo e equipes multidisciplinares. De acordo com os autores, muitas vezes, o sucesso pode levar à falta de disposição da empresa em se adaptar, de reconhecer novos desafios e de responder a estes desafios através da geração de conhecimento novo.

Para Lemos (2000), existem possibilidades concretas de acesso e transferência de informações, propiciadas principalmente pelas novas tecnologias de informação e comunicação. Entretanto, o acesso às informações não é suficiente para que um

indivíduo, empresa ou país se adapte às condições técnicas e de evolução do mercado. É imprescindível que estes agentes mantenham interação social com outros. As mudanças são muito rápidas e somente aqueles que estão envolvidos na criação do conhecimento possuem possibilidades reais de acesso aos seus resultados.

Um dos desafios da gestão do conhecimento é assegurar que compartilhar o conhecimento seja mais lucrativo do que enclausurá-lo. É importante a criação de uma cultura na qual as pessoas estejam aprendendo, crescendo e se desenvolvendo, pois esta legitimará o relacionamento entre os colaboradores e a organização (COVEY, 2005). Segundo o autor, quando o colaborador compartilha o que aprende, ele se compromete socialmente em aplicá-lo e, mais ainda, quando este sabe que tem que ensinar o que aprende, ele aprende melhor e mais rápido.

Sob o ponto de vista acadêmico, Teixeira (2005) afirma que a gestão do conhecimento é um campo novo na confluência entre teoria da organização, estratégia gerencial e sistemas de informação. Desta forma, é comum encontrar na literatura especializada as questões de gestão de conhecimento associadas à organização do aprendizado, reengenharia de processos, corporações virtuais, novas formas de organização, educação para o trabalho, criatividade, inovação e tecnologia da informação.

Quando se trata de Tecnologia de Informação (T.I.), o autor sugere que a tecnologia deve apoiar a comunicação empresarial e a troca de idéias e experiências, que facilitem e incentivem as pessoas a se unir, a participar, a tomar partes em grupo, e a se renovar em redes informais (TEIXEIRA, 2005).

Segundo Lemos (2000), com o potencial oferecido pelos novos meios técnicos disponibilizados com as tecnologias de informação, intensifica-se a geração e absorção de conhecimento e as possibilidades de implementação de inovações. As exigências de especialização ao longo da cadeia de produção se tornam cada vez maiores. As novas tecnologias geram, assim, tanto os meios para a cooperação, como a necessidade de criação de mais intensivas e variadas formas de interação e aprendizado intensivo.

Terra (2000) aponta que são evidentes os casos de sucesso de empresas com saltos significativos na sua competitividade a partir do uso de ferramentas de T.I., mas no entanto, sua utilização, apesar de necessária, não é suficiente. Os grandes benefícios oriundos destas tecnologias estão relacionados ao maior grau de conectividade entre as

pessoas, maior disseminação das informações ao longo dos vários níveis hierárquicos, bem como com os parceiros comerciais, fornecedores e clientes.

Segundo o autor, as aplicações e tecnologias desenvolvidas têm como principal objetivo a obtenção de informações e conhecimentos que possam ser efetivamente utilizados nos diversos processos e atividades da organização. Caso este propósito não seja atendido, a utilização das mesmas começa a ser questionada, chegando, até mesmo, a serem abandonadas. O que se vê atualmente nas organizações é uma enorme quantidade de dados que não são interpretados e nem proporcionam informações relevantes para a tomada de decisões empresariais. Tal fato pode gerar frustrações e descrédito quanto à utilização das tecnologias de informação em geral (TERRA, 2000).

Independentemente de suas características e qualidades, é importante estar ciente de que nenhuma ferramenta de tecnologia da informação atenderá a todas as necessidades das organizações. O gerenciamento do conhecimento não é obtido a partir do uso de uma determinada tecnologia por si só, e sim a partir da utilização de uma série de tecnologias de forma compartilhada e conciliada às atividades exercidas pela organização, bem como a existência de uma estrutura organizacional voltada para o conhecimento.

II.2 - Gestão por Competências

O termo competência está envolvido diretamente com os sentidos de resultado e êxito (RESENDE, 2003). Segundo o autor, em uma sociedade cada vez mais competitiva, pessoas e organizações buscam por melhores resultados, necessitando, portanto, desenvolver novas competências. Por outro lado, o conceito de competência também vem sendo utilizado para qualificar manifestações espirituais e comportamentos a elas relacionados, como no caso da cidadania. Sendo assim, o termo competência mostra uma grande variedade de significados e aplicações e o requisito competência se manifesta de muitas maneiras, nas pessoas e nas organizações.

Para Ruas (1999), a competência não se reduz ao saber, nem tampouco ao saber-fazer, mas sim à capacidade de mobilizar e aplicar esses conhecimentos e capacidades. Zarifian (2001) define competência como "uma combinação de conhecimentos, de saber-fazer, de experiências e comportamentos, que se exerce em um contexto preciso".

Segundo Resende (2003), nos últimos anos o tema competência, seu desenvolvimento e sua gestão entraram na pauta das discussões acadêmicas e empresariais, associado as diferentes instâncias de compreensão, seja no nível da pessoa (a competência do indivíduo), das organizações (*core competences*) ou dos países (sistemas educacionais e formação de competências).

Desta forma, o sucesso de uma organização depende de seus produtos e serviços, mas seu crescimento perene está ligado às competências de seus profissionais. Para Fleury & Fleury (2000) competência é uma palavra de senso comum, utilizada para julgar uma pessoa qualificada a realizar algo. É saber agir de forma responsável e reconhecida, que implica mobilizar, integrar, transferir conhecimentos e habilidades que agreguem valor à organização e valores sociais ao indivíduo. O seu oposto, ou o seu antônimo, não implica apenas a negação desta capacidade, mas guarda um sentimento pejorativo, depreciativo.

Entender o que é competência é o ponto de partida para predizer o futuro das organizações, dos profissionais e da sociedade. Segundo Resende (2003), o conceito de competência começou a adquirir novos significados a partir de estudos feitos por David C. McClelland, no inicio da década de 70, quando participava de um processo de seleção de pessoal para o Departamento de Estado americano. McClelland publicou em 1973 um artigo com o título "*Testing for Competence Rather Than Intelligence*" (McCLELLAND, 1973). Neste texto, o autor afirma que os testes tradicionais de conhecimento e inteligência utilizados na seleção de pessoal não eram capazes de predizer sucesso no trabalho e na vida e ainda favoreciam preconceitos contra minorias, mulheres e pessoas de nível socioeconômico inferior.

A competência, segundo o autor, é uma característica subjacente a uma pessoa que é casualmente relacionada com desempenho superior na realização de uma tarefa ou em determinada situação. Diferenciava assim competência de aptidões: talento natural da pessoa, o qual pode vir a ser aprimorado, de habilidades, demonstração de um talento particular na prática e conhecimentos, o que as pessoas precisam saber para desempenhar uma tarefa (McCLELLAND, 1973). Em um segundo momento, Boyatzis (1982), um colaborador em suas pesquisas, afirmava que competência no trabalho é uma destacada característica de um empregado (que pode ser motivação, habilidade, conhecimento, auto-imagem, função social) que resulta em desempenho efetivo e/ou superior.

Estas constatações levaram ao encontro de métodos de avaliação que permitissem identificar variáveis de competência que pudessem predizer êxito nas atuações no trabalho e na vida e selecionar pessoas sem discriminação de sexo, cor ou condição social. McClelland (1973) desenvolveu e aplicou uma técnica especial para destacar as variáveis de comportamento que explicassem o motivo pelo qual determinados diplomatas do Departamento de Estado tinham sucesso e outros não, em suas difíceis missões em países onde havia forte rejeição à presença dos EUA. Os resultados dessa pesquisa indicaram claramente diferenças de habilidades, aptidões e atitudes entre os melhores e os piores embaixadores.

A identificação dessas características pessoais que faziam diferença no resultado das ações desses dois grupos permitiu a McClelland e sua equipe enfatizarem o novo conceito para competência, confirmando suas suposições relativas às limitações dos testes tradicionais para identificar e avaliar competências, bem como o fato de discriminarem grupos sociais diferenciados (McCLELLAND, 1973).

Considerando às condições atuais das empresas, essa preocupação com o acertar do prognóstico do desempenho é muito importante porque evita, por exemplo, as conseqüências negativas de custos, produtividade e clima de trabalho numa possível seleção de pessoal. Evita ainda que pessoas sejam colocadas em um posto de trabalho para o qual não possuem os requisitos de qualificações, aptidões e habilidades. Observa-se portanto, quantas conseqüências negativas podem ser geradas por não se atender ao princípio básico da seleção e da promoção: "colocar a pessoa certa no lugar certo." Princípio este que continua valendo mesmo quando se cogita necessitar de policompetências para atuar em trabalhos multifuncionais (RESENDE, 2003).

Lawler (1996) também argumentou sobre esta linha de raciocínio, mostrando que trabalhar com o conjunto de habilidades e requisitos definidos a partir do desenho do cargo poderia não atender mais às demandas de uma organização complexa, mutável em um mundo globalizado. O autor afirmou ainda que, em tais situações, o conceito de qualificação propiciava o referencial necessário para se trabalhar a relação profissional indivíduo x organização, de forma que a qualificação estaria usualmente definida pelos requisitos associados à posição, ou ao cargo, ou pelos saberes ou estoque de conhecimentos da pessoa, os quais podem ser classificados e certificados pelo sistema educacional.

Notadamente, o conceito de competência se aplica a uma característica ou um conjunto de características ou requisitos. Conhecimento ou uma só habilidade ou aptidão, por exemplo, indicados como uma condição capaz de produzir efeitos ou resultados na solução de problemas pode ser chamada de competência. Ter conhecimento e experiência e não saber aplicá–los em favor de um objetivo, de uma necessidade, de um compromisso, significa não ser competente, no sentido aqui destacado. Competência é, portanto, resultante da combinação de conhecimentos com comportamentos. Conhecimentos incluem formação, treinamento, experiência, auto-desenvolvimento. Comportamento engloba habilidades, interesse, vontade (RESENDE, 2003).

Segundo Rabaglio (2001), todo profissional tem um perfil que compreende competências técnicas e comportamentais. Segundo a autora, o grande desafio do indivíduo consiste em desenvolver competências como flexibilidade, criatividade, inovação e empreendorismo.

Rabaglio (2001) evidencia que Gestão por Competências deve ser um mecanismo que permita a criação de um modelo de competência para cada função dentro da empresa, elaborando um mapeamento que faça parte das estratégias de competitividade e diferenciação no mercado de trabalho.

Para Gramigna (2002), a Gestão por Competências é um programa que se instala por meio de blocos de intervenção, ou seja, etapas que se sucedem de forma simultânea ou por uma seqüência de passos. A autora define 5 etapas: sensibilização, definição de perfis, avaliação de potencial e formação do banco de talentos, capacitação e gestão do desempenho.

Para Resende (2003), Gestão por Competências significa programar planos com aplicação de princípios e técnicas de gerenciamento para desenvolver competências específicas que o executivo, equipes ou áreas precisam adquirir e aplicar, como: trabalho em equipe, competência logística, negociação de contratos e competência na utilização de terceiros, por exemplo.

Neste contexto, iniciativas organizacionais devem ser elaboradas para suplantar deficiências de competências, aumentando e aferindo o grau de competência da instituição. Chiavenato (2002) destaca que a integração dos funcionários na organização é uma das tarefas básicas da administração de recursos humanos e um dos

pré-requisitos para obtenção da produtividade e otimização almejada pelas empresas. Rifkin (2004) reflete que a automação e a reengenharia tomaram conta do trabalho humano em muitas áreas relacionadas, havendo, portanto, a necessidade de que o elemento humano se destaque cada vez mais através da qualificação profissional.

No âmbito deste trabalho, as competências correspondem apenas a uma forma de categorização dos documentos avaliados. Como se tratam de competências técnicas, o texto não faz diferenciação entre habilidades e competências.

II.3 - Qualificação e Certificação Profissional

Com as transformações técnico-organizacionais, a qualificação profissional integra-se ao novo perfil dos profissionais, satisfazendo a demanda atual do sistema produtivo e econômico. A qualificação profissional é a compreensão dos conhecimentos e habilidades necessários para a realização do trabalho, fazendo com que o trabalhador se torne apto e possibilitando-lhe interagir com freqüentes mudanças, além de estimular a criatividade e agilidade nos momentos necessários (PINHEIRO *et al*, 1996). Segundo Pinheiro (1996), essas competências fazem com que o profissional compreenda instruções complexas para que possa intervir adequadamente no processo de trabalho.

A qualificação, segundo Rocha (2002), está associada aos seguintes fatores:

- Conhecimento que as novas diretrizes organizacionais requerem dos profissionais, desde a capacidade de abstração, passando pela formação, até as chamadas competências "tácitas";
- Conduta do profissional face às novas formas de organização do trabalho (responsabilidade, confiabilidade, envolvimento, participação, capacidade de interação e de trabalho em grupo);
- Articulação entre os diversos saberes do profissional que a prática laboral estabelece (por exemplo: Polivalência, Policognição Tecnológica e Politecnia);
- Relação entre perfis de qualificação e a formação de identidades grupais;
- Relação entre os novos hábitos, habilidades e comportamentos e a possibilidade da "expansão do indivíduo humano".

O mercado de trabalho busca profissionais com diversas competências, que conheçam o processo de trabalho e que possam socializar seus conhecimentos de forma

simples, objetiva e ética com o seu grupo de trabalho. Além disso, as novas tecnologias exigem do usuário o domínio de conteúdos variados e conhecimentos básicos mais amplos, mesmo quando o usuário se encontra apenas marginalmente integrado ao mercado, porque ainda assim é atingido pelas tecnologias. Desta forma, o profissional deve estar disposto a constantes aprendizagens e busca de novos conhecimentos para inserir-se no mercado atual, onde existem grandes mudanças e aprimoramentos constantemente (PINHEIRO, 1996).

Conforme Pinheiro (1996), é essencial a capacitação em conhecimentos básicos da qualificação requerida, desenvolvimento de competências, incluindo atitudes, valores éticos e hábitos, aos quais serão somadas às habilidades específicas ao desempenho da função visada. Além disso, o mercado de trabalho busca profissionais com diferencial, exigindo que estes tenham auto-gerenciamento, flexibilidade, raciocínio, sejam comunicativos, tenham capacidade de ler, interpretar, saibam trabalhar em grupo, desenvolvam relações interpessoais saudáveis, tenham conhecimentos técnicos, científicos, sociais e econômicos. Sendo assim, o profissional deve assimilar as competências operacionais e privilegiar o desenvolvimento das competências cognitivas e sócio-comunicativas para atender a este mercado. E para isto, a gestão do conhecimento na organização é imprescindível.

Segundo Gonçalves *et al* (2004), para as competências que o mercado exige, conforme relatado, é necessária a ocorrência de um direcionamento na aprendizagem. Preparar o profissional para ter competências polivalentes não significa prepará-lo para diversos postos. O ideal é fazer com que o profissional aprenda a dominar as técnicas no âmbito intelectual, mediando o conhecimento das bases técnicas e científicas que fundamentam a prática, utilizando diferentes e complexos instrumentos, tornando-se mais crítico e sintonizando-se com as exigências criadas pela modernização do setor produtivo.

Por outro lado, é importante ressaltar que é bastante sugestivo que a qualificação possa ser mensurada e diagnosticada. Neste contexto, processos como Certificações indicam a autenticação do conhecimento em determinado contexto.

O tema certificação está em evidência nas diversas organizações espalhadas pelo mundo. Entretanto, ainda há muito a discutir e, na verdade, somente grandes empresas de ponta e que se encontram à frente do seu tempo estão desenvolvendo um trabalho sério e responsável e, através dele, vem beneficiando os seus usuários a partir da

implantação desse novo modelo de ensino-aprendizagem. Dessa forma, a certificação profissional, desde que atestada por entidade reconhecida, é fator fundamental para o currículo de qualquer pessoa, principalmente, quando este profissional passa a ser aceito internacionalmente, pelas qualificações e habilidades que ostenta em seu currículo (FICO & ALONSO, 2006).

Fico e Alonso (2006) indicam que a certificação profissional vem sendo cada vez mais exigida pelas empresas na Europa e nos Estados Unidos. Esta exigência é para a contratação de um profissional ou para especificar a titulação requerida por uma pessoa para exercer determinada função ou elaborar um tipo de serviço.

Ainda segundo os autores, na Inglaterra e nos Estados Unidos a certificação foi concebida para reparar questões ligadas a deficiências do plano educacional. Na França se buscou estabelecer o que seria uma pedagogia das competências entendida não apenas como as práticas de transmissão na escola, mas toda atividade social que englobe a seleção dos saberes a serem transmitidos pela escola, sua organização, sua distribuição em uma instituição diferenciada e hierarquizada, sua transmissão por agentes especializados e sua avaliação por métodos adequados ou uma lógica das competências, a que a certificação estaria a serviço (FICO & ALONSO, 2006).

No Brasil, o Conselho Nacional de Educação (CNE) foi ambicioso ao dar ao Ministério da Educação (MEC) o encargo de criar um sistema nacional. Desde então, o MEC tem variado em instâncias e em concepções, sem ter chegado ainda a nenhum caminho definido para conceber tal sistema. Na verdade, a Lei de Diretrizes e Bases (LDB), em seu artigo 41, estabelece que "todo o conhecimento adquirido na educação profissional, inclusive no trabalho, poderá ser objeto de avaliação, reconhecimento e certificação, para prosseguimento ou conclusão de estudos" (BRASIL, 1996). Foi com base nesse artigo que o CNE (Resolução 04/99) (BRASIL, 1999) atribuiu ao MEC "a organização de um Sistema Nacional de Certificação Profissional", com a participação de "representantes dos trabalhadores, dos empregadores e da comunidade educacional". Diz ainda que o MEC "fixará normas para o credenciamento de instituições, para o fim específico da certificação profissional". O CNE argumenta que "não é cabível nos dias atuais a postura de desconsideração pelas habilidades, conhecimentos e competências adquiridas por qualquer pessoa por meio de estudos não-formais ou no próprio trabalho" (BRASIL, 1997).

Entendida em seu sentido mais amplo, a certificação é a comprovação formal dos conhecimentos e competências do trabalhador, requeridos pelo sistema produtivo e definidos em termos de padrões ou normas acordadas previamente.

Em um mundo caracterizado pelo desenvolvimento tecnológico, pelo predominante uso da informação e pelo forte incremento do comércio internacional, este recurso tem se tornado cada vez mais eficiente para sintetizar conhecimentos adquiridos de maneira dispersa ao longo da vida acadêmica e profissional, reorganizar o mercado de trabalho e promover a produtividade.

Uma pesquisa do *Institute Data Corporation* Brasil (IDC) divulgada em 2004, demonstrou que as chances do profissional certificado conseguir um emprego aumentaram em 53% naquele ano em relação a profissionais que não possuem este título, podendo esse índice torna-se ainda mais elevado de acordo com a categoria de certificação possuída. Os salários são de 10% a 100% superiores à média que o mercado paga a profissionais sem certificação que ocupam as mesmas funções (IDC, 2004).

Para Francischini (2005), a certificação também é importante para o mercado de trabalho porque proporciona informação objetiva e oportuna sobre o candidato a emprego, facilitando e reduzindo custos do processo de recrutamento e seleção. O autor reforça ainda que, para se tornar um profissional certificado, não é obrigatório que o candidato participe de treinamentos. O indivíduo pode ser autodidata e obter sucesso nas avaliações cabíveis; entretanto, o treinamento é a forma mais rápida e segura de preparação. Além disso, sem descartar a formação acadêmica formal, uma grande alternativa para o profissional ganhar competitividade no mercado de trabalho, num curto espaço de tempo, é buscar uma certificação profissional das empresas líderes.

O reconhecimento da competência adquirida fora dos bancos escolares colabora para que a certificação profissional torne-se um instrumento de educação profissional permanente, comprometida em proporcionar ao colaborador mais autonomia e maior capacidade de gerir o seu destino profissional. Pode ainda proporcionar-lhe um amplo leque de opções e oportunidades (FRANCISCHINI, 2005).

Por outro lado, encontrar as competências internas é sempre um desafio para as instituições. Uma alternativa de mapeamento de competências é a associação deste procedimento a mecanismos de descoberta de conhecimento. Muitos destes

mecanismos estão dispostos na área de Recuperação de Informações e Extração de Conhecimento, elucidados na seção a seguir.

III - EXTRAÇÃO DE CONHECIMENTO

O mapeamento de competências está fortemente ligado à descoberta de conhecimento. Muitas informações estão dispostas na forma de dados estruturados, facilitando a aplicação de métodos de extração tradicionais ou, comumente denominados mineradores. Ainda devem ser levados em consideração os dados de textos não-estruturados que muitas vezes estão indisponíveis no formato eletrônico, pois analisá-las automaticamente nunca havia sido pensado.

Entretanto, a cada dia cresce o número de documentos armazenados eletronicamente (TAN, 1999). Esta tendência é mantida nas empresas que, cada vez mais, disponibilizam seus documentos em meio digital. Muitas consultas a esses documentos são necessárias e esta busca é freqüentemente lenta devido à variedade de temas tratados e à quantidade de locais diferentes de armazenamento. Além disso, é comum uma pesquisa retornar um conjunto considerável de documentos de interesse específico, onde apenas uma pequena parte é realmente relevante.

Tecnologias são necessárias para acelerar a análise, examinando de forma automatizada estes documentos e aferindo aquilo que é verdadeiramente significativo. É possível também, a partir da análise de um resumo de um grupo de documentos, descobrir relações importantes entre estes que antes não seriam percebidas. Uma dessas tecnologias é a Mineração de Textos (*Text Mining*).

Com a disseminação crescente do uso de computadores, cada vez mais documentos eletrônicos estão sendo armazenados e colocados à disposição das pessoas. Em sua grande maioria, estes documentos contêm informações codificadas em forma textual, tais como dicionários, manuais, enciclopédias, guias e mensagens de correio eletrônico. Alguns estudos já afirmavam, no final da década passada, que 80% da informação de uma companhia estão contidos em documentos textuais (TAN, 1999).

Encontrar tal informação é uma tarefa árdua. A evolução da área de Recuperação de Informações teve como conseqüência o surgimento da área de Descoberta de Conhecimento em Textos (*Knowledge Discovery from Text* - KDT). O termo foi utilizado pela primeira vez por Feldman & Dagan (1995) para designar o processo de encontrar algo interessante em coleções de textos (artigos, histórias de revistas e jornais, mensagens de *e-mail*, páginas *Web*, entre outros). Sinônimos como

Text Mining ou *Text Data Mining* foram introduzidos e utilizados para o mesmo fim (TAN, 1999).

Pode-se então definir *Text Mining* como sendo o processo de extrair padrões ou conhecimento, interessantes e não triviais, a partir de documentos textuais (TAN, 1999). Assim, ao invés de encontrar os textos que contenham informações e deixar que o usuário procure o que lhe interessa, esta área se preocupa em encontrar informações dentro dos textos e tratá-las de forma a apresentar ao usuário algum tipo de conhecimento útil e novo. Mesmo que o conhecimento novo não seja a resposta direta às indagações do usuário, tal conhecimento deve contribuir para satisfazer as suas necessidades de informação.

III.1 - O PROCESSO DE MINERAÇÃO DE TEXTOS

Para um melhor entendimento é interessante apresentar o esquema básico de um processo de *Text Mining* com os passos que compõem cada etapa. Os principais procedimentos são mostrados na Figura 2. Inicialmente, será apresentado o pré-processamento dos dados que prepara o conjunto de dados textuais para as fases posteriores de execução das tarefas de processamento dos dados e análise de resultados obtidos. Em seguida, será mostrado o conjunto de tarefas que podem ser realizadas a partir dos textos, mostrando em que situações podem ser utilizadas e o que se espera como saída. Depois, apresentam-se formas de avaliação da qualidade dos resultados advindos das etapas anteriores, para que os mesmos possam ser efetivamente empregados. Algumas ferramentas disponíveis atualmente serão também comentadas.

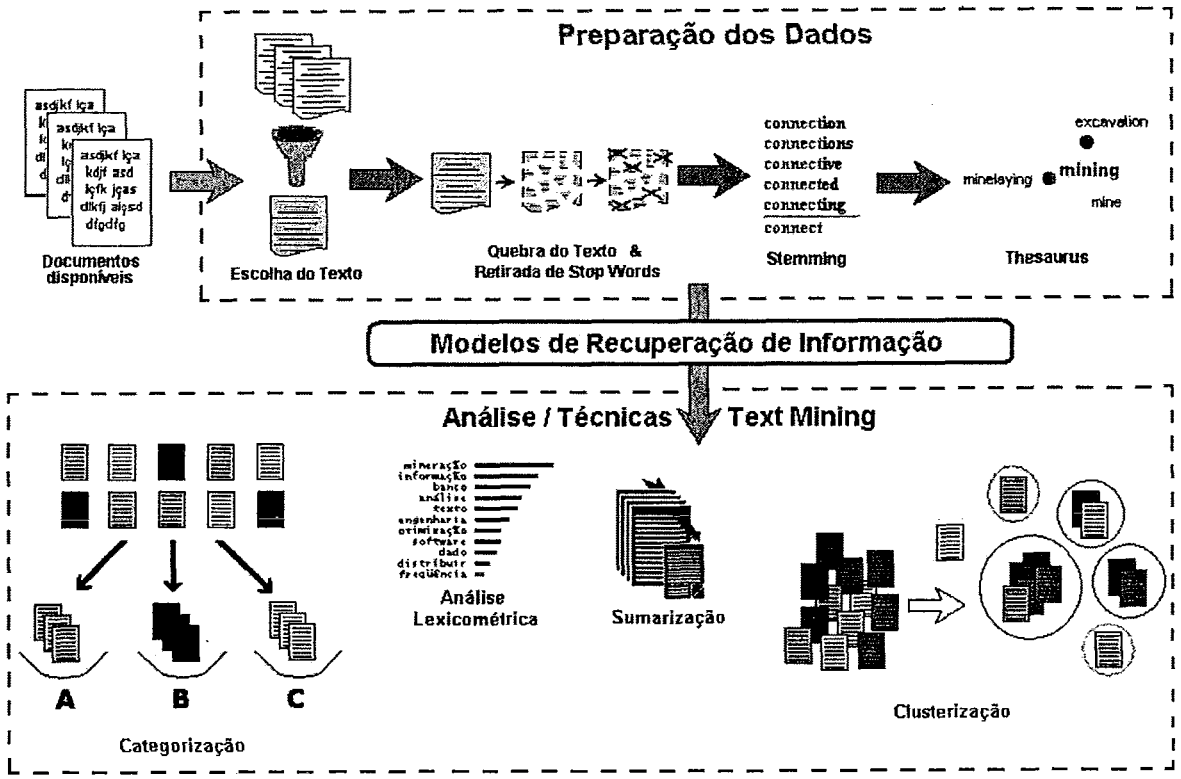


Figura 2 - Etapas de um processo de Mineração de Textos (elaborada pelo autor)

III.2 - PREPARAÇÃO DOS DADOS

A manipulação de arquivos texto requer diversos procedimentos. A primeira tarefa envolve o próprio formato dos textos, geralmente com pouco ou nenhuma estruturação. Esta falta de formatação dificulta a utilização de várias técnicas de mineração de dados. Um problema comum diz respeito ao tamanho dos arquivos, comumente da ordem de milhares de palavras ou termos. Além disso, muitas dessas palavras são irrelevantes, repetidas ou expressam o mesmo significado. Estas circunstâncias devem ser trabalhadas para aumentar a eficiência das técnicas de mineração e, consequentemente, melhorar o processo de extração de conhecimento.

A preparação dos textos é a primeira etapa neste contexto e envolve a seleção daqueles que constituirão os insumos do processo de mineração. Além disso, alguns algoritmos podem ser utilizados para reduzir a dimensão dos arquivos através da retirada de informações irrelevantes e identificar similaridades em função do significado dos termos.

III.2.1 - ESCOLHA DO TEXTO

A tarefa de extração de conhecimento começa exatamente na escolha do insumo que será utilizado. Esta etapa é bem simples, consiste na definição de cada texto que será enviado ao processo de mineração. A partir da manipulação de cada texto, outros procedimentos poderão ser executados, conforme serão apresentados nas seções seguintes.

III.2.2 - RETIRADA DE STOP WORDS

Nem todas as palavras são relevantes ao contexto de uma mineração. Termos que aparecem em todos os documentos, por exemplo, não são necessariamente passíveis de serem representantes de alguma categoria, e palavras que aparecem raras vezes, também não. Muitas destas palavras podem ser descartadas, são as chamadas *stop words*. Em geral, as *stop words* são compostas por artigos, pronomes, preposições, advérbios e outras palavras que são comuns à maioria dos textos.

No processo de preparação dos dados, após a quebra do texto em termos, são eliminadas as *stop words* com o objetivo de diminuir a estrutura do índice e melhorar o desempenho do sistema.

Para realizar esta tarefa, basta elaborar uma lista com todas as *stop words* referentes ao domínio que o sistema irá tratar. Esta lista é chamada de *stoplist* e pode ser elaborada manualmente, definindo-se as palavras que não devem aparecer no índice. Esta lista também pode ser elaborada de forma automática, a partir das palavras com grande freqüência em todos os textos, por exemplo.

O processo apresentado na Figura 3, exemplifica a etapa de preparação do texto com uma *stoplist* que contém artigos, preposições, conjunções e algumas seqüências de caracteres. As palavras que aparecem riscadas são as *stop words* presentes na *stoplist* definida no sistema (WIVES, 1997).

... ~~Na~~ maioria ~~das~~ vezes ~~os~~ documentos retornados ~~pelas~~ ferramentas ~~de~~ recuperação ~~de~~ informações envolvem ~~um~~ contexto ~~mais~~ amplo fazendo ~~com que o~~ usuário tenha ~~que~~ garimpar, ~~ou seja,~~ especificar ~~ou~~ filtrar ~~estes~~ documentos, ~~o que~~ demanda tempo e conhecimento ~~a~~ fim ~~de~~ obter ~~as~~ informações ~~que ele~~ realmente necessita.

Figura 3- Remoção de Stop Words (WIVES, 1997)

Entretanto, a remoção de *stop words* pode apresentar problemas, principalmente em sistemas que assumem palavras de maior e menor freqüência como sendo palavras irrelevantes. Segundo Riloff (1995), a remoção destes termos pode comprometer o processo de mineração na medida em que parte destas palavras pode ter influência no domínio escolhido, o que justificaria seu armazenamento para futuras análises.

III.2.3 - UTILIZAÇÃO DE ALGORITMOS DE RADICALIZAÇÃO (STEMMING)

Diferentes palavras podem possuir o mesmo radical, como por exemplo: *texto*, *textos*, *textual* e *textualmente*. Contudo, no processo de recuperação de informações, é sugestivo que a contabilização da freqüência destes termos seja feita pelo radical e não pela palavra original. Para minimizar este problema, algoritmos de radicalização (ou *stemming*) podem ser utilizados.

Esta seção aborda dois algoritmos propostos para radicalização de palavras na Língua Inglesa ou Portuguesa respectivamente. O primeiro corresponde ao mais conhecido – o algoritmo de Porter. O segundo corresponde a uma proposta de algoritmo de *stemming* para a Língua Portuguesa baseada no algoritmo de Porter.

Nas seções referentes ao desenvolvimento da ferramenta SMiner, serão apresentadas as modificações feitas no *software* para adequar ao contexto proposto nesta dissertação.

III.2.3.1 - O radicalizador de Porter (Porter Stemmer)

O radicalizador de Porter foi desenvolvido por Martin Porter na Universidade de Cambridge em 1980. O radicalizador é baseado na idéia de que sufixos da Língua Inglesa são na maioria das vezes feitos de uma combinação de sufixos menores e mais simples. Este *stemmer* tem 5 passos, aplicando regras dentro de cada passo.

Em cada passo, se uma regra de sufixo coincide com uma palavra, então as condições associadas àquela regra são testadas sobre o que seria o radical resultante caso aquele sufixo fosse removido. Uma vez que uma regra passe suas condições, o sufixo é removido e o controle vai para o próximo passo. Se a regra não é aceita, então a próxima regra no passo é testada, até uma outra regra deste passo ser aceita ou até não existirem mais regras. Este processo continua por todos os cinco passos, retornando o radical resultante após a execução do quinto passo (PORTER, 1980).

- **1°. Passo** – remove o plural, incluindo casos especiais tais como "sses" e "ies".
- **2°. Passo** – une padrões com alguns sufixos, tais como:
- "ational"->"ate" "tional" -> "tion" "enci" -> "ence"
- "anci"->"ance" "iser" ->"ize" "abli" -> "able"
- "alli"->"al" "entli" -> "ent" "eli" -> "e"
- "ousli"->"ous" "ization" -> "ize" "isation" -> "ize"
- "ation"->"ate" "ator" -> "ate" "alism" ->"al"
- "iveness"->"ive" "fullness" -> "ful" "ousness" -> "ous"
- "aliti"->"al" "iviti" -> "ive" "biliti" -> "ble"

Nessas transformações, removem-se os sufixos e substituem-nos por suas raízes.

- **3°. Passo** – ocorre a manipulação das transformações necessárias para algumas palavras especiais, como por exemplo:
- "icate" -> "ic" "ative" -> "" "alize" -> "al"
- "alise" -> "al" "iciti" -> "ic" "ical" -> "ic"
- "ful" -> "" "ness" -> ""
- **4°. Passo** – a palavra analisada é checada perante mais sufixos, considerando que palavra pode ser composta; exemplos de sufixos analisados:

"al"	"ance"	"ence"	"er"	"ic"
"able"	"ible"	"ant"	"ement"	"ment"
"ent"	"sion"	"tion"	"ou"	"ism"
"ate"	"iti"	"ous"	"ive"	"ize"

- **5°. Passo** – verifica se a palavra termina em vogal, fixando-a apropriadamente. Este procedimento serve também para analisar e separar afixos e alguns prefixos simples, tais como: "kilo", "micro", "milli", "intra", "ultra", "mega", "nano", "pico", e "pseudo".

Baseados nos procedimentos de Porter, muitos outros algoritmos foram desenvolvidos, inclusive para outras línguas.

III.2.3.2 - O Radicalizador da Língua Portuguesa (Portuguese Stemmer)

O algoritmo *Portuguese Stemmer* foi baseado no algoritmo disponibilizado por Porter (PORTER, 1997) e adaptado à realidade da Língua Portuguesa (ORENGO & HUYCK, 2001). Este algoritmo leva em conta as classes morfológicas, executando uma série de passos de remoção de sufixos conhecidos.

Cada passo tem um conjunto de regras. As regras dentro de um passo são examinadas em seqüência e somente uma regra pode ser aplicada em cada passo. O possível sufixo mais longo é sempre removido primeiro por causa da ordem das regras dentro de um dado passo. Por exemplo, o sufixo de plural -es deve ser testado antes do sufixo -s. Este algoritmo define cerca de 200 regras.

Cada regra estabelece:

- O sufixo a ser removido;
- O tamanho mínimo do radical, para evitar remover um sufixo quando o radical é muito curto.
- Um sufixo substituto para ser anexado ao radical, se aplicável;
- Uma lista de exceções, que identificam palavras que terminam no sufixo indicado, mas que não devem ser reduzidas.

A seguir, estão detalhadas as regras utilizadas.

1. Redução do Plural

A forma plural na Língua Portuguesa habitualmente termina em -s. Entretanto, nem todas as palavras terminadas em -s denotam plural, por exemplo, *lápis*, *mais* e *ademais*. Este passo consiste basicamente em remover o final -s das palavras que não estão listadas na lista de exceções. Contudo, algumas vezes ajustes são necessários, como por exemplo, palavras terminadas em -ns devem ter o sufixo substituído por -m como em bons → bom. Palavras terminadas em -ões devem ter o sufixo substituído por -ão como em ações → ação.

2. Redução do Feminino

Os substantivos e adjetivos na Língua Portuguesa possuem um gênero. Este passo consiste em transformar palavras que estão no gênero feminino em suas correspondentes no gênero masculino. Somente palavras terminadas em -a são testadas neste passo, mas nem todas são transformadas, apenas as que terminam em sufixos mais comuns, por exemplo chinesa → chinês, vilã → vilão.

3. Redução do Advérbio

Este passo trata do sufixo que denota advérbios: -mente. Nem todas as palavras com esta terminação são advérbios, por exemplo, a palavra *experimente* termina em -mente, mas não é um advérbio, por isso uma lista de exceções é necessária.

4. Redução do Aumentativo e Diminutivo

Não são todos os sufixos aumentativos que se juntam ao radical de um substantivo. Há derivações feitas sobre adjetivos (ricaço, de rico; sabichão, de sábio) e também sobre radicais verbais (chorão, de chorar; mandão, de mandar).

As palavras têm aumentativo, diminutivo e formas superlativas, por exemplo, casinha = "casa pequena", onde - inha é o sufixo que indica um diminutivo. Esses casos são tratados neste passo.

5. Redução de Formas Nominais

Este passo testa as palavras contra terminações de substantivos e adjetivos. Por exemplo, palavras terminadas com o sufixo -ista devem ter este sufixo removido como acontece em realista → real, do mesmo modo que palavras terminadas com o sufixo -ismo devem ter este sufixo removido como em realismo → real. Observe que são duas palavras diferentes que conservam uma relação de sentido com o mesmo radical. Se um sufixo é removido neste passo, os passos de redução de verbos e redução da vogal temática não são executados.

6. Redução das Terminações Verbais

Enquanto os verbos regulares da Língua Inglesa possuem quatro variações (talk, talks, talked, talking), os verbos regulares da Língua Portuguesa possuem mais de 50 formas diferentes (CUNHA & CINTRA, 2001). Os verbos apresentam as variações de número, de pessoa, de modo, de tempo, de aspecto e de voz.

Foi verificado que as estruturas das formas verbais se relacionam pelo radical. Além disso, esse radical verbal se junta, em cada forma, a uma terminação, da qual

participa pelo menos um dos elementos. É justamente neste passo que as formas verbais são reduzidas ao seu radical.

7. Redução da Vogal Temática

Neste passo, o algoritmo remove a última vogal das palavras que não sofreram o processo de radicalização nos passos de redução de formas nominais e de terminações verbais.

8. Remoção dos Acentos

Remover acentos é necessário porque existem casos em que algumas formas diferentes de palavras são acentuadas e algumas não, como por exemplo, psicólogo e psicologia. Depois desse passo, estas palavras assumiriam o radical *psicolog*.

III.2.3.3 - Comparações entre Stemming Português e Inglês

Encontrar radicais em Português é mais complexo do que em Inglês. Basicamente, as diferenças entre os algoritmos são fundamentadas pela disparidade morfológica entre as línguas portuguesa e inglesa. Algumas particularidades podem ser destacadas:

- **Exceções:** Talvez a maior dificuldade de implementação do algoritmo de *Stemming* para Português seja o número de exceções que devem ser tratadas. A diferença neste contexto entre este algoritmo e o de Porter é presença da lista de exceções, dada a simplicidade de formação das palavras inglesas.

- **Homônimos**[1]**:** Há vários exemplos que podem ser citados, principalmente envolvendo conjugação de verbos e plural como, por exemplo, a palavra *"atuais"*, que pode ser a 2ª pessoa do plural do verbo *"atuar"* no presente ou o plural da palavra *"atual"*. Neste caso, o algoritmo irá tratar desta palavra logo na 1ª regra e substituirá o sufixo *"-ais"* por *"al"*, retornando o radical *"atual"*. No algoritmo de Porter isto ocorre da mesma forma.

- **Nomes próprios:** Nomes próprios não deveriam ser modificados, contudo, assim como o algoritmo de Porter, neste sistema estas palavras são verificadas. A princípio, se poderia imaginar uma lista de exceções contendo nomes próprios.

[1] Homônimos são palavras com escrita idêntica, porém com diferentes significados.

Entretanto, existem sobrenomes que se confundem a substantivos, por exemplo, *madeira* substantivo é homônima de *Madeira* sobrenome.

III.2.4 - *THESAURUS*

Como forma de minimizar os problemas de homonímia e sinonímia no vocabulário dos textos a serem minerados, é interessante integrar um *thesaurus* neste processo.

Um *thesaurus* é um conjunto de termos, palavras ou frases com relações entre elas. Os principais objetivos do uso de *thesaurus* concentram-se na substituição dos termos com o mesmo significado pelo termo procurado ("andar" → "caminhar"), ou ainda substituir hierarquias de conceitos tais como substituições de generalização (MILLER, 1996) onde todos os termos são generalizados para os termos adequados de mais alto nível de acordo com a hierarquia de conceitos descritas no *thesaurus* ("futebol" → "esporte").

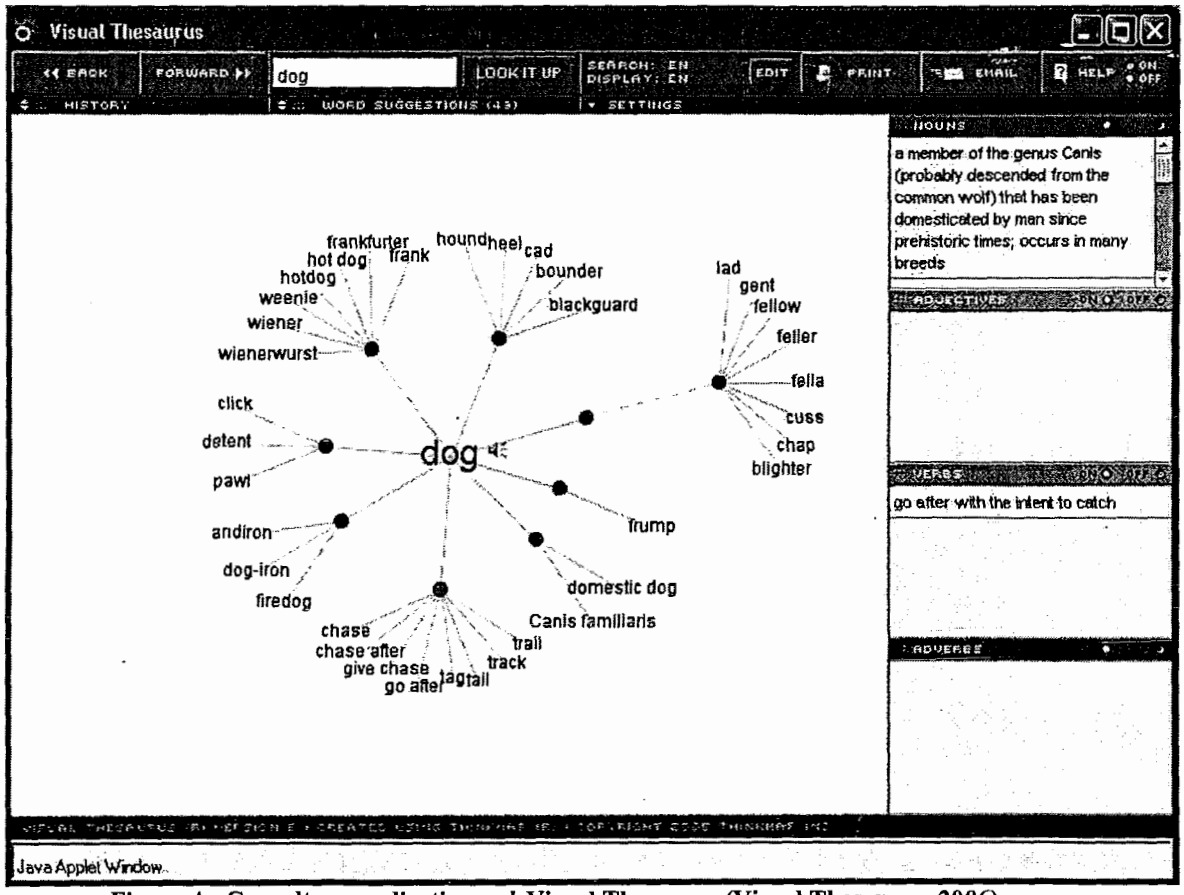


Figura 4 - Consulta no aplicativo *web* Visual Thesaurus (Visual Thesaurus, 2006)

A Figura 4 mostra o aplicativo Visual Thesaurus (2006) com o termo *dog* sendo aplicado à ferramenta. Nas folhas da árvore de termos, os possíveis significados encontrados.

Embora o uso do *thesaurus* tenha ajudado com problemas de sinonímia, este método ainda sofre de homonímia.

III.3 - Modelos de Recuperação de Informação

A área de Recuperação de Informação (*Information Retrieval*) desenvolveu modelos para a representação de grandes coleções de textos que identificam documentos sobre tópicos específicos. Embora esse seja um campo bastante extenso, neste trabalho o interesse se restringe à representação e identificação de documentos sobre conjuntos de assuntos específicos.

Desta forma, um sistema de Recuperação de Informações atua como se fosse um filtro sobre um conjunto de documentos, retornando ao usuário o resultado de um problema ou consulta particular. Esta seção dá uma visão geral de métodos usados.

III.3.1 - Coincidência de Palavras

O usuário especifica sua necessidade de informação através de um conjunto de palavras. Alguns documentos poderiam satisfazer a busca na medida em que a cadeia especificada pelo usuário exista nos documentos. Este método é um dos mais simples procedimentos, entretanto, sofre de três problemas (CROFT & DAS, 1990):

- Homonímia: o significado de uma palavra depende do contexto no qual ela aparece;
- Sinonímia: palavras tendo o mesmo significado;
- Tempo de resposta normalmente ruim.

III.3.2 - Modelo Booleano

O modelo *booleano* é um dos modelos clássicos que considera uma consulta como uma expressão booleana convencional, que liga seus termos através de conectivos lógicos AND, OR e NOT. Neste modelo, um documento é considerado relevante ou não relevante a uma consulta, não existe resultado parcial e não há informação que permita a ordenação do resultado da consulta (BELKIN & CROFT, 1992).

Exemplo: Numa consulta com 3 termos t1, t2 e t3, onde t significa estar contido enquanto t' significa não estar contido, as possibilidades de ocorrência destes termos em documentos, pertence a uma das seguintes opções:

$m_1 = t_1\, t_2\, t_3$ $\qquad$ $m_5 = t_1\,'t_2\,'t_3$

$m_2 = t_1\,'t_2 t_3$ $\qquad$ $m_6 = t_1 t_2\,'t_3\,'$

$m_3 = t_1 t_2\,'t_3$ $\qquad$ $m_7 = t_1\,'t_2 t_3\,'$

$m_4 = t_1 t_2 t_3\,'$ $\qquad$ $m_8 = t_1\,'t_2\,'t_3\,'$

Onde m1..m8 são min-termos, conjuntos que descrevem todas as possibilidades para o conjunto resposta da consulta. Com 4 termos isso fica bem mais complicado, e assim por diante. O número de min-termos cresce exponencialmente (nº de min-termos = 2^n, onde n é o número de termos da consulta).

As consultas são construídas como uma combinação dos min-termos. Geralmente, para n termos, temos $k=2^n$ min-termos e 2^k consultas. Além disso, o tamanho da base de dados afeta tanto as estratégias de consultas quanto os resultados obtidos usando o método *booleano*.

Vantagens do modelo *booleano*:

- Expressividade completa se o usuário souber exatamente o que quer.
- É facilmente programável e exato.

Desvantagens do modelo *booleano*:

- A formulação de uma consulta adequada, isto é, a seleção dos termos para a consulta é difícil, especialmente se o domínio não é bem conhecido.
- O tamanho da saída não pode ser controlado. O conjunto resultante tanto pode conter nenhum como milhares de itens. Além disso, sem um grau de comparação parcial, não se pode saber o que foi deixado de fora da definição da consulta.
- Uma vez que não há um grau de comparação, não é possível ordenar os resultados de acordo com a relevância.

III.3.3 - Modelo Espaço Vetorial

Ao examinar estratégias para Recuperação de Informação, Bates (1986) concluiu que, para obter sucesso, o usuário que procura informação deve usar uma variedade de termos tão grande quanto à variedade produzida no momento da indexação. Este tipo de redundância permite identificar termos comuns usados pelo autor ou indexador e pelo

usuário, no momento de expressar idéias e conceitos. Assim, a eficiência na identificação dos conceitos é maior porque mais termos foram cobertos.

Desta forma, um conjunto suficiente de termos ou palavras deve ser utilizado para representar cada conceito. Neste contexto, os termos descritores de um conceito podem incluir sinônimos, quase-sinônimos (palavras semanticamente relacionadas), variações léxicas (conjugações verbais, verbos e substantivos correlatos, variações em grau e gênero) e outros. Os termos funcionam como "tokens", então não é necessário que o termo tenha um significado universal. Assim, podem ser usados nomes próprios, abreviações e siglas específicas do domínio.

No método de similaridade de vetores, as classes são representadas por vetores (conjuntos) de palavras (denominados centróides). O documento é comparado com o vetor descritivo de cada classe. A classe que apresentar maior similaridade com o documento é tomada como classe do documento (LOH *et al*, 2000).

No *Vector Space Model* ou modelo espaço de vetores, cada conceito é representado por um vetor de termos simples, ilustrado na Figura 5. Neste caso, não há relação direta entre os termos e todos são considerados do mesmo nível (vetor não-ordenado e sem conexões entre os termos). A razão desta escolha é que este modelo é o mais simples e facilita as tarefas de definição e identificação dos conceitos. Associado ao termo, pode haver um peso no vetor, descrevendo o grau de importância do termo para descrever ou identificar o conceito. De acordo com Chakrabarti (1993), o vetor com pesos é melhor que o modelo binário (sem pesos) porque aumenta a precisão.

Os pesos devem ser normalizados para uma escala entre um e zero, para indicar a força relativa do termo descritor. Por exemplo, para representar o conceito "futebol", o termo "futebol" pode receber um grau maior que "jogador", uma vez que a presença deste termo indica fortemente a presença do conceito "futebol". Já o segundo termo pode aparecer em outros conceitos semelhantes, como "vôlei" e "basquete" e, portanto, deve receber um peso menor. Feldman & Dagan (1995) defendem o uso de estruturas simples porque permitem que as tarefas sejam apoiadas por ferramentas automatizadas e porque geram menos esforço. Entretanto, o problema do modelo espaço de vetores é que o contexto dos termos não é analisado e isto pode levar a uma interpretação errada. Por exemplo, o termo "não" pode alterar completamente o significado de uma expressão.

Cada conceito deve ter somente um conjunto de descritores, mas um termo pode aparecer em mais de um conceito. No momento, somente termos simples são permitidos nesta proposta. Entretanto, sabe-se que o uso de pares de termos e expressões complexas melhora os métodos. O uso de termos simples não deve influenciar demais nos resultados, pois o uso exclusivo de termos simples é relativamente eficiente em contrapartida ao uso exclusivamente de pares de termos que implica em resultados mais pobres (APTÉ *et al*, 1994).

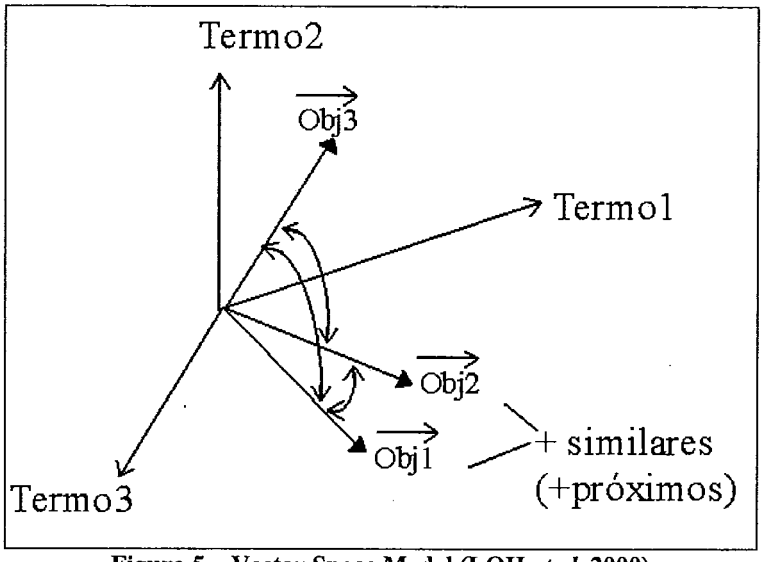


Figura 5 – Vector Space Model (LOH *et al*, 2000)

Tornando o esquema mais intuitivo e denominando o vetor de documentos como D e o vetor da consulta como Q, para mensurar a relevância entre D e Q, pode-se utilizar a distância euclidiana entre D e Q, onde:

$SC(Q, Di) = Q \cdot DI$

Adicionalmente, o coseno do ângulo entre Q e D é freqüente usado:

$SC(Q, Di) = Q \cdot DI / ||DI||$.

O mais difícil neste modelo é identificar os componentes dos vetores.

III.3.4 - Modelo Vetor de Contexto

O modelo vetor de contexto é uma extensão do modelo espaço vetorial apresentado no tópico anterior. Os vetores dos sistemas de recuperação de informação baseados no modelo espaço vetorial são compostos apenas da listagem dos termos e de

suas freqüências. Estes termos são considerados no momento de se medir a similaridade com outros vetores, impondo assim limites ao resultado, pois não consideram os relacionamentos semânticos que porventura existirem entre os termos informados e os outros termos existentes nos documentos.

Isso significa que documentos relevantes à consulta em questão poderão ficar de fora do resultado (BILLHARDT *et al.*, 2002). Assim sendo, o modelo espaço vetorial foi expandido para agregar aos termos do vetor, que representa um documento, o contexto em que ele está inserido, ou seja, outros termos que podem indicar alguma relação semântica.

Shütze (1992) considera o significado semântico dos termos e seus contextos como vetores em um espaço vetorial em que as dimensões correspondem aos termos. Tais vetores são denominados "vetores de contexto" por possuírem o contexto onde cada termo está inserido no documento. Esse contexto é obtido através dos termos próximos, dentro do texto, considerando uma janela de termos que indica quantos deles antes e depois serão considerados na definição do contexto. Assim, para cada termo pode ser gerado um vetor contendo os termos próximos no contexto, e o vetor de contexto de um documento é formado por esses pequenos vetores de contexto relacionados a cada termo (CAID; CARLETON,1994).

III.3.5 - MODELO PROBABILÍSTICO

O modelo probabilístico possui esta denominação justamente por trabalhar com conceitos provenientes da área de probabilidade e estatística. Neste modelo os termos indexados dos documentos e das consultas não possuem pesos pré-definidos. A ordenação dos documentos é calculada pesando dinamicamente os termos da consulta relativamente aos documentos (RIJSBERGEN, 1999).

É baseado no princípio da ordenação probabilística (*Probability Ranking Principle*). Neste modelo, busca-se saber a probabilidade de um documento D ser ou não relevante para uma consulta Q_a. Tal informação pode ser obtida assumindo-se que a distribuição de termos na coleção seja capaz de informar a relevância provável para um documento qualquer da coleção.

Princípio da Ordenação Probabilística:

- $+R_a \rightarrow$ o documento é relevante para a consulta Q_a

- $-R_a$ → o documento não é relevante para a consulta Q_a
- D → documento
- $P(+R_a / D)$→ probabilidade que o documento D seja relevante para a consulta Q_a
- $P(-R_a / D)$→ probabilidade que o documento D não seja relevante para a consulta Q_a

Assumindo que a relevância de um documento é independente da relevância de todos os outros (isso não é verdade), um documento D é relevante a uma consulta Q_a quando: $P(+R_a /D) > P(-R_a /D)$.

Assim, dada uma consulta Q_a, o modelo probabilístico atribui a cada documento D (como medida de similaridade) um peso $W_{D/Qa}$, como sendo:

$$W_{D/Q_a} = \frac{P(+R_a / D)}{P(-R_a / D)}$$

Essa fórmula calcula a probabilidade de observação aleatória de D que pode ser tanto relevante quanto irrelevante. A teoria de Bayes auxilia a identificar para cada termo da consulta o grau de relevância e de irrelevância do documento, selecionando o mais adequado (o que produz menor erro) para o somatório final, já que o grau final de probabilidade de relevância é dado pelo somatório dos graus de relevância de cada termo.

Assim, aplicando a regra de Bayes:

$$W_{D/Q_a} = \frac{P(D/+R_a) \times P(+R_a)}{P(D/-R_a) \times P(-R_a)}$$

Onde:

- $P(D / +R_a)$ → probabilidade que, dado um documento relevante para Q_a, este seja D
- $P(D / -R_a)$ → probabilidade que, dado um documento não relevante para Q_a, este seja D
- $P(+R_a)$ → probabilidade que um documento é relevante
- $P(-R_a)$ → probabilidade que um documento não é relevante

Como os termos indexados nos documentos são apenas presentes ou não presentes, para calcular $P(D /+R_a)$ e $P(D / -R_a)$ o documento pode ser representado pelo

vetor: $D=\{x_1, x_2, ..., x_n\}$, $x_k \in \{0,1\}$. Ou seja, o peso para o termo indexado x_1 pertence ao conjunto $\{0,1\}$. Colocando isso na fórmula, reescreve-se:

$$P(D/+R_a) = \prod_{k=1}^{n} P(x_k/+R_a)$$

Onde:

- $P(x_k/+R_a)$ → probabilidade que, dado que o documento D, este é relevante para a consulta Q_a, se o evento descrito em x_k (presença ou ausência do termo k no documento D) ocorre.

Seja:

- $r_{ak}=P(x_k=1/+R_a)$ → probabilidade que, dado que o documento D, este é relevante para a consulta Q_a, se o termo k está presente em D.

Então a fórmula pode ser reescrita da seguinte forma:

$$P(D/+R_a) = \prod_{k=1}^{n} r_{ak}^{xk} (1-r_{ak})^{1-xk}$$

Analogamente, pode-se derivar a uma expressão similar para $P(D/-R_a)$, seguindo os mesmos passos, onde:

- $s_{ak}=P(x_k=1/-R_a)$ → probabilidade que, dado que o documento D, este não é relevante para a consulta Q_a, se o termo k está presente em D.

Conclui-se que:

$$P(D/-R_a) = \prod_{k=1}^{n} s_{ak}^{xk} (1-s_{ak})^{1-xk}$$

Substituindo as duas últimas expressões na primeira (regra de Bayes) e tomando os logs, podemos recalcular os pesos da seguinte forma:

$$w_{D/Q_a} = \sum_{k=1}^{n} x_k \times w_{ak} + C$$

$$x_k \in \{0,1\}$$

$$w_{ak} = \log \frac{r_{ak}}{1-r_{ak}} + \log \frac{1-s_{ak}}{s_{ak}}$$

$$C = \log \frac{P(+R_a)}{P(-R_a)} + \sum_{k=1}^{n} \log \frac{1-r_{ak}}{1-s_{ak}}$$

Desta forma, percebe-se que, para avaliar um documento, é preciso simplesmente avaliar os pesos para os termos da consulta (w_{ak}), que também estão

presentes nos documentos (x_k=1). A constante C, que é a mesma para qualquer documento, vai variar de consulta para consulta, mas pode ser interpretada como o valor de corte para a função de recuperação. Por esta razão, a equação final pode ser escrita simplesmente na forma:

$$sim(D,Q_a) = W_{D/Q_a} = \sum_{k=1}^{n} x_k \times w_{ak}$$

$W_{D/Qa}$ é a medida de similaridade entre a consulta Q_a e o documento D. Note que w_{ak} é o peso para o termo k da consulta, enquanto x_k é o peso para o termo k no documento. Uma vez que o valor de x_k é binário ($x_k \in \{0, 1\}$), pode-se dizer que o modelo probabilístico não atribui pesos aos termos nos documentos, ou seja, o modelo ordena os documentos apenas pela medida dos pesos dos termos da consulta (w_{ak}) (BAEZA-YATES; RIBEIRO, 1999).

Vantagens do Modelo *Probabilístico*:

- Por obrigar o Princípio da Ordenação Probabilística, o modelo comporta-se otimamente (os documentos são ordenados de forma decrescente por suas probabilidades de serem relevantes).
- Algumas evidências parecem indicar que este modelo tem um desempenho melhor que o do modelo vetorial.

Desvantagens do Modelo *Probabilístico*:

- Assume a independência entre os termos.
- Não há como calcular r_{ak} (dado um documento este é relevante para a consulta se o termo está presente) ao iniciar a execução do sistema (uma vez que os documentos ainda não são conhecidos).
- O modelo não faz uso da freqüência dos termos no documento.

III.4 - ANÁLISE E PROCESSAMENTO DOS DADOS

Independente da área de aplicação, os métodos e técnicas de mineração de textos podem ser os mesmos. Entretanto, em virtude do tema proposto, nesta seção são destacados alguns métodos e técnicas encontrados nas diversas ferramentas de *Text Mining* que podem auxiliar na descoberta de conhecimento e competências.

III.4.1 - ANÁLISE LEXICOMÉTRICA

A análise lexicométrica é uma das técnicas de descoberta de conhecimento mais simples que existe. Consiste na identificação da freqüência de palavras (características) presentes nos documentos. Esse tipo de análise serve para que o usuário identifique o conteúdo tratado em um documento ou conjunto de documentos. A listagem de palavras por ordem de freqüência (partindo da mais freqüente para a menos freqüente) permite a identificação das palavras mais relevantes de um documento e, conseqüentemente, seu conteúdo ou assunto mais importante.

Com essa análise também é possível (rapidamente) identificar novos termos ou palavras que, eventualmente, apareçam nas listagens. Do mesmo modo, são identificados os centros de interesse, tópicos mais relevantes e os objetos envolvidos (pessoas, institutos, países).

Aplicando-se essa técnica em diferentes conjuntos de documentos relativos a períodos ou épocas diferentes é possível realizar uma análise de tendências, identificando, por exemplo, que determinado pesquisador está abordando um determinado tema ou que determinada tecnologia está passando a ser utilizada em alguma área ou ramo de atividade especifica.

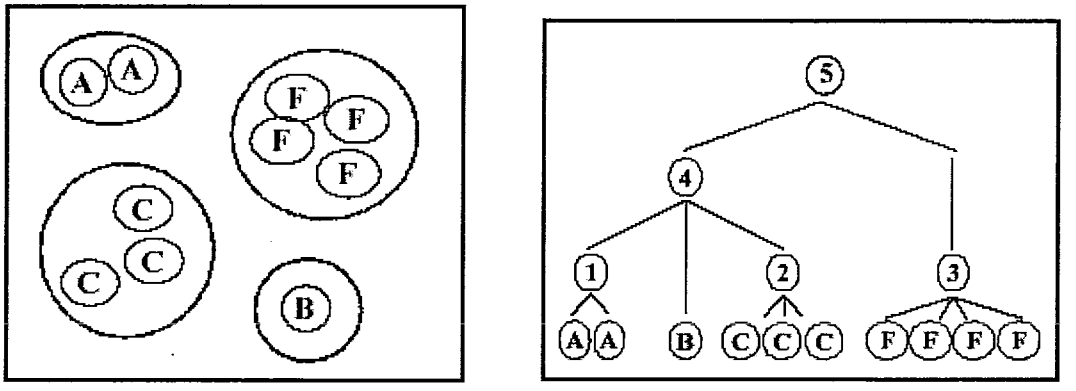
Praticamente todas as ferramentas de *Text Mining* utilizam análise lexicométrica. Ainda que esta análise não seja oferecida ao usuário final, internamente é utilizada como base para a aplicação de outros métodos mais complexos.

III.4.2 - CLUSTERIZAÇÃO

O modelo de clusterização (*clustering model*) (FRAKES, 1992) utiliza técnicas de agrupamento de documentos. A idéia consiste em identificar documentos que tratem de assuntos similares e, em seguida, armazená-los ou indexá-los em um mesmo grupo. Esta identificação é feita pela quantidade de palavras similares e freqüentes que os textos possuem.

Esta técnica é geralmente utilizada antes de um processo de classificação, facilitando a definição de classes, pois o especialista pode analisar os co-relacionamentos entre os elementos de uma coleção de documentos e identificar a melhor distribuição de classes para os objetos em questão. Isso significa que não há a necessidade de se ter conhecimento prévio sobre os assuntos dos documentos ou do

contexto dos documentos. Os assuntos e as classes dos documentos são descobertos automaticamente pelo processo de agrupamento.



Partição disjunta Partição hierárquica

Figura 6 - Tipos de Agrupamento (WIVES, 2004)

O *clustering* pode gerar topologias de grupos *isolados* ou *hierárquicos*, como mostra a Figura 6. No primeiro caso, um algoritmo de *partição* é aplicado à coleção de documentos e estes são colocados em grupos distintos, geralmente não havendo espécie alguma de relacionamento entre os grupos identificados. Já no segundo, pode haver algum relacionamento ou ligação entre grupos. Nesse caso, o processo de identificação de grupos é aplicado recursivamente e acaba gerando uma espécie de árvore onde as folhas representam os grupos mais específicos e os nodos intermediários representam grupos mais abrangentes (WIVES, 2004).

Cada uma destas topologias possui suas vantagens e desvantagens. No primeiro caso, disjunto, não há estruturas que indiquem o co-relacionamento entre os grupos, impossibilitando o usuário de identificar os assuntos mais específicos e os mais abrangentes. Esse problema é solucionado pelo segundo caso, que oferece estruturas de navegação hierárquica entre os grupos, facilitando a localização de informações. Essa vantagem exige um tempo de processamento maior, já que o algoritmo de *clustering* deve passar a analisar os grupos identificados várias vezes, tornando-se uma desvantagem. Outra desvantagem do método hierárquico diz respeito à manutenção dos *clusters*, que é mais complexa (KOWALSKI, 1997).

Em geral, o *clustering,* hierárquico ou não, possui diversas aplicações. O *clustering* pode ser utilizado para facilitar a organização e a recuperação de informações

ou em outros processos de análise textual que visam descobrir o conhecimento a partir de textos.

A recuperação de informações é facilitada porque o método desenvolvido consegue processar uma grande quantidade de documentos (de assuntos diversos) e agrupá-los em *clusters* de documentos de assuntos similares. Os grupos de documentos similares são armazenados em um mesmo local no arquivo de dados e indexados de forma que todo um *cluster* seja recuperado quando um dos documentos que fazem parte dele for considerado relevante a uma consulta.

Em relação à área de descoberta de conhecimento em textos, o agrupamento é comumente utilizado no processo de descoberta de associações entre palavras, facilitando o desenvolvimento de dicionários e *thesaurus*. Esses dicionários podem ser utilizados em ferramentas de busca ou editoração de documentos, expandindo consultas ou padronizando o vocabulário dos documentos em edição.

Os grupos identificados também podem ser utilizados em alguns processos de identificação de características relevantes (espécie de sumarização), capazes de identificar o padrão e, em diferentes períodos de tempo, as tendências dos grupos (ou seja, as características que mudam com o decorrer do tempo).

III.4.3 - CLASSIFICAÇÃO

A classificação é um processo de aprendizado em que um objeto é mapeado em uma das classes pré-definidas (FAYYAD *et al*, 1996). A partir de um conjunto de atributos previamente escolhidos, o algoritmo de classificação procura estabelecer relações entre os dados, classificando os registros de acordo com as características de cada um, confrontando-os com as características das classes previamente determinadas.

A classificação pode, então, identificar a qual classe este objeto pertence, a partir de seu conteúdo. Para tal, é necessário que as classes tenham sido previamente descritas, expressando suas características por meio de definições, fórmulas e/ou atributos.

Os sistemas de classificação de objetos geralmente utilizam uma das seguintes técnicas (WIVES, 2000).

- **Regras de inferência** – baseadas em um conjunto de características que devem ser encontradas no objeto para que esse seja identificado como pertencendo a uma

determinada categoria. Necessitam de muito tempo para serem elaboradas e devem ser adaptadas caso o domínio mude. Geralmente são desenvolvidas para uma tarefa e domínio específicos. O conhecimento modelado é facilmente compreendido e seus resultados são, na maioria dos casos, melhores do que os apresentados pelos outros métodos.

- **Modelos conexionistas** – esses sistemas induzem automaticamente um modelo matemático ou um conjunto de regras a partir de um conjunto de objetos de treinamento. Podem ser colocados em prática rapidamente e são capazes de se adaptar as mudanças do ambiente de dados. Os modelos conexionistas não necessitam de um especialista ou pessoa para a análise do domínio. Por outro lado, necessitam do conjunto de treinamento e seu modelo ou regras não são tão facilmente compreensíveis.
- **Método de similaridade de vetores ou de centróides** – nesse caso, as classes são representadas por vetores de palavras, denominadas centróides. O documento é comparado com o vetor descritivo de cada classe. A classe que apresentar maior similaridade com o documento é tomada como classe do documento.
- **Árvores de decisão** – uma abordagem parecida com a primeira, porém, utiliza técnicas de aprendizado de máquina para induzir as regras. Para cada classe, uma árvore é criada.
- **Classificadores de Bayes** – semelhantes aos conexionistas, porém possuem como base a teoria de probabilidades. Os classificadores de Bayes conseguem informar a probabilidade de determinado objeto pertencer a uma determinada classe.

Um exemplo de classificação: uma empresa que atua no setor de concessão de crédito pretende avaliar o risco associado a empréstimos que realiza aos seus clientes. Os clientes são classificados como *bons* ou *indesejáveis* dependendo se o crédito é recuperado ou não pela empresa. A partir de informações encontradas em cada registro, pode-se utilizar um algoritmo de classificação para testar os valores destas variáveis, e classificar desta forma, os clientes em uma das duas possibilidades – *bons* ou *indesejáveis*.

III.4.4 - SUMARIZAÇÃO

Sumarização é uma técnica que identifica as palavras e frases mais importantes de textos, gerando um resumo ou sumário. Este procedimento permite uma visualização geral do documento. O objetivo é separar as partes mais importantes e interessantes, possibilitando a identificação rápida do assunto abordado.

Existem diversas abordagens para esta técnica, uma delas é apresentada por Miike (1994), que consegue gerar resumos em tempo de execução através de interações com o usuário. O tamanho do resumo e as partes que vão compô-lo podem ser definidos pelo usuário, dependendo do seu interesse.

A análise do texto é feita sobre sua organização (seções, parágrafos, títulos, sub-títulos), sobre as sentenças que o compõem (análise morfológica e sintática com uso de um dicionário), sobre a estrutura do texto (conectivos lógicos, expressões idiomáticas entre parágrafos e frases) e com a extração de papéis ou funções semânticas por *tags* (para cada função, são definidos *tags* específicos; por exemplo, para achar tópicos, procurar pelo *tag* "... é explicado").

McKeown e Radev (1995) apresentam técnicas e ferramentas para analisar diversos artigos sobre um mesmo evento e criar um resumo em linguagem natural. São extraídas informações de partes dos textos (por técnicas tradicionais de extração de informações), as quais são estruturadas em *slots* (pares atributo-valor, representando internamente conceitos).

A saída em linguagem natural é gerada em formatos padrões preenchidos com os *slots*. Entretanto, ao invés de unir simplesmente as frases, são utilizados conectivos lógicos e palavra-chave (termos lingüísticos) para formar resumos mais complexos.

As representações dos textos (em *slots*) são analisadas para encontrar similaridades e diferenças de informações. Então, para combinar as informações extraídas de artigos diferentes, são aplicados operadores semânticos sobre os *slots* para exprimir, por exemplo, contradição, adição, refinamento de informação, concordância e falta de informação.

Tais operadores também decidem quais dados serão incluídos no resumo final, com base em graus de importância, determinados segundo critérios tais como: informações que aparecem em mais artigos tem maior grau. A ordem de apresentação

das frases também é definida automaticamente, respeitando as restrições de espaço definidas pelo usuário da ferramenta (algumas frases podem não ser apresentadas).

III.5 - Ferramentas de Extração de Conhecimento

É interessante notar a ascensão de aplicativos que utilizam técnicas de mineração em textos. Desta forma, serão destacados a seguir, alguns sistemas que utilizam parcialmente ou integralmente este tipo de tecnologia.

III.5.1 - SAS Text Miner

SAS se intitula um dos líderes em *softwares* para Business Intelligence. A empresa, com 30 anos de experiência tem como um dos principais produtos o SAS® Enterprise Miner™, que é uma solução *data mining* atualmente na versão 5.2 (SAS TECHNOLOGIES, 2006).

Além disso, ainda na linha de mineração, SAS possui o SAS Text Miner®, que é composto por um conjunto de ferramentas que permite a descoberta e extração de conhecimento a partir de documentos textuais (Figura 7). Os recursos de mineração de textos possibilitam a classificação de documentos em categorias pré-definidas, a descoberta de relacionamentos ou associações entre documentos, clusterização de documentos, além de integrar informações com base em texto com dados estruturados para realizar um processo de *data mining* aperfeiçoado.

O SAS Text Miner® se promove como a primeira solução comercial que efetivamente integra informações baseadas em texto com dados estruturados e, consequentemente, provê relatórios qualitativamente analíticos que suportam a tomada de decisão (SAS TECHNOLOGIES, 2003).

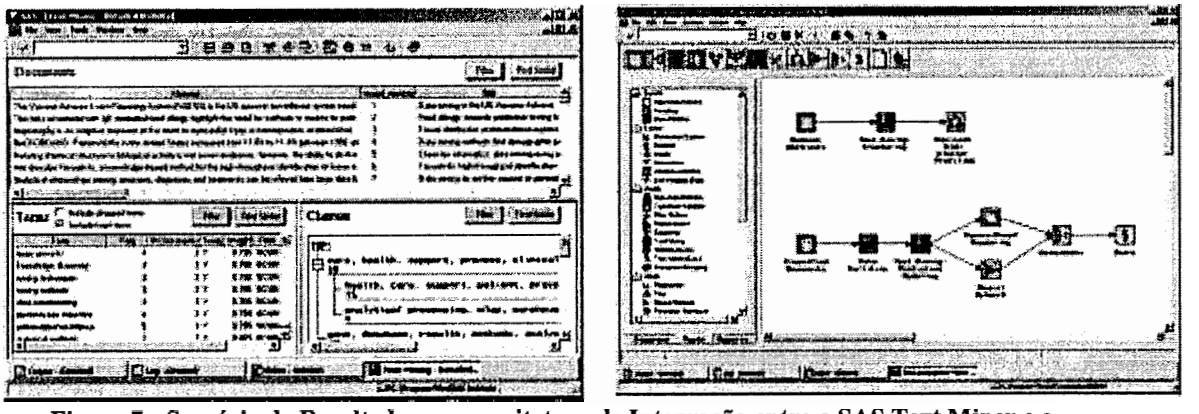


Figura 7 - Sumário de Resultados e a arquitetura de Integração entre o SAS Text Miner e o Enterprise Miner (SAS TECHNOLOGIES, 2006)

No mundo, os casos de sucesso do SAS Text Miner destacam-se em empresas como a HP-Hewlett Packard e seu CRM com SAS Text Miner (SAS TECHNOLOGIES, 2005), a American Honda com a análise de satisfação apurada (SAS TECHNOLOGIES, 2005), a Universidade de Louisville com a mineração e pesquisas em documentos hospitalares (CERRITO, 2004), entre outros.

No Brasil, a SAS apresenta Vivo, Cemig, Losango e ABN-AMRO como grandes utilizadores das ferramentas de análise analítica (SAS TECHNOLOGIES, 2005). Além disso, no meio acadêmico brasileiro, o SAS destaca o auxílio às análises analíticas na UFRJ, onde o departamento de geologia usa esta tecnologia em aplicações de técnicas estatísticas, na análise de dados geológicos e na criação de modelos de testes (SAS TECHNOLOGIES, 2005).

Segundo relatório técnico (SAS TECHNOLOGIES, 2003), a ferramenta SAS Text Miner® permite:

- Integração da mineração de textos com os processos de *data mining*;
- Aceita como insumo textos em PDF, textos em formato ASCII, HTML, Word, entre outros;
- Suporta e identifica textos em Inglês, Francês e Alemão;
- Possui lista de Stop words customizada;
- Realiza stemming;
- Análise de contexto;
- Identificação de substantivos;
- Lista de sinônimos;
- Desmembramento e composição de palavras em subtermos (especialmente para Alemão);
- Clusterização com algoritmos próprios;
- Categorização de documentos;
- Relatórios interativos.

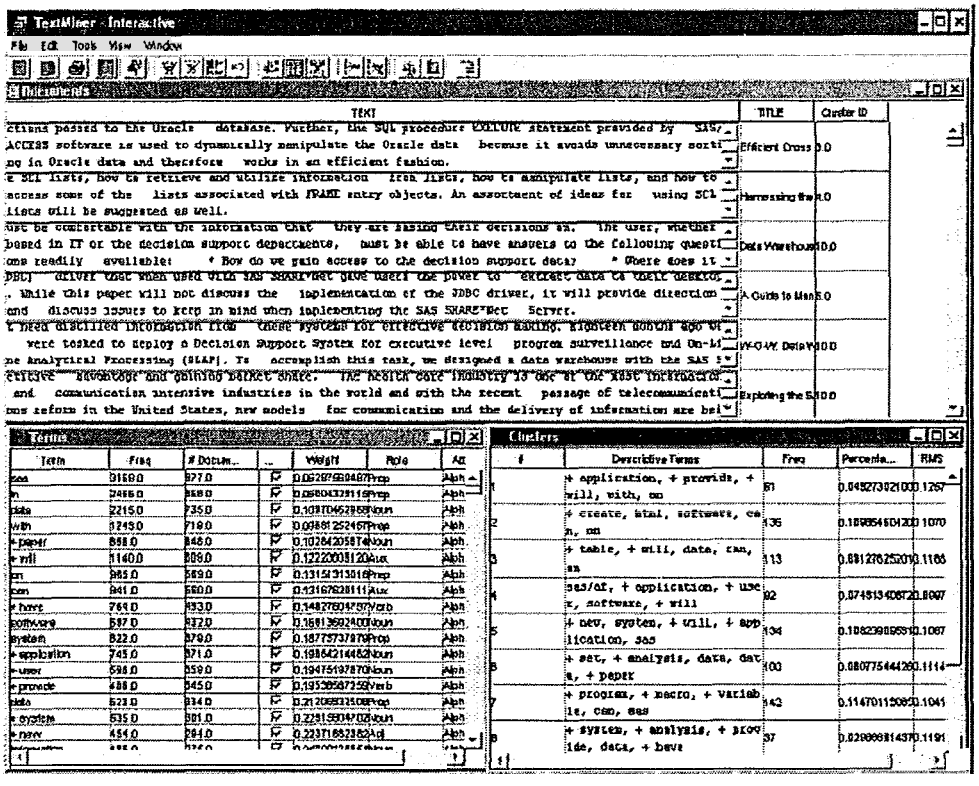


Figura 8 – Relações entre conceitos e documentos no Navegador de Resultados Interativos do SAS (SAS TECHNOLOGIES, 2006)

A Figura 8 mostra a ferramenta de análise interativa, apresentando os termos extraídos do texto bem como sua freqüência, peso, tipo (preposição e substantivo, por exemplo). A Figura 9 exibe a visualização do recurso de clusterização.

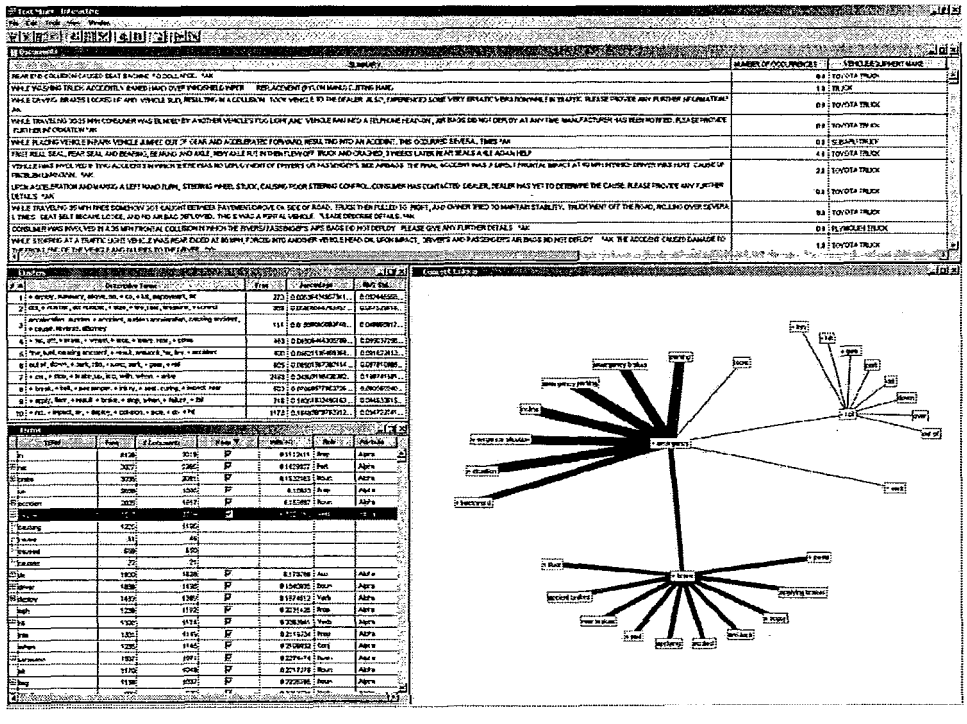


Figura 9 - Documentos analisados e a lista dos respectivos clusters no SAS Text Miner (SAS TECHNOLOGIES, 2006)

III.5.2 - TEXTANALYST

Atualmente na versão 2.3, a ferramenta Text Analyst é distribuída pela Megaputer Intelligence e utilizada por muitas empresas, instituições acadêmicas, governamentais, entre outros (MEGAPUTER, 2006). A entrada para o aplicativo é um texto, do qual as informações relevantes são extraídas. As opções iniciais da ferramenta estão exibidas na Figura 10, onde o usuário pode escolher entre analisar novos textos e criar a base de conhecimento, abrir uma base de conhecimento existente ou ainda utilizar o tutorial.

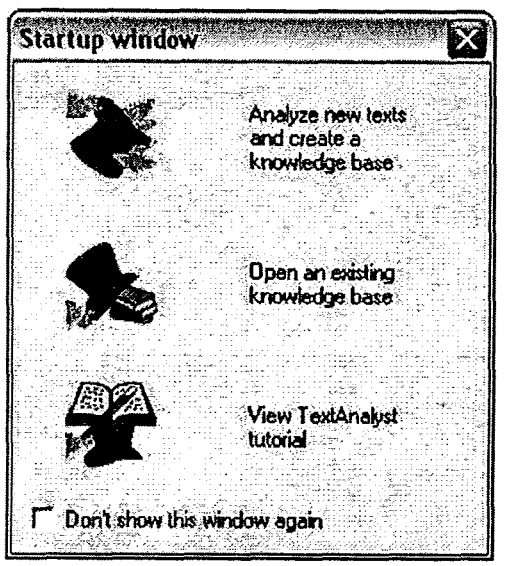


Figura 10 - Caixa de diálogo inicial do Text Analyst (MEGAPUTER, 2006)

Escolhendo a opção Analisar novos textos, o aplicativo determinará conceitos, ou seja, palavras e combinações de palavras que são mais importantes no contexto. Os algoritmos dispostos na ferramenta determinam a importância relativa de um conceito, conforme exemplo da Figura 11.

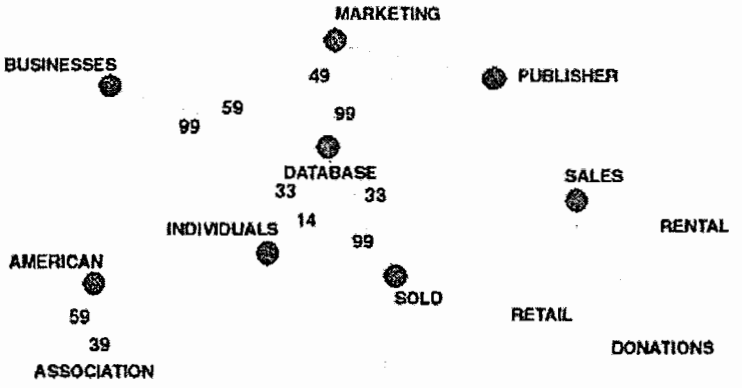


Figura 11 - Rede semântica (MEGAPUTER, 2006)

Deste modo, o Text Analyst cria uma rede semântica sem utilizar um conhecimento prévio do assunto. Entretanto, o tema pode ser adicionado pelo usuário através de um dicionário externo, quando desejado.

Para cada conceito, são obtidos dois valores que representam os pesos semânticos dos conceitos em relação ao conceito-pai e os pesos semânticos dos conceitos em relação ao documento.

O Text Analyst permite ainda a criação de sumários de documentos baseando-se nos pesos semânticos. Além disso, possibilita a exportação de resultados para arquivos HTML ou Excel, ilustrado na Figura 12.

	A	B	C	D	E	F	G	H
1	Parent	Frequency	Weight	Subordinate				
2	industry	9	67	==========				
3	industry	2	49	database				
4	industry	2	49	regulation				
5	industry	3	66	datum				
6	industry	2	49	many				
7	source	4	46	==========				
8	source	2	65	database				
9	source	2	65	datum				
10	source	2	65	common				
11	advertiser	3	10	==========				
12	advertiser	2	71	publisher				
13	consumer	14	97	==========				
14	consumer	3	57	privacy				
15	consumer	4	68	database				

ExportBase

Figura 12 - Resultados exportados para o Excel (MEGAPUTER, 2006)

III.5.3 - ONTOWEB

ONTOWEB é um sistema de análise de informações na Internet, que possibilita uma pesquisa contextualizada nas fontes acessadas. É uma solução desenvolvida com

tecnologias digitais de tratamento textual, com destaque para a Pesquisa Contextual Estruturada – PCE, a Representação do Conhecimento Contextualizado Dinamicamente – RC2D e a Engenharia de Ontologias (ESTADO DE SÃO PAULO, 2006).

Figura 13 - Processo de consulta do OntoWeb (ONTOWEB, 2006)

Esta tecnologia permite a realização de consultas com grandes volumes de texto, onde semântica e ontologias trabalham juntas para incrementar o processo de busca de informações relevantes em documentos digitais. Quanto maior o texto de entrada, melhor é a qualidade da resposta do ONTOWEB.

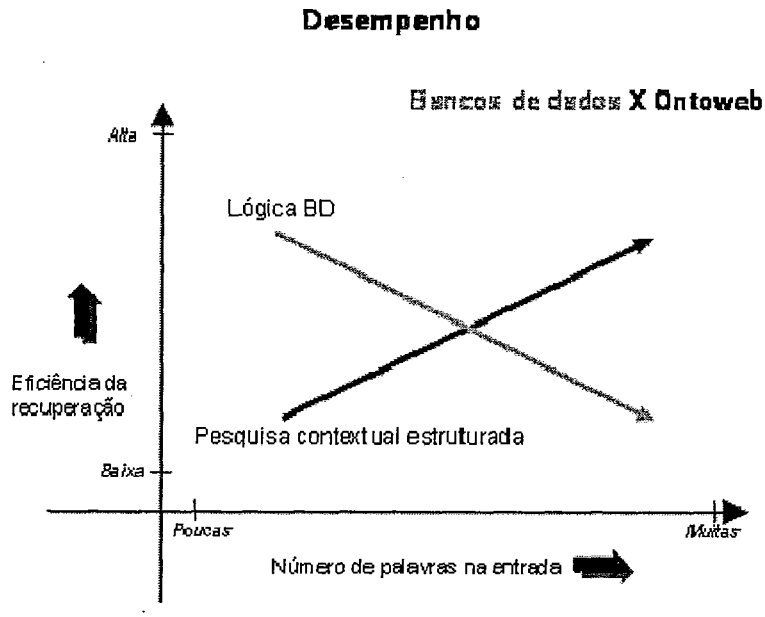


Figura 14 - Desempenho, segundo o OntoWeb (ONTOWEB, 2006)

A utilização de ontologias permite a este sistema considerar o contexto do assunto que está sendo pesquisado. As ontologias formam uma rede pré-existente de conceitos inter-relacionados que expandem o conceito utilizado, indicando ao sistema o cenário em que ele se enquadra, com base no processo de RC^2D. A partir daí, o *software* localiza, automaticamente, quais registros em sua base guardam mais semelhança com o texto digitado.

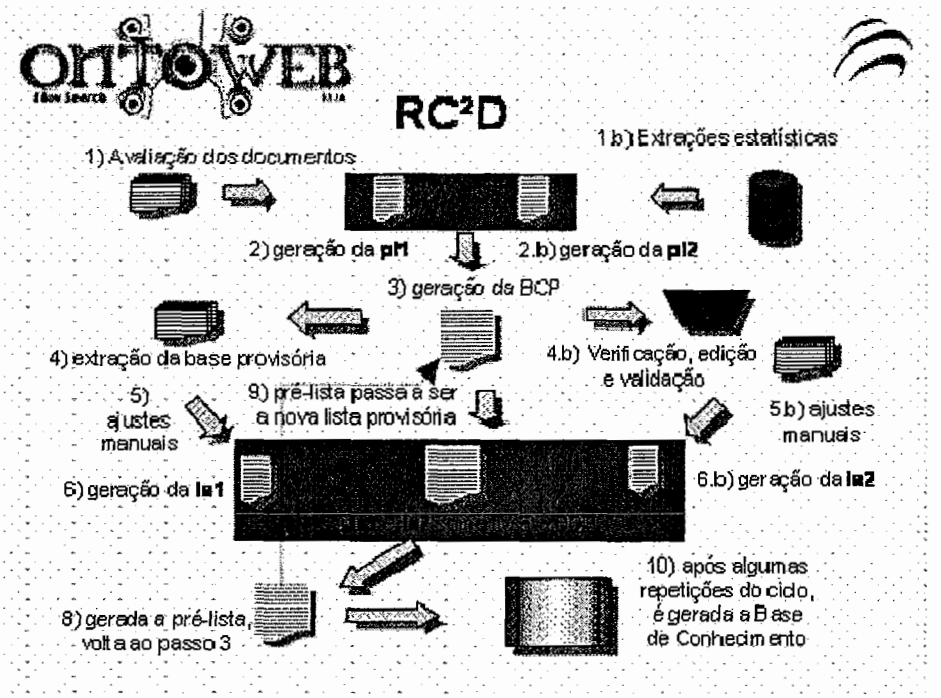


Figura 15 - O RC2D no OntoWeb (ONTOWEB, 2006)

Como as fontes de informação são selecionadas previamente, evita-se a recuperação de itens irrelevantes ou mesmo de lixo, apresentando resultados eficientes e organizados por ordem de similaridade (ONTOWEB, 2006).

O ONTOWEB está disponível através do site http://www.ontoweb.com.br.

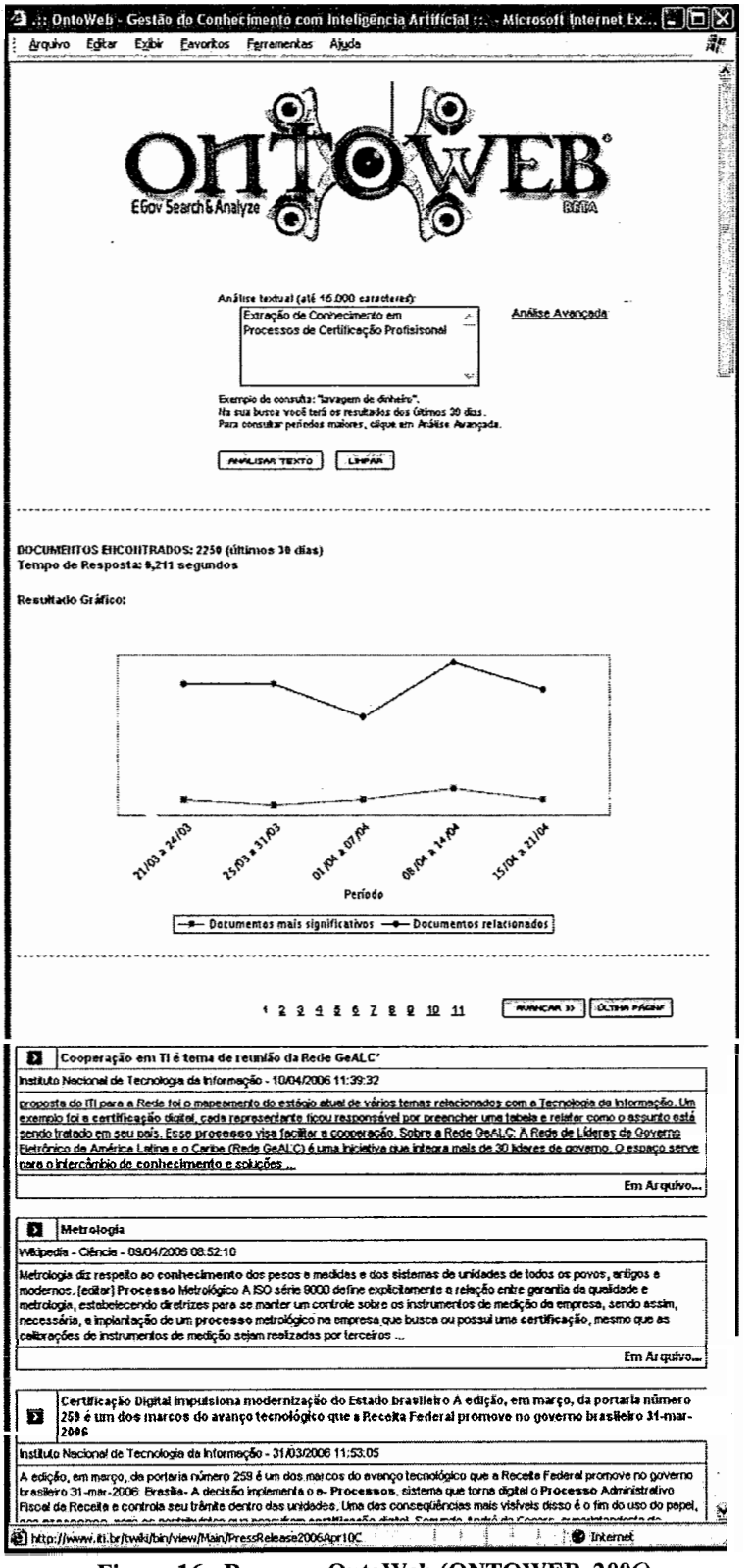


Figura 16 - Busca no OntoWeb (ONTOWEB, 2006)

III.5.4 - EUREKHA

Eurekha é um *software* desenvolvido no Instituto de Informática da Universidade Federal do Rio Grande do Sul cuja finalidade é de auxiliar o processo de análise e recuperação de informações provenientes de bases de dados textuais.

Basicamente, o que o Eurekha faz é analisar o conteúdo de textos e identificar aqueles que contêm o mesmo assunto. Estes documentos com conteúdo similar são atribuídos a um único *cluster* (grupo). Ao final do processo de análise, o *software* oferece ao usuário os diferentes grupos (*clusters*) encontrados e seus respectivos documentos, como apresentado na Figura 17.

Deste modo, tem-se uma distribuição dos documentos por assunto. Isso facilita a análise de uma quantidade muito grande de informações, pois basta analisar as palavras principais de cada grupo (fornecidas pela ferramenta) para identificar se o assunto contido nos textos do cluster em questão é relevante. Caso positivo, uma análise mais profunda pode ser feita no cluster e os demais documentos podem ser ignorados. Essa característica torna o processo de busca e recuperação de informações muito mais prática (WIVES & RODRIGUES, 2000).

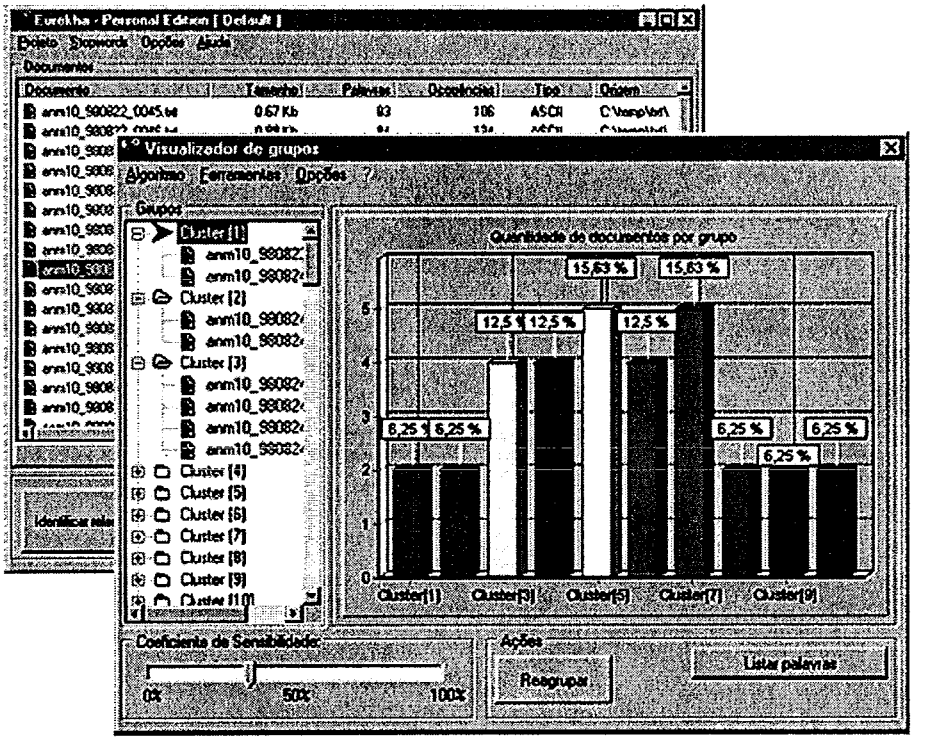


Figura 17 - Interfaces do Eurekha (WIVES & RODRIGUES, 2000)

III.5.5 - SMINER DESKTOP

Essencialmente, o SMiner Desktop surgiu a partir de um projeto final dedicado exclusivamente à mineração de textos no contexto acadêmico (RODRIGUES *et al*, 2004). O projeto foi desenvolvido em C/C++, utilizando SQL Server 2000 como banco de dados. O escopo deste projeto restringiu-se à utilização de publicações como insumo do sistema.

Nesta seção será abordada a utilização do minerador através de um exemplo de extração, desde o processamento do texto à obtenção das palavras relevantes.

Como pode ser observada na Figura 18, a entrada do minerador corresponde ao conjunto de textos a serem minerados. Este texto passará por um crivo que irá retirar os termos insignificantes (*stop words*) e terá como saída uma lista contendo as palavras-chave.

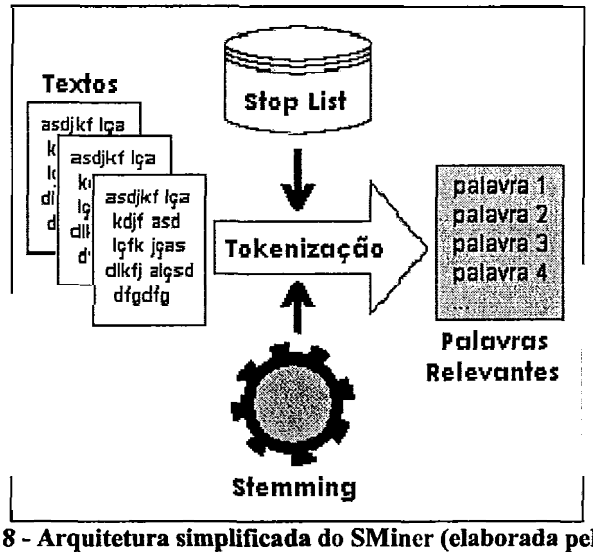


Figura 18 - Arquitetura simplificada do SMiner (elaborada pelo autor)

O sistema conta com um banco de dados que possibilita ao usuário escolher e acrescentar, quando necessário, as listas de *stop words* que deverão ser filtradas. Além disto, o minerador conta com um conjunto de configurações que auxiliam a extração, como: i) escolha do algoritmo de indexação; ii) definição do *threshold* (patamar), indicando a partir de que freqüência uma palavra se torna relevante; iii) pré-definição de termos importantes no contexto apresentado; iv) definição das palavras que devem ser consideradas diferencialmente relevantes; v) alimentação dinâmica dos termos a partir do histórico de consultas (MIRANDA *et al*, 2004).

O exemplo a seguir mostra, passo a passo, como o algoritmo apresentado se comporta no sistema desenvolvido.

Texto de exemplo: *TextMining consiste na extração de padrões ou conhecimentos interessantes e não-triviais a partir de documentos textuais. A tecnologia de TextMining pode ser aplicada para formalizar e explorar conhecimento tácito. O conhecimento disponível com pessoas pode ser armazenado em textos, os quais serão analisados para se entender seu significado, ou seja, do que tratam os textos. Depois, pode-se explorar o conhecimento extraído dos textos para gerar novos conhecimentos. Também se pode combinar este conhecimento com o conhecimento explícito armazenado em bases estruturadas.*

PALAVRAS	FREQÜÊNCIA		PALAVRAS	FREQÜÊNCIA	
Analisados	1	1,9231%	Interessantes	1	1,9231%
Aplicada	1	1,9231%	Não-triviais	1	1,9231%
Armazenado	2	3,8462%	Novos	1	1,9231%
Bases	1	1,9231%	Padrões	1	1,9231%
Combinar	1	1,9231%	Partir	1	1,9231%
Conhecimento	7	13,4615%	Pessoas	1	1,9231%
Consiste	1	1,9231%	Pode	3	5,7692%
Disponível	1	1,9231%	Pode-se	1	1,9231%
Documentos	1	1,9231%	Seja	1	1,9231%
Em	2	3,8462%	Ser	2	3,8462%
Entender	1	1,9231%	Serão	1	1,9231%
Estruturadas	1	1,9231%	Significado	1	1,9231%
Explícito	1	1,9231%	Tácito	1	1,9231%
Explorar	2	3,8462%	Também	1	1,9231%
Extração	1	1,9231%	Tecnologia	1	1,9231%
Extraído	1	1,9231%	Textmining	2	3,8462%
Formalizar	1	1,9231%	Textos	3	5,7692%
Gerar	1	1,9231%	Textuais	1	1,9231%
Interessantes	1	1,9231%	Tratam	1	1,9231%

Tabela 1 - Lista de palavras extraídas do texto (sem retirada de stop words)

A comparação entre a Tabela 1 e a Tabela 2 mostra a quantidade de *tokens* irrelevantes no contexto do texto aplicado. Isto é perfeitamente compreensível dado que a língua portuguesa possui uma infinidade de termos e conjunções que não interessam neste caso. Na Tabela 2, a saída é com base na freqüência das palavras consideradas relevantes, retirando-se as *stop words*.

PALAVRAS	FREQÜÊNCIA	ORIGINAL
Armazenado	2	3,8462%
Conhecimento	7	13,4615%
Explorar	2	3,8462%
Textmining	2	3,8462%
Textos	3	5,7692%

Tabela 2 - Lista de palavras após o processo de retiradas as *stop words*

Este exemplo apresentado foi concebido utilizando pesos iguais para todas as palavras e *threshold* igual a 2. Desta forma, é compreensível porque a palavra *"padrões"*, que de certa forma é significativa neste pequeno texto, não apareceu na saída do sistema, como mostra na Tabela 2. Para que esta palavra fosse considerada, o usuário poderia indicar seu peso apropriado previamente.

Uma outra aplicabilidade implementada é a redução de radicais através do *Stemming*. Este procedimento possibilita que palavras como *"textos"* e *"textuais"* sejam tratadas de forma análoga, através de seus radicais, no caso *"text"*.

Completado o processo de extração das palavras-chaves, o sistema armazena esta análise, a identificação da publicação bem como seus autores. Estas informações serão vitais no processo de formação das equipes (de trabalho ou comunidades, por exemplo). Cada equipe tem por si só um conjunto de competências previamente mapeadas e, de acordo com a extração do texto, a ferramenta apresenta o grupo de pesquisadores mais condizente com o conhecimento extraído (OLIVEIRA *et al*, 2004).

Desta forma, pode-se observar que a concepção inicial do projeto contemplava prioritariamente o meio acadêmico, sobretudo, a Gestão de Conhecimento da COPPE. Entretanto, em função da possibilidade de aferir resultados similares em um contexto organizacional, muitas mudanças seriam necessárias no *software*. Optou-se, portanto, pela reformulação da ferramenta, migrando para um contexto totalmente *web*, com uma infra-estrutura adequada ao contexto organizacional. Estas mudanças estruturais no *software* serão apresentadas nos capítulos seguintes.

III.5.6 - COMPARATIVO ENTRE AS FERRAMENTAS

Como forma de mensurar a motivação de se expandir o desenvolvimento do SMiner Desktop, é importante verificar as principais características de outras

ferramentas que possuem Text Mining como recurso. A Tabela 3 apresenta um comparativo com as principais características de cada sistema.

Para melhor visualizar o produto desta dissertação, foi incluída uma coluna contendo a versão atual do SMiner, que serve como um bom comparativo da sua evolução frente as ferramentas de mercado.

Características	SAS Text Miner	TextAnalyst	Onto Web	Eurekha	SMinerDesktop	Sminer
Tecnologia nacional			X	X	X	X
Insumos						
Analisa entradas / arquivos ascii	X	X	X	X	X	X
Analisa entradas / arquivos pdf	X				X	X
Analisa arquivos html	X					X
Analisa arquivos do pacote office (doc, xls, ppt)	X					X
Preparação dos Dados						
Uso de Stop List	X	X	X	X	X	X
Stemming Inglês	X	X*	X*		X	X
Stemming Português			X*		X	X
Uso de Thesaurus	X	X	X			X
Modelos de Recuperação de Informação						
Coincidência de Palavras	X*		X			X
Modelo Booleano	X*		X*		X	X
Modelo Espaço Vetorial	X	X	X	X	X	X
Modelo Vetor de Contexto	X	X	X	X		X
Modelo Probabilístico	X		X			
Técnicas para análise dos dados						
Análise Lexicométrica	X*	X*			X	X
Clusterização	X	X		X		
Classificação	X	X	X			X
Sumarização		X				
Outras ferramentas						
Web Crawler			X			

Tabela 3 - Comparativo entre ferramentas de mineração de textos

** Não foi encontrada na documentação e nos testes realizados referências diretas sobre a presença deste item, entretanto, pelas próprias tecnologias utilizadas, estima-se que a ferramenta contemple internamente o tópico abordado.*

III.6 - Considerações Finais

Esta seção apresentou os principais algoritmos, técnicas e ferramentas utilizadas na área de Mineração de Textos. Muitos destes algoritmos / técnicas dependem do contexto analisado. Por exemplo, o uso de *thesaurus* é fortemente acoplado ao tema em que se está trabalhando. Desta forma, alguns destes algoritmos e técnicas foram modificados em função do contexto desta dissertação.

As particularidades do processo abordado neste trabalho serão abordadas nas seções seguintes como forma de melhor elucidar a motivação do direcionamento no desenvolvimento do aplicativo SMiner.

IV - O Ambiente Organizacional Analisado

O ambiente organizacional estudado neste trabalho corresponde ao modelo de Formação, Qualificação e Certificação Profissional utilizado em uma grande indústria brasileira. Esta empresa é dividida em vários departamentos e unidades de negócio espalhadas pelo Brasil. Conforme dito na Introdução, a caráter de confidencialidade, será denominado neste texto os nomes EmpresaXPTO e DepartamentoA, que apelidam o nomes reais do caso de uso abordado. Para um melhor entendimento, será chamado de Operador o funcionário que trabalha na área de operação das unidades de negócios (indústrias) da EmpresaXPTO.

A EmpresaXPTO possui uma grande diversidade de documentos. A área do DepartamentoA realiza há 6 anos o Programa de Qualificação e Certificação Profissional da sua Força de Trabalho. Muitos documentos foram gerados neste período: livros didáticos, exercícios, apresentações sobre diversos temas, críticas e sugestões, questões de prova, documentos de referência pedagógica, relatórios diversos, entre outros. Não há um padrão de formatação específico, entretanto, todos estão em formato textual e podem ser tratados de tal forma.

A partir da manipulação desta série de insumos será possível realizar buscas qualitativas que poderão auxiliar à criação de materiais didáticos para os diversos cursos oferecidos pela Empresa. Estes cursos estão dispostos nos diversos Processos de Aprendizagem que seguem fluxos particulares e serão elucidados nas seções seguintes.

IV.1 - Os Processos de Aprendizagem

Os programas de Formação, Qualificação e Certificação da Força de Trabalho da EmpresaXPTO, coordenado pela área de Recursos Humanos do DepartamentoA, tem por objetivo treinar (qualificação e/ou certificação) a Força de Trabalho da EmpresaXPTO, que é formada, atualmente, por aproximadamente 5000 funcionários do *chão de fábrica*[2] distribuídos pelas diferentes Unidades de Negócios da Empresa.

Para realização deste processo, o setor de Recursos Humanos conta hoje com uma rede de 25 instituições credenciadas para a realização do processo de qualificação

2 Jargão utilizado para denominar a força de trabalho utilizada prioritariamente na operação de máquinas das Unidades de Negócio da Empresa, espalhadas pelo Brasil.

e/ou certificação. As empresas podem ser credenciadas como instituições que possuem a tarefa de formar, qualificar e aplicar exames aos candidatos (IFQ's – Instituição de Formação e Qualificação) ou instituições que possuem a tarefa de desenvolvimento dos recursos didáticos e instrumentos de avaliação (IDFQ's – Instituição de Desenvolvimento de Formação e Qualificação). Estas instituições auxiliam o gerenciamento dos programas de capacitação da Empresa.

Embora a denominação *Programa de Formação, Certificação e Qualificação Profissional* seja muito utilizada, há uma necessidade proeminente de maiores esclarecimentos sobre este contexto. Como o próprio nome sugere, este Programa é divido em 3 grandes áreas: Formação de Novos Operadores, Certificação Profissional e Qualificação Profissional. Cada um destes temas será elucidado nas seções seguintes.

IV.1.1 - FORMAÇÃO DE NOVOS OPERADORES

A EmpresaXPTO realiza frequentemente concursos de admissão de novos funcionários. Como a maioria das empresas, a EmpresaXPTO possui funcionários de diversos níveis de instrução, formados pelas diversas instituições educacionais (por exemplo: escolas, cursos técnicos e profissionalizantes e universidades) espalhadas pelo Brasil.

Entretanto, devido a grande peculiaridade do trabalho praticado, os candidatos que pleiteiam as vagas de Operador na empresa devem realizar, independente da formação acadêmica, um curso padrão, denominado Curso de Formação de Novos Operadores.

Este curso possui uma carga horária em torno de 500 horas e é dividido em cerca de 40 módulos (disciplinas) que, normalmente, são realizados em 4 meses de aulas (EMPRESAXPTO, 2003a). Desta forma, como o Curso de Formação de Operadores tem caráter reprobatório, um candidato somente será um Operador após a conclusão deste curso com êxito (Figura 19).

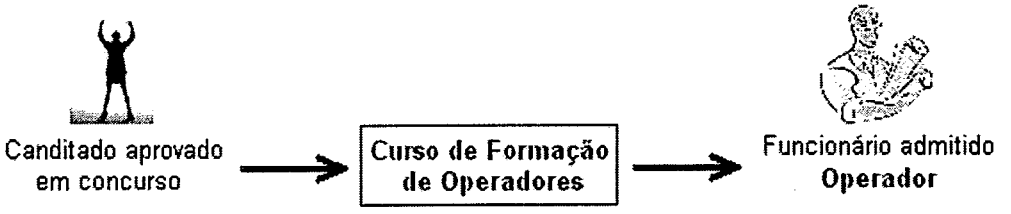


Figura 19 - Formação de Novos Operadores (elaborada pelo autor)

O Curso de Formação é um programa composto por aulas presenciais, provas intermediárias e uma prova final que irá avaliar se o candidato terá condições ou não de ser um Operador. Este programa é oferecido desde o ano 2000 como parte das iniciativas de capacitação profissional da empresa.

Entretanto, para que esta idéia de capacitação abrangesse os operadores que ingressaram antes do ano 2000, e tornar o programa acessível a toda força de trabalho, foi concebido o programa de Certificação Profissional. Este programa tem como objetivo diagnosticar os operadores "antigos" e propiciar cursos de qualificação específicos, quando necessário.

IV.1.2 - CERTIFICAÇÃO PROFISSIONAL

O objetivo do programa é propiciar aos operadores, em exercício da função, condições de rever conceitos, atualizar e/ou aperfeiçoar conhecimentos visando a certificação profissional nacional da empresa, que aborda diferentes especialidades técnicas.

O programa é destinado aos operadores do quadro de pessoal das Unidades de Negócio do Brasil. O perfil profissional dos operadores induz a um dinamismo em função do contexto de trabalho onde atuam. Trata-se de um processo de produção contínua, com alta exigência de qualidade e produtividade, num mercado dinâmico e competitivo.

Os operadores representam cerca de 40% do efetivo de pessoal e exercem funções vitais para a continuidade da produção. Utilizando um sistema de revezamento em turnos e por intermédio de ações coordenadas, adotadas pela empresa, os operadores mantêm o sistema em constante operação. No cotidiano de suas funções, precisam tomar decisões importantes, antever problemas ou situações de risco, ter alta capacidade de concentração, raciocínio e análise, facilidade de adaptação às novas tecnologias e novas formas de organização do trabalho. Na execução das tarefas mais simples até as mais complexas, a conjugação do conhecimento teórico e da prática profissional se faz necessária.

Assim, para esse perfil profissional, o aperfeiçoamento e a atualização constantes são indicadores de um profissional altamente qualificado, capaz de exercer de forma competente a sua função e sem trazer riscos à segurança, saúde e ao meio ambiente (EMPRESAXPTO, 2003b).

A Figura 20 apresenta o fluxo do Programa de Certificação Profissional, que é dividido em 3 etapas:

- Etapa Básica: Conclusão do Curso de Formação de Operadores ou Qualificação Básica[3];
- Qualificação Específica[4];
- Recertificação.

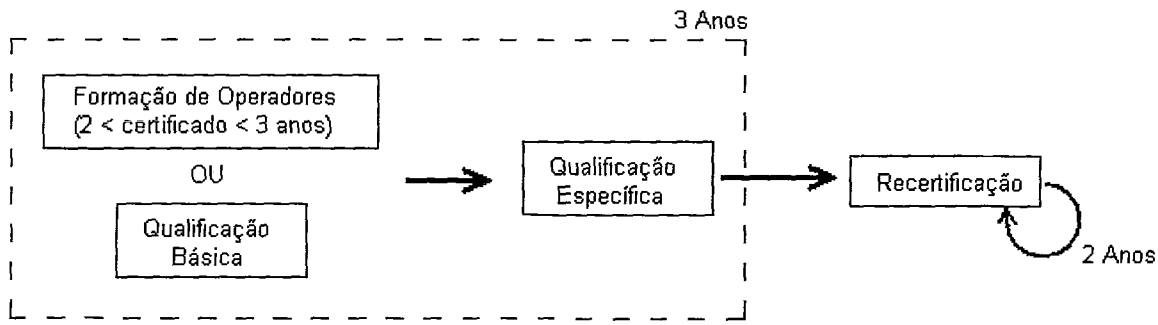


Figura 20 - Certificação Profissional (elaborada pelo autor)

O objetivo da etapa básica é propiciar aos operadores condições de rever conceitos, atualizar e/ou aperfeiçoar conhecimentos elementares de sua função. Esta, portanto, é uma etapa que precede a Qualificação Específica, e pode ser considerada concluída em duas situações:

1. O operador é relativamente novo na empresa e possui o certificado de conclusão do Curso de Formação de Operadores emitido entre 2 a 3 anos da data corrente;
2. O operador efetuou a Qualificação Básica.

Tendo feito a etapa básica, o operador poderá se especializar fazendo um dos diversos cursos da Qualificação Específica. Neste programa, o operador será diagnosticado e fará cursos em que sua capacidade técnica demonstrou-se pouco expressiva nas avaliações propostas.

---

[3] São os requisitos de qualificação comuns a todas as especialidades da carreira de operadores. Ex.: conhecimento sobre equipamentos, unidades de medidas, aspectos básicos de Segurança, Meio-ambiente e Saúde, ente outros (EMPRESAXPTO, 2003c).

[4] São os requisitos de qualificação específicos de cada especialidade da carreira de operadores.

Ao completar a etapa específica, o operador estará automaticamente certificado na área escolhida. O operador possui 3 anos para completar as etapas básica e específica. Se, neste período não conseguir finalizar todos os módulos, seu histórico é expirado e o operador deverá reiniciar todo o processo.

Por outro lado, uma vez certificado, o operador de 2 em 2 anos deverá confirmar sua certificação realizando a etapa de Recertificação. Esta etapa consiste na realização de uma prova mais complexa que abrange questões dos programas de qualificação básica e específica.

IV.1.3 - QUALIFICAÇÃO PROFISSIONAL

A EmpresaXPTO, por sua dimensão e pela própria distribuição geográfica possui, ao redor de suas Unidades de Negócio, comunidades (a maioria de baixa renda) que complementam a mão-de-obra de produção.

Faz parte da visão de negócios da empresa uma atuação responsável com o objetivo de cumprir a função social. O Programa de Qualificação Profissional é destinado às comunidades próximas às unidades de negócios, espalhadas por todo país. Os requisitos de seleção são: experiência profissional na área pretendida, escolaridade mínima exigida pelo curso e residir em comunidades nas proximidades das Unidades de Negócios.

O programa se insere numa perspectiva de inclusão social, dando oportunidade aos profissionais selecionados, a realização de cursos de atualização e formação profissional gratuitos, respeitando cada área de interesse.

Os perfis profissionais, bem como os conteúdos formativos do programa, foram elaborados por um grupo de especialistas das referidas áreas, técnicos da empresa e por profissionais da área, tendo como referência o Catálogo Brasileiro de Ocupações – CBO – 2002 e o Programa Nacional de Qualificação e Certificação de Pessoal na Área de Manutenção – PNQC – da Associação Brasileira de Manutenção – ABRAMAN.

Desta forma, o programa de formação e atualização profissional objetiva propiciar aos profissionais das comunidades adjacentes às Unidades de Negócio da EmpresaXPTO condições de realizar com qualidade, segurança e respeito ao meio ambiente as atividades inerentes à sua área de atuação visando o aumento da empregabilidade. São exemplos de ocupações disponibilizadas: soldador, caldeireiro,

eletricista industrial de manutenção, mecânico industrial de manutenção, montador de andaimes, operador de máquinas, pedreiro refratarista, instrumentista e pintor industrial.

No contexto desta dissertação, parte do material didático deste programa será utilizada nos diversos testes do estudo caso apresentado nas seções seguintes.

IV.2 - A MANUTENÇÃO DOS PROCESSOS

Diversos são os procedimentos que envolvem a manutenção dos processos de aprendizagem da empresa. Assim como uma instituição de ensino, a EmpresaXPTO oferece cursos para seus funcionários e utiliza provas para avaliá-los. Analogamente, a empresa dispõe de um sistema de cadastramento de críticas e sugestões aos processos. A seguir, estão detalhados os principais itens referentes à manutenção destes processos.

IV.2.1 - CURSOS E DIAGNÓSTICOS

Nos programas de Qualificação Básica e Específica o operador sempre tem a possibilidade de cursar módulos de especialidades (ou disciplinas) distintas. Entretanto, cada curso efetuado recorre em um ônus significativo tanto para a EmpresaXPTO, que terá que arcar com todas as despesas, quanto para o operador, que poderá estar despendendo tempo fazendo um curso cujo conteúdo já lhe é profundamente conhecido.

A solução encontrada neste caso foi diagnosticar o operador antes da realização dos cursos, ou seja, verificar se há a real necessidade de atualização. Desta forma, o operador tem a possibilidade de eliminar algumas especialidades sem a necessidade de fazer cursos, evitando a saída do seu posto de trabalho. Por outro lado, se o operador necessitar de um aprimoramento através de cursos, ele ainda deverá ter que fazer outro diagnóstico para verificar o seu aprendizado pós-curso.

O diagnóstico então, consiste na realização de provas de qualificação que indicarão se aluno possui a capacidade plena para o desempenho de atividades de trabalho previamente definidas. Além disso, também são considerados diagnósticos pela empresa, as provas intermediárias feitas pelos alunos dos cursos de formação de operadores.

As informações de alguns cursos e diagósticos também serão utilizadas no contexto desta dissertação como forma de viabilizar comparações qualitativas entre as questões elaboradas e aplicadas nos diagnósticos e o material didático apresentado nos cursos.

IV.2.2 - DOCUMENTOS DE REAÇÃO

O documento de reação é um documento de opinião pessoal composto por um conjunto de questões objetivas e uma observação subjetiva. Este documento é respondido após a finalização de cada curso e visa avaliar qualitativamente os cursos e diagnósticos oferecidos pela empresa, bem como mensurar os pontos fortes e fracos dos processos de formação, qualificação e certificação.

Item a observar	Fraco	Regular	Bom	Muito Bom
1 QUANTO À ORGANIZAÇÃO				
1.1 Carga horária x conteúdo do módulo	O	O	O	O
1.2 Tempo para realização da avaliação	O	O	O	O
1.3 Condição das instalações para realização do módulo	O	O	O	O
1.4 Condição das instalações para realização da avaliação	O	O	O	O
2 QUANTO AO CONTEÚDO DO MÓDULO				
2.1 Contribuição para o desempenho profissional da turma	O	O	O	O
2.2 Qualidade e quantidade das informações atenderam as expectativas da turma	O	O	O	O
2.3 Metódo utilizado	O	O	O	O
2.4 Equilíbrio entre teoria e/ou prática	O	O	O	O
3 QUANTO À AVALIAÇÃO				
3.1 Quanto ao comportamento dos convocados para realização da avaliação em sala de aula	O	O	O	O
3.2 Quanto às condições do ambiente para realização da avaliação (sem interrupções, com tranquilidade)	O	O	O	O
3.3 Tempo para realização da avaliação	O	O	O	O
4 QUANTO A ATUAÇÃO DA TURMA				
4.1 Absorção do conteúdo	O	O	O	O
4.2 Participação	O	O	O	O
4.3 Relacionamento docente - aluno	O	O	O	O
5 MATERIAL DIDÁTICO				
5.1 Existe relação entre os temas abordados e o contexto profissional do operador	O	O	O	O
5.2 Conteúdos são apresentados de forma atraente e atrativa	O	O	O	O
5.3 Qualidade dos módulos (encadernação, impressão e ilustração)	O	O	O	O
6 ATUAÇÃO DO DOCENTE-ILDEFONSO MARTINS DOS SANTOS				
6.1 Clareza na exposição dos assuntos	O	O	O	O
6.2 Facilidade em esclarecer dúvidas	O	O	O	O
6.3 Relacionamento com a turma	O	O	O	O
6.4 Assíduo e cumpre horário de trabalho	O	O	O	O
6.5 Utilizou o kit de material didatico padrão	O	O	O	O
6.6 Utilizou plano de atividade	O	O	O	O

Comentários e Sugestões 3000

Figura 21 - Documento de Reação (elaborada pelo autor)

A avaliação eletrônica, por meio de uma ferramenta computacional, destes documentos é fundamental para o entendimento das lacunas expostas pelos interlocutores do processo. Esta avaliação terá como insumo a extração de conhecimento a partir do campo "Comentários e Sugestões" do formulário disposto na Figura 21.

IV.3 - A Elaboração de Materiais Didáticos

A EmpresaXPTO dispõe de um grande número de funcionários que utilizam a Certificação para aprimorar seus conhecimentos técnicos. Dentre estas pessoas, existe um grupo (cerca de 5%) que atua diretamente na elaboração de materiais.

Normalmente, a elaboração dos materiais é feita a partir de *workshops*. Nestes *workshops* os funcionários da EmpresaXPTO ficam "confinados" em um ambiente propício e são constantemente motivados a externalizar os seus conhecimentos tácitos. É interessante ressaltar que a EmpresaXPTO detém conhecimentos únicos no Brasil, informações que, muitas vezes, só os funcionários possuem. Por este motivo, é muito difícil contratar pessoas externas (por exemplo: técnicos, mestres e/ou doutores) para a exclusiva elaboração de materiais. Desta forma, os *workshops* da Certificação (como são comumente chamados) correspondem a uma tática de sucesso para elaborar conteúdos para as Certificações da EmpresaXPTO.

Além disso, para cada tema de Certificação existe um conjunto de funcionários especializados. Por exemplo, na certificação X, existe a Banca de Certificação X, responsável por criar materiais didáticos para a especialidade X.

O processo de elaboração de cada insumo segue um fluxo particular. Nos tópicos a seguir, apresenta-se uma breve explanação de como é criado cada um dos itens acima abordados.

IV.3.1 - Elaboração de Livros

Os livros podem ser produzidos independentemente dos *workshops*. Os funcionários são divididos por subtemas e seus textos são encaminhados a um profissional responsável pela análise pedagógica. A cada sugestão feita pelo pedagogo, os funcionários efetuam as ponderações e retornam para nova avaliação. Este procedimento é efetuado até a conclusão do conteúdo em comum acordo.

Após a finalização do conteúdo, o texto segue para a análise gráfica e editoração, onde são acertadas as questões de interface de forma que o material fique o mais intuitivo e agradável possível. Ao finalizar a editoração, um protótipo é enviado para cada membro da banca e são reavaliados até a conclusão final do trabalho.

A cada criação de um novo livro, o processo de disseminação do conhecimento continua na medida em que a EmpresaXPTO distribui, gratuitamente, exemplares para

cada Unidade de Negócio da empresa. Além disso, a EmpresaXPTO incentiva e viabiliza cursos corporativos, fornecendo os materiais didáticos necessários.

Em cada curso, os alunos e/ou professores e coordenadores também podem participar do processo de melhoria destes materiais. Através dos documentos de sugestão, cada aluno poderá ponderar favoravelmente ou não sobre itens relacionados aos livros. As ponderações são variadas, desde aspectos gráficos a conteúdos específicos.

No contexto desta dissertação, os aspectos mencionados são importantes porque definem dois itens que servirão como insumos para o processo de mineração – os livros e os documentos de sugestões ponderados pelos membros envolvidos (alunos, professores e coordenadores dos cursos).

IV.3.2 - Elaboração das Questões dos Diagnósticos

Habitualmente, em processos de certificação existem conjuntos de diagnósticos (provas) que devem ser feitos para aferir o conhecimento do candidato à Certificação almejada. O processo de elaboração das questões que compõem os diagnósticos possui similaridades com o processo de elaboração de livros, contudo, as questões são elaboradas somente nos *workshops* da Certificação.

Assim como na elaboração dos livros, existe uma banca para cada especialidade. Esta banca, composta por funcionários de Unidades de Negócio diferentes, se reúne nos *workshops* e criam as questões baseadas nos livros elaborados. Este é um ponto extremamente peculiar – a orientação é que as questões sejam baseadas nos livros.

Esta orientação tem um propósito tão forte que gera inclusive discussões nos diversos cursos de certificação. No momento em que o aluno se depara com uma questão que, supostamente, não se encontra no livro, há uma reação imediata. Esta reação é justificada porque o contexto motiva isto. Basta imaginar um candidato tenso, ansioso e furioso ao perceber que uma determinada questão de seu diagnóstico foge as regras estabelecidas.

O desgaste provocado por uma insatisfação deste gênero é comumente externalizado nos documentos de sugestão de questões. Por outro lado, através das experiências de campo, pôde-se perceber que uma ferramenta de apoio poderia auxiliar os membros da banca a gerar questões mais compatíveis com os livros. Neste intuito,

foram adicionados como insumo a esta dissertação, os documentos de sugestão de questões, as questões e suas respectivas respostas.

IV.4 - Considerações Finais

A partir do entendimento destes processos é possível compreender a origem dos documentos e, desta forma, corroborar a necessidade de uma ferramenta capaz de extrair informações relevantes dos contextos apresentados.

Sob o ponto de vista dos profissionais envolvidos na criação dos materiais didáticos, há um grande ganho de tempo com a possibilidade de encontrar documentos a partir de buscas diretas, ou ainda, verificar a probabilidade de uma questão estar contida em um material didático, por exemplo. Estas e outras facilidades serão possíveis a partir da criação de um Sistema de Mineração de Textos – objeto desta dissertação e tecnicamente apresentado na seção a seguir.

V - IMPLEMENTAÇÃO DO SISTEMA

Como foi abordado, na maior parte das organizações, o volume de documentos textuais é bastante significativo. Muitas consultas a esses documentos são necessárias e esta busca é freqüentemente lenta devido à variedade de temas tratados e à quantidade armazenada, por vezes em múltiplos locais. Além disso, é comum uma pesquisa retornar um conjunto grande de documentos de interesse específico, onde apenas uma pequena parte é realmente relevante.

Tecnologias são necessárias para acelerar a análise, examinando de forma automatizada estes documentos e aferindo aquilo que é verdadeiramente significativo. É possível também, a partir da análise de um resumo de um grupo de documentos, mensurar relações importantes entre eles e que antes não seriam percebidas. Uma dessas tecnologias é a Mineração de Textos (*Text Mining*), como apresentado.

No âmbito dos ambientes de aprendizagem, sobretudo, nas certificações profissionais, muitos insumos podem ser considerados para uma análise qualitativa, entre eles: livros, apostilas, documentos de referência, documentos de opinião (críticas e sugestões), questões de provas, simulados e dicionários técnicos.

Desta forma, uma proposta de *software* foi concebida para diminuir possíveis divergências dos processos de certificação e os insumos nela utilizados. A ferramenta desenvolvida neste contexto é baseada na tecnologia de mineração de textos bem como nas particularidades proferidas nos processos de aprendizagem e certificação anteriormente citados.

Nas seções a seguir serão apresentados os processos de preparação dos insumos bem como os principais algoritmos utilizados para a manipulação dos mesmos.

V.1 - PREPARAÇÃO DOS INSUMOS

A rigor, qualquer texto pode sofrer uma tentativa de avaliação por um sistema de mineração de textos, contudo, no contexto desta dissertação, optou-se por utilizar os insumos encontrados nos processos de certificação profissional da EmpresaXPTO:

- Livros – parte do material didático utilizado em possíveis cursos;
- Documentos de Referência – documentos que regulamentam os cursos, normalmente acessados pela coordenação;

- Material de Apoio Complementar – Apostilas, artigos, apresentações e/ou textos complementares aos livros;
- Questões – pontos de avaliação, compostos em sua maioria, de itens objetivos com uma opção de resposta correta, cujo conteúdo é baseado no material didático (livros) disponível;
- Dicionário de Termos Técnicos – termos e seus respectivos significados, encontrados em diversos processos de aprendizagem e certificação;
- Documentos de Opinião – documentos de opinião pessoal, contendo críticas e sugestões do processo.

Como podem ser observados na lista apresentada, os insumos contemplados correspondem a um conjunto valiosíssimo para análise qualitativa. Entretanto, a simples ação de ler cada um destes itens demanda um custo de tempo inviável nas rotinas institucionais. Além disso, seria interessante uma análise paralela, o que demandaria um tempo ainda maior.

O *Capítulo IV - O Ambiente Organizacional Analisado* apresenta a motivação para o desenvolvimento dos módulos descritos nas seções seguintes. Ainda neste capítulo, abordou-se como cada insumo é gerado. Desta forma, é possível compreender que estes insumos (documentos) exercem um papel imprescindível na busca. Entretanto, estes materiais tiveram que ser preparados de uma forma muito particular, o que será elucidado nas seções seguintes.

V.1.1 - Livros, Documentos de Referência e Materiais Complementares

Os livros, os documentos de referência profissional e os materiais de apoio didático complementar estão, em sua maioria, dispostos em PDF e DOC e foram carregados a partir dos cadastros da própria ferramenta, descritos na seção *VI.2 - Manutenção do Sistema.*

Neste caso, não houve um tratamento específico, dado que a ferramenta contempla estes formatos e tem por objetivo justamente a manipulação de arquivos texto não estruturados.

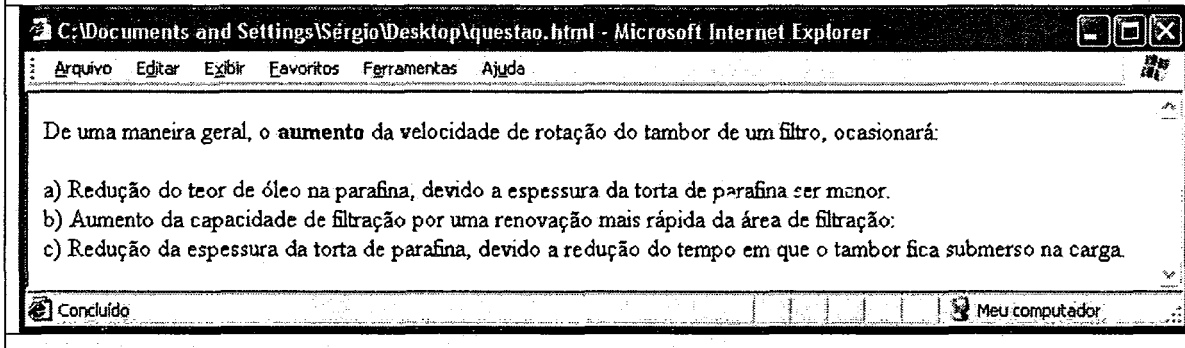
V.1.2 - QUESTÕES E RESPOSTAS

Originalmente, as questões e respostas estão armazenadas em um banco de dados da EmpresaXPTO. Uma parte dessa massa de dados foi migrada para o SMiner de forma que as buscas pudessem contemplar este insumo.

Embora estivessem em formato texto, as questões e respostas continham *tags* HTML que eram utilizadas para formatar as questões no momento da apresentação. Desta forma, um mecanismo de "limpeza" foi desenvolvido para retirar estas *tags* e viabilizar, assim, a entrada do texto no minerador.

A Figura 22 apresenta o tratamento efetuado no HTML das questões para que estas pudessem ser utilizadas como insumo no SMiner.

Apresentação da Questão no Navegador



Texto em HTML referente à questão

```
De uma maneira geral, o aumento da velocidade de rota&ccedil;&atilde;o do tambor de
um filtro, ocasionar&aacute;: <br/>a) Redu&ccedil;&atilde;o do teor de
&oacute;leo na parafina, devido a espessura da torta de parafina ser
menor. <br/>b) Aumento da capacidade de filtra&ccedil;&atilde;o por uma
renova&ccedil;&atilde;o mais r&aacute;pida da &aacute;rea de
filtra&ccedil;&atilde;o: <br/>c) Redu&ccedil;&atilde;o da espessura da torta de
parafina, devido a redu&ccedil;&atilde;o do tempo em que o tambor fica submerso na
carga. <br/>
```

Texto após tratamento

```
De uma maneira geral, o aumento da velocidade de rotação do tambor de um filtro,
ocasionará:
a) Redução do teor de óleo na parafina, devido a espessura da torta de parafina ser
menor.
b) Aumento da capacidade de filtração por uma renovação mais rápida da área de
filtração:
c) Redução da espessura da torta de parafina, devido a redução do tempo em que o
tambor fica submerso na carga.
```

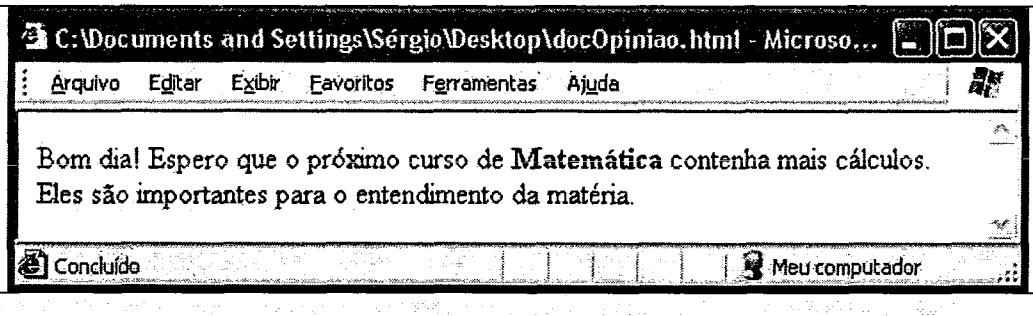
Figura 22 - Preparação das Questões (elaborada pelo autor)

Como pode ser observado, o texto após o tratamento segue um padrão ASCII suportado pelo processo de extração de informações.

V.1.3 - DOCUMENTOS DE OPINIÃO

Analogamente às questões, os documentos de opinião (críticas & sugestões) estavam dispostos em formato HTML e tiveram que ser tratados. O processo foi exatamente o mesmo – utilizou-se um mecanismo que tratasse das *tags* HTML e transformasse o texto em ASCII, como mostra a Figura 23.

Apresentação da Questão no Navegador



Texto em HTML referente à questão

```
<p>Bom dia!  Espero que o pr&oacute;ximo curso de <b>Matem&aacute;tica</b>
contenha mais c&aacute;lculos.</br>  Eles s&atilde;o importantes para o
entendimento da mat&eacute;ria.</p>
```

Texto após tratamento

```
Bom dia! Espero que o próximo curso de Matemática contenha mais cálculos.
Eles são importantes para o entendimento da matéria.
```

Figura 23 - Preparação dos Documentos de Opinião (elaborada pelo autor)

V.1.4 - TERMOS TÉCNICOS

Os termos técnicos foram inseridos na base de dados a partir da manipulação de um dicionário de termos técnicos. Este dicionário era composto por um conjunto de páginas HTML, onde cada arquivo (HTML) correspondia ao glossário de um livro.

Como havia uma formatação previamente definida, bastava recuperar as informações a partir de cada arquivo (HTML) e classificá-la de acordo com o livro abordado. Foi desenvolvido então, um Categorizador para manipular estes arquivos que, baseados na formatação padrão, extraiu os termos técnicos e armazenou na base de dados do SMiner, conforme apresentado na Figura 24.

Foi observado ainda que a formatação padrão seguia uma hierarquia bem definida. Desta forma, optou-se por armazenar os termos técnicos em uma estrutura do tipo árvore que, viabilizasse os mecanismos de consulta bem como a ilustração do dicionário de termos a um usuário qualquer (*seção VI.3.1 - Árvore de Conhecimento dos Processos*).

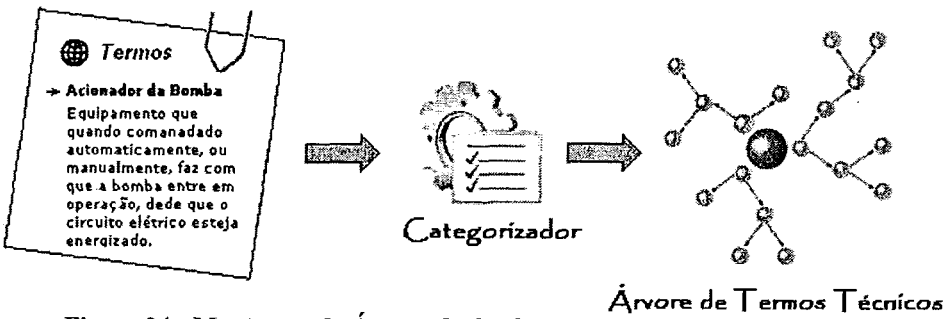


Figura 24 - Montagem da Árvore de Conhecimento (elaborada pelo autor)

A compreensão dos procedimentos de preparação dos insumos é importante porque correspondem às entradas dos processos de extração de informações, apresentados na seção a seguir.

V.2 - EXTRAÇÃO DAS INFORMAÇÕES TEXTUAIS

Inicialmente, esta seção aborda uma proposta de arquitetura de extração de palavras relevantes a partir um texto qualquer. Na verdade, o termo relevante é plenamente dependente do contexto apresentado, bem como nos algoritmos utilizados na arquitetura proposta.

Como pode ser observado na Figura 25, os insumos do procedimento são exatamente os textos a serem minerados. Os principais procedimentos adotados para extração são os chamados: Tokenização, Retirada da *Stop List*, *Stemming* e, como resultado, a separação das palavras relevantes.

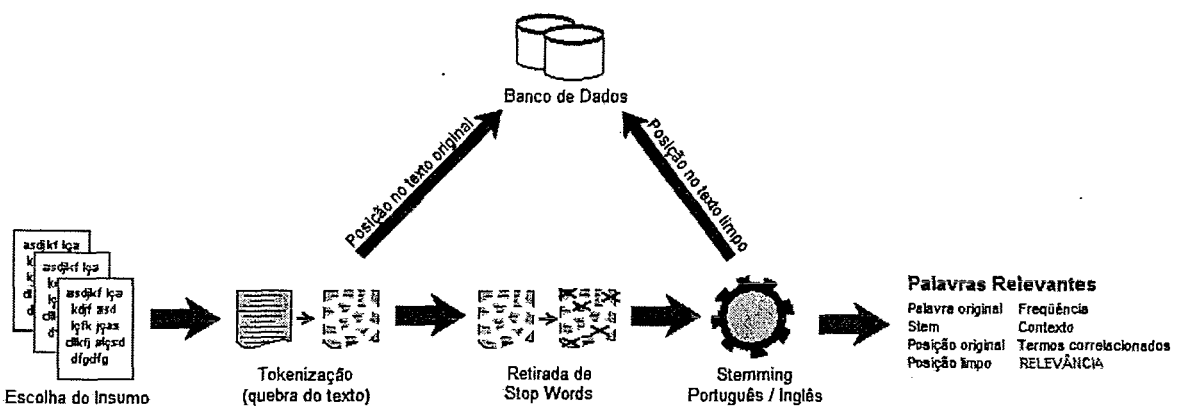


Figura 25 - Extração de palavras relevantes (elaborada pelo autor)

Em um processo típico, primeiramente são escolhidos os textos (insumos) com os quais o sistema irá trabalhar. Tendo uma base significativa de textos, o processo pode ser iniciado. Para cada texto é criada uma chave identificadora que será

armazenada em um Banco de Dados. Esta identificação possibilitará consultas futuras sem necessidade de uma nova aplicação do algoritmo para o insumo em questão (a não ser que os parâmetros de extração sejam alterados).

Paralelamente, o texto é submetido ao algoritmo de geração de *tokens*. A Tokenização consiste em uma regra de identificação de palavras (os *tokens*). Esta técnica sugere que os *tokens* sejam definidos como uma *string* de caracteres alfanuméricos sem espaços, podendo incluir hífens e letras acentuadas. Desta forma, o espaço em branco é o recurso utilizado para separar palavras. Aparentemente, a aplicação deste algoritmo parece simples, contudo, alguns problemas podem ser encontrados ao aplicar esta técnica:

- Pontuação – as palavras podem estar anexadas a uma pontuação representada por vírgula, ponto-e-vírgula e ponto final. A principio, pode parecer fácil o reconhecimento desta pontuação, contudo, nos casos de abreviatura, por exemplo, o ponto final é problemático.

- Espaço em branco – algumas vezes o espaço em branco não indica uma separação entre o sentido geral do conjunto de palavras. Por exemplo, no conjunto de palavras Rio de Janeiro, o algoritmo deveria sinalizar de que se trata apenas de uma palavra (Rio de Janeiro), e não apresentar como duas palavras relevantes (rio e janeiro), que, separadas, pouco tem a ver com o contexto. Para minimizar este problema, a ferramenta armazena a posição de cada palavra no texto, antes e após a retirada de *stop words*.

Como pode ser observado na Figura 26, após a quebra do texto em palavras (*tokens*), o algoritmo retira as palavras que não possuem relevância significativa no texto – as chamadas *Stop Words*. O conjunto de *Stop Words* que serão retirados do texto compõe a *Stop List*. Esta lista de palavras irrelevantes é profundamente dependente da língua e do contexto utilizados. O aplicativo proposto nesta dissertação permite a manipulação das *Stop Words* através de interfaces de manutenção, que possibilitam a criação de palavras irrelevantes no contexto apresentado.

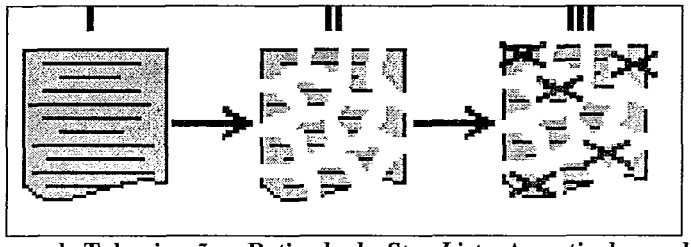


Figura 26 - Processos de Tokenização e Retirada da *Stop List* – A partir de um determinado texto (I) obtém-se as palavras em separado - Tokenização (II) e, então, retira-se as *stop words* (III). (elaborada pelo autor)

Com a retirada das *Stop Words*, as palavras restantes são consideras filtradas e devem passar por um novo processo de triagem. Nesta fase, o procedimento seguinte é a criação de pesos para cada tipo de palavra. Um artifício mais simples é indicar que todas as palavras possuem o mesmo peso em relação umas as outras, desta forma, o grau de relevância de cada *token* é dado a partir da freqüência em que este aparece no texto. A alternativa mais significativa sugere a criação de uma lista de palavras e seus respectivos pesos. Neste caso, o algoritmo além de contabilizar a freqüência dos *tokens*, analisa também se as palavras recuperadas possuem relevância no contexto. É interessante observar que o fato de uma palavra ter uma alta freqüência não indica claramente que esta seja significativa no contexto.

Além disto, diferentes palavras podem possuir o mesmo radical, como por exemplo: *texto*, *textos*, *textual* e *textualmente*. Entretanto, a contabilização da freqüência destes *tokens* deve ser feita pelo radical e não pela palavra original. Para minimizar este problema, procedimentos baseados no algoritmo de Porter podem ser utilizados.

V.3 - RADICALIZAÇÃO DE PALAVRAS

O algoritmo de Radicalização (*Stemming)* consiste em converter cada palavra para seu radical ("stem"). Por exemplo, as palavras "learning" e "learned" são ambas convertidas para o *stem* "learn" (PORTER, 1980).

Os algoritmos de *stemming* utilizados neste trabalho são baseados no algoritmo de Porter (1997) e no algoritmo do *Portuguese Stemmer* (ORENGO & HUYCK, 2001), descritos na seção *III.2.3 - Stemming*.

A partir destes procedimentos, foi implementado um algoritmo para fazer a radicalização das palavras na ferramenta proposta.

Embora haja suporte para a radicalização de palavras em inglês (implementado a partir dos conceitos da seção *III.2.3.1 - Porter Stemmer*), a implementação do

radicalizador para língua Portuguesa merece um maior destaque dado que é o mais utilizado no contexto desta dissertação.

A Figura 27 apresenta os 8 passos utilizados na implementação do radicalizador de palavras em português.

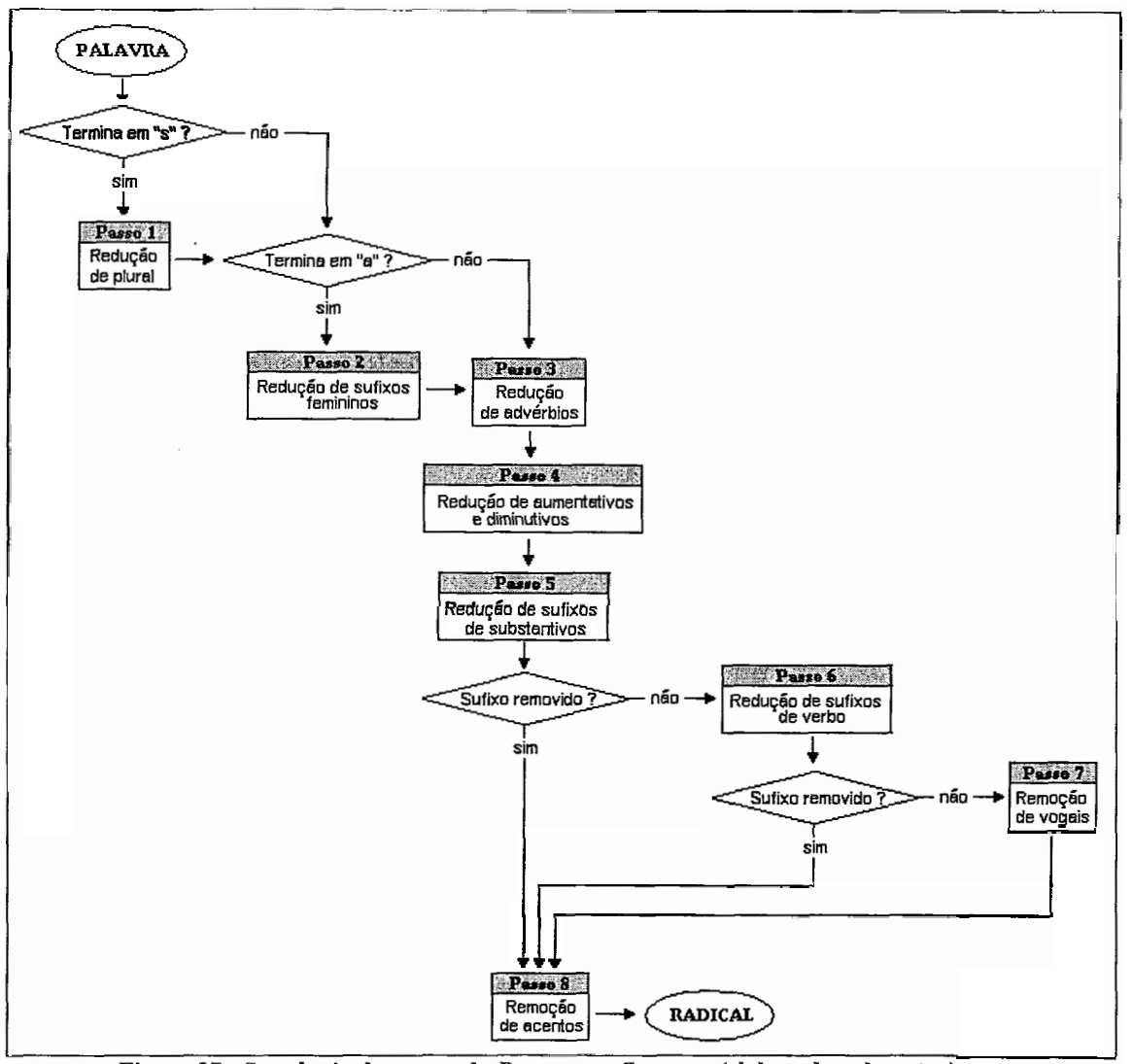


Figura 27 - Sequência de passos do *Portuguese Stemmer* (elaborada pelo autor)

Cada passo contém uma série de regras, que são executadas seqüencialmente. Os sufixos maiores são retirados primeiramente e os últimos passos contemplam melhorias na obtenção dos radicais. Por exemplo, o sufixo *–es* deve ser testado antes do sufixo *–s*. Atualmente, cerca de 250 regras similares a Tabela 4 são verificadas. A configuração de entrada no algoritmo também deve ser observada:

- Sufixo a ser retirado;

- Menor tamanho do radical: para evitar a remoção do sufixo quando o radical é muito pequeno;
- Sufixo que pode ser adicionado ao radical, se possível;
- Tratamento a partir da lista de exceções: para cada regra definida, existem várias exceções. Desta forma, quando uma palavra encontra-se na lista de exceções, a regra não é aplicada.

"inho", 3, "", {"caminho", "carinho", "cominho", "golfinho", "padrinho", "sobrinho", "vizinho"}

Tabela 4 - Exemplo de configuração de uma regra

Onde:

- *"inho"* – representa um sufixo que indica um diminutivo;
- O tamanho mínimo para o radical é 3: se o algoritmo encontrar a palavra *linho*, por exemplo, não irá aplicar o *stemming*, dado que o resultado apresentaria apenas a letra *l* neste caso; desta forma, somente palavras com 7 letras ou mais serão tratadas nesta regra;
- As demais palavras entre chaves correspondem às exceções dispostas à regra em questão.

Embora esteja baseado no algoritmo de Porter, este algoritmo foi estendido em virtude da diversidade de termos da língua portuguesa. Os 8 passos a seguir demonstram o seu funcionamento:

Passo 1: Redução de Plural

Com raras exceções, as formas plurais em Português terminam em *–s*. Entretanto, nem todas as palavras que terminam com *–s* indicam plural: *lápis*, por exemplo. Este passo consiste basicamente em remover o "s" final das palavras que não estão listadas nas exceções. Em alguns casos, os radicais necessitam de modificação: palavras que terminam com *–ns* devem ter o sufixo substituído por *"m"*, como por exemplo, *bons* → *bom*. Outros exemplos podem ser vistos na Tabela 5.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"ns"	1	"m"	bons → bom
"ões"	3	"ão"	balões → balão

"ães"	1	"ão"	capitães → capitão
"ais"	1	"al"	normais → normal
"eis"	2	"el"	papéis → papel
"eis"	2	"el"	amáveis → amável
"óis"	2	"ol"	lençóis → lençol
"is"	2	"il"	barris → barril
"lês"	3	"l"	males → mal
"res"	3	"r"	mares → mar
"s"	2		casas → casa

Tabela 5 - Regras de redução de plural

Passo 2: Redução de sufixos femininos

Todos os substantivos e adjetivos em Português possuem um gênero. Este passo transforma as formas femininas em sua correspondência masculina. Como pode ser visto na Tabela 6, somente palavras que terminam em *–a* são testadas neste passo. Entretanto, nem todas são convertidas, somente aquelas que possuem sufixos femininos conhecidos, como por exemplo, *chinesa → chinês*.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"ona"	3	"ão"	chefona → chefão
"ã"	2	"ão"	vilã → vilão
"ora"	3	"or"	professora → professor
"na"	4	"no"	americana → americano
"inha"	3	"inho"	sozinha → sozinho
"esa"	3	"ês"	inglesa → inglês
"osa"	3	"oso"	famosa → famoso
"íaca"	3	"íaco"	maníaca → maníaco
"ica"	3	"ico"	prática → prático
"ada"	2	"ado"	cansada → cansado
"ida"	3	"ido"	mantida → mantido
"ima"	3	"imo"	prima → primo
"iva"	3	"ivo"	passiva → passivo
"eira"	3	"eiro"	primeira → primeiro

Tabela 6 - Regras de redução de sufixos femininos

Passo 3: Redução de Advérbios

Este passo trata somente dos sufixos *–mente;* salvo a palavra *experimente* que, no caso, é uma exceção. A Tabela 7 apresenta um exemplo.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"mente"	4		felizmente → feliz

Tabela 7 – Regras de redução de advérbios

Passo 4: Redução de Aumentativos e Diminutivos

Substantivos e adjetivos em Português apresentam muitas variações. As palavras podem possuir aumentativo, diminutivo e superlativo, como por exemplo: casinha, onde *inha* indica um diminutivo. Este é um típico caso tratado neste passo, como mostra a Tabela 8.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"íssimo"	3		fortíssimo → fort
"ésimo"	3		centésimo → cent
"érrimo"	4		chiquérrimo → chiqu
"zinho"	2		pezinho → pe
"quinho"	4	"c"	maluquinho → maluc
"uinho"	4		amiguinho → amig
"adinho"	3		cansadinho →cansad
"inho"	3		carrinho → carr
"alhão"	4		grandalhão→ grand
"uça"	4		dentuça → dent
"aço"	4		ricaço→ ric
"adão"	4		casadão→ cas
"ázio"	3		corpázio→ corp
"arraz"	4		pratarraz→ prat
"arra"	3		bocarra → boc
"zão"	2		calorzão→ calor
"ão"	3		meninão → menin

Tabela 8 - Regras de redução em sufixos aumentativos e diminutivos

Passo 5: Redução de sufixos de substantivos

A Tabela 9 apresenta o passo que analisa os sufixos de substantivos e adjetivos. Se o sufixo é removido neste momento, os passos 6 e 7 não são executados.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"encialista"	4		existencialista → exist
"alista"	5		minimalista → minim
"agem"	3		contagem → cont
"iamento"	4		genrenciamento → gerenc
"amento"	3		monitoramento →monit
"imento"	3		nascimento → nasc
"alizado"	4		comercializado→comerci
"atizado"	4		traumatizado → traum
"izado"	5		alfabetizado→alfabet
"ativo"	4		associativo → associ
"tivo"	4		contraceptivo→contracep
"ivo"	4		esportivo→esport
"ado"	2		abalado → abal
"ido"	3		impedido → imped
"ador"	3		ralador → ral
"edor"	3		entendedor→entend
"idor"	4		cumpridor→cumpr
"atória"	5		obrigatória→obrig
"or"	2		produtor→produt
"abilidade"	5		comparabilidade→compar
"icionista"	4		abolicionista→abol
"cionista"	5		intervencionista→interven
"ional"	4		profissional→profiss
"ência"	3		referência→refer
"ância"	4		repugnância→repugn
"edouro"	3		abatedouro→abat
"queiro"	3	"c"	fofoqueiro→fofoc
"eiro"	3		brasileiro→brasil
"oso"	3		gostoso→gost

"alizaç"	5		comercializaç→comerci
"ismo"	3		consumismo→consum
"izaç"	5		concretizaç→concret
"aç"	3		alegaç→aleg
"iç"	3		aboliç→abol
"ário"	3		anedotário→anedot
"ério"	6		ministério→minist
"ês"	4		chinês → chin
"eza"	3		beleza→ bel
"ez"	4		rigidez→rigid
"esco"	4		parentesco→parent
"ante"	2		ocupante→ocup
"ástico"	4		bombástico→bomb
"ático"	3		problemático→problem
"ico"	4		polêmico→polêm
"ividade"	5		produtividade→produt
"idade"	5		profundidade→profund
"oria"	4		aposentadoria→aposentad
"encial"	5		existencial→exist
"ista"	4		artista→art
"quice"	4	"c"	maluquice→maluc
"ice"	4		chatice→ chat
"íaco"	3		demoníaco→demon
"ente"	4		decorrente→decorr
"inal"	3		criminal→crim
"ano"	4		americano→ americ
"ável"	2		amável→ am
"ível"	5		combustível→combust
"ura"	4		cobertura→cobert
"ual"	3		consensual→consens
"ial"	3		mundial→mund
"al"	4		experimental→experiment

Tabela 9 - Regras de redução de substantivos

Passo 6: Redução de sufixos de verbo

Notadamente, o Português possui muitos conjugações verbais. Em contraste com a língua inglesa, que possui apenas 4 variantes (por exemplo: *talk, talks, talked, talking*), na língua portuguesa este número ultrapassa a 50 diferentes formas, segundo Macambira (1999). Cada uma destas variantes possui seu sufixo e podem ser verificados de acordo com seu tempo, pessoa, número e modo. A estrutura das formas verbais é representada como:

Radical + vogal temática + tempo verbal + pessoa → *and + a + ra + m* → *andaram*

Desta forma, as formas verbais serão reduzidas ao seu radical, como mostra a Tabela 10.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"aríamo"	2		cantaríamo → cant
"ássemo"	2		cantássemo → cant
"eríamo"	2		beberíamo → beb
"êssemo"	2		bebêssemo → beb
"iríamo"	3		partiríamo → part
"íssemo"	3		partíssemo →part
"áramo"	2		cantáramo→ cant
"árei"	2		cantárei→ cant
"aremo"	2		cantaremo → cant
"ariam"	2		cantariam→ cant
"aríei"	2		cantaríei→ cant
"ássei"	2		cantássei→ cant
"assem"	2		cantassem→ cant
"ávamo"	2		cantávamo→ cant
"êramo"	3		bebêramo→ beb
"eremo"	3		beberemo→ beb
"eriam"	3		beberiam→ beb
"eríei"	3		beberíei→ beb
"êssei"	3		bebêssei→ beb

"essem"	3	bebessem→ beb
"íramo"	3	partiríamo→ part
"iremo"	3	partiremo→ part
"iriam"	3	partiriam→ part
"iríei"	3	partiríei→ part
"íssei"	3	partíssei→ part
"issem"	3	partissem→ part
"ando"	2	cantando→ cant
"endo"	3	bebendo→ beb
"indo"	3	partindo→ part
"ondo"	3	propondo→ prop
"aram"	2	cantaram→ cant
"arde"	2	cantarde → cant
"arei"	2	cantarei→ cant
"arem"	2	cantarem→ cant
"aria"	2	cantaria→ cant
"armo"	2	cantarmo→ cant
"asse"	2	cantasse → cant
"aste"	2	cantaste→ cant
"avam"	2	cantavam→ cant
"ávei"	2	cantávei→ cant
"eram"	3	beberam→ beb
"erde"	3	beberde→ beb
"erei"	3	beberei→ beb
"êrei"	3	bebêrei→ beb
"erem"	3	beberem→ beb
"eria"	3	beberia→ beb
"ermo"	3	bebermo→ beb
"esse"	3	bebesse→ beb
"este"	3	bebeste→ beb
"íamo"	3	bebíamo→ beb
"iram"	3	partiram→ part
"íram"	3	concluíram→ conclu
"irde"	2	partirde→ part
"irei"	3	partírei→ part
"irem"	3	partirem→ part

"iria"	3	partiria→ part
"irmo"	3	partirmo→ part
"isse"	3	partisse→ part
"iste"	4	partiste→ part
"amo"	2	cantamo→ cant
"ara"	2	cantara→ cant
"ará"	2	cantará→ cant
"are"	2	cantare→ cant
"ava"	2	cantava→ cant
"emo"	2	cantemo→ cant
"era"	3	bebera→ beb
"erá"	3	beberá→ beb
"ere"	3	bebere→ beb
"iam"	3	bebiam→ beb
"íei"	3	bebíei→ beb
"imo"	3	partimo→ part
"ira"	3	partira→ part
"irá"	3	partirá→ part
"ire"	3	partire→ part
"omo"	3	compomo→ comp
"ai"	2	cantai→ cant
"am"	2	cantam→ cant
"ear"	4	barbear→barb
"ar"	2	cantar→ cant
"uei"	3	cheguei → cheg
"ei"	3	cantei→ cant
"em"	2	cantem→ cant
"er"	2	beber→ beb
"eu"	3	bebeu→ beb
"ia"	3	bebia→ beb
"ir"	3	partir→ part
"iu"	3	partiu→ part
"ou",	3	chegou→ cheg
"i"	3	bebi→ beb

Tabela 10 - Regras de redução de sufixos verbais

Passo 7: Remoção de vogais

Este passo consiste em remover as últimas vogais (*"a"*, *"e"* ou *"o"*) das palavras que não foram tratadas pelos passos 5 e 6. A Tabela 11 mostra o exemplo da palavra *menino* que não sofre nenhuma modificação pelos passos anteriores, no entanto, o passo 7 irá remover o final *–o*.

Sufixo a ser removido	Tamanho mínimo	Modificador	Exemplo
"a"	3		menina → menin
"e"	3		grande → grand
"o"	3		menino → menin

Tabela 11 - Regras de remoção de vogais

Passo 8: Remoção de acentos

A remoção dos acentos é necessária porque há casos em que algumas variações da palavra são acentuadas e outras não, por exemplo, *psicólogo* e *psicologia*, ambas com o radical *psicolog*. Este passo é feito somente no final do processo devido à quantidade de regras que precisam ser verificadas previamente.

V.4 - Utilização das Palavras Relevantes

Conforme visto na Figura 25, a extração terá como resultado a lista de palavras relevantes de um determinado texto. A partir destas palavras, o minerador irá manipular as consultas e estimar a relevância de cada resultado obtido. As palavras relevantes são armazenadas em dois formatos: original, tal como está no texto e radicalizada, resultado da aplicação do algoritmo de *stemming*. Desta forma, as consultas tendem a obter graus de precisão mais acurados na medida em que palavras como *texto*, *textos* e *textuais* são considerados da mesma forma.

Este mesmo tratamento de "limpeza" dado aos textos também é realizado nas palavras de consulta. Por exemplo, quando o sistema verifica se uma questão de prova está contida em um determinado livro, há a extração de palavras relevantes da questão assim como a extração de palavras relevantes do livro. Desta forma, ambos os conjuntos de palavras poderão ser comparados sem divergências de tratamento.

O passo seguinte à extração de palavras relevantes é a utilização do algoritmo que trará os resultados da consulta. Os algoritmos utilizados neste trabalho são baseados nos Modelos de Recuperação de Informação apresentados na seção *III.3 - Modelos de Recuperação de Informação*.

Como forma de melhorar a apresentação dos resultados, sobretudo no cálculo das relevâncias, a fórmula de Salton (1987) foi utilizada para normalizar os resultados. Os detalhes deste cálculo bem como as formas de apresentação e utilização das ferramentas desenvolvidas serão apresentados na seção seguinte, que mostra ainda o funcionamento das interfaces do objeto desta dissertação – o SMiner.

VI - FERRAMENTAS DO SMINER

Como foi abordado anteriormente, os dados pertinentes ao contexto analisado (por exemplo: livros, questões, dicionários, críticas e sugestões) tiveram que ser manipulados e tratados para servirem de insumos das interfaces de buscas.

Desta forma, surgiu o SMiner, um *software web*, desenvolvido em Java, utilizando a arquitetura EJB (Enterprise Java Beans) e banco de dados Oracle 9i. O grande trabalho de preparação dos dados do contexto analisado e a criação de interfaces simples fornecem ao SMiner uma flexibilidade de utilização, na medida em que o sistema mostra-se eficiente em consultas simples e avançadas bem como nas buscas contextuais, onde o posicionamento e a proximidade das palavras são igualmente relevantes no resultado final.

A partir da criação dos módulos de consultas, onde o usuário pode encontrar vários tipos de documentos com o uso de palavras chave, foi desenvolvido um conjunto de mecanismos específicos para a comparação de questões de prova e os respectivos livros de referência.

Nas seções a seguir, serão apresentadas as principais funcionalidades do SMiner. Inicialmente, são descritas as interfaces de consultas e sua utilização bem como o relatório de lacunas entre questões e livros. Em seguida, os mecanismos de manutenção e configuração do sistema serão elucidados. E, por fim, são apresentadas algumas ferramentas extras que auxiliam a compreensão do processo de mineração.

VI.1 - CONSULTAS

O SMiner provê mecanismos de consulta aos insumos cadastrados na base de dados. As buscas são realizadas sobre os Livros, Documentos de Referência, Questões de Prova, Materiais de Apoio Pedagógico, Críticas e Sugestões. Existem 2 tipos de consultas básicas: por palavra chave, que verifica exatamente o texto procurado, sem qualquer tratamento e avançada, utilizando recursos que ampliam o espectro da consulta.

Na verdade, um usuário mais habituado com a ferramenta poderá utilizar operadores como E, OU, NÃO na busca por palavras chave e obter, desta forma, os

mesmos recursos dispostos na busca avançada. A diferenciação entre estas buscas é apenas para tornar a utilização destas ferramentas mais intuitiva aos mais leigos.

Além disso, um mecanismo baseado no posicionamento e proximidade das palavras buscadas foi desenvolvido – a Busca Contextual. Esta ferramenta possibilita que as palavras de uma consulta "química geral aplicada" sejam encontradas somente se estiverem próximas, ou até mesmo respeitando a seqüência entre elas. Isto impossibilita, por exemplo, que consultas por "Rio de Janeiro" retornem "... em janeiro o rio costuma encher...".

Com a criação dos algoritmos para a busca contextual foi possível gerar uma busca por lacunas entre questões e livros. As probabilidades de acerto desta busca estão diretamente ligadas à forma como estão dispostas as palavras relevantes, o que torna o procedimento mais difícil. Entretanto, foi concebida a utilização de um pequeno *thesaurus* destinado ao contexto desta dissertação e também serão abordados no decorrer do texto.

Desta forma, esta seção apresenta as características de utilização das consultas seguida, naturalmente, pela explanação sobre a usabilidade das buscas Simples, Avançada e Contextual. Após esta explanação, a ferramenta de lacunas entre questões e livros é abordada e como são calculadas as relevâncias de cada documento encontrado.

VI.1.1 - CARACTERÍSTICAS

Nesta seção, poderão ser observadas as características comuns às buscas do SMiner. Algumas características estão intrínsecas ao processo de extração de conhecimento, entretanto, há procedimentos dependentes da vontade do usuário, como por exemplo, a utilização de operadores avançados.

VI.1.1.1 - Procedimentos Básicos

Como a maioria das ferramentas de busca, para realizar uma consulta basta digitar uma ou mais palavras e clicar no botão Procurar. Os resultados desta busca são organizados pela relevância dos textos em relação à busca desejada. Esta relevância é baseada na freqüência dos termos da consulta em relação aos insumos da base de dados, o que será melhor apresentado nas seções seguintes.

São insumos do processo: Livros, Documentos de Referência, Termos Técnicos, Questões e Respostas, Documentos de Opinião (críticas & sugestões) e outros Materiais de Apoio Didático.

São exemplos de pesquisa:

- **'A B'** retorna somente os insumos onde podem ser encontradas as palavras A B juntas.
- **'A & B'** retorna os insumos onde as palavras A e B estão contidas no texto, independente de sua localização.
- **'A | B'** retorna os insumos onde é possível encontrar a palavra A ou a palavra B.
- O mesmo procedimento pode ser utilizado com um conjunto maior de palavras.

VI.1.1.2 - Palavras Irrelevantes e Radicalizador

Na consulta, a ferramenta propicia ao usuário escolher se ignora ou não as palavras e caracteres comuns, conhecidos como Irrelevantes. A opção, quando checada, retira automaticamente palavras como: preposições, artigos, termos comuns, entre outros. Estas palavras podem ser comumente descartadas por não afetarem significativamente o resultado da busca, além de torná-la consideravelmente mais lenta.

Além disto, as buscas contam com a radicalização de palavras. Este procedimento retira os sufixos dos termos aumentando as chances de encontrar o item procurado. Palavras como 'texto', 'textos' e 'textuais' assumiriam o radical 'text'. Desta forma, a probabilidade de encontrar um destes termos é aumentada significativamente.

VI.1.1.3 - Maiúsculas & minúsculas

As buscas no sistema não são sensíveis a Maiúsculas e minúsculas. Por exemplo, pesquisar por 'espanhol', 'ESPANHOL' ou 'EsPAnhoL' irá mostrar os mesmos resultados.

VI.1.1.4 - Acentuação

Por padrão, as pesquisas não são sensíveis a acentos. Ou seja, 'freqüência' e 'frequencia' são procurados da mesma forma.

VI.1.1.5 - Operadores

A ferramenta possibilita a manipulação dinâmica das entradas para as consultas. É comum tentarmos inferências avançadas, onde as palavras possuem uma ordem específica de procura. Por exemplo, um usuário pode querer encontrar documentos que contenham as palavras 'física' ou 'química', ou ainda, que contenham 'português' mas não possuam 'matemática' no seu texto. Nestes tipos de situação, alguns artifícios podem ser utilizados:

& Operador E

Este símbolo é utilizado para realizar consultas onde as palavras associadas devem, obrigatoriamente, aparecer no texto.

Ex: português & matemática → serão retornados os insumos que contenham as palavras 'português' e 'matemática'.

| Operador OU

Este símbolo é utilizado para realizar consultas onde pelo menos uma das palavras procuradas aparece no texto.

Ex: português | matemática → serão retornados os insumos que contenham a palavra 'português' ou 'matemática'. Os textos que possuam ambas as palavras também serão naturalmente apresentados.

~ Operador NÃO

É possível excluir uma palavra da busca colocando um sinal negativo ("-") imediatamente na frente do termo que você queira evitar.

Ex: física ~ eletricidade → retorna todos os insumos que contenham a palavra 'física' mas não possuam a palavra 'eletricidade'.

VI.1.1.6 - Visualização dos Resultados

Os resultados das buscas incluem as inferências encontradas nos diversos insumos do sistema. Para melhor compreensão, estes resultados são classificados por grau de relevância (do maior para o menor). Além disto, são separados por tipos de insumo (livros, termos técnicos e questões, por exemplo).

Em cada resultado, são apresentados o nome do insumo, sua relevância e um link para visualização do conteúdo completo do texto ou arquivo. O sistema contempla diversos tipos de arquivo, como por exemplo: arquivos texto (TXT), Adobe Portable Document Format (PDF), Microsoft Word (DOC), entre outros.

VI.1.2 - BUSCA SIMPLES, POR COINCIDÊNCIA DE PALAVRAS

A Busca Simples é o mecanismo de consulta mais simples desta ferramenta. Esta funcionalidade simplesmente compara as palavras digitadas no campo de busca, respeitando os procedimentos básicos, e retorna o conjunto de insumos encontrados.

O exemplo da Figura 28, apresenta uma consulta contendo a expressão "eletrodo revestido" como termo de busca. Os resultados são apresentados conforme seu grau de relevância entre os insumos.

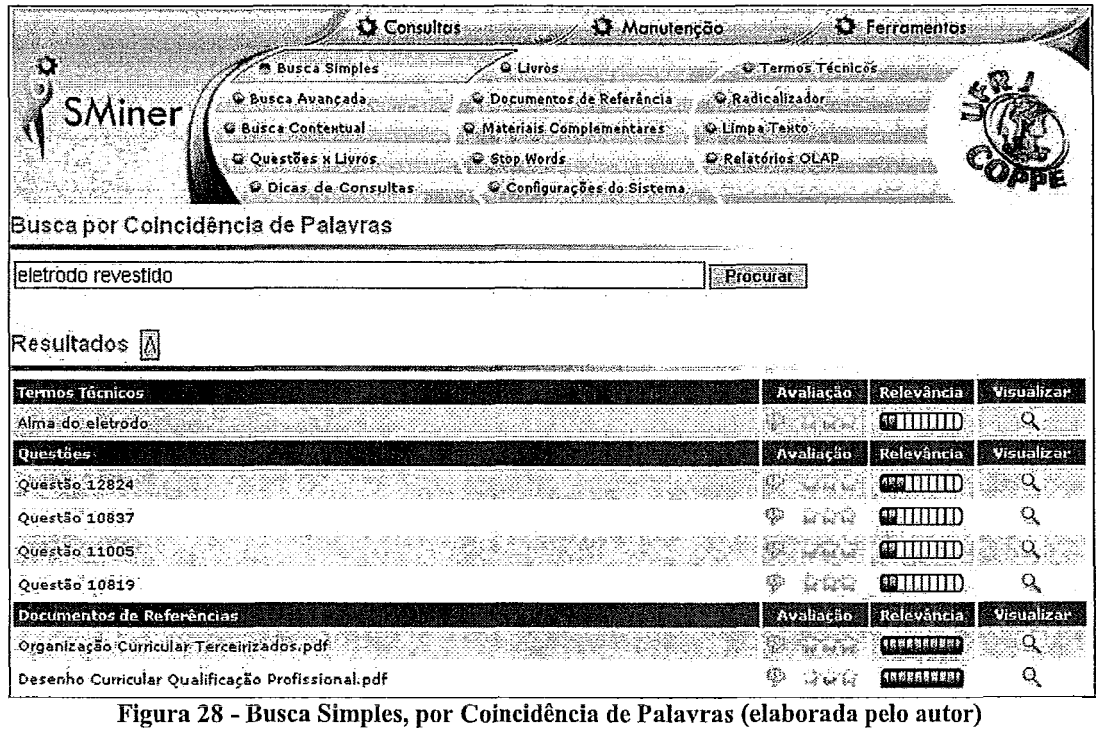


Figura 28 - Busca Simples, por Coincidência de Palavras (elaborada pelo autor)

É importante ressaltar que o termo da busca é procurado no interior (conteúdo) dos documentos, o que explica não haver a palavra "eletrodo" na descrição dos insumos encontrados. A Figura 29 apresenta um exemplo de visualização do Termo Técnico "Alma do eletrodo", onde pode ser visto a expressão "eletrodo revestido" no destaque.

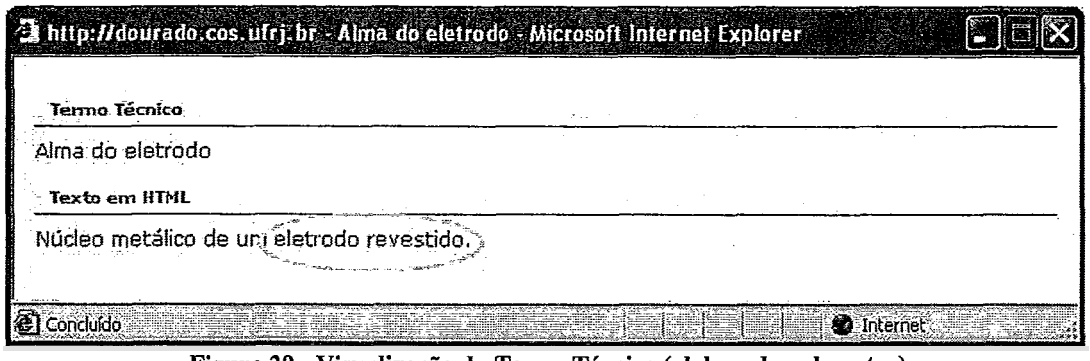


Figura 29 - Visualização do Termo Técnico (elaborada pelo autor)

Por outro lado, pode-se tornar esta busca um pouco mais avançada utilizando operadores e/ou caracteres especiais, conforme descrito na seção anterior. A Figura 30 apresenta um exemplo de busca neste sentido, onde são procurados insumos que contenham os termos "windows" e "powerpoint" no mesmo conteúdo.

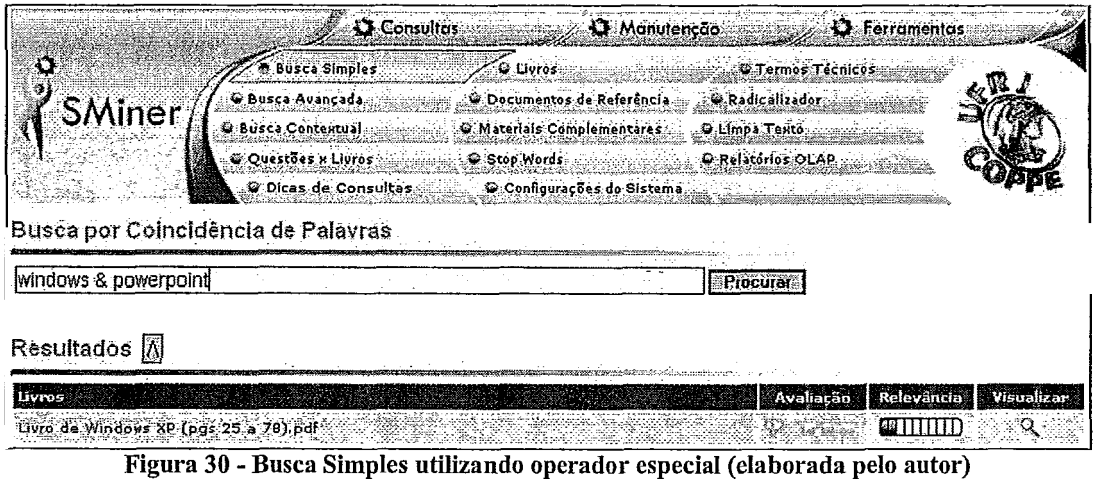


Figura 30 - Busca Simples utilizando operador especial (elaborada pelo autor)

VI.1.3 - BUSCA AVANÇADA

Para um usuário leigo, adicionar palavras na busca simples irá reduzir o âmbito da busca até encontrar o que pretende. No entanto, se o usuário não tiver afinidades com Operadores ou Caracteres especiais, este refinamento pode se tornar muito difícil ou até mesmo inviável.

No intuito de viabilizar buscas avançadas, foi desenvolvida uma interface amigável para ajudar o usuário em refinamentos específicos. Desta forma, o SMiner também fornece possibilidades de busca que:

- Contenham todas as palavras definidas pelo usuário;
- Contenham exatamente a expressão definida pelo usuário;

- Apareça qualquer uma das palavras definidas pelo usuário;
- NÃO possuam as palavras definidas pelo usuário;

Além disto, é possível ainda:

- Restringir a pesquisa apenas aos insumos determinados;
- Retirar palavras irrelevantes e aplicar radicalização na busca;

A ferramenta de busca avançada permite aplicar facilmente estas preferências à pesquisa através de uma interface amigável baseada em alguns pontos da Pesquisa Avançada do Google (2006), conforme pode ser visto na Figura 31.

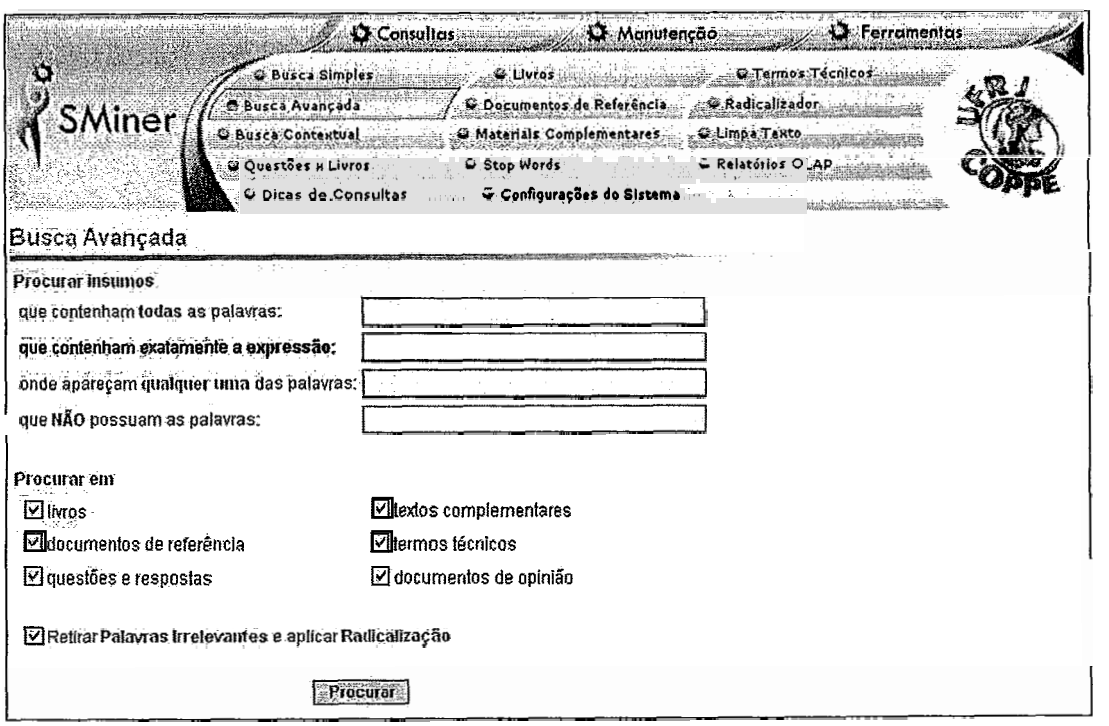


Figura 31 - Interface de Busca Avançada (elaborada pelo autor)

È importante ressaltar que é possível realizar a combinação das opções disponíveis na Busca Avançada, conforme visto na Figura 32.

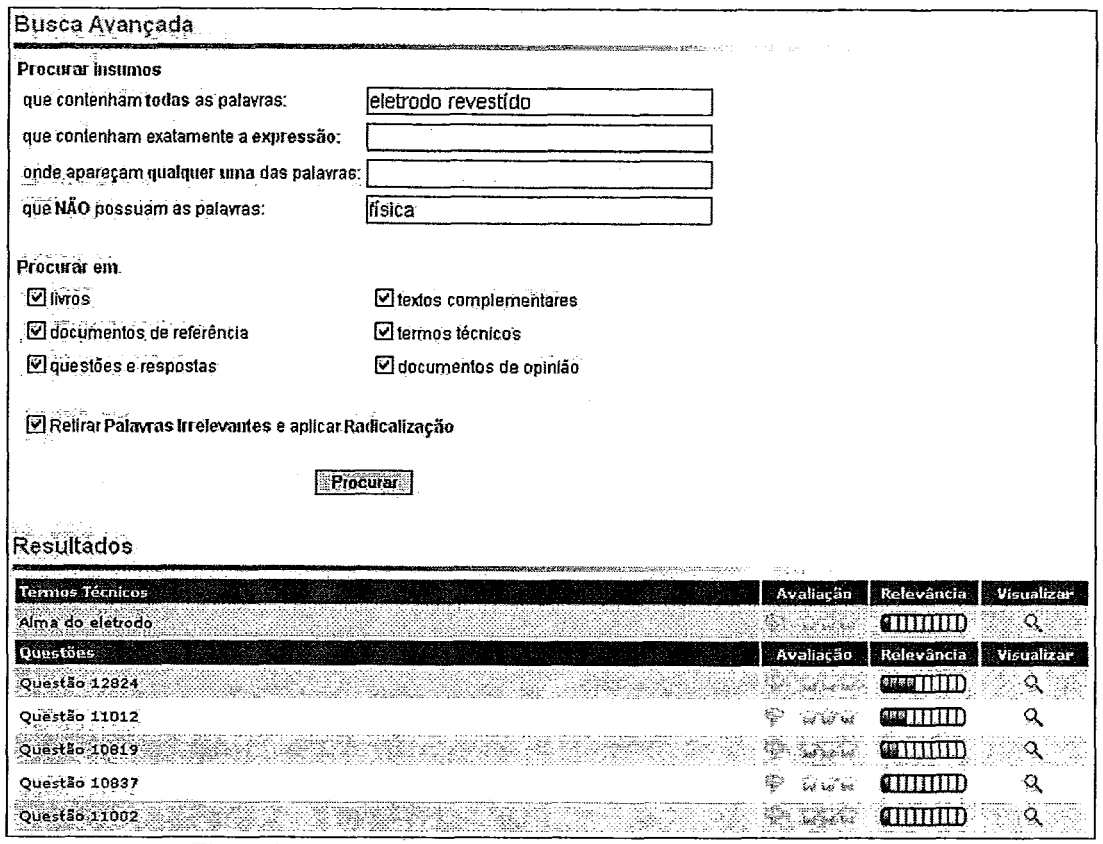


Figura 32 - Resultados de Busca Avançada (elaborada pelo autor)

VI.1.4 - BUSCA CONTEXTUAL

A busca contextual foi desenvolvida tendo em vista os problemas de vocabulário e posicionamento de palavras na busca. São exemplos: procurar por "Física" e não obter os documentos que contenham palavras como "Eletricidade", "Eletrodo" ou "Termodinâmica", por exemplo. Ou ainda, procurar por "Rio de Janeiro" e obter como resposta um documento que contenha "... em Janeiro o rio estava seco...".

Através da utilização de busca por contexto, espera-se determinar graus de relação mais adequados que identifiquem melhor quais termos são realmente importantes em um conjunto de documentos (SALTON & BUCKLEY, 1987).

A abordagem mais usual, definida por Chen (1996), define o contexto como sendo um conjunto de palavras que representam o assunto ou a área do conhecimento, que podem ser auxiliadas com o uso de Thesaurus, dicionários de sinônimos, redes semânticas e clusters (SALTON, 1983).

O SMiner combina a utilização de um pequeno *thesaurus* e o posicionamento das palavras extraídas dos documentos para determinar o contexto da busca. A interface

contém as opções de escolha de insumos (*Procurar em*) e a possibilidade de *Retirar Palavras Irrelevantes e aplicar Radicalização*. A novidade fica por conta da avaliação a partir da posição das palavras e a opção de utilização de *thesaurus* (Figura 33).

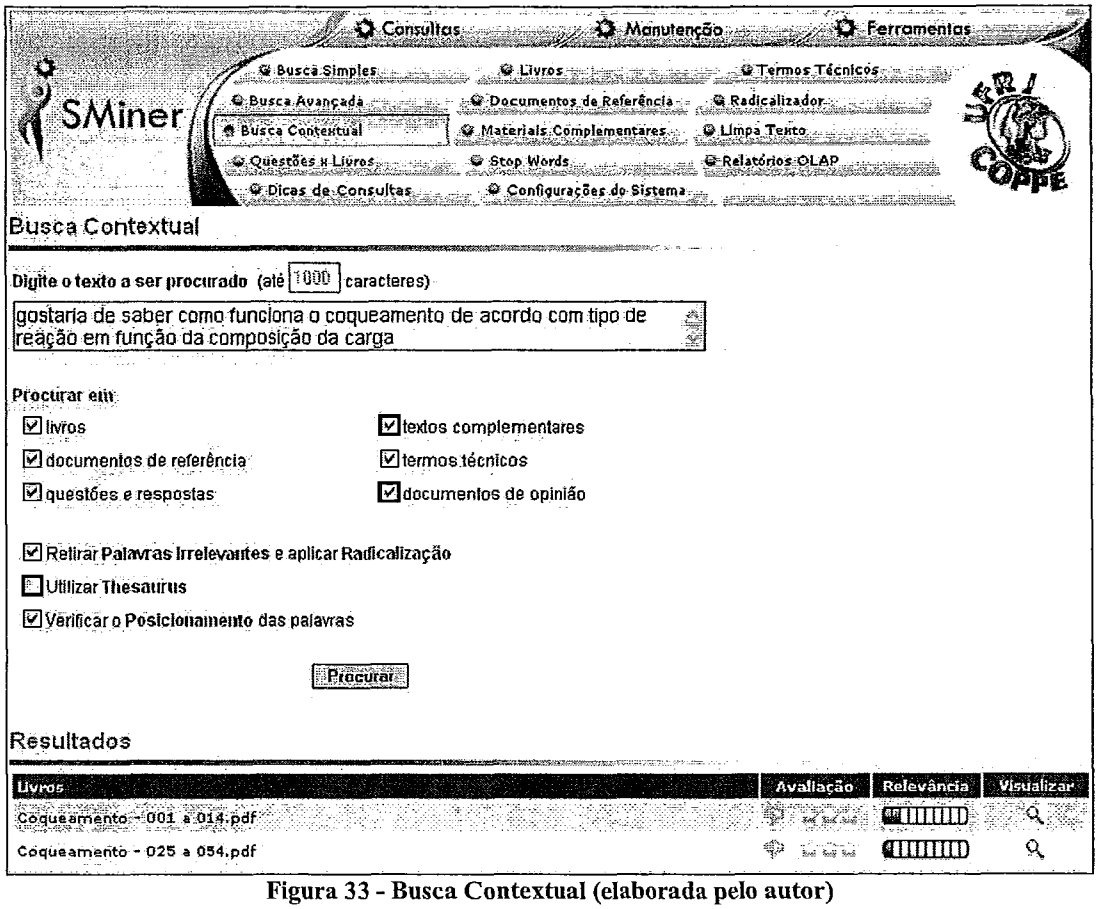


Figura 33 - Busca Contextual (elaborada pelo autor)

Tecnicamente, o algoritmo de busca contextual é ligeiramente diferente das demais buscas. A busca verifica se as palavras contidas no texto procurado estão próximas, havendo ainda a possibilidade de verificar se estão na seqüência em que foram digitadas. A Figura 34 apresenta o fluxo básico da Busca Contextual.

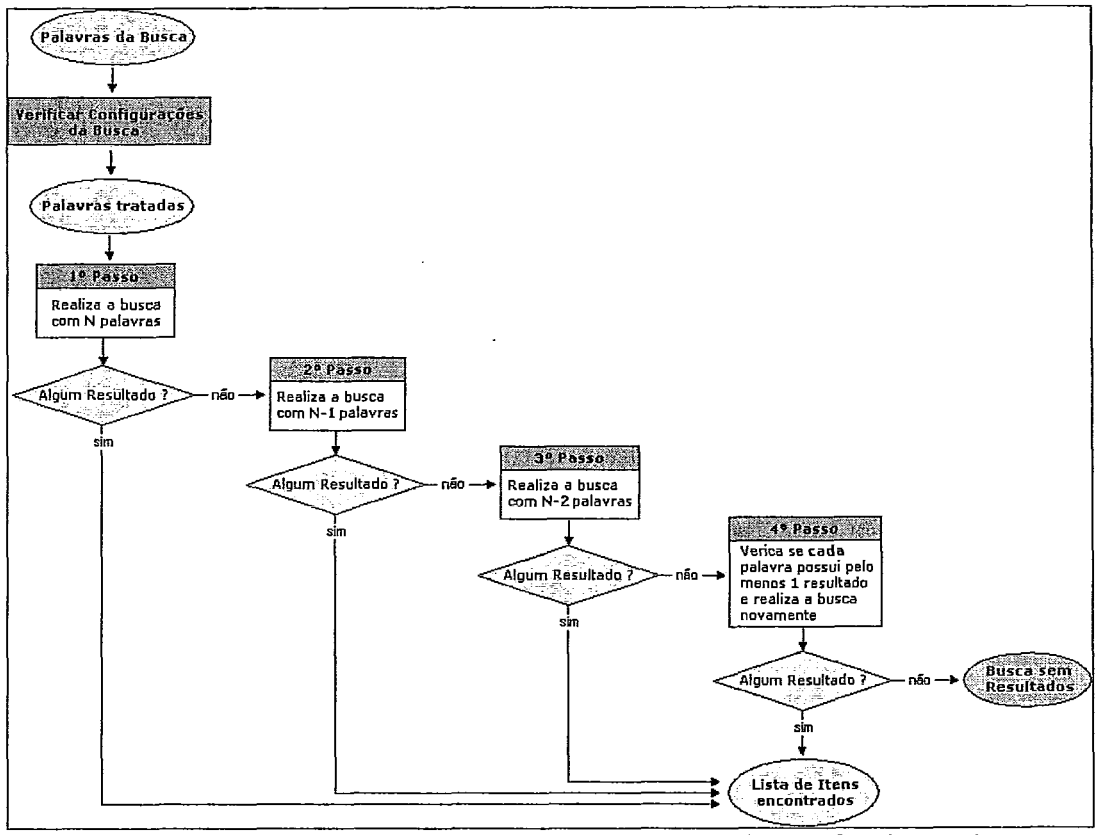


Figura 34 – Fluxo e possíveis passos da Busca Contextual (elaborada pelo autor)

O algoritmo da busca contextual inicia-se recuperando a lista de palavras (palavras "soltas" ou frases) digitadas pelo usuário. Em seguida, são verificadas as opções de configuração da busca:

- Procurar em – conforme visto na Busca Avançada, este recurso permite que o algoritmo manipule somente os itens escolhidos pelo usuário;
- Retirar Palavras Irrelevantes e Aplicar Radicalização – igualmente abordado, esta opção desconsidera Palavras Irrelevantes ("o", "a", "de", entre outros) e aplica a Radicalização ("textos" → "text", por exemplo) nas palavras originalmente procuradas;
- Utilizar Thesaurus – esta funcionalidade manipula cada palavra da busca de forma a incluir os itens cadastrados no Thesaurus, por exemplo, se na busca contiver a palavra "Física", os termos "Eletricidade", "Termodinâmica", "Mecânica", entre outros, serão igualmente adicionadas à consulta, o que aumenta significativamente o espectro da busca.

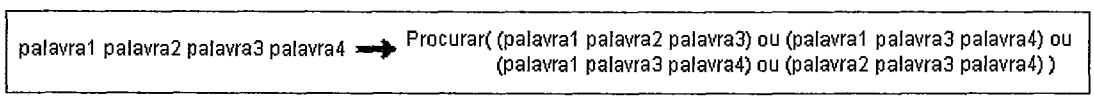
- Verificar o posicionamento das palavras – o algoritmo verificará a necessidade ou não de conferir se as palavras contidas na busca estão igualmente seqüenciadas nos itens encontrados.

Após a Verificação das Configurações da Busca, o algoritmo passa a trabalhar com as N Palavras Tratadas e inicia o fluxo interno da busca:

- 1° Passo: Realiza a busca com N palavras – o algoritmo irá procurar pelos insumos que contenham exatamente as N palavras (tratadas) da busca.

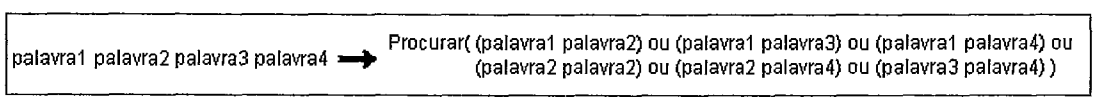
Se o algoritmo não retornar resultados, o 2° passo é acionado;

- 2° Passo: Realiza a busca com N–1 palavras – o algoritmo irá procurar pelos insumos que contenham N–1 palavras da busca. Neste caso, haverá uma combinação entre as palavras de forma que cada palavra esteja ausente na busca.



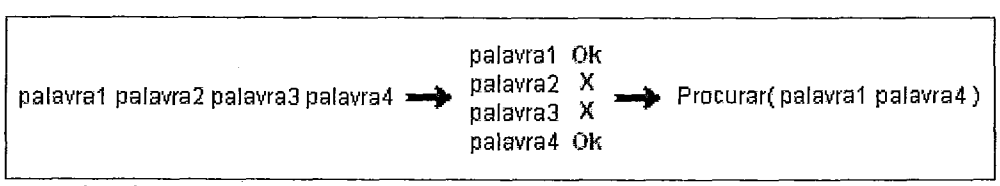
Se o algoritmo não retornar resultados, o 3° passo é acionado.

- 3° Passo: Realiza a busca com N–2 palavras – o algoritmo irá procurar pelos insumos que contenham N–2 palavras da busca. Neste caso, haverá uma combinação entre as palavras de forma que 2 palavras estejam ausentes.



Se o algoritmo não retornar resultados, o 4° passo é acionado.

- 4° Passo: Verifica se cada palavra possui pelo menos 1 resultado – este mecanismo verifica se todas as palavras possuem, independentemente, resultados associados. Após a verificação de cada palavra, o algoritmo desconsidera aquelas que não estejam contidas em documento algum.



Se o algoritmo não retornar resultados, a busca é finalizada sem sucesso.

É importante observar que, independente das configurações iniciais, a busca contextual trabalha com a proximidade entre as palavras procuradas. A configuração

padrão sugere que as palavras (tratadas) estejam a uma distância de até 100 palavras uma das outras. O algoritmo considera ou não a ordem de posicionamento de acordo com as configurações estabelecidas pelo usuário. As buscas Simples e Avançada não possuem estes recursos.

VI.1.5 - LACUNAS – QUESTÕES X LIVROS

A utilização da interface de visualização das Lacunas da Certificação é motivada pelas possíveis divergências entre as questões de prova os livros nos quais se baseiam a criação das questões. Conforme mencionado na seção *IV.3.2 -Elaboração das Questões dos Diagnósticos*, o contexto desta dissertação assume que as provas sejam baseadas nos livros. Desta forma, o objetivo desta interface é estimar a probabilidade de um conjunto de questões estarem contidos em um determinado livro.

A utilização da ferramenta é bem simples. O usuário pode optar pela digitação de uma possível questão e escolher o respectivo livro que aborde o assunto (Figura 35) ou definir a especialidade das questões / livros que serão manipulados (Figura 36). Internamente, o algoritmo que efetua esta assimilação é análogo ao da Busca Contextual.

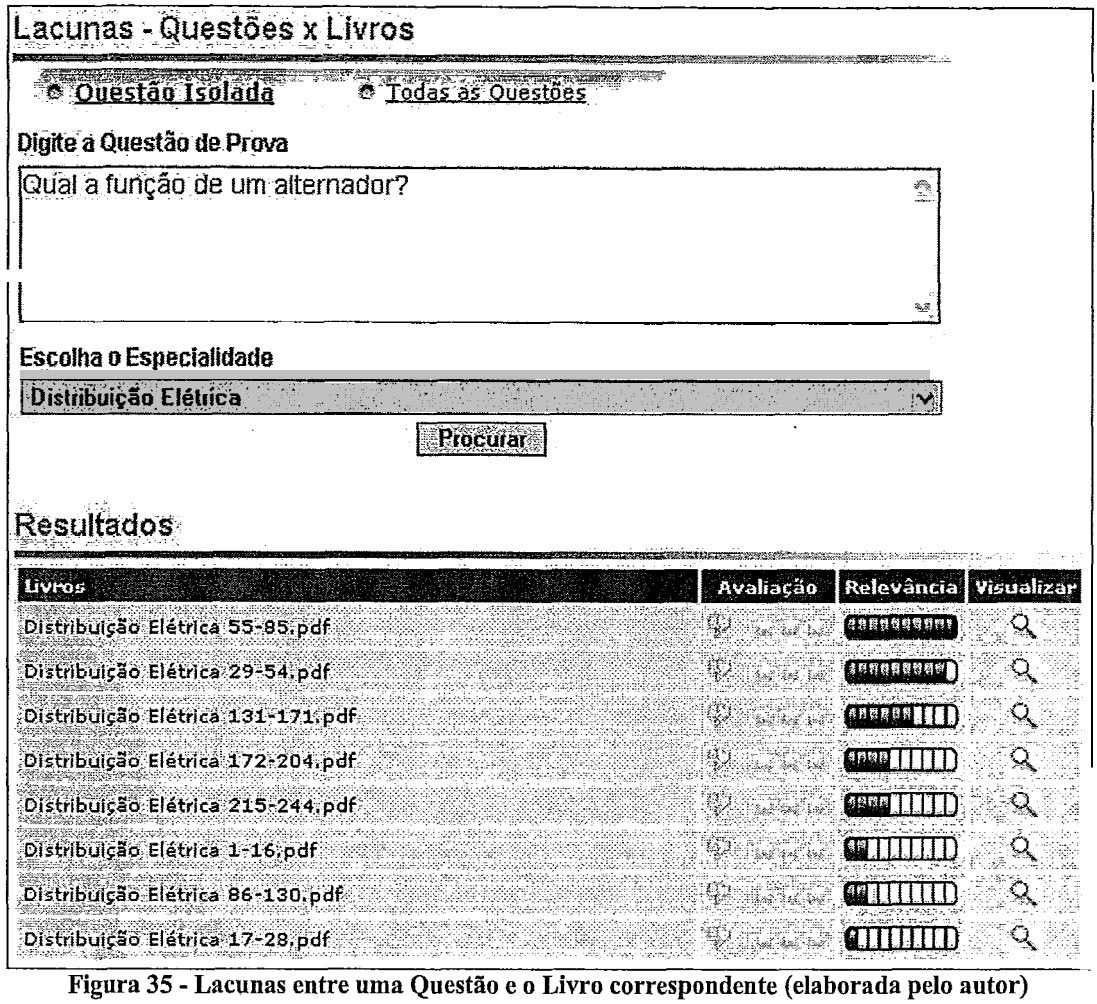


Figura 35 - Lacunas entre uma Questão e o Livro correspondente (elaborada pelo autor)

A Figura 35 ilustra a procura de similaridades entre a Questão digitada pelo usuário e o respectivo Livro. Neste caso, o sistema irá realizar uma busca contextual considerando as palavras tratadas (Retirada de Palavras Irrelevantes, por exemplo) e tendo como item de busca somente o livro especificado.

Por outro lado, o usuário pode desejar avaliar todas as questões de uma **determinada competência (ou especialidade) cadastradas no banco de dados.** Para realizar esta verificação, basta definir a Especialidade das questões / livros que serão manipulados (Figura 36).

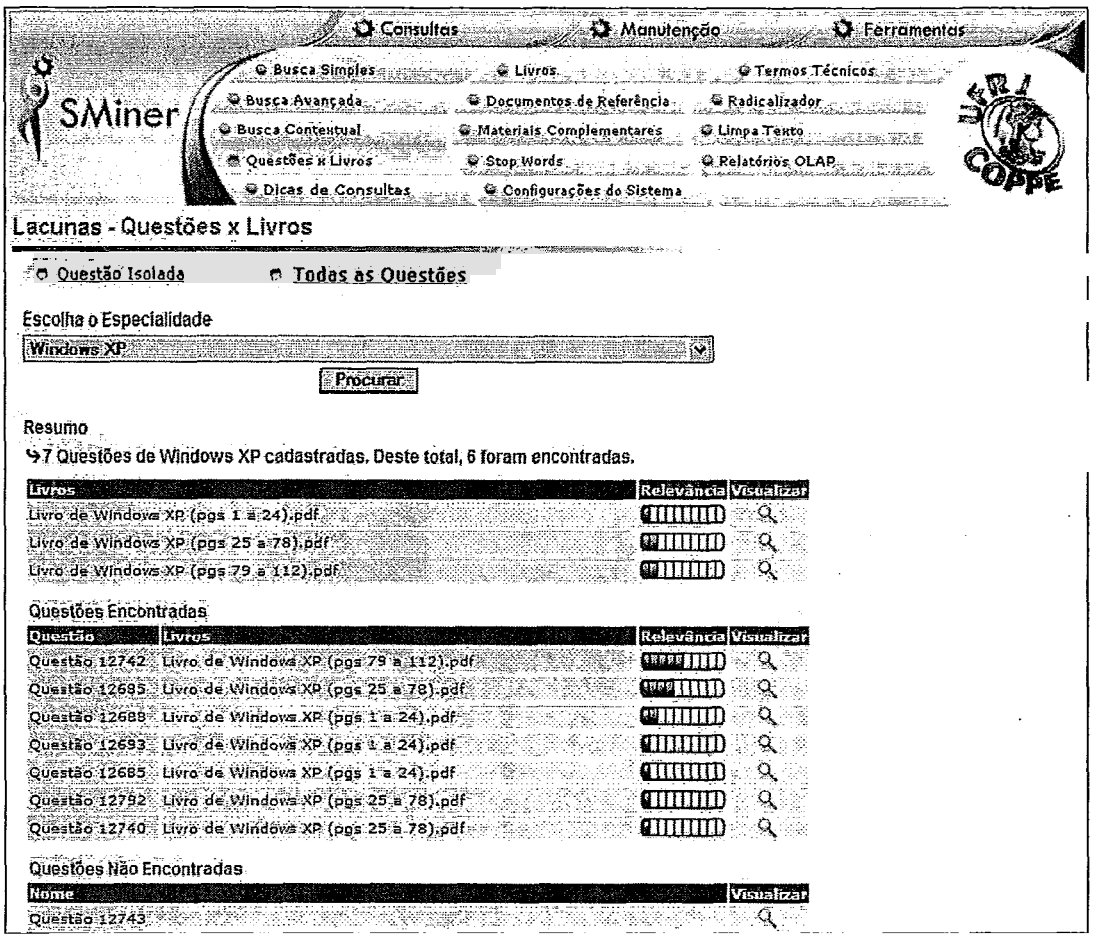


Figura 36 - Lacunas entre um conjunto de Questões e o Livro correspondente (elaborada pelo autor)

No exemplo ilustrado na Figura 36, o sistema irá efetuar buscas contextuais entre cada questão da especialidade "Windows XP" e o respectivo livro. Desta forma, todas as questões desta especialidade serão manipuladas e o sistema irá retornar a probabilidade de cada questão estar no livro referente.

Como se trata de um conjunto de questões há um campo denominado "Questão", onde o usuário pode verificar o texto da questão apresentada, além do habitual *link* "Visualizar", que permite a visualização do insumo encontrado.

Os cálculos das relevâncias dispostos nas buscas, bem como na interface de lacunas são similares. Este cálculo é normalizado para facilitar a compreensão do usuário e será melhor elucidado na seção a seguir.

VI.1.6 - CÁLCULO DA RELEVÂNCIA

O cálculo da relevância apresentada nos resultados das buscas foi baseado na fórmula de Salton (1987):

$$3 * f * \left(1 + \log\left(\frac{N}{n}\right)\right)$$

Figura 37 - Fórmula de Salton (elaborada pelo autor)

Onde:

- f = a freqüência do termo procurado no documento;
- N = o número de documentos existentes;
- n = o número de documentos que contém o termo procurado.

Como forma de facilitar a visualização no sistema, a relevância pode estar entre 0 e 100. Para encontrar este valor a partir da consulta, o sistema utiliza o cálculo por freqüência inversa baseada na fórmula de Salton (1987), presumindo que termos com freqüência em muitos documentos podem ser considerados de baixa relevância. Desta forma, para um documento ser considerado relevante, os termos de consulta devem possuir boa freqüência no documento em questão e pouca freqüência nos demais documentos da base de dados.

A Tabela 12 mostra um exemplo aproximado de cálculo de relevância assumindo apenas 1 documento contém o termo consultado.

Número de documentos cadastrados	Ocorrência no documento para obter Relevância = 100
1	34
5	20
10	17
50	13
100	12
500	10
1.000	9
10.000	7
500.000	5
50.000.000	4
10.000.000.000	3

Tabela 12 - Cálculo de relevância onde termo de consulta consta em apenas 1 documento

Como pode ser observado, se houver apenas um documento na base, o termo deverá ter uma freqüência igual a 34. Por outro lado, se a base contiver 500.000 documentos, basta 5 repetições do termo no único documento que o contém para atingir a relevância igual a 100.

VI.2 - MANUTENÇÃO DO SISTEMA

Algumas interfaces são necessárias para garantir a manutenibilidade do Sistema:

- Manutenção de Livros;
- Manutenção de Documentos de Referência;
- Manutenção de Textos Complementares;
- Manutenção de Palavras Irrelevantes;
- Configurações do Sistema;

As interfaces de manutenção são precedidas de uma listagem e possibilitam o cadastro, a alteração e a exclusão do item referido. A interface de configurações possibilita a manutenção de alguns itens de configuração bem como a reindexação dos índices que auxiliam nas buscas.

VI.2.1 - MANUTENÇÃO DE LIVROS

Os livros são os principais materiais utilizados como referência de estudo. Para acessar a área de manutenção dos livros, a ferramenta apresenta inicialmente uma lista com todos os livros cadastrados no sistema. A Figura 38 apresenta a listagem onde o usuário poderá cadastrar, alterar ou excluir um livro da base de dados.

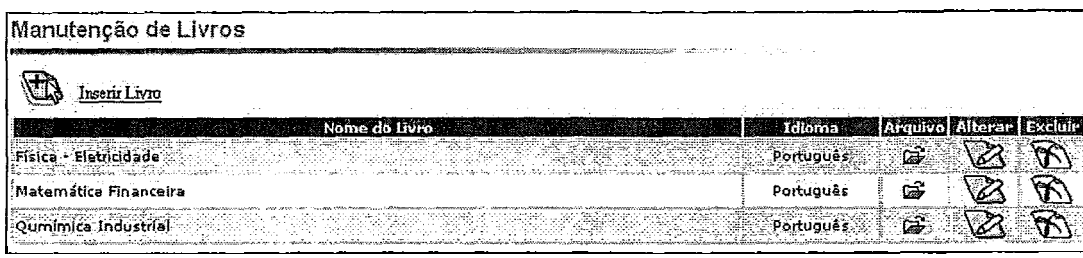


Figura 38 - Listagem de Livros (elaborada pelo autor)

O cadastro dos livros, apresentado na Figura 39, segue um fluxo peculiar em função da mineração do texto no momento do seu cadastramento. Este procedimento pode incorrer em uma demora na efetivação do cadastro, o que motivou a possibilidade de indexação a posteriori, informada em uma opção nas configurações do sistema. Desta forma, o processo de mineração (retratado na seção *V.2 - Extração das Informações Textuais*) pode ser feito no momento do cadastro ou através da funcionalidade Indexação de Livros, disposta na área de Configurações do Sistema.

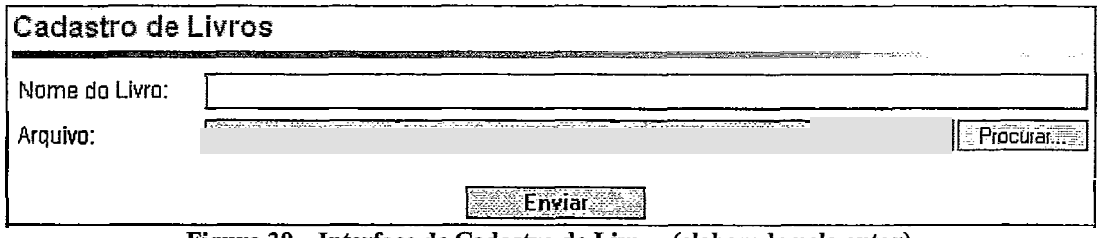


Figura 39 – Interface de Cadastro de Livros (elaborada pelo autor)

O sistema conta ainda com a possibilidade de alteração de algumas informações dos livros, bem como a exclusão na base de dados.

VI.2.2 - MANUTENÇÃO DE DOCUMENTOS DE REFERÊNCIA

Os Documentos de Referência contém informações que regulamentam os cursos; são documentos inerentes aos procedimentos de cada curso, contemplando o conteúdo formativo de cada área abordada. Em geral, são informações utilizadas no âmbito de Coordenadores e Gerentes.

Assim como os livros, a manutenção de documentos de referência é iniciada com uma interface de listagem. Através desta interface, o usuário poderá incluir, alterar ou excluir estes tipos de documentos. A Figura 40 e a Figura 41 mostram as telas de listagem e cadastro, respectivamente.

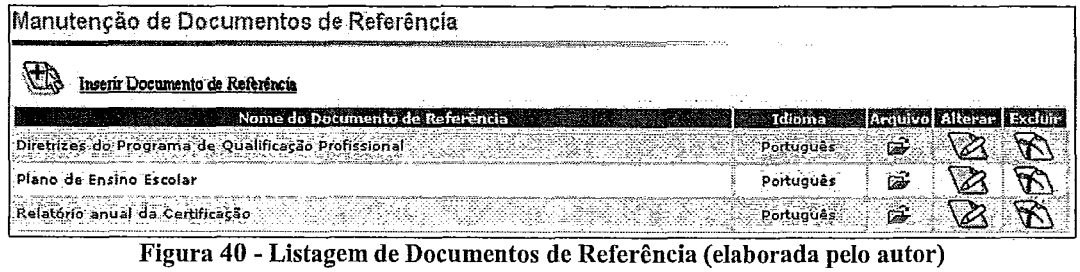


Figura 40 - Listagem de Documentos de Referência (elaborada pelo autor)

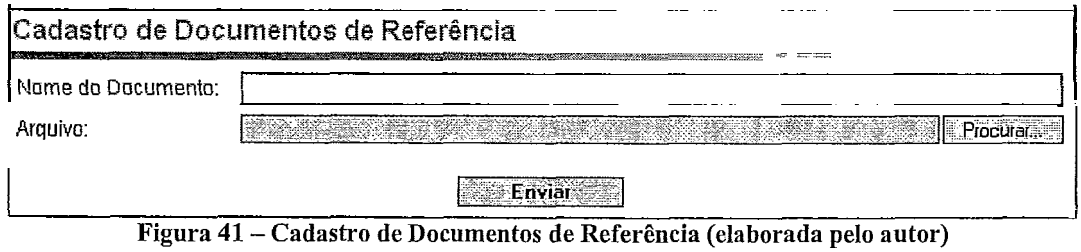


Figura 41 – Cadastro de Documentos de Referência (elaborada pelo autor)

VI.2.3 - MANUTENÇÃO DE MATERIAIS COMPLEMENTARES

Os Materiais Complementares correspondem aos Materiais de Apoio Didático complementares aos Livros, como por exemplo: apostilas, apresentações e exercícios. Analogamente às interfaces anteriores, a manutenção destes itens é feita a partir de uma

interface de listagem, que possibilita a inclusão, alteração e exclusão destes itens. A Figura 42 e a Figura 43 ilustram as interfaces de listagem e cadastramento.

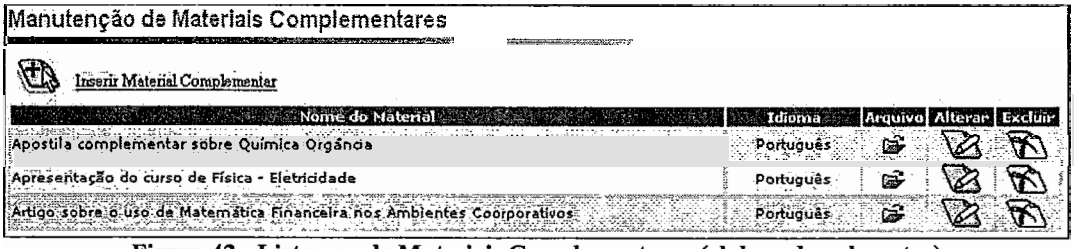


Figura 42 - Listagem de Materiais Complementares (elaborada pelo autor)

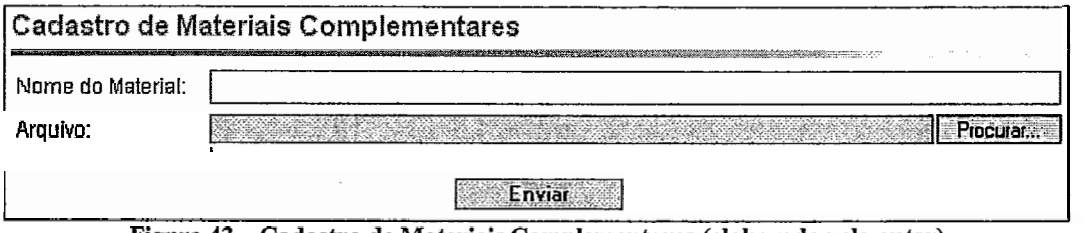


Figura 43 – Cadastro de Materiais Complementares (elaborada pelo autor)

VI.2.4 - MANUTENÇÃO DE STOP WORDS

As palavras irrelevantes, também conhecidas por *Stop Words* são importantíssimas no processo de mineração. Como visto anteriormente, existem palavras que podem ser descartadas em função da grande periodicidade em que elas aparecem em todos os textos, ou pela sua própria função, como por exemplo: adjetivos, advérbios e pronomes.

A manutenção segue o padrão visto nas seções anteriores, contendo uma listagem inicial, com opções de inclusão, alteração e exclusão de palavras irrelevantes. A Figura 44 ilustra a interface de cadastramento de *stop words*.

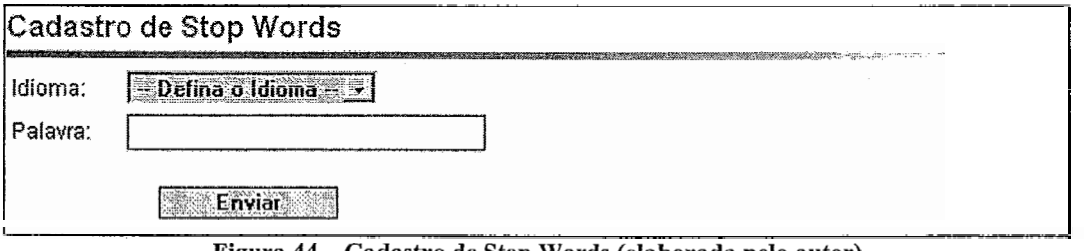


Figura 44 – Cadastro de Stop Words (elaborada pelo autor)

VI.2.5 - CONFIGURAÇÕES DO SISTEMA

A interface de Configurações do Sistema auxilia na forma como os insumos são procurados e na ativação do modo de avaliação do Sistema. Como pode ser visto na

Figura 45, ao salvar estas informações, os módulos de busca são automaticamente impactados. Além disso, um conjunto de estatísticas são apresentadas..

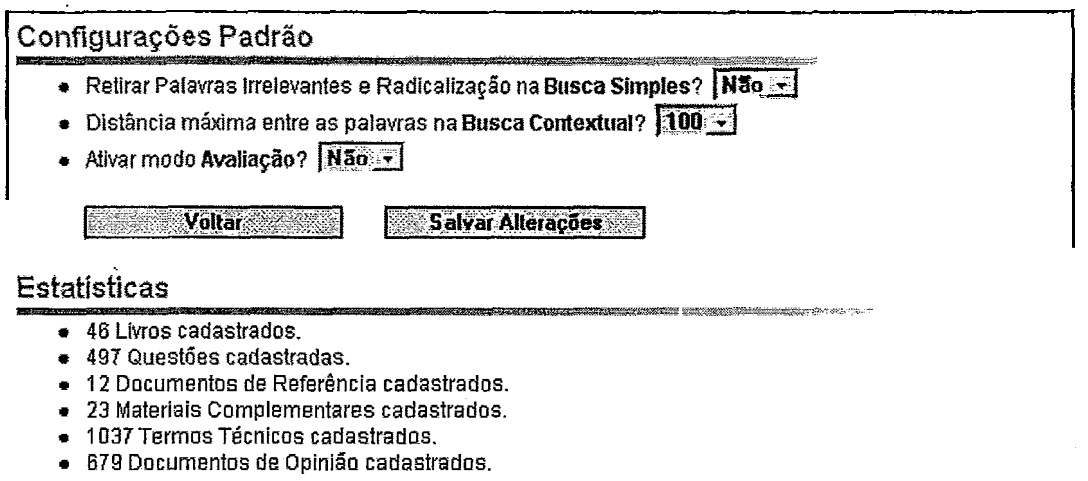


Figura 45 - Interface de Configurações do Sistema (elaborada pelo autor)

VI.3 - FERRAMENTAS EXTRAS

Como forma de ajudar a compreensão de alguns mecanismos utilizados na mineração dos textos, foi desenvolvido um módulo extra no SMiner contendo alguns aplicativos simples que ilustram o funcionamento das funcionalidades que auxiliam o processo como um todo.

VI.3.1 - ÁRVORE DE CONHECIMENTO DOS PROCESSOS

Conforme apresentado na seção *V.1.4 - Termos Técnicos*, a Árvore de Conhecimento dos Processos foi montada a partir de um conjunto de glossários dispostos em arquivos HTML.

O objetivo inicial era apenas extrair os termos técnicos para serem utilizados nas inferências dos processos de mineração de textos. Entretanto, notou-se que o armazenamento em uma estrutura hierárquica era possível. Desta forma, verificou-se que seria interessante a apresentação gráfica desta árvore, de forma que o usuário pudesse compreender a sua estrutura.

Em meio aos insumos recuperados, foi criada uma hierarquia para organizar os ramos desta árvore. Esta hierarquia possui 9 níveis e está formatada da seguinte forma:

- 1º Nível – Nome do Projeto
- 2º Nível – Tipos de Programa

- 3° Nível – Programas
- 4° Nível – Módulos
- 5° Nível – Especialidades
- 6° Nível – Livros
- 7° Nível – Capítulos
- 8° Nível – Termos
- 9° Nível – Definição e Referências dos Termos

Em termos de visualização, várias opções são possíveis. Na maioria destas visões, são apresentados 3 níveis: nó central, nó intermediário e folhas. As figuras a seguir são referentes à conjuntura desta dissertação, entretanto, qualquer estrutura de tópicos similar poderia ser representada desta forma.

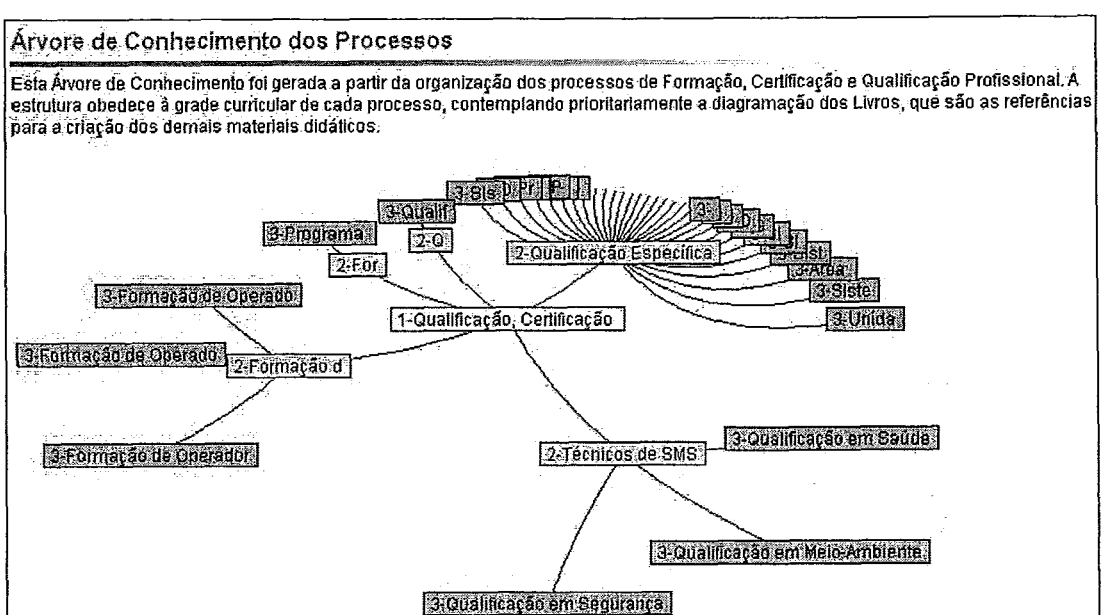


Figura 46 - Visão geral da Árvore de Conhecimento (1°, 2° e 3° níveis) (elaborada pelo autor)

A Figura 46 apresenta a visão inicial da árvore, contendo os 3 primeiros níveis. A ferramenta, neste caso, permite ao usuário um entendimento da dimensão e do fluxo de trabalho em cada processo abordado. Ao selecionar um nó desta árvore, a ferramenta automaticamente o desmembra em conteúdos específicos, referentes aos subprocessos do nó escolhido. É importante notar, que o nó almejado torna-se o centro da árvore para

facilitar a visualização dos demais subprocessos associados, conforme pode ser visto na Figura 47.

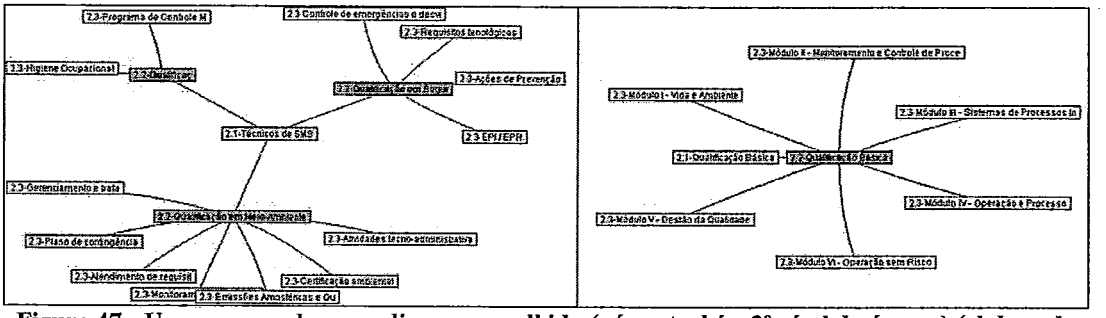


Figura 47 - Um processo de aprendizagem escolhido (nó central é o 2º nível da árvore) (elaborada pelo autor)

Seguindo a mesma linha de raciocínio, é possível continuar navegando na árvore escolhendo novos nós para ocupar o centro. A estrutura desta árvore permite a visualização desde o macro processo principal até a definição dos termos. A Figura 48 apresenta mais um exemplo de navegação em profundidade realizado.

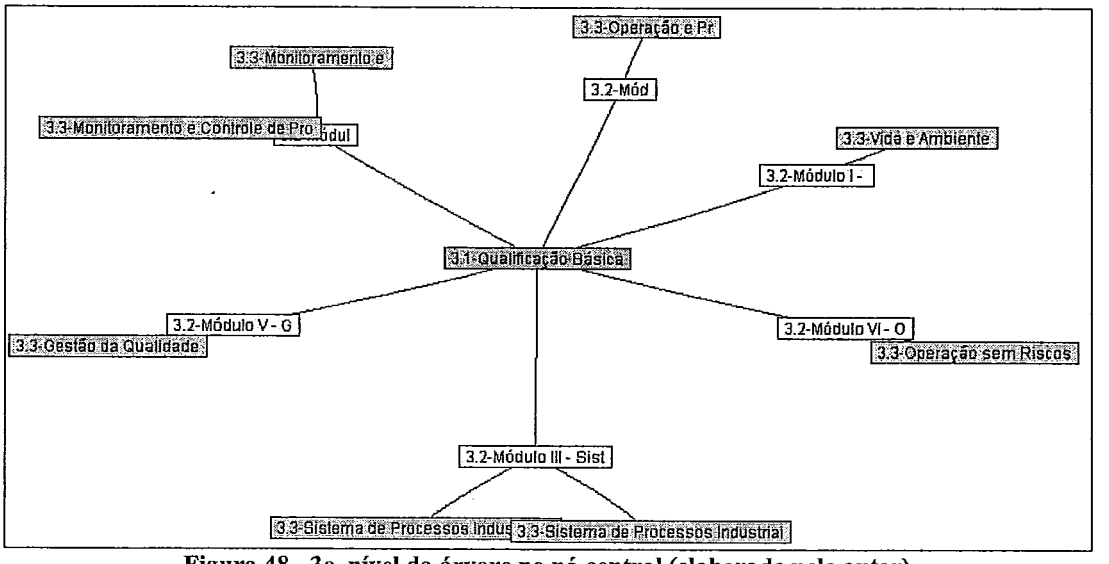


Figura 48 - 3o. nível da árvore no nó central (elaborada pelo autor)

Como pode ser observada, a navegabilidade desta árvore é bastante intuitiva. Como processo interno, a mineração de textos usa esta árvore para correlacionar os insumos do processo de mineração com os documentos extraídos.

É importante ressaltar que esta árvore foi montada de acordo com as especificidades do estudo de caso, entretanto, qualquer outro contexto poderia ser facilmente acoplado ao processo, bastando, naturalmente, a inclusão da árvore adequada.

VI.3.2 - RADICALIZADOR

Como foi abordado anteriormente (seção *III.2.3 - Stemming*), no processo de mineração é sugestivo que a contabilização da freqüência dos termos seja feita pelo radical e não pela palavra original, o que evita que palavras como *texto*, *textos*, *textuais* e *textual* sejam tratadas diferentemente.

O processo de radicalização é utilizado internamente no SMiner, entretanto, uma interface foi desenvolvida para ilustrar o funcionamento deste mecanismo.

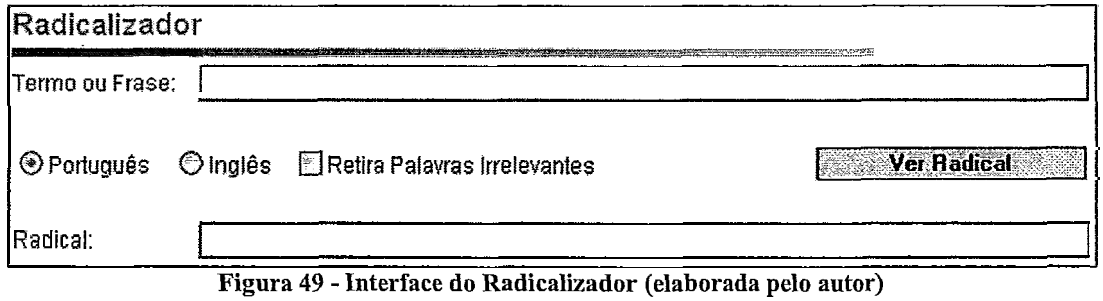


Figura 49 - Interface do Radicalizador (elaborada pelo autor)

A partir da interface ilustrada na Figura 49, é possível escrever uma frase e o sistema retornará as palavras radicalizadas.

VI.3.3 - LIMPA TEXTO

Em geral, os insumos passam por uma "limpeza" no processo de mineração de textos, como a retirada de caracteres especiais, a eliminação de palavras irrelevantes e, quando dispostos em formato HTML, necessitam de um tratamento para remoção das *tags*.

A retirada de acentuação e caracteres especiais é um procedimento simples, mas se não efetuada pode diminuir bastante o espectro das consultas. Além disso, um outro mecanismo que acelera a busca é a retirada de palavras irrelevantes, na medida em que a ausência destas palavras torna a tarefa de indexação mais ágil.

Para validar e aprimorar continuamente estes mecanismos, foi desenvolvida uma interface onde é possível verificar o funcionamento deste processo de "limpeza", como mostra a Figura 50.

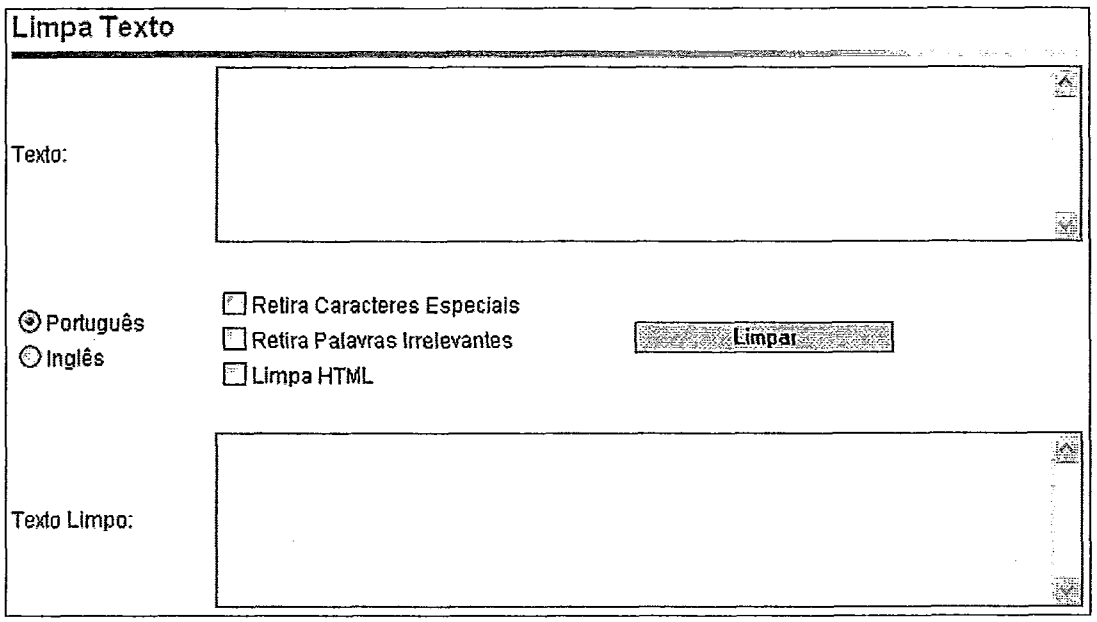


Figura 50 - Interface do Limpa Texto (elaborada pelo autor)

VI.4 - CONSIDERAÇÕES FINAIS

Esta seção apresentou as ferramentas do sistema SMiner. Essas funcionalidades fornecem condições para a realização de consultas qualitativas em processos que envolvam, por exemplo, livros, questões, críticas e sugestões.

O SMiner mostra-se como uma alternativa de busca de informações que otimiza as respostas a consultas em documentos disponíveis em processos de aprendizagem, sobretudo, na certificação profissional por possibilitar desde consultas por palavras chave a correlações entre os livros didáticos e as respectivas questões. Este trabalho só foi possível com a utilização de tecnologias de mineração de textos.

Apesar de a ferramenta apresentar resultados satisfatórios nos testes realizados, faz-se necessário que as avaliações sejam realizadas em um conjunto maior de usuários. Desta forma, um modelo de avaliação foi concebido e os resultados obtidos nesta aferição podem ser vistos na seção a seguir.

VII - AVALIAÇÃO DOS RESULTADOS

Existem medidas para avaliar a qualidade dos resultados de uma busca, e essas medidas representam mais heurísticas do que verdadeiras medidas. Isso porque o problema de avaliar a qualidade dos resultados é tão difícil como o próprio problema de encontrar resultados relevantes. Desta forma, o objetivo desta seção é apresentar uma maneira sucinta de avaliação dos resultados obtidos nas buscas do sistema.

Em muitos casos, a avaliação da qualidade dos resultados se faz manualmente por um especialista do sistema. No caso das buscas, por exemplo, um conjunto de pessoas que conheçam o contexto pode observar os resultados produzidos e avaliar se eles têm sentido ou não.

Segundo Kontogiannis (1997), a avaliação manual tem características boas e ruins:

- Adaptável: Muitas vezes não existe uma única boa maneira de fazer alguma coisa. A avaliação manual pode julgar a qualidade de um resultado, independentemente do que se esperava. Ao contrário, uma medida automática, dificilmente aceitaria resultados bons diferentes do padrão.
- Adaptada às necessidades: Na medida em que especialistas são as pessoas que trabalham com o contexto avaliado, sabem o que é preciso ou não para o trabalho e podem julgar as verdadeiras qualidades de um resultado.
- Imprecisa: Devido à quantidade de insumos, é impossível para uma pessoa entender o contexto inteiro, assim fica difícil também julgar a qualidade exata de um resultado.
- Subjetiva: Duas pessoas, ou uma pessoa em dois momentos, podem ter opiniões muito diferentes sobre um resultado. Isso não facilita as comparações.

Diante da especificidade do tema e a possibilidade de utilização de especialistas para observar os resultados, a avaliação manual foi escolhida. Métricas como *precision recall* e *fallout* também foram estudadas para quantificar a avaliação dos resultados, e serão elucidados nas seções seguintes.

VII.1 - MEDIDAS DE AVALIAÇÃO DOS RESULTADOS

Precision e *Recall* são medidas baseadas na noção de documentos relevantes de acordo com uma determinada necessidade de informação. *Recall* é a proporção de documentos relevantes de uma coleção que foram recuperados e *precision* é a proporção dos documentos recuperados em uma busca que são relevantes. Em geral, *precision* e *recall* são calculados usando uma coleção de consultas, documentos e julgamentos de relevâncias conhecidos (RIJSBERGEN, 1999).

Outras medidas utilizadas são a medida F, a medida E e o *fallout* (RIJSBERGEN, 1999). A decisão de quais medidas utilizar em uma avaliação depende da aplicação e há sempre discussões sobre a confiabilidade de tais medidas (SU, 1998). Um exemplo é o artigo de GWIZDKA & CHIGNELL (1999), aonde se discute como avaliar máquinas de busca. Não é claro, por exemplo, o quanto pequenas diferenças na *precision* e *recall* têm algum efeito no sucesso na busca de um usuário.

Segundo OARD & MARCHIONINI (1996), a medida de *fallout* mede a eficácia de rejeição, sendo calculado pelo número de documentos não-relevantes recuperados divididos pelo número de documentos não-relevantes presentes na coleção.

SPARCK-JONES & WILLET (1997) discutem a necessidade de novas medidas para avaliar o conhecimento relevante fornecido ao usuário, dependendo de seu interesse ou objetivo e de seus conhecimentos prévios. Entretanto, o próprio conceito de relevância é bastante subjetivo e traz à tona discussões profundas. GREISDORF (2000) aborda este tema, indicando que a avaliação da relevância tem a ver com a efetividade da comunicação, ao assunto em questão e depende do que o usuário já sabe.

Neste trabalho, mecanismos baseados em precisão (*precision*), cobertura (*recall*) e desacerto (*fallout*) foram estudados, dado que são medidas clássicas para avaliar a qualidade de métodos de pesquisa "imprecisa", ou seja, a busca é baseada em itens com propriedades distintas. Um exemplo típico de "pesquisa imprecisa" poderia ser procurar por "alguma coisa sobre física" no sistema de busca.

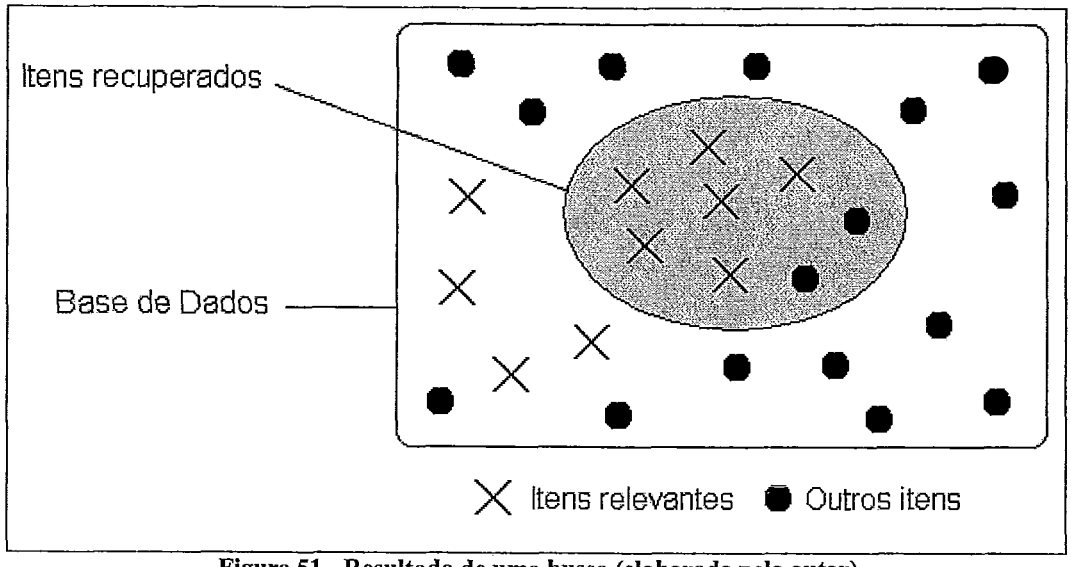


Figura 51 - Resultado de uma busca (elaborada pelo autor)

A Figura 51 apresenta um exemplo de resultado de busca. Para calcular *Precision, Recall* e *Fallout* basta utilizar as seguintes fórmulas:

$$\text{Precision} = \frac{\text{Quantidade de itens relevantes recuperados}}{\text{Quantidade de itens recuperados}}$$

$$\text{Recall} = \frac{\text{Quantidade de itens relevantes recuperados}}{\text{Quantidade de itens relevantes na base de dados}}$$

$$\text{Fallout} = \frac{\text{Quantidade de itens não-relevantes recuperados}}{\text{Quantidade de itens não-relevantes na base de dados}}$$

Neste exemplo e utilizando as fórmulas destacadas é possível chegar aos seguintes resultados:

$$\text{Precision} = \frac{6}{8} = 75\% \qquad \text{Recall} = \frac{6}{8} = 60\% \qquad \text{Fallout} = \frac{2}{16} = 12{,}5\%$$

Foram realizados testes para verificar a usabilidade destas medidas, entretanto, a natureza dos testes realizados não permitiu uma mensuração precisa sobre o conjunto de documentos relevantes ou não da base de dados. Para se obter de forma aceitável estas medidas, experimentos exaustivos com outras ferramentas deveriam ser contemplados. Como a proposta dos testes aborda somente o ponto de vista dos usuários, os únicos dados que foram efetivamente usados, neste caso, correspondem a quantidade de itens recuperados e a quantidade de itens da base de dados, que permitem estivar a precisão (*Precision*).

Desta forma, foi adicionado ao SMiner um mecanismo que possibilita aferir a relevância dos resultados obtidos. A avaliação é feita pelo próprio usuário a partir das interfaces de visualização dos resultados. Além disso, para aumentar o espectro de análises, o SMiner possui um módulo OLAP (*On Line Analytic Processing* ou Processo Analítico Online) que permite análises qualitativas sob diversas perspectivas. Estas ferramentas serão elucidadas na seção a seguir.

VII.2 - AVALIAÇÃO DAS BUSCAS

Conforme foi abordado, o usuário da ferramenta poderá avaliar as buscas a partir da própria interface de visualização dos resultados (Figura 52). A disponibilização desta avaliação é indicada na seção *VI.2.5 -Configurações do Sistema.*

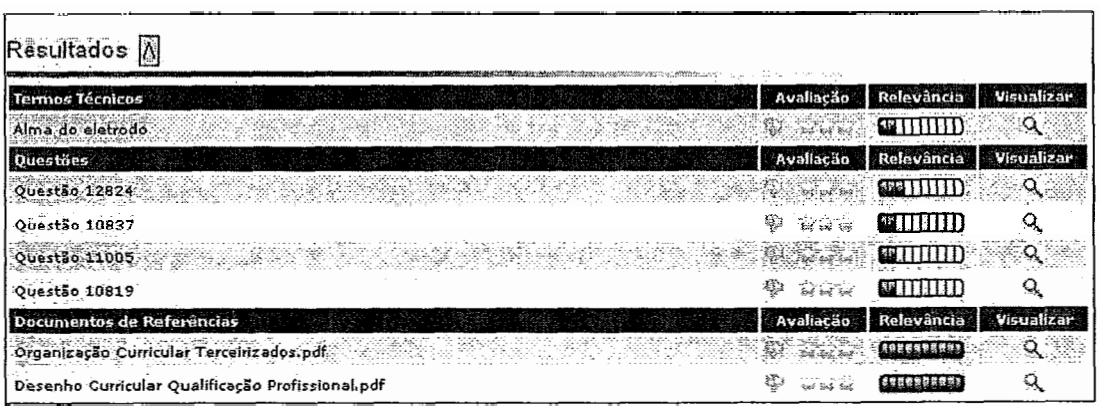


Figura 52 - Avaliação dos Resultados (elaborada pelo autor)

O modelo de avaliação proposto na diagramação desta interface é baseado em GWIZDKA & CHIGNELL (1999) e permite que o usuário indique os níveis de relevância (satisfação, segundo o ponto de vista do usuário) entre 0 e 3. Este esquema de pontuação permite indicar inclusive se o item encontrado é irrelevante.

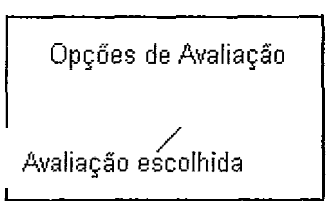


Figura 53 - Possibilidades de Avaliação dos Resultados (elaborada pelo autor)

Em cada aferição são armazenados na base de dados: os termos de consulta, o item avaliado e a avaliação proferida. Os tipos de avaliação estão dispostos na Figura 54 e são armazenados segundo o seguinte critério:

- 0: Irrelevante – significa que o item não retornado não atende as expectativas do usuário.
- 1: Pouco Relevante – embora seja um retorno relevante, é pouco significativo para o usuário.
- 2: Relevante – em geral, o usuário está satisfeito com o resultado obtido.
- 3: Muito Relevante – neste caso, o retorno atende exatamente as expectativas do usuário.

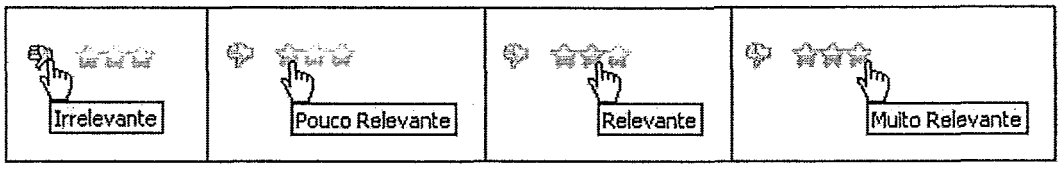


Figura 54 - Tipos de Avaliação (elaborada pelo autor)

Futuramente, com um maior volume de medições, será possível calibrar a ferramenta indicando graus de relevância mais precisos, na medida em que o Sistema levará em consideração as avaliações dos resultados anteriores.

Estas informações contribuem para a geração das estatísticas dos processos. Os testes realizados contaram com a participação de usuários com pouco, bom e muito conhecimento dos processos. O exame destes testes pode ser visto na seção a seguir.

VII.2.1 - TESTES REALIZADOS

Segundo TRAVASSOS *et al* (2002), durante a avaliação, uma maior atenção deve ser prestada aos participantes, ou seja, à seleção da população. Os problemas ou riscos relacionados à validade da avaliação dependem fundamentalmente dos participantes, dado que estes podem ficar cansados ou desanimados ou ainda aprender ao longo do estudo, por exemplo.

O autor afirma ainda que os grupos de participantes podem produzir resultados diferentes por causa do comportamento ou habilidades diferentes. Além disso, o ser humano sempre está tentando parecer melhor quando está sendo avaliado, de forma que os pesquisadores podem afetar os resultados (viés) projetando o estudo baseado naquilo que eles esperam do experimento (TRAVASSOS *et al*, 2002).

Foram realizadas aferições com 32 pessoas, entre Especialistas, Operadores e Usuários Comuns. O grupo dos Especialistas é composto por Instrutores e/ou

Elaboradores de materiais didáticos da especialidade (competência ou disciplina) avaliada. Os Operadores são os estudantes de Formação e Certificação Profissional e os Usuários Comuns são, em geral, estudantes de informática que avaliaram a ferramenta sob um foco mais técnico.

Naturalmente, a opção de avaliar ou não um resultado é exclusiva do usuário da ferramenta. Foi solicitado aos participantes que avaliassem as buscas com o propósito de melhorar o desempenho da ferramenta, entretanto, não houve obrigatoriedade de avaliação de todos os resultados. Além disso, optou-se pelo caráter anônimo dos participantes para que o grau de criticidade fosse o mais realista possível.

Uma vez definidos os grupos de participantes e os insumos modelados para o experimento, alguns objetivos foram estabelecidos:

- Analisar o nível de aceitação das ferramentas de busca;
- Verificar possíveis discrepâncias entre a percepção de cada grupo de usuários;
- Verificar a viabilidade de utilização da ferramenta de correlação entre questões e livros;
- Mensurar precisão (*precision*) sob o ponto de vista de cada grupo de usuários;
- Mensurar a média de itens relevantes e/ou irrelevantes sob o ponto de vista de cada grupo de usuários.
- Recuperar, a partir das análises, melhorias de desenvolvimento futuro.

Diante destes objetivos, um conjunto de insumos foi modelado para nortear o estudo dos resultados obtidos. Os itens disponibilizados para as avaliações foram:

- 25 Livros, totalizando 139.695 palavras, sendo 10.313 distintas, compostas por 4.995 radicais.
- 4 Documentos de Referência, totalizando 38.143 palavras, sendo 3.911 distintas, compostas por 2.460 radicais.
- 10 Materiais de Apoio Complementar, totalizando 14.012 palavras, sendo 2.349 distintas, compostas por 1.535 radicais.
- 9671 Questões de Prova, totalizando 308.617 palavras, sendo 15.115 distintas, compostas por 7.940 radicais.

- 5258 Termos Técnicos, totalizando 169.548 palavras, sendo 12.903 distintas, compostas por 6.581 radicais.
- 1309 Críticas e Sugestões, totalizando 29.563 palavras, sendo 1.888 distintas, compostas por 1.163 radicais.

A partir destes itens, foram armazenadas as avaliações e condensados os resultados, que serão vistos na seção a seguir.

VII.2.2 - VISUALIZAÇÃO DOS RESULTADOS

Para uma melhor análise dos resultados, foi desenvolvida uma ferramenta OLAP (*On Line Analytic Processing* ou Processo Analítico Online), que permite a visualização dos dados em diversas dimensões. A Figura 55 ilustra esta implementação.

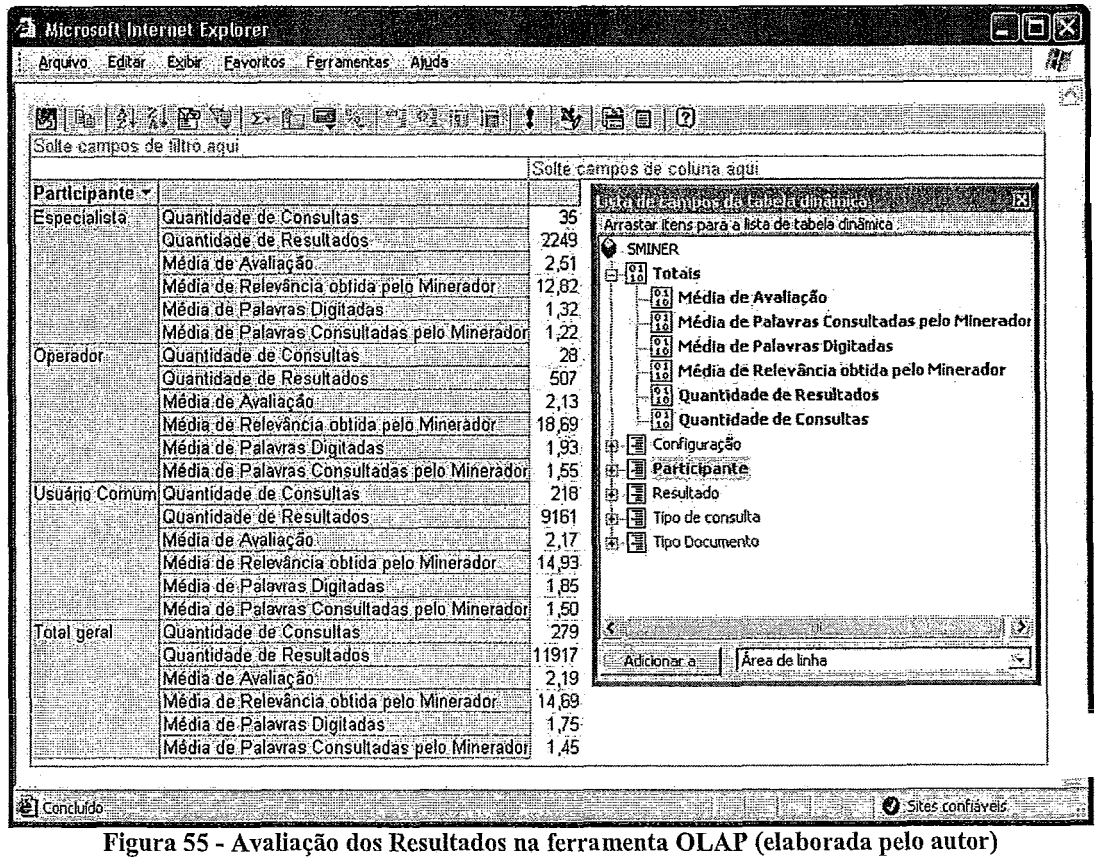


Figura 55 - Avaliação dos Resultados na ferramenta OLAP (elaborada pelo autor)

As ferramentas OLAP são comumente utilizadas em Data Warehouse – ambiente de suporte à decisão que utiliza dados de diferentes fontes e os organiza para os tomadores de decisão, independente de seu nível de qualificação técnica, ou seja, é uma tecnologia de gestão e análise de dados (SINGH, 2001).

Ainda neste contexto estão os Data Marts, que são implementações específicas de um departamento ou assunto em particular. Segundo Cazarini (2002), a principal diferença entre Data Mart e Data Warehouse é que os Data Marts são voltados somente para uma determinada área, enquanto o Data Warehouse é voltado para assuntos de toda empresa.

Desta forma, para esta dissertação foi elaborado um Data Mart cuja modelagem proporciona a análise dos resultados através da ferramenta OLAP disponibilizada no SMiner. O modelo do Data Mart pode ser visto no Anexo III.

Em linhas gerais, a Tabela 13 apresenta algumas informações quantitativas dos resultados obtidos:

Características		Quantidade	Total
Participantes	Especialistas	4	32
	Operadores	10	
	Usuários Comuns	18	
Documentos	Livros	25	11.917
	Documentos de Referência	4	
	Materiais de Apoio Complementar	10	
	Questões de Prova	9.671	
	Termos Técnicos	5.258	
	Críticas e Sugestões	1.309	
Palavras tratadas distintas	Irrelevantes (*Stop Words*)	479	27.149
	Relevantes	26.670	
	Radicais	13.418	13.418
Avaliação dos Resultados	Muito Relevantes	410	11.917
	Relevantes	170	
	Pouco Relevantes	119	
	Irrelevantes	71	
	Não avaliados	11.147	

Tabela 13 - Resultados Obtidos na Avaliação das Ferramentas de Busca (elaborada pelo autor)

Com o cruzamento dos dados a partir da ferramenta OLAP, foi possível encontrar algumas particularidades. Os itens que chamaram mais atenção podem ser vistos nas seções a seguir.

VII.2.2.1 - Participação dos Avaliadores

Os testes totalizaram 279 consultas, obtendo 11.917 resultados e, conforme pode ser visto na Tabela 13, foram avaliados 770 resultados. Estes números mostram uma média de 43 resultados por consulta realizada, sendo que cada participante realizou cerca de 9 consultas e avaliou 24 resultados em média.

Participante	Quantidade de Consultas	Quantidade de Resultados	Quantidade de Pessoas
Especialista	35	2249	4
Operador	28	507	10
Usuário Comum	218	9161	18
Total geral	279	11917	32

Figura 56 - Participação dos Avaliadores (elaborada pelo autor)

A Figura 56 mostra que os Usuários Comuns foram os que mais utilizaram as ferramentas, obtendo uma média de 12 consultas. Os especialistas realizaram 9 consultas em média, enquanto os operadores realizaram 3 consultas em média.

Em relação à quantidade de resultados obtidos, os Especialistas obtiveram, em média, 64 resultados por consulta, enquanto os Operadores obtiveram 18 e os Usuários Comuns 42. Isto sinaliza que os Especialistas consultaram por assuntos mais amplos, ao passo que os Operadores foram mais específicos em suas buscas.

VII.2.2.2 - Resultados por Participante

A Figura 57 mostra um comparativo entre as Avaliações dos Resultados por Participante. Nesta figura são apresentados apenas os resultados avaliados.

		Avaliação dos Resultados				
		Muito Relevante	Relevante	Pouco Relevante	Irrelevante	Total geral
Participante						
Especialista	Quantidade de Resultados	69,23%	16,67%	10,26%	3,85%	100,00%
Operador	Quantidade de Resultados	56,13%	16,13%	12,26%	15,48%	100,00%
Usuário Comum	Quantidade de Resultados	50,09%	24,58%	17,13%	8,19%	100,00%
Total geral	Quantidade de Resultados	53,25%	22,08%	15,45%	9,22%	100,00%

Figura 57 - Porcentagem de Resultados por Participante (elaborada pelo autor)

Como pode ser visto na Figura 57, as melhores avaliações foram dadas pelos Especialistas, onde cerca de 70% apontaram como Muito Relevante os itens recuperados pela ferramenta. Por outro lado, os Operadores foram os mais críticos, indicando 15% dos resultados eram Irrelevantes.

Em geral, as avaliações foram bastante satisfatórias apontando que 75% dos entrevistados avaliaram os resultados como Muito Relevante ou Relevante, enquanto 15% apontaram como Pouco Relevante e apenas 9% como Irrelevante.

VII.2.2.3 - Avaliação x Relevância

Pode-se notar ainda a correlação entre o nível da avaliação dada pelo usuário e a relevância normalizada pelo minerador.

Avaliação dos Resultados	Média de Relevância obtida pelo Minerador
Muito Relevante	36,49
Relevante	24,35
Pouco Relevante	17,34
Irrelevante	23,96
Não Avaliado	13,65
Total geral	14,69

Figura 58 - Avaliação dos usuários x Relevância média obtida pelo Minerador (elaborada pelo autor)

A partir da Figura 58, verifica-se que a relevância média obtida pelo minerador para os itens considerados Muito Relevantes foi de 36,49 onde 100 é o valor máximo na normalização. Para os resultados classificados pelos usuários como Relevante a média obtida pelo minerador foi de 24,35 em 100, enquanto as classificadas como Pouco Relevante obteve um valor de 17,34 em 100.

Estes dados mostram uma convergência, indicando que quanto maior é a relevância normalizada, maiores são as avaliações dos usuários. Entretanto, os itens Irrelevantes fogem a esta tendência na medida em que a média obtida foi de 23,96 em 100. Este dado foi principalmente afetado por consultas que, embora tenham relação com as palavras da consulta, não atendem as expectativas do usuário, como por exemplo, os desenhos curriculares e a contracapa dos livros que contêm a descrição de todas as especialidades e grade curricular, mas não são conteúdos didáticos.

Uma informação interessante é a média dos itens não avaliados. Esta média é a menor (13,65) dentre as analisadas, o que indica que os usuários avaliam prioritariamente os materiais com maiores relevâncias.

VII.2.2.4 - Tratamento das Palavras Consultadas

A língua portuguesa possui muitas variedades lingüísticas. A própria forma de conjugação de verbos, os vários tipos de sufixos e a vasta acentuação de palavras são exemplos que dificultam o tratamento de uma consulta em um conjunto de documentos.

Para demonstrar este tipo de dificuldade e como o Sistema tratou estas particularidades, foi adicionada ao conjunto de estatísticas a comparação entre quantidade de palavras digitadas pelo usuário e a quantidade de palavras efetivamente aproveitadas pelo minerador, como pode se visto na Figura 59.

Avaliação dos Resultados	Média de Palavras Digitadas	Média de Palavras Consultadas pelo Minerador
Muito Relevante	3,24	2,28
Relevante	2,89	2,22
Pouco Relevante	3,34	2,54
Irrelevante	2,90	2,21
Não Avaliado	1,65	1,39
Total geral	1,75	1,45

Figura 59 - Comparação entre os Resultados e o tratamento das palavras consultadas (elaborada pelo autor)

O resultado desta análise foi bastante satisfatório na medida em que 60% das consultas sofreram algum tipo de tratamento e, em média, foram digitadas 1,75 palavras e efetivamente consultadas 1,45 palavras. Estes números indicam que 18% das palavras foram descartadas para uma melhor aferição da consulta.

Uma análise interessante é observar somente os itens avaliados, ou seja, retirando das estatísticas os itens não avaliados. A Figura 60 mostra que a média de palavras digitadas sobe para 3,15 enquanto a média de palavras consultadas pelo minerador sobe para 2,30 e, consequentemente, o número de palavras descartadas sobe para 27%.

Estes dados indicam que os participantes optam por avaliar as consultas mais detalhadas, com um grau maior de especificidade; ou ainda, uma tentativa de calibrar ou se ambientar à ferramenta.

Avaliação dos Resultados	Média de Palavras Digitadas	Média de Palavras Consultadas pelo Minerador
Muito Relevante	3,24	2,28
Relevante	2,89	2,22
Pouco Relevante	3,34	2,54
Irrelevante	2,90	2,21
Total geral	3,15	2,30

Figura 60 - Comparação entre os itens avaliados e o tratamento das palavras (elaborada pelo autor)

Ainda sob este contexto, pode ser notado na Figura 61 que as consultas que retiraram palavras irrelevantes (*stop words*) e utilizaram radicalização de palavras (texto limpo) foram as mais utilizadas, correspondendo a 57% do total.

Configuração	Quantidade de Resultados	Média de Avaliação
Sem configuração	43,10%	2,28
Texto Limpo	44,55%	2,08
Limpo e usa Posicionamento	4,76%	2,40
Limpo, usa Thesaurus e Posicionamento	7,59%	2,08
Total geral	100,00%	2,19

Figura 61 - Avaliação dos Especialistas em função das Configurações utilizadas (elaborada pelo autor)

A Figura 61 indica que as melhores avaliações coincidem com a utilização de Posicionamento, indicando que o algoritmo que verifica a proximidade das palavras (utilizado na Busca Contextual) foi bastante eficaz.

Por outro lado, não houve grandes mudanças de avaliação quanto ao uso de *thesaurus*. Apenas 8% das consultas utilizaram este mecanismo e não obtiveram, nestas avaliações, grandes melhorias em relação às demais configurações.

Ao que tudo indica, esta análise é reflexo da pequena base de dados que foi acoplada ao Sistema. Diante disto, pode-se afirmar que, embora seja de suma importância a infra-estrutura criada para utilização de *thesaurus*, ainda é incipiente sua aplicabilidade em função da quantidade de sinônimos cadastrados.

Entretanto, como era esperado, verificou-se que as consultas que utilizam este mecanismo retornaram mais resultados. Enquanto a média de retorno por consulta foi de 43 resultados, as consultas que utilizaram *thesaurus* tiveram, em média, um retorno de 100 resultados.

VII.2.2.5 - Tipos de Consulta

O SMiner permite 4 tipos de consulta: Simples (ou Busca por Palavra Chave), Avançada, Contextual e Lacunas - Livros x Questões. A análise sobre o a avaliação dos resultados por tipo de consulta foi bastante linear, como mostra a Figura 62.

		Consulta				
		Por Palavra Chave	Avançada	Contextual	Questões x Livros	Total geral
Participante						
Especialista	Média de Avaliação	2,34	2,56	2,84	2,44	2,51
Operador	Média de Avaliação	2,19	2,13	1,97	2,29	2,13
Usuário Comum	Média de Avaliação	2,31	2,14	2,13	2,04	2,17
Total geral	Média de Avaliação	2,28	2,15	2,16	2,17	2,19

Figura 62 - Tipos de Consulta (elaborada pelo autor)

A Figura 63 mostra a predileção dos Especialistas em verificar discrepâncias entre livros x questões em relação aos demais participantes. Isto é justificado pelo fato dos especialistas participarem da elaboração destes materiais.

		Consulta				
		Por Palavra Chave	Avançada	Contextual	Questões x Livros	Total geral
Participante						
Especialista	Quantidade de Consultas	34,29%	11,43%	28,57%	25,71%	100,00%
Operador	Quantidade de Consultas	39,29%	14,29%	32,14%	14,29%	100,00%
Usuário Comum	Quantidade de Consultas	24,77%	29,82%	27,06%	18,35%	100,00%
Total geral	Quantidade de Consultas	27,24%	25,81%	27,96%	19,00%	100,00%

Figura 63 - Consultas por Participante (elaborada pelo autor)

A Figura 63 mostra ainda uma diferença entre a preferência dos Usuários Comuns ao uso da busca Avançada em relação aos Especialistas e Operadores.

VII.2.2.6 - Questões x Livros

Encontrar lacunas entre os materiais didáticos, sobretudo, entre os livros e as questões de prova é, sem dúvida, o maior desafio deste trabalho. O que se propõe, neste caso, é encontrar a probabilidade de um documento conter conteúdo suficiente para responder as questões consultadas.

Em geral, as avaliações foram boas, indicando que 83% dos Especialistas avaliaram em Muito Relevante ou Relevante os resultados das correlações, como mostra a Figura 64.

Tipo de consulta: Questões x Livros						
		Avaliação dos Resultados				
		Muito Relevante	Relevante	Pouco Relevante	Irrelevante	Total geral
Participante						
Especialista	Quantidade de Resultados	61,11%	22,22%	16,67%		100,00%
Operador	Quantidade de Resultados	64,29%	14,29%	7,14%	14,29%	100,00%
Usuário Comum	Quantidade de Resultados	49,02%	15,69%	25,49%	9,80%	100,00%
Total geral	Quantidade de Resultados	54,22%	16,87%	20,48%	8,43%	100,00%

Figura 64 - Avaliação do Módulo de Questões x Livros (elaborada pelo autor)

VII.2.3 - Considerações Finais

Como pode ser visto, a ferramenta OLAP possibilita a interseção de diversos dados nas diversas dimensões fornecidas, como por exemplo: participantes, resultados recuperados, tipos de consulta, tipos de documento e tipos de configuração.

É importante observar que foram destacadas as principais análises sob o ponto de vista do autor desta dissertação. Entretanto, a ferramenta possibilita um conjunto imenso de combinações, dependendo da linha de interesse de cada observador.

Uma análise interessante foi notar que as melhores qualificações foram dadas no final dos testes, quando o usuário já estaria "acostumado" com a ferramenta, o que

sugere uma "calibragem" do sistema para cada usuário. É bastante provável que em testes mais exaustivos haverá uma curva de aceitação ainda maior entre os usuários do minerador.

CONCLUSÕES E TRABALHOS FUTUROS

Como pode ser visto no decorrer deste trabalho, há uma dificuldade latente das instituições de ensino em avaliar a relevância dos seus materiais didáticos, sobretudo quando há criação de sistemas de diagnóstico a partir de materiais previamente selecionados.

Esta dificuldade mostra a necessidade de processos cada vez mais automatizados para auxiliar na descoberta de padrões desconhecidos de dados reais. A habilidade de manusear grandes quantidades de textos é um atenuante em ambientes que envolvam muitos livros, questões de prova e documentos diversos, por exemplo.

Neste trabalho, foram descritas algumas técnicas de mineração de textos e sua aplicabilidade, apresentando ferramentas que suportam esta tecnologia e propondo uma ferramenta de auxilio a processos de certificação profissional.

Como foi abordado, os processos de certificação envolvem muitos materiais didáticos e há um constante problema em encontrar informações e, consequentemente, disseminar conhecimento devido a este volume de dados.

Através do estudo de caso, foram analisadas situações comuns e adversas, envolvendo um considerável conjunto de pessoas e materiais. Neste estudo, os avaliadores foram divididos em grupos (Especialistas, Operadores e Usuários Comuns) para tornar a avaliação mais qualitativa e as inferências na ferramenta proposta permitiam uma grande combinação de configurações nas consultas realizadas.

Diante desta extensa quantidade de atributos de avaliação, as possibilidades de correlação entre as perspectivas de análise apontavam um imenso grupo de combinações. A partir deste preceito, foi utilizada uma ferramenta OLAP para uma completa análise da avaliação dos resultados obtidos.

Com a modelagem do problema e a melhor visualização das avaliações na ferramenta OLAP e muitos pontos puderam ser observados.

Trabalhos Futuros

Este trabalho abre portas para outras pesquisas na medida em que a utilização de tecnologias de mineração de textos aplicados a processos de certificação ainda são pouco explorados. Por exemplo, uma maior ênfase na medição de desempenho associado à expansão de itens da ontologia poderia ser realizada, ênfase esta que não pôde ser dada como completa devido à pequena quantidade informações que foram disponibilizadas para a geração da ontologia neste trabalho. Além disso, a avaliação do minerador poderia ser incrementada a partir da comparação dos resultados obtidos no SMiner com outras ferramentas disponíveis no mercado, respeitando naturalmente a necessidade de autorização da EmpresaXPTO.

A expectativa do autor desta dissertação é que os trabalhos de continuidade deste projeto priorizem sempre a usabilidade do Sistema. A partir dos resultados obtidos ao longo deste processo, alguns itens podem ser destacados como principais trabalhos futuros:

- Criar mecanismos de realimentação de informações por perfil, onde a relevância dos resultados obtidos use heurísticas baseadas nas avaliações aferidas e em *thesaurus* associados a cada usuário distinto;
- Aumentar da base do *thesaurus*, o que tornará a busca por insumos mais acurada na medida em que aumentará o espectro da consulta sobre um mesmo conjunto de dados;
- Desenvolver algoritmos para manipular a semântica das palavras, possibilitando uma aferição melhor nos resultados das buscas.
- Desenvolver mecanismos que garantam uma maior portabilidade da ferramenta, independente de plataforma de utilização ou banco de dados;
- Incrementar a base de insumos, abordando principalmente assuntos *extra* certificação, de forma que sistema envolva novas áreas de aprendizagem;
- Criar um *website*, de domínio público, de forma a compartilhar e propiciar a colaboração de outros indivíduos neste trabalho.

Entende-se ainda como trabalho futuro o aumento constante de toda base de dados, as naturais melhorias nos algoritmos, bem como o acompanhamento das

tendências de interfaces. Desta forma, há uma grande probabilidade de que este *software* seja sempre algo atrativo e intuitivo ao usuário final.

Notas do Autor

Não quero dizer que este trabalho está concluído, tampouco, pensar na possibilidade de descontinuidade. Sendo assim, posso afirmar que muitos foram os desafios no desenvolvimento desta ferramenta até *este momento*. Certamente a Figura 65 é o que melhor expressa a relação entre o que apresento e o real trabalho efetuado.

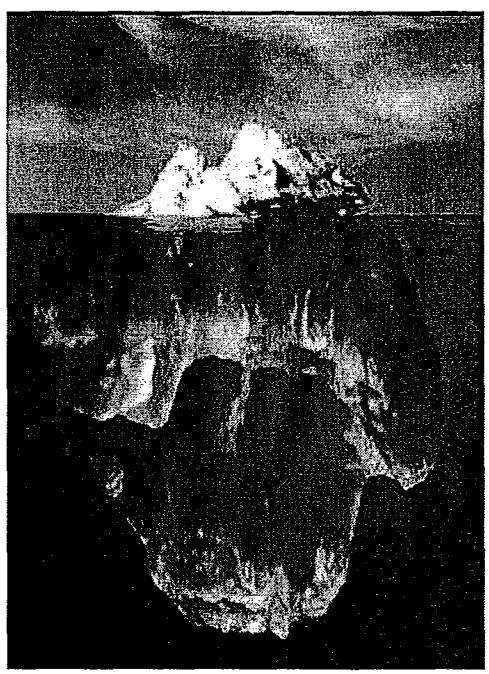


Figura 65 - O dilema do iceberg (elaborada pelo autor)

Os trabalhos iniciais de preparação de insumos geraram diversos re-trabalhos em função do contexto específico, principalmente a manipulação de informações que não constavam em banco de dados, como os livros, as apostilas, os exercícios, as apresentações e os termos técnicos. Programas computacionais específicos foram criados para tratar cada item, como por exemplo, os limpadores de *tags* HTML, os classificadores dos termos técnicos e os mecanismos responsáveis por migrar os documentos de opinião (críticas e sugestões) e questões de prova.

A geração do pequeno *thesaurus* que, embora não tenha atendido a todas as expectativas, mostrou-se um desenvolvimento dispendioso e meticuloso, sobretudo pela especificidade das competências de cada certificação analisada. Um atenuante foi o fato

de não haver, a todo o momento, um especialista de cada competência atestando se a árvore fora perfeitamente carregada ou não.

O desenvolvimento do minerador ainda contemplou mecanismos de radicalização de palavras, algoritmos de retirada de palavras irrelevantes e mecanismos para tratamento de textos em HTML. Nas buscas, algoritmos baseados em modelos vetoriais e de contexto foram utilizados e, posteriormente, tiveram seus resultados normalizados para uma melhor visualização do usuário final.

A opção por avaliar a ferramenta com 3 perfis (Especialista, Operador e Usuário Comum) gerou resultados enriquecedores, mas necessitou de uma logística de agendamentos muitas vezes inusitada.

As avaliações dos resultados das buscas foram bastante satisfatórias, mas as restrições de acesso a determinados dados impostas pela EmpresaXPTO limitaram algumas análises.

Os módulos de consultas, especialmente, o módulo de lacunas tiveram que ser amplamente testados e, a cada pergunta imposta, verificava-se a necessidade de mudanças estruturais ou inclusão de *stop words* em função da linguagem coloquial muitas vezes utilizada.

Quando tudo parecia perfeito, veio a constatação de que simples *selects* no banco de dados não iriam retornar mais do que uma análise quantitativa. Somente uma visualização multidimensional poderia aumentar a qualidade do tratamento das avaliações realizadas. Naquele momento, optou-se pela modelagem e construção do módulo OLAP. Conseqüência imediata: mais trabalho, porém excelentes formas de combinação entre as diversas variáveis de avaliação.

Enfim, muitas foram as dificuldades convergidas em desafios no momento em que apareceram. Não há precedentes na minha vida que possam explicar o crescimento pessoal e profissional que obtive no decorrer deste trabalho

A todos que participaram de forma direta ou indireta desta obra, MUITO OBRIGADO!!!

REFERÊNCIAS BIBLIOGRÁFICAS

ANTONELLI, C.; QUÉRÉ, M. (2004). *The governance of the generation and dissemination of localized technological knowledge.* Itália: Università di Torino and Fondazione Rosselli, 2004. Disponível em <http://www.fondazionerosselli.it/The governance of the generation and dissemination of localized technological knowledge.doc>. Acessado em 12/06/2005.

APTÉ, C.; DAMERAU, F.; WEISS, S. M. (1994). "Automated learning of decision rules for text categorization". *ACM Transactions on Information Systems*, New York, v.12, n.3, p.233-251, 1994.

BAEZA-YATES, R.; RIBEIRO B. N. (1999). *Modern Information Retrieval.* England: Pearson Education Limited, 1999.

BATES, M. J. (1986). "Subject access in online catalogs: a design model". *Journal of the American Society for Information Science*, New York, v.37, n.6, p.357- 376, 1986.

BELKIN, N.J.; CROFT, B.W. (1992). "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Communications of the ACM*, 1992.

BILLHARDT, H. et al (2002). "A context vector model for information retrieval". *Journal of the American Society for Information Science and Technology*, Hoboken, v. 53, Issue 3, p. 236, 2002.

BOYATZIS, R. E (1982). The competent manager: a model for effective performance. New York, John Wiley, 1982.

BRASIL (1996). "Lei n. 9.394, de 20 de dezembro de 1996". *Leis, Decretos.* Documenta, Brasília, n. 423, p. 569-586, dez. 1996. Publicado no D.O.U de 23.12.96. Seção I, p. 1-27.841. Estabelece as Diretrizes e Bases de Educação Nacional. Art. 41.

BRASIL (1997). "Parecer 17, aprovado em 03 de dezembro de 1997". *Conselho Nacional de Educação - Câmara de Educação Básica.* Documenta, Brasília, n. 435, p.

29-38, dez. 1997. Institui as Diretrizes Curriculares Operacionais para a Educação Profissional em nível nacional.

BRASIL (1999). "Resolução CEB 4/99, aprovado em 08 de dezembro de 1999". *Conselho Nacional de Educação - Câmara de Educação Básica.* Documenta, Brasília, n. 459, p. 277-306, dez. 1999. Institui as Diretrizes Curriculares Nacionais para a Educação Profissional de nível técnico. Art. 16.

CAID, W. R.; CARLETON, J. L. (1994). "Context Vector-Based Retrieval". *In: 4th IEEE DUALUSE CONFERENCE*, 1994, Utica, NY. Proceedings... Utica, NY, May 1994.

CAZARINI, A. (2002). *Auxílio do Data Warehouse e suas ferramentas à estratégia do CRM Analítico.* São Carlos: Dissertação de Mestrado, Escola de Engenharia de São Carlos, Universidade de São Paulo, 2002.

CERRITO, P. (2004). *Louisville Hospitals Advance with SAS Text Miner.* DataWarehouse.com. Disponível em <http://www.datawarehouse.com/article/?articleId=3239>. Publicado em 13/02/2004. Visualizado em 30/05/2006.

CHAKRABARTI, S. (1993). *Data mining for hypertext: a tutorial survey.* SIGKDD Explorations, New York, v.1, n.2, p.1-11, Jan. 2000.

CHEN, H. at al (1996). "A concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System". Disponível em <http://ai.bpa.arizona.edu/papers>. Acesso em 05/07/2004.

CHIAVENATO, I. (2002). *Gestão de Pessoas.* São Paulo, Editora Campus, 2002.

COVEY, S. R. (2005). *Os 7 hábitos das pessoas altamente eficazes.* Rio de Janeiro, Editora Best Seller, 2005.

CROFT. W.B.; DAS, R. (1990). "Experiments with query acquisition and use in document retrieval systems". *In: Proceedings of the ACM SIGIR*, Conference on

Research and Development in Information Retrieval, 349-368, (1990).

CUNHA, C.; CINTRA, L.F.L. (2001). *Nova gramática do Português contemporâneo.* 3a. Edição. Rio de Janeiro, Editora Nova Fronteira, 2001.

DAFT, R L (2002). *Organizações: teorias e projetos*. Tradução de Cid. Knipel Moreira. São Paulo: Pioneira, 2002.

DAVENPORT, T.; PRUSAK, L. (1998). *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, 1998.

DRUCKER, P. F. (1996). *Administrando para o Futuro: ao anos 90 e a virada do século*. 6ª Edição, São Paulo, Editora Pioneira, 1996.

EMPRESAXPTO (2003a), Desenho Curricular - Programa de Formação de Novos Operadores. EmpresaXPTO, 2003.

EMPRESAXPTO (2003b), Desenho Curricular - Programas de Qualificação de Operadores, versão 4, EmpresaXPTO, 2003.

EMPRESAXPTO (2003c), Processo de Formação, Qualificação e Certificação da Força de Trabalho, EmpresaXPTO, 2003.

ESTADO DE SÃO PAULO (2006). "Ontoweb: A nova era das ferramentas de busca". *Consultor Jurídico*. São Paulo, 2006. Disponível em <http://conjur.estadao.com.br/static/text/41493>. Acesso em 20/03/2006.

FAYYAD, U. et al. (1996). *Advances in Knowledge Discovery and Data Mining*. 1a. Edição, Cambridge – Massachusetts, Mit Press, 1996.

FELDMAN, R.; DAGAN, I. (1995). "Knowledge discovery in textual databases (KDT)". *INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY*, Montreal, 1995.

FICO, C.; ALONSO. P. (2006). "Certificação Profissional". *Monitor Mercantil* -

Opinião. Disponível em <http://www.monitormercantil.com.br>. Acesso em 12/01/2006.

FLEURY, A.; FLEURY, M. T. L. (2000). *Estratégias Empresariais e Formação de Competências*. São Paulo, Editora Atlas, 2000.

FRAKES, W. B. (1992). "Introducion to information storage and retrieval systems". In: FRAKES, William B.; BAEZA-Yates, Ricardo *A. Information Retrieval: Data Structures & Algorithms*. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.1-12.

FRANCISCHINI, P. (2005). "Vantagens da Certificação Profissional". *Revista TI-Master*. Disponível em <http://www.timaster.com.br/revista/artigos/main_artigo.asp?codigo=1056>. Acesso em 05/03/2006.

GONÇALVES, M. H. B.; BOTINI, J.; PINHEIRO, B. A. et al (2004). *Referências para Educação Profissional. Rio de Janeiro, 2004.* In: IBGE. 12 milhões atrás de emprego. Instituto Brasileiro de Geografia e Estatísticas. Disponível em: http://www.ibge.org.br. Acesso em 18/10/2004.

GOOGLE (2006). *Características do Google*. Disponível em <http://www.google.com.br/intl/pt_BR//features.html>. Acessado em 07/10/2006.

GRAMIGNA, M. R. (2002). *Modelo de Competências e Gestão de Talentos*. São Paulo, Editora Makron Books, 2002.

GREISDORF, H. (2000). *Relevance: An Interdisciplinary and Information Science Perspective. Special Issue on Information Science Research*. Volume 3, No 2, 2000. Disponível em <http://www.inform.nu/Articles/Vol3/IndexV3.htm>. Acessado em 28/12/2006.

GWIZDKA, J.; CHIGNELL, M. (1999). *Towards Information Retrieval Measures for Evaluation of Web Search Engines*. 1999. Disponível em http://anarch.ie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf. Acessado em

10/11/2006.

IDC (2004). "Pesquisa sobre Certificação Profissional", *Institute Data Corporation Brasil*. 2004.

JACOBSEN, P. (1992). "O Intrapreneur e as Oportunidades". In ABRH Rio. *Recursos Humanos: foco na modernidade: textos selecionados*. Rio de Janeiro: Qualitymark Ed, 1992.

KONTOGIANNIS, K. (1997). "Evaluation experiments on the detection of programming patterns using *software* metrics". *In: Working Conference on Reverse Engineering*, pages 44-54. IEEE, IEEE Comp. Soc. Press, Oct. 1997.

KOWALSKI, G. (1997). *Information Retrieval Systems: Theory and Implementation*. Boston, Kluwer Academic Publishers, 1997.

LAVIN, M. R. (1992). Business Information: how to find it, how to use it. Phoenix: Oryx, 1992

LAWLER, E. (1996). *From the ground up*. S.F.: Jossey Bass Publishers, 1996.

LEMOS, C. (2000). "Inovação na era do conhecimento". In: *Parcerias estratégicas*, n. 08, p.157-179, 2000

LOH, S. et al, (2000). "Concept-based knowledge discovery in texts extracted from the *web*". *ACM SIGKDD EXPLORATIONS*, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, v.2, n.1, p.29-39. 2000.

MACAMBIRA, J. R. (1999). A Estrutura Morfo-Sintática do Português: aplicação do estruturalismo linguístico. São Paulo, Ed. Pioneira, 1999.

McCLELLAND, D. C. (1973). *Testing for competence rather than intelligence*. American Psychologist, 1973.

McKEOWN, K.; RADEV, D.R. (1995). "Generating summaries of multiple news

articles". *In: SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL*, SIGIR, 1995. Proceedings... New York: Association for Computing Machinery, 1995.

MEGAPUTER Intelligence Inc, MicroSystem Co. (2006). 1997-2006. *Documentação on-line.* Disponível em <http://www.megaputer.com>. Acesso em 16/03/2006.

MIIKE, S. et al. (1994). "A full-text retrieval system with a dynamic abstract generation function". *In: SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL*, SIGIR, VII. Proceedings... London: Springer-Verlag. 1994.

MILLER, G. (1996). "Wordnet: An online lexical database", *International Journal of Lexicography*, 3(4):235-312, 1996.

MIRANDA, R.; RODRIGUES, S.; SAMPAIO, J.O.; SOUZA,J.M.(2004). "Recomendação Automática de Conteúdo para Integrantes de Comunidades Virtuais". In: *IV Simpósio de Desenvolvimento e Manutenção de Software da Marinha,* 2004, Rio de Janeiro. Anais do IV Simpósio de Desenvolvimento e Manutenção de *Software* da Marinha, 2004.

NONAKA, I.; TAKEUCHI, H. (1997). Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro, Editora Campus, 1997.

OARD, D.W.; MARCHIONINI, G.(1996). *A conceptual framework for text filtering. Technical Report*, University of Maryland, 1996. Disponível em <http://www.ee.umd.edu/medlab/filter>. Acessado em 20/05/2005.

OLIVEIRA, J.; RODRIGUES, S.; SOUZA, J. M. (2004). "Competence mining for virtual scientific community creation". *International Journal of web Based Communities*, Grã-Bretanha, v. 1, n. 1, p. 90-102, 2004.

ONTOWEB (2006). *Gestão do Conhecimento com Inteligência Artificial.* Disponível em <http://www.ontoweb.com.br>. Acesso em 20/03/2006.

ORENGO, V.M.; HUYCK, C.R. (2001). "A Stemming Algorithm for the Portuguese

Language". *In: Proceedings of the SPIRE Conference.* Laguna de San Raphael, pg 13-15, 2001.

PINCHOT, G (1989). "Intrapreneuring". São Paulo: Ed. Harbra, 1989.

PINHEIRO, B. A. et al. (1996). *Formação profissional Senac.* Rio de Janeiro, Senac, 1996.

PORTER, M.F. (1980). *The Porter Stemming Algorithm.* Disponível em: <http://www.tartarus.org/~martin/PorterStemmer> Acesso em Fevereiro de 2004.

PORTER, M.F. (1997). "An algorithm for suffix stripping". *In: Readings in Information Retrieval*, 313-316. Morgan Kaufmann, 1997.

RABAGLIO, I. (2001). *Seleção por competências.* São Paulo. Editora Educator, 2001.

RESENDE, E. (2003). O livro das competências: desenvolvimento das competências - a melhor auto-ajuda para pessoas, organizações e sociedade. 2ª Edição, Rio de Janeiro, Editora Qualitymark, 2003.

RIFKIN, J. (2004). Fim dos Empregos: o Contínuo Crescimento do Desemprego em Todo o Mundo. São Paulo, Editora Makron Books, 2004.

RIJSBERGEN, C. J. (1999). *Information Retrieval.* Disponível em <http://www.dcs.gla.ac.uk/~iain/keith/>. Acessado em 15/01/2005.

RILOFF, E. (1995). "Little words can make a big difference for text classifications. Proceedings", *ACM SIGIR'95*, Washington. ACM PRESS, p. 130-136, 1995.

ROCHA, A. F. (2002). "Qualificações e competências no mutável ambiente das organizações: um estudo inesgotável". In: GOULART, Íris B. *Psicologia Organizacional e do Trabalho.* São Paulo: Casa do Psicólogo, 2002.

RODRIGUES, S.; OLIVEIRA, J.; SOUZA, J. M. (2004). "Competence Mining for Virtual Scientific Community Creation". *In: IADIS International Conference web Based*

Communities Creation, 2004, Lisboa. Anais do IADIS International Conference *web* Based Communities Creation. Lisboa, 2004.

RODRIGUEZ y RODRIGUEZ, M. V. (2001). Organização do conhecimento: a implantação de universidades corporativas. Florianópolis: v.3, n.6, 2001.

RUAS, R. L. (1999). "A problemática do desenvolvimento de competências e a contribuição da aprendizagem organizacional". *In: SEMINÁRIO INTERNACINAL COMPETITIVIDADE BASEADA NO CONHECIMENTO*, São Paulo, 1999.

SALTON, G. (1983). "Introduction to Modern InformationRetrieval". MCGRAW-HILL, 1983.

SALTON, G.; BUCKLEY, C. (1987). Term Weighting Approaches in automatic Text Retrieval. Technical Report. New York: Department of Computer Science, Cornell University. 1987.

SAS TECHNOLOGIES (2003). SAS Text Miner – Capitalize on the value hidden in textual information, SAS Institute Inc, 2003

SAS TECHNOLOGIES (2005). "Mining more than gold". *SAS Customer Success*. Disponível em <http://www.sas.com/success/index.html>. Visualizado em 29/12/2005.

SAS TECHNOLOGIES (2006). "Getting Started with SAS Enterprise Miner 5.2". *SAS Publishing*, 156 pg, Abril 2006.

SINGH, H. (2001). Data Warehouse: Conceitos, Tecnologias, Implementação e Gerenciamento. São Paulo, Makron Books, 2001.

SHÜTZE, H. (1992). "Dimensions of Meaning". *In: SUPERCOMPUTING (IEE)*, 1992, Mineapolis. Proceedings... Mineaolis, p. 787-796, 1992.

SPARCK-JONES, K.; WILLET, P. (1997). *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997.

SU, L T. (1998). "Value of search results as a whole as the best measure of information retrieval performance". *Information Processing and Management.* Vol.34, nº 5, 557-579. 1998.

TAN, A. (1999). "Text Mining: the state of the art and the challenges". *In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases – PAKDD'99*, p.65-70, Beijing, April 1999. Disponível em <http://ww.ewastrategist.com/papers/text_mining_kdad99.pdf>. Acesso em 23/01/2004.

TEIXEIRA, J. F. (2005). "Tecnologia da Informação para a Gestão do Conhecimento". *HSM Management.* Disponível em <http://www.intermanagers.com.br>. Acesso em 26/09/2005.

TERRA, J. C. C. (2000). Gestão do Conhecimento: o grande desafio empresarial - uma abordagem baseada no aprendizado e na criatividade. São Paulo, Negócio Editora, 2000.

TRAVASSOS, G.; GUROV, D., AMARAL, D. (2002). "Introdução à Engenharia de *Software* Experimental". Relatório Técnico – Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, 52 pg, 2002.

WAH, L. (2000). "Muito além de um modismo". *HSM Management*, ano 4, n. 22, 2000. Disponível em <http://www.intermanagers.com.br>. Acesso em 17/11/2005.

WIVES, L. K. (1997). *Indexação de Documentos Textuais*. Porto Alegre: Universidade Federal do Rio Grande do Sul, Instituto de Informática, 1997.

WIVES, L. K. (2000). *Tecnologias de descoberta de conhecimento em textos aplicadas à Inteligência Competitiva.* Porto Alegre: Universidade Federal do Rio Grande do Sul, Instituto de Informática, 2000.

WIVES, L. K.; RODRIGUES, N. A. (2000). "Eurekha". *READ - Revista Eletrônica da Administração*. Porto Alegre: Universidade Federal do Rio Grande do Sul, v. 6, n. 5, 2000.

WIVES, L. K. (2004). Utilizando Conceitos como descritores de Textos para o processo de identificação de conglomerados (clustering) de documentos. Tese (doutorado) -- Instituto de Informática, UFRGS, Porto Alegre, 2004

ZARIFIAN, P. (2001). *Objetivo competência: por uma nova lógica*. São Paulo, Editora Atlas, 2001.

APÊNDICE I – BASE DE DADOS

O modelo de dados foi desenhado no *software* ERWin, da Computer Associates®. A seguir, estão dispostos o Modelo Físico e o Dicionário de Dados, onde estão contempladas a definições das tabelas e o significado dos devidos campos.

Modelo de Dados

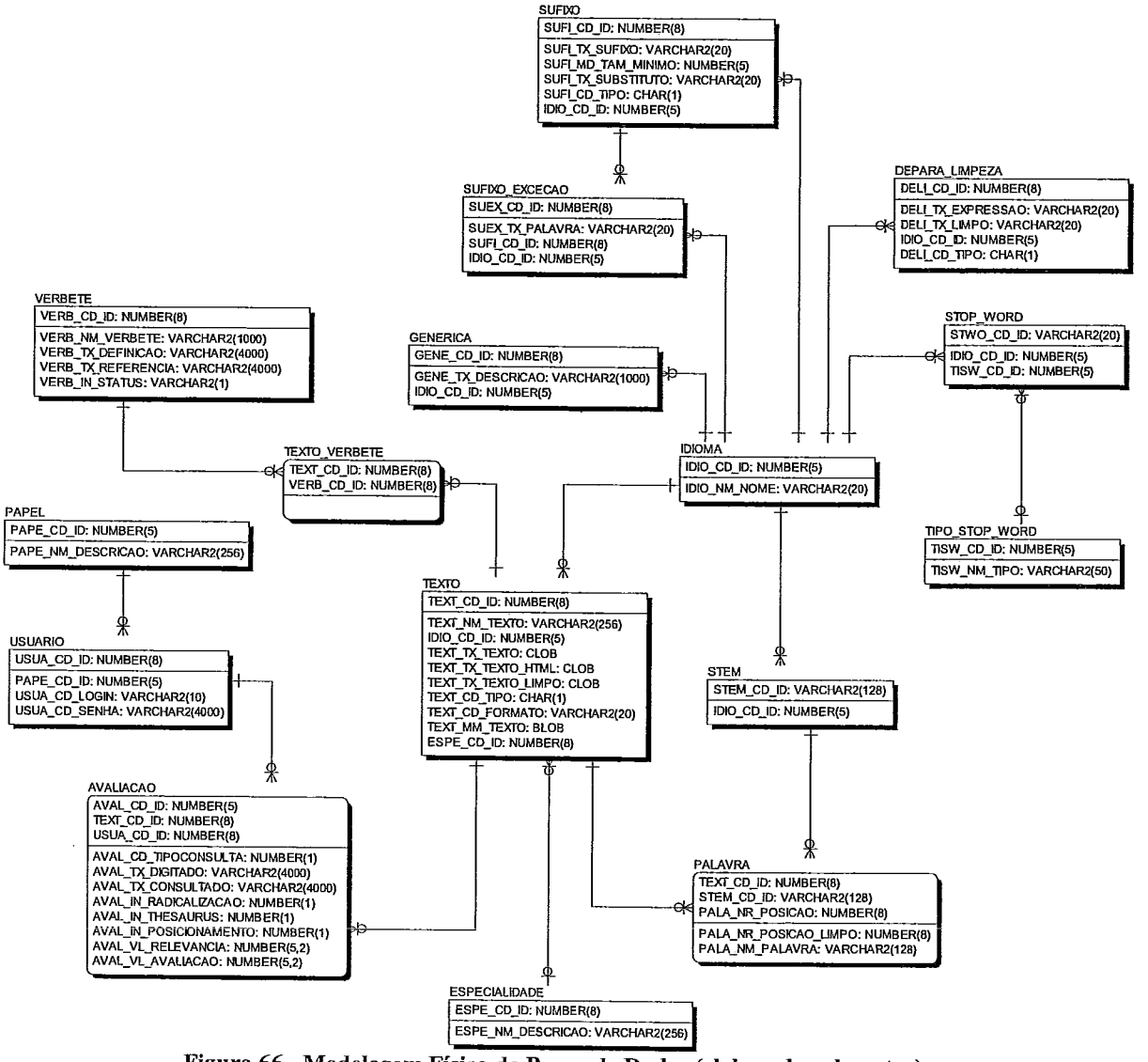


Figura 66 - Modelagem Física do Banco de Dados (elaborada pelo autor)

Dicionário de Dados

Tabelas	
Nome	**Comentários**
AVALIACAO	Entidade utilizada durante o processo de avaliação dos resultados da busca.
DEPARA_LIMPEZA	Entidade utilizada durante o processo de limpeza de textos antes de seu processamento propriamente dito. Em linhas gerais, esta entidade armazena palavras que quando presentes no texto devem ser substituídas por outra pré-definida visando a limpeza do texto. Um exemplo de substituição entre palavras indicada pela tabela é a troca de 'é' por 'e'.
ESPECIALIDADE	Entidade que armazena as Especialidades que irão compor os Módulos de Qualificação. Uma especialidade corresponde a uma das diversas "matérias" que são lecionadas e diagnosticadas nos diversos módulos da qualificação profissional.
GENERICA	Tabela que armazena itens de descrições genéricos utilizados principalmente na apresentação de mensagens e palavras que necessitem de armazenamento no banco de dados. Exemplo de itens genéricos são: (1) a frase utilizada quando um aluno está inscrito em uma turma e alguém tenta inscrevê-lo em outra 'Este aluno já está inscrito em outra turma.'; ou (2) a palavra 'Domingo' utilizada para denotar o dia da semana.
IDIOMA	Entidade que armazena idiomas utilizados no sistema como, por exemplo, 'português' e 'espanhol'.
MENU_ITEM	Entidade que armazena os itens de menu.
PALAVRA	Entidade utilizada para o armazenamento de informações sobre palavras contidas nos textos processados pelo minerador do módulo de lacunas. Esta entidade armazena as palavras presentes no texto, seus respectivos radicais e as respectivas posições tanto no texto limpo após processamento (sem palavras irrelevantes), quanto no texto original.
PAPEL	Entidade que armazena o Papel de acesso ao sistema (Especialista, Operador e Usuário comum são exemplos).
PAPEL_MENU_ITEM	Entidade que armazena os itens de menu pertinentes a um determinado perfil (papel) de acesso.
STEM	A entidade STEM armazena o produto resultante da aplicação do algoritmo de radicalização (Stemming) sobre as palavras presentes nos textos cadastrados no sistema para mineração. Tal produto nada mais é que as palavras em sua forma 'raiz', ou seja, palavras sem prefixos e sufixos e com as substituições realizadas e indicadas na tabela depara_limpeza.
STOP_WORD	A entidade STOP_WORD é a entidade que representa todas as palavras ditas irrelevantes para o contexto da mineração de textos. Tais palavras são consideradas irrelevantes por serem muito comuns e não influenciarem de forma significativa no processo de indexação ou em mecanismos de busca. Alguns exemplos de palavras irrelevantes são: advérbios, artigos, preposições, etc.
SUFIXO	Entidade que armazena os sufixos que deverão ser retirados ou substituídos de palavras dos textos cadastrados no momento de seu processamento (radicalização). Os sufixos podem ser dos seguintes tipos: plural, feminino, advérbio, aumentativo, substantivo, verbo e vogal temática. São exemplos de substituições são: a troca de 'ões' por 'ão'; e a retirada do 's' indicativo de plural do final das palavras.
SUFIXO_EXCECAO	A entidade SUFIXO_EXCECAO armazena as possíveis exceções a serem avaliadas durante o processo de retirada/substituição de sufixos das palavras contidas nos textos cadastrados. Durante a retirada de um sufixo, é necessário determinar se o sufixo que está sendo retirado/substituído é de fato um sufixo ou é parte constituinte de alguma palavra que possui um final igual ao sufixo. Um exemplo de exceção para o sufixo 'is' está contido na palavra 'depois'. Durante o processo de radicalização o sufixo 'is' seria substituído por 'il', o que para a palavra 'depois' não faz sentido.
TEXTO	Representa todos os textos cadastrados no sistema.
TEXTO_VERBETE	Entidade de relaciona o verbete com o respectivo texto cadastrado.
TIPO_STOP_WORD	Entidade que indica os tipos de stop words cadastrados no sistema, por exemplo: advérbios, numeração, substantivos gerais, etc.
USUARIO	Entidade que armazena os logins e senhas cadastrados no sistema
VERBETE	Entidade que armazena os termos técnicos que formarão o glossário.

Tabela AVALIACAO			
Nome do Campo	Tipo de Dados	Opção Null	Comentário
AVAL_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela AVALIACAO.
TEXT_CD_ID	NUMBER(8)	NOT NULL	Chave primaria da tabela TEXTO. Indica o texto avaliado.
USUA_CD_ID	NUMBER(8)	NOT NULL	Identificador (chave primária) da tabela Usuário. Indica o usuário que realizou a Avaliação
AVAL_CD_TIPOCONSULTA	NUMBER(1)	NOT NULL	1:Simples, 2:Avancada, 3:Contextual, 4:Lacunas.
AVAL_TX_DIGITADO	VARCHAR2(4000)	NOT NULL	Texto digitado pelo usuario.
AVAL_TX_CONSULTADO	VARCHAR2(4000)	NOT NULL	Texto consultado pela ferramenta.
AVAL_IN_RADICALIZACAO	NUMBER(1)	NOT NULL	Retirar Palavras Irrelevantes e aplicar Radicalizacao? 0:Nao, 1:Sim
AVAL_IN_THESAURUS	NUMBER(1)	NOT NULL	Utilizar Thesaurus? 0:Nao, 1:Sim
AVAL_IN_POSICIONAMENTO	NUMBER(1)	NOT NULL	Verificar o Posicionamento das palavras? 0:Nao, 1:Sim
AVAL_VL_RELEVANCIA	NUMBER(5,2)	NOT NULL	Valor da Relevancia obtida pela ferramenta.
AVAL_VL_AVALIACAO	NUMBER(5,2)	NULL	Null:Nao Avaliado, 0:Irrelevante, 1:Pouco Relevante, 2:Relevante, 3:Muito Relevante

Tabela DEPARA_LIMPEZA			
Nome do Campo	Tipo de Dados	Opção Null	Comentário
DELI_CD_ID	NUMBER(8)	NOT NULL	Identificadro único da tabela DEPARA_LIMPEZA. Representa também a palavra a ser substiuída durante o processo de limpeza.
DELI_TX_EXPRESSAO	VARCHAR2(20)	NOT NULL	Representa a palavra a ser substiuída durante o processo de limpeza.
DELI_TX_LIMPO	VARCHAR2(20)	NOT NULL	Representa a palavra substituta a ser utilizada durante o processo de limpeza.
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador único (chave

			primária) da tabela IDIOMA.
DELI_CD_TIPO	CHAR(1)	NOT NULL	Representa o tipo de limpeza que está sendo feita. A princípio os tipos serão: A - Acentuação O - Outros

Tabela ESPECIALIDADE			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
ESPE_CD_ID	NUMBER(8)	NOT NULL	**Identificador (chave primária)** da tabela ESPECIALIDADE
ESPE_NM_DESCRICAO	VARCHAR2(256)	NULL	Descrição da Especialidade

Tabela GENERICA			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
GENE_CD_ID	NUMBER(8)	NOT NULL	Identificador único (chave primária) da tabela GENERICA.
GENE_TX_DESCRICAO	VARCHAR2(1000)	NOT NULL	Valor de cada item da tabela GENERICA.
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador único (chave primária) da tabela IDIOMA.

Tabela IDIOMA			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Idioma. Quando chave estrangeira, define o idioma dos registros.
IDIO_NM_NOME	VARCHAR2(20)	NOT NULL	Nome do Idioma: Português, Espanhol, Inglês

Tabela MENU_ITEM			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
MEIT_CD_ID	NUMBER	NOT NULL	**Identificador (chave primária)** da tabela Menu_Item
MEIT_NM_IMAGEM	VARCHAR2(100)	NOT NULL	Imagem do item de menu
MEIT_NM_HREF	VARCHAR2(200)	NULL	Referencia do link do menu

Tabela PALAVRA			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
TEXT_CD_ID	NUMBER(8)	NOT NULL	**Identificador único (**chave primária) da tabela TEXTO.
STEM_CD_ID	VARCHAR2(128)	NOT NULL	Identificador único (chave primária) da tabela STEM. Representa também a forma

			básica (raiz) de uma palavra.
PALA_NR_POSICAO	NUMBER(8)	NOT NULL	Indica a posição da palavra no texto completo.
PALA_NR_POSICAO_LIMPO	NUMBER(8)	NULL	Posição da palavra no texto limpo (sem stop words).
PALA_NM_PALAVRA	VARCHAR2(128)	NOT NULL	A palavra completa que deu origem ao stem indicado.

Tabela PAPEL

Nome do Campo	Tipo de Dados	Opção Null	Comentário
PAPE_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Papel
PAPE_NM_DESCRICAO	VARCHAR2(256)	NULL	Indica o nome do Papel (Especialista, Operador ou Usuário Comum, por exemplo)

Tabela PAPEL_MENU_ITEM

Nome do Campo	Tipo de Dados	Opção Null	Comentário
PAPE_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Papel
MEIT_CD_ID	NUMBER	NOT NULL	Identificador (chave primária) da tabela Menu_Item

Tabela STEM

Nome do Campo	Tipo de Dados	Opção Null	Comentário
STEM_CD_ID	VARCHAR2(128)	NOT NULL	Identificador único (chave primária) da tabela STEM. Representa também a forma básica (raiz) de uma palavra.
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Idioma. Quando chave estrangeira, define o idioma dos registros.

Tabela STOP_WORD

Nome do Campo	Tipo de Dados	Opção Null	Comentário
STWO_CD_ID	VARCHAR2(20)	NOT NULL	Identificador único (chave primária) da tabela STOP_WORD. Este atributo armazena a stop_word propriamente dita.
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Idioma. Quando chave estrangeira, define o idioma dos registros.
TISW_CD_ID	NUMBER(5)	NULL	Identificador (chave primária)

			da tabela TIPO_STOP_WORD.

Tabela SUFIXO			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
SUFI_CD_ID	NUMBER(8)	NOT NULL	Identificador único (chave primária) da tabela SUFIXO.
SUFI_TX_SUFIXO	VARCHAR2(20)	NOT NULL	Representa os sufixos passíveis de serem encontrados no algoritmo STEMMER, os quais deverão receber tratamento especial.
SUFI_MD_TAM_MINIMO	NUMBER(5)	NOT NULL	Tamanho mínimo com o qual o STEM deve ficar após a retirada do sufixo. Caso esse tamanho não seja satisfeito, o mesmo não poderá ser extirpado da palavra.
SUFI_TX_SUBSTITUTO	VARCHAR2(20)	NULL	Novo sufixo que irá substituir o sufixo em avaliação.
SUFI_CD_TIPO	CHAR(1)	NOT NULL	Tipo do sufixo sendo avaliado. A princípio este tipo poderá ser: P - Plural F - Feminino A - Advérbio U - Aumentativo S - Substantivo V - Verbo T - Vogal Temática
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Idioma. Quando chave estrangeira, define o idioma dos registros.

Tabela SUFIXO_EXCECAO			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
SUEX_CD_ID	NUMBER(8)	NOT NULL	Identificador único (chave primária) da tabela SUFIXO_EXCECAO.
SUEX_TX_PALAVRA	VARCHAR2(20)	NOT NULL	Representa a palavra através da qual será avaliado se o caso em questão se trata ou não de uma exceção.
SUFI_CD_ID	NUMBER(8)	NOT NULL	Identificador único (chave primária) da tabela SUFIXO.
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador único (chave primária) da tabela IDIOMA.

Tabela TEXTO			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
TEXT_CD_ID	NUMBER(8)	NOT NULL	Identificador único (chave primária) da tabela TEXTO.
TEXT_NM_TEXTO	VARCHAR2(256)	NOT NULL	Nome do texto.
IDIO_CD_ID	NUMBER(5)	NOT NULL	Identificador único (chave primária) da tabela IDIOMA.
TEXT_TX_TEXTO	CLOB	NULL	Texto completo em formato simples (CLOB), sem formatação.
TEXT_TX_TEXTO_HTML	CLOB	NULL	Texto completo em formato HTML, com formatação.
TEXT_TX_TEXTO_LIMPO	CLOB	NULL	Texto já limpo, sem stop words.
TEXT_CD_TIPO	CHAR(1)	NULL	Identifica o tipo de texto cadastrado no sistema, podendo assumir os valores V - Verbete, Q - Questões, R - Documentos de Referência, O - Documentos de Opinião, L - Livros, C - Textos Complementares.
TEXT_CD_FORMATO	VARCHAR2(20)	NOT NULL	Formato do texto. Utilizado para processamento do INSO_FILTER. Pode assumir os valores: text ou binary.
TEXT_MM_TEXTO	BLOB	NOT NULL	Texto completo, em seu formato original (possivelmente binário).
ESPE_CD_ID	NUMBER(8)	NULL	Identificador (chave primária) da tabela ESPECIALIDADE

Tabela TEXTO_VERBETE			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
TEXT_CD_ID	NUMBER(8)	NOT NULL	Identificador (chave primária) da tabela Texto
VERB_CD_ID	NUMBER(8)	NOT NULL	Identificador (chave primária) da tabela Verbete

Tabela TIPO_STOP_WORD			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
TISW_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela TIPO_STOP_WORD.
TISW_NM_TIPO	VARCHAR2(50)	NOT NULL	Descrição do Tipo de Stop Word (Números, Meses,

			Substantivos gerais, etc.)

Tabela USUARIO			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
USUA_CD_ID	NUMBER(8)	NOT NULL	Identificador (chave primária) da tabela Usuário
PAPE_CD_ID	NUMBER(5)	NOT NULL	Identificador (chave primária) da tabela Papel. Indica se o usuário é Especialista, Operador ou Usuário Comum
USUA_CD_LOGIN	VARCHAR2(10)	NOT NULL	Login de acesso ao Sistema
USUA_CD_SENHA	VARCHAR2(4000)	NOT NULL	Senha de Acesso ao Sistema

Tabela VERBETE			
Nome do Campo	**Tipo de Dados**	**Opção Null**	**Comentário**
VERB_CD_ID	NUMBER(8)	NOT NULL	Identificador (chave primária) da tabela Verbete
VERB_NM_VERBETE	VARCHAR2(1000)	NOT NULL	Nome do Verbete
VERB_TX_DEFINICAO	VARCHAR2(4000)	NULL	Definição geral do verbete, ou seja, seu significado.
VERB_TX_REFERENCIA	VARCHAR2(4000)	NULL	Referências Bibliográficas
VERB_IN_STATUS	VARCHAR2(1)	NULL	Indica se o Verbete já foi carregado no sistema de extração de informações

APÊNDICE II – STOP WORDS

a	dom	nosso	se
á	domingo	nossos	seg
abr	dos	nov	segunda
abril	doze	nove	segunda-feira
absolutamente	ducentesimo	novecentos	segundo
acaso	dum	novembro	seis
ago	duma	noventa	seiscentesimo
agora	dumas	num	seiscentos
agosto	duns	numa	sem
ah	duzentos	numas	Sempre
ai	e	nunca	Senhor
ainda	é	nuns	Senhora
algo	efetivamente	o	Senhoria
alguem	eia	oba	Senhorita
algum	ela	octingentesimo	septingentesimo
alguma	elas	octogesimo	septuagesimo
algumas	ele	oh	sessenta
alguns	eles	oitava	set
ali	em	oitavo	sete
alo	en	oitenta	setecentos
alto	entao	oito	setembro
amanha	então	oitocentos	setenta
amanhã	essa	ola	setima
ante	essas	onde	setimo
anteontem	esse	ontem	seu
antes	esses	onze	seus
ao	esta	opa	sex
aonde	estas	os	sexagesimo
aos	este	ou	sexcentesimo
apos	estes	outra	sexta
após	eu	outras	sexta-feira
aquela	f	outrem	sexto
aquelas	fev	outro	si
aquele	fevereiro	outrora	silencio
aqueles	fielmente	outros	sim
aqui	g	outubro	sob
aquilo	h	oxala	sobre
as	hem	p	sua
às	hoje	para	t
assaz	hum	pára	tais
assim	i	pela	tal
ate	ih	pelas	talvez
até	isso	pelo	tanta
avante	isto	pelos	tantas
b	j	perante	tanto
basta	ja	pior	tantos
bastante	já	por	tao
bem	jamais	porem	tarde
bilhao	jan	porém	te
bilionesimo	janeiro	porventura	ter

bilionesimos	jul	possivelmente	terca
bis	julho	pouca	terça
bravo	jun	poucas	terca-feira
breve	junho	pouco	terça-feira
c	k	poucos	terceira
cada	l	primeira	terceiro
calmamente	levemente	primeiramente	teu
cedo	lhe	primeiro	teus
cem	lhes	provavelmente	ti
centesimo	logo	psit	toda
certa	m	psiu	todas
certamente	mai	puxa	todo
certas	maio	q	todos
certo	mais	qua	tras
certos	mal	quadragesimo	trás
chi	mar	quadrigentesimo	trecentesimo
cinco	marco	quais	tres
cinquenta	março	quaisquer	treze
com	mas	qual	trezentos
comigo	me	qualquer	trigesimo
como	melhor	quanta	trinta
conosco	menos	quantas	tu
consigo	meu	quanto	tudo
contigo	meus	quantos	u
contra	mil	quao	ue
convosco	milesimo	quarenta	uh
coragem	milesimos	quarta	ui
corretamente	milhao	quarta-feira	ultimamente
cuja	milionesimo	quarto	um
cujo	milionesimos	quase	uma
d	mim	quatorze	umas
da	muita	quatro	uns
das	muitas	quatrocentos	v
de	muito	que	vamos
debalde	muitos	quem	varia
decima	n	qui	varias
decimo	na	quica	vario
demais	nada	quingentesimo	varios
depois	nao	quinhentos	vigesimo
depressa	não	quinquagesimo	vinte
desde	nas	quinta	viva
devagar	nenhum	quinta-feira	vóce
deveras	nenhuma	quinto	você
deverás	nenhumas	quinze	vos
dez	nenhuns	r	vós
dezembro	ninguem	realmente	vossa
dezenove	no	s	vosso
dezesseis	nogentesimo	sab	vossos
dezessete	nonagesimo	sáb	w
dezoito	nono	sabado	x
Do	nos	sábado	y
Dois	nós	são	z

APÊNDICE III – MODELAGEM DO CUBO OLAP

O cubo OLAP utilizado para visualizar a avaliação dos resultados é composto por 5 Dimensões e 6 medidas de Fatos.

A Figura 67 apresenta a modelagem do cubo SMiner.

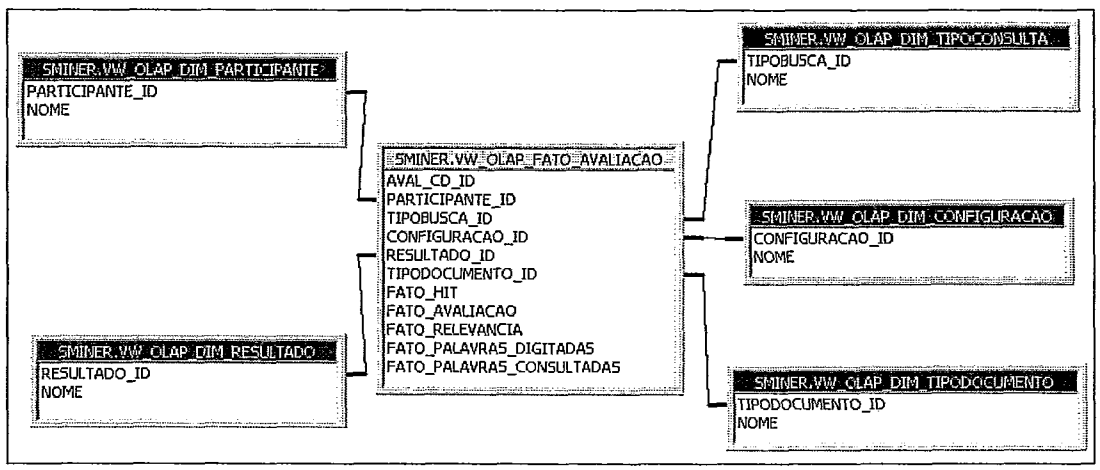


Figura 67 - Modelagem do Cubo OLAP (elaborada pelo autor)

As medidas da tabela Fato correspondem aos indicadores numéricos que serão analisados. As medidas de Fato deste cubo são:

- Quantidade de consultas;
- Quantidade de avaliações;
- Média das avaliações;
- Média das Relevâncias obtidas pelo Minerador;
- Média de palavras digitadas na busca;
- Média de palavras consultadas pelo minerador.

A partir destas medidas é possível manipular o cubo através das dimensões.

A Dimensão é a propriedade do cubo responsável por qualificar o Fato a ser analisado, ou seja, são os ângulos pelos quais as informações contidas em um Fato podem ser analisadas. As dimensões proporcionam liberdade para realizar combinações entre elas, gerando relatórios bem mais qualitativos. Essa riqueza de combinações permite também que o usuário descubra padrões ocultos, já que é possível sumarizar as informações de forma completamente livre.

As dimensões utilizadas neste cubo foram:

- Tipos de Participante;
- Tipos de Consulta;
- Tipos de Documento;
- Tipos de Resultado;
- .Tipos de Configuração.