Instituto Superior de Engenharia do Porto

Modelo Generalista de páginas Web para motores de pesquisa com interface 3D

Daniel Jorge Rangel Tavares Gonçalves

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática

Área de especialização em

Tecnologias do Conhecimento e Decisão

Orientador: Doutor Nuno Filipe Fonseca Vasconcelos Escudeiro

JÚRI:

Presidente:

Doutora Maria de Fátima Coutinho Rodrigues, Professora Coordenadora, Instituto Superior de Engenharia do Porto

Vogais:

Doutora Paula Maria de Sá Oliveira Escudeiro, Professora Adjunta, Instituto Superior de Engenharia do Porto

Doutor Nuno Filipe Fonseca Vasconcelos Escudeiro, Equiparado Assistente 2ºTrienio D/M, Instituto Superior de Engenharia do Porto

Porto, Novembro 2010

Dedico esta tese à minha namorada e
aos meus pais, por toda a força que me deram.

Agradecimentos

Agradeço, em primeiro lugar, aos meus pais, por todo o apoio que me deram ao longo da minha vida, e em especial, da minha vida académica. Por todos os conselhos e força que nunca recusaram dar, e o incentivo de procurar sempre fazer melhor e chegar mais longe. A ambos dedico especialmente este trabalho.

Ao meu orientador, o Engenheiro Nuno Escudeiro, pelo apoio e disponibilidade que me deu para realizar este trabalho. Muito obrigado por tudo.

A todos os meus verdadeiros amigos, que sempre me ajudaram a ultrapassar as dificuldades que foram surgindo e me apoiaram ao longo de toda a minha vida académica.

E por último, mas não menos importante, à minha namorada, Alexandra Alves, por todo o apoio, paciência, força e carinho que me deram forças para completar mais este capítulo da minha vida. Um obrigado muito especial.

Resumo

A Web tornou-se uma ferramenta indispensável para a sociedade moderna. A capacidade de aceder a enormes quantidades de informação, disponível em praticamente todo o mundo, é uma grande vantagem para as nossas vidas. No entanto, a quantidade avassaladora de informação disponível torna-se um problema, que é o de encontrar a informação que precisamos no meio de muita informação irrelevante. Para nos ajudar nesta tarefa, foram criados poderosos motores de pesquisa online, que esquadrinham a Web à procura dos melhores resultados, segundo os seus critérios, para os dados que precisamos. Actualmente, os motores de pesquisa em voga, usam um formato de apresentação de resultados simples, que consiste apenas numa caixa de texto para o utilizador inserir as palavras-chave sobre o tema que quer pesquisar e os resultados são dispostos sobre uma lista de hiperligações ordenada pela relevância que o motor atribui a cada resultado.

Porém, existem outras formas de apresentar resultados. Uma das alternativas é apresentar os resultados sobre interfaces em 3 dimensões. É nestes tipos de sistemas que este trabalho vai incidir, os motores de pesquisa com interfaces em 3 dimensões. O problema é que as páginas Web não estão preparadas para serem consumidas por este tipo de motores de pesquisa.

Para resolver este problema foi construído um modelo generalista para páginas Web, que consegue alimentar os requisitos das diversas variantes destes motores de pesquisa.

Foi também desenvolvido um protótipo de instanciação automático, que recolhe as informações necessárias das páginas Web e preenche o modelo.

Palavras-Chave: Motores de pesquisa, Interfaces gráficos 3D, Modelo generalista, Standards de Metadata.

Abstract

The Web has become an indispensable tool for modern society. The ability to access vast amounts of information available in almost the entire world is a big advantage in our lives. However, the overwhelming amount of available information becomes a problem, which is, to find the information we need in an ocean of irrelevant information. To help us in this task, powerful online search engines were created that scour the Web looking for the best results, according to its criteria, for the information we need. Currently, these search engines, use a simple format, consisting only in a text box upon which the user enters keywords about the topic he wants to search, and the results are arranged on a list of links ordered by the relevance that the engine assigns to each result.

However, other types of result visualization can be considered. One alternative is to present the results on three dimensional interfaces. It is in these types of systems that this work will focus, the search engines with three dimensional interfaces. The problem is that Web pages are not ready to be consumed by these types of search engines.

To solve this problem, a generic model was built for the web pages, which can feed the requirements of the different variants of these search engines.

An automated instantiation prototype was also developed, which collects the necessary information from Web pages and fills the model.

Keywords: Search engines, 3D Interfaces, Generalist Model, Metadata Standards.

Índice

Índice de Tabelas

Índice de Figuras

Notação e Glossário

DC	Dublin Core
HTML	Hypertext Markup Language
API	Application Programming Interface
XML	Extensible Markup Language

1. Introdução

A Web no mundo actual é considerada um serviço indispensável para grande parte da sociedade moderna. Desde que foi criada, a Web tem vindo a expandir exponencialmente e os websites que a compõem já ascendiam às centenas de milhões no final de Dezembro de 2008. (Netcraft, 2008). Com tanta informação disponível levanta-se a problemática de a organizar para que seja útil aos utilizadores.

O método mais comum de procurar informação na internet nos dias que correm é por meio de motores de pesquisa online. Estes sistemas indexam os websites segundo um sistema de relevância e através de um conjunto de palavras-chave fornecidas pelo utilizador devolvem um conjunto de páginas que à luz do critério de relevância particular do motor de pesquisa em questão, parecem mais adequadas à pesquisa efectuada. O interface destes motores de pesquisa normalmente é minimalista, consistindo apenas de uma caixa de texto, onde o utilizador insere as palavras-chave sobre a informação que procura, e um botão para iniciar a pesquisa. A nível de apresentação dos resultados, estes são dispostos numa lista de hiperligações que conectam às páginas Web ou documentos respectivos, ordenados pela relevância que o motor de pesquisa atribui a cada um desses resultados.

Existem porém motores de pesquisa com interfaces gráficos alternativos aos motores de pesquisa tradicionais. Estes motores têm uma forma diferente de apresentar a informação aos utilizadores, proporcionando uma experiência de navegação diferente dos modelos tradicionais.

Este documento centra-se neste tipo de motores de pesquisa, e em particular, na especificação de um modelo generalista para a representação de páginas Web que suporte qualquer motor de pesquisa gráfico em 3 dimensões.

O desafio enfrentado, em concreto, passa por inicialmente criar um modelo de páginas Web que suporte qualquer motor de pesquisa com interfaces gráficos em 3 dimensões (3D). Apresentar uma metodologia para a instanciação desse modelo. Criar um protótipo que implementa essa metodologia. Finalmente, fazer uma avaliação sobre o desempenho da instanciação do modelo.

1.1. Enquadramento

A relevância de documentos face a uma determinada necessidade de informação, é uma área muito complexa sobre a qual se fazem vários estudos, com o objectivo de determinar qual a melhor forma de o fazer. Uma das principais dificuldades desta área consiste na escolha dos critérios de avaliação da relevância dos documentos. Estes critérios podem variar entre comparar o número de palavras que os documentos têm, o número de ocorrências de certas palavras-chave num documento, os títulos dos textos, o servidor onde estão armazenados, etc. Uma agravante a esta dificuldade é a escolha entre optar por seleccionar apenas um critério ou um conjunto de critérios para determinar a relevância de uma página. Se optarmos pela primeira opção, temos como vantagem o facto de haver menos informação a considerar. Mas a escolha de apenas um critério pode resultar em avaliações desniveladas. Se escolhermos, por exemplo, o critério como sendo o número de total palavras de um documento, onde consideramos que quanto maior for o total, maior vai ser a sua relevância, estamos a impor a condição que os documentos de maiores dimensões são os mais relevantes, e este factor, dependendo do objectivo do utilizador, pode estar completamente incorrecto. Se por outro lado, optarmos por escolher um conjunto de critérios, o cruzamento dos seus valores deverá resultar numa avaliação mais nivelada e correcta. Uma técnica que pode ser aplicada neste caso é a de atribuir pesos a cada um dos atributos. Desta forma os atributos mais importantes, dependentes do objectivo do utilizador, vão contribuir mais na determinação da relevância de um documento.

Os motores de pesquisa com interfaces gráficos permitem apresentar a informação de uma forma diferente dos motores de pesquisa tradicionais (que apresentam os resultados em forma de listas de hiperligações). Este motores tradicionais, embora sejam os mais usados, por vezes não são muito eficazes a apresentar a informação, e obrigam o seu utilizador a despender muito tempo à procura da informação que pretende no meio de uma lista enorme de hiperligações. A utilização de motores de pesquisa com interfaces gráficos mais avançados, permite apresentar a informação de uma forma mais agradável e aliciante ao utilizador. Podemos usar uma metáfora simples para apresentar a informação, como por exemplo uma árvore onde apresentamos os documentos como ramos da árvore, sendo os ramos mais próximos do tronco os que mais se aproximam da pesquisa efectuada. Este tipo de “truques” facilitam a análise visual do utilizador, que ao olhar para os documentos, sobre um interface com uma metáfora aplicada, conseguem mais facilmente decifrar quais são

os melhores documentos, ou caminho a tomar que o vão ajudar a encontrar a informação que necessita.

No entanto, cada metáfora, ou seja, cada formato de interface, vai requerer um conjunto de atributos específico para conseguir apresentar os resultados com o maior nível de precisão e eficácia. O problema, actualmente, é que nas páginas Web, os atributos requisitados por estes sistemas não são facilmente extraídos. O propósito deste documento é precisamente encontrar uma solução a este problema através da criação de um modelo que consiga reunir de cada página todos os atributos chave necessários ao funcionamento de qualquer um destes sistemas, independentemente do formato que estejam a utilizar. Desta forma, será criada uma estrutura de informação comum que possa ser utilizada por qualquer motor de pesquisa com interface em 3D.

Ao longo deste documento é realizado um estudo sobre motores de pesquisa com interfaces 3D actuais, quais os requisitos que estes precisam para funcionarem e finalmente é apresentado um modelo generalista para páginas Web optimizado para ser usado por este tipo de sistemas.

1.2. Tecnologias Utilizadas

O protótipo de instanciação do modelo generalista de páginas Web foi desenvolvido sobre a linguagem .NET em conjunção com as APIs (*Application Programming Interface*) AlchemyAPI e Google Ajax Language API. Foi usada a ferramenta Visual Studio 2010 para o desenvolvimento do código do protótipo.

1.3. Contributos deste trabalho

Este trabalho incide sobre a temática dos motores de pesquisa com interfaces gráficos em 3D. O seu objectivo é providenciar informações sobre estes sistemas, seja sobre a forma como funcionam ou sobre as dificuldades que enfrentam actualmente. Em relação às dificuldades é focado o problema da recolha dos atributos necessários à avaliação de relevância das páginas Web e apresentada uma solução ao mesmo.

Como resultados deste trabalho destacam-se os seguintes contributos:

- Estudo do estado de arte sobre motores de pesquisa com interfaces em 3D.

- Estudo e análise dos mais recentes Standards de Metadata.
- Proposta de modelo generalista de representação de uma página Web com vista à sua recuperação com recurso a motores de pesquisa com interface em 3D.
- Protótipo de instanciação do modelo referido.

1.4. Organização do documento

Este trabalho encontra-se estruturado por capítulos para melhor compreensão do leitor.

No capítulo 1 deste documento introduz-se o problema proposto descrevendo-se as ideias e objectivos gerais.

No capítulo 2 é feito um estudo sobre os motores de pesquisa, focado principalmente nos motores com interfaces 3D. É feito um levantamento do estado de arte deste tipo de sistemas. São ainda recolhidos os requisitos principais necessários ao bom funcionamento dos motores de pesquisa.

No capítulo 3 são abordados os standards de metadata. Estes standards foram estudados com o intuito de encontrar uma solução viável para definir um modelo generalista para páginas Web a ser usado pelos motores de pesquisa a fim de melhorar a sua eficiência.

É apresentada, no capítulo 4, uma avaliação aos standards apresentados no capítulo anterior. Após a avaliação foi escolhido o standard que melhor se adequa aos requisitos impostos pelos motores de pesquisa e apresentado um modelo com base nesse standard.

No capítulo 5 são estudados vários métodos para extrair informação de páginas Web com o intuito de instanciar, de forma automática, o modelo proposto no capítulo anterior. É também apresentado um protótipo de instanciação do modelo generalista desenvolvido com recurso a uma das ferramentas estudadas.

A avaliação do protótipo de instanciação do modelo generalista está exposta no capítulo 6.

Por fim, no capitulo 7, são apresentadas as conclusões obtidas da realização deste projecto assim como as recomendações e linhas de trabalho futuro.

2. Motores de pesquisa com interfaces 3D

2.1. Definição

Os Motores de Pesquisa facilitam a exploração das vastas quantidades de informação existentes na Internet. Sem estes motores, encontrar a informação que pretendemos seria um enorme problema face à quantidade de informação disponível. Para dar resposta a este elevado volume de dados que continua a crescer [REF16], os motores de pesquisa têm sofrido várias evoluções. Desde há vários anos que investigadores têm procurado a melhor forma de pesquisar, organizar e apresentar a informação para que seja útil para o utilizador. O método de apresentação utilizado actualmente pelos principais motores de pesquisa como o Google [REF10], Bing [REF11] e Yahoo [REF12] é o da lista em 2 dimensões (2D) com hiperligações para páginas onde se encontra a informação que procuramos.

No entanto já foram levados a cabo vários estudos sobre métodos de visualização em 3D, onde se tentou arranjar formas inovadoras de apresentar a informação. Estes métodos quando acoplados a motores de pesquisa visam tornar a experiencia de navegação pelos resultados obtidos mais gratificante e esclarecedora para os seus utilizadores. Através destes métodos de visualização também se espera obter informações extra, como relações entre resultados, que de outra forma não seria possível visualizar.

2.2. Critérios de relevância

Segundo estudos efectuados por vários investigadores e instituições [REF42], a forma mais utilizada de pesquisa de informação na Web consiste no uso de motores de pesquisa. O paradigma dos motores de pesquisa passa por 3 fases. (1) Especificação do interesse de informação através de um conjunto de palavras-chave que são indicadas pelo utilizador. (2) Estas palavras-chave são posteriormente usadas pelo sistema de pesquisa para identificar os documentos mais relevantes (segundo algum critério, ou conjunto de critérios desconhecidos pelo utilizador). (3) Finalmente, é apresentado ao utilizador por ordem decrescente da sua relevância (conforme foi determinado pelo sistema de pesquisa escolhido).

Este modelo de pesquisa levantou um novo problema, o da selecção das páginas Web relevantes entre as centenas, milhares, ou até mesmo milhões que podem ser obtidas como resposta a uma pesquisa. Para filtrar resultados indesejados pode-se recorrer a mais palavras-chave, a expressões mais específicas, ou a operadores booleanos. No entanto a diminuição do número de páginas retornadas aumenta o risco de perder páginas que poderiam ser mais relevantes para a consulta do que as seleccionadas com os novos parâmetros.

Uma dificuldade típica com que os motores de pesquisa se debatem é conhecida por "semantic gap", que consiste na dificuldade em perceber a real necessidade do utilizador com base exclusivamente num conjunto de palavras-chave. Esta dificuldade decorre naturalmente da natureza ambígua da língua natural.

A questão que se coloca é: quais são os critérios que determinam que umas páginas sejam consideradas, pelos motores de pesquisa, como mais relevantes do que outras. Os motores de pesquisa de primeira geração baseavam-se no seguinte: [Bateira, 2006]

- A frequência absoluta ou relativa, tendo ou não em consideração o tamanho da página Web, da palavra-chave ou da expressão nas páginas Web e, eventualmente, o seu destaque mediante um tipo especial de letra;
- A posição da palavra-chave ou da expressão nas páginas Web, nomeadamente a sua colocação em lugares estratégicos como o título, o subtítulo, a secção inicial, a metadata, etc;
- O peso relativo de certos termos nas páginas Web que contêm as palavras-chave ou as expressões, tendo em consideração factores como a presença de termos não habituais ou não comuns, o desprezo das chamadas *stopwords* (palavras que contribuem pouco no significado geral de um texto, tais como artigos, preposições e advérbios).
- A proximidade das palavras-chave ou das expressões em relação a certos termos que, por isso mesmo, serão também considerados relevantes.

Por outro lado, a utilização destes critérios traz vários problemas, nomeadamente a sua grande permeabilidade em relação às diversas técnicas de spam [REF27], a dificuldade ou mesmo impossibilidade de lidar com fenómenos típicos da linguagem natural como a sinonímia (relação que se estabelece entre duas palavras ou mais que apresentam significados iguais ou semelhantes), a homonímia (relação que se estabelece entre palavras que se escrevem da mesma maneira, se dizem da mesma

maneira, mas têm significados diferentes) ou a derivação das palavras (alterar a forma da terminação das palavras para exprimir uma variação de significado).

A consequência destes problemas era que, nos motores de "primeira geração", o resultado de uma pesquisa continha algumas páginas Web relevantes no meio de uma vastidão de páginas irrelevantes face à consulta especificada; as respostas tinham uma precisão muito baixa [Leighton e Dr. Srivastava, 1997].

Na tentativa de resolver estes problemas, os motores de "segunda geração" utilizam critérios de relevância que se agrupam em duas grandes categorias: [Bateira, 2006]

- Critérios que determinam a relevância da páginas Web em função de um conceito ou campo semântico (sugerindo significados ou termos alternativos aos actuais de forma a encontrar a melhor solução para a consulta efectuada), de tal forma que são consideradas como relevantes todas as páginas circunscritas a tal conceito ou campo semântico.
- Critérios que determinam a relevância das páginas Web em função do comportamento dos utilizadores da mesma.

Porém, estes motores de "segunda geração" não conseguiram eliminar todos os problemas. Um problema que estes motores de pesquisa herdaram dos seus predecessores foi o carácter global relativo dos critérios de relevância, em que um determinado documento pode ser muito relevante para um motor de pesquisa, mas pouco relevante para outro motor de pesquisa concorrente. Podia-se optar por usar vários motores de pesquisa na tentativa de discutir melhor a relevância de um documento, embora essa solução acabasse por agravar o problema que se procura resolver. Embora se obtenha uma média de relevância para cada resultado, se os motores de pesquisa tiverem critérios de relevância muito dispares podem provocar uma variação substancial da relevância dos resultados. O utilizador pode ficar baralhado quanto à precisão dos resultados, além de se gastar muito tempo na avaliação de cada resultado.

2.3. Interfaces 3D VS Interfaces 2D

Quando se planeia um interface em 3D temos de considerar as vantagens que este nos vai trazer em relação aos interfaces 2D tradicionais. Embora existam pesquisas que comprovam que interfaces 3D melhoram a experiência do utilizador, como por exemplo em simulações e design gráfico em 3D CAD\CAM [Shneidman, 2003],

quando estes são aplicados na visualização dos resultados das pesquisas de um motor de pesquisa online isso não acontece de forma imediata. Como os utilizadores já estão muito familiarizados como o formato 2D dos seus motores de pesquisa, quando lhes é apresentado um novo interface existe sempre uma resistência e confusão sobre o que lhes está a ser apresentado. Estudos feitos sobre esta problemática [Sutcliffe e Pattel, 1996 ; Risden, 2000] mostram que estes interfaces em 3D inibem os utilizadores e tendem a ser mais confusos que os interfaces em 2D, dificultando a tarefa principal que é apresentar resultados fáceis de decifrar pelos utilizadores.

Dos aspectos que têm de ser considerados na construção de interfaces 3D, a oclusão de resultados e o custo da profundidade de visualização devem ser especialmente considerados. Estes aspectos são muitos importantes no desenvolvimento do interface e devem ser bem estipulados e planeados pois levantam vários problemas: Onde vai ser o ponto de oclusão? Qual a quantidade de resultados a mostrar antes de se tornar confuso? Estaremos a esconder os resultados certos? Todas estas perguntas têm de ser cuidadosamente estudadas para obter um equilíbrio entre a quantidade e a qualidade da informação se está a apresentar.

Cerca de 25% dos utilizadores têm dificuldade em usar interfaces 3D quando são apresentados em aparelhos 2D como monitores de PC. [Modjeska, 2000]

Outros problemas associados aos interfaces em 3D é que são mais difíceis e confusos de usar para pessoas menos experientes nesta área. Um utilizador casual de motores de pesquisa adapta-se mais facilmente a um interface 2D simples do que a um interface 3D. [Sebrechts, 1999]

“The task of managing and accessing large information spaces is a problem in large scale cognition. Emerging technologies for 3D visualization and interactive animation offer potential solutions to this problem, especially when the structure of the information can be visualized. We describe one of these Information Visualization techniques, called the Cone Tree, which is used for visualizing hierarchical information structures. The hierarchy is presented in 3D to maximize effective use of available screen space and enable visualization of the whole structure. Interactive animation is used to shift some of the user's cognitive load to the human perceptual system.” [Robertson et al., 1991]

O que se pode retirar destes estudos é que os interfaces aplicados a motores de pesquisa têm de ser simples e intuitivos. Um utilizador, principalmente se não tiver

muita experiencia, vai ter mais dificuldade a fazer as pesquisas e visualizar os resultados se o interface não for intuitivo e fácil de usar. Logo, um interface em 3D que tenha demasiadas funcionalidades ou animações vai perder para um interface 2D em termos de usabilidade e em vez de auxiliar o utilizador a encontrar o que procura pode fazer exactamente o oposto, o que não é aceitável.

2.4. Interfaces em 3D? O que são?

Um interface em 3D implica a adição de mais uma dimensão, a profundidade, ao interface tradicional (2D) de uma aplicação. Logo vai proporcionar uma nova forma de visualizar um conjunto de dados, com uma nova perspectiva, diferente da que os principais motores de pesquisa da actualidade usam e a que estamos habituados. Com a adição da profundidade, a forma como são apresentados os resultados de uma pesquisa apresenta novas componentes que antes estavam escondidas e que podem ajudar o utilizador a melhor entender e assimilar a informação. A posição em termos de profundidade de um resultado em relação a outro pode por si só revelar uma relação que antes não era possível observar nos interfaces em 2D.

Mas o que é que podemos considerar como interface 3D?

Podem existir algumas variações ao que os utilizadores consideram um interface como sendo em 3D. No caso dos motores de pesquisa um interface pode ser em 3D por níveis ou como um todo. Em relação ao interface de um motor de pesquisa propomos ter em conta 3 níveis:

- **O nível da pesquisa:** onde ainda não temos resultados e vamos inserir os parâmetros que possuímos para obter a informação;
- **O nível da apresentação:** de resultados, aqui é onde é mostrado ao utilizador os resultados da pesquisa que efectuou;
- **O nível da informação:** onde o utilizador está a aceder já a um dos resultados obtidos a fim de retirar a informação que pretende.

Num motor de pesquisa convencional, como o Google, os níveis acima referidos são representados da seguinte forma:

- Nível de pesquisa: è apresentado um interface em 2D com uma caixa de texto onde o utilizador insere as palavras-chave para a sua pesquisa e um botão para iniciar a pesquisa de informação.

- Nível de apresentação: Os resultados são apresentados em forma de uma lista de hiperligações, com um pequeno extracto do texto por baixo, também esta num plano em 2D.
- Nível de informação: Após o utilizador aceder a um dos resultados da lista de resultados é reencaminhado para uma nova página ou documento onde se encontra alojada a informação relativa ao resultado escolhido, que pode estar ou não representada em 3D, não sendo responsabilidade do motor de pesquisa a forma como é apresentada essa informação.

O conceito de um interface 3D para um motor de pesquisa fica dividido, entre um interface 3D nivelado, onde pelo menos um dos níveis tem de ser apresentado em 3D, ou um interface 3D puro, onde todos os níveis são apresentados em 3D.

Um exemplo de um interface 3D nivelado seria de um motor que apenas tivesse um nível de apresentação em 3D, isto é, os resultados de uma pesquisa fossem apresentados sobre um plano 3D, usando a profundidade como factor de relevância de um determinado resultado. No entanto os níveis de pesquisa e de informação seriam apresentados de uma forma convencional.

Para os interfaces 3D puros teríamos motores de pesquisa onde todos os níveis seriam representados em 3D. Neste caso seria necessário criar um ambiente virtual. Um possível exemplo de um ambiente virtual para um motor de pesquisa seria uma biblioteca onde a informação era representada por livros numa estante. A estante seria o nível de apresentação de resultados que conforme os parâmetros inseridos pelo utilizador, seria preenchida por livros que por sua vez representam os resultados retornados pela pesquisa, ordenados na estante por ordem de relevância da esquerda para a direita e de cima para baixo. Quando um livro fosse seleccionado, era executada uma animação que abria o livro virtual e mostrava nas suas páginas a informação referente ao resultado escolhido.

Um motor de pesquisa com um interface 3D pode ser uma boa ferramenta que nos permite visualizar a informação de uma forma alternativa e vislumbrar relações inesperadas entre resultados. Existem várias formas de apresentar os resultados num ambiente em 3D, como foi referido no subcapítulo anterior, mas dependendo da informação que se esteja a pesquisar, algumas formas tem mais sucesso que outras. Se for utilizado um interface inadequado à pesquisa que um utilizador está a realizar, em vez de facilitar e melhorar a experiencia da pesquisa só vai dificultar e baralhar o utilizador.

2.5. Metáforas de Interfaces em 3D

Uma das particularidades que um motor de pesquisa com interface 3D oferece é a forma como apresenta a informação e altera a percepção sobre a mesma por parte do utilizador. De modo a melhorar a percepção e cativar visualmente os utilizadores, estes sistemas recorrem a metáforas visuais idealizadas para ajudar e guiar o utilizador a encontrar a informação que procura.

É necessário analisar estas metáforas, e consequentemente os sistemas que as utilizam, com o objectivo de perceber quais são os atributos/características dos documentos que são utilizados para determinar a sua relevância e determinar a posição onde serão apresentados na amostragem de resultados. Esta análise é um dos pontos fulcrais deste trabalho, pois é daqui que se retira os atributos necessários para construir o modelo generalista.

2.5.1. Montanha de dados

Nesta metáfora os resultados da pesquisa são dispostos sobre uma paisagem em 3D. Os resultados semelhantes são aglomerados e conforme a sua relevância para um determinado conceito comum formam elevações na paisagem. Deste modo, aglomerados maiores formam elevações mais preponderantes que se destacam na paisagem e os resultados menos relevantes são representados em volta das elevações ou dispersos pelas zonas mais baixas. Estudos sobre a utilização deste tipo de representação de informação provaram ser ineficiente. [Cockburn e McKenzie, 2001]

Figura 1 - Data Mountain [FG 1]

2.5.1.1. VxInsight

Esta ferramenta foi desenhada para trabalhar sobre bases de dados extensas. Os dados são apresentados como objectos geométricos num plano e o seu posicionamento representa a sua relação com os outros objectos. A proximidade entre objectos é baseada na medição das suas similaridades com os outros objectos. Quando muitos objectos ficam agregados estamos normalmente na presença de um contexto que os interliga e graças a este tipo de representação gráfica fica fácil de identificar esse contexto. Isto ajuda o utilizador a perceber a estrutura implícita da informação em vez de apenas o contexto onde ela se encontra.

O VxInsight após calcular o plano em 2D dispõe-no sobre uma paisagem virtual em 3D que se assemelha a uma cadeia montanhosa. A altura de cada montanha é proporcional à densidade dos objectos que se encontram por baixo desta.

O VxInsight permite ao utilizador olhar para a informação como um todo e depois fazer zoom para ver os detalhes até à rede de relações entre os dados. Se a base de dados representar dados que foram acumulados ao longo de um período de tempo, o VxInsight permite também que o utilizador veja como as representações do terreno variaram ao longo do tempo à medida que o conjunto da informação cresceu até atingir o ponto actual.

Vários princípios básicos foram usados para desenhar esta ferramenta:

- O sistema visual humano é excelente a identificar padrões, tendências e anomalias.

- Os humanos interpretam melhor a informação se forem utilizadas metáforas familiares.
- Uma ferramenta de navegação útil que permita, intuitivamente e facilmente, percorrer as vistas detalhadas e de alto nível.
- A compreensão de grandes bases de dados é facilitada pela habilidade de explorar a informação em resoluções diferentes, como por exemplo, fazer zoom *in* e *out* da representação. Para melhorar a interpretação, a resolução deve variar continuamente.
- Como não existe nenhuma representação óptima para todos os tipos de questões, deve ser criada a opção de criar vários tipos de amostragens e mecanismos de interacção.

O verdadeiro poder da ferramenta VxInsight é revelado quando o utilizador faz questões usando a interface SQL do sistema à base de dados. Normalmente estas questões são do tipo booleanas que são passadas à base de dados. Depois a ferramenta usa os resultados da pesquisa para marcar os elementos visualmente no mapa. Faz isso pondo pontos coloridos sobre os elementos na paisagem 3D. Os resultados de cada pesquisa são representados com cores diferentes. Desta forma, podemos fazer várias questões, e através desta representação com cores, podemos ver quais as questões que retornam elementos iguais ou aproximados. Este tipo de representação dos dados e das pesquisas ajudam o utilizador a gerar hipóteses, testa-las e actualizadas e por vezes levar a novas pesquisas que ainda não foram idealizadas. Tudo isto em tempo real.

Os atributos utilizados por esta ferramenta para determinar a similaridade dos documentos são os seguintes:

- Palavras-chave comuns entre os documentos.
- Vocabulário idêntico entre os documentos.
- Citações ligadas por hiperligações entre os documentos.
- Hiperligações directas.
- Registos de transacções financeiras entre corporações.
- A associação em organizações comuns entre os indivíduos.

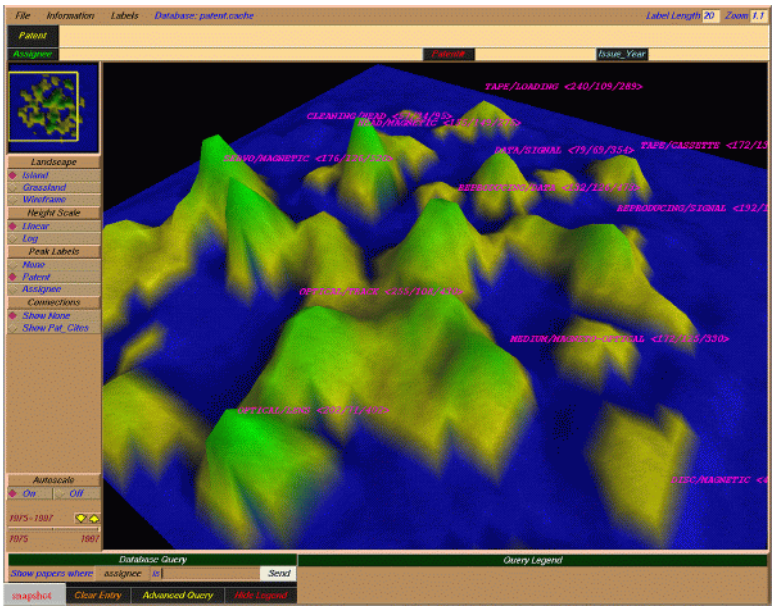

Figura 2 - Interface do VxInsight [FG 2]

2.5.2. Árvores Hiperbólicas

O uso de estruturas em árvore para representar informação já é usado há muito tempo. Os casos mais comuns são a representação de hierarquias em árvores binárias simples onde é fácil de gerir e visualizar a hierarquia quando esta é reduzida. O problema surge quando a hierarquia que se está a visualizar é extensa o que representa um aumento exponencial nos nós da arvore. À medida que vamos expandindo os nós da árvore para revelar mais informação, vamos ocupando mais e mais espaço visual. Isto é um problema pois rapidamente podemos encher o espaço visual disponível. Foi então criada uma técnica de visualização, recorrendo a árvores, para lidar com este problema, a árvore hiperbólica.

As árvores hiperbólicas são representadas através de grafos inspirados na geometria hiperbólica [Cannon, 1997]. Estas árvores estão inseridas num plano hiperbólico que tem mais espaço utilizável que o plano euclidiano usado pelas suas antecessoras.

Os resultados são apresentados através de uma visão “olho de peixe” que foca no centro do espaço visual o nó que estamos a visualizar, ficando este mais destacado e distante dos outros nós, enquanto os restantes nós estão fora de foco e são aglomerados nas pontas. [Munzner, 1997].

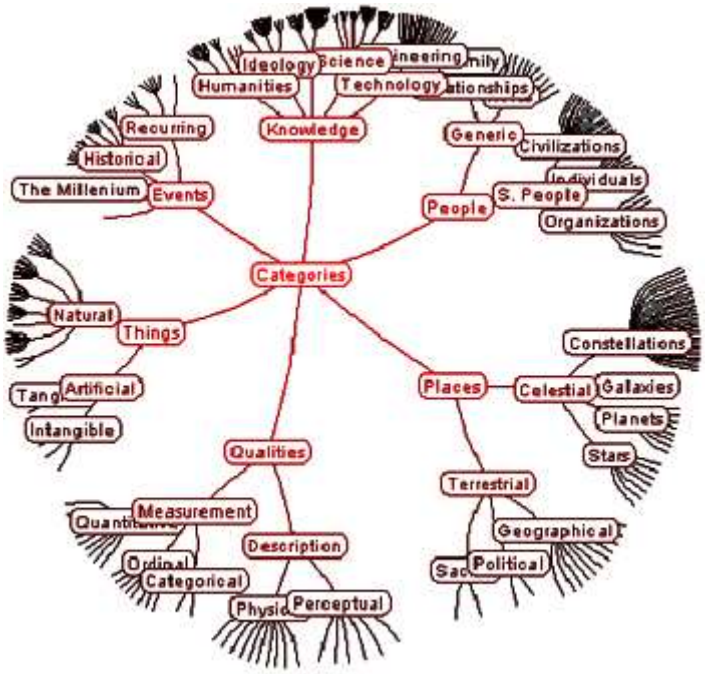

Figura 3 - Exemplo de árvore hiperbólica [FG 3]

2.5.2.1. Árvore hiperbólica da Embrapa

Um exemplo deste tipo de modelo de visualização é usado na **Embrapa** – Empresa Brasileira de Pesquisa Agropecuária. O seu modelo é alimentado com termos referentes à agropecuária. São-nos apresentados vários termos iniciais que irão ser o ponto de partida para a nossa pesquisa. Quando seleccionamos um termo o gráfico hiperbólico é desenhado com esse termo focado no centro. Para visualizarmos os outros resultados relacionados com o tema central utilizamos a técnica de "clique e arraste", que faz com que o tema que estamos a visualizar fique focado no centro e o anterior seja empurrado para um dos lados. Este tipo de visualização é bastante intuitivo, visto que o utilizador percebe facilmente quais são os temas que estão mais próximos uns dos outros, e facilmente navega pela árvore à procura da informação de deseja.

Neste exemplo, os documentos ainda são analisados manualmente por pessoal especializado, que determina, com base no conteúdo de cada documento, a qual categoria este pertence e quais são as categorias predecessoras. Após o documento ser analisado é inserido na árvore na posição correspondente.

Os atributos necessários para construir a árvore são:

- Texto do documento

- Categoria do documento
- Categoria dos predecessores

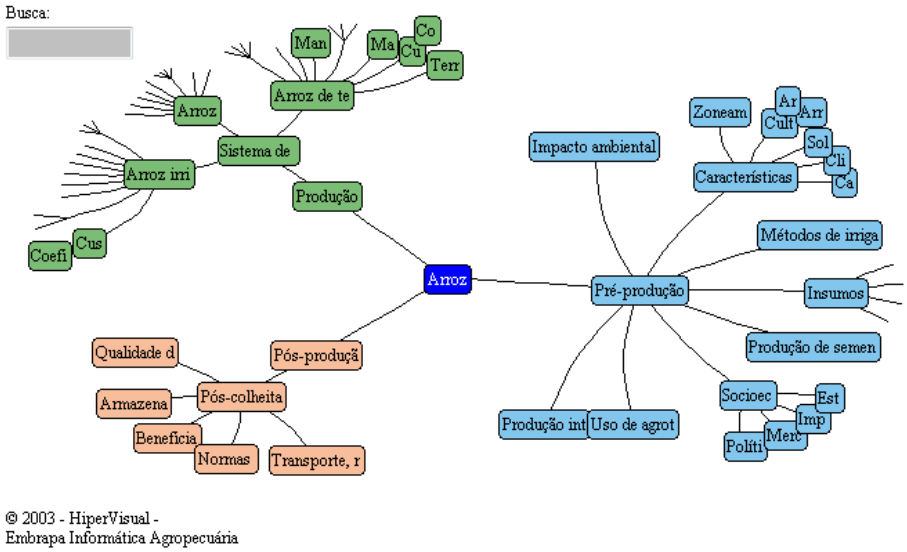

Figura 4 - Árvore Hiperbólica do arroz da Embrapa [FG 4]

2.5.3. Árvores Cónicas

Esta representação de informação foi desenvolvida para se visualizar estruturas hierárquicas de dados em 3D de uma forma mais interactiva e apelativa aos utilizadores. A ideia por trás desta técnica é usar a profundidade para aumentar o espaço visível, da árvore cónica, no ecrã e estimular a capacidade de percepção visual dos utilizadores. Os utilizadores podem rodar a árvore para visualizarem a informação que está escondida, e ver a informação de vários ângulos dando uma melhor percepção da hierarquia que está a ser visualizada. Um estudo sobre está técnica revelou que os utilizadores mostraram-se entusiásticos sobre esta nova forma de visualização e que achavam que melhorava a assimilação da informação que estavam a pesquisar. [Cockburn & McKenzie, 2000]

No entanto este método foi também considerado muito confuso por muitos outros utilizadores. Segundo um estudo realizado pela Universidade de Canterbury da Nova Zelândia [Cockburn & McKenzie, 2000], os utilizadores têm mais dificuldade em

navegar pela hierarquia do que num sistema com uma árvore vertical, e à medida que a hierarquia vai crescendo, mais confusa se torna. [Robertson et all, 1991].

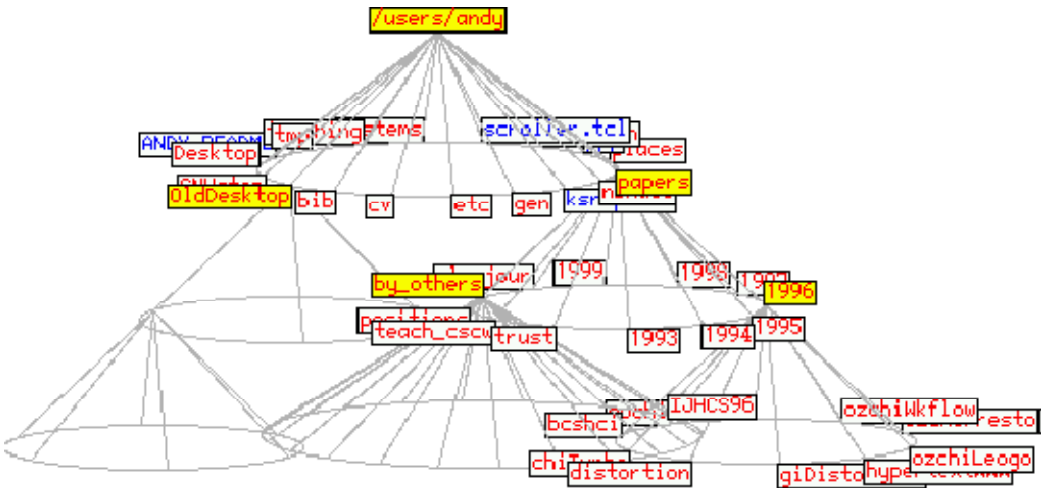

Figura 5 - Exemplo de uma árvore cónica [FG 6]

2.5.3.1. VR-VIBE

O VR-VIBE é um ambiente virtual criado para apresentar e pesquisar informação. É um sistema multi-utilizador onde é possível visualizar colecções de documentos ou referencias a documentos num ambiente virtual 3D. Esta ferramenta usa *Points of Interest* (POI) para estruturar e localizar a informação que está a apresentar. A posição espacial de um documento indica a sua afinidade e proximidade aos diferentes POIs e através desses relacionamentos, é possível determinar termos de temática similar. Logo, quanto mais próximo estiver um documento de um POI, mais este está relacionado com o assunto ao qual o POI se refere. No entanto um documento pode estar perto de diversos POI, o que significa que é relevante ou contém informações relevantes sobre vários temas.

Para indicar a relevância dos documentos o VR-VIBE usa um sistema de cores e tamanhos. Quanto mais relevante for um documento para um determinado tema, maior é o objecto que o representa no mundo virtual, ganhando assim mais relevo visual em relação aos outros. Para além disso, a sua cor torna-se mais brilhante. Isto ajuda o utilizador a encontrar os documentos mais relevantes ao(s) tema(s) que está a pesquisar.

Para determinar o tamanho que cada objecto deve ter, o VR-Vibe usa um algoritmo simples que procura, no texto dos documentos, padrões de similaridade com os temas (contando o número de ocorrências de palavras-chave, texto dos títulos e o corpo do

documento). Desta forma determina também a distância a que o documento vai estar dos POIs.

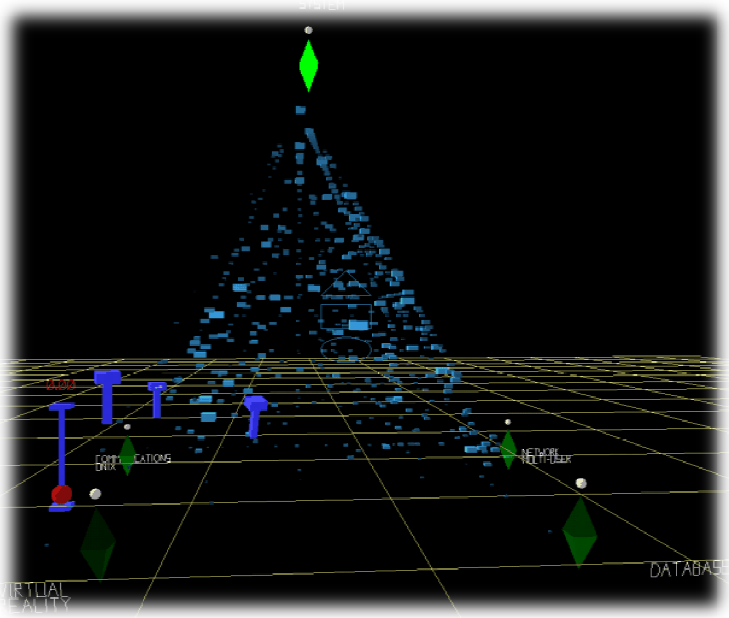

Figura 6 - Exemplo do VR-VIBE [FG 5]

Na figura acima apresentada temos 5 POIs representados pelos octaedros verdes. As esferas brancas por cima dos octaedros indicam que o POI está activo. Os blocos azuis representam os documentos.

Os atributos utilizados por este sistema são:

- Texto dos documentos.
- Palavras-chave.

2.5.3.2. Cat-a-Cone

O Cat-a-Cone consiste num interface interactivo para a especificação de pesquisas e visualização de resultados usando hierarquia de categorias. Esta aplicação utiliza extensas colecções de documentos, previamente categorizados, e apresenta-os ao utilizador num ambiente interactivo. Um exemplo, que é mostrado nas imagens, é a colecção MEDLINE que consiste em documentos sobre medicina e estão associados com a *Medical Subject Headings* (MESH) onde estão guardadas aproximadamente

16.000 categorias. O interface para apresentar a informação é separado em duas partes. A primeira parte consiste numa árvore cónica, ou várias entrelaçadas, exposta na horizontal, onde são mostradas as categorias. Aqui o utilizador pode navegar sobre as várias categorias, rodando os ramos da árvore, e permite visualizar as hierarquias às quais as categorias pertencem. Quando o utilizador selecciona uma categoria, o sistema desenha uma estante de livros em 3D no canto inferior esquerdo, sendo esta a segunda parte do interface do Cat-a-Cone. Na estante são representados por livros virtuais, todos os documentos relacionados com a categoria seleccionada. Quando se selecciona um dos livros é mostrado um painel com informações gerais sobre esse documento e o conteúdo deste, dando a impressão que pegamos num livro real.

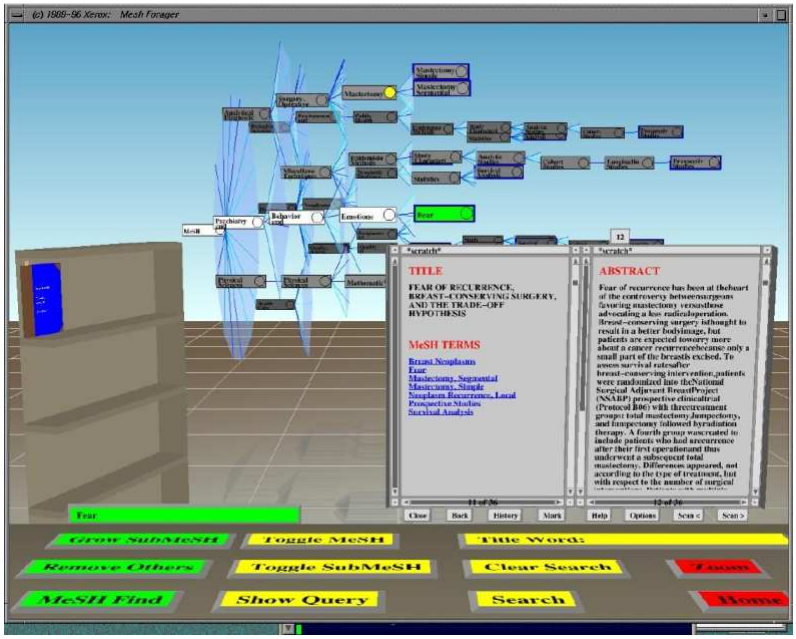

Figura 7 - Vista geral do interface Cat-a-Cone [7]

O Cat-a-Cone é um sistema que se aproxima da noção de interface 3D puro já que todos os níveis são apresentados em 3 dimensões.

Os atributos utilizados por este sistema são:

- Categoria dos documentos
- Texto dos documentos
- Palavras-chave
- Vocabulário Idêntico
- Hiperligações Directas.

2.5.4. Globo Terrestre

Nesta metáfora os resultados são expostos na superfície de um globo ou esfera. A posição dos resultados no globo é determinada pela sua relevância para com os parâmetros de pesquisa que foram introduzidos. Dependendo do sistema que esteja a ser usado, a latitude e longitude dos resultados é calculada de forma a dispor os resultados em áreas de relevância, ou não, no globo.

2.5.4.1. NIRVE (NIST Information Retrieval Visualization Engine)

O propósito deste motor de pesquisa é permitir a visualização e manipulação de conjuntos de documentos retornados de uma pesquisa primária a um motor de pesquisa externo. Esta pesquisa é feita com recurso a palavras-chave. O NIRVE permite consolidar estas palavras-chave em conjuntos mais específicos de conceitos para depois organizar os documentos em grupos, chamados *clusters*, baseados nos perfis conceptuais dos documentos. Após os *clusters* terem sido criados, são apresentados num ambiente 3D onde o utilizador pode interagir com estes e com os documentos que os compõem. O utilizador pode optar ainda por esconder certos *clusters* para focar-se na informação mais relevante para a sua pesquisa, facilitando e melhorando assim a experiencia de utilização.

O NIRVE está dividido em três áreas:

- Controlo NIRVE: área que contem os controlos principais da aplicação, como por exemplo, a inserção das palavras-chave da pesquisa, esconder e mostrar os clusters, filtrar resultados, etc.
- Espaço Documental: área onde é apresentada a representação gráfica dos resultados. Aqui vão aparecer os *clusters* que foram criados sobre os documentos retornados pela pesquisa e a aplicação dos conceitos escolhidos pelo utilizador. Os *clusters* são representados por ícones, e são expostos na superfície de um globo. A latitude dos ícones é determinada pelo número de conceitos que um determinado conjunto possui. Quantos mais conceitos um *cluster* possuir, mais perto este fica do pólo Norte. A longitude não tem nenhum significado intrínseco, apenas é usada para manter *clusters* que tenham perfis idênticos mais próximos uns dos outros. Os conjuntos que se diferenciem por apenas um conceito são conectados por um arco cuja cor representa a diferença conceptual entre eles. Exemplo: se um conjunto A tem como

conceitos ‘*boat’*, ‘*sink’* e ‘*ocean’* e o conjunto B tem como conceitos ‘*boat’*, ‘*sink’*, ’*ocean*’ e ‘*storm’*, então eles vão ser conectados por um arco com a cor atribuída ao conceito ‘*storm’*. A espessura do ícone do *cluster* é proporcional ao número de conceitos que possui. Quando um *cluster* é aberto, um rectângulo 2D é apresentado, onde são apresentados todos os títulos dos documentos contidos no *cluster*. Os títulos por sua vez são arranjados de forma que títulos similares tenham posições horizontais aproximadas. As posições verticais dos títulos são definidas pelo *ranking* que foi retornado pelo motor de pesquisa externo.

- Controlo de Conceitos: Esta área é usada para o utilizador consolidar as palavras-chave usadas na pesquisa em conjuntos mais específicos de conceitos. Tipicamente, sinónimos seriam consolidados como conceitos disjuntivos, por exemplo ‘ensinar’ e ‘educar’ e nomes próprios como conceitos conjuntivos, por exemplo *'internal'* e *'revenue'* e *'service'* (IRS). São atribuídas cores aos conceitos inseridos que depois são apresentados no fundo do ecrã de visualização de resultados. É ainda criado um conceito com o valor ‘UNUSED’ que serve para albergar o conjunto de palavras-chave às quais não se atribuíram conceitos.

Os atributos utilizados por este sistema são:

- Palavras-chave.
- *Ranking* retornado pelo motor de pesquisa externo para determinar a posição dos resultados.

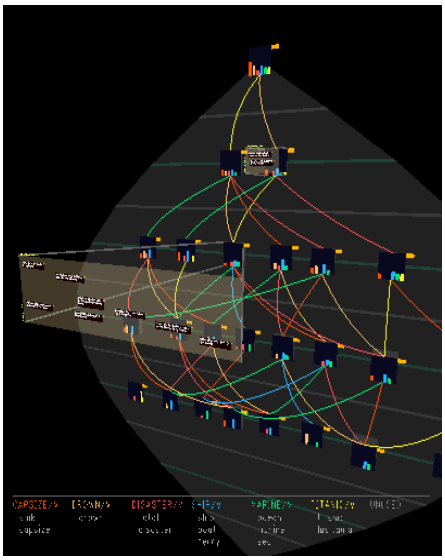

Figura 8 - Interface do NIRVE [FG 8]

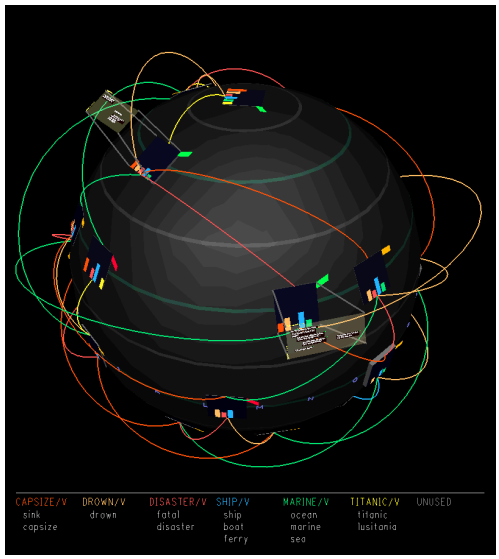

Figura 9 - Vista Global do interface do NIRVE [FG 9]

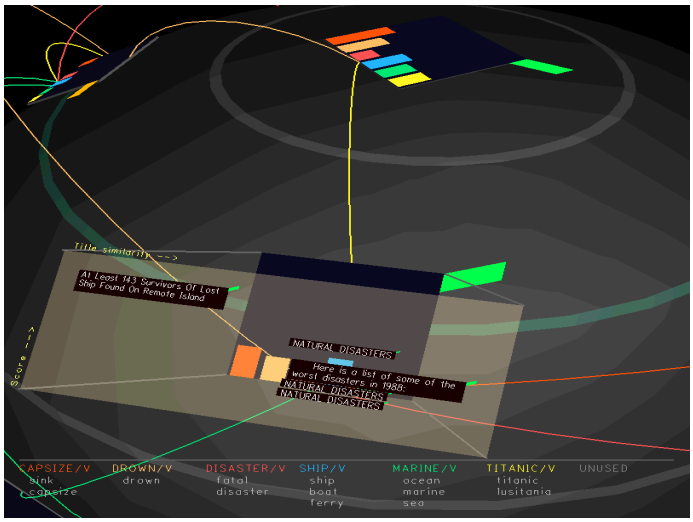

Figura 10 - Vista detalhada de um resultado no NIRVE [FG 10]

2.5.5. Tile Browsing

Nesta metáfora os resultados são representados por “mosaicos”, ou seja, representações compactas dos resultados dentro de pequenos quadrados. Geralmente, estes mosaicos são representados com uma imagem do ecrã principal do respectivo resultado. Alguns motores de pesquisa tradicionais, como o Google, já testaram esta metáfora ao colocarem estes mosaicos junto às hiperligações da sua lista de resultados, mas na vertente a 2D. Aqui, os mosaicos são dispostos num espaço 3D, onde a profundidade é usada para dar relevância aos resultados, ou

apenas para facilitar a navegação, puxando os resultados que o utilizador está a inspeccionar para a frente, enquanto relega os restantes para trás.

2.5.5.1. SpaceTime

O SpaceTime é uma aplicação gratuita que mostra os resultados de outros motores de pesquisa online sobre um interface 3D. Podemos fazer pesquisas usando o Google como o provedor das pesquisas ou o Ebay [REF20] se estivermos à procura de produtos para comprar. Além destes, ainda podemos pesquisar usando o Youtube [REF17], Amazon [REF18], ou outro fornecedor que disponibilize informação via RSS [REF19]. O que a aplicação faz é apresentar os resultados recolhidos pelo motor seleccionado num interface 3D simples e intuitivo de usar.

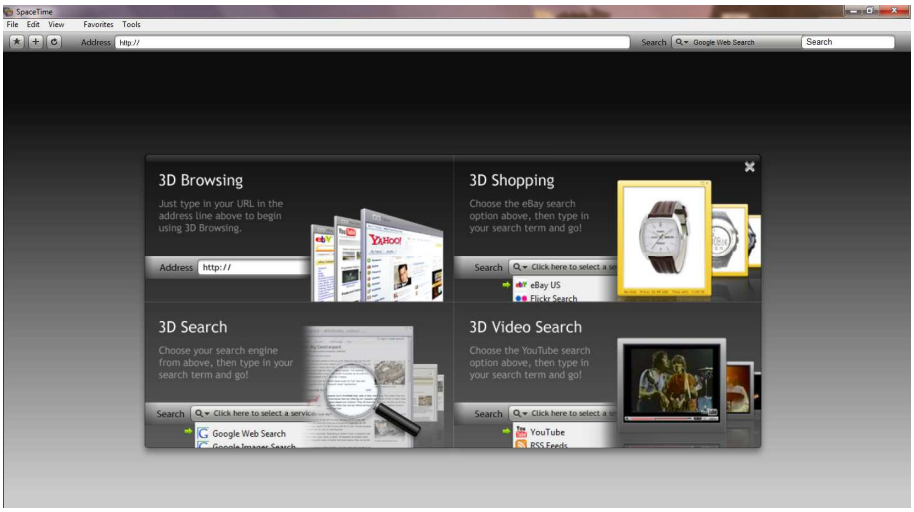

Figura 11 - Ecrã inicial do SpaceTime

Mas o SpaceTime permite também navegar a Web com este interface. Para tal basta digitar o endereço da página pretendida na barra de endereços e a página é carregada usando o mesmo interface gráfico em 3D.

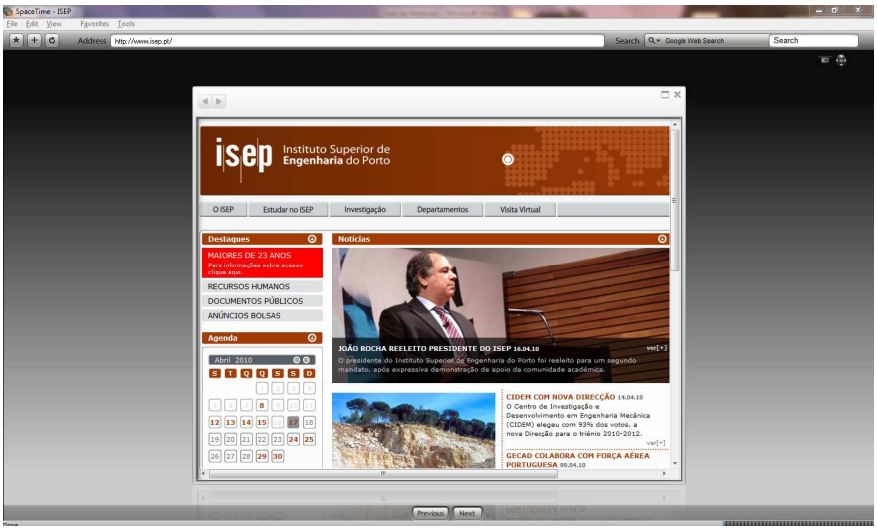

Figura 12 - Visualização de sites no SpaceTime

Quando fazemos uma pesquisa usando o Google como provedor dos resultados, obtemos a lista de resultados de uma forma diferente da que estamos habituados. Em vez da tradicional lista de hiperligações que o Google nos apresenta, o SpaceTime organiza essas hiperligações num espaço tridimensional e mostra pré-visualizações dos sites retornados pela pesquisa. Os sites mais relevantes são apresentados à frente dando uma noção de profundidade à pesquisa. Isto dá ao utilizador uma sensação mais realista. Os resultados menos relevantes parecem estar mais longe da informação que procuramos.

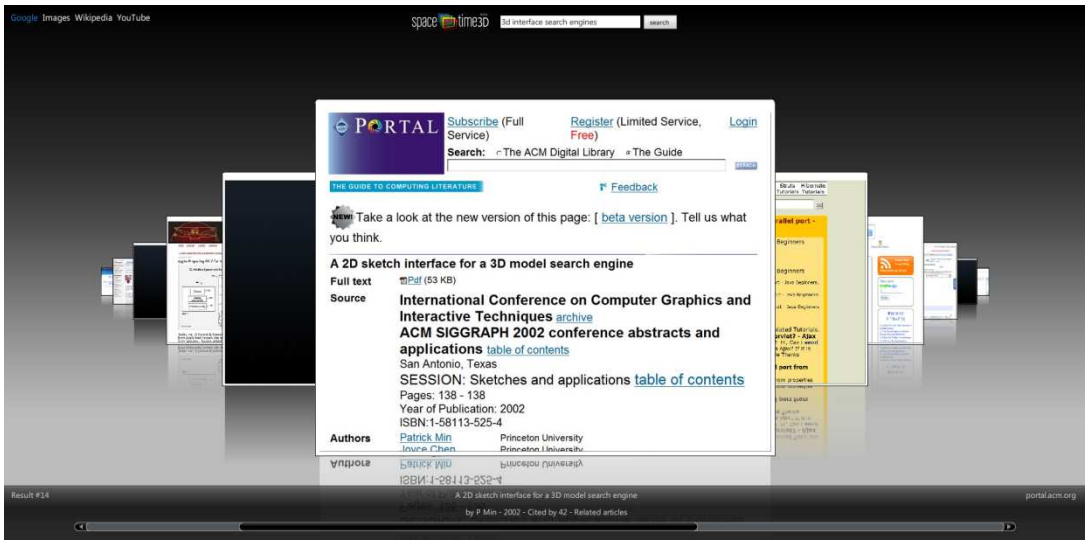

Figura 13 - Pesquisa Google no SpaceTime

Esta aplicação também apresenta resultados de pesquisas no Ebay. Permite ver os resultados da pesquisa no Ebay no mesmo interface 3D com as pré-visualizações das imagens dos produtos e na parte inferior mostra o estado actual do leilão para o produto apresentado.

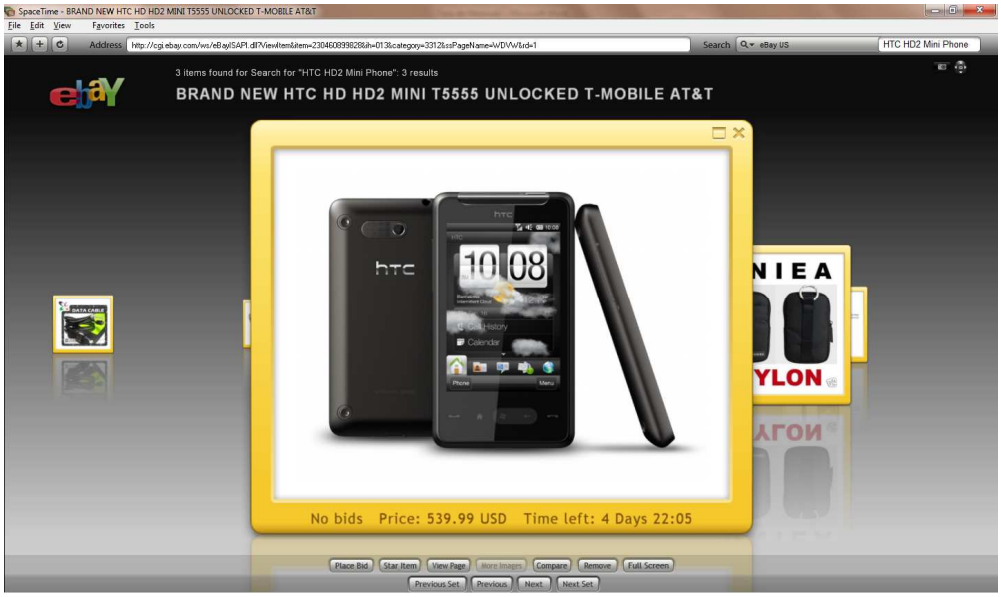

Figura 14 - Pesquisa de produtos no Ebay usando o SpaceTime

Os atributos utilizados por este sistema são:

- Palavras-chave.
- Texto do documento.

- *Ranking* retornado pelo motor de pesquisa externo para determinar a posição dos resultados.

2.5.5.2. Cooliris

O Cooliris é uma aplicação freeware desenvolvida pela Cooliris, Inc. É um *plugin* para os navegadores Web que proporciona uma nova forma de vermos e navegarmos sobre a Web. Para já, apenas está disponível para os seguintes navegadores Web: Firefox, Chrome, Safari e Internet Explorer. O único conteúdo que se pode visualizar no Cooliris são imagens e vídeos e é compatível com os seguintes sites ou provedores: Bing Images, Google Image Search, Yahoo! Image Search, Ask.com Images, AOL Image Search, Flickr, Photobucket, Corbis, Picasa, Fotki, FotoTime, deviantART, SmugMug, Facebook, MySpace, Bebo, hi5, Friendster, YouTube (para vídeos), Gallery, Craigslist, Amazon.com, e qualquer outro site que implemente *tags* mediaRSS nas suas páginas HTML. Além do conteúdo que se pesquisa na Web o Cooliris também consegue mostrar imagens que estejam no nosso computador.

A apresentação dos resultados no Cooliris é feita numa parede tridimensional, onde são dispostos mosaicos com as imagens ou vídeos encontrados na pesquisa que efectuamos. O utilizador tem de escolher qual o provedor que pretende para efectuar a pesquisa, e providenciar as palavras-chave. Após a recolha dos resultados a aplicação mostra os mesmos em formato de mosaicos, ordenados por relevância, da esquerda para a direita. O utilizador pode agora usar o rato para navegar pela parede de mosaicos, e pode fazer zoom para ver um item mais detalhadamente ou ter uma vista alargada de todos os resultados. Do lado esquerdo do ecrã é apresentado um menu com uma lista de canais pré-definidos, que são colecções de determinadas categorias generalistas. Entre estes canais estão por exemplo: Musica, TV e Cinema ou Imagens do nosso computador. Quando escolhemos um destes canais, a aplicação faz uma pesquisa pré-definida sobre o tema e retorna os resultados mais populares, segundo os seus critérios, que encontre sobre o tema escolhido.

Para navegar pelos resultados o utilizador apenas tem de fazer o movimento de arrasto para a direita ou para a esquerda e os resultados representados na parede fluem na direcção pretendida. Esta animação é bastante fluida e intuitiva, o que pode ser uma mais-valia para atrair os utilizadores a utilizar esta ferramenta. O zoom aos mosaicos pode ser feito usando a roda de *scroll* do rato.

Figura 15 - Visão geral do interface do CoolIris

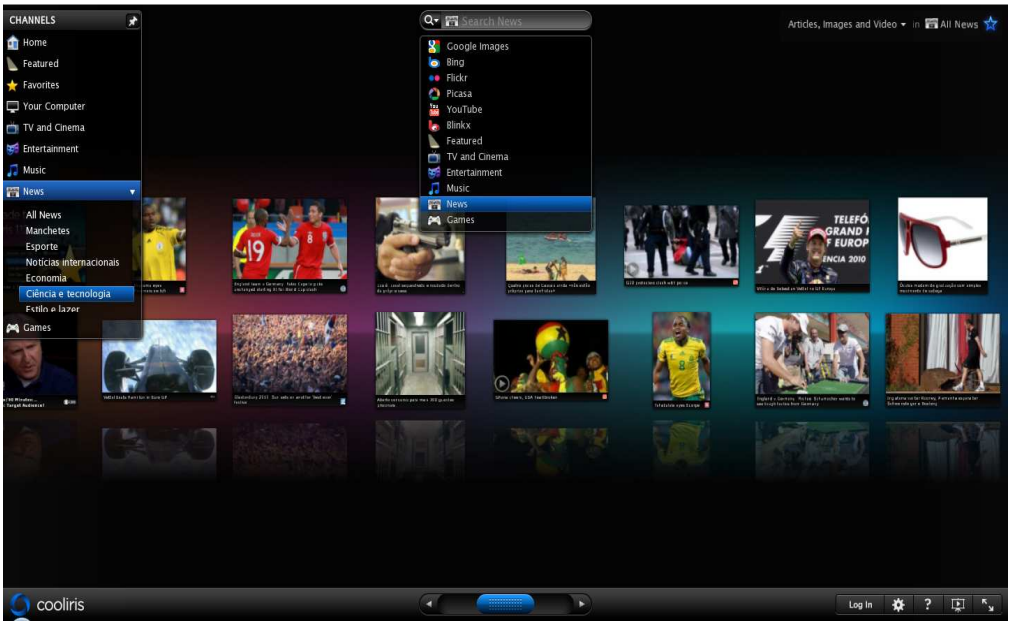

Figura 16 - Motores de pesquisa suportados no CoolIris

Os atributos utilizados por este sistema são:

- Palavras-chave.
- Texto do documento.
- *Ranking* retornado pelo motor de pesquisa externo para determinar a posição dos resultados.

2.5.6. Pivot Tables

2.5.6.1. Microsoft Pivot

Esta aplicação, ainda muito recente, foi desenvolvida pela Microsoft e apresenta uma ideia fresca e inovadora à forma como podemos visualizar a informação. O Microsoft Pivot foi desenvolvido para responder a um problema que tem vindo crescer e a tornar-se cada vez mais comum a todos os utilizadores, “Como é que organizo uma enorme colecção de dados, e tiro partido dela?” [Flake, 2010].

Com o Pivot, os dados são apresentados na forma de colecções de imagens acompanhadas de informação textual. Uma das opções, disponibilizada pelo sistema, é a consulta da Wikipédia, onde são criados pequenos mosaicos das páginas onde a informação está contida que é complementada com um texto identificativo do tema ao qual a página se refere. O utilizador pode agora navegar pelos mosaicos, fazer ‘zoom in’ sobre a zona de imagens que interessa para observar com mais detalhe os mosaicos que contenham a informação que procura, ou ‘zoom out’ para ver a colecção inteira. O Pivot aposta nesta forma simples e intuitiva de visualizar informação, e o seu criador, Gary Flake, espera que esta ajude a que a informação intrínseca dos dados simplesmente venha a superfície visualmente, sem que o utilizador tenha de gastar muito tempo a investigar.

O poder do Pivot está mesmo nesta interface de utilizador, que foi desenhada para facilitar a identificação de padrões de semelhança entre os dados em grande conjuntos de informação, interligando assim toda a informação em vez de termos apenas uma enorme pilha de dados desorganizados e dispersos.

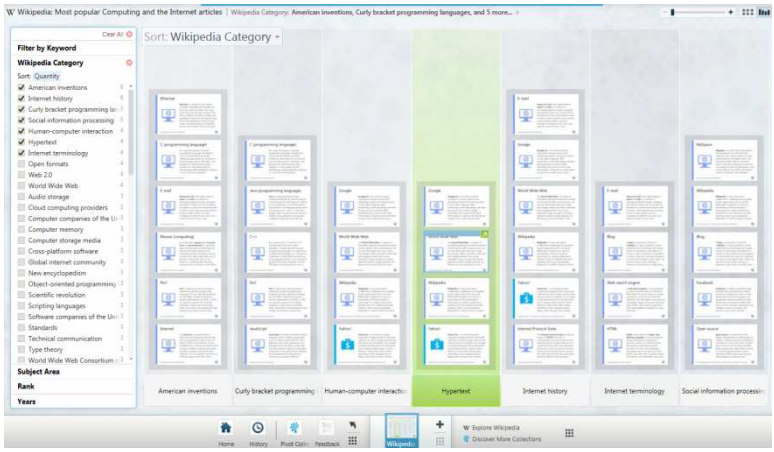

Figura 17 - Pesquisa na Wikipédia no Microsoft Pivot

"You can interact with the data in a way that's not quite browsing and not quite searching," [Flake, 2010].

A tecnologia que impulsiona o Pivot é o Microsoft *SeaDragon*, que é um software criado para lidar com grandes quantidades de informação visual. Este software permite que naveguemos por vastas colecções de imagens, aplicar zooms sem termos enormes tempos de carregamento. Isto tudo também é possível pois só recentemente é que os computadores têm capacidade de processamento para aguentarem este tipo de aplicações.

Outro ponto forte do Pivot são as colecções. O uso de colecções permite que os dados fiquem ainda mais estruturados e organizados de uma forma intuitiva, facilitando a navegação por parte do utilizador sobre os dados. As colecções combinam grandes grupos de itens similares na Internet, de modo a que possamos ver as relações que existem entre os vários pedaços de informação de uma forma completamente nova e diferente para nós. Isto permite vislumbrar padrões novos, escondidos na informação enquanto estamos a interagir com centenas de coisas ao mesmo tempo. A equipa de desenvolvimento do Pivot já fornece com a aplicação um conjunto de colecções predefinidas, mas o cada utilizador pode e deve fazer as suas próprias colecções de modo a tirar o máximo partido das potencialidades desta aplicação.

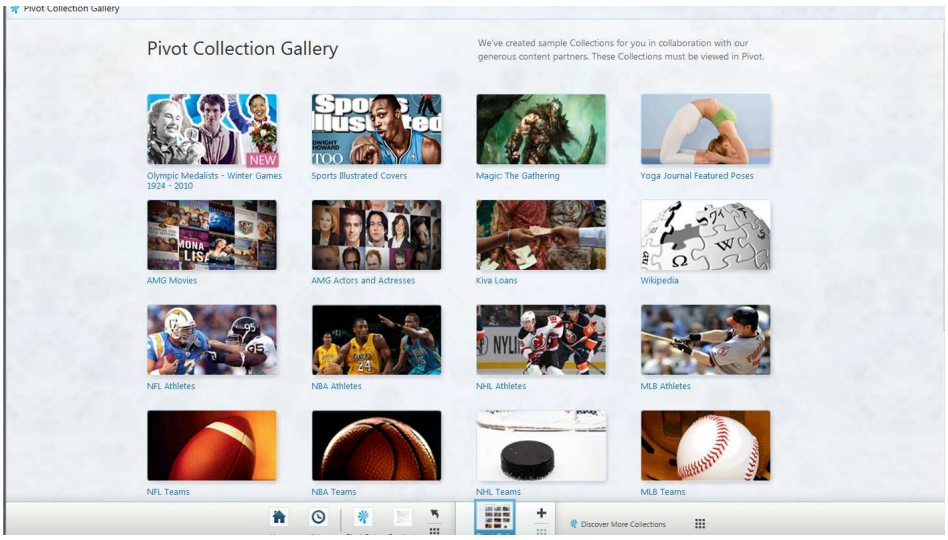

Figura 18 - Galeria de colecções do Microsoft Pivot

Foi ainda criado um controlo em Silverlight [REF2] que nos dá a opção de embeber as colecções directamente nas nossas páginas Web.

Os atributos utilizados por este sistema são:

- Categoria dos documentos.
- Texto dos documentos.
- Palavras-chave.
- *Ranking* retornado pelo motor de pesquisa externo para determinar a posição dos resultados.

2.6. Atributos utilizados na classificação e indexação de documentos

Fazendo a análise das aplicações acima referidas, podemos retirar um conjunto de atributos chave que são usados para a organização, classificação e indexação dos modelos 3D usados no seu interface. Dependendo da aplicação em questão a classificação dos documentos pode ser feita com base num só atributo, ou num conjunto combinado de atributos para determinar melhor a posição do documento no espaço virtual.

Da análise anterior retiramos o seguinte conjunto de atributos requisitados pelas aplicações:

Tabela 1 - Atributos requisitados pelos motores de pesquisa

Atributo	Descrição
Categoria:	Categoria geral do conteúdo da página Web.
Texto do documento:	Refere-se ao próprio conteúdo textual da página Web.
Palavras-Chave:	Refere-se às palavras-chave contidas numa página Web.
Vocabulário idêntico:	Sinónimos e/ou traduções das palavras-chave e/ou texto do documento.
Hiperligações de citação	Hiperligações usadas para referir a origem de algum conteúdo da página Web.
Hiperligações directas	Hiperligações de conexão a páginas com conteúdos semelhantes.

Transacções Financeiras	Dados sobre transacções financeiras do conteúdo da página. Por exemplo, no caso de o conteúdo ser acerca de um determinado livro, ter uma referência ao numero de cópias vendidas e/ou valor monetário arrecadado das vendas.
Predecessor	Referência a documentos predecessores do documento actual.
Ranking Externo	Posição ocupada no *Ranking* de um motor de pesquisa externo.

Nem todos os sistemas providenciaram uma descrição clara e concreta dos atributos em que se baseavam para apresentar os resultados. Nesses casos, os atributos retirados são baseados em assumpções feitas sobre a análise visual do interface do motor aliada com a escassa informação disponibilizada. Algumas destas assumpções podem estar incompletas, ou mesmo incorrectas.

De seguida será apresentado um quadro comparativo, entre os atributos recolhidos e os sistemas estudados. Este quadro vai-nos ajudar a determinar quais são os atributos comuns entre os sistemas e ter uma imagem geral dos atributos que teremos de considerar prioritários na construção do modelo generalista.

Tabela 2 – Atributos necessários para a classificação de documentos

	Cat-a-Cone	VR-VIBE	VxInsight	Embrapa	NIRVE	SpaceTime	Cooliris	Pivot
Categorias	✓			✓				✓
Texto do documento	✓	✓		✓		✓	✓	✓
Palavras-Chave	✓	✓	✓	✓	✓	✓	✓	✓
Vocabulário idêntico	✓		✓					
Hiperligações de citação			✓					
Hiperligações directas	✓		✓					
Transacções Financeiras			✓					
Predecessor				✓				
Ranking Externo	✓				✓	✓	✓	✓

2.6.1. Outros atributos de classificação de documentos

Nesta secção serão apresentados alguns dos atributos, que na nossa opinião, podem ajudar os motores de pesquisa a atingir uma classificação mais equilibrada e precisa das páginas Web.

Nº de palavras: Número total de palavras que o documento possui; Este atributo determina o tamanho, relativamente ao texto, de um documento; O tamanho de um documento pode estar directamente ligado à sua qualidade e relevância perante um determinado tema e às necessidades do utilizador. Um utilizador pode necessitar apenas de uma pequena definição ou um documento detalhado sobre o tema que está a pesquisar.

Nº de imagens: Número total de imagens que o documento possui; A inclusão de imagens e diagramas nos documentos podem ajudar á melhor compreensão dos seus conteúdos, Se um documento apresentar muitas imagens poderá ter maior relevância que outros.

Nº de vídeos: Número total de vídeos que o documento possui; A inclusão de vídeos numa página pode ser usada como medida de relevância da mesma.

Servidor:

Localização: Localização do servidor; Esta informação pode, por exemplo, ser usada para restringir resultados a uma determinada área geográfica.

Autor: Autor do documento; Saber quem foi o autor dos conteúdos incluídos numa página pode aumentar consideravelmente a sua relevância caso o autor seja considerado uma autoridade sobre o tema que estamos a pesquisar.

Nº de ocorrências de palavras-chave: O número de ocorrências das palavras-chave seleccionadas para uma pesquisa num determinado documento pode ajudar a determinar a sua relevância.

Total: Número total de vezes que as palavras-chave aparecem no documento.

Nos títulos: Número total de vezes que as palavras-chave aparecem no (s) titulo (s) do documento.

No texto: Nos títulos: Número total de vezes que as palavras-chave aparecem no (s) texto (s) do documento.

Data de documentos: Data de criação e/ou actualização dos documentos; A indicação da data de criação e actualização dos documentos pode ser um factor de avaliação útil para determinar se a informação está obsoleta ou não, e determinar também o espaço temporal onde se enquadra.

Referencias: São representadas normalmente por hiperligações nas páginas Web.

Internas: são referências a outros documentos; Este factor pode ser um bom indicador de um documento apropriado a ser o ponto de partida para o inicio da pesquisa.

Externas: são referências de outros documentos ao documento em questão; Se um documento for referenciado por vários outros documentos pode indicar que este seja mais relevante.

Língua: Língua do documento; Os documentos que estejam escritos na língua nativa do utilizador poderão ser mais úteis do que documentos escritos noutras línguas.

Número de visitas: diárias/semanais/mensais; a afluência de visitantes a um determinado documento pode ser um indicador da relevância do mesmo. Um documento que é visitado regularmente, por um grande número de utilizadores, poderá ser considerado mais relevante que outro que não recebe quase visitas nenhumas.

3. Standards de Metadata

Uma forma de garantir que a informação é bem categorizada, é dar de antemão essa informação aos motores de pesquisa no próprio recurso, a página Web. Logo, desde a construção das páginas onde vamos expor a informação devemos ter o cuidado de fornecer o máximo de detalhes descritivos para que seja associada a colecções de dados que realmente estejam em consonância com o seu conteúdo. Neste momento já existem alguns campos, dentro da estrutura de uma página Web, apropriados para este fim. Um bom exemplo é o da *tag* HTML <meta> [REF8] que pode ser usada para guardar dados descritivos sobre o conteúdo da página Web, a fim de ajudar os motores de pesquisa a encontrarem informação requisitada pelos seus utilizadores. Este é um dos métodos mais simples de guardar informação sobre o conteúdo de uma página Web, mas muitas vezes não está disponível, pode não ser fiável e não consegue armazenar toda a informação que precisamos.

Uma forma de resolver este problema é recorrer a estruturas mais poderosas que consigam armazenar todos os dados descritivos que necessitamos. Já foram criados vários standards para propósitos semelhantes, principalmente na área bibliográfica onde existe a necessidade de ter informação pormenorizada sobre as obras contidas nas colecções das bibliotecas. De seguida são apresentados os standards de metadata analisados.

3.1. Dublin Core

O standard de metadata Dublin Core (DC) [REF28] é um conjunto, eficaz mas ao mesmo tempo simples, para descrever uma grande gama de recursos interligados. O standard DC inclui dois níveis: Simples e Qualificado. O nível Simples é composto por quinze elementos e o nível Qualificado possui três elementos adicionais ao anterior (*Audience, Provenance* e *RightsHolder*) e também um grupo de elementos de refinamento, chamados qualificadores que refinam a semântica dos elementos de forma a serem úteis na descoberta dos recursos. A semântica do DC foi estabelecida por um grupo internacional de profissionais de várias áreas: bibliotecários, ciências computacionais, processamento de textos, comunidades de museus e outros campos académicos e práticos.

Outra forma de olhar para o DC é como uma pequena linguagem para fazer descrições específicas sobre recursos. Nesta linguagem existem duas classes de termos. Os elementos que são representados por nomes e os qualificadores que são classificados por adjectivos que podem ser dispostos num padrão simples de declarações. Os recursos são os sujeitos implicados na linguagem.

"The association of standardized descriptive metadata with networked objects has the potential for substantially improving resource discovery capabilities by enabling field-based (e.g., author, title) searches, permitting indexing of non-textual objects, and allowing access to the surrogate content that is distinct from access to the content of the resource itself." [Weibel e Lagoze, 1997]

Objectivos do Dublin Core:

- Simplicidade de criação e manutenção

O conjunto de elementos do DC foi mantido o mais pequeno e simples possível para permitir que utilizadores casuais consigam criar registos descritivos simples para fontes de informação de uma forma fácil e sem custos, enquanto ao mesmo tempo dá aos sistemas uma forma efectiva de pesquisar e retornar a informação contida nas fontes.

- Semântica comum reconhecida

A pesquisa de informação na internet muitas vezes depara-se com um problema complicado de ser ultrapassado, as várias terminologias e praticas descritivas usadas por diferentes culturas sobre um mesmo tema. O DC tem por missão ajudar o utilizador a encontrar o caminho certo fornecendo-lhe um conjunto de elementos comuns que são universalmente aceites e suportados. Por exemplo, quando um utilizador estiver interessado em encontrar artigos publicados por um dado autor, ou estudantes de artes estão interessados por um artista particular, todos concordam na importância do elemento '*creator*'. O uso destes elementos comuns aumenta a visibilidade e acessibilidade de todos os recursos relacionados com um determinado tema e até temas semelhantes.

- Projecção internacional

O conjunto de elementos do DC foi desenvolvido em Inglês, mas versões em outras línguas estão a ser criadas, entre estas está o Português, Espanhol, Alemão, Japonês, etc.

- Extensibilidade

Enquanto tenta balancear a necessidade de simplicidade na descrição digital de recursos de informação com a necessidade de retorno exacto de resultados, a equipa do DC reconheceu a importância de incluir um mecanismo que permita estender o conjunto de elementos para necessidades extra de pesquisa. É esperado que especialistas de outras comunidades de metadata criem e administrem novos conjuntos de elementos especializados baseados nas necessidades dessas comunidades. Os elementos destes novos conjuntos podem ser usados com o DC para colmatar a necessidade da interoperabilidade.

Conjunto base de elementos Dublin Core:

Tabela 3 - Conjunto Base de elementos Dublin Core

Conteúdo	Propriedade Intelectual	Instanciação
Coverage	**Contributor**	**Date**
Description	**Creator**	**Format**
Type	**Publisher**	**Identifier**
Relation	**Rights**	**Language**
Source		
Subject		
Title		

Definições dos elementos retiradas da versão portuguesa da secção de atributos da DCMI. [REF3]

Elemento: Título

Nome:	***Title***
Definição:	O nome dado ao recurso.
Comentário:	Tipicamente, um Título será o nome pelo qual o recurso é formalmente conhecido.

Elemento: Criador

Nome:	***Creator***
Definição:	A entidade responsável em primeira instância pela existência do recurso.
Comentário:	Exemplos de Criador incluem uma pessoa, uma organização, ou um serviço. Tipicamente, o nome de um Criador deve ser usado para indicar uma entidade.

Elemento: Assunto

Nome:	***Subject***
Definição:	Tópicos do conteúdo do recurso.
Comentário:	Tipicamente, um Assunto deverá ser expresso por palavras-chave, frases, ou códigos de classificação que descrevem o conteúdo do recurso. Como boa prática recomenda-se a selecção de termos de vocabulários controlados, ou de sistemas de classificação formais.

Elemento: Descrição

Nome:	**Description**
Definição:	Uma descrição do conteúdo do recurso.
Comentário:	Descrições podem incluir, sem estarem limitadas a tal: um resumo, um índice, uma referência a uma representação gráfica do conteúdo, ou uma descrição textual.

Elemento: Editor

Nome:	***Publisher***
Definição:	Uma entidade responsável por tornar o recurso acessível.

Comentário:	Exemplos de um Editor incluem uma pessoa, uma organização ou um serviço. Tipicamente, o nome de um Editor deve ser usado para indicar a entidade.

Elemento: Outro Contribuinte

Nome:	***Contributor***
Definição:	Uma entidade responsável por qualquer contribuição para o conteúdo do recurso.
Comentário:	Exemplos de Outro Contribuinte incluem uma pessoa, organização ou serviço. Tipicamente, o nome de um Outro Contribuinte deve ser usado para indicar a entidade.

Elemento: Data

Nome:	***Date***
Definição:	Uma data associada a um evento do ciclo de vida do recurso.
Comentário:	Tipicamente, uma Data deve ser associada à criação ou disponibilidade do recurso. Como boa prática recomenda-se para codificação de valores de datas um perfil da norma ISO 8601 [REF31], segundo o formato AAAA-MM-DD.

Elemento: Tipo

Nome:	***Type***
Definição:	A natureza ou género do conteúdo do recurso.
Comentário:	Tipos incluem termos descrevendo categorias genéricas, funções, géneros, ou níveis de agregação para o conteúdo. Recomenda-se como boa prática a selecção de valores a partir de vocabulários controlados (por exemplo, a lista do documento de trabalho "*Dublin Core Types*" [REF29]). Para descrever a manifestação física ou digital do recurso, deve ser usado o elemento Formato.

Elemento: Formato

Nome:	***Format***
Definição:	A manifestação física ou digital do recurso.
Comentário:	Tipicamente, o Formato deve incluir o tipo de meio do recurso, ou as

	suas dimensões. Este elemento deve ser usado para determinar as aplicações informáticas ou qualquer tipo de equipamento necessário para reproduzir ou operar com o recurso. Exemplos de dimensões incluem tamanho e duração. Como boa prática recomenda-se a selecção de valores a partir de vocabulários controlados (como por exemplo a lista de "*Internet Media Types*" [REF30] definindo formatos e meios).

Elemento: Identificador

Nome:	***Identifier***
Definição:	Uma referência não ambígua ao recurso, definida num determinado contexto.
Comentário:	Como boa prática recomenda-se a identificação do recurso por meio de uma cadeia de caracteres ou por um número de acordo com um sistema de identificação formal. Exemplos de sistemas de identificação formais incluem o "*Uniform Resource Identifier*" (URI) (incluindo o "*Uniform Resource Locator*" (URL)), o "*Digital Object Identifier*" (DOI) e o "*International Standard Book Number*" (ISBN).

Elemento: Fonte

Nome:	***Source***
Definição:	Uma referência a um recurso de onde o presente recurso possa ter derivado.
Comentário:	O presente recurso pode ter derivado do recurso Fonte na sua totalidade ou apenas em parte. Como boa prática recomenda-se a referência ao recurso fonte através de um identificador em conformidade com um sistema de identificação formal.

Elemento: Língua

Nome:	***Language***
Definição:	A língua do conteúdo intelectual do recurso.
Comentário:	Como boa prática recomenda-se para valores do elemento Língua a utilização do RFC 1766 [RFC1766], o qual inclui um código de língua de duas letras (retirado da norma ISO 639 [ISO639]), seguido opcionalmente por um código de duas letras para o país (retirado da

	norma ISO 3166 [ISO3166]). Por exemplo, 'en' para Inglês, 'fr' Francês, ou 'en-uk' para o Inglês do Reino Unido.

Elemento: Relação

Nome:	***Relation***
Definição:	Uma referência a um recurso relacionado.
Comentário:	Como boa prática recomenda-se referir o recurso através de uma cadeia de caracteres ou número em conformidade um sistema de identificação formal.

Elemento: Cobertura

Nome:	***Coverage***
Definição:	A extensão ou alcance do recurso.
Comentário:	Cobertura inclui tipicamente uma localização espacial (o nome de um lugar ou coordenadas geográficas), um período no tempo (a sua designação, data, ou intervalo de tempo), ou jurisdição (o nome de uma entidade administrativa). Como boa prática recomenda-se a selecção de valores de vocabulários controlados (como por exemplo o "*Thesaurus of Geographic Names*" [TGN]), devendo ainda ser usados, quando apropriado, preferencialmente nomes de lugares e designações de períodos no tempo, em vez de identificadores numéricos tais como coordenadas ou intervalos de datas.

Elemento: Direitos

Nome:	***Rights***
Definição:	Informação de direitos sobre o recurso ou relativos ao mesmo.
Comentário:	Tipicamente, este elemento deverá conter uma declaração de gestão de direitos sobre o recurso, ou uma referência a um serviço que fornecerá essa informação. Tal poderá compreender informação sobre direitos de propriedade intelectual, direitos de autor, ou outros. A ausência deste elemento não permite formular qualquer hipótese válida sobre quaisquer direitos que possam incidir sobre o recurso.

Cada elemento Dublin Core é definido usando um conjunto de 10 atributos do standard ISO/IEC 11179 [ISO11179] para a descrição de elementos de dados. Estes são:

- **Nome** - A etiqueta atribuída ao elemento de dado
- **Identificador** - O identificado único atribuído ao elemento de dado
- **Versão** - A versão do elemento de dado
- **Autoridade de Registo** - A entidade autorizada a registar o elemento de dado
- **Língua** - A linguagem na qual o elemento de dado é definido
- **Definição** - Uma afirmação que representa claramente o conceito e a natureza do elemento de dado
- **Obrigação** - Indica se o elemento de dados é sempre obrigatório ou não (contém um valor)
- **Tipo dos Dados** - Indica o tipo de dados que podem ser representados no valor do elemento de dado
- **Máxima Ocorrência** - Indica qualquer limite à repetição do elemento de dado
- **Comentário** - Uma nota relativa à aplicação do ao elemento de dado

Seis dos referidos dez atributos são comuns a todos os elementos DC. Esses são, com valores exemplificativos:

Versão	1.1
Autoridade de Registo	Dublin Core Metadata Initiative
Língua	pt ("en" para a versão original)
Obrigação	Opcional
Tipo dos Dados	Cadeia de caracteres
Máxima Ocorrência	Ilimitada

3.2. Text Encoding Initiative

A "*Text Encoding Initiative*" (TEI) é um projecto internacional desenvolvido para estabelecer standards na forma como se cataloga textos electrónicos como romances, peças de teatro, poesia, etc. Para além de definir a forma como se codifica os textos, a TEI também especifica um espaço no cabeçalho de um texto onde é inserida metadata acerca do documento. O cabeçalho TEI, assim como o resto do projecto TEI, é

definido como um SGML [REF21] DTD (*document type definition*) – um conjunto de etiquetas e regras definidas em sintaxe SGML que descrevem a estrutura e elementos do documento. Estes elementos SGML ficam embebidos no próprio texto. No entanto, o TEI DTD é bastante extenso e complexo. Para ser usado em grandes conjuntos de textos foi criado o TEI Lite, que é uma versão mais simplista do anterior.

É assumido que os textos codificados usando a TEI são cópias electrónicas dos textos impressos. Por esta razão o cabeçalho TEI pode conter informação bibliográfica tanto do seu formato electrónico mas também do formato não electrónico original. A informação bibliográfica básica é similar à usada na categorização das bibliotecas e pode ser mapeada de e para o formato MARC [REF24]. O MARC, cujo nome é derivado de "*MAchine-Readable Cataloging"*, é um formato de dados que permite ser lido e interpretado por máquinas com o objectivo de catalogar documentos. No entanto, existem elementos novos usados para definir informações adicionais sobre o texto, como a forma como o texto foi transcrito e editado, como é que a catalogação foi feita, que revisões foram feitas e outros factos não bibliográficos. As bibliotecas tendem a usar a codificação TEI quando possuem textos com codificação SGML. O SGML é um standard internacional usado para a definição de representações de documentos em formato electrónico. [NISO, 2001]

3.3. Metadata Encoding and Transmission Standard

O "*Metadata Encoding and Transmission Standard*" (METS) foi desenvolvido para responder à necessidade de criar uma estrutura de dados standard que conseguisse descrever objectos complexos de bibliotecas digitais. O METS é um esquema XML para criar instâncias de documentos XML para expressar a estrutura dos objectos digitais, a sua metadata descritiva e administrativa e os nomes e locais dos ficheiros que compõem o objecto. A metadata utilizada para organizar e utilizar este objectos digitais é mais extensa e diferente da metadata usada para fazer o mesmo com textos impressos e outros materiais físicos. É necessária uma metadata estrutural para assegurar que diferentes ficheiros digitalizados, como por exemplo diferentes páginas de um livro, são estruturados apropriadamente.

Um documento METS tem sete secções principais:

- Cabeçalho METS: cotem informação acerca do documento, incluindo informações sobre o autor, editor, etc.

- Metadata Descritiva: Apontador para metadata descritiva externa ao documento ou um apontador para metadata descritiva embebida no documento ou ambos.
- Metadata Administrativa: Aqui está a informação sobre como o documento foi criado e está armazenado, quem possui direitos de autor sobre o documento, qual foi a fonte física do documento, etc.
- Secção de Ficheiro: Lista de ficheiros que formam o documento digital.
- Mapa Estrutural: Delimita uma estrutura hierárquica para o objecto digital e liga os elementos dessa estrutura aos ficheiros e à metadata relativa a cada elemento.
- Hiperligações Estruturais: Permitem aos criadores do METS registarem os nós na estrutura hierárquica no mapa estrutural.
- Comportamento: Associa comportamentos de execução ao conteúdo do objecto METS.

O cabeçalho METS, secção de ficheiro, mapa estrutural, hiperligações estruturais e comportamento são definidos dentro do esquema do METS. Quanto à metadata descritiva e metadata administrativa, o METS é menos prescritivo e deixa isso em extensões ao esquema base. O METS apoia três esquemas de metadata descritiva: Dubin Core simples, MARCXML [REF25] e MODS. Para metadata técnica o METS disponibiliza no seu site esquemas para texto e imagens digitais fixas. O último é um esquema específico para imagens intitulado MIX (*Metadata for Images in XML Schema*). Estão ainda a ser desenvolvidos esquemas para áudio, vídeo e websites. [NISO, 2001]

3.4. Metadata Object Description Schema

O *Metadata Object Description Schema* (MODS) [REF41] é um esquema de metadata descritiva derivado do formato bibliográfico MARC 21 [REF40]. Este esquema pode ser usado para uma variedade de propósitos, mas é usado principalmente em aplicações bibliográficas. Inclui um subconjunto de campos MARC e usa etiquetas (*tags*) baseadas em linguagem em vez de baseadas em números como no MARC 21. Como é um derivado do MARC permite carregar informação já existente em registos MARC 21 ou permitir criar novos registos descritivos. Tal como o METS, o MODS é expressado em esquemas XML.

Ainda que o standard MODS possa ser usado sozinho, pode também ser usado para complementar outros standards. Por causa da sua flexibilidade e o uso do XML pode ser usado como uma extensão do METS ou um conjunto de metadata usada para recolher e criar registos de metadata original em sintaxe XML. Um dos principais objectivos do MODS é providenciar registos descritivos ricos e completos, o que se torna numa vantagem sobre outros standards de metadata. Os elementos MODS tendem a ser mais ricos que os do Dublin Core; e também a ser mais compatíveis com os dados das bibliotecas pois é mais simples de aplicar o formato MARC 21. [NISO, 2001]

De seguida será apresentada a lista dos elementos principais, e a sua descrição, que compõem o MODS na versão 3.4:

- **titleInfo**
- **name**
- **typeOfResource**
- **genre**
- **originInfo**
- **language**
- **physicalDescription**
- **abstract**
- **tableOfContents**
- **targetAudience**
- **note**
- **subject**
- **classification**
- **relatedItem**
- **identifier**
- **location**
- **accessCondition**
- **part**
- **extension**
- **recordInfo**

Elemento: Titulo

Nome:	***titleInfo***
Definição:	Elemento que pode ser qualificado por vários atributos, que além de outras informações, pode indicar qual é o tipo do título que está a ser usado, a língua em que está escrito ou se foi transliterado de outro texto.
Comentário:	Este elemento é composto por um conjunto de 5 sub-elementos, dos quais, o sub-elemento *title* é de preenchimento obrigatório que é usado para guardar o título principal do item a descrever.

Elemento: Nome

Nome:	***name***
Definição:	È o equivalente ao elemento *Creator* ou *Contributor* no Dublin Core. O elemento *name* é usado para descrever as pessoas ou organizações responsáveis pela criação do conteúdo intelectual do item a ser descrito, ou descrever quem contribuiu em alguma forma para a criação.
Comentário:	Os nomes podem ser separados em componentes (nome de família, nome próprio). O papel de cada pessoa ou organização na criação do documento pode ser descrito de diversas formas, incluindo *relator codes* (como os códigos MARC), ou em texto simples. Pode ser ainda acrescentada uma descrição textual para descrever uma pessoa ou organização em mais detalhe.

Elemento: Origem da Informação

Nome:	***originInfo***
Definição:	Este elemento é usado para descrever informações sobre a origem ou publicação do item.
Comentário:	Contem alguns sub-elementos que podem conter a data de origem do item, que pode ser a data da sua publicação, criação (no caso de documentos não publicados ou manuscritos), ou captura no caso de secções de documentos originais. Podem ainda ser descritos detalhes sobre o editor do documento, ou se o documento é monográfico ou continuo, e no ultimo caso, qual a periodicidade das publicações.

Elemento: Descrição Física

Nome:	***physicalDescription***
Definição:	Este elemento contém também uma variedade de sub-elementos que são usados para providenciarem uma descrição básica das características físicas do documento.
Comentário:	Estas características físicas normalmente apenas são relevantes quando se trata de recursos electrónicos: estes podem incluir *internetMediaType*, que guarda o formato em qual a data é descrita (normalmente representado num formato MIME como por exemplo

	"text/html"), *reformattingQuality* que indica a qualidade (em termos de resolução e *bit-depth)* que um item foi digitalizado, e *digitalOrigin* que indica se um documento é originalmente electrónico ou se é uma cópia digital de outro género de item. Os itens em formatos mais tradicionais podem ser descritos de forma mais limitada, principalmente através do sub-elemento *extent*, onde é guardada informação relativa ao número de páginas, ilustrações, etc e pelo sub-elemento note que pode guardar informação não estruturada sobre as características físicas do item.

Elemento: Assunto

Nome:	***subject***
Definição:	Este elemento é usado para descrever o conteúdo intelectual do item por meio de termos retirados de uma taxonomia. O elemento *subject* divide-se em componentes usados para cobrirem diferentes tipos de termos, tais como nomes, termos geográficos ou intervalos temporais.
Comentário:	Um dos sub-elementos, *hierarchicalGeographic* pode definir uma hierarquia de termos geográficos, permitindo pesquisas de termos mais gerais (como continentes) até termos mais específicos (como cidades). Outro sub-elemento, *cartographics*, permite guardar informação detalhada de coordenadas espaciais, em adição à escala e projecção usadas em mapas. Em suma, o MODS providencia uma rica e extremamente flexível variedade de termos para descrições de assuntos, onde pode ser incorporada um qualquer número de taxonomias.

Elemento: Item Relacionado

Nome:	***relatedItem***
Definição:	Este elemento é muito útil no contexto de uma colecção de itens que partilhem qualquer forma de relação. Aqui podem ser referenciados itens que possuam conteúdos relacionados com o conteúdo do item em questão.
Comentário:	Este elemento permite que registos MODS inteiros de itens relacionados possam ser embebidos dentro do registo MODS do item

	a ser descrito. Isto permite aceder através de apenas 1 item a toda a informação sobre uma colecção de itens. O sub-elemento *type* contem uma lista fixa de valores (incluindo antecessor, predecessor, original ou parte) que especifica o tipo de relação entre os itens. Este elemento tem a mesma função que o elemento *Relation* no Dublin Core, mas é muito mais flexível na sua utilização.

Elemento: Extensão

Nome:	***extension***
Definição:	Embora o MODS ofereça um conjunto de elementos muito mais extensível que o Dublin Core inqualificável é possível que não satisfaça todos os requisitos de metadata necessários para um dado objecto. Nesses casos, oferece a facilidade de estender o seu conjunto de elementos, permitindo que seja guardada metadata em esquemas alternativos que depois são embebidos no registo MODS.
Comentário:	Esta metadata adicional tem de ser descrita num espaço XML diferente (um mecanismo construído de modo a permitir que atributos com o mesmo nome não entrem em conflito uns com os outros entre os vários esquemas no mesmo documento).

Outros elementos MODS

Os restantes elementos MODS são auto-descritivos, pelo que não existe necessidade de os examinar tão detalhadamente. [Gartner, 2003]

- ***typeOfResource***: Tipo de objecto a ser descrito, por exemplo, texto, cartográfico, multimédia. Estes termos são retirados de uma lista pré-definida.
- ***genre***: É um termo mais específico do que o elemento *typeOfResource*. Permite definir termos relativamente ao género (retirados de uma fonte especificada, como por exemplo uma taxonomia ou thesaurus) que caracterizam o documento.
- ***language***: Aqui é definida a língua global do documento, por meio de códigos providenciados por uma de duas listas de autoridade aprovadas para o efeito.
- ***abstract***: Uma descrição do conteúdo intelectual ou uma hiperligação para uma descrição deste tipo.

- ***tableOfContents***: uma lista dos conteúdos do item, que podem ser providenciadas explicitamente ou através de uma hiperligação para uma lista adequada.
- ***targetAudience***: Um termo que descreve para que tipo de audiência o item foi criado (por exemplo, adultos, adolescentes, investigadores) retirado de preferência de uma lista controlada MARC. (disponível em http://www.loc.gov/marc/sourcecode/target/targetlist.html)
- ***note***: Usado para definir informação adicional que não se enquadre em nenhum dos elementos mencionados.
- ***classification***: Número da classificação para um recurso sobre um esquema aprovado como por exemplo a *Library of Congress Subject Headings* [REF4] ou *Dewey Decimal Classification* [REF5].
- ***identifier***: um número único ou código designado em acordo com um esquema aprovado, como por exemplo o ISBN ou número ISSN para uma monografia ou documento periódico respectivamente.
- ***location***: registo da localização física do item, incluindo o repositório e a prateleira onde o item se encontra fisicamente.
- ***accessRestriction***: informação sobre restrições de acesso ao item, incluindo informações sobre os direitos de autor.
- ***recordInfo***: conjunto de elementos que contêm informação sobre a criação do próprio registo MODS, incluindo a data de criação, números de controlos, etc.

3.5. Encoded Archival Description

A *Encoded Archival Description* (EAD) [REF26] começou com um projecto realizado pela biblioteca da Universidade da California, Berkeley, em 1993. O objectivo desse projecto era investigar a aceitação e a viabilidade de desenvolver um standard, não proprietário, de codificação capaz de ser interpretado por máquinas, para ajudas de pesquisa (*finding aids*) como inventários, registos, indexadores ou outros documentos criados por arquivos, bibliotecas, museus e repositórios de manuscritos para suportarem a utilização dos seus textos. Os directores do projecto repararam que o uso da Web estava a ter um papel cada vez maior na pesquisa de informação e acesso aos dados dos repositórios de informação, e apostaram que este deveria ser o caminho a seguir. Foi então desenvolvido o standard que permite organizar os dados de forma a estes serem pesquisados e apresentados online.

Em arquivos e colecções especiais as ajudas de pesquisa são uma ferramenta importante para descreverem recursos. As ajudas de pesquisa diferem dos registos de catalogação por serem muito mais extensas, mais narrativas e explicitas além de serem melhor estruturadas hierarquicamente. Normalmente começam por uma descrição da colecção de dados como um todo, indicando o tipo de informação contida. Se por acaso a colecção consiste em documentos pessoais de um certo individuo, esta descrição pode conter também uma extensa biografia sobre essa pessoa. As ajudas de pesquisa descrevem a serie à qual a colecção de dados pertence, como por exemplo, correspondência, registos comerciais, documentos pessoais, discursos, e no fim acaba com uma discriminação dos conteúdos das caixas fixas e pastas que formam a colecção.

Tal como o cabeçalho TEI, o EAD é definido como um SGML DTD. Começa com uma secção de cabeçalho que descreve a ajuda de pesquisa (por exemplo, quem a criou) e depois continua com a descrição da colecção de dados como um todo. Depois continua com a informação cada vez mais detalhada sobre os registos ou series dentro da própria colecção. Se certos itens da colecção que estejam a ser descritos existirem em formato digital, a EAD pode incluir apontadores para esses documentos digitais. [NISO, 2001]

3.6. Objectos Visuais – CDWA e VRA

A metadata necessária para descrever objectos visuais como pinturas ou esculturas requer uma aproximação especial. A *Art Information Task Force* (AITF) [REF22], desenvolveu uma Framework conceptual para descrever e aceder a informação sobre objectos e imagens chamada *Categories for the Description of Works of Art* (CDWA) [REF23]. Foram definidas perto de 30 categorias, muitas com subcategorias associadas. Alguns dos elementos descritivos especializados incluem: Orientação, Dimensões, Estado, Inscrições, Tratamento de conservação e Históricos de exibições e empréstimos.

Tipicamente, as colecções de recursos visuais usados no ensino de história da arte e assuntos semelhantes não contêm as peças originais mas apenas slides ou fotografias das obras originais. A metadata para estes materiais tem então de acomodar a descrição de múltiplos níveis de recursos relacionados, como a obra original, um slide da obra e uma imagem digital do slide. As categorias VRA Core foram embutidas na

CDWA para definirem um único conjunto de elementos de metadata que possa ser usado para descrever a própria obra assim como as imagens da mesma.

A versão 3.0 das categorias VRA Core consiste em 17 elementos: *Record Type, Type, Title Measurements*, *Material*, *Technique, Creator*, *Date*, *Location*, *IDNumber*, *Style/Period*, *Culture*, *Subject*, *Relation*, *Description*, *Source*, e *Rights*. Tal como o Dublin Core, o esquema do VRA Core não especifica nenhuma sintaxe particular ou regras para representar o seu conteúdo.
Tanto o CDWA como o VRA dão importância ao uso de vocabulários controlados para os elementos especificados. Já são fornecidos alguns conjuntos de vocabulários pronto a ser usados, mas as comunidades de artes são encorajadas a desenvolverem novos conjuntos à medida que forem necessários. [NISO, 2001]

4. Modelo Generalista Proposto

Com base nos motores de pesquisa que foram analisados no capítulo 2, nos atributos que estes necessitam para funcionarem e nas novas tendências na pesquisa de informação online de seguida será apresentada uma proposta para um modelo generalista de uma página Web. Este modelo tem de armazenar o máximo, ou no melhor dos casos todos, de atributos descritivos na sua estrutura para que os motores de pesquisa que possuam interfaces 3D possam retirar a informação que necessitam para representar a informação á luz do seu próprio paradigma.

4.1. Análise dos standards de metadata

Com base na informação recolhida no ponto anterior, podemos agora comparar os atributos disponibilizados pelos standards analisados com os atributos requisitados pelos motores de pesquisa para construir os seus interfaces 3D.

Tabela 4 - Tabela de comparação dos standards de metadata com os requisitos dos interfaces 3D

	Dublin Core	TEI	METS	MODS	EAD	CDWA/VRA
Categorias	✔	✘	✔	✔	✘	✔
Texto do documento	✔	✔	✘	✔	✔	✘
Palavras-Chave	✔	✔	✔	✔	✔	✘
Vocabulário idêntico	✘	✘	✔	✔	✘	✘
Hiperligações de citação	✔	✘	✘	✔	✘	✘
Hiperligações directas	✔	✘	✘	✔	✘	✘
Transacções Financeiras	✔	✘	✔	✔	✔	✘
Predecessor	✔	✘	✘	✔	✘	✘
Ranking Externo	✘	✘	✘	✘	✘	✘

Como se pode verificar nos resultados obtidos na comparação, existem dois standards, Dublin Core e MODS, que conseguem responder a grande parte dos requisitos impostos pelas aplicações. Ambos os standards conseguem providenciar através dos seus atributos, informação suficiente para satisfazer a grande maioria de requisitos requeridos pelas aplicações, e como ambos são flexíveis em termos de adição de atributos, pode-se colmatar as falhas com a adição dos atributos em falta.
Sendo assim, apenas faz sentidos considerar estes dois standards para serem a base do modelo a propor. No ponto seguinte é apresentada uma comparação dos dois standards a fim de determinar qual será a melhor opção para o projecto proposto.

4.1.1. Dublin Core versus Metadata Object Description Schema

O Dublin Core é um dos standards mais amplamente conhecidos. Isto faz com que muitos sistemas o usem e estejam familiarizados com este. Mas a sua simplicidade começou a causar limitações e provou não estar a altura dos problemas encontrados. Por esta razão, foi desenvolvido outro standard, o MODS, que eclipsou, em parte, o Dublin Core com o seu conjunto de elementos mais rico.

O Dublin Core foi desenvolvido para guardar metadata sobre recursos que possam ser armazenados na Web, incluindo documentos em modo texto, imagem, vídeo ou em particular páginas Web. Possui um conjunto de elementos base, como Contribuidor, Editor ou Língua, que normalmente guardam valores bastante simples, como o nome do criador do elemento que está a ser descrito. Existem porém, elementos mais complexos, como o *Coverage*, usados para descrever o conteúdo do documento e não o documento em si.
Um dos problemas do Dublin Core é o facto de não estar claramente definido. Os valores usados para preencher os elementos muitas vezes não são retirados de colecções de valores bem definidas. Logo, se os valores dos elementos podem significar coisas diferentes para utilizadores diferentes, isto levanta um grave problema. Por consequência, falha na tarefa para a qual foi desenhado, a de providenciar termos bem definidos que possam ser usados por todos e que sejam uniformemente interpretados.

O MODS é um standard desenvolvido mais recentemente e que usa a linguagem XML para representar a sua estrutura. O MODS é frequentemente aclamado como o sucessor do Dublin Core.
O MODS possui mais elementos, e ainda mais importante, mais elementos claramente definidos, com a inclusão do acesso a listas de valores fechadas e aprovadas por autoridades reconhecidas globalmente. Como utiliza a habilidade do XML de expressar estruturas aninhadas ou embebidas, pode reter muita mais informação do que o Dublin Core. Com o acesso a termos fornecidos por autoridades, os valores dos elementos deixam de ser ambíguos e passam a ser globalmente interpretados da forma correcta, evitando assim qualquer tipo de confusão ou erro.

Exemplo do MODS:

```
<name type=”personal”>
      <namePart type=”family”>King</namePart>
      <namePart type=”given”>Bugs</namePart>
</name>
```

Neste exemplo, conseguimos retirar a informação sobre um indivíduo chamado Bugs King, que é claramente descrito no MODS como sendo uma pessoa com o primeiro nome Bugs e nome de família King.

Comparemos agora com o elemento *Contributor* do Dublin Core.

```
<dc:contributor>Bugs King</dc:contributor>
```

Apenas com o valor “Bugs King”, não se consegue determinar concretamente se nos estamos referir a uma pessoa, a uma organização ou outro tipo de contribuidor. [King, 2009]
O standard MODS é visto como uma solução às falhas do Dublin Core face ao problema de estandardizar a metadata descritiva. O MODS oferece, com um sucesso considerável, uma ponte entre a necessidade de diferentes requisitos de interoperabilidade e precisão, isto é, conseguir representar a informação de uma forma precisa e ao mesmo tempo garantir que o maior número de utilizadores, provenientes de diferentes áreas, a possa utilizar. [Gartner, 2003]

Pode-se então concluir que o standard MODS, devido a ao seu conjunto de elementos mais ricos e uso de valores estandardizados por autoridades oficiais, reduzindo assim

os problemas de ambiguidade e aumentando a legibilidade e utilidade da informação, é superior ao Dublin Core.
Foi então escolhido o standard MODS para servir de base do modelo proposto.

4.2. Definição do modelo

Nesta secção do documento é apresentada a proposta do modelo generalista de páginas Web. Como foi determinado anteriormente o modelo será baseado no standard MODS. Visto que este standard disponibiliza o melhor conjunto de campos necessários para construir interfaces gráficos em 3D e a sua estrutura é representada em XML, a sua interacção com as páginas Web será bastante facilitada, sendo que a linguagem XML é reconhecida globalmente.

Agora é necessário cruzar os atributos requisitados pelos motores de pesquisa com os elementos do standards MODS para ver quais são os campos que vão ser utilizados e se existe a necessidade de fazer alterações, acrescentar parametrizações ou até acrescentar campos novos.

4.2.1. Elementos

Nesta secção será explicado onde é que os vários atributos recolhidos da análise aos motores de pesquisa vão encaixar no nosso modelo. Cada atributo será mapeado para um, ou vários elementos do modelo conforme as necessidades apresentadas.

- **Categorias**

O campo eleito para representar as categorias no standard MODS será o elemento *genre*. Este elemento era usado pelo standard para guardar categorias que representam um determinado estilo, forma ou conteúdo de um documento, o que se enquadra na informação que pretendemos guardar.

Sobre este campo é preciso ter uma especial atenção no que toca aos motores de pesquisa que utilizem interfaces com árvores hiperbólicas. Para se construir uma árvore hiperbólica é necessário ter uma hierarquia de categorias bem definida e estruturada. Para esse efeito será necessário criar uma taxonomia com as categorias e restringir os valores do elemento *genre* com os valores da referida taxonomia. O standard MODS suporta este requisito com o uso de autoridades, que definem uma lista de valores fechada que um determinado campo pode reter. Desta forma podemos efectivamente forçar que os valores definidos fazem parte de uma hierarquia

previamente estabelecida e que pode ser representada pelas árvores hiperbólicas sem problemas.

- **Texto do documento**

O campo eleito para representar o texto do documento será o elemento *abstract*.

- **Palavras-Chave**

O campo eleito para representar o texto do documento será o elemento *subject* no seu subelemento *topic*. Neste último elemento encontram-se as palavras-chave mais importantes sobre o recurso.

- **Vocabulário idêntico**

Para representar este atributo o standard MODS faculta a possibilidade de definir várias vezes o mesmo elemento com valores diferente e em línguas diferentes. Logo podemos, por exemplo, ter um titulo principal definido em português e definir depois o mesmo titulo por outras palavras ou noutras línguas.

- **Hiperligações**

O campo eleito para representar as hiperligações presentes numa página Web no standard MODS será o elemento *relatedItem*.

- **Transacções Financeiras**

O standard MODS não tem nenhum elemento específico para guardar esta informação. Poderá ser utilizado o elemento *note* e inclui-se esta informação de uma forma não estruturada ou outra opção será usar o elemento *extension* que permite representar dados que não são suportados pelo standard MODS. Se a segunda opção for utilizada é possível definir um esquema em XML que represente a informação necessária e embeber essa estrutura dentro do elemento *extension*, estendendo assim o standard para que possa abranger a informação necessária de uma forma estruturada.

- **Predecessor**

O predecessor é um campo requisitado pelos interfaces das árvores hiperbólicas, mas como já foi referido anteriormente, com o uso de autoridades sobre o elemento *genre* é possível definir qual o lugar dentro da hierarquia de conceitos à qual o recurso pertence.

- ***Ranking* Externo**

O ranking externo refere-se à posição ocupada na lista de resultados, ou seja, à relevância de uma página, determinada por um motor de pesquisa externo. A relevância atribuída a uma página altera-se consoante o motor de pesquisa e os parâmetros de pesquisa utilizados. Logo, não faz sentido este atributo ser incluído no nosso modelo, já que não podemos atribuir uma relevância geral a uma página Web, sem sabermos qual é a base da pesquisa e sobre que motor externo esta é feita. Optámos por excluir este atributo do nosso modelo devido a esta ambiguidade. Os motores que utilizem este atributo específico terão de o continuar a usar como o fazem actualmente em conjunção com a informação fornecida pelo nosso modelo.

4.2.2. Apreciação global do modelo

Agora que todos os atributos requisitados foram mapeados para os respectivos elementos, já se pode construir a estrutura básica do nosso modelo.

Como já foi referido anteriormente, a informação relativa aos atributos utilizados pelos motores de pesquisa foi inferida do estudo que se fez sobre os mesmos. A informação disponibilizada nem sempre continha dados concretos sobre os atributos que os motores de pesquisa usam nas suas representações tridimensionais.

Foi necessário fazer algumas assumpções com base na análise visual dos interfaces dos motores e alguns indícios da sua documentação. Estas assumpções podem estar incompletas, ou até incorrectas, em alguns dos casos. Como consequência directa, o nosso modelo poderá ser insuficiente nalguns detalhes, atendendo a que o seu objectivo é conseguir suportar todo o tipo de motores de pesquisa com interface 3D.

No entanto, esta dificuldade pode ser contrariada. Se for detectada alguma dependência num atributo que não foi considerado, será uma questão de encontrar um outro elemento disponibilizado pelo standard MODS, onde o nosso modelo é baseado, e guardar lá a informação relativa ao novo atributo. No caso de não haver nenhum atributo MODS nativo que dê resposta às necessidades o modelo também tem a flexibilidade de colmatar essa limitação. Tal como já foi idealizado para o atributo de transacções financeiras, poderá ser usado o elemento *extension* para estender o nosso modelo de modo a ser embebida uma estrutura capaz de satisfazer os requisitos do novo atributo.

A capacidade de extensão é um dos pontos fortes do modelo, pois qualquer alteração dos requisitos dos motores de pesquisa poderá ser acrescentada para responder às novas necessidades.

Como todos os atributos foram mapeados com sucesso, podemos assumir que o nosso modelo vai conseguir suportar todos os motores de pesquisa com interface 3D estudados, cumprindo assim um dos principais objectivos deste trabalho.

5. Método de Instanciação do Modelo

Neste capítulo é abordado o problema de instanciar o modelo proposto. Agora que sabemos quais os campos chave que necessitamos para classificar uma página e temos um modelo onde guardar esses dados, é preciso primeiro proceder à extracção da informação das páginas. Isto não é tarefa simples, pois este é um dos principais obstáculos ao bom funcionamento do nosso modelo. Se os dados não forem correctamente recolhidos, ou não se conseguir recolher parte, ou no pior dos casos não se recolher nenhum dos dados precisos e o nosso modelo pouco ou nada vai servir para ajudar a classificar as páginas.

5.1. Técnicas de extracção de informação

Com o crescimento cada vez mais acelerado da Web, e consequentemente o aumento considerável do volume de informação disponível, é necessário criar métodos viáveis para extrair a informação relevante no meio de tanta oferta. A classificação por métodos convencionais já há muito que se tornou insustentável dado o volume de informação que tem de ser analisado.

A solução para este problema passa pela criação de ferramentas automatizadas que analisam os sites e extraem os pedaços de informação mais relevantes a fim de os poder catalogar. São usadas técnicas como o *Text Mining* [Even-Zohar, 2002; Rodrigues, 2008], e mais especificamente o *Web Mining* [Scime, 2004; Chang et all, 2001] dado o ambiente online onde se encontra a informação, para se extrair a informação em bruto das páginas Web. Após a extracção dos dados em formato bruto usam-se vários algoritmos, sejam eles de análise sintáctica, semântica, estatística, etc, para refinar os dados extraídos em informação que tenha valor para os utilizadores e aplicações.

5.2. Ferramentas de extracção de informação online

5.2.1. AeroText

O AeroText é um software de extracção de informação. Foi desenvolvido pela Lockheed Martin para dar resposta aos problemas de lidar com grandes quantidades

de informação não estruturada. O que este software faz é extrair informação relevante, como entidades, relações e eventos de textos não estruturados. Segundo os proprietários do sistema, este consegue identificar e extrair informação relevante de um texto com uma precisão igual ou superior á capacidade humana, além de encontrar relações ou eventos que de outra forma passariam despercebidas. Suporta múltiplas línguas e permite a interacção com outras ferramentas de descoberta de conhecimento. [REF13]

5.2.2. CATPAC

O CATPAC é um programa inteligente capaz de sumarizar textos e extrair os seus conceitos principais. O software não precisa de configurações adicionais e suporta várias línguas. Possui um registo significativo nas pesquisas sobre publicações dos maiores jornais mundiais.

A rapidez de processamento faz com que seja uma mais-valia quando se pretende analisar grandes quantidades de texto e não se dispõe de tempo nem de recursos para o fazer. Sendo totalmente automatizado o CATPAT não precisa de codificação extra para fazer o seu trabalho.

O CATPAC suporta qualquer língua que possa ser codificada em ASCII ou RTF. Foi desenvolvido por cientistas de renome internacional e é usado por todo o mundo em empresas, governos e pesquisas científicas. [REF14]

5.2.3. AlchemyAPI

A AlchemyAPI é um produto da Orchester8 [REF9]. É uma API que fornece aos detentores de conteúdos e programadores Web um conjunto de análise de conteúdos rico e ferramentas de anotação de metadata. Com a AlchemyAPI conseguimos expor, até um certo ponto, a informação semântica escondida nos textos, usando a extracção de entidades, frases e termos, categorização de documentos, detecção de língua, etc.

A AlchemyAPI usa tecnologias de processamento estatístico de linguagem natural e algoritmos de *machine learning* para analisar os conteúdos. Com isto é capaz de extrair a metadata semântica, como por exemplo, informação sobre pessoas, lugares, empresas, tópicos, línguas e etc. A API disponibiliza *endpoints*, acessíveis através da internet para fazer a análise dos conteúdos pretendidos, que podem estar em HTML, texto simples ou imagens digitalizadas de documentos.

Para usar a AlchemyAPI é necessário requisitar uma chave de acesso através do seu website.

Algumas das funcionalidades disponíveis na AlchemyAPI:

- Extracção de entidades: Identifica pessoas, empresas, organizações, eventos, cidades, características geográficas e outros tipos de entidades presentes em páginas HTML, documentos ou excertos de texto.
- Identificação de conceitos: Capacidade identificar conceitos de uma forma similar à humana. É capaz de fazer abstracção dos termos que encontra nos documentos e retirar um conceito comum entre estes (José Sócrates + Durão Barroso + Cavaco Silva = Governo Português ou Primeiros Ministros).
- Extracção de palavra-chave: Extrai os termos e tópicos mais importantes como palavras-chave para um documento.
- Categorização de tópicos / Classificação de textos: Capacidade de classificar um documento com um tópico geral. Isto permite classificar o conteúdo de um documento mediante uma taxonomia.
- Identificação automática da língua: Capacidade de identificar a língua de um texto ou conteúdo Web.
- Suporta 97 línguas.
- Extracção de Texto / Limpeza de página Web: Limpa automaticamente uma página Web de todo o conteúdo que não interessa, como hiperligações de navegação, publicidade, etc. Extrai artigos e textos chave de um documento, retornando desta forma o conteúdo que realmente interessa aos utilizadores.

A AlchemyAPI fornece diversos tipos de soluções dependendo das necessidades dos seus utilizadores. A AlchemyAPI FREE é a solução grátis e permite usar todos os serviços disponíveis na API, mas está limitada a 30.000 utilizações por dia. As soluções mais avançadas oferecem um limite superior de utilizações diárias, acordos de suporte e outras vantagens mas estão sujeitas ao pagamento de uma licença.

5.3. Protótipo de instanciação do modelo

Após a análise de algumas das ferramentas de extracção de informações disponíveis no mercado, optou-se por usar a AlchemyAPI como base do protótipo de instanciação do nosso modelo de dados. A AlchemyAPI foi escolhida por disponibilizar muitas das funcionalidades necessárias para preencher o modelo com a grande vantagem de ter

uma versão gratuita. O protótipo foi desenvolvido em .NET C# com recurso às classes disponibilizadas no SDK da AlchemyAPI da mesma linguagem.

A função principal do protótipo é extrair informação de uma página Web, utilizando os serviços disponibilizados pela AlchemyAPI, e guardar os resultados da extracção num ficheiro XML com a estrutura do modelo proposto.

De seguida será explicado como é que cada um dos atributos é extraído, tratado e posteriormente guardado no modelo XML.

- Categoria

A categoria da página é extraída usando uma funcionalidade disponibilizada pela AlchemyAPI, a categorização de tópicos. Esta funcionalidade permite dois tipos de entrada, o endereço da página Web ou um texto, e retorna a categoria respectiva. No que respeita às categorias suportadas pela API, estamos um pouco limitados, já que apresenta uma lista bastante curta e genérica. Isto faz com que algumas páginas não sejam classificadas correctamente ou nem sequer consigam ser classificadas.

A lista de categorias suportadas pela AlchemyAPI é a seguinte.

Tabela 5 - Categorias suportadas pela AlchemyAPI

Categoria	Descrição
Arts & Entertainment	Artes (Pintura, Escultura, etc.) & Entretenimento (Filmes, Musica, etc.) Noticias & Discussões.
Business	Negócios & Finanças, Comercio electrónico, etc.
Computers & Internet	Tecnologias de Informação (Computadores, Internet, Telecomunicações, etc.) Noticias & Discussões.
Culture & Politics	Politica, Cultura e Sociedade. Noticias & Discussões.

Gaming	*Gaming* (Jogos de computador, Jogos de vídeo, Jogos de tabuleiro) Noticias & Discussões.
Health	Saúde (Medicações, Tratamentos, Doenças, etc.) Noticias & Discussões.
Law & Crime	Leis e Crimes. Noticias & Discussões.
Religion	Religião. Noticias & Discussões.
Recreation	Actividades Recreativas (Navegar, etc.)
Science & Technology	Ciência (Física, Astronomia, etc.) Noticias & Discussões.
Sports	Desportos. Noticias & Discussões.
Weather	Meteorologia. Noticias & Discussões.

Os resultados para este atributo são guardados no elemento *genre* no modelo proposto.

Exemplos:

```
<mods:genre>sports</mods:genre>

<mods:genre>computer_internet</mods:genre>

<mods:genre>religion</mods:genre>

<mods:genre>science_technology</mods:genre>
```

- Texto do documento

Para este atributo a API também disponibiliza uma funcionalidade, que analisa uma página Web e extrai só os textos relevantes, descartando publicidades e outros textos irrelevantes.

Os resultados para este atributo são guardados no elemento *abstract* no modelo proposto.

Exemplos:

```
<mods:abstract xml:lang="en">Buy academic journals, books and online media at Springer. Choose from thousands of scientific, technology medical and business titles and view our range of services for authors, booksellers and librarians.</mods:abstract>

  <mods:abstract xml:lang="es">Comprar revistas académicas, libros y medios de comunicación en línea en Springer. Elige entre miles de títulos de la tecnología científica, médica y de negocios y ver nuestra gama de servicios para los autores, libreros y bibliotecarios.</mods:abstract>
```

- Palavras-Chave

Quanto às palavras-chave, são usados dois métodos distintos para as retirar. Numa primeira fase é usada a funcionalidade que extrai as palavras-chave e/ou terminologias dos textos de uma página Web. Após termos os resultados da primeira fase, a API disponibiliza uma funcionalidade que extrai conceitos de textos, que é usada para retirar os conceitos da combinação das palavras-chave e do texto extraídos previamente. Desta forma não só temos as palavras-chave normais, como também se retiram conceitos novos através da combinação das palavras-chave e do texto da página Web.

Os resultados para este atributo são guardados no elemento *topic* no modelo proposto.

Exemplos:

```
<mods:subject authority="keyword">
    <mods:topic>Copyright</mods:topic>
    <mods:topic>Book Series</mods:topic>
    <mods:topic>Artificial Intelligence</mods:topic>
    <mods:topic>Computer Science</mods:topic>
```

```
    <mods:topic>International Conference</mods:topic>
    <mods:topic>Volume Issue Page</mods:topic>
    <mods:topic>books</mods:topic>
...
```

- Vocabulário Idêntico

Para este atributo, a AlchemyAPI não oferece nenhuma funcionalidade. Termos idênticos e/ou traduzidos só fazem sentido para os campos de categoria, palavras-chave e texto da página. Para encontrar termos idênticos não foi encontrada nenhuma solução viável, mas com recurso à "*Google AJAX Language API*" [REF15], foi criada uma funcionalidade capaz de traduzir texto para outras línguas, desde que seja suportada pelo sistema de tradução do Google. Esta funcionalidade é usada apenas para traduzir o texto extraído da página, enquanto a categoria e palavras-chave podem incluir termos que não devem ou não podem ser traduzidos e não há meio de os distinguir dos restantes. Por exemplo, se uma palavra-chave for o nome de uma pessoa, esta não deve ser traduzida, mas a funcionalidade de extracção de palavras-chave não fornece esse tipo de distinção. Desta forma, o texto extraído da página é traduzido para outras línguas o que faz com que pesquisas feitas sobre o mesmo assunto mas numa língua diferente considerem esta página como um resultado válido.

Os resultados para este atributo são guardados no elemento *abstract* marcados com a língua para que foram traduzidos.

Exemplos:

```
<mods:abstract xml:lang="en">MODS Schemas (Metadata Object Description Schema: MODS)</mods:abstract>

<mods:abstract xml:lang="fr">Schémas MODS (Metadata Object Description Schema: MODS)</mods:abstract>

<mods:abstract xml:lang="ar">المخططات إنتاجيـــة (إنتاجيـــة : القطعـــة وصـف الفوقيـــة مخطط)</mods:abstract>

<mods:abstract xml:lang="ja">
モッズスキーマ（メタデータオブジェクトの概要スキーマ：モッズ）</mods:abstract>

<mods:abstract xml:lang="pt-PT">Esquemas MODS (Metadata Object Description Schema: MODS)</mods:abstract>
```

```
<mods:abstract xml:lang="ru">MODS схем (Схема метаданных Описание объекта:
MODS)</mods:abstract>
```

- Hiperligações

É usada a funcionalidade respectiva da API para extrair as hiperligações de uma página Web. Esta funcionalidade apenas retorna o texto da hiperligação e o seu endereço, o que por si só não transmite nenhuma informação concreta sobre o conteúdo da página referente à hiperligação. Agora que se extraiu os endereços das hiperligações presentes na página, podemos usa-los para recolher informação sobre as páginas que estes representam. Ao recolher esta informação extra, podemos ligar a página principal com tópicos que estão presentes nas páginas secundárias acessíveis desde a primeira. Desta forma uma página pode ganhar mais relevância, pois para além de nos basearmos apenas nos tópicos desta, temos também acesso aos tópicos das suas hiperligações e estabelecer relações entre todas as páginas, obtendo assim resultados mais precisos para a pesquisa efectuada.

Os resultados para este atributo são guardados no elemento *relatedItem* no modelo proposto. Em cada elemento vai ser guardada a informação recolhida a partir do endereço da hiperligação que está a ser tratada. São preenchidos todos os elementos acima referidos para cada hiperligação, excepto a para o elemento *relatedItem*, pois não faz muito sentido recolher dados de hiperligações de uma hiperligação, dado que podemos estar a recolher dados de uma página que já nada tem a ver com a página principal e isso poderá produzir maus resultados.

Exemplos:

```
<mods:relatedItem displayLabel="Guidance for MODS record creation"
xlink:href="http://www.loc.gov/standards/mods/mods-guidance.html"><mods:titleInfo>
    <mods:title>Guidance: Metadata Object Description Schema: MODS (Library of
Congress)</mods:title>

  </mods:titleInfo>

  <mods:genre>computer_internet</mods:genre>

  <mods:subject authority="local">

    <mods:topic>Term Source Codes</mods:topic>
```

```
    <mods:topic>Term List</mods:topic>
    <mods:topic>MODS User Guidelines</mods:topic>
    <mods:topic>XML Documents Book</mods:topic>
    <mods:topic>Controlled vocabulary</mods:topic>
   </mods:subject>
   <mods:language>
    <mods:languageTerm type="code" authority:="iso639-2b">en</mods:languageTerm>
    <mods:languageTerm type="text">english</mods:languageTerm>
   </mods:language>
   <mods:name type="organization">
    <mods:namePart>Congress</mods:namePart>
   </mods:name>
   <mods:name type="country">
    <mods:namePart>Us</mods:namePart>
   </mods:name>
   <mods:name type="printmedia">
    <mods:namePart>Computer</mods:namePart>
   </mods:name>
   <mods:name type="technology">
    <mods:namePart>Web</mods:namePart>
   </mods:name>
   <mods:name type="personal">
    <mods:namePart type="family">Genre</mods:namePart>
    <mods:namePart type="given">MARC</mods:namePart>
   </mods:name>
   <mods:abstract xml:lang="af">Riglyne (Metadata Object Beskrywing Schema:
MODS)</mods:abstract>
```

```
   <mods:abstract xml:lang="ar">القطعـة وصـف مخطط الفوقيـــة) التوجيهيـــة الخطـوط :
إنتاجيـــة)</mods:abstract>

   <mods:abstract xml:lang="el">Κατευθυντήριες γραμμές (Metadata Schema Περιγραφή
Αντικειμένου: MODS)</mods:abstract>

   <mods:abstract xml:lang="en">Guidelines (Metadata Object Description Schema:
MODS)</mods:abstract>

   <mods:abstract xml:lang="pt-PT">Orientações (de objeto de esquema de metadados Descrição:
MODS)</mods:abstract>

  </mods:relatedItem>
```

6. Avaliação de resultados

Neste capítulo é avaliado o desempenho do protótipo de instanciação proposto.

A instanciação do modelo é um processo crítico e é através dela que a página instanciada se torna, ou não, visível, através do sistema de pesquisa, para o utilizador final. Quantos mais dados forem recolhidos com sucesso, de uma página, e transferidos para o modelo, melhor a probabilidade de este conter a informação que se está a pesquisar e consequentemente eleger a página como um resultado válido para a pesquisa. Outro ponto importante é a precisão dos dados recolhidos. Se os dados estiverem errados isso pode produzir resultados que não pertencem ao domínio do que se está a pesquisar, baixando assim os níveis de confiança do motor de pesquisa o que se tem de evitar a todo o custo.

O protótipo de instanciação desenvolvido consegue extrair, com sucesso, muita da informação necessária, mas em alguns casos a precisão dos dados recolhidos não é a melhor, devido a limitações da API usada, e pode levar a alguns erros de classificação.

Uma adição interessante foi a inclusão de um sistema de tradução para múltiplas línguas dos textos extraídos da página. Isto possibilita que pesquisas feitas com palavras-chave noutras línguas consigam encontrar referências em páginas que devido à barreira linguística não iriam ser consideradas. Depois, o utilizador se não domina-se a língua em que a página resultante está expressa poderia usar um dos vários tradutores de páginas disponíveis online para fazer a tradução para a sua língua nativa.
Abaixo é apresentado um exemplo da tradução dos textos, começando com a língua em qual inicialmente o texto foi escrito.

- Inglês

```
<mods:abstract xml:lang="en">Buy academic journals, books and online media at Springer. Choose from thousands of scientific, technology medical and business titles and view our range of services for authors, booksellers and librarians. </mods:abstract>
```

- Francês

<mods:abstract xml:lang="fr">Acheter des revues spécialisées, livres et médias en ligne chez Springer. Choisissez parmi des milliers de scientifiques, de la technologie médicale et de titres d'affaires et de voir notre gamme de services pour les auteurs, les libraires et les bibliothécaires.</mods:abstract>

- Espanhol

<mods:abstract xml:lang="es">Comprar revistas académicas, libros y medios de comunicación en línea en Springer. Elige entre miles de títulos de la tecnología científica, médica y de negocios y ver nuestra gama de servicios para los autores, libreros y bibliotecarios.</mods:abstract>

- Português

<mods:abstract xml:lang="pt-PT">Compre revistas acadêmicas, livros e mídia online no Springer. Escolha entre milhares de artigos científicos, títulos de tecnologia médica e de negócios e visualizar a nossa gama de serviços para os autores, livreiros e bibliotecários.</mods:abstract>

- Coreano

<mods:abstract xml:lang="ko">학술지, 도서 및 스프링어에서 온라인 미디어를 구입하세요. 과학, 기술, 의학 및 비즈니스 타이틀의 수천에서 선택 및 저자, 서점과 사서에 대한 서비스의 범위를 볼 수 있습니다.</mods:abstract>

Estas são apenas uma amostra das traduções feitas sobre uma página, no limite poderão ser traduzidas até 114 línguas suportadas pelo tradutor da Google.

Nota: A tradução foi feita com recurso ao tradutor automático disponibilizado gratuitamente pela Google, e poderão ocorrer alguns erros de tradução. Mas para a função que irá desempenhar não se prevê que isso possa ser considerado um problema.

6.1. Problemas e limitações

A limitação mais grave é a identificação da categoria da página. Visto que a AlchemyAPI tem uma lista de categorias muito limitada e genérica, e apenas retorna uma categoria por página, iremos ter categorias repletas de páginas que na realidade não estão a ser bem classificadas pois não foi encontrada uma categoria ideal para a página em questão. Este problema torna-se ainda mais grave quando estamos a lidar com motores de pesquisa que utilizem árvores hiperbólicas, que na sua natureza necessitam de uma boa estrutura de categorias hierárquicas para construir a sua estrutura. Uma solução para este problema seria usar uma ferramenta de *Text Mining* que conseguisse extrair categorias mais precisas para as páginas, e que permitisse passar uma taxonomia apropriada para o tipo de pesquisa que se está a fazer.

Vejamos agora um exemplo onde a classificação falhou:

```
<mods d1p1:mods="http://www.loc.gov/mods/v3" d1p2:xmlns:="http://www.w3.org/1999/xlink"
xmlns:="http://www.loc.gov/mods/v3" d1p3:xmlns:="http://www.w3.org/2001/XMLSchema-
instance" d1p4:xmlns:="http://www.loc.gov/standards/mods/v3/mods-3-2.xsd" ID:="MODS"
version:="3.2" xmlns:d1p4="schemaLocation" xmlns:d1p3="xsi" xmlns:d1p2="xlink"
xmlns:d1p1="mods">
 <mods:titleInfo>
  <mods:title>Nuno Escudeiro</mods:title>
 </mods:titleInfo>
 <mods:Location>
  <url dateLastAccessed="01-11-2010 18:18:26">http://www.dei.isep.ipp.pt/~nuno/</url>
 </mods:Location>
 <mods:genre>sports</mods:genre>
…
```

No exemplo exposto, foi extraída a informação de uma página pessoal de um docente do departamento de Engenharia do Instituto Superior do Porto, o Eng. Nuno Escudeiro. O conteúdo da página é maioritariamente sobre conceitos informáticos, sobre as actividades do docente na instituição de ensino e informações sobre publicações feitas pelo próprio. Como se pode ver no exemplo, o protótipo de instanciação catalogou incorrectamente o conteúdo da página como pertencendo à categoria de desportos. Isto demonstra a funcionalidade de catalogação de página da AlchemyAPI não é 100% fiável, provavelmente muito por culpa do seu pequeno leque de escolha no que toca às categorias.

Outro dos problemas encontrados foi na extracção das hiperligações de uma página. Neste ponto o ideal seria extrair apenas as hiperligações relevantes para o conteúdo principal e/ou conteúdos similares da página. No protótipo desenvolvido estão a ser extraídas todas hiperligações sem ter em consideração o destino final para onde estas apontam. Desta forma podem haver hiperligações que apontam para páginas que nada têm a ver com os conceitos da página principal, como páginas de publicidade ou de registo de utilizador, e que não trazem nenhum valor adicional para melhorar a relevância da página.

A única filtragem que foi implementada foi a de eliminar hiperligações que apontem para a mesma página para evitar redundância de informação. No entanto isto não resolve os problemas de incluir hiperligações irrelevantes e teve de ser pensada uma solução.

A solução encontrada foi a de incluir com cada hiperligação, uma descrição detalhada, dentro do possível, da página à qual se refere. A informação das hiperligações é extraída da mesma forma que para a página principal, excepto na parte da extracção das hiperligações. Desta forma acompanhamos as hiperligações de um detalhe, onde nos podemos basear para avaliar a relevância destas e consequentemente melhorar a relevância da página principal tornando também o modelo mais rico em informação.

Abaixo é apresentado um extracto da informação recolhida para uma hiperligação.

```
<mods:relatedItem displayLabel="Conversions"
xlink:href="http://www.loc.gov/standards/mods/mods-conversions.html">

   <mods:titleInfo>

     <mods:title>Conversions: Metadata Object Description Schema: MODS (Library of
Congress)</mods:title>

   </mods:titleInfo>

   <mods:genre>computer_internet</mods:genre>

   <mods:subject authority="local">

     <mods:topic>Dublin Core</mods:topic>

     <mods:topic>MODS Version</mods:topic>

     <mods:topic>MODS Official Web</mods:topic>

…

<mods:language>

     <mods:languageTerm type="code" authority:="iso639-2b">en</mods:languageTerm>

     <mods:languageTerm type="text">english</mods:languageTerm>
```

```
    </mods:language>

    <mods:name type="city">

      <mods:namePart>Dublin</mods:namePart>

    </mods:name>

    <mods:name type="organization">

      <mods:namePart>Congress</mods:namePart>

    </mods:name>

    <mods:name type="country">

      <mods:namePart>Us</mods:namePart>

    </mods:name>

    <mods:name type="fieldterminology">

      <mods:namePart>HTML</mods:namePart>

    </mods:name>

    …

    <mods:abstract xml:lang="en">Conversions (Metadata Object Description Schema:
MODS)</mods:abstract>

    <mods:abstract xml:lang="es">Conversiones (metadatos de objetos de esquema Descripción:
MODS)</mods:abstract>

    <mods:abstract xml:lang="it">Conversioni (Metadata Object Description Schema:
MODS)</mods:abstract>

     <mods:abstract xml:lang="pt-PT">Conversões (Metadata Object Description Schema:
MODS)</mods:abstract>

</mods:relatedItem>
```

7. Conclusões

O objectivo principal deste trabalho consistiu em desenvolver um modelo generalista para páginas Web a ser usado por motores de pesquisa com interfaces 3D, tendo em conta os requisitos que estes sistemas necessitam para funcionarem o mais eficientemente possível.

Durante o desenvolvimento deste trabalho foram analisados vários motores de pesquisa com interfaces em 3D com o intuito de descobrir quais os requisitos que estes necessitam para funcionarem correctamente, estudar os seus interfaces e comportamentos e ter uma noção do ponto de evolução deste tipo de tecnologias. Além das pesquisas sobre os motores de pesquisa foi necessário fazer um estudo sobre standards de metadata, que estão disponíveis actualmente, numa tentativa de retirar ideias e guias para a construção do modelo generalista funcional e que ao mesmo tempo seguisse um conjunto de normas que fosse aceite por vários sistemas de naturezas diferentes.

Com base nos estudos referidos acima, chegou-se à conclusão que o conjunto de requisitos que os motores de pesquisa necessitavam era melhor representado pelo standard de metadata MODS, ou melhor, numa versão ligeiramente modificada para acomodar as novas necessidades. Foi então construída a estrutura do nosso modelo generalista tendo como base o standard MODS e os atributos requisitados foram mapeados para os elementos do novo modelo.

Agora que a estrutura do modelo generalista estava definida, o próximo passo foi o de desenvolver um protótipo de instanciação capaz de ler a informação de uma página Web, extrair a informação relevante e carregar os dados para dentro do nosso modelo. Isto não foi tarefa fácil, pois embora existam várias ferramentas disponíveis que conseguem extrair informação das páginas Web, ou necessitam de licenças pagas para as poder utilizar, ou produzem resultados aquém das expectativas. Após alguma procura foi encontrada uma API que faz muito do trabalho que era necessário em termos de extracção de informação. Esta API, de seu nome AlchemyAPI, tinha também a grande vantagem de ter uma solução gratuita e uma boa documentação de suporte que ajudou imenso no processo de codificação do protótipo desenvolvido. Tendo agora uma ferramenta que permitia extrair a informação das páginas Web, começou o árduo trabalho de encontrar a melhor forma de preencher os vários elementos do nosso modelo de forma a providenciarem o máximo de informação

relevante possível para os motores de pesquisa. No final do desenvolvimento do protótipo, todos os atributos requisitados são extraídos e instanciados com sucesso no modelo, embora devido a algumas limitações da API usada não tenham, em alguns casos, os valores mais indicados ou correctos.

Com este modelo, os motores de pesquisa com interfaces 3D têm agora uma estrutura bem definida de onde podem extrair a informação necessária para apresentarem os resultados das suas pesquisas.

7.1. Limitações e trabalho futuro

Como já foi referido no capítulo anterior o protótipo de instanciação possui algumas limitações, sendo a mais grave a lista limitada de categorias na qual as páginas Web são classificadas. O modelo generalista em si poderia também ser estendido com ainda mais elementos que providenciassem mais informação sobre a página em questão, como por exemplo, a média de visitantes, o número de hiperligações, etc. Neste caso, seria também necessário modificar o protótipo de instanciação para que recolhesse informação para os novos elementos, o que poderia ser extremamente difícil, ou em alguns casos impossível de o fazer.

Em relação a trabalhos de melhoramento futuros, o principal seria melhorar o sistema de extracção de informação das páginas, principalmente no que toca às categorias. De preferência, como opção, o utilizador escolher uma taxonomia de categorias previamente e a categoria da página seria encontrada dentro da taxonomia providenciada. Dependendo dos resultados obtidos pelos motores de pesquisa, podia-se acrescentar mais elementos ao modelo que facilitassem a tarefa de apresentação dos resultados.

8. Bibliografia

Allen, J. Thomas. "*Researching Information*", Furman University. 2009

Bateira, Marcelo. "Visualização de consultas em motores de busca na Web". 2006.

Beckett, Dave. et all. "Expressing Simple Dublin Core in RDF/XML", 2001. http://dublincore.org/documents/dcmes-xml/ (último acesso em 06 de Novembro de 2010).

Brin, Sergey e Page, Lawrence. "*The Anatomy of a Large-Scale Hypertextual Web Search Engine*". Computer Science Department, Stanford University, Stanford. http://infolab.stanford.edu/~backrub/google.html, 1998.

Cannon, James W. "*Hyperbolic Geometry*", 1997

Cazella, Sílvio César. "Introdução a Mineração de Textos (KDT)". Unisinos, 2007.

Chang, George et all. "*Mining the World Wide Web: an information search approach*". 2001.

Cockburn e McKenzie. "*An Evaluation of Cone Trees*". People and Computers XIV-Usability Or Else! British Computer Society Conference on Human Computer Interaction, 2000

Cockburn e McKenzie. "*3D or not 3D?: evaluating the effect of the third dimension in a document management system*". Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01), 2001.

Cugini, John V. et all. "*Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces*". National Institute of Standards & Technology (NIST), 1999.

Even-Zohar, Yair. "*Introduction to Text Mining*". Automated Learning Group, National Centrer for Supercomputing Applications, Universidade de Illinois, 2002.

Flake, Gary. "*Making Sense of Mountains of Data"*, 2010. http://www.techreview.com/web/24645/ (último acesso em 6 de Novembro de 2010).

Gartner, Richard. "*MODS: Metadata Object Description Schema*". Oxford University, 2003.

Hearst, Marti A. "*Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy*". Xerox Palo Alto Research Center, 2002.

Hillmann, Diane I. "*Using Dublin Core*". Project Manager & Metadata Specialist National Science Digital Library Project at Cornell Department of Computer Science Cornell University Ithaca, New York, USA, 2000. http://dublincore.org/documents/2000/07/16/usageguide/, último acesso em 6 de Novembro de 2010.

Jaju, Ravindra. "*An Introduction to Text Mining*", 2004.

King, Roger. "The Dublin Core and the Metadata Object Description Schema: a look at namespaces", 2009. http://itknowledgeexchange.techtarget.com/semantic-web/the-dublin-core-and-the-metadata-object-description-schema-a-look-at-namespaces/ (último acesso em 06 de Novembro de 2010).

Kunze, John "*Encoding Dublin Core metadata in HTML*", 1999.

Kokkelink, Stefan e Schwdnzl, Roland. "Expressing Qualified Dublin Core in RDF / XML", 2002. http://dublincore.org/documents/dcq-rdf-xml/ (último acesso em 06 de Novembro de 2010).

Leighton, H. Vernon e Dr. Srivastava, Jaideep. "Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos", 1997.

NISO. "*Understanding Metadata*", NISO Press, 2001 - ISBN 1-880124-62-9.

Powell, Andy e Johnston, Pete. "*Guidelines for implementing Dublin Core in XML*". http://dublincore.org/documents/dc-xml-guidelines/ (último acesso em 06 de Novembro de 2010).

Robertson, G.G. et all. "*Conne Trees: Animated 3D Visualizations of Hierarchical Information"*. UIR-R-1991-06.

Rodrigues, Fátima. "Descoberta do Conhecimento – Text Mining". Departamento de Engenharia de Informática (DEI/ISEP), 2008.

Scime, Anthony. "*Web mining: applications and techniques*", 2004.

Weibel, S. L., e Lagoze, C. "*An element set to support resource discovery*". *International Journal on Digital Libraries, 1*, 1997.

Wiza, Wojciech. "*ADAPTIVE 3D INTERFACES FOR SEARCH RESULT VISUALIZATION*". Department of Information Technologies, The Poznan University of Economics, 2003.

Referências a sites:

[REF1] http://en.wikipedia.org/wiki/Hyperbolic_tree, último acesso em 23 de Outubro de 2010.

[REF2] http://www.silverlight.net/, último acesso em 23 de Outubro de 2010.

[REF3] http://purl.pt/201/1/, último acesso em 23 de Outubro de 2010.

[REF4] http://memory.loc.gov/ammem/awhhtml/awgc1/lc_subject.html, último acesso em 23 de Outubro de 2010.

[REF5] http://www.oclc.org/dewey/, último acesso em 23 de Outubro de 2010.

[REF6] http://www.loc.gov/standards/sourcelist/name-title.html, último acesso em 23 de Outubro de 2010.

[REF7] http://www.loc.gov/standards/iso639-2/, último acesso em 23 de Outubro de 2010.

[REF8] http://www.w3schools.com/tags/tag_meta.asp, último acesso em 23 de Outubro de 2010.

[REF9] http://www.orchestr8.com/, último acesso em 23 de Outubro de 2010.

[REF10] http://www.google.com, último acesso em 05 de Maio de 2010.

[REF11] http://www.bing.com, último acesso em 05 de Maio de 2010.

[REF12] http://www.yahoo.com, último acesso em 05 de Maio de 2010.

[REF13] http://www.lockheedmartin.com/products/AeroText/products.html, último acesso em 25 de Outubro de 2010.

[REF14] http://www.galileoco.com/N_catpac.asp, último acesso em 25 de Outubro de 2010.

[REF15] http://code.google.com/intl/pt-PT/apis/ajaxlanguage/documentation/, último acesso em 31 de Outubro de 2010.

[REF16] http://www.zakon.org/robert/internet/timeline/, último acesso em 05 de Maio de 2010.

[REF17] http://www.youtube.com, último acesso em 05 de Maio de 2010.

[REF18] http://www.amazon.com, último acesso em 05 de Maio de 2010.

[REF19] http://www.infowester.com/rss.php, último acesso em 05 de Maio de 2010.

[REF20] http://www.ebay.com, último acesso em 05 de Maio de 2010.

[REF21] http://www.w3.org/MarkUp/SGML/, último acesso em 06 de Novembro de 2010.

[REF22] “THE ART INFORMATION TASK FORCE”, http://eclipse.wustl.edu/~listmgr/imagelib/Oct1994/0003.html, último acesso em 06 de Novembro de 2010.

[REF23] “Categories for the Description of Works of Art”, http://www.getty.edu/research/conducting_research/standards/cdwa/, último acesso em 06 de Novembro de 2010.

[REF24] en.wikipedia.org/wiki/MARC_standards, último acesso em 06 de Novembro de 2010.

[REF25] http://www.loc.gov/standards/marcxml/, último acesso em 06 de Novembro de 2010.

[REF26] http://www.loc.gov/ead/, último acesso em 06 de Novembro de 2010.

[REF27] SPAM - ORIENTAÇÔES, http://en.wikipedia.org/wiki/Spam_(electronic), último acesso em 06 de Novembro de 2010.

[REF28] Dublin Core. http://dublincore.org/documents/usageguide/#introduction, último acesso em 06 de Novembro de 2010.

[REF29] Type Vocabulary. http://dublincore.org/documents/dcmi-type-vocabulary/, último acesso em 06 de Novembro de 2010.

[REF30] Internet Media Types. http://www.isi.edu/in-notes/iana/assignments/media-types/media-types, último acesso em 06 de Novembro de 2010.

[REF31] Date and Time Formats, W3C Note. http://www.w3.org/TR/NOTE-datetime, último acesso em 06 de Novembro de 2010.

[REF32] “Dublin Core Metadata Element Set, Version 1.1: Reference Description” - http://www.dublincore.org/documents/1999/07/02/dces/, último acesso em 06 de Novembro de 2010.

[REF33] HyperText Markup Language (HTML). http://www.w3.org/MarkUp/, último acesso em 06 de Novembro de 2010.

[REF34] Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://dublincore.org/documents/dces/, último acesso em 06 de Novembro de 2010.

[REF35] DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/, último acesso em 06 de Novembro de 2010.

[REF36] XHTML 1.1: Module-based XHTML W3C Recommendation, 2001. http://www.w3.org/TR/xhtml11, último acesso em 06 de Novembro de 2010.

[REF37] Namespace Policy for the Dublin Core Metadata Initiative (DCMI) http://dublincore.org/documents/dcmi-namespace/, último acesso em 06 de Novembro de 2010.

[REF38] A Proposed Convention for Embedding Metadata in HTML http://www.w3.org/Search/9605-Indexing Workshop/ReportOutcomes/S6Group2.html, último acesso em 06 de Novembro de 2010.

[REF39] AGLS Metadata Standard. http://www.naa.gov.au/records-management/create-capture-describe/describe/AGLS/index.aspx, último acesso em 06 de Novembro de 2010.

[REF40] MARC Standards. http://www.loc.gov/marc/, último acesso em 06 de Novembro de 2010.

[REF41] Metadata Object Description Schema. http://www.loc.gov/standards/mods/v3/mods-userguide-intro.html, último acesso em 06 de Novembro de 2010.

[REF42] 10x Marketing, http://www.10xmarketing.com/Learning-Center/Internet-Statistics/Search-Engine-Statistics.html, último acesso em 06 de Novembro de 2010.

Bibliografia de Imagens

[FG 1] - Data Mountain - http://www.infovis.net/imagenes/T1_N75_A2_DataMountain.jpg, último acesso em 06 de Novembro de 2010.

[FG 2] - Interface do VxInsight - http://www.infovis.net/imagenes/T1_N168_A947_VxInsight.gif, último acesso em 06 de Novembro de 2010.

[FG 3] - Exemplo de árvore hiperbólica - http://vw.indiana.edu/ivsi2004/jherr/hyperbolic.png, último acesso em 06 de Novembro de 2010.

[FG 4] - Árvore Hiperbólica do arroz da Embrapa - http://www.agencia.cnptia.embrapa.br/gestor/arroz/arvore/arroz.html, último acesso em 06 de Novembro de 2010.

[FG 5] - Exemplo do VR-VIBE - http://www.crg.cs.nott.ac.uk/research/technologies/visualisation/vrvibe/vrvibe-multiuser1-small.gif, último acesso em 06 de Novembro de 2010.

[FG 6] - Exemplo de uma árvore cónica - http://www.crg.cs.nott.ac.uk/research/applications/pits/mapper1.gif, último acesso em 06 de Novembro de 2010.

[FG 7] - Vista geral do interface Cat-a-Cone - http://www2.parc.com/istl/projects/ia/papers/cac-sigir97/catacone.jpg, último acesso em 06 de Novembro de 2010.

[FG 9] - Vista Global do interface do NIRVE - http://www.itl.nist.gov/iaui/vvrg/cugini/uicd/gallery/sph-3d-ship.gif, último acesso em 06 de Novembro de 2010.

[FG 10] - Vista detalhada de um conjunto no NIRVE - http://www.itl.nist.gov/iaui/vvrg/cugini/uicd/gallery/sph-3d-ship-detail.gif, último acesso em 06 de Novembro de 2010.