Universidade Federal do Espírito Santo
Centro Tecnológico
Programa de Pós-Graduação em Informática

Anderson Poltronieri

Modelo Gráfico de Recuperação de Informação Semântica

Vitória –ES

Anderson Poltronieri

Modelo Gráfico de Recuperação de Informação Semântica

Dissertação submetida ao corpo docente do Programa de Pós-graduação em Informática da Universidade Federal do Estado do Espírito Santo, como requisito parcial para a obtenção do título de Mestre em Informática.
Orientador: Prof. Dr. Elias de Oliveira

Vitória – ES, 16 de dezembro de 2006.

FOLHA DE ROSTO

SUBSTITUIR PELO DA UFES

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Poltronieri, Anderson, 1976-
P779m Modelo gráfico de recuperação de informação semântica / Anderson Poltronieri. – 2006.
109 f. : il.

Orientador: Elias de Oliveira.
Dissertação (mestrado) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Recuperação da informação. 2. Indexação. 3. Ferramentas de busca. 4. Linguagem de programação (Computadores) - Semântica. 5. Documentos - Classificação.
I. Oliveira, Elias de. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 004

Agradecimentos

Agradeço a Deus, meu amigo fiel e companheiro. Meu amigo da noite e do dia, das tempestades e das calmarias. Amigo que me deu força e me guiou até aqui em todos os caminhos da minha vida.

Agradeço a Glau, minha esposa, minha companheira e minha amiga. Obrigado pela sua paciência, pelo seu amor e compreensão quando minhas ausências foram necessárias. Sem você eu não teria conseguido.

A Victoria e Pedro, minhas razões de lutar.

Não poderia deixar de agradecer ao meu mestre Elias. Chamo-o de mestre no sentido real da palavra: *Pessoa que ensina, Homem de saber.* Parabéns e obrigado por tudo Elias.

E finalmente agradeço a meus pais, sem o exemplo de lutas com vitórias e derrotas eu não teria condições de enfrentar minhas próprias batalhas.

Figuras

Sumário

Resumo

Desde os primórdios da história das civilizações, a humanidade tem buscado transmitir informações e desenvolver mecanismos de armazenamento de tais informações. No entanto, o volume de informações produzidas tem exigido que o homem utilize cada vez mais ferramentas computacionais para permitir a recuperação das informações contidas nos seus repositórios de dados. Neste trabalho apresentamos as diversas etapas necessárias para automatizar o processo de coleta, indexação e classificação de documentos e propõe um modelo gráfico tridimensional(3D) para visualização de consultas destes documentos através de um sistema de busca semântica. Para chegarmos ao modelo gráfico 3D proposto, descrevemos toda a arquitetura do sistema que foi estruturado em 5 camadas: alimentação, indexação, comparação, agrupamento e visualização. No estudo destas camadas apresentamos modelos e propostas de diferentes autores em cada uma delas. Focando nossos trabalhos nos modelos de visualização apresentamos um modelo de representação 3D chamado MOGRIS(*Modelo Gráfico de Representação de Informação Semântica)*. Por fim, apresentamos um sistema protótipo chamado SIGRIS*(Sistema Gráfico de Recuperação de Informação Semântica)* que utiliza o MOGRIS, e é baseado na arquitetura de 5 camadas proposta em nossa modelagem.

Abstract

Since the beginning of the history of the civilizations, the mankind has searched to transmit information and to develop mechanisms of storage of such information. However, the volume of produced information has demanded that the man uses each time more computational tools to allow the recovery of the information contained in its repositories of data. In this work we present the diverse stages necessary to automatize the collection process, document indexation and classification and considers a three-dimensional graphical model (3D) for visualization of consultations of these documents through a search system semantics. To arrive at the graphical model 3D considered, we describe all the architecture of the system that was structuralized in 5 layers: feeding, indexation, comparison, grouping and visualization. In the study of these layers we present models and proposals of different authors in each one of them. Focusing our works in the visualization models we present a representation model 3D called GMRIS(Graphical Model of Representation of Information Semantics). Finally, we present a prototype system called GSRIS (Graphical System of Recovery of Information Semantics) that it uses the GMRIS and based on the architecture of 5 layers proposal in our modeling.

Capítulo 1

Introdução

Desde os primórdios da história das civilizações, a humanidade tem buscado transmitir e desenvolver mecanismos de armazenamento de informações. Já na pré-história o homem utiliza-se das pinturas rupestres para deixar registrada sua passagem por certas regiões bem como, deixar registrada a informação para os que viessem posteriormente.

O advento da escrita, por volta de 6000 a.C surgindo independentemente no Egito, na Mesopotâmia e na China[3], tornou possível armazenar conhecimentos em uma escala muitas vezes superior àquela até então existente. Surgida entre diferentes povos, a escrita passou a atuar como instrumento de aproximação cultural e social[3]. Então, através de entalhes em pedra, madeira ou placas de barro o homem começou a criar os seus primeiros documentos rudimentares.

Avançando alguns milhares de anos na história, encontramos documentos produzidos em papiros, pergaminhos e papel. Existem registros de inscrições em papiro datadas de 2200 a.C e em pergaminhos por volta de 2000 a.C. [2], e mais tarde em papel, por volta do ano 105d.C. A invenção de papel é atribuída a T'sai Lun na China, fabricado a partir de fibras de cânhamo trituradas e revestidas de uma fina camada de cálcio, alumínio e sílica[3].

Naquele tempo a informação entre os povos trafegava via cavalos, barcos e longas caminhadas. Informações como ordens imperiais, resultados de campanhas militares e notícias de amigos distantes demoravam dias ou até meses para chegar ao seu destino.

De lá para cá o homem vem desenvolvendo mecanismos para agilizar o processo de transmissão de informação e melhorar a forma de armazenamento das mesmas. Com a invenção da imprensa de caracteres móveis, datada de 1440, por Johannes Gutenberg[1], o homem deu o primeiro passo no caminho da produção de documentos em massa no formato com que estamos familiarizados, tanto em livros impressos como em mídia digital. No entanto a informação ainda continuava demorando a chegar ao destino, seja ela em forma de cartas, livros ou comunicados oficiais.

E foi buscando agilizar a comunicação à distância que o homem fez, durante muitos anos da história, uso de recursos naturais para tal, como fogueiras em montes, troncos de árvores para comunicação em meio a florestas e pombos correios. No entanto, somente o surgimento de processos elétricos fez com que a transmissão de informações fosse realmente agilizada. Um dos pioneiros deste processo foi o médico espanhol Francisco Salvá, de Barcelona. Em 1795 transmitiu mensagens por meio da descarga de um condensador[4], sendo o primeiro modelo de telégrafo desenvolvido. A partir daí a velocidade com que os dados passaram a chegar aos seus destinos aumentou rapidamente.

Em paralelo ao processo de desenvolvimento de mecanismos de comunicação, o número de documentos produzidos nas diferentes áreas do conhecimento aumentava cada vez mais. Grandes quantidades de documentos eram produzidas por filósofos, matemáticos, escritores e cientistas. Tais documentos compostos por livros e documentos em couro, papiro e pergaminhos eram organizados em bibliotecas que já exigiam profissionais especializados em catalogar todos os documentos e dispô-los de modo a facilitar o acesso aos mesmos.

Voltando à preocupação com a transmissão da informação, em 1875 o escocês Alexander Graham Bell terminou a construção do primeiro telefone. A invenção do telefone foi um dos grandes marcos da comunicação para a humanidade. Hoje, passados cerca de 130 anos, o telefone tornou-se para muitas pessoas um objeto imprescindível de trabalho.

Outro marco surgiu no início do século XX, o advento dos primeiros computadores já sinalizava para o “boom” tecnológico que invadiriam nossas vidas. Milhares de aplicações puderam sair do papel tomando corpo, e como um processo em cadeia, gerando mais tecnologia.

Os computadores trouxeram ao homem a capacidade de trabalhar com volume imenso de informações que lhe permitiam calcular, pesquisar e obter resultados jamais pensados em trabalhos manuais. Com os computadores também vieram programas que nos permitiam organizar e catalogar documentos em formato digital. Não demorou muito e surgiu a necessidade de compartilhar estes acervos de dados, agora também armazenados em grandes computadores, entre vários usuários. Tentando realizar esta tarefa de compartilhamento, no final da década de 60, a ARPANET foi lançada pelo governo americano visando criar uma rede de interligação de projetos estratégico-militares.

A demanda por informação e compartilhamento de dados era crescente, e o número de usuários da ARPANET crescia dia a dia. Em 1983, o governo americano decidiu separar as redes de conteúdo civil da militar, dando origem à Internet.

Com esta nova forma de propagação da informação, e de comunicação, em questão de minutos tornou-se possível saber de um atentado, de um acidente, conversar com um amigo distante ou vizinho. Em questão de segundos podemos saber se a bolsa de valores caiu ou subiu provocando uma mudança instantânea de vida. Em instantes sabemos de desastres e momentos alegres. A informação está cada vez mais rápida e presente, já não se usam mais cavalos, ou mensageiros, mas computador e e-mails. Os documentos, que antes estavam no formato tradicional, impresso sobre os moldes da imprensa criada no século XIV, agora se encontram num formato digital.

Assim como a quantidade de informações que crescia rapidamente, também crescia o volume de materiais a serem armazenados(*e-books*, jornais, revistas, artigos). Milhões de páginas, documentos, artigos, livros estão disponíveis *on-line*. Mas como localizar estas informações? Como organizar todos estes arquivos? Como classificar estes arquivos? O que antes era feito por um bibliotecário, tornou-se uma tarefa quase impossível diante da grandiosidade de produções.

1.1) Motivação

Tentando num primeiro momento agilizar a busca de conteúdos, sites de busca como Google, Yahoo, Cadê, Altavista, dentre outros, foram criados para permitir ao usuário chegar às informações relacionadas a uma palavra chave ou frase com maior rapidez. No entanto, uma grande quantidade de resultados das buscas nestes sites não possui conteúdo com o significado desejado, muito embora possuam os termos pesquisados. Outro fator prejudicial à pesquisa *on-line* é que o número de resultados encontrados a partir de um termo tem crescido a cada dia, o que dificultada a localização de documentos de conteúdo confiável e contextualizado do que se deseja procurar.

Em outro foco de pesquisa, sistemas para controle de bibliotecas digitais também têm sido estudados e desenvolvidos. Algoritmos de classificações automáticas e categorização têm ocupado grande quantidade de pesquisadores no intuito de otimizar o processo de

identificação do conteúdo e das áreas de conhecimento de um documento visando auxiliar os bibliotecários e arquivista em parte importante dos seus trabalhos.

1.2) Objetivos

Neste trabalho apresentamos um conjunto de mecanismos de visualização e classificação de documentos. Estes mecanismos podem ser utilizados no desenvolvimento de aplicações que visem ajudar os profissionais da área de documentação e navegantes desejosos por informações.

Dentre o conjunto de ferramentas adotadas, apresentaremos dois modelos de visualização montados sobre a mesma base de informações. Esta base de informações montada e classificada poderá ser visualizada através de dois modelos: o modelo *Orbital* de navegação em um ambiente 3D na *web*, e o modelo de *Links Clusterizados* semanticamente, num padrão típico adotado por diversos portais de busca.

O modelo Orbital, trás as informações, resultado de uma consulta do usuário, organizadas num espaço tridimensional. Esta forma de visualização nos permitirá navegar entre as informações de documentos e entre as informações dos autores que compõem o acervo, sem termos que mudar de ambiente e de notação gráfica. Isto facilita o entendimento e a navegação do usuário.

Na estrutura de *Links Clusterizados* apresentamos a resposta a uma consulta ao acervo em forma de *links* organizados em níveis de visualização. Os níveis de refinamento nos permitirão navegar entre categorias, documentos e autores. Com isto o usuário pode, a qualquer momento, mudar o foco de sua consulta entre documentos e autores.

1.3) Contribuições

Diferente dos modelos apresentados na literatura, os modelos gráficos de visualização demonstrados neste trabalho, nos oferece a possibilidade de, em um único modelo gráfico ou textual, representarmos informações de estruturas distintas, documentos e autores, utilizando apenas uma única notação, comum a ambos os elementos. Esta notação simplificada favorece a familiarização do usuário com o modelo de consulta utilizado, uma vez que não há modificação das representações visuais quando mudamos do contexto da busca.

Ainda neste trabalho, visando desenvolver um modelo de visualização, apresentamos

uma arquitetura que nos guia no processo de criação de um ambiente de recuperação de informação. Através desta arquitetura, qualquer pessoa, que se interessar em trabalhar nesta área, pode adotá-la como referência dos passos que devemos tomar para construção de um ambiente que vai desde a obtenção do documento, até o processo de visualização do acervo. Mostramos que todas as atividades do processo de recuperação de informação podem ser desenvolvidas independentemente dos algoritmos e processos adotados em cada um dos níveis. Com isto, o desenvolvedor pode trabalhar na melhoria de cada um destes níveis de forma isolada, sem afetar os demais níveis da arquitetura.

1.4) Organização do trabalho

Com o objetivo de detalhar nosso trabalho, este documento está estruturado em 7 capítulos, conforme descrito abaixo.

No capítulo 2, chamado "*Visualização de Documentos*" apresentamos um conjunto de ferramentas para visualização de documentos digitais existentes e propostos na literatura. Neste capítulo também fazemos uma apresentação dos mecanismos de classificação de documentos que dão suporte a estas ferramentas de visualização de dados.

No capítulo 3, chamado "*Classificação de Documentos*" apresentamos as técnicas de classificação de documentos propostas por diversos autores. A classificação de documentos é um pré-requisito fundamental para termos um sistema de visualização de documentos confiável e coerente.

No capítulo 4 – "*Visualização de Grupos Semânticos"*, são apresentadas duas propostas de representação gráfica. Uma baseada em *links* textuais tradicionais como utilizados por diversos *sites*. A segunda forma baseia-se no modelo gráfico Orbital que é uma das propostas deste trabalho.

O quinto capítulo – "*Arquitetura do Sistema*", apresenta a construção do protótipo utilizado

experimento realizado. Serão exibidos os resultados obtidos através de dois conjuntos distintos de documentos e com processos de classificação também distintos. Estes dois conjuntos de documentos nos permitem mostrar que podemos utilizar os mecanismos de visualização independentemente dos mecanismos de classificação adotado para o acervo.

O capítulo 6 apresenta os comentários finais sobre o trabalho, destacando as dificuldades apresentadas e as soluções para que tivéssemos sucesso em nossos experimentos. Neste capítulo também apresentamos as qualidades principais do trabalho e descrevemos alguns pontos a serem melhorados para que o sistema possa a ser implantado em um ambiente real de produção.

Capítulo 2

Visualização de Documentos

A visualização de documentos de um acervo e as relações contidas entre eles têm sido objetos de muitos estudos nas áreas de biblioteconomia, ciência da informação e na área da computação. Cada tipo de visualização ou combinação de tipos de visualização pode nos levar a um entendimento melhor do conteúdo de um acervo.

Algumas ferramentas de visualização são baseadas em conteúdo textual dos documentos ou grupos de documentos fazendo uma análise simples de palavras chaves ou processamentos complexos de classificação para elaborar os gráficos da ferramenta[33][34]. Outras ferramentas focam a utilização das referências bibliográficas e citações para encontrar relações entre os documentos do acervo e gerar uma metáfora de visualização para eles[39].

Em outra linha, pesquisadores visam entender o conteúdo temático de um acervo, sem a preocupação de permitir o acesso a documentos individuais. Nestes casos, busca-se permitir a evolução da temática e a importância de cada tema durante a existência do acervo[37].

Em alguns modelos a ordem temporal é uma importante relação entre os documentos. Esta relação ajuda em algumas tarefas analíticas, visto que é uma dimensão natural para o ser humano e que pode facilmente ser interpretado graficamente. A interpretação desta relação nos permite identificarmos tendências e relações entre conteúdos publicados por autores contemporâneos ou não.

Neste capítulo, nós apresentaremos alguns modelos de representação espacial de documentos ou de assuntos relevantes de um acervo. Estes modelos tiveram um papel importante no entendimento das semânticas de representação de documentos e de agrupamentos de documentos apresentados nos capítulos seguintes.

2.1) NIRVE(*NIST Information Retrieval Visualization Engine)*

O Projeto *NIRVE(NIST Information Retrieval Visualization Engine)* [33][34]é uma iniciativa do *NIST(National Institute of Standards and Technology).* Sua proposta é permitir a

navegação pelo resultado e a manipulação de um conjunto documentos digitais resultantes de uma consulta inicial de um usuário através de uma interface tridimensional ou simplesmente numa interface típica em uma página HTML[34]. A consulta do usuário é um conjunto de palavras chaves que o usuário deseja recuperar associado a um conjunto de conceitos que o próprio NIRVE oferece ao usuário permitindo-o até mesmo configurar o peso destes conceitos através de um marcador em frente aos itens selecionados. A figura 1 nos mostra a tela utilizada para consulta pelo NIRVE.

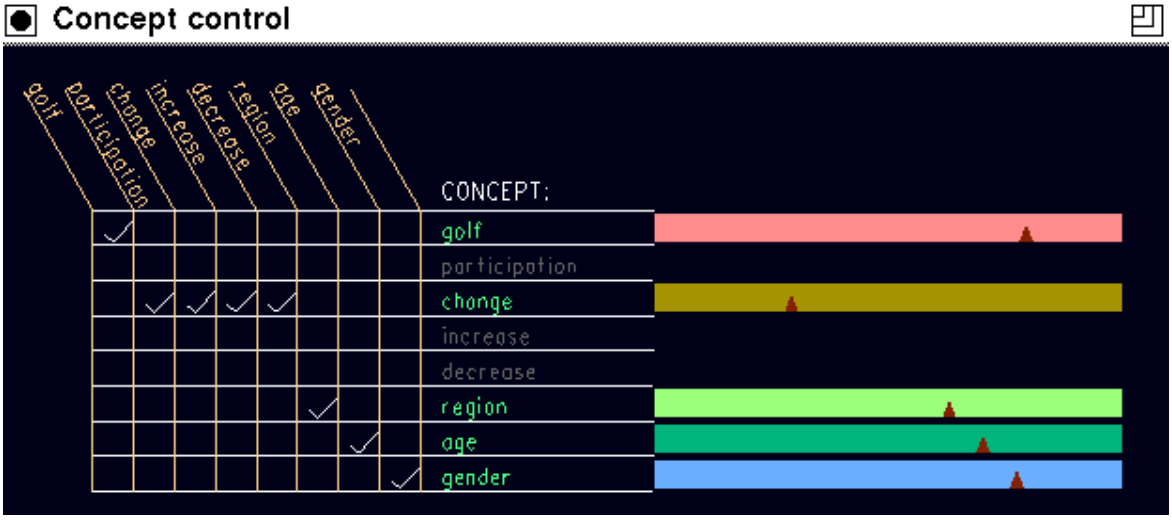


Figura 1 - Seleção dos conceitos no NIRVE

Podemos observar na figura 1, algumas palavras na horizontal. Na vertical temos um conjunto de conceitos. A grade formada permite ao usuário indicar a quais conceitos o termo que deseja consultar está relacionado. Na seqüência, o usuário indica o peso que este conceito deve ter nos documentos pesquisados. Esta indicação é feita arrastando a seta indicadora pela linha colorida formada. Então, de posse dos conceitos e seus pesos, o sistema realiza a busca que o usuário solicitou e então organiza os documentos em grupos baseados nestes conceitos. O sistema permite a pesquisa em nível de documentos e em nível de grupos de documentos[33] [35].

O processo de mapeamento de palavras chaves em conceitos visa melhorar a relevância semântica dos documentos buscados. Isto permiti a associação de palavras que possuem mesmos significados com o intuito de buscar mais documentos similares dentro do universo de documentos digitais.

O agrupamento de documentos realizado pelo NIRVE é baseado em alguns elementos: o tamanho do documento, o número de ocorrências da palavra-chave no documento e o mapeamento de palavra-chave com um conceito estabelecido[34]. Com estes três elementos surge então o conceito de *Força* de uma palavra chave para um documento baseado num valor normalizado (entre 0 e 1).Este valor é calculado como a raiz quadrada do número das ocorrências de cada palavra-chave dividido pelo tamanho original do documento[34][35]. Um documento é então mapeado para um vetor onde cada posição representa a *força* de uma palavra-chave do documento. Este vetor é chamado de *Profile de Conceito* que caracteriza o documento, sendo ele a principal fonte de informação semântica do documento. Este *profile* é interpretado como a posição do documento em um espaço *n*-dimensional onde n é o número de conceitos ativados pelas palavras chaves da consulta do usuário[33].

No NIRVE a visualização dos grupos ou dos documentos pode ser feita de algumas maneiras distintas. A seguir apresentamos os modelos utilizados no sistema.

2.1.1) Modelo em Espiral

Este primeiro modelo é exibido na figura 2. O modelo em espiral tenta preservar a estrutura seqüencial do ranking de documentos retornados pela consulta realizada. Os documentos no topo do ranking, ou seja, os mais significativos ficam posicionados mais ao centro da espiral. Os documentos são representados por ícones contendo um pequeno gráfico em seu interior indicando os três conceitos mais significativos de cada documento. No modelo em espiral existe também uma legenda de cores que indicam os fatores de pesos de cada palavra-chave da consulta sobre cada documento ou grupo de documentos[33]. Na figura 2 esboçamos o layout de como o NIRVE propõe o modelo em espiral.

Um dos grandes problemas deste modelo ocorre quando dois ou mais documentos possuem o mesmo grau de relevância. Isto causa uma sobreposição de ícones que impedem a identificação precisa dos documentos representados. Por conta deste problema o NIRVE desenvolveu outros modelos buscando uma melhor representação.

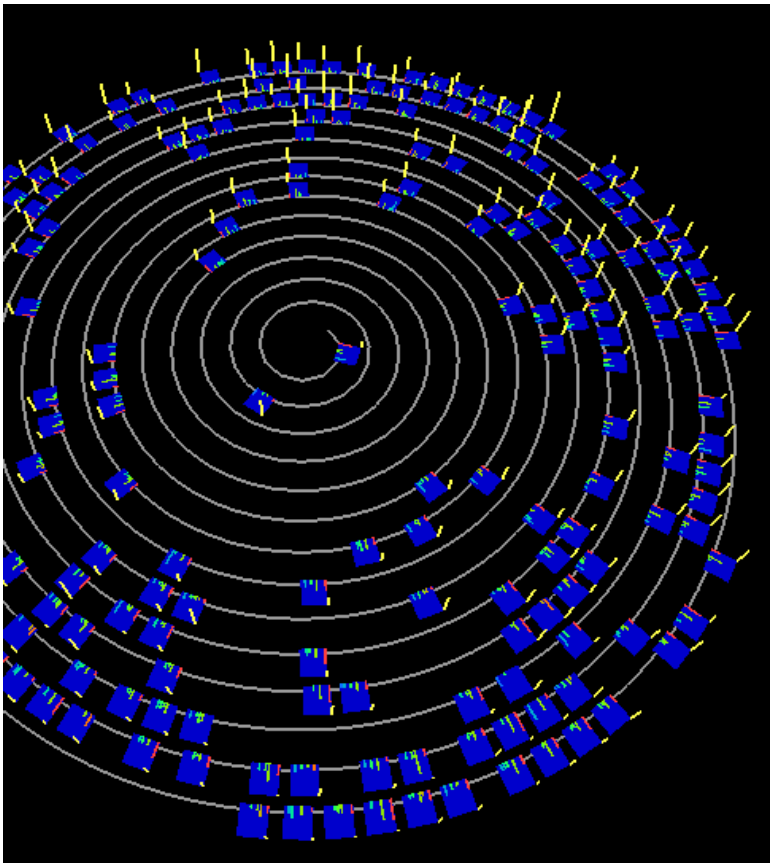


Figura 2-Representação de um modelo em Espiral

2.1.2) Modelo em Eixo 3D

O modelo em Eixo 3D foi desenvolvido na mesma época que o modelo espiral. A figura 3 exibe o modelo gerado pelo sistema. Nesta interface o usuário pode selecionar dinamicamente três palavras chaves nos modelo apresentado na figura 1 para serem apontadas como eixos X, Y e Z e que corresponderiam a três componentes dos *profiles* dos documentos. O modelo foi extendido posteriormente visando entender um conjunto de palavras chaves em cada eixo, visto que com apenas 3 palavras o sistema apresentava uma significativa limitação na semântica da consulta[33]. Cada ícone neste modelo, seguindo o modelo adotado pelo NIRVE, apresentava as colunas coloridas que representavam os conceitos mais importantes do documento e o peso de cada um.

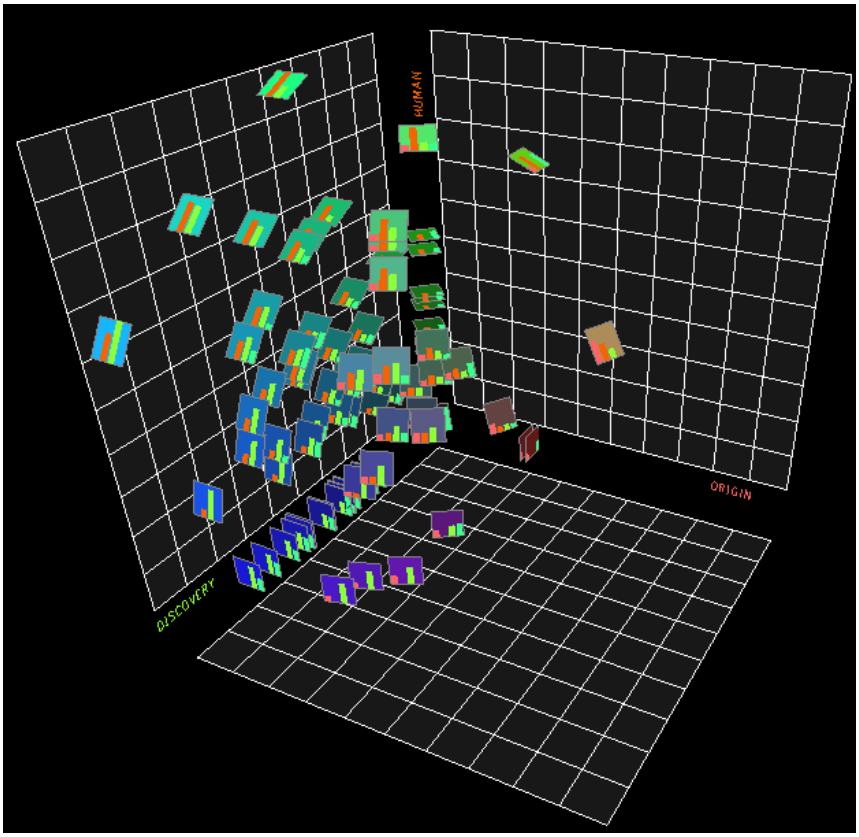


Figura 3- Modelo em 3D

Este modelo em Eixo 3D, apresentou deficiências devido a aglomeração de ícones em pontos específicos do gráfico 3D que prejudicavam a navegação. Muitos documentos ficavam escondidos devido a valores iguais a zero para algum dos eixos e até mesmo, valores negativos, permitidos pelas configurações do usuário. Estes valores negativos faziam com que documentos simplesmente sumissem do gráfico. Outra característica negativa deste modelo está no fato de que não existe uma identificação visual precisa do grau de importância de cada documento[33].

2.1.3) Modelo do Raio e da Roda

Neste modelo os ícones que representam documentos e grupos são dispostos em formato de círculo, conforme a figura 4. O mapeamento de palavras chaves em conceitos, conforme já vimos neste capítulo permite uma redução no universo semântico, permitindo o agrupamento de documentos que possuem palavras distintas, porém, com a mesma significância facilitando a visualização dos documentos. O NIRVE permite que esta

associação seja feita pelo próprio usuário que determina um peso para cada conceito. Estes pesos determinam o grau de importância do conceito na diferenciação de dois documentos.

Neste modelo existe a figura do grupo de documentos e do documento de maneira que cada documento e cada conjunto são posicionados de acordo com a distância lógica entre o perfil de cada documento.

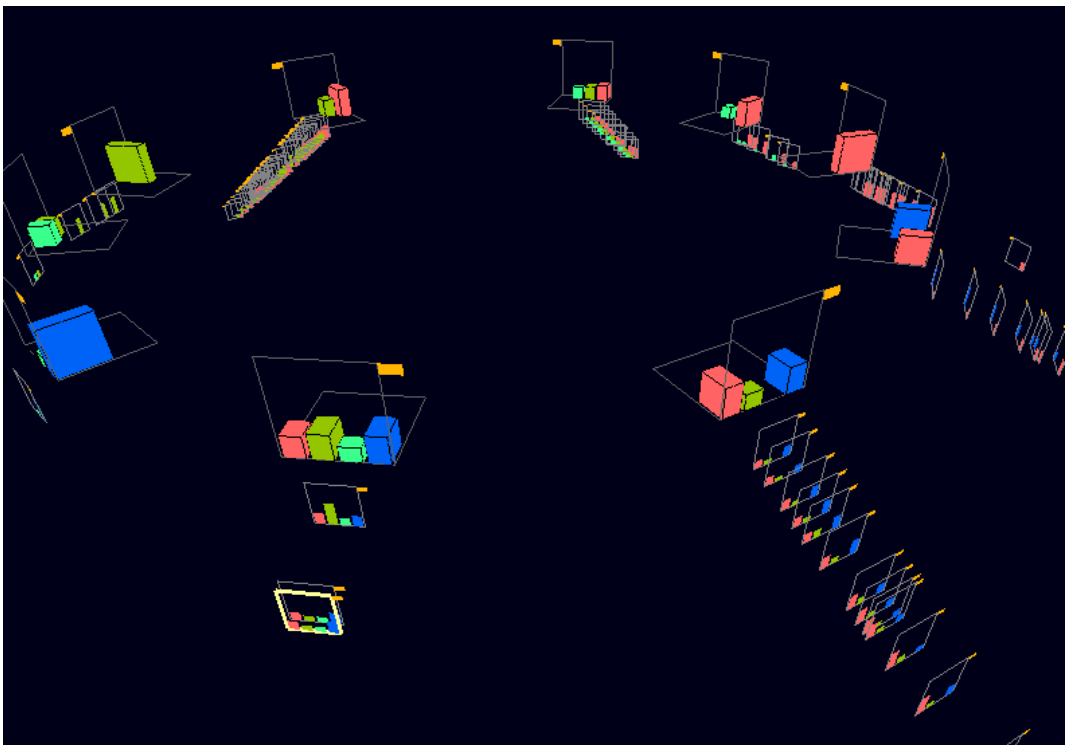


Figura 4- Modelo do raio e roda

Este modelo também representa o julgamento dos usuários com relação ao grupo ou ao documento, através de uma bandeira colorida que varia de vermelho(ruim) a verde(bom). Tal característica permite que seja realizado um filtro sobre o atributo para seleção de documentos desejados.

No modelo do Raio e Roda, exibido na figura 4, podemos ver os identificadores de grupos em disposição circular com suas colunas de assuntos mais significativos em gráfico de colunas identificados por cores. Atrás de cada identificador de grupo existe uma coluna de documentos que os compõem organizados em forma radial onde mais ao centro estão os documentos mais significativos para os conceitos pesquisados.

2.1.4) Modelo de Globo de Conceito

O modelo do Globo surgiu da necessidade de criar grupos de documentos que tivessem o mesmo conjunto de ocorrências de conceitos com valores no *profile* diferente de 0. Definindo os conjuntos de documentos, os mesmos foram agrupados sobre a superfície de um globo. O ícone do grupo é agora uma caixa cuja altura representa o número de documentos que contém, e cuja face mantém o gráfico de conceitos para identificar o perfil do conceito[33] como mostrado na figura 5.

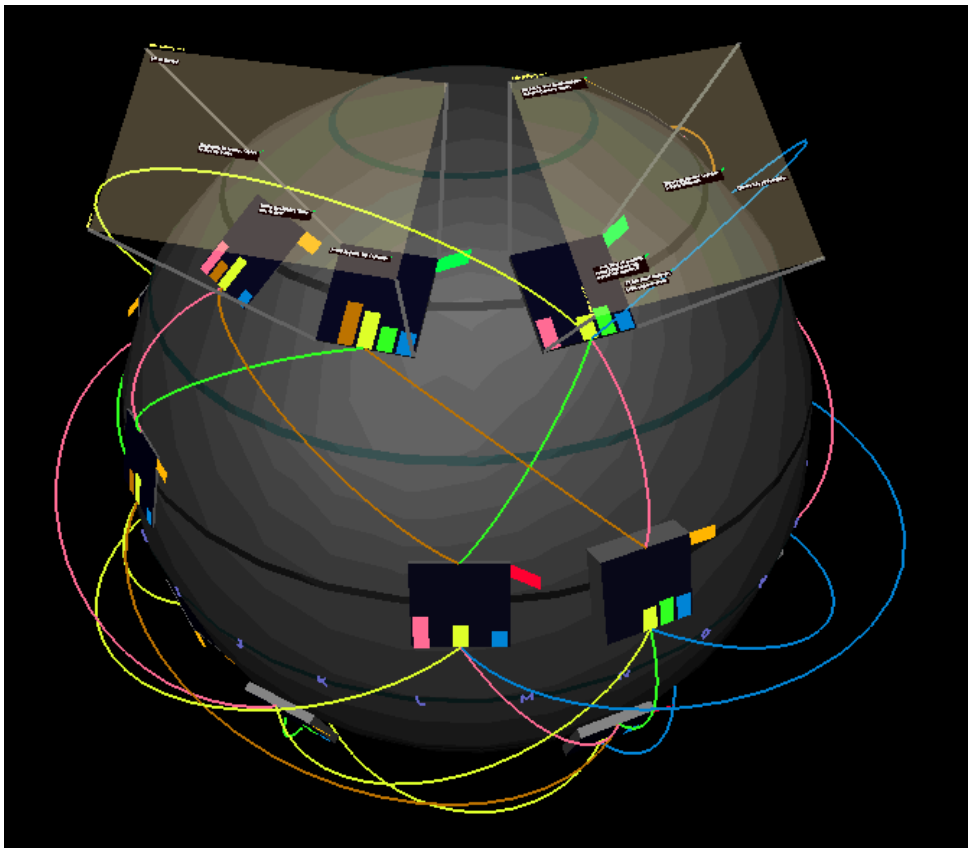


Figura 5- Modelo de Globo conceitual

No Globo de Conceito a exibição inicial nos traz apenas os grupos de documentos e apresenta as barras dos conceitos envolvidos no grupo. Os ícones que representam cada grupo ficam distribuídos em quatro hemisférios sobre o globo. Cada grupo está conceitualmente ligado a outros através de uma representação direta da ligação entre eles. O usuário poderá ver maiores detalhes do documento clicando sobre ele. Quando isto ocorre o sistema exibe um plano com os documentos que compõem este conjunto, permitindo com isto o acesso a cada um deles através do título do mesmo.

2.2) BiblioViz

O sistema BiblioViz é um sistema de visualização de bibliotecas que visa oferecer o máximo número de dados com um número mínimo de visualizações[36]. Para realizar a exibição de dados, o BiblioViz necessita de um conjunto de dados de entrada:

- Lista de artigos(ano, palavras chaves, áreas de pesquisa, autores e resumo)
- Lista de Palavra chaves(nome e área de pesquisa)
- Lista de autores(primeiro e último nome)
- Lista de áreas de pesquisa(nome)

De posse deste conjunto de dados de entrada o sistema apresenta os documentos através de dois modelos: a visão de Tabelas e a visão de Rede.

O modelo de visão de Tabelas é utilizado para visualização de informações temporais tais como documentos publicados num ano, autores que mais publicaram num período e outros conjuntos de dados. Estes conjuntos de dados são refinados conforme o usuário acessa os itens do gráficos apresentados.

Nos gráficos da visão de Tabelas cada retângulo representa um documento e cada cor representa um autor do artigo. Este modelo também pode ser utilizado para visualização dos documentos que compõem uma área de pesquisa ou os autores que a compõem num determinado período, sempre utilizando o eixo x como referência temporal do gráfico.

O modelo de Redes projetado no BiblioViz foi desenvolvido visando identificar as relações existentes entre os objetos do acervo, ou seja, autores, documentos, categorias. O que será representado neste modelo é escolhido pelo usuário para exibição no gráfico de rede. No modelo em rede, de acordo com a escolha do usuário, é montada uma rede onde cada nó representa um objeto do tipo escolhido. Cada ligação entre os nós representa um elo de ligação existente entre cada um deles obtido das informações do acervo. Podemos dizer, por exemplo, que se os nós da rede forem os autores, os documentos serão os elos de ligação. Se os documentos forem os nós, as áreas de pesquisa são as interligações. No software BiblioViz, ao clicarmos num nó, toda a sua rede é colocada em foco sobre as demais estruturas do gráfico, visando obter informações sem a influência das demais redes no plano de visualização.

No modelo em rede a variação de cor dos nós também nos permite representar informações. Quando, por exemplo, os nós correspondem a documentos a cor dos nós representa os autores de cada documento.

O BiblioViz é uma ferramenta que também permite a exibição dos dois modelos numa única tela de forma a permitir que um conjunto de informações possa complementar o outro para a identificação de um documento desejado.

2.3) ThemeScape

ThemeScape[38] é uma ferramenta de visualização de documentos que lê automaticamente os textos, reconhecendo o conteúdo da informação e criando grupos de documentos por tópico num mapa de grupos. Através de uma ferramenta de ampliação o mapa pode ser detalhado até o nível de documento para seu acesso.

O mapa de coleções ou grupos de documentos nos permite acessar documentos e enviá-los para outras pessoas. *Themescape* apresenta uma forma de organizar informação que não utiliza nem diretórios de arquivos, nem árvores hierárquicas. Ao invés disso, cada documento é representado como um pequeno ponto no mapa topológico. Com isto podemos dizer que quanto mais próximos estiverem dois pontos, maior será sua similaridade.

Este mapa de documentos é representado através de uma metáfora de mapas topográficos. Nestes mapas, as coleções com grandes quantidades de documentos formam picos de alta concentração pontos que representam cada documento. Assim como a distância entre cada ponto representa a similaridade entre dois documentos, a distância entre dois picos representa o grau de proximidade de duas áreas. Com esta informação visual o usuário pode intuitivamente conhecer informações como assuntos relacionados a uma coleção de documentos e correlação entre eles.

Como cada ponto do mapa corresponde a um documento, passando o mouse sobre um documento uma pequena descrição do documento é exibida simplificando o processo de busca realizado pelo usuário.

2.4) ThemeRiver

O *ThemeRiver*, ou metáfora do Rio, exibe as variações temáticas de uma coleção de documentos ao longo do tempo. As mudanças temáticas são apresentadas em um contexto temporal, ou seja, temas podem ter mais destaque em certos períodos que em outros.

Nesta metáfora de rio, cada tema discutido num ano é representado por uma corrente dentro de um rio. Esta corrente pode ser mais intensa numa altura(tempo) do rio que em outras.

O uso da metáfora do rio nos permite representar algumas informações importantes sobre o acervo:

- Evolução da coleção de documentos com o tempo;
- Evolução de um conteúdo temático selecionado no transcorrer do tempo;
- A influência de um tema selecionado em épocas diferentes;

A direção do fluxo da esquerda para a direita é interpretado como o movimento do tempo, ou seja, quanto mais a direita do gráfico mais recente será o assunto discutido. Cada corrente possui uma cor distinta representando um tema. A largura vertical de uma corrente indica a força de um tema num determinado momento. Obviamente, a ferramenta que provê esta visualização não mostra todos os assuntos possíveis como correntes no rio. Os assuntos exibidos são selecionados antecipadamente e então são analisados no período também informado.

Um dos efeitos possíveis é o desaparecimento temporário de uma corrente no rio. Isto ocorre quando num determinado período o tema selecionado não aparece em nenhum documento do acervo. No entanto, ao reaparecer, o tema retornar como uma corrente com a mesma cor que possuía anteriormente.

Outro efeito que podemos encontrar é o efeito de rio seco. Este efeito acontece quando todos os temas selecionados para serem exibidos não aparecem num determinado período.

A metáfora do Rio é muito útil quando desejamos mapear como um determinado evento pode influenciar a produção de documentos numa área. Uma análise inversa também é

permitida através deste modelo, ou seja, nos permite entender ou localizar quais eventos fizeram com que um assunto se tornasse visível em certa época.

2.5) Visualização Baseada em Citações e Semântica Latente

Em [6], Chen nos apresenta um conjunto de ferramentas integradas para permitir a visualização de um acervo digital utilizando a semântica do documento e as citações e referências bibliográficas existentes entre os documentos do acervo.

A elaboração do modelo de visualização utiliza como ferramenta de classificação o LSI(*Latent Semantic Indexing*). A matriz de similaridade documento-documento é submetida a uma técnica de análise de proximidade de dados utilizada na psicologia chamada *Pathfinder Network Scaling(PFNets)[39]*. Esta técnica simplifica uma representação de uma complexa rede em um modelo mais conciso e significativo contendo as ligações mais importantes da rede original.

O *PathFinder* trabalha com o que chama de triângulo de desigualdade. Este triângulo baseia-se no princípio de que se um documento A está associado a um documento B, e um documento B está associado a um documento C, então podemos inserir uma ligação direta entre A e C sem passarmos por B[39].

Uma vez determinadas as ligações mais importantes do PFNets sobre a matriz de similaridade obtida com o LSI, o gráfico em 3D é gerado. Cada documento é representado por uma esfera cujo o raio determina o tamanho do documento e a cor determina a fonte de dados ou o ano de publicação.

As ligações entre as esferas são representadas por cilindros cujos raios representam a similaridade semântica entre os documentos interligados.O comprimento do cilindro exibe a distância semântica entre os documentos.

O modelo de visualização proposto por Chen[6] ainda apresenta um outro cilindro perpendicular ao plano da rede. A altura deste cilindro apresenta o grau de relevância da palavra-chave pesquisada no documento. A cor do cilindro indica qual a palavra-chave representada pelo cilindro.

A estrutura do gráfico de Chen possui um anel central formado por documentos inter-relacionados. Neste anel, vários ramos estão associados pendurados em documentos deste anel. Cada ramo deste anel possui um conjunto documentos similares entre si.

No processo de indexação dos documentos, são cadastradas algumas informações importantes do documento como ano de publicação, autores citados, resumos e lista de palavras-chaves.

Com estas informações um segundo modelo é apresentado: o mapa de Co-citações de autores. O mapa de co-citações nos permite caracterizar o impacto das publicações de um pesquisador num campo de pesquisa assim como a relação existente entre pesquisadores de áreas, grupos de estudos, universidades e/ou culturas afins.

Os mapas de co-citações que são gerados são também baseados nos resultado da análise do *PathFinder Networks* levando-se em consideração para este mapa apenas as citações dos documentos com o objetivo de gerar as ligações e os nós. No modelo proposto por Chen[5][6], os autores são os nós do gráfico e as relações entre eles são determinadas pelas ligações entre estes nós. A área dos autores é definida através da análise dos textos dos artigos realizada com o uso do LSI. No entanto, as áreas apenas indicam as regiões onde estão posicionados os grupos de autores não compondo a malha de interligações da estrutura gerada pelo *Pathfinder*..

Este modelo, integrando ferramentas de análise semântica, com ferramentas de análise de citações nos oferece uma gama de informações sobre o acervo. Estas informações nos permitem não apenas identificar os documentos que o compõe, mas também identificar as relações existentes entre seus autores numa área de pesquisa com o passar do tempo. Conseguimos também visualizar as influências de cada trabalho sobre o universo de pesquisa num campo científico.

2.6) MVAB - Visualização de Galáxias

O projeto *MVAB(Multidimensional Visualization and Advanced Browsing)* visa explorar algumas técnicas de visualização de documentos a fim de permitir uma melhor análise da informação textual. Uma das metáforas exploradas é a de Galáxias.

A representação de galáxias exibe grupos de documentos inter-relacionados que são plotados num gráfico 2D como estrelas que aparecem no céu à noite. Os grupos são os pontos de aglomeração de estrelas(galáxias). Tal representação permite-nos identificar os documentos que possuem assuntos em comum. A proximidade entre galáxias nos leva a uma proximidade entre as áreas de interesse, ou grupamentos de documentos baseados no conceito de similaridade de documentos baseada no contexto e conteúdo dos mesmos.

Este modelo de representação espacial pode ser refinado utilizando marcadores temporais que permitem ao usuário refinar sua busca e reduzir o universo de pesquisa e a quantidade de estrelas no cluster.O usuário neste modelo poderá informar o ano, mês, dia horas e minutos criando janelas temporais de visualização de documentos nas suas respectivas galáxias.

Uma das vantagens deste modelo é a clara distribuição dos documentos por área ou contexto, formando grupos explícitos de documentos de mesma área de interesse que podem ser acessados diretamente.

2.7) Lighthouse

O *Lighthouse* [40] é um sistema que possui uma interface voltada para integrar uma lista de documentos ordenada por relevância através de uma visualização de *cluster*. Cada documento selecionado na lista é exibido como uma esfera no espaço. Cada esfera está distribuída com uma certa distância entre si baseada na similaridade entre dois documentos quaisquer. A figura 6 nos mostra um resultado de busca do sistema.

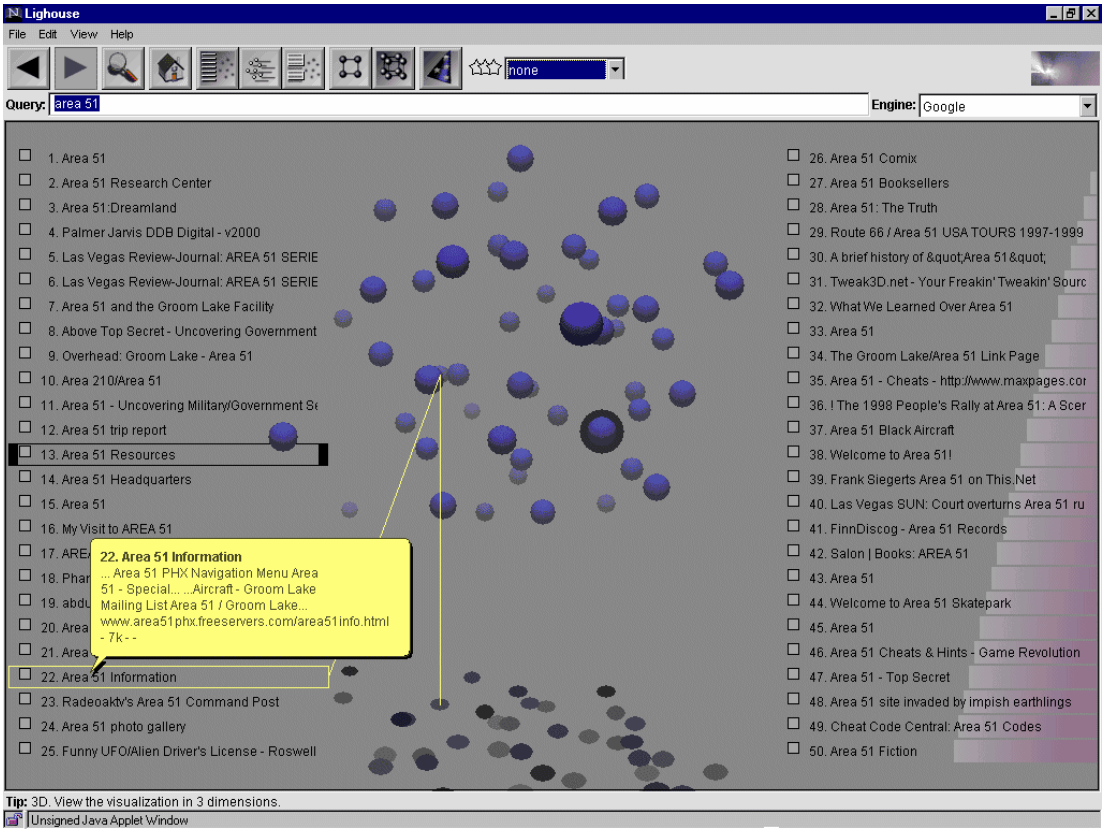


Figura 6-Tela do Ligthhouse

O funcionamento do sistema baseia-se em um mecanismo de busca que retorna os documentos relacionados a uma consulta. Estes documentos são listados baseados num grau de relevância e divididos em duas colunas, localizando-se entre estas colunas, as esferas representando cada documento.

O usuário pode definir o grau de relevância para um documento visitado e selecionar cores distintas para as esferas com o intuito de diferenciar as esferas utilizando este conceito.

Quando dois ou mais documentos(esferas) são pintados pelo usuário com tons de cores muito próximas, o usuário pode ter dificuldade de diferenciar os considerados mais relevantes. Para resolver este problema um segundo elemento de diferenciação foi inserido no modelo. O usuário poderá inserir estrelas ao lado das esferas. Quanto maior o número de estrelas ao lado do documento, mais significativo é o documento dentro dentre aqueles de mesma cor.

Quando o usuário passa o mouse sobre um documento, uma caixa de mensagem apresenta o título do documento, data de criação, grau de relevância fornecido pelo site de

busca utilizado, autores e endereço web relacionado, permitindo ao usuário localizar seu documentos conteúdo apenas navegando pelas esferas.

Apesar de sua aparência moderna da facilidade de navegação pelo ambiente o *LightHouse* nos oferece poucas informações contidas na imagem 3D, sendo a maior parte das informações inseridas via caixas de texto.

2.8) Ontoweb

O ONTOWEB[1] é um sistema de análise de informações na Internet, que possibilita uma pesquisa contextualizada nas fontes acessadas. É uma solução desenvolvida com a última geração das tecnologias digitais de tratamento textual. A tecnologia adotada permite a realização de consultas com grandes volumes de texto e destaca-se, na qual semântica e ontologias trabalham juntas para incrementar o processo de busca de informações relevantes em documentos digitais.

Apesar do uso de recursos como ontologias, o resultado da busca depende da existência de pelo menos um dos termos da pesquisa dentro de um documento a ser listado. Esta lista é então classificada e ordenada pela ferramenta utilizando seu mecanismo de ordenação[1]. Podemos ver um exemplo na figura 7.

Figura 7- Listagem de consulta de pelo termo "Presidente da República"

Na figura 7, vemos um resultado classificado. A ordem dos textos selecionados e a cor do ícone em frente ao título do texto indicam o grau de representatividade de cada documento

listado, ou seja, quanto mais ao topo, mais significativo é o documento na lista obtida. Vemos na figura 7, uma consulta pelos termos *"Presidente da República"*. Navegando pela listagem de 37932 documentos, observamos que todos os documentos apresentavam pelo menos um dos dois termos principais da consulta *"Presidente"* ou *"República"*.

Um diferencial desta ferramenta é o fato de que seu retorno é classificado de acordo com um critério de importância e representatividade de cada documento em relação aos termos consultados.

2.9) Conclusão

Neste capítulo, mostramos alguns modelos de visualização espacial de documentos digitais de um acervo. Podemos perceber que metáforas familiares ajudam o usuário a compreender mais facilmente os dados apresentados de forma intuitiva e diretamente ligada às habilidades cognitivas e perceptivas dos usuários. No entanto, cada modelo apresentado neste capítulo limita-se a apresentar os acervos com apenas um foco, ou seja, documentos, ou autores ou assunto. Nenhum destes modelos nos oferece um elo entre duas ou mais características.

Então, buscando criar uma metáfora familiar e que permita visualizar mais de uma característica do acervo. Desta forma desenvolvemos um modelo baseado no conceito de Sistema Planetário, que nos permite identificar um conjunto de relações entre os documentos de um acervo e seus autores utilizando um único modelo representativo.

Para desenvolvermos este modelo de visualização, alguns procedimentos tiveram que ser executados sobre o acervo visando fornecer as informações necessárias para a criação de uma metáfora representativa consistente. Estes procedimentos incluem indexação e classificação dos documentos digitais.

A classificação é um dos processos mais importante desta preparação do acervo. A escolha de uma boa ferramenta de classificação nos permite exibir grupos de documentos similares com alta ou baixa credibilidade. No próximo capítulo, trataremos exclusivamente de ferramentas e algoritmos de classificação, apresentando suas principais características e descrevendo seus funcionamentos.

Capítulo 3

Classificação de Documentos

Todos nós ao lermos um artigo, tese ou qualquer outro documento científico algumas vezes nos perguntamos: *Seria este documento uma cópia?* , ou talvez, *Quem escreveu algo parecido?* , *Quem são os autores mais importantes neste assunto?* , ou mesmo, *Em que classe de documentos eu coloco este novo documento que chegou às minhas mãos?* .

Com o grande aumento de publicações, sejam elas no mundo digital ou mesmo nos acervos de bibliotecas, estas dúvidas têm surgido com maior freqüência, visto a dificuldade, nas ferramentas de recuperação de documentos, de se obter algo pertinente aos nossos anseios de busca. O número crescente de publicações digitais tem nos oferecido novos algoritmos, metodologias e técnicas que nos permitem, particularmente no mundo digital, um melhor tratamento da verificação de similaridades de documentos, categorização e recuperação dos mesmos com mais rapidez e precisão.

No contexto dos documentos digitais, vários modelos para indexação e classificação foram propostos na literatura. Alguns modelos propostos tornaram-se clássicos com o passar dos anos como Booleano, Vetorial e Probabilísticos. Existem também alguns modelos dinâmicos como Redes Bayesianas [23]; *Latent Semantic Indexing* (LSI) [12] e Redes Neurais Artificiais [25], dentre outras técnicas de tratamento da informação em meio digital.

Neste capítulo, veremos como funcionam alguns destes modelos existentes.

3.1) Modelo Booleano

O modelo Booleano[23] é um dos modelos mais utilizados, na recuperação de informação. Este modelo baseia-se na teoria dos conjuntos e na álgebra de *Boole*. A álgebra de Boole é baseada na utilização de apenas 2 valores: 1 ou 0, Verdadeiro ou Falso.

As consultas utilizando o modelo *Booleano* são elaboradas através de expressões de busca que combinam termos de indexação e operadores booleanos (*and*, *or* e/ou *not*).

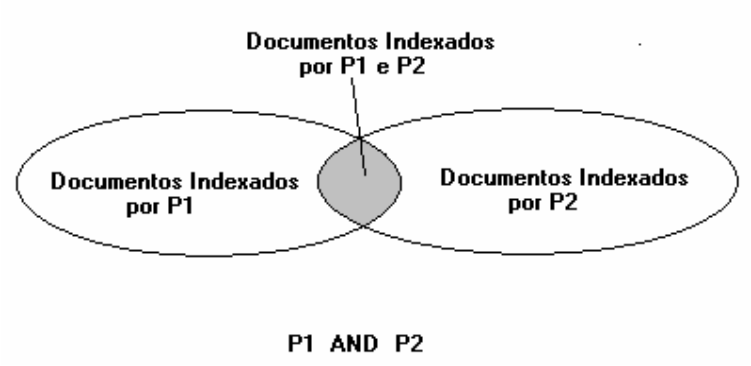


Figura 8 - Exemplo de Busca do modelo Booleano

Na figura 8, podemos identificar a área hachurada como sendo a área de interesse resultado da consulta de *P1 AND P2*, ou seja, P1(palavra 1) e P2(palavra 2). No exemplo da figura 8, supomos que o usuário fez uma consulta utilizando os termos P1 e P2. Com uma indexação utilizando o sistema booleano, temos um conjunto de documentos cujo termo P1 fez parte da indexação, enquanto noutro conjunto documentos o termo P2 esteve no conjunto termos indexados. Como a consulta do usuário pesquisa apenas documentos que possuam os termos P1 e P2, temos um pequeno conjunto de documentos cuja indexação possui ambos os termos. Este conjunto corresponde a área hachurada no gráfico da figura 8.

Na figura 9, apresentamos um modelo real de consulta mais elaborado utlizando o modelo booleno de consulta. Nesta figura apresentamos uma consulta de documentos que possuam os termos (*Recuperação* e *Informação*) ou *Visualização* de acordo com sentença lógica. Com isto, a consulta a um acervo retornaria os documentos localizados nas áreas hachuradas do gráfico.

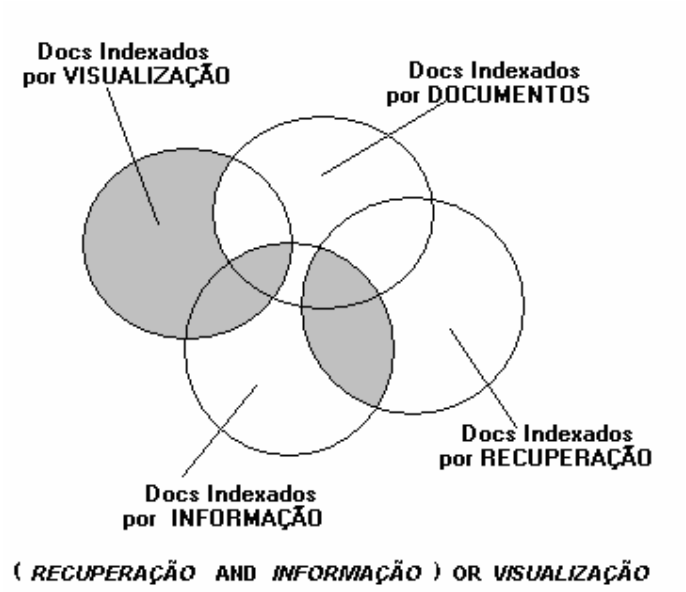


Figura 9 - Modelo Booleano. Nas áreas cinza encontram-se os documentos desejados

Apesar de muito utilizado, graças à sua simplicidade, o uso deste modelo nos apresenta algumas deficiências como[26]:

- Nenhuma ordenação de documentos é fornecida, sendo retornado apenas um conjunto de documentos que atendem à consulta e os que não atendem, sem que haja uma informação de quão relevante é um documento à consulta;
- A passagem da necessidade de informação do usuário à expressão booleana é considerada complicada, uma vez que o usuário necessita conhecer as regras e notações da lógica como “E” e “OU” para a montagem das suas consultas.

3.2) Modelo Vetorial

O modelo Vetorial[23] tem como objetivo representar um documento como um vetor multidimensional onde cada posição do vetor corresponde ao peso de um termo (palavra do texto) no documento e a um eixo no espaço. Para montarmos cada um dos vetores para representação do documento é preciso que um processo de indexação faça um filtro das palavras que não serão significativas em um documento. Estas palavras não significativas são denominadas *Stop-words.* Apesar de não serem significativas para o documento, sua presença pode causar distorção nas comparações de documentos, uma vez que as *stop-words* normalmente aparecem em grande quantidade nos textos, fazendo com que documentos

distintos sejam classificados como semelhantes. No idioma português podemos chamar de *stop-words* artigos(a,as,os,os, um,...), pronomes(eu, seu, meu, tu...) , advérbios(ali, aqui, acolá, ...), preposições, dentre outros. Com isto, temos um conjunto muito grande de palavras que devem ser retiradas do cálculo de pesos que são armazenados no vetor. O peso de um termo *i* num documento *d* (W*id)* no vetor é dado por:

$$W_{id} = freq(t_i,d) \text{ x } idf_i ,$$

onde *freq(ti, d)* é o número de ocorrências de um termo *i* no documento *d* e *idfi(inverse document frequency)*[27] pode ser calculado por :

$$Idf\ i = log\ (\ N / ni),$$

Nesta equação, *N* representa o número de documentos do acervo e *ni* representa o número de documentos que possuem o termo i.

A figura 10 nos mostra um esboço deste vetor para dois documentos *D1* e *D2* com 3 termos cada um. O vetor do documento *D1* apresentado é formado pelos termos *t1*,*t2* e *t3*. O gráfico nos indica os pesos de cada um destes termos para o documento como se cada termo fosse um eixo do vetor. Estes valores são exibidos na coluna de valores *D1*. O mesmo acontece para o vetor *D2*. Com isto quanto mais próximo estão dois vetores(documentos) podemos dizer que os dois documento são mais parecidos, uma vez que os termos tendem a ter o mesmo peso.

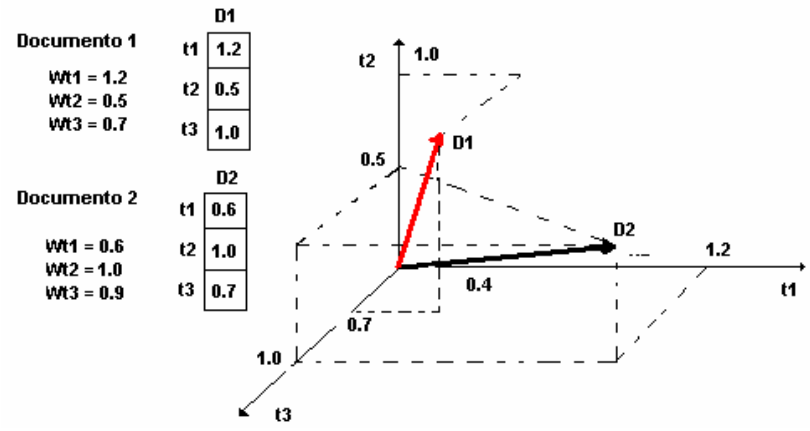


Figura 10- Representação Vetorial de 2 documentos de três termos

Como um documento normalmente possui uma quantidade muito grande de termos, fica impossível representarmos visualmente em um gráfico cartesiano um documento real, como na figura 10. No entanto podemos representar um conjunto de documentos como uma matriz bidimensional onde as linhas são documentos e as colunas os pesos dos termos dentro do documento. A consulta do usuário, neste caso, deve ser incorporada à matriz como se fosse um documento que pode ser comparado com todos os demais documentos da matriz.

	t1	t2	t3	t4	t5		tj
D1	0.5	0.7	1.3	5.2	3.1	...	3.0
D2	0.7	0	0.3	4.8	2.4	...	1.5
D3	1.0	0.1	0.2	0	2.5	...	2.5
.							
.							
.							
Di	0	0.6	1.2	3.0	2.2	...	0

Tabela 1-Documentos X W(i,j)

A matriz de pesos resultante é também utilizada em outros modelos de cálculo de similaridade de documentos mais complexos conforme veremos adiante.

Inserindo uma consulta *q* na matriz da mesma maneira como fazemos com um documento, ou seja, calculando os pesos dos termos podemos obter o grau de similaridade(*sim*) do documento *d* com esta consulta *q* utilizando a equação de similaridade provida pela notação vetorial onde:

$$sim(d,q) = \frac{\sum_{i=1}^{t} w_{id} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{id}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

Equação 1- Equação de cálculo de similaridade entre dois documentos

Esta equação pode também ser utilizada para calcularmos o grau de similaridade entre dois documentos. Este grau de similaridade nos permite agrupar todos os documentos que possuem similaridade dentro de um limite desejado, podendo com isto gerar o conceito de categorias.

Diferente do modelo booleano, o uso desta técnica vetorial nos oferece um resultado que pode ser ordenado pelo grau de similaridade entre os documentos, permitindo-nos

restringir o número de documentos da resposta de acordo com a nossa exigência de similaridade. Outra vantagem sobre o modelo booleano é que a consulta pode ser feita em sem o uso de operadores lógicos(e,ou, etc.).

Baseado no modelo conceitual vetorial[23], Gerard Salton desenvolveu o projeto SMART(*System for the Manipulation and Retrieval of Text)*[28] que teve início em 1961 e que até hoje é uma referência no desenvolvimento de sistemas de recuperação de informação.

3.3) Modelo Probabilístico

A principal ferramenta matemática do modelo probabilístico é o teorema de Bayes. O teorema de Bayes é usado na inferência estatística para atualizar estimativas da probabilidade de que diferentes hipóteses sejam verdadeiras, baseado nas observações e no conhecimento de como essas observações se relacionam com as hipóteses[29].

A probabilidade direta de uma hipótese chamada (*H)* condicionada à um corpo de dados chamado (*E),* nos produz a probabilidade *P(H|E)* está relacionada ao inverso da probabilidade dos dados *E* condicionados à hipótese *H*, *P(E|H)* .

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Equação 2 - Equação de Bayes

Veja o exemplo abaixo:

	Bebeu antes	Não Bebeu	
Sofreu acidente de carro	40	8	48
Não sofreu acidente de carro	10	60	70
	50	68	118

Tabela 2-Entrevista de pessoas que sofreram e não sofreram acidentes

Na tabela 2 temos um exemplo de dois grupos de pessoas entrevistadas. No primeiro grupo de pessoas temos 50 beberam e dirigiram. Destas 50 pessoas, 40 sofreram acidente. Num outro grupo de 70 pessoas que dirigiram e não sofreram acidentes, apenas 10 haviam bebido. Logo uma das perguntas que podemos responder com o teorema de Bayes é: qual a

probabilidade de que ao beber e dirigir eu sofra um acidente? Façamos a seguinte consideração:

E: Pessoas que sofreram acidentes

H : Pessoas que beberam

Então o resultado que nos interessa é P(H/E). Utilizando o cálculo do teorema de Bayes devemos antes calcular o percentual dos sofreram acidente e beberam P(E/H).

P(E/H) = P(E e H) / P(H) = (40/118)/(50/118) = 0,8 ou 80%, ou seja 80% das pessoas que beberam sofreram acidentes no universo entrevistado. Logo, para saber o percentual de chance de alguém que beber sofrer acidente é calculado da seguinte maneira:

$$P(H/E) = P(E/H) \times P(H)/P(E) = 0,8 \times (50/118) / (48/118) = 0,83333 ,$$

ou seja, existe cerca de 83,3% de chance de uma pessoa que bebeu sofrer acidente de carro.

O modelo de classificação probabilístico é baseado no princípio probabilístico de ordenação (*Probability Ranking Principle*), que estabelece que este modelo pode ser usado de forma ótima. Este princípio é baseado na hipótese de que a relevância de um documento para uma determinada consulta é independente de outros documentos. O princípio é o seguinte: “Se a resposta de um sistema de recuperação de referência a cada requisição, é uma ordem de documentos classificada de forma decrescente pela probabilidade de relevância para o usuário que submeteu a requisição, onde as probabilidades são estimadas com a melhor precisão com base nos dados disponíveis, então a efetividade geral do sistema para o seu usuário, será a melhor que pode ser obtida com base naqueles dados”[27]. Desta forma, como teoricamente temos um conjunto ideal de documentos vejamos como obter este conjunto:

1) Um conjunto inicial de documentos é recuperado. Para realização deste processo o conjunto ideal é modelado em termos probabilísticos, onde são dados uma consulta *q* e um conjunto de documentos *dj*. Com isto estima-se a probabilidade que o usuário considere o documento *dj* interessante, isto é, relevante. O modelo assume que a probabilidade de relevância depende das representações da consulta *q* e dos documentos *dj*. Um documento *dj* e uma consulta *q* são modelados de forma similar ao modelo vetorial, porém os valores possíveis dos pesos *w* são apenas 0s e 1s. A figura 11 ilustra

este exemplo. Nela um documento *D1* que contem os temos *t1,t2* e *t3* apresenta um vetor com os pesos de cada um dos temos, segundo o modelo probabilístico.

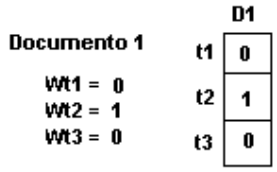


Figura 11- Vetor do documento

2) O conjunto resposta ideal é denotado por $\underline{R}$ e deve maximizar a probabilidade de relevância, e conter os documentos previstos como relevantes. O cálculo que nos permite identificar a relevância de um documento é :

$$sim(q,dj) = P(dj \text{ relevante-para } q) / P(dj \text{ não-relevante-para } q)$$

Por definição temos que :

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

$P(R \mid dj)$: probabilidade que o documento seja relevante;

$P(\sim R \mid dj)$: probabilidade que o documento seja não relevante.

Desta forma temos que a relevância é dada por:

$$\begin{aligned} sim(q,\underline{dj}) &= P(R \mid \underline{dj}) / P(\sim R \mid \underline{dj}) \\ &= \frac{[P(dj \mid R) * P(R)]}{[P(\underline{dj} \mid \sim R) * P(\sim R)]} \\ &\sim \frac{P(dj \mid R)}{P(\underline{dj} \mid \sim R)} \end{aligned}$$

Onde $P(\underline{dj} \mid R)$ é probabilidade de selecionar randomicamente o documento dj do conjunto R de documentos relevantes. De posse de um termo ti um índice qualquer temos:

$$\begin{aligned} sim(q,\underline{dj}) &= [(\Pi\, g(dj)=1\ P(ti|R)) * (\Pi\, g(dj)=0\ P(\sim ti|R))] / \\ &\quad [(\Pi\, g(dj)=1\ P(ti|\sim R)) * (\Pi\, g(dj)=0\ P(\sim ti|\sim R))] \end{aligned}$$

$P(ti|R)$: probabilidade do índice ti pertencer a um $d \in R$;

$P(\sim ti|R)$: probabilidade do índice ti não estar presente em $d \in R$.

$P(ti|\sim R)$, $P(\sim ti|\sim R)$: idem para $d \in R$.

Usando *log* e considerando $P(ti|R)+P(\sim ti|R)=1$:

$$sim(q,dj) \quad \sim= \sum_{i=1,T} wiq * wij * \{\log [P(ti|R)/(1-P(ti|R))] + \log [(1-P(ti|\sim R)) / P(ti|\sim R)]\}$$

Inicialmente $P(ti|R) = 0,5$ e $P(ti|\sim R)= ni / N$

ni : número de documentos. que contém *ti* ;

N : número total de documentos.

Se *V* é conjunto de documentos inicialmente recuperados e *Vi* é o subconjunto de *V* que contém *ti* :

$P(ti|R) = Vi / V$; e

$P(ti|\sim R)= (ni - Vi) / (N - V)$.

3) De posse de um conjunto inicial dado como resultado, o usuário seleciona aqueles documentos que considera mais importantes.O sistema aprende com as indicações do usuário, ou seja, o modelo incorpora explicitamente o conceito *Relevance Feedback* [28], visto que, é o usuário o responsável por indicar a relevância de um documento num conjunto. Para cada conjunto selecionado pelo usuário, o processo se repete refinando o resultado e ordenando-o automaticamente pela similaridade calculada.

Alguns pontos tornam os resultados do modelo probabilístico questionáveis. Dentre elas podemos citar[26][27][28]:

- O fato de os pesos não levarem em consideração as freqüências dos termos e sim um valor binário;
- Ignora a filtragem de informação.

3.4) Modelo de Redes Neurais em Recuperação da Informação

O modelo de redes Neurais é um modelo dinâmico de recuperação de informação. Uma de suas características é reconhecer a importância do usuário no processo de classificação do documento. Isto permite ao usuário adaptações na representação do documento e na relevância do mesmo.

Num processo de recuperação de informação identificamos três estruturas básicas: as *expressões de busca*, os *documentos* e os *termos de indexação*. Esta estrutura pode ser mapeada numa rede neural onde teríamos uma camada de entrada (as expressões de busca), uma camada de saída(os documentos) e uma camada central formada pelos termos de indexação.

As três camadas interagem tendo inicio com os termos de busca que disparam o processo de inferência dos respectivos termos de indexação. Os termos da expressão de busca que não fizeram parte do conjunto de termos indexados são automaticamente descartados, pois não ativam nenhum termo de indexação. Os termos de indexação ativados disparam sinais para os documentos que são multiplicados pelos pesos dos termos indexados. Com isto os documentos ativados enviam novamente mensagens aos termos de indexação, que novamente enviam aos documentos. Este processo se repete até que o sinal torne-se fraco o suficiente para suspender a propagação, ou seja, todos os termos de um documento tenham sido visitados. Podemos entender melhor o funcionamento através da seqüência de passos indicada à frente.

Após o processo, os documentos ativados pelos termos de busca e pelos termos de indexação são listados pelo grau de ativação destes documentos, ou como também podemos chamar, grau de relevância. Nesta lista, alguns documentos que não possuem os termos pesquisados podem ser listados, porém foram inferidos durante o processo de busca e podem possuir um certo grau de relacionamento com o que o usuário deseja. Resumindo o processo temos a seqüência de passos:

1. Os termos de busca ativam os termos de indexação correspondentes;
2. Os termos ativam documentos aos quais estão ligados;
3. Os documentos ativam novos termos de indexação;
4. Estes novos termos ativam novos documentos ou reforçam a importância de outros já ativados;
5. Estes novos documentos respondem para seus termos de indexação, fortalecendo os termos já visitados.

6. O processo se repete até que nenhum novo documento seja mais ativado.

7. Ao final do processo temos um conjunto de documentos ordenados pelo grau de ativação do mesmo no processo, ou seja, documentos que foram ativados poucas vezes são menos relevantes que documentos muito ativados.

3.5) *Latent Semantic Indexing* (LSI)

Quando utilizamos os modelos clássicos de recuperação de informação podemos obter um conjunto de documentos com uma qualidade muito baixa. Isto significa que documentos não relacionados ao assunto podem estar incluídos no conjunto resposta e documentos relevantes que não contém nenhum termo da consulta foram deixados de fora. Podemos dizer que o processo de recuperação baseado simplesmente na indexação de termo é vaga e com distorções e não leva em conta a semântica do documento.

Quando tratamos os documentos apenas pelas palavras chaves ou termos indexados, encontramos dois problemas a serem resolvidos a *polissemia* e a *sinonímia*. A polissemia é a propriedade que determina que uma mesma palavra pode apresentar vários significados, por outro lado a sinonímia é a ocorrência de várias com o mesmo significado. Quando trabalhamos apenas com palavras chaves, ignoramos estas duas possibilidades.

O LSI(*Latent Semantic Indexing*)[12] visa tratar a necessidade de informação do usuário de acordo com os conceitos e idéias e não somente baseado em índices. Isto implica em dizer que uma expressão de busca pode ter como resultado um documento que apresenta a mesma idéia retornada, mas não possui nenhum termo da expressão digitada pelo usuário. Então, se dois documentos A e B não têm palavras em comum, mas contém várias palavras em comum com um documento C, então A e B podem ser considerados similares.

O processo de identificação da similaridade tem início com a montagem da matriz de pesos e termos de forma similar à matriz do modelo vetorial. A partir desta matriz, que podemos chamar de um conjunto de vetores de documentos, o método realiza uma modificação no espaço vetorial de forma a evidenciar as relações existentes entre as palavras de um documento e as relações existentes entre documentos.

Uma vez montada a tabela ou matriz de *Documentos X Termos* devemos utilizá-la como entrada para a técnica de Decomposição de Valores Singulares(DVS)[32]. O DVS é uma técnica utilizada em muitas áreas de pesquisa, por exemplo, em processamento de imagens[31], recuperação de informação[12], geofísica[30], dentre outros.

Com o uso do DVS, decompomos a matriz em três novas matrizes de acordo com o teorema de Decomposição de Valores Singulares[32]. Suponhamos que X seja nossa matriz *Documentos X Termos* com *t* linhas e *d* colunas, então teremos :

$$X = US\,V^T ,$$

(U) é a matriz dos autovetores derivada de $(X)(X)^t$

$(V)^t$ é a matriz dos autovetores derivados de $(X)^t(X)$

(S) é a *r x r* matriz diagonal dos valores singulares, onde r é o posto da matriz X <= min(t,d).

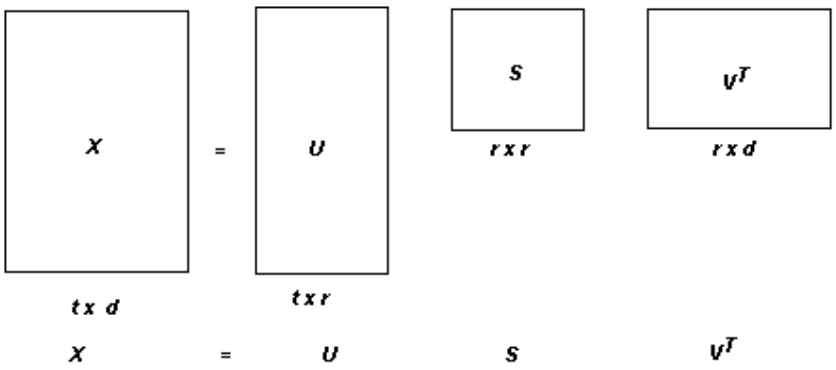


Figura 12 - Decomposição DVS

A figura 12 nos mostra um esboço da divisão realizada pelo método DVS. Faremos agora um exemplo um exemplo real de redução de uma matriz utilizando o DVS.

	Documentos				
	D1	D2	D3	D4	D5
Palavra 1	1	2	1	0	3
Palavra 2	4	0	3	0	4
Palavra 3	4	2	0	4	2
Palavra 4	3	0	2	2	0
Palavra 5	3	5	9	3	4
Palavra 6	6	4	3	0	8

Tabela 3-Matriz Documento X Termo original(X)

O processo de decomposição DVS nos gera as seguintes matrizes:

-0.198	-0.139	-0.207	-0.343	0.311
-0.333	-0.218	-0.078	0.714	0.568
-0.265	-0.178	0.809	-0.350	0.313
-0.166	0.130	0.472	0.467	-0.524
-0.622	0.752	-0.131	-0.119	0.046
-0.605	-0.565	-0.239	-0.133	-0.454

Tabela 4 - Matriz U

17.362	0.000	0.000	0.000	0.000
0.000	6.377	0.000	0.000	0.000
0.000	0.000	4.982	0.000	0.000
0.000	0.000	0.000	3.257	0.000
0.000	0.000	0.000	0.000	0.695

Tabela 5 -Matriz S0

-0.495	-0.372	-0.515	-0.188	-0.563
-0.387	0.136	0.712	0.283	-0.495
0.464	-0.082	-0.279	0.760	-0.351
0.417	-0.772	0.387	-0.253	-0.126
-0.466	-0.491	0.025	0.492	0.546

Tabela 6 -Matriz Vt

Assim, como propõe Deerwester[12], esta decomposição nos permite reduzir as dimensões das matrizes resultantes de forma tal que, possamos remontar a matriz *Documentos X Termos* eliminando detalhes insignificantes da matriz original e acentuando a importância de determinadas referências e citações.

A redução das dimensões das matrizes *U, S* e V^T é feita da seguinte forma:

- Na matriz diagonal S de dimensões *r x r*, permanecem somente os *K* maiores valores singulares, formando uma matriz S_k.
- Na matriz U são mantidas as *K* primeiras colunas e na matriz V^T são mantidas as *K* primeiras linhas. Formando as matrizes U_k e $V^T_{k.}$
- A nova matriz *Documentos X Termos*, aqui chamada de X_k é dada por:

$$X_k = U_k S_k V^T_{k,}$$

Onde *K* é a dimensionalidade do espaço conceitual sendo K < r. A escolha do valor de *K* deve ser grande o suficiente para permitir o transporte das informações significativas durante a redução de ordem e pequeno o suficiente para eliminar detalhes irrelevantes contidos na matriz original. Veja a figura abaixo:

Com o uso de um *k* adequado ao universo de documentos trabalhado, a redução das dimensões das matrizes originais decompostas e posterior remontagem da matriz *Documentos X Termos* nos permite identificar a importância de alguns termos para documentos em que eles não aparecem. Utilizando a matriz de exemplo da tabela 3 e suas matrizes de decomposição, faremos a redução para um valor escolhido de *K= 2.*

-0.198	-0.139
-0.333	-0.218
-0.265	-0.178
-0.166	0.130
-0.622	0.752
-0.605	-0.565

Tabela 7 -Matriz T reduzida a t x K

17.362	0.000
0.000	6.377

Tabela 8 - Matriz diagonal reduzida a K x K

-0.494	-0.372	-0.515	-0.188	-0.563
-0.387	0.136	0.712	0.283	-0.495

Tabela 9 - Matriz Vt reduzida a Kx d

Após reduzir as matrizes, refazemos a multiplicação das mesmas de forma a obter uma a matriz original. Veja a tabela 10.

2.045	1.159	1.140	0.394	2.379
3.392	1.958	1.982	0.689	3.942
2.710	1.555	1.561	0.542	3.151
1.106	1.186	2.077	0.777	1.216
3.479	4.662	8.968	3.382	3.705
6.584	3.415	2.843	0.950	7.701

Tabela 10 - Matriz X remontada a partir da redução dimensional de U, S_0 e V^T

Podemos observar que todas as colunas que possuíam o valor 0 agora possuem valores. Estes valores correspondem à influência de outros termos sobre a composição do documento observado. Vejamos a "*Palavra 2"* no documento *D2*. Ela aparece como 0 na matriz inicial. No entanto, esta palavra aparece em outros documentos ao redor do documento *D2*. O resultado da matriz X após a redução dimensional nos mostra que D2 possui um peso de 1.9575 e não mais de 0.Como *D2* possui outras palavras que são comuns a outros documentos e estes outros documentos possuem a palavra *D2*, podemos dizer que a *Palavra 2* possui um grau de relevância para documento *D2* por ser, por exemplo, sinônimo de uma palavra que aparece em *D2*.

Por outro lado, algumas palavras terão seus pesos reduzidos após a redução da matriz X. Tal redução pode ser explicada pelo efeito provocado pela *Polisemia*, ou seja, uma palavra aparece *n* vezes num certo contexto, e aparece outras *m* vezes em outro contexto. Embora as palavras sejam escritas da mesma maneira, as demais palavras que a circundam não dão a ela o mesmo significado. Isto implica dizer que se dois documentos têm uma palavra em comum e as demais diferentes, a palavra em questão pode apresentar significados diferentes em cada um dos documentos.

Utilizando os dois raciocínios apresentados nos parágrafos anteriores o *LSI* realiza o tratamento da *polisemia* e *sinonímia* que tanto dificulta a recuperação de documentos apenas baseando-se em palavras chaves.

Uma vez obtida a matriz *X* reduzida, que podemos considerar um conjunto de vetores de documentos, realizamos o cálculo de similaridade entre estes documentos, ou vetores, utilizando a mesma fórmula apresentada no modelo Vetorial, onde *d* e *q* são um documento e uma consulta ou a simples comparação entre os vetores de dois documento.

O resultado da Equação 1 nos permite comparar todos os documentos um a um. Este resultado é que nos permite tomar decisões de como lidar com um determinado documento considerando-o ou não um elemento importante para nossa consulta de acordo com um valor referência que pode ser configurado pelo usuário.

3.6) Conclusão

Neste capítulo exibimos alguns modelos clássicos de recuperação de informação de informação. Estes modelos nos oferecem condições para classificarmos os documentos que serão exibidos através dos diversos modelos de visualização.

O modelo Vetorial, o modelo Booleano e o modelo Probabilístico têm como objetivo representar um documento como um vetor multidimensional onde cada posição do vetor corresponde a um valor representativo dos termos existentes. Uma busca realizada nestes modelos depende da existência do termo consultado no vetor *Documentos X Termos.* Desta forma podemos dizer que estes modelos não nos oferecem uma análise semântica do conteúdo.

O LSI e o modelo em Redes Neurais visam tratar a necessidade de informação do usuário de acordo com os conceitos e idéias e não somente baseado em índices. Isto implica em dizer que uma expressão de busca pode ter como resultado um documento que apresenta a mesma idéia retornada, mas não possui nenhum termo da expressão digitada pelo usuário.

Visto os modelos de identificação de similaridades, o próximo capítulo apresentará o modelo gráfico de visualização proposto. Este modelo pressupõe a existência de um mecanismo de classificação que o permita exibir os dados de modo que o que está sendo mostrado se aproxime ao máximo da classificação real dos grupos.

Capítulo 4

MoGRIS - Modelo Gráfico de Recuperação de Informação por Semântica

4.1) Introdução

No capítulo 2, apresentamos diversas metodologias de visualização de acervos de documentos. Algumas destas metodologias trabalhavam com mecanismos convencionais baseados em *links* textuais. Esta forma de representação e acesso a documentos é a que, a maioria de nós, está familiarizado, devido a nossas "navegações" por ambientes *Web* e pelo conceito de interface de software que prevalece na grande maioria das aplicações de hoje.

Outros autores como Chen[5][6] e Noel[7] nos oferecem visões mais "complexas" de representação de acervos. Representações baseadas em um universo de navegação tridimensional onde cada característica dos objetos gráficos nos leva a uma característica do acervo ou do documento. Quando chamados esta forma de representação de complexa, nos referimos a um modelo que não nos remete ao utilizado no dia a dia da maioria das pessoas, ou seja, um texto com *links*. Estes modelos complexos nos permitem acessar um documento através da visualização do universo onde o documento está inserido. Isto nos permite obter informações visuais, como posicionamento estratégico no universo em que está inserido, grau de relacionamento de um documento com outros ao seu redor, grau de relacionamento entre os diversos autores, etc. Enfim, temos condições de analisar todo um contexto que muitas vezes não pode ser representado apenas por palavras.

Numa outra frente, autores apresentam trabalhos que visam representar características específicas do acervo através de resultados gráficos gerados a partir de indicadores bibliométricos [8][9][10].

Neste capítulo apresentaremos dois modelos para visualização de um acervo classificado. O primeiro é um modelo tradicional baseado em *links* textuais e o segundo é um modelo, o MoGRIS*(Modelo Gráfico de Recuperação de Informação por Semântica)*, baseado numa interface tridimensional diferente dos mecanismos de exibição encontrados nos

sistemas atuais. Ambos os mecanismos se baseiam num mecanismo de agrupamento semântico de documentos.

O mecanismo de *Grupos* Semânticos que apresentamos é uma variação dos modelos de visualização encontrados em muitos mecanismos de busca disponíveis em sites como Google, Cadê e outros. Nestes mecanismos de busca o resultado é uma seqüência de *links* que nos direcionam diretamente para a página ou documento acessado. Estes mecanismos, apesar de muito eficientes no que diz respeito a pesquisa de conteúdo, nos apresentam uma infinidade de apontamentos para endereços dos mais diversos temas interligados pela palavra-chaves digitadas pelo usuário.

Uma proposta de classificação de documentos é oferecida no site *OntoWeb.com*[1], contudo, o mesmo também só nos oferece resultados baseados nas palavras do texto digitado pelo usuário sem a capacidade de identificar e agrupar documentos que não possuam tais elementos. Este se diferencia dos demais pelo uso de ontologias para verificar a importância de um documento no resultado da pesquisa[1] através de regras de engenharia de ontologias.

Seguindo o conceito de Grupos, adotado pelos *sites* já citados, desenvolvemos nosso primeiro mecanismo de visualização, dividindo a apresentação de nossos resultados de busca em cinco partes que veremos adiante.

Embora utilizemos o sistema de grupos, já implementado em alguns sites, nosso processo de construção destes grupos é distinto dos serviços disponíveis já relatados. Enquanto os serviços de agrupamentos atuais adotam palavras-chave para montar os grupos dinamicamente, nossos grupos podem conter documentos que sequer possuam tais palavras-chave. Esta capacidade pode ser muito boa quando o mecanismo de classificação semântica dos documentos é um mecanismo capaz de garantir uma alta taxa de acertos fazendo com que a grande maioria dos documentos listados sejam realmente associados ao que o usuário deseja. Por outro lado, se o mecanismo de classificação for falho, teremos uma grande quantidade de documentos listados não associados ao que o usuário buscou. Em nosso mecanismo de agrupamento, utilizamos o seguinte algoritmo:

1) Definimos o limite de similaridade ***S*** que permitirá a formação de grupos. Este valor de similaridade será utilizado como valor de corte no processo de inserção de documentos num grupamento.

2) Geramos a matriz ***N x N*** de similaridades entre os documentos, onde ***N*** é o número de documentos do sistema. Cada posição da matriz nos diz a similaridade existente entre dois documentos do acervo. Esta matriz pode ser gerada através de diversos processos como vimos no capítulo 2. Em nosso modelo, utilizamos o LSI[12] para gerarmos nossa matriz. O uso do LSI, nos permite fazer agrupamentos baseados numa similaridade semântica, conforme vimos no capítulo 3, no entanto podemos utilizar qualquer critério de similaridade que nos gere uma entrada conforme a figura 13.

Modelo de Matriz de Similaridade						
	D1	D2	D3	D4	D5	D6
D1	1	0,934	0,223	0,344	0,898	0,977
D2	0,934	1	0,645	0,526	0,898	0,911
D3	0,223	0,645	1	0,985	0,652	0,856
D4	0,344	0,526	0,985	1	0,586	0,778
D5	0,898	0,898	0,652	0,586	1	0,874
D6	0,977	0,911	0,856	0,778	0,874	1

Figura 13- Modelo da matriz de similaridade utilizada como entrada

3) Iniciamos a varredura da matriz pegando o primeiro elemento e criando uma categoria. Observando a figura 13 note que a varredura pode ser feita utilizado apenas uma das partes da matriz, a parte superior a diagonal ou a parte inferior devido a simetria apresentada pela mesma. Então ao pegarmos o segundo elemento, verificamos o valor de similaridade deste segundo documento com o documento da primeira categoria. Caso a similaridade entre os documento seja maior que ***S,*** este documento será agrupado ao documento 1 formando um cluster com dois elementos, caso contrário será criado um novo cluster apenas com o segundo documento.

4) Ao começarmos a ler as similaridades do n-ésimo documento, passamos por cada categoria criada. Caso o documento ***N*** tenha similaridade superior ao valor **S** com todos os documentos do cluster, ele é também inserido no cluster. Se um documento do cluster existente não estiver na faixa de similaridade, então criamos um novo cluster com o documento ***N*** e todos os documentos que eram similares a ele na categoria lida.

5) Após passar por todos os grupos existentes, o documento ***N*** pode ter sido inserido em nenhum ou em vários grupos. Logo poderemos ter categorias bem próximas

6) com diferenças de até um documento. É exatamente nestas pequenas diferenças entre grupos que poderemos identificar áreas de conhecimento próximas, identificadas exatamente pela interseção do conjunto de documentos entre elas. A figura 14 nos apresenta um exemplo de possíveis grupos. Temos os grupos *"Engenharia de Software"* e *"Programação OO"* que possuem como interseção os documentos *Doc1* e *Doc3*. Estes documentos em comum nos permitem dizer que estas categorias apresentam um elo de ligação, ou uma proximidade evidenciada pelo compartilhamento destes dois documentos em seus grupos. O mesmo podemos dizer para as áreas *"Medicina"* e *"Biotecnologia"*.

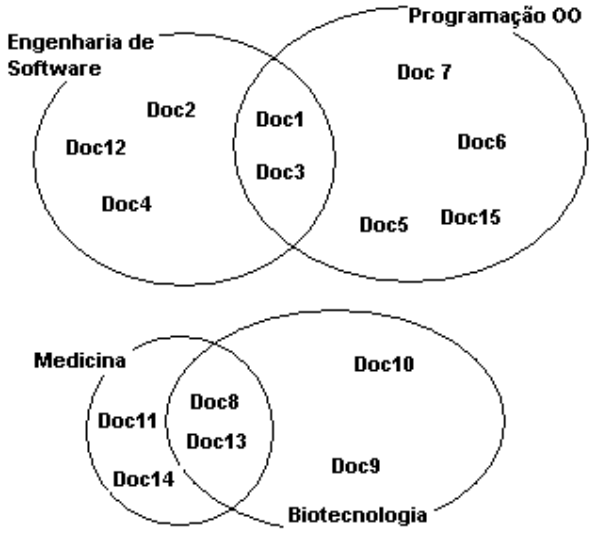


Figura 14- Exemplo de formação dos Grupos

Um Grupo Semântico de documentos é então composto por documentos que tratam de assuntos similares. Sua montagem independe do termo digitado pelo usuário num mecanismo de busca, pois o processo de montagem é feito no ato da indexação de um novo documento ao acervo. Com isto, quando um usuário faz uma busca sobre nossa base de grupos, ele não recebe como resultado apenas o documento que possui o termo de pesquisa desejado, mas também diversos outros documentos que tratam do mesmo assunto embora não utilizem as palavras digitadas pelo usuário. É sobre este conjunto de grupos semânticos que executaremos o mecanismo visualização de dados.

4.2) Visualização de Grupos Semânticos

Ainda que os grupos não sejam montados baseando-se apenas nas palavras chaves do texto digitado pelo usuário em um mecanismo de busca, desenvolvemos um modelo de exibição dos mesmos que pesquisa todos aqueles grupos que possuam os termos digitados pelo usuário. Uma vez encontrados todos os grupos que possuam os termos pesquisados pelo usuário, o sistema os ordena por ordem decrescente de quantidade de ocorrência dos termos e por ordem crescente de volume de documentos no grupo, ou seja, aquele grupo que apresentar maior ocorrência de termo em uma quantidade menor de documentos será o grupo mais relevante da pesquisa. A diferença para os demais sistemas de buscas já apresentados é que a busca pelas palavras chaves não acontece diretamente nos documentos e sim nos *grupos* montados previamente.

A realização de um processamento semântico dos documentos em tempo de consulta seria praticamente inviável devido ao volume de processamento de dados para encontrarmos tais documentos. Uma forma de realizar a consulta sem processar todos os documentos seria a existência de um dicionário léxico-semântico de associação de termos. Este dicionário nos permitiria, através de uma palavra, encontrar documentos similares baseados em palavras associadas ao termo pesquisado no dicionário. Como nosso objetivo é realizar a busca semântica sem ajuda de dicionários, em nosso projeto esta busca é realizada sobre uma base pré-classificada, e não em uma classificação realizada em tempo de consulta.

Buscamos criar um mecanismo de visualização de documentos para um sistema automático de classificação de documentos baseado no LSI[12]. Com isto, não é possível através de um número reduzido de palavras encontrar uma semelhança direta que não passe pelas próprias palavras. Isto significa que, se um usuário pesquisa por *"Carro"*, para encontrarmos todos os documentos que falam de *"veículos"*, temos que encontrar um documento que fale de "*Carro"* e a partir dele encontrar os seus similares. Isto se deve ao fato de que, se nós submetermos apenas o termo *"Carro"* ao mecanismo de classificação utilizado, não teríamos sucesso, pois o termo *"Carro"* sozinho não é capaz de criar um contexto semântico à sua volta. Como a própria definição nos diz: "*semântica é o estudo dos sentidos das frases e das palavras que a integram*" [11]. Ainda que o usuário digitasse uma frase com um contexto para um processo de consulta, o tempo de processamento e recebimento da resposta seria extremamente longo devido ao imenso volume de documentos

que estariam sujeitos a estas comparações. Com isto a matriz LSI gerada seria muito grande para ser processada rapidamente, inviabilizando o processo de consulta.

Uma vez definido o motivo de consultarmos por termos em uma busca semântica, partimos para a segunda etapa de exibição do resultado. Listadas as categorias ou grupos, o usuário seleciona a categoria desejada para a listagem de todos os documentos da mesma. Neste momento estamos exibindo o resultado semântico da consulta. O processo de localização de documentos e relações entre eles é trabalhado em cinco níveis conforme exibimos num modelo de consulta na figura 15. E é sobre esta estrutura que desenvolvemos nossos mecanismos de visualização.

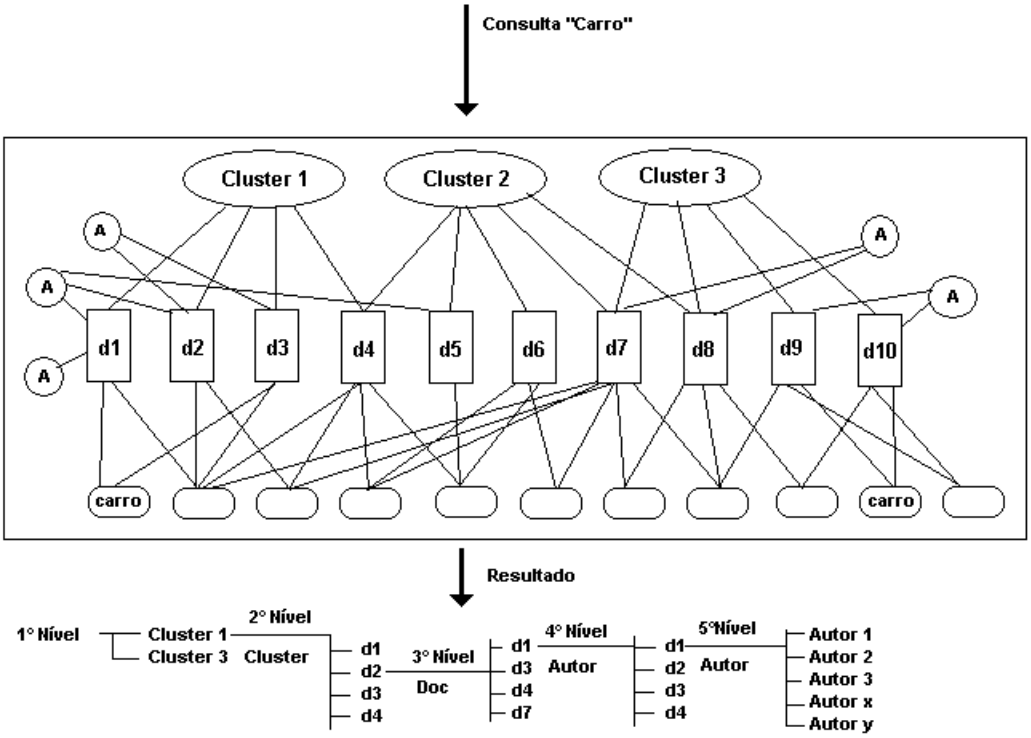


Figura 15- Exibição dos 5 níveis de consulta em documentos

Na figura 15, nos indica que no primeiro nível de resultado de consulta teremos os grupos(*clusters*) de documentos. No segundo nível trabalharemos com os documentos de um grupo. No terceiro nível apresentamos os documentos similares a um documento de segundo nível. No nível 4, apresentamos os documentos escritos por um autor selecionado no nível 2 ou 3. Por fim, o nível 5 nos mostra todos os autores similares a um autor selecionado.

4.3) Modelos de Visualização de Grupos Semânticos

Uma vez definido o projeto de navegação, partimos para a representação visual destes documentos. Quando tratamos da representação visual, abstraímos como os grupos foram gerados. Pensamos apenas que existe uma massa de dados que pode ser representada.

Neste trabalho apresentamos dois modelos de visualização da informação: um baseado num modelo trivial apresentado por muitos dispositivos de busca e softwares de controle de documentos e outro num modelo em 3D. Veremos agora a semântica de cada uma delas com suas características, suas vantagens, desvantagens.

4.4.1) Modelo *Links* Agrupados

Este modelo é baseado no formato padrão utilizado na grande maioria das ferramentas de navegação. Trabalhamos aqui com informação textual simples *casada* com informações transmitidas por posicionamento dos *links* e referência de cores.

Como em nosso projeto de navegação pelos grupos utilizamos 5 etapas de visualização das informações começaremos apresentando as características da primeira etapa de navegação:

4.4.1.1) Navegação em 1° Nível (Categorias)

O primeiro nível de navegação utilizando o sistema de texto nos retorna todas as categorias que possuam o termo digitado. A ordem com que as categorias aparecem no resultado nos indica a significância desta categoria com relação ao termo pesquisado. A figura 16 nos mostra o resultado de uma consulta pelo termo *"Futebol"*. Nesta figura, a listagem exibida nos mostra a lista de categorias mais relevantes para os termos consultados. Podemos ver que todas as categorias têm algo em comum com o termo consultado *"Futebol"*. Vejamos a seguir as informações exibidas no resultado de primeiro nível.

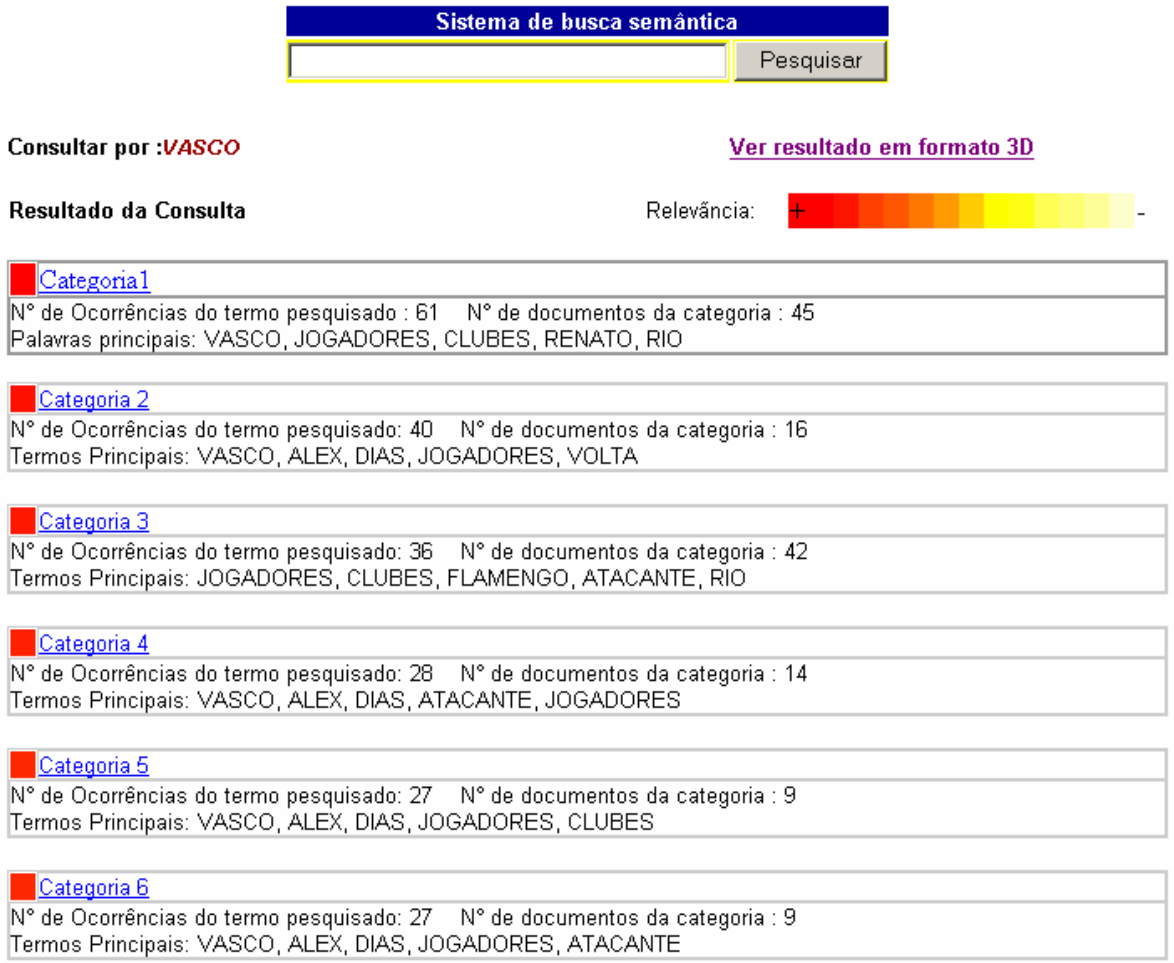


Figura 16- Tela de visualização dos grupos ou categorias

N° de ocorrências do termo pesquisado: este número totaliza a ocorrência de cada uma das palavras dos termos pesquisados na categoria. Por exemplo, se pesquisássemos a seqüência *"Futebol Brasileiro"*, contaríamos o total de ocorrências do termo "*Futebol*" e de "*Brasileiro*" que aparecem na categoria e somaríamos os dois. Como não pesquisamos diretamente sobre o arquivo com o texto e o índice sobre a qual trabalhamos é montado termo a termo, não temos nesta implementação a possibilidade de buscarmos a seqüência de termos inteira. Por este motivo, totalizamos a ocorrência de cada termo dentro de cada categoria e exibimos a ordenação das categorias dentro de acordo com este total.

N° de documentos da categoria: este número totaliza a quantidade de documentos existentes numa categoria num determinado instante. Esta quantidade pode variar de consulta para consulta de acordo com a inserção de novos documentos no acervo.

Palavras chaves das categorias: uma vez listada a categoria, o sistema pesquisa as cinco palavras com maior ocorrência nos documentos da categoria. Estes cinco termos são exibidos em ordem decrescente de importância(quantidade) dentro da categoria visando auxiliar na identificação do assunto de que a categoria trata. Eventualmente duas categorias poderão apresentar a mesma seqüência de cinco termos, isto pode nos parecer que as categorias são iguais. No entanto, quando estudamos as categorias mais detalhadamente e observamos seus documentos, podemos ver que, embora tenham os mesmos elementos identificadores, elas apresentam nuances que nos permitem classificá-las como bem próximas, mas não como categorias iguais. Isto acontece porque, embora os cinco primeiros termos sejam semelhantes, a categoria é formada por todo um universo de palavras que compõem os documentos e que podem diferir em muitos outros termos.

Posicionamento das categorias: este item identifica quais categorias apresentam maior quantidade de ocorrência dos termos pesquisados em uma menor quantidade de documentos. Com isto consideramos como categorias mais relevantes aquelas que se encontram no topo desta classificação. Por exemplo, se duas categorias apresentam dez ocorrências de um termo, o sistema considerará mais relevante aquela que contiver os termos numa quantidade menor de documentos.

Cor de identificação: Um outro efeito visual que nos permite identificar o grau de relevância de um cluster no resultado de uma pesquisa é o retângulo colorido que aparece ao lado esquerdo do texto da categoria. A cor deste elemento varia de vermelho a branco,

passando pelo amarelo de acordo com a distância da categoria dos termos pesquisados, ou seja, as categorias que apresentam maior número de termos pesquisados são as que apresentam menor distância dos termos pesquisados. Um modelo similar e que nos serviu de inspiração é apresentado no site *Ontoweb.com*[1] onde a cor indica o grau de relevância de um documento dentro do resultado de busca e não de um agrupamento. No caso do Ontoweb a relevância é calculada através da engenharia de ontologias[1] e não da forma como estamos propondo.

Estes são elementos informativos contidos no primeiro nível de consulta. Uma vez exibidos as categorias, o usuário pode acessar os documentos da que lhe for mais interessante.

4.4.1.2) Navegação em 2° Nível (Documentos da categoria)

Neste momento o usuário escolheu o cluster que deseja conhecer. Começa aqui nosso resultado da pesquisa semântica. O sistema então seleciona dentro da categoria o documento com a maior quantidade de termos pertencentes a consulta. Este documento será o ponto de referência para a ordenação da semântica. Como sabemos que nosso cluster possui documentos que estão dentro de uma mesma faixa de similaridade, ordenamos estes documentos de acordo com a similaridade com relação ao documento principal escolhido para esta consulta.

A figura 17 apresenta um modelo da listagem do resultado. Esta listagem possui alguns itens que descreveremos agora:

Nome do documento: Título do documento

Grau de similaridade: Este elemento nos retorna o grau de similaridade entre o documento e o documento mestre selecionado para a categoria. O grau de similaridade entre os documentos de uma categoria não pode ser inferior a 80% por cento visto que, definimos na construção dos grupos, que vimos no capítulo 3, que este seria o valor mínimo aceitável de similaridade para considerarmos uma categoria.

Número de acesso: número que indica a quantidade de acessos do documento até o momento.

Cor do marcador: a cor do marcador localizado a frente do documento indica o grau de intensidade com que um documento em questão tem sido acessado. Esta informação é obtida através de uma comparação do número de acessos do documento mais acessado(R) com o número de acessos do documento exibido na linha(N).

$$Intensidade\ da\ Cor = N / R$$

Quanto mais vermelho, mais próximo de ter alcançado uma intensidade máxima de acessos. Quanto mais claro, menos intensamente o documento tem sido acessado, sofrendo, portanto uma variação de vermelho a branco.

Nome dos Autores: Nome dos autores do documento. Este item nos permite acessar, através dos *links* que acompanham o autor, mais duas informações significativas: quais outros documentos que este autor produziu e os autores que escrevem documentos na mesma área do autor assinalado.

Número de documentos similares: o número de documentos similares é obtido buscando-se na base todos os documentos cuja similaridade esteja dentro da faixa de similaridade configurada. Este número não depende da categoria em que o documento está inserido, mas sim das categorias das quais ele faz parte. Suponhamos que o mecanismo de busca esteja configurado para agrupar todos os documentos que possuam uma similaridade superior 0.8, como explicado no LSI. Ao selecionarmos um documento, buscamos todos os documentos que possuem uma similaridade maior este valor. Isto não depende das categorias em que os documentos estão inseridos.

Este segundo nível pode nos permitir acessar diretamente o documento ou novamente ver os documentos similares ao documento desejado utilizando a interface de *links* ou por meio da interface em 3D.

Figura 17- Exibição de consulta em 2 ª fase

4.4.1.3) Navegação em 3° Nível (Documentos Similares)

O terceiro nível de visualização é exibido quando o usuário seleciona a opção *"Ver Similares"*. Ao fazer esta escolha o usuário acessa o maior nível de abstração semântica do mecanismo de busca. Ao descer até o 3° nível, o usuário não está mais ligado diretamente ao termo consultado e sim a semântica do termo consultado.

Quando acessamos o primeiro nível, estamos visualizando categorias que estão diretamente ligadas aos termos das consultas. O segundo nível nos permite visualizar os documento da categoria escolhida no primeiro nível. Ainda neste nível, nosso documento mestre de comparação é aquele que mais possui os termos pesquisados, embora os documentos listados não precisem ter em seu corpo nenhuma ocorrência dos elementos da consulta, visto que estes documentos já apresentam entre si uma relação semântica. Neste

terceiro nível, o usuário pode acioná-lo através de um documento que não apresentou o termo pesquisado em seu conteúdo no segundo nível. A nova listagem, então originada do termo de consulta, passa a ter ligação puramente semântica com o termo da consulta, podendo eventualmente possuí-los em seu conteúdo, mas já não há ligação real entre estes termos do documento e o digitado pelo usuário. A figura 18 nos mostra a uma consulta em terceiro nível.

Figura 18 -Navegação no 3° nível (documentos similares)

A semântica de terceiro nível apresenta todas as característica da semântica de segundo nível, acrescentando-se apenas a informação de qual documento é o de referência da tela. Todas as similaridades apresentadas são calculadas em função deste documento referência. Observe que neste ponto o termo inicial da consulta, não aparece mais na tela, pois

a navegação agora ocorre num ambiente semântico e não mais no ambiente do termo da consulta.

Todas as etapas nos apresentam a possibilidade de exibirmos o resultado da consulta em um ambiente tridimensional. E é justamente esta forma tridimensional que apresentaremos no segundo modelo que proporemos adiante.

4.4.1.4) Navegação em 4° Nível (Documentos dos Autores)

O que chamamos de quarto nível de navegação atua num outro aspecto relevante dos documentos que são os autores do mesmo. Cada documento listado nos níveis 2 e 3, podem ou não ter seus autores listados. Quando isto acontecer, o nome do autor nos remete a outro conjunto de documentos que são de autoria do autor. Com isto podemos verificar a área de preferência do autor, os assuntos mais trabalhados e outros. Temos disponível uma gama de documentos que nos permitem conhecer o autor nos mais diversos aspectos.

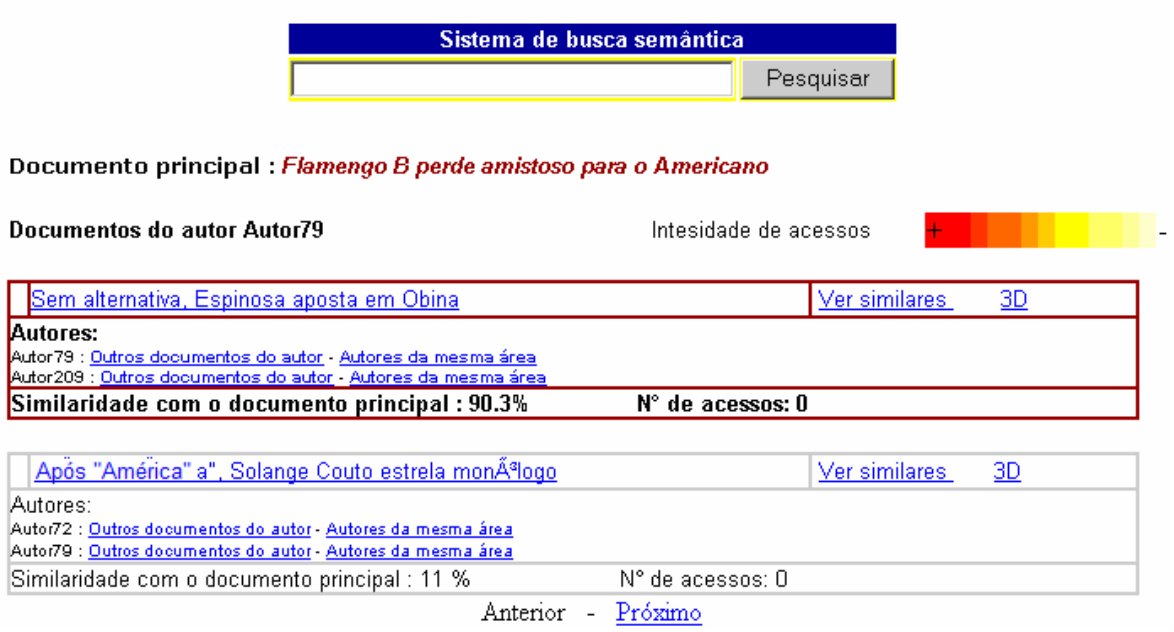


Figura 19- Visualização dos documentos produzidos por um autor

Os documentos produzidos pelo autor selecionado são apresentados segundo a ordem de similaridade com o documento de origem acessado e que se tornou a referência da consulta. A figura 19 nos mostra uma listagem de autores. A partir desta lista podemos acessar os documentos da mesma forma que no nível 3 visto que a etapa apresenta as mesmas características apresentadas no 3° Nível de visualização.

4.4.1.5) Navegação em 5° Nível (Autores de mesmas áreas)

O quinto e último nível de visualização contém informações que nos afastam do conceito puro de similaridade de documentos. Embora utilizemos os mesmos cálculos para o agrupamento dos documentos, neste nível exibimos todos os autores que possuem mesma área de atuação e interesse de acordo com o estudo de seus textos. O *link "Autores de mesma área"* área nos níveis 2,3 e 4 nos remete a uma lista de todos os autores que escreveram documentos que se encontram em Grupos afins. Ou seja, dizemos que se um autor escreve sobre um assunto e vários outros autores também escrevem sobre este mesmo assunto, através da comparação de seus documentos produzidos podemos dizer que todos estes autores possuem uma certa relação em comum, que se dá pela área em comum de estudo.

Para obtermos os autores similares a um outro selecionado como referência, realizamos o seguinte procedimento:

- Selecionamos todos os grupos de documentos dos quais o autor selecionado como referência participa;
- De posse de todos estes grupos, buscamos todos os autores que também escreveram para a área.
- Para cada autor encontramos o total de acesso obtido pelo autor somando-se todos os acessos realizados em documentos dos quais o mesmo é autor.
- Para cada autor totalizamos o número de documentos escritos que fazem parte das áreas do autor referência.
- De posse da quantidade de acessos que um autor possui e do número de documentos que o mesmo escreveu numa área podemos realizar uma comparação entre autores de forma a indicar os autores mais relevantes, mais visitados e mais próximos do autores utilizado como referência de comparação.

Este nível nos permite identificar grupos de autores com o mesmo interesse e identificarmos através de seus dados, pólos científicos em certas áreas, países e redes de relacionamentos.

Neste quinto nível, podemos retornar ao nível de Categorias, ou seja, o primeiro nível. Isto nos permite identificar todas as áreas das quais um autor tem participado. Também podemos, através dele, chegar a todos os exemplares escritos por um autor. Podemos observar estas características no modelo da figura 20 a seguir.

Figura 20 - Visualização de quinto nível

4.4.2) MoGRIS

Visando provar que é possível representarmos num gráfico único o conjunto de informações representado pelo sistema de *Links Clusterizados*, apresentado no item anterior, nesta seção apresentamos o MoGRIS. Este modelo de visualização é baseado no modelo orbital de planetas. Com o este modelo surgiu da necessidade de mapearmos em um ambiente gráfico um universo de documentos e algumas relações existentes entre eles e que foram apresentadas no modelo apresentado anteriormente. Como vimos no capítulo 2, esta tem sido uma área onde muitos pesquisadores têm dedicado seu tempo. Trabalhos como os de Chen[5][6], Merkl[14], Casado[16] , Raphael[17], Paula[9] buscam representar graficamente

documentos e ligações existentes entre eles como ano, autores, assunto, nacionalidades dentre outras.

Com o objetivo de criar mais uma metodologia que possa auxiliar na extração destas relações implícitas, desenvolvemos o modelo Orbital. Este modelo é uma metáfora de um sistema planetário aplicado ao contexto de visualização de grupos, documentos e suas relações, como mostra a figura 21. Para isto representamos o universo de categorias ou documentos como ***Planetas*** que orbitam em torno de um elemento central(***Sol***), a consulta do usuário. Quando o sistema estiver representando documentos, ainda teremos a representação dos Autores dos documentos, mapeados por ***Luas*** ou ***Satélites*** em torno dos documentos.

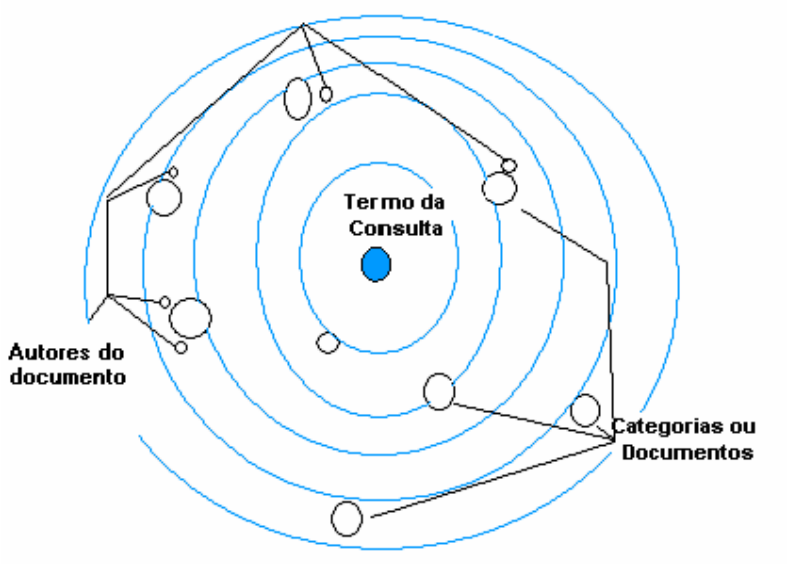


Figura 21- Metáfora da representação Orbital.

Este modelo de visualização de dados apresentado na figura 20 trabalha sobre a mesma base de dados montada pelos processos de agrupamento vistos anteriormente e utilizado no modelo de *Links Clusterizados* apresentado no item anterior. No entanto, no modelo Orbital trabalhamos com a visualização de documentos em apenas quatro etapas. A primeira etapa é responsável por exibir todos os grupos associados a uma consulta de um termo, neste caso não teremos os objetos Satélites, visto que não identificaremos documentos específicos. Nos demais níveis, nós representamos os documentos de uma categoria específica com base nos termos de consulta. Nestes níveis exibimos a figura das luas, pois poderemos ter autores cadastrados. Veremos agora mais detalhes sobre cada um dos níveis de visualização.

4.4.2.1) Navegação em 1° Nível (Visualização dos Grupos)

Da mesma forma que trabalhamos no modelo de *Links* Agrupados, o primeiro nível nos mostra os grupos de documentos similares a uma consulta, no entanto o resultado é apresentado de uma maneira bem diferente. Vejamos agora o que significa cada elemento que compõe a figura 22 que representa o resultado de uma busca.

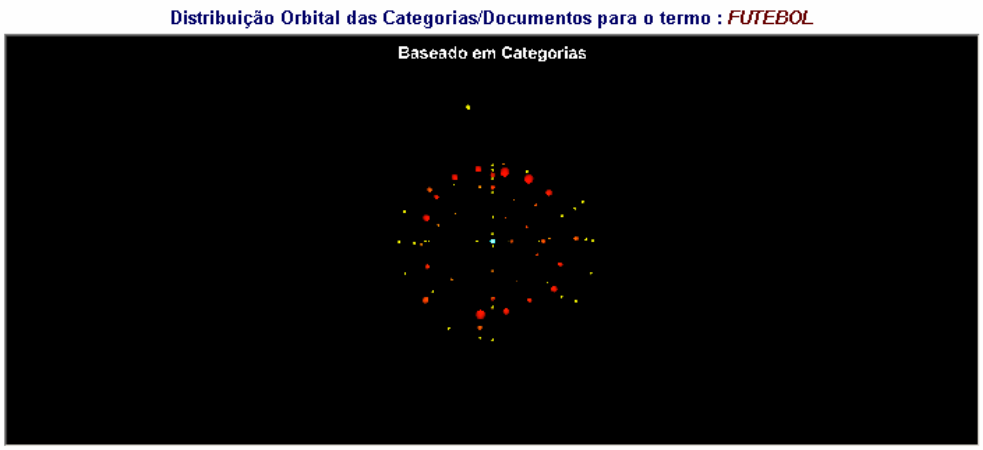


Figura 22 - 1° Nível do modelo Orbital –Categorias

Núcleo: No centro do sistema encontramos uma esfera azul. Esta esfera de diâmetro constante representa os termos da consulta do usuário, ou seja, sua consulta.

Planetas: Cada planeta representa uma categoria de documentos. A distribuição destas categorias e suas características são listadas abaixo:

a) Distância do Planeta ao Centro: A distância do planeta ao centro do sistema corresponde ao grau de relevância dos termos da consulta que aparecem na categoria. Consideramos como grau de relevância ou fator de relevância o número de ocorrências de termos dividido pelo número de documentos da categoria.

Fator de Relevância = N° de Ocorrência / N° de Documentos

Este fator de relevância nos permite identificar categorias com poucos documentos e que o número de ocorrência dos termos de consulta são proporcionalmente mais significativos que em categorias com grandes ocorrências, mas também grande quantidade de documentos.

Esta distância também será exibida quando passarmos o mouse sobre o objeto juntamente com as cinco palavras mais significativas do documento.

b) Cor do planeta: identifica a quantidade de termos da consulta na categoria. A cor varia de vermelho a branco. No padrão RGB(*Red Green Blue*) esta variação seria de um RGB(*FF0000*) a RGB(*FFFF00*). Com isto temos que os documentos mais vermelhos são aqueles que possuem mais termos relacionados com a consulta. Este item não leva em consideração o fator de relevância, mas apenas as ocorrências dos termos.

c) Volume do planeta: identifica a quantidade de documentos da categoria. Como podemos ter categorias com quantidade muito grande de documentos, criamos faixas de valores que são representadas por raios definidos. Isto impede que tenhamos planetas com tamanhos desproporcionais e que prejudiquem a visualização do modelo.

d) Hint: o recurso do hint*(textos que surgem quando passamos o mouse sobre o objeto)* é utilizado para permitir ao usuário identificar algumas informações que não são possíveis de visualizarmos em um objeto puramente gráfico. Nestes hints exibimos ao usuário as 5 palavras chaves principais da categoria. Esta informação é a mesma passada ao usuário no modelo de *Link*s clusterizados, explicado como primeiro modelo de visualização.Também exibimos no hint a distância do planeta(categoria) até o objeto central(consulta). Exibimos este valor pois trabalhamos num modelo tridimensional onde os ângulos de perspectiva podem dar uma falsa noção de proximidade de acordo com o ponto de referência em que fazemos a visualização.

A figura 23 nos mostra um zoom do objeto em 3D, exibido na figura 8, para uma busca utilizando o termo *“Futebol”*.

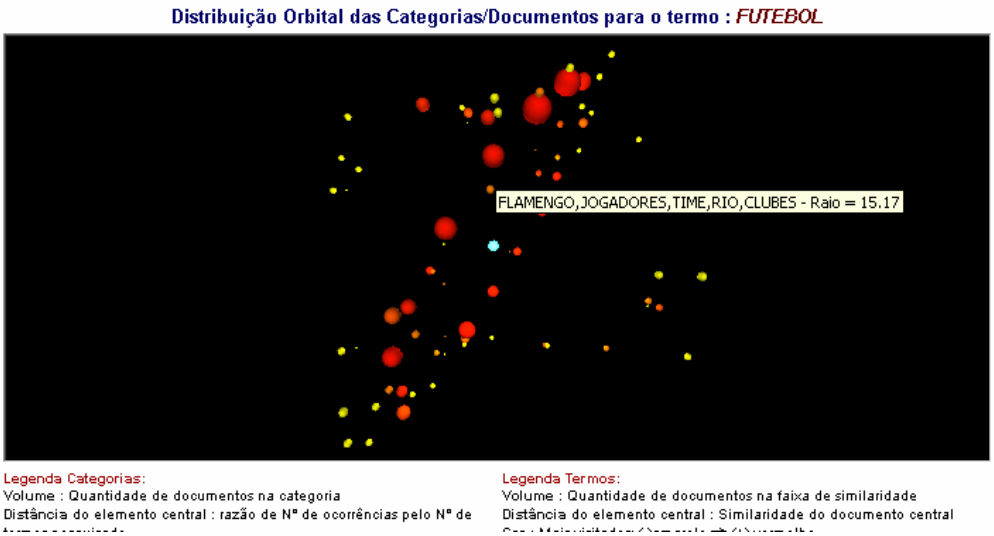


Figura 23 - Zoom do 1° Nível de busca em 3D- Categorias

4.4.2.2) Navegação em 2° Nível (Visualização dos documentos da categoria)

O segundo nível de navegação no ambiente 3D é acessado através das categorias no primeiro nível de consulta. Ao acessarmos uma categoria, acessamos todos os documentos que se encontram classificados nesta categoria.

No momento do acesso à categoria, obtemos o documento mais significativo da categoria com relação aos termos da consulta, ou seja, realizamos uma contagem da quantidade de ocorrências dos temos nos documentos e pegamos aquele documento com maior ocorrência dos termos consultados. Este documento de referência estará no meio do sistema planetário e substituindo o termo da consulta que era exibido no primeiro nível. Os demais documentos da categoria são exibidos numa distribuição espacial relacionada com a similaridade de cada documento em relação ao documento de referência. Com isto cada documento passa a ser um *planeta* orbitando em torno do documento referência*(sol)*.

Uma terceira figura que pode aparecer neste nível são as *Luas* ou *Satélites.* Estes objetos representam os autores de cada documento. As *Luas* também apresentam características próprias que nos trazem mais informações sobre os autores que veremos a seguir. A figura 24 nos mostra um modelo montado sobre a consulta da palavra *"Futebol"*.

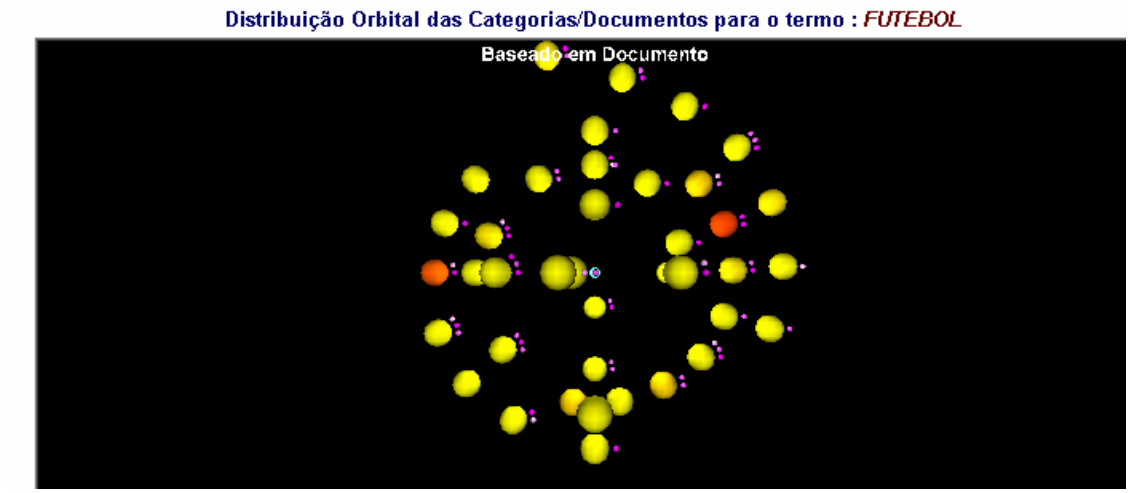


Figura 24 - 2° Nível (Documentos)

O modelo da figura 24 nos mostra o resultado de um consulta em 3D, vamos agora identificar as características importantes do modelo:

Termo Central: O termo central em cor azul representa o documento escolhido para ser o ponto de partida para a navegação em uma categoria. A escolha do documento é feita baseada em parâmetros não semânticos, pois apenas verificamos os documentos em que mais ocorreram os termos das consultas e selecionamos o que será o elemento central. Apesar desta escolha não ser baseada em cálculos semânticos, o fato deste documento estar nesta categoria é um fato semântico, visto que ele foi agrupado através de um mecanismo de análise semântica, logo todos os demais documentos possuem um grau de similaridade dentro do limite em que foi configurada a similaridade semântica.

Planetas: Os planetas são elementos que representam os documentos da categoria ou cluster. Os planetas possuem algumas características especiais que exibimos abaixo:

Volume do planeta: esta característica indica os documentos que possuem maior número de documentos similares. No entanto, para evitarmos que pudéssemos encontrar documentos com tamanhos quase imperceptíveis ou documentos muito grandes, nós truncamos estas medidas para que os raios variem de 0.2 a 3.0 unidades de medida de acordo com o número de documentos similares, ou seja, de 6 a 90 documentos. Fora destes extremos, os documentos são representados com as medidas citadas.

Cor do planeta: esta característica nos aponta a intensidade de acesso de um documento dentro do acervo utilizando como referência o documento que teve mais acessos.

Isto provocará uma variação de cor que vai do amarelo ao Vermelho. Sendo que os documentos*(Planetas)* com cores mais avermelhadas são os que tiveram maior número de acessos. Conforme outros documentos são acessados, os gráficos apresentarão alterações de cores em cada objeto.

$$Intensidade\ do\ vermelho = 1 - (M - N) / M$$

M:Maior n° de acessos a um documento
N: N° de acessos do documento a ser exibido

Distância do planeta ao sol: esta característica indica o grau de similaridade do documento com o documento de referência. O grau de similaridade é calculado utilizando o LSI[12]. Embora o documento de referência possa ter muitos outros documentos similares, os documentos exibidos são exclusivamente os contidos na categoria selecionada, assim como é feito no modelo de *Link*s Agrupados. O grau de similaridade varia de 0 a 1. Como estamos trabalhando sobre um valor de corte para montagem das categorias, podemos dizer que nossa faixa de similaridade é de 0,8 a 1. Estes valores são mapeados em distâncias do centro. Estas distâncias não devem ser muito pequenas para não criarem uma massa de documentos embolada com o objeto central e nem muito grandes a ponto de se perderem no mapa.

Luas: Podemos visualizar na figura 24 que alguns planetas possuem Luas orbitando ao seu redor. Cada uma destas luas representa um autor do documento. As Luas possuem características próprias que são variação de cor e volume. Vejamos agora o que cada uma delas representa:

Cor das Luas: A cor das luas representa a intensidade com que um documento é acessado. Esta intensidade é medida comparando-se o acesso a todos os documentos do autor na categoria com o autor de maior acesso na categoria. A cor das luas varia de RGB FF00FF(um tom de vinho) a FFFFFF(branco), ou seja, quanto mais escuro, mais acessado. O algoritmo para cálculo da cor segue os seguintes passos:

- Contamos o número de acessos de cada documento do cluster e obtemos o maior valor *M (Máximo da Categoria)*;

- Para cada autor encontrado, verifica-se cada documento do autor que pertença ao cluster visualizado;

- Totalizamos a quantidade de acessos para cada autor *T(Total Autor);*

- Com estes valores de *T* e *M* calcularemos a intensidade da cor como:

$$Intensidade\ da\ cor = 1 - (M - T)/M$$

Volume das Luas: O volume das luas representa o volume de documentos produzidos por um autor quando comparado com o autor que mais produziu documentos no acervo. Isto nos permite identificar se o autor, quando comparado com outros autores do acervo, possuem uma alta produção de textos. O raio, fator que determina os volume das luas, varia de 0.25 a 0.75, onde o segundo nos diz que o autor possui um número de produções muito próximo ou igual ao número máximo de documentos produzidos por um autor no acervo. O primeiro nos diz que a produção é muito pequena quando comparada à quantidade máxima de documentos produzidos por um autor.

Hint das Luas: Os textos que são exibidos ao passarmos o mouse sobre uma *Lua* nos informam o nome do autor e o número de documentos produzido pelo mesmo. Podemos ver um exemplo na figura 25. Estes textos têm a finalidade de facilitar o entendimento dos objetos gráficos determinando claramente, por exemplo, o número de documento, que no gráfico apresentará variação muito pequena estas quantidades forem muito grandes.

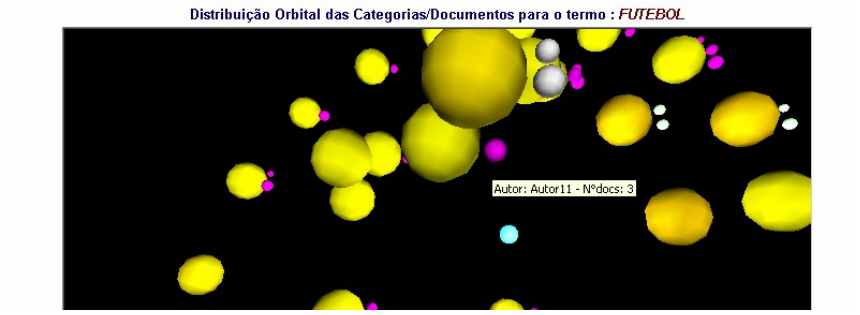


Figura 25- Zoom de uma consulta em 2° Nível

4.4.2.3) Navegação em 3° Nível (Visualização de Similaridade de Autores)

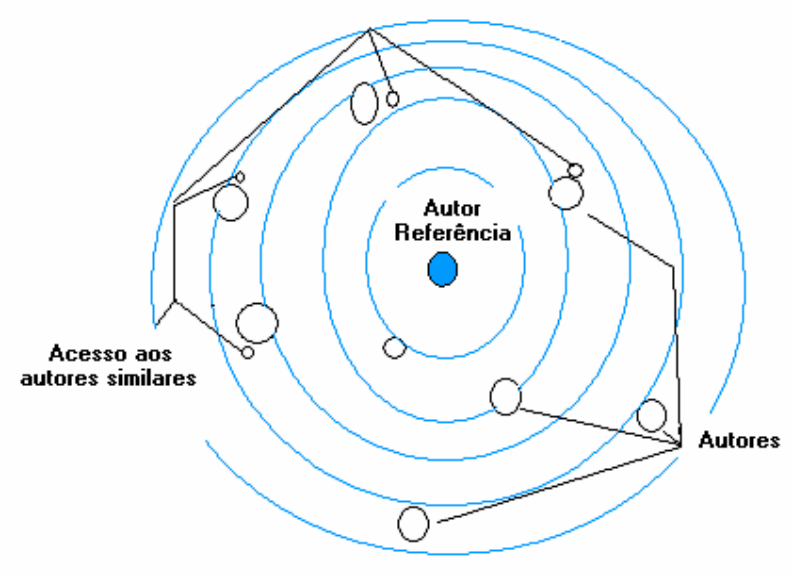


Figura 26- Modelo orbital para Similaridade de autores

O terceiro nível de visualização gráfica é composto pela similaridade de autores. Ao escolhermos no segundo nível um autor, abrimos um novo gráfico onde o autor escolhido torna-se o centro da comparação(sol), veja figura 26. A partir deste elemento, são pesquisados todos os autores que escreveram documentos nas mesmas áreas que o autor selecionado. Vejamos agora os demais elementos que compõem este gráfico.

Planetas: Os planetas são elementos que representam os autores similares ao autor referência. A seleção de um planeta nos levará à todas as áreas ou cluster de um autor, também representados através da metáfora planetária de forma semelhante ao primeiro nível. Os planetas possuem algumas características especiais que exibimos abaixo:

Volume do planeta: neste nível, esta característica indica o volume de documentos escrito pelo autor dentro das áreas que o autor referência pertence. Isto quer dizer que planetas com um volume muito grande indicam que o autor possui muitos documentos dentro da área em questão. Para evitarmos que pudéssemos encontrar planetas com tamanhos imperceptíveis ou muito grandes, nós truncamos estas medidas para que os raios variem de 0.2 a 3.0 unidades de medida de acordo com o número de documentos escritos pelo autor.

Cor do planeta: esta característica nos aponta a intensidade de acesso de um autor utilizando como referência o autor que teve mais acessos dentro das categorias relacionadas ao autor de referência. Isto provocará uma variação de cor que vai do amarelo ao Vermelho. Sendo que os documentos*(Planetas)* com cores mais avermelhadas são os que tiveram maior número de acessos. Conforme outros documentos do autor são acessados, os gráficos apresentarão alterações de cores em cada objeto.

Distância do centro: esta característica nos mostra o grau de proximidade das áreas de interesse de um autor com relação às áreas de interesse do autor referência. A distância baseia-se em uma pontuação calculada segundo a regra abaixo:

$$Distância = (C * 5) + (A)$$

C : Número de documentos onde o autor é co-autor do autor referência

A : Número de documentos da mesma área que o autor referência sem co-autoria

O fator de multiplicação por 5 para documentos de co-autoria visa acentuar a importância de autores que trabalharam junto do autor utilizado como referência, o que nos garante a existência de um elo de ligação entre os autores. Quando analisamos documentos de mesma área sem as co-autorias, temos um elo estatístico de áreas de interesse, visto que se dois autores escrevem documentos de mesma área de conhecimento, teoricamente são autores com mesmos interesses de pesquisa.

Luas: Os planetas são elementos que representam os autores similares ao autor referência. Neste terceiro nível cada planeta conterá apenas uma Lua. Esta lua é apenas um elo de ligação com um novo gráfico comparativo entre o autor a que pertence, ou seja, onde órbita, e os autores similares ao autor relacionado com a lua. A cor das luas nos indica a quantidade de autores similares existentes para o autor em questão, ou seja, muitos autores similares nos exibirão uma lua com a cor muito acentuada (RGB FF00FF) e para poucos autores similares teremos uma lua tendendo ao branco(RGB FFFFFF). Ao selecionarmos uma lua, abrimos novamente o gráfico de similaridade, figura 27, de autores alterando o autor de referência.

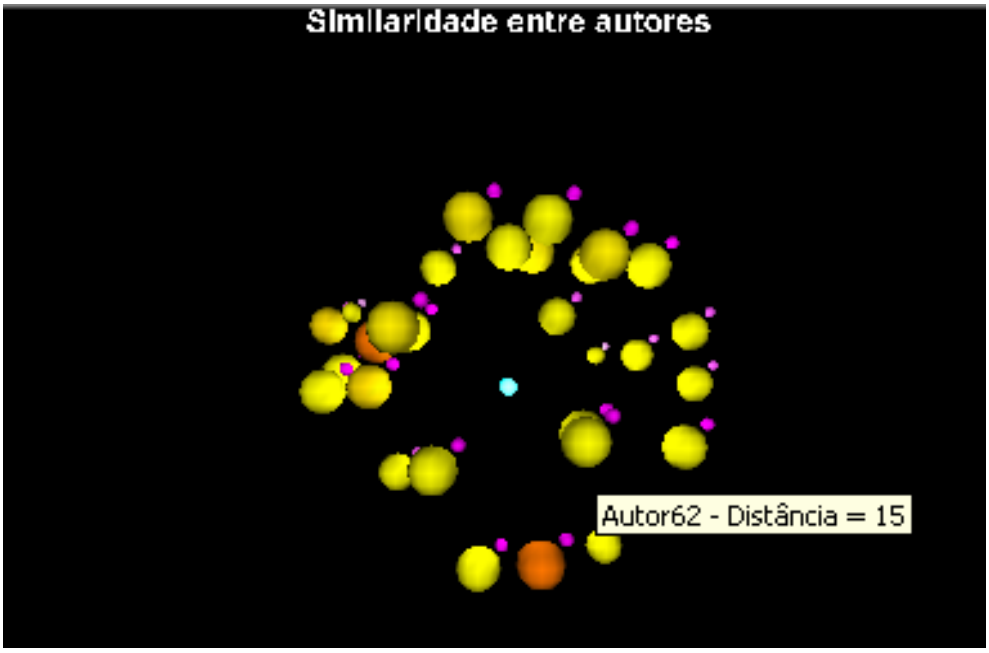


Figura 27 - Modelo Orbital para comparação entre autores

4.4.2.4) Navegação em 4° Nível (Visualização de Grupos de um Autor)

O quarto e último nível de visualização gráfica é composto por todas as áreas ou grupos de um autor selecionado no nível anterior. Ao escolhermos um autor no terceiro nível, abrimos um novo gráfico onde o autor escolhido torna-se o centro da comparação(*sol*) do modelo de primeiro nível. Veja modelo da figura 27.

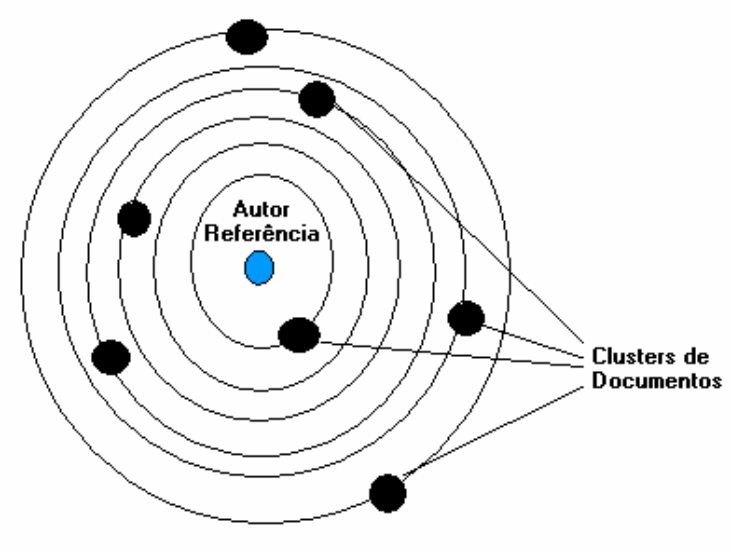


Figura 28 - Layout da Visualização em 5° Nível

No primeiro nível de navegação trabalhamos sobre a referência da consulta. Quando acessamos o quarto nível, a representação nos direciona para todas as categorias de um determinado autor. A figura 22 ainda continua sendo uma representação do modelo utilizado neste nível, porém termos algumas alterações semânticas que descrevemos abaixo:

Planetas: Cada planeta representa uma categoria de documentos. A distribuição destas categorias e suas características são listadas abaixo:

a) Distância do Planeta ao Centro: A distância do planeta ao centro do sistema corresponde ao grau de relevância do autor referência na categoria. Consideramos como grau de relevância ou fator de relevância o número de documento do autor dividido pelo número de documentos da categoria.

Fator de Relevância = N° de Ocorrência do autor / N° de Documentos da categoria

Este fator de relevância nos permite identificar categorias onde os documentos do autor são proporcionalmente mais significativos.

Esta distância também será exibida quando passamos o mouse sobre o objeto juntamente com as cinco palavras mais significativas do documento.

b) Cor do planeta: identifica a quantidade de documentos do autor na consulta na categoria. A cor varia de vermelho a branco. No padrão RGB(*Red Green Blue*) esta variação seria de um RGB(*FF0000*) a RGB(*FFFF00*). Com isto temos que as categorias mais vermelhas são aquelas que possuem mais documentos do autor em questão. Este item não leva em consideração o fator de relevância, mas apenas a quantidade de documentos existentes.

c) Volume do planeta: o volume dos planetas neste nível possui o mesmo significado que nos primeiro nível, ou seja, quanto maior o planeta, maior o número de documentos que o mesmo possui.

d) Hint: os hints deste nível também conterão as mesmas informações contidas no primeiro nível de navegação descrito no item 4.3.2.1

O modelo de quarto nível nos faz retornar ao ponto de partida de nosso gráfico, a exibição de categorias de documentos, porém neste momento abstraindo-se do termo inicial da consulta. Com isto temos nosso processo de navegação entrando num ciclo que poderá se repetir pelas várias etapas, sempre mostrando informações atualizadas de acordo com os documentos selecionados.

4.4) Conclusão

Neste capítulo apresentamos um modelo de visualização semântica de documentos. O modelo apresentado se divide em um modelo gráfico 3D e em um modelo textual baseado em *links.* Em ambos os modelos apresentados, o grande diferencial com relação ao demais modelos apresentados no capítulo 3, é a condição de utilizar uma notação única que nos permite encontrar relações entre documentos e entre os autores destes documentos. Para montarmos estes modelos, os documentos precisam estar classificados de acordo com alguma regra pré-definida que crie grupos de documentos e ligações entre cada um deles.

Nos capítulos 5 e 6 estaremos apresentando nosso modelo de visualização semântica executado sobre um acervo classificado através de análise de textos e sobre um acervo classificado através das referências bibliográficas de cada documento, respectivamente. No capítulo 5, mostramos toda a arquitetura que dá suporte ao modelo. O capítulo 6 apresenta um protótipo desta arquitetura e os resultados obtidos quando executado sobre um acervo real.

Capítulo 5

SiGRIS
Sistema Gráfico de Recuperação de Informação por Semântica

O modelo de visualização baseado numa metáfora planetária, MoGRIS, descrito no capítulo 4, é apenas parte de um sistema modular desenvolvido para oferecer aos usuários *web* uma alternativa diferente dos modelos tradicionais de consulta e visualização disponíveis na Internet.

Para oferecer este serviço, projetamos um conjunto de ferramentas que trabalham de forma independente, porém integradas por um modelo de dados em comum que permite com que cada ferramenta visualize o resultado das demais e utilize a informação desejada. A este sistema integrado denominamos SiGRIS*(Sistema Gráfico de Recuperação de Informação por Semântica)*

A arquitetura do sistema é exibida na figura 29.

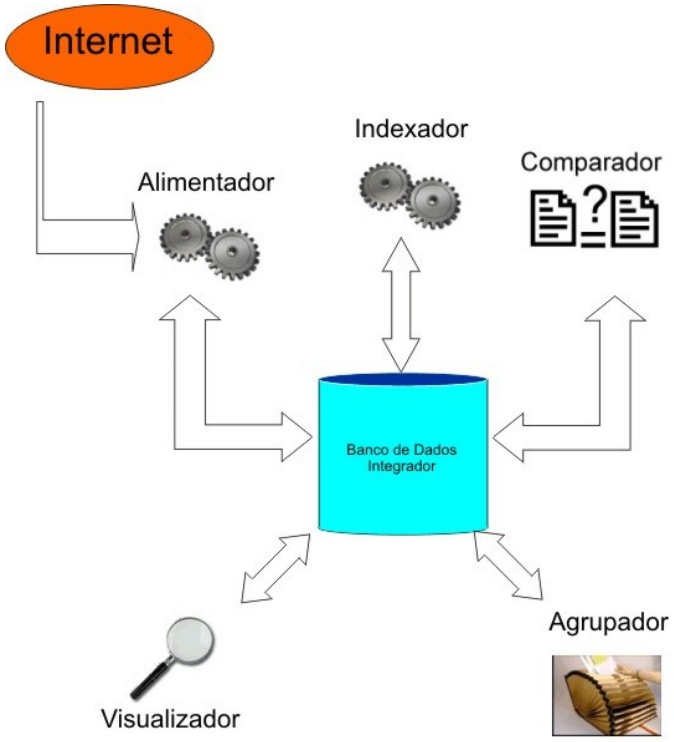


Figura 29- Arquitetura do sistema do projeto de visualização

A arquitetura proposta para o SiGRIS, nos permite melhorar ou até mesmo substituir cada um dos aplicativos que colaboram para o projeto de visualização. Cada ferramenta pode ser considerada uma caixa preta que deve necessariamente gerar um resultado em determinado formato para que a próxima ferramenta possa visualizá-lo. Para que a ferramenta de visualização proposta neste trabalho funcione, podemos substituir todas as outras ferramentas, desde que, o conjunto de novas ferramentas gere o modelo no formato necessário de entrada para o módulo de visualização.

O sistema completo foi desenvolvido inicialmente em cinco módulos:

- Módulo Alimentador: Este módulo é responsável por inserir novos arquivos para serem visualizados.
- Módulo Indexador: Este módulo pega os novos arquivos e os indexa na base de dados gerando um mapa de documentos e termos.
- Módulo Comparador: o módulo é responsável por gerar um resultado comparativo entre todos os documentos dois a dois.
- Módulo Agrupador: Esta ferramenta analisa os resultados comparativos dos documentos e gera os *grupos* de documentos similares que darão base para a representação segundo o modelo de visualização.
- Módulo Visualizador: A ferramenta de visualização é responsável por entender os grupos e gerar as informações para o usuário final, seja ela em formato texto ou em formato de gráfico em três dimensões.

Veremos agora cada um dos módulos e as características de implementação de cada um deles.

5.1. Módulo Alimentador

Este módulo é o responsável por adicionar novos documentos no acervo digital que será manipulado e visualizado. Em nosso protótipo este módulo foi desenvolvido para buscar arquivos no formato XML(*Extensible Markup Language)*[41] associados com *RSS(Rich Site Summary)* de portais de notícias.

O RSS foi desenvolvido em conformidade com a especificação *Resource Description Framework* (RDF) do *World Wide Web Consortium* (W3C)[41]. A tradução da sigla RSS, no entanto, é controversa. A própria *Netscape*, que criou o RSS para ser usado no portal *My Netscape Network*, rebatizou o formato para *Rich Site Summary*, ao passar da versão 0.9 para a versão 0.91 e incorporar novos elementos, alheios ao *RDF*. Há ainda quem chame o *RSS* de *Really Simple Syndication.*

A ferramenta foi desenvolvida visando obter conteúdo de um conjunto diverso de fonte de dados com o objetivo de permitir uma maior variedade de temas. Com o uso de RSS para obter o conteúdo a ser analisado, temos um mecanismo de comparação de nossos resultados finais de pesquisa com o processo de classificação humana de conteúdo, visto que os textos apontados pelos *RSS's* são pré-classificados em canais de notícias.

A ferramenta alimentadora busca os *RSS's* cadastrados num banco de dados, e onde a pré-classificação é registrada no ato do cadastro do endereço de *RSS*. Todos os RSS's são então baixados e têm seus conteúdos atualizados são lidos. Durante a leitura todos os endereços de notícias apontados no RSS têm suas páginas baixadas e todas as *flags* HTML*(HyperText Markup Language)* são limpas, permanecendo apenas um arquivo com um conteúdo textual referente a notícia que se deseja indexar futuramente. Cada arquivo é então, cadastrado no banco de dados para posterior indexação pelo módulo Indexador. A figura 30 apresenta o fluxo do processo alimentador.

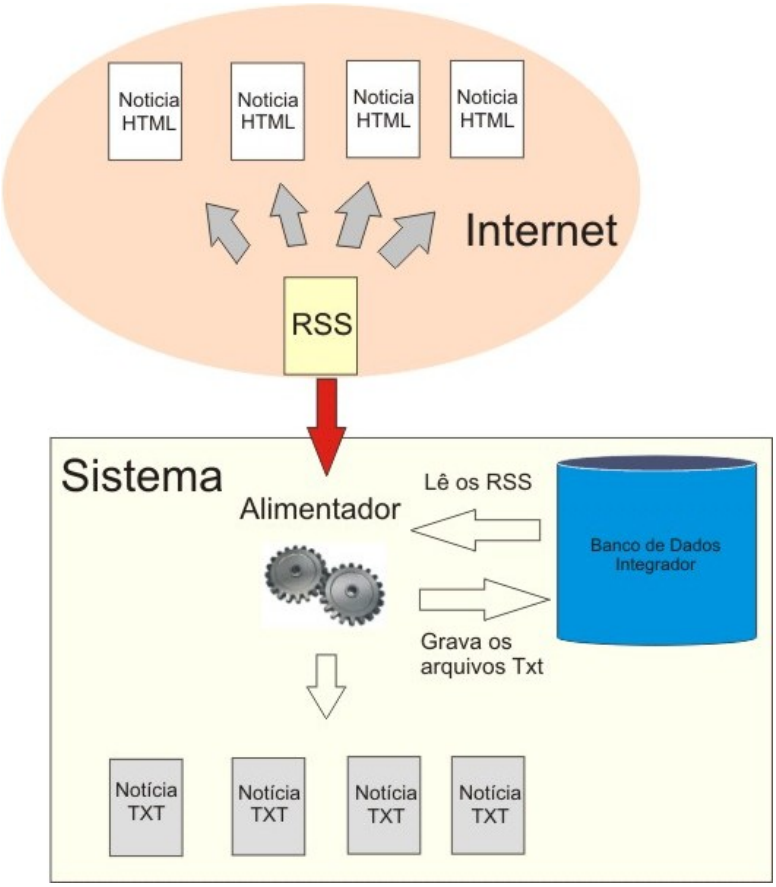


Figura 30- Fluxo do módulo alimentador

Como protótipo este modelo funcionou bem e atendeu as necessidades de captura de documentos. Como trabalhamos com *sites* onde os endereços indicados nos RSS são bloqueados a partir de uma data específica de acordo com a fonte, tivemos que carregar todos os conteúdos para arquivos locais, simulando com isto, um ambiente de biblioteca digital.

O protótipo foi desenvolvido em PHP 5.0 e é acessado através de comando de linha sem a necessidade de uso do navegador. A aplicação lê os *RSS's* cadastrados no banco de dados *Mysql Server 5.0.1* gerando os arquivos lidos numa pasta do sistema. Esta ferramenta foi testado sobre a plataforma *Linux Ubuntu 5.0* e *Windows 2000 Server*, funcionando sem restrições em ambas plataformas.

Outros tipos de mecanismos de busca como softwares robôs que varrerão a Internet em busca de documentos a serem indexados podem assumir o papel de ferramenta de alimentação. Neste caso, como o volume de documentos seria muito grande, o sistema pode

armazenar os endereços e o mecanismo de indexação faria a leitura diretamente das fontes sem criar uma cópia local do documento.

Num ambiente de biblioteca, este módulo pode ser substituído por uma ferramenta de cadastro e controle de acervo. Isto quer dizer que a alimentação de arquivos pode ocorrer manualmente de forma que o usuário cadastre os dados de um documento e insira o arquivo digitalizado na biblioteca digital. Utilizando este recurso o sistema pode obter informações do documento que, no modelo atual de busca, não são possíveis como por exemplo: nome de autores e referências bibliográficas, ano, editora e outros. Desta forma os arquivos cadastrados pelo usuário seriam a entrada para o módulo 2, que é responsável pela indexação.

5.2. Módulo Indexador

A ferramenta de indexação é a responsável por realizar a varredura dos arquivos informados pelo módulo de alimentação armazenar no banco a relação existente entre um documento e um termo ou palavra. Esta relação é determinada pelo número de ocorrências da palavra no documento lido.

Como esta ferramenta acessa um documento através de um endereço apontado pelo alimentador, se este endereço for um endereço na Internet, o indexador também poderá realizar a leitura, no entanto o processo se tornará significativamente mais demorado.

Para encontrarmos as relações documentos-termos significativas, o indexador tem que ser capaz de identificar e remover do processo todas as palavras que podem distorcer o processo de comparação entre os documentos. Estas palavras removidas são os chamados *stop-words* que vimos no capítulo 3.

Outra consideração que fazemos nesta ferramenta é com relação à “compressão semântica”. Como num processo de classificação semântica a busca por palavras chaves não nos retorna todos os documentos associados ao assunto e sim aos termos, fazemos um agrupamento semântico de palavras. Agrupamos sobre uma mesma palavra a representação de palavras que se encontram com diferenças de gênero, número, grau, gerúndios e até mesmo palavras distintas com o significados similares. Podemos dar o seguinte exemplo:

Um primeiro tipo de agrupamento pode ser exemplificado com a palavra *"Atores"* que representa sobre seu texto as palavras *"Ator"*, *"Atriz"*, *"Atrizes"*.

Outro exemplo pode ser dado pela palavra *"Furacão"*, que agrupa as seguintes palavras: *furacões, tufão, tufões, tempestades, tempestade.* Este significados foram agrupados utilizando o sentido real e não metafórico expresso no dicionário Aurélio[11]. Vejamos a tradução obtida em [11]:

Furacão:
. Ciclone (2) que se forma nas regiões do Atlântico Norte, do mar do Caribe, do golfo do México e na costa nordeste da Austrália, e no qual a velocidade dos ventos pode atingir até 300km/h: & [Cf. tornado e tufão.]
. Fig. Tudo que destrói com violência e rapidez; turbilhão, vórtice.
. Fig. Ímpeto muito veemente

Tufão:

. Ciclone (2) que se forma nas regiões oeste e norte do Pacífico, e sul do mar da China.

[Cf.furacão e tornado.]

.Vento fortíssimo e tempestuoso; vendaval.

. V. pé-de-vento (2). [Cf. furacão e tornado.]

Como estamos preocupados numa busca semântica, qualquer uma das palavras agrupadas, quando aparecer num documento, colocará um ou mais documentos no mesmo contexto no trecho onde elas aparecem.

Um usuário ao pesquisar por *Tufão,* de acordo com seus significados, não deverá surpreender-se se lhe aparecer um documento falando sobre *Furacão* ou vice-versa, principalmente quando este conceito pode não estar bem definido na mente do usuário.

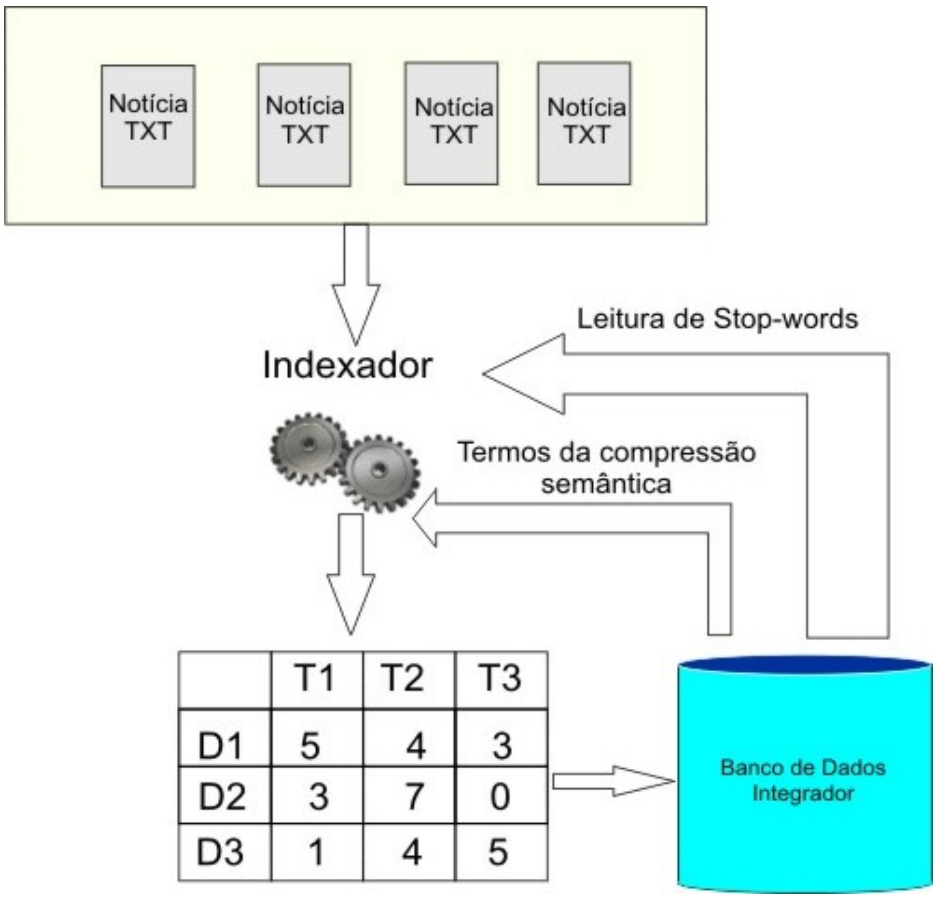


Figura 31 - Fluxo de Dados do Indexador

Para realizarmos a compressão semântica na indexação criamos manualmente um cadastro de palavras e dos grupos que elas representam. De posse deste cadastro, o sistema ao indexar as palavras, substitui as palavras do grupo pela palavra que o representa na montagem do índice que será montado para comparação de documentos. Este mesmo cadastro é utilizado para entender as palavras que os usuários informam na ferramenta de visualização, que falaremos adiante. Isto reduz o espaço semântico da indexação e aproxima documentos similares que ficariam afastados numa comparação simples de palavras.

Uma vez selecionadas apenas as palavras consideradas importantes semanticamente o sistema indexador deve gerar o resultado em forma de uma tabela onde as colunas são o número de palavras e as linhas os identificadores dos documentos montando uma matriz de *documentos x termos* que será utilizada pelas demais ferramentas do sistema. Na figura 25 temos o modelo da ferramenta indexadora.

A implementação do indexador foi realizada em C/C++ e roda sobre a plataforma Linux(Ubuntu 5.0). A indexação de 423 documentos utilizado no experimento foi executada com e sem compressão semântica gerando resultados distintos de desempenho. A tabela abaixo mostra o desempenho da aplicação em ambos os casos.

Resultado de Execução do Indexador		
	Sem compressão	Com compressão
N° de Documentos	423	423
N° de Termos gravados	8356	7389
Tempo de Indexação	56 s	87 s
Tempo de Gravação no Banco	71 s	58 s
Duração total do processo	127 s	145 s

Tabela 11-Resultado demonstrativo de desempenho do Indexador

O resultado de desempenho do indexador mostra que a indexação utilizando compressão semântica é mais demorada que a indexação normal. No entanto, considerando todo o processo de preparação dos documentos para visualização, inclusive os módulos posteriores a este, o tempo final torna-se consideravelmente menor que o processo realizado sem compressão semântica. Estes resultados serão mostrados mais adiante neste capítulo.

5.2.1) Indexação por citação

Um modelo de indexação proposto em [45] não utiliza nenhum mecanismo de mapeamento dos termos digitados. Este modelo de indexação baseia-se na contagem de citações às referências bibliográficas e geração de uma matriz *documentos X referências bibliográficas* onde cada posição da tabela corresponde ao número de vezes que uma referência foi citada no documento.

Esta técnica nos oferece um mesmo modelo de matriz que será enviada para o módulo comparador. Em [45] a ferramenta de comparação também é o *LSI*, utilizado neste trabalho. A técnica de análise de citação pode ser uma ferramenta capaz de ajudar a resolver problemas de classificação que, apenas com a análise semântica não foi possível resolvermos. Sendo com isto um método complementar a análise semântica.

No modelo baseado em citações nos permite mapear redes de conhecimento que se propagam na linha do tempo de autor para autor. Como nosso objetivo neste trabalho é propor uma ferramenta de visualização que seja independente do mecanismo de indexação e

classificação, podemos substituir toda a estrutura de entrada do módulo de visualização, que hoje é semântica, pelo modelo de citações. Com isto teríamos a representação das ligações de conhecimento entre autores indicadas no MOGRIS de forma mais acentuada e confiável. E um conjunto de documentos de menor volume classificado.

Contudo, um dos grandes problemas do modelo de indexação por citações está na dificuldade em obtermos tais informações. Em [45] o teste de classificação foi realizado em 70 documentos, onde para montar a matriz inicial foi necessário realizar uma contagem manual das citações em cada documento utilizado no experimento. Diante disto torna-se muito difícil a classificação de um documento utilizando esta técnica sem a criação de uma ferramenta capaz de identificar os mais diversos tipos de citação e contá-las automaticamente.

5.3. Módulo Comparador

Uma vez gerada a matriz de *documentos x termos*, uma outra ferramenta entra em ação o Comparador. Este comparador é responsável por comparar e identificar o grau de similaridade entre dois documentos. De posse desta matriz utilizamos a fórmula de peso de um termo *i* no documento para obtermos uma nova matriz de *documento x pesos_dos_termos.* A partir desta matriz o grau de similaridade pode ser calculado utilizando-se qualquer mecanismo de comparação como LSI, probabilístico, vetorial, redes neurais, dentre outros. A figura 31 nos mostra a arquitetura do processo de comparação de documentos.

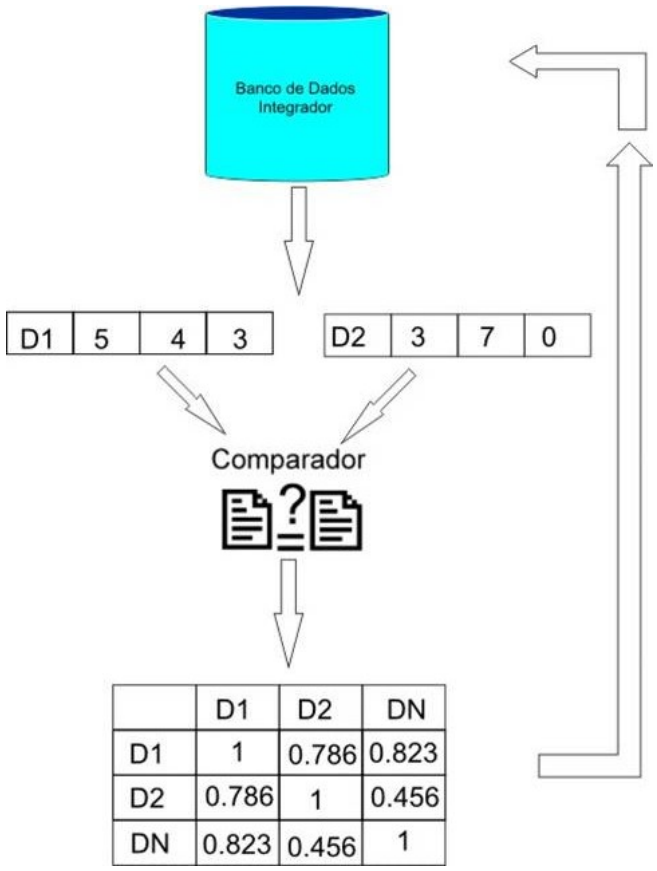


Figura 32- Arquitetura do módulo Comparador

Em nossa ferramenta optamos pelo uso do modelo LSI por dois motivos:

- A existência de uma ferramenta capaz de realizar o cálculo matemático principal deste modelo o *DVS(Decomposição de Valores Singulares)* já apresentado no capítulo 3. A ferramenta GSL*(GNU Scientific Library)*[42] é uma biblioteca matemática para programas em C e C++ capaz de realizar cálculos vetoriais e matriciais necessárias para a implementação do DVS.

- O modelo LSI nos permite realizar cálculos semânticos comparativos entre dois documentos gerando um resultado que nos permite comparar cada documento e agrupá-los com facilidade segundo critérios pré-definidos sem que o usuário interfira no processo de classificação.

- A entrada para o LSI é facilmente implementada indexando os documentos e gerando uma matriz bidimensional.

- Sua saída é uma matriz com índices de comparação entre cada um dos documentos envolvidos no processo. Este índice de comparação é então lido e interpretado pelo próximo processo com o objetivo de criar os agrupamentos de documentos.

5.4. Módulo Agrupador

O mecanismo de categorização de documentos é o módulo responsável por entender os dados comparativos de documentos armazenados numa matriz *documentos X documentos* e gerar grupos de documentos similares. Para realizar o agrupamento, nosso algoritmo de agrupamento necessita de um fator comparativo fixo que limitará o grau de similaridade permitido dentro da categoria.

Os valores do fator de classificação estão contidos no intervalo de 0 a 1, o que significa percentualmente valores de 0 a 100 por cento de semelhança entre documentos. Este valor pode variar de acordo com a margem de erro de classificação aceita pelo organizador do acervo. Em nosso trabalho com 423 documentos, utilizamos um fator de 80% que nos deu uma margem de erro 0.

Este valor pode ser calibrado pelo usuário através de análises feitas sobre o acervo classificado. Caso perceba-se que o número de documentos que divergem da categoria está alto, o processo pode ser executado novamente com um fator de agrupamento mais alto. Isto reduz o número de documentos classificados, mas reduz a margem de erro de documentos classificados equivocadamente.

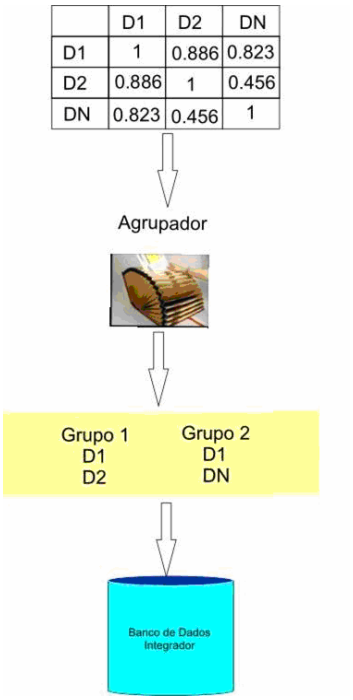


Figura 33- Estrutura de Categorização de documentos

A lógica de classificação de documentos é exibida no item 4.1 no capítulo 4. A figura 33 nos mostra o fluxo de dados no sistema. Observe que o módulo exige apenas uma matriz comparativa de documentos. Isto permite que todo o processo de alimentação, indexação e comparação possam ser substituídos sem impacto para os passos seguintes, desde que esta matriz seja gerada como resultado para o Agrupador. Após a execução do processo de agrupamento, temos todas as categorias de documentos contendo documentos cujo grau de similaridade entre eles é maior ou igual ao grau de similaridade aceito pelo usuário e indicado anteriormente.

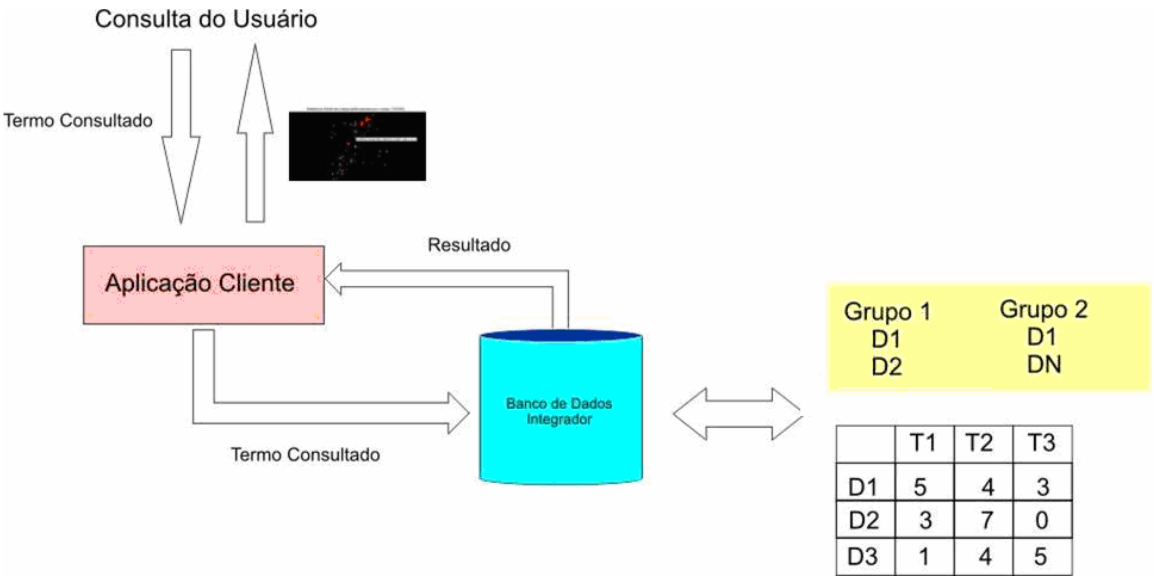
5.5. Módulo Visualizador

O módulo visualizador, em nosso trabalho, é o responsável pela exibição de dados em 3D. É este módulo que entenderá os grupos gerados e seguindo especificações próprias exibirá os dados de acordo com uma especificação definida.

Neste módulo, os dados de entrada devem estar em um acervo digital indexado onde possa ser identificada a relação *documento X termo* e onde cada arquivo que compõe o acervo possua um cadastro básico contendo título, autores, ano, editora. É importante lembrar que

quanto mais informações sobre um documento existirem cadastradas no sistema, maior é a gama de informações que podem ser representadas através de um mecanismo de visualização.

Outro elemento de entrada para nosso mecanismo de classificação está no resultado de um processo de classificação e agrupamento. Para nosso mecanismo de visualização, proposto no capítulo 4, devemos ter uma lista de grupos e documentos que os compõem. Esta lista nos permite criar os universos de visualização em 3D e listas de links agrupados.



	T1	T2	T3
D1	5	4	3
D2	3	7	0
D3	1	4	5

Figura 34 - Módulo Visualizador

Em nosso trabalho, desenvolvemos como Visualizador, uma ferramenta de consulta para internet. Esta ferramenta permite ao usuário digitar um termo a ser consultado e receber como resultado uma listagem de links desejados ou uma imagem em 3D no qual o usuário poderá navegar.

Ao fazer a consulta, o sistema encaminha os dados a um servidor Web que processa a consulta, identificando os documentos que são de interesse do usuário utilizando o índice *documentos X termos*, a lista de categorias e os dados do documento. Após a montagem do resultado da consulta, o sistema modela este resultado para um modelo de visualização. Este processo foi desenvolvido em PHP 5.1.4 e Mysql 5.0 e testado sobre um servidor Apache 2.2 sobre Windows 2000 Server.

O mecanismo de visualização em 3D foi gerado através da linguagem *VRML(Virtual Reality Modeling Language)* versão 2000. A VRML é um método utilizado para exibição de objetos em 3D na web.

Para uso do VRML em navegadores é necessário instalar um *plugin*. Para os testes do sistema utilizamos o *Blaxxun Contact 5.3*[44] sobre um navegador Internet Explorer 6.0.

5.6. Modelo de Dados do SiGRIS

5.6.1) Modelo de dados do Alimentador

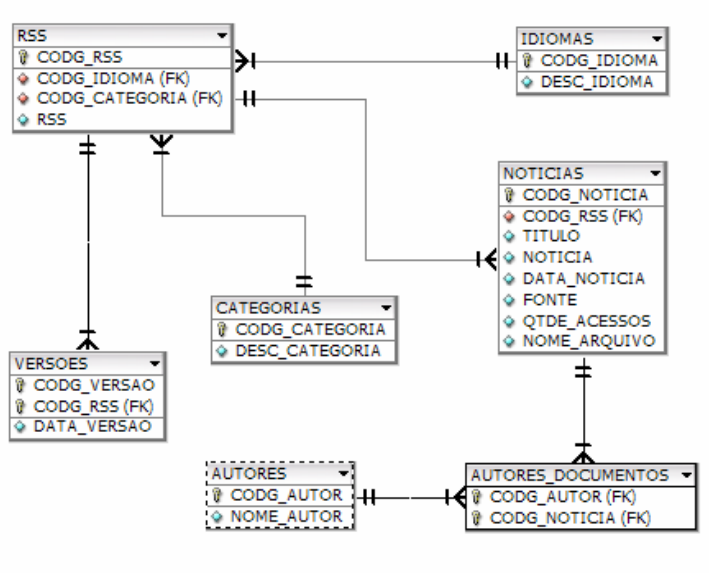


Figura 35. Visualização do MOGRIS

A figura 35 nos mostra o modelo de dados utilizado pelo módulo alimentador. A tabela RSS armazena o endereço e categoria do RSS. Esta categoria será utilizado como referência para cada registro de notícia contido na tabela NOTICIAS e será utilizado ao final do processo para conferência de grau de correção do processo de classificação realizado pelo SIGRIS. Cada notícia, ou documento, pode conter um ou vários autores. É importante ressaltar que o modelo de dados já está preparado para aceitar documentos de diferentes idiomas.

5.6.2) Modelo de dados do Indexador

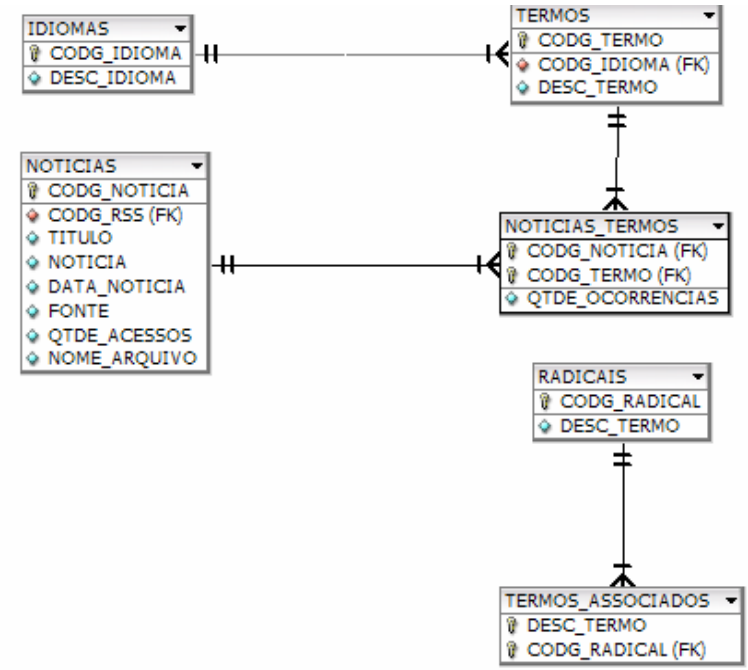


Figura 36 - Modelo de dados do Indexador

Uma vez obtidos os documentos, o sistema lê cada registro de notícia e busca o arquivo a ser indexado e que se encontra indicado no campo NOTICIAS.NOME_ARQUIVO, conforme figura 36.

De posse do arquivo o indexador grava na tabela TERMOS as palavras não existentes e cria uma associação com o documento através da tabela NOTICIAS_TERMOS. Os termos que serão gravados na tabela TERMOS são todos aqueles que se não se encontram na tabela TERMOS_ASSOCIADOS ou então gravará um termo associado à tabela RADICAIS.

Após o processo de indexação, todos os documentos possuem a contagem dos seus termos registrada no sistema.

5.6.3) Modelo de dados do Comparador

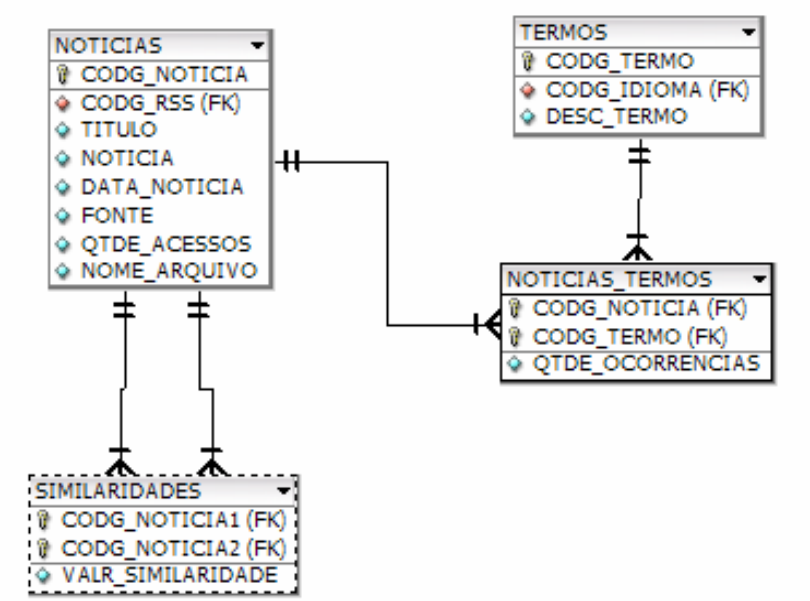


Figura 37-Modelo de Dados do Comparador

A figura 37 nos mostra a base de dados que deve ser utilizada pelo sistema comparador. Com estas tabelas o sistema monta a matriz *documentos X termos* e posteriormente calcula a similaridade entre cada um dos documentos armazenando-os na tabela SIMILARIDADES. Esta tabela será posteriormente consultada pela estrutura de agrupamento para criação das categorias.

5.6.4) Modelo de dados do Agrupador

De posse das comparações dois a dois de documentos, o módulo Agrupador acessa a tabela de similaridades e cria os grupos de documentos. Estes grupos são armazenados nas tabelas DOCUMENTOS_GRUPOS e GRUPOS que estão associadas a tabela de notícias. A figura 38 nos mostra o modelo das tabelas.

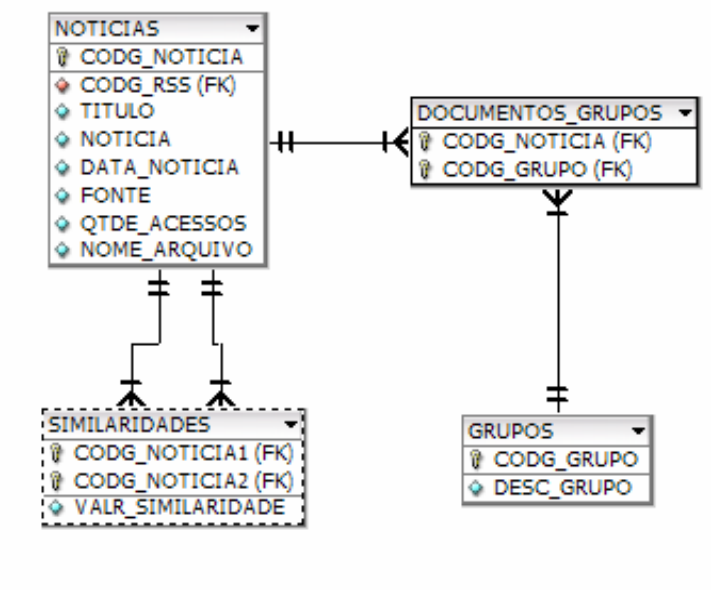


Figura 38- Modelo de Dados do Agrupador

5.6.5) Modelo de dados do Visualizador

O modelo de dado acessado pela ferramenta de visualização, exibido na figura 39, nos permite verificar que o modelo de visualização acessa dados de todas as fases anteriores do projeto. Com isto temos um sistema com módulos integrados e que devem trabalhar em conjunto para um objetivo em comum sobre a mesma base dados.

A ferramenta de visualização busca dados das notícias(documentos), seus termos, seus grupos e monta um modelo em 3D ou um modelo de link. Nestes mdoelos, conforme o usuário navega pelos documentos o campo NOTICIAS.QTDE_ACESSOS é atualizado contabilizando um novo acesso e modificando as próximas visualizações na característica que representa o número de acessos ao documento.

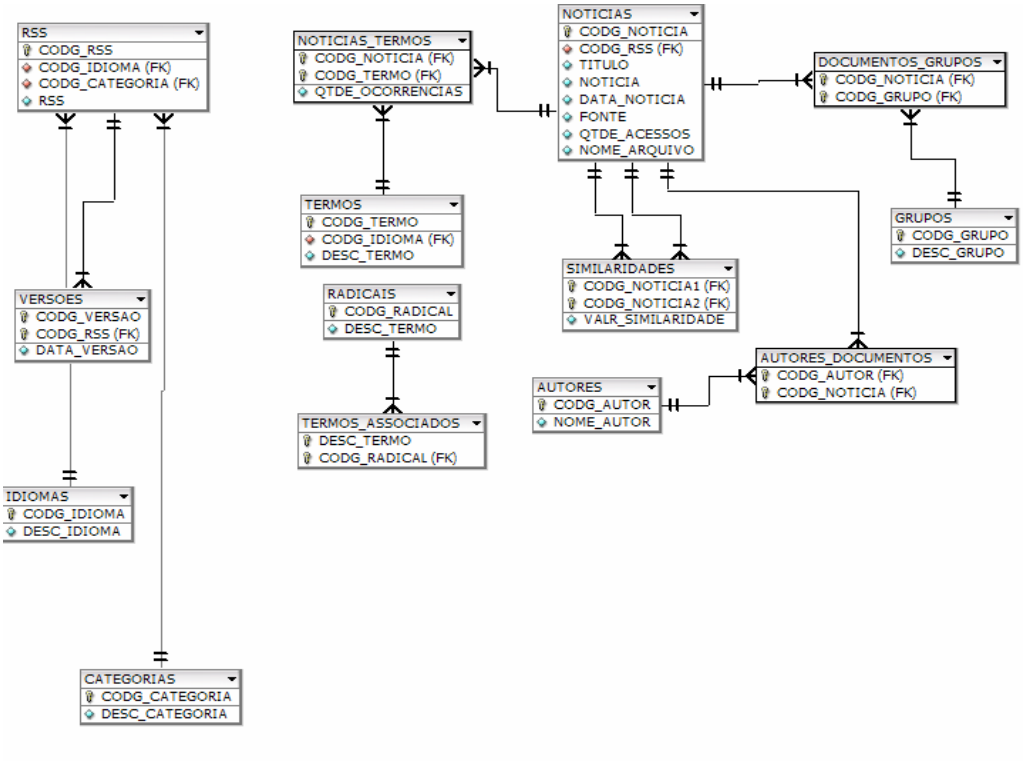


Figura 39- Modelo de Dados do Visualizador

5.7. Conclusão

Neste capítulo exibimos a arquitetura do sistema e o modelo de dados que permite nos permite a realização dos experimentos desejados. Uma das grandes facilidades oferecida por este modelo está em sua modularização. Isto permite ao desenvolvedor alterar cada uma das partes de forma independentes.

Na implementação do visualizador uma grande facilidade é encontrada na implementação das consultas ao banco. Cada consulta pode ser alterada seguindo as regras definidas pelo projetista sem a interferência no modelo de interface proposto.No próximo capítulo apresentamos os resultados desta implementação e o modelos gerados através da análise do acervo proposto.

Capítulo 6

Resultados Obtidos

Descreveremos nesta seção os resultados obtidos e os dados que foram utilizados como massa de teste para nosso sistema. Todos os testes foram executados em um micro dotado de um processador AMD- Duron 1.7 MHz e 512 MB de memória RAM. Os teste foram divididos em duas partes: uma parte executada em sistema Linux UBUNTU 5.0, utilizando como servidor *web* o Apache 2.0, e banco de dados Mysql 5.0; a segunda parte correspondente a aplicação de consulta *web* foi testada sobre uma plataforma Windows 2000 Server com o servidor *web* Apache versão 2.2 e o banco de dados Mysql 5.0. Em ambas as fases foi utilizado a linguagem PHP versão 5.1.4.

Nossos experimentos tiveram inicio com a leitura de 423 notícias retiradas de canais de RSS do portal Terra(www.terra.com.br/rss) pré-classificados em 7 categorias e um RSS com conteúdo de informática obtida no site da revista Info(http://info.abril.com.br/aberto/infonews/rssnews.xml). Como cada RSS possui num determinado instante cerca de 10 notícias, fizermos várias leituras dos arquivos em momentos distintos nos dias 4 a 19 de janeiro de 2006, onde obtivemos um número de 423 notícias que serão indexadas, comparadas e classificadas para posterior visualização. A tabela 12 nos indica os endereços e o número de notícias(documentos) que foram obtidos.

Categorias	**RSS**	**Nº de Documentos**
Moda	http://rss.terra.com.br/0,,EI1119,00.xml	72
Ciência	http://rss.terra.com.br/0,,EI238,00.xml	80
Vasco-Futebol	http://rss.terra.com.br/0,,EI1979,00.xml	47
Música	http://rss.terra.com.br/0,,EI1267,00.xml	71
Automobilismo	http://rss.terra.com.br/0,,EI1874,00.xml	20
Arte e Cultura	http://rss.terra.com.br/0,,EI3615,00.xml	39
Flamengo- Futebol	http://rss.terra.com.br/0,,EI1977,00.xml	57
Informática(Revista info)	http://info.abril.com.br/aberto/infonews/rssnews.xml	37

Tabela 12-Número de documentos por categoria de RSS

6.1. Indexação

Uma vez obtido todos os documentos pelo módulo alimentador, executamos o processo de indexação destes documentos. O processo de indexação foi realizado de duas

maneiras, em primeiro sem compactação semântica e posteriormente com compactação semântica.

O processo que chamamos de compactação semântica representamos com uma única palavra um conjunto de palavras associadas como palavras que possuem diferenças de gênero, número, grau e palavras derivadas. Criando um conjunto inicial, nosso sistema é capaz de representar 2357 palavras com um universo de 661 palavras. A tabela 13 representa os tempos de execução dos processos e o número de termos indexados.

	Sem redução	**Com redução**
Nº de Documentos	423	423
Nº de Termos	8356	7389
Tempo(s)	61s	52s

Tabela 13- Resultado da Indexação

A redução de 9 segundos no tempo de indexação pode parecer pequena para um processo de indexação. No entanto, a redução em quase 1000 termos indexados, provocado pelo redução, fará com que os tempos de cálculos das matrizes de similaridades no próximo módulo seja reduzido significativamente como veremos a seguir.

6.2. Comparador e Agrupador

Após feita a indexação com redução o processo de comparação é executado utilizando a matriz *documento X termos* e posteriormente é criada a matriz *documento X pesos_termos*, conforme citado nos capítulos 3 e 4. Com isto, a matriz gerada pela indexação possui dimensões de *423 x 7389.*

De posse desta matriz, executamos o processo de comparação utilizando o LSI . Ao término do LSI, o algoritmo de agrupamento é executado para alguns valores de similaridade que variam de 95% a 75%. A tabela 14 nos mostra alguns dados coletados e que explicaremos a seguir.

% de similaridade mínimo	**95**	**90**	**85**	**80**	**75**
N° de Documentos	423	423	423	423	423
N° de Categorias	151	292	337	383	422
N° de Documentos agrupados	181	282	332	366	376
% de Erro	0%	0%	0%	0%	3%
% de documentos Classificados	43%	67%	78%	87%	89%

Tabela 14 - Resultado obtidos para faixas de similaridades distintas

Através da tabela 14 definimos para nosso projeto que a entrada do nosso mecanismo de visualização seria a matriz de similaridade com margem mínima de similaridade maior ou igual a 80% . Ao utilizarmos taxas de similaridade muito alta, restringimos o grau de similaridade de documentos que podem pertencer a mesma categoria, reduzindo com isto o número de documentos agrupados. Por outro lado quando baixamos a 75% o número de documentos classificados aumenta, porém, surge uma margem de erro de 3 por cento. Consideramos como erro todo o documento que foi classificado de forma diferente da classificação humana definida pelos canais de RSS e que após análise textual do mesmo, não foi encontrada relação real com os demais documentos da categoria.

Quando analisamos os documentos classificados com uma margem igual ou superior 80 por cento de similaridade, encontramos alguns documentos agrupados com outros de categoria distinta segundo a classificação humana. O documento *"(Pearl Jam Lança Livro de fotos em seu site)"* que estava num canal de música do terra, continua sendo agrupado com documentos que também são do canal de música. No entanto, nosso mecanismo de classificação também o enquadrou no grupo de documentos que eram originários do canal Arte e Cultura e que tratavam exclusivamente de literatura. Um outro caso de encontro de identificação correta feita pelo sistema ocorreu com a notícia *"(SANDISK lanca player para competir com IPOD Nano)"*. Esta notícia estava no canal música. Nosso mecanismo de classificação conseguiu agrupá-la também com notícias do canal Informática proveniente de outra fonte de RSS que é a Info.

Outro agrupamento de documentos de canais distintos ocorreu em Vasco-Futebol e Flamengo-Futebol que se misturaram em alguns grupos criados pelo sistema. Esta mistura já era esperada visto que ambos os canais tratam do assunto Futebol, porém diferenciando apenas nos clubes.

Um elo não esperado mas real que encontramos, foi entre os canais moda e futebol. O documento *"(Ronaldinho irá lançar marca de artigos esportivo...)"* é proveniente do canal Moda e foi classificado juntamente com alguns documentos de futebol. Como Ronaldinho está intimamente ligado ao esporte e principalmente a futebol, a ligação não pode ser considerada como um elo errado.

Os demais documentos classificados se juntaram com documentos do próprio canal de RSS. O sistema foi capaz de criar grupos mais específicos dentro dos próprios canais, como

por exemplo, um grupo que contém apenas documentos relacionados a literatura, onde todos são provenientes do canal Arte e Cultura.

Podemos afirmar pela classificação realizada que todos os documentos classificados estão em grupos com documentos similares. No entanto, com o valor de similaridade 80% , não conseguimos agrupar alguns documentos automaticamente com pelo menos mais um documento. Este número foi de 56 documentos, dos 423 documentos utilizados, o que equivale a cerca de treze por cento do acervo total.

Com relação ao tempo de processamento nesta etapa, todo o cálculo de similaridade e agrupamento durou cerca de 17 minutos e 20 segundos. Com isto, o tempo total do processo de Indexação, cálculos de similaridade e atualização do banco demorou cerca de 19 minutos. Depois de realizado todo este processo o banco de dados fica disponível para consultas pelo SIGRIS. O tempo de consulta e montagem de uma tela em 3D gira entre 5 e 10 segundos, dependendo da capacidade de processamento do servidor *web* e do servidor de banco de dados. Caso a consulta já tenha sido realizada por algum usuário e esteja ainda no *buffer* do servidor de banco de dados o retorno é quase que imediato.

6.3. Ajuste para o modelo de visualização de autores

Nosso modelo de visualização, apresentado no capítulo 4, nos permite obter várias informações sobre os autores dos documentos e os relacionamentos existentes entre eles. Porém, como em nosso experimento não utilizamos artigos técnicos ou outro documento que nos permitisse identificar o autor, tivemos que simular a existência de autores nas notícias dos canais. Para isto criamos um conjunto de autores que fictícios e os distribuímos pelos documentos, de forma que um artigo poderia ter no máximo 3 autores e um autor poderia ter escrito no máximo 5 artigos aleatoriamente. Com isto criamos uma malha de influências que utilizaremos para exibir nossos modelos.

6.4. Visualização do MOGRIS

O mecanismo de visualização está contido numa página de internet onde o usuário digita o termo a ser consultado. Inicialmente o protótipo realiza a busca de apenas um termo, o que pode ser ajustado numa versão futura. Ao digitar o termo da consulta e clicar em

"Pesquisar" o SIGRIS busca todas as categorias que possuem pelo menos uma ocorrência do termo digitado ordenando por número de ocorrências dos termos. Isto significa contar todas as ocorrências do termo nos documentos que compõem a categoria. Com este número e quantidade de documentos de uma categoria calculamos o grau de relevância das categorias (raio do planeta com relação ao centro-termo) conforme especificado no capítulo 4.

Uma busca pelo termo *Futebol* nos retornou a imagem da figura 40.

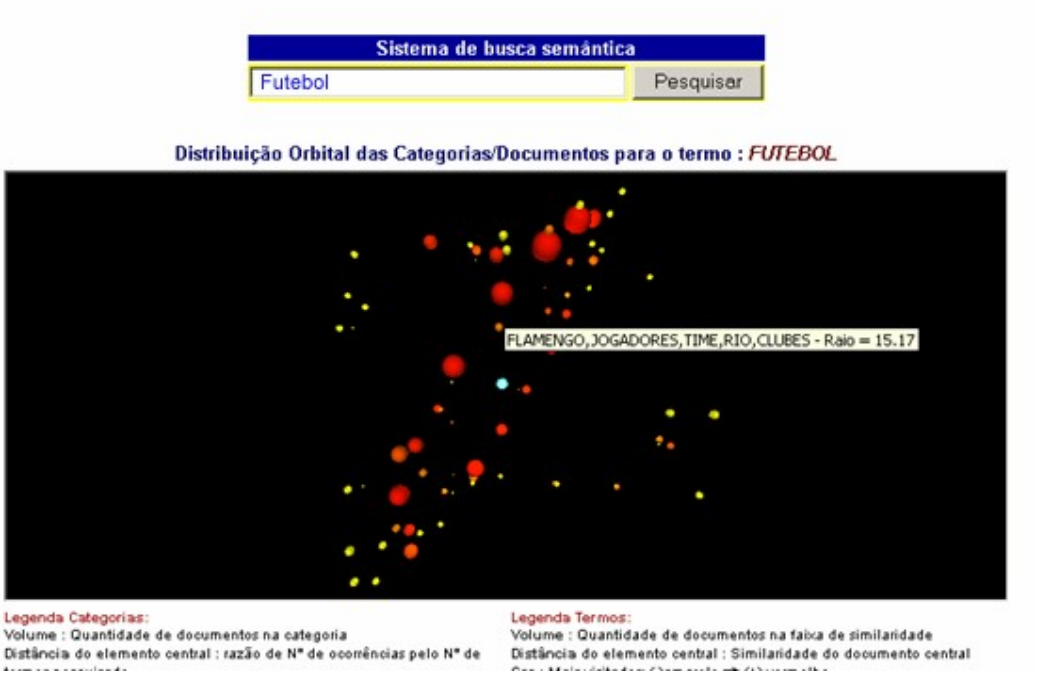


Figura 40- Resultado em 3D da Consulta de categorias sobre Futebol

Podemos perceber na imagem que ao centro da imagem, ou universo de categorias, existe uma esfera azul. Esta esfera representa o termo "Futebol". O afastamento dos planetas em relação ao centro significa o grau de relevância deste grupo com relação ao termo consultado. O volume de cada esfera representa uma quantidade maior ou menor de documento na categoria. Podemos perceber que as categorias(planetas) possuem cores que variam do vermelho ao amarelo. Com isto podemos dizer que os planetas mais vermelhos tiveram um número maior de acesso que os planetas mais claros. Isto pode nos ajudar a encontrar as categorias mais relevantes segundo a classificação do usuário.

Cada categoria, como podemos ver na figura 40, possui um texto visível quando passamos o mouse sobre a mesma. Este texto possui as 5 palavras chaves mais relevantes do grupo por ordem de importância e o raio de afastamento do centro do sistema.

Os grupos, representados por esferas dentro do universo, são elos de ligações com os sistemas planetários (nível de planetas). Ao clicarmos numa das categorias, o sistema exibe todos os documentos da categoria, em torno do termo consultado. Como os grupos são montados através de comparação semântica, alguns documentos do grupo selecionado podem não conter a palavra consultada, mas certamente terá seu assunto no contexto de futebol.

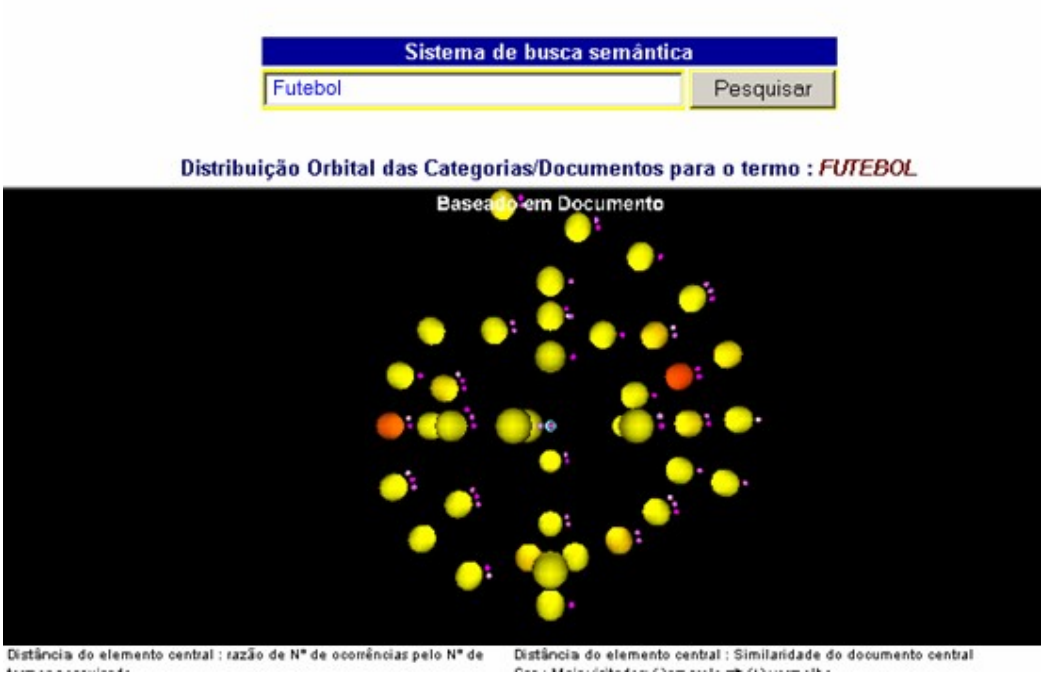


Figura 41- Segundo nível da consulta pelo termo Futebol.

No segundo nível, exibido na figura 41, podemos observar que cada planeta(documento) é orbitado por luas(autores). Cada lua é um *link* para visualizarmos informações sobre os autores do documento. As luas também podem assumir também variações de tamanho de acordo com o número de documentos publicado pelo autor que ela representa. A cor das luas mais brancas ou mais roxas indicam autores com menos ou mais autores na mesma área de interesse, respectivamente. A figura 42 nos apresenta uma aproximação da imagem a exibindo com mais detalhes os planetas.

Figura 42 - Aproximação do sistema de documentos

Ainda no segundo nível, ao clicarmos sobre um planeta abrimos o documento para leitura, no entanto, ao clicarmos sobre uma lua, mudamos o foco de nossa pesquisa de *Futebol* para autores de futebol relacionados com o autor clicado. A figura 43 nos mostra o resultado da navegação onde o autor selecionado passa a ser o foco da navegação, ou seja, a esfera azul ao centro do sistema que será a referência dos demais planetas, neste caso representando autores. Como explicado no capítulo 4, cada planeta(autor) possuirá apenas uma Lua. Ao clicarmos nesta lua, o autor clicado, passa a ser o foco da imagem e uma nova consulta é realizada. Ao clicarmos no planeta, voltamos ao nível de categorias, semelhante à figura 40. A diferença está no fato de que as categorias são todas as categorias onde o autor já publicou um documento.

Figura 43- Busca pelo autor

6.5. Visualização do MOGRIS

Neste capítulo mostramos o resultado de nossos experimentos e de nosso protótipo para obtenção de um modelo de visualização de documentos baseado em consulta a grupos de documentos classificados semanticamente. Exibimos para isto, o resultado dos processos intermediários de obtenção de documentos, comparação e classificação dos mesmos e que nos ofereceram obter uma massa de dados que nos permitiu demonstrar o modelo proposto.

No próximo capítulo descrevemos algumas propostas de trabalhos futuros baseados no conjunto de ferramentas apresentados neste projeto, e nossos comentários finais sobre este trabalho.

Capítulo 7

Conclusões

Neste trabalho apresentamos um modelo de visualização gráfica de documentos digitais e todas as atividades de tratamento de informações que devem ser executadas antes do mecanismo de visualização utilizá-las para exibição. O uso do MOGRIS em um sistema de busca oferece uma alternativa visual diferente dos modelos tradicionais de exibição de documentos. Este modelo nos permite, através da visualização de um objeto dentro de um contexto espacial, identificar características de um documento e as relações existentes entre eles, sejam elas indicadas pelo usuário(cores dos planetas) ou através de cálculos automáticos, como a distância ao centro do sistema. Outra grande vantagem do MOGRIS é que podemos com uma notação única encontrar relações entre autores dos documentos, e entre os documentos. As ligações entre autores nos permitem identificar comunidades de estudo sobre determinados temas e os autores mais representativos de uma determinada área de conhecimento.

Cada uma das atividades que desenvolvemos e que compõem o SIGRIS, sistema que utiliza o MOGRIS como modelo de visualização gráfica, podem ser evoluídas em trabalhos futuros para melhorar o desempenho do sistema nos seus diversos níveis. Citaremos a seguir alguns possíveis trabalhos que podem vir a ser executadas visando dar continuidade a este projeto.

Alimentador: trabalhamos com um mecanismo de alimentação de documentos utilizando RSS. Em trabalhos futuros este mecanismo pode ser substituído por robôs de busca na internet ou por um sistema de entrada manual de documentos. Este sistema de entrada manual pode permitir ao usuário cadastrar um conjunto de informações dos documentos e também inserir o documento digital. Esta última forma é mais adequada para acervos pertencentes a bibliotecas e que devam ser digitalizados.

Indexador: O indexador foi desenvolvido para documentos em inglês e português, no entanto, não é capaz de gerar vínculos entre documentos de idiomas diferentes. Uma ferramenta de indexação que seja capaz de gerar vínculos entre documentos de idiomas distintos seria um grande avanço para o sistema, pois permitiria a comparação e visualização de documentos dentro de uma mesma visualização. Outra melhoria que pode ser feita é no

indexador está no mecanismo de redução de espaço semântico, onde o indexador pode ser otimizado para reduzir ainda mais o número de termos necessários para representação semântica dos documentos.

Comparador: em nossa ferramenta de comparação utilizamos o LSI, como vimos nos capítulos anteriores, no entanto, novas técnicas de comparação, novos algoritmos e novas implementações podem melhorar o processo de comparação reduzindo o tempo de comparação entre eles. Hoje o sistema re-classifica todos os documentos sempre que um conjunto de novos documentos é inserido no sistema. Um mecanismo estatístico pode ser adotado de forma que, ao serem inseridos novos elementos, apenas um conjunto teoricamente mais significativos dos documentos classificados fossem utilizados. Isto evitaria que a cada novo documento, todas as categorias fossem recriadas com o processamento de um número crescente de documentos.

No SIGRIS, para efeito de protótipo, a consulta feita pelo usuário deve conter apenas um termo. No entanto, num sistema em produção, o mesmo deve ser capaz de identificar mais de um termo dentro do modelo. Como nosso objetivo era criar uma base ferramental para gerar os modelos de visualização, este é um dos elementos que pode ser alterado para permitir maior poder nos mecanismos de consulta.

A criação de um modelo de visualização de documentos nos levou a pesquisar conceitos de diversas áreas de estudo que envolvem desde a parte de tecnologias com ferramentas como XML, PHP, servidores *web*, passando por algoritmos de indexação, classificação e chegando aos modelos de visualização de documentos.

Analisando o esforço cognitivo para associação da metáfora proposta no MOGRIS com acervos ou autores representados, diversas alterações no modelo podem ser realizadas visando agregar mais informações dentro de um modelo de visualização cada vez mais simples.

O desenvolvimento do MOGRIS não visa criar um modelo que vá substituir todos os modelos de visualização já existentes. O que pretendemos foi criar um modelo que seja mais uma opção de ferramenta que possa ser inserida nos mecanismos de busca, sejam em ferramentas públicas na internet ou em ferramentas corporativas. Através deste trabalho visamos oferecer novos horizontes em ferramentas de visualização em 3D e contribuir para o avanço da área de classificação automática de documentos e representação gráfica de

documentos de forma a automatizar os processos manuais de classificação realizados por profissionais das mais diversas áreas da ciência da informação.

Referências Bibliográficas

[1] **http://www.ontoweb.com.br** , acessado em 10 de junho de 2006

[2] **http://www.numaboa.com.br/criptologia/escrita/papel.php** , acessado em 22 de fevereiro de 2006

[3] **http://www.aracruz.com.br/web/pt/curiosidades/curios_histpap.htm** , acessado em 22 de fevereiro de 2006

[4] **http://geocities.yahoo.com.br/jcc5001pt/museutelegrafo.htm**, acessado em 22 de fevereiro de 2006

[5]Chen C., "*Information Visualization Research: Citation em Co-Citation Highlights*" Drexel University,1999.

[6] Chen, C., *Visualising Semantic Spaces and Author Co-citation Networks in Digital Libraries* , Department of Information Systems & Computing, Brunel University, 1999.

[7] Noel, S. ; Chu, C. H.; Raghavan, V. *"Visualization of Document Co-Citation Count"* , Center for Secure Information Systems George Mason University

[8] Yamashita,Y. ; Yoshiko, *O. ; Patterns of Scientific Collaboration between Japan and France: Inter-sectoral analysis using Probabilistic Partnership Index(PPI)*. 10ªInternational Conference of the Society for Scientometrics and Infometrics,2005.

[9] Paula, Aldair B. *Uma Representação Gráfica do Conhecimento: Abordagem Bibliográfica* – UFES , 2005

[10] Vaughan, L., You, J. ; *Mapping Business Positions Using Web Co-link Analysis*. 10ªInternational Conference of the Society for Scientometrics and Infometrics,2005.

[11] DICIONÁRIO AURÉLIO eletrônico; século XXI. Rio de Janeiro, Nova Fronteira e Lexicon Informática, 1999, CD-rom, versão 3.

[12] Deerwester, S. ;Dumais, S.T.; Furnas, G.W; Landauer, T.K.; Harshman,R.; *Indexing by Latent Semantic Analysis* , Bell Communications Research

[13] Navega, Sergio C.; *Manipulação Semântica de Textos Os Projetos Wordnet e LSA* , Intelliwise AI Research – Agosto 2004

[14]Merkl, D. (1998). *Text classification with self-organizing maps: Some lessons learned*. Neurocomputing, 21(1--3):61--77.

[15] Hoeschi, H. C; *A nova era das ferramentas de busca*. **http://www.javafree.org/javabb/viewtopic.jbb?t=853570** – Acessado em 10 de março de 2006.

[16] Casado, E.S.; Balseiro,C.S. ; Maestro, I.I.; Pau, M.R.S ; Cuesta,J.P; *Bibliometric Study of Scientific Research on Prion Diseases Encephlopathy and Creutzfeldt-Jakob Disease, 1973-2002* . 10ªInternational Conference of the Society for Scientometrics and Infometrics,2005.

[17] Raphael, S.T ; *"Interactive Visualizations for Text Exploration Using SVG to navigate large collections of unstructured documents"*, National Institute for Technology and Liberal Education [http://www.nitle.org/] <

[18]Machado, L; *"Não há Inteligências Múltiplas"* http://www.cidadedocerebro.com.br/newsletter_inteligencias_multiplas.asp

[19] Gardner, H. "*Multiple Intelligences: The Theory in Practice*". New York: Basic Books,1993.

[20] Bruner, J. "*Going Beyond the Information Given"*, New York: Norton ,1973.

[21]Vygotsky, L. S. *"Thought and Language."* Cambridge, MA: The M.I.T. Press, 1985.

[22] Nitzke**,** J. A.; Campos, M. B ; Lima , M. F.P *"Teoria de Piaget"* **http://penta.ufrgs.br/~marcia/teopiag.htm**

[23] Baeza-Yates R, Ribiero-Neto B. "*Modern Information Retrieval"*. 1st ed. New York: Addison-Wesley; 1998.

[24] Neapolitan, R.E. "*Learning Bayesian Networks"*. New Jersey, USA: Pearson & Prentice-Hall; 2004.

[25] Haykin S. "*Neural Networks – A Comprehensive Foundation"*. Pearson Education; 1998.

[26] Kaestner, C. A. A , *"Recuperação Inteligente de Informação"* . Programa de Pós-Graduação em Informática Aplicada – PUC-Paraná.

[27] Cardoso, O. N. P. *"Recuperação da Informação"* UFLA – Universidade Federal de Lavras DCC – Departamento de Ciência da Computação

[28] Ferneda, E. *"Recuperação da Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação."* USP, São Paulo -2003

[29] Teorema de Bayes - **http://www.propus.com.br/articles/alt/1/html/x24.html** , acessado em 10 de junho de 2006

[30] Freire, Sérgio L. M.; *Aplicações do método de decomposição em valores singulares no processamento de dados sísmicos.* UFBA, 1986.

[31] Doretto,G.; Chiuso, A. ;Soatto, S. ; WU, Y.N. ; *Modeling temporal stationarity*. Vision Lab- UCLA , 2000

[32] Poole, D.; *"Álgebra Linear"* , Thomson Learning Inc. , São Paulo-SP , 2004.

[33] Cugini, J.; Laskowski, S. S.; Sebrechts, M.; "Design of 3D Visualization of Search Results: Evolution and Evaluation", *Proceedings of IST/SPIE's 12th Annual International Symposium: Electronic Imaging 2000: Visual Data Exploration and Analysis* **(SPIE 2000),** San Jose, CA, 23-28 January 2000.

[34] Cugini,J.; Piatko,C.; Laskowski,S.; *"Document Clustering in Concept Space: The NIST Information Retrieval Visualization Engine (NIRVE)"*, CODATA Euro-American Workshop on Visualization of Information and Data, Paris, France, June 1997.

[35] Cugini,J.; Piatko,C.; Laskowski,S.; *"Interactive 3D Visualization for Document Retrieval"*, Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation , ACM Conference on Information and Knowledge Management (CIKM '96), November 1996.

[36] Shen,Z.; Ogawa,M.; Teoh,S.T.; *"BiblioViz: A System for Visualizing Bibliography Information"*, 2006

[37] Havre, S., Hetzler,B., Nowell,L., *"ThemeRiver: Visualizing Theme Changes over Time"* Infovis, p. 115, IEEE Symposium on Information Vizualization 2000, 2000.

[38] Dodge, M., "NewsMaps: Topographic Mapping of Information", 2000.

[39] Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V. "V*isualizing the nonvisual: Spatial analysis and interaction with information from text documents*." IEEE Information Visualization 95, 1995

[40]Leuski, A., Allan, J. , "*Lighthouse: Showing the Way to Relevant Information"* , Center for Intelligent Information Retrieval , Department of Computer Science - University of Massachusetts Amherst, MA 01003 USA

[41]**http://www.w3.org/Xml** , acessado em 04 de março de 2005.

[42]http://www.gnu.org/software/gsl/ , acessado em 10 de abril de 2005.

[43]http://cic.nist.gov/vrml/vbdetect.html ,acessado em 20 de março de 2005

[44]http://www.blaxxun.com, , acessado em 14 de junho de 2005.

[45] Poltronieri, A., Oliveira,E., " Finding Related Articles by a Bibliometric Approach", 9th World Congress on Health Information and Libraries, Salvador - BA, Brazil, 2005

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo