



Universidade do Porto
Faculdade de Engenharia
FEUP



Artur Jorge da Silva Correia

O Uso de Metodologias Baseadas em *Fuzzy Sets* e *Rough Sets* na Recuperação de Informação

MGI
Porto, 2003

UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA

**O uso de metodologias baseadas em *fuzzy sets* e
rough sets na recuperação de informação**

Artur Jorge da Silva Correia

Licenciado em Ciência de Computadores pela Faculdade de Ciências da
Universidade do Porto

Dissertação submetida para satisfação parcial dos requisitos do grau de mestre em
Gestão de Informação

Dissertação realizada sob a supervisão de
Professora Doutora Maria Cristina de Carvalho Alves Ribeiro
e de
Professor Doutor Manuel António Cerqueira da Costa Matos

Porto, 22 de Junho de 2004

À Cristina e aos meus pais, Manuel e Laura

Resumo

A teoria dos conjuntos difusos e a teoria dos conjuntos aproximativos não são tópicos novos na área da recuperação de informação. No entanto, a combinação das duas teorias não tem sido objecto de um estudo aprofundado na área. A motivação desta pesquisa foi a exploração desta combinação num sistema experimental de recuperação de informação no qual foram aplicadas, em diferentes fases do processo, as teorias dos conjuntos difusos e dos conjuntos aproximativos. Uma estratégia de indexação difusa foi usada para a descrição de documentos, baseada num tesouro criado a partir da colecção de documentos com recurso a uma medida estatística. Os documentos recuperados foram agrupados em clusters usando uma abordagem no domínio da teoria dos conjuntos aproximativos. Os algoritmos foram testados com recurso a colecções de teste. Defende-se nesta dissertação que as teorias citadas são aplicáveis à resolução dos problemas de recuperação em vários dos seus estádios. Apresentam-se os fundamentos das duas teorias em estudo, uma análise dos modelos de recuperação actuais, a arquitectura de um sistema de recuperação proposto, os detalhes da sua implementação, os resultados obtidos e as conclusões para as diferentes abordagens.

Os resultados obtidos não foram inteiramente satisfatórios, tendo sido efectuados desenvolvimentos sobre o protótipo inicial que conduziram a uma melhoria na precisão obtida. No entanto, esta continua a ser muito baixa comparativamente a outros métodos. As conclusões apontam para a necessidade de serem efectuados testes suplementares com recurso a outras colecções de teste que apresentem menor uniformidade temática e a outras medidas que produzam uma melhor definição da semântica dos termos. Foi estabelecido que as teorias dos conjuntos difusos e aproximativos são aplicáveis na recuperação de informação, devendo um sistema de recuperação fundamentado nestas teorias apresentar uma maior eficácia quando usado em colecções em que documentos de dimensão adequada sejam descritos por termos suficientemente heterógeneos.

Abstract

Rough sets and fuzzy sets are not a new topic in Information Retrieval (IR) and have been applied in different ways in the field, in order to understand their appropriateness to IR models. However, the combinations of these two theories in this domain have not been extensively studied. The motivation of this research is to explore one such combination in an experimental IR system. An information retrieval system has been designed and implemented applying the theories of fuzzy sets and rough sets in different parts of the retrieval process. A fuzzy indexing strategy has been used to describe documents, based on a fuzzy thesaurus generated from the document collection using a statistical measure; retrieved documents were clustered using a rough set approach. The algorithms were tested using standard information retrieval test collections. This thesis presents an outline of both theories, the current logical models proposed in the IR field, the architecture of the system, implementation details, test results, and conclusions for the different approaches.

The conclusions point out that further testing should be made using different measures for the fuzzy indexing scheme and different testing collections, in order to induce a more adequate term semantic, since the results obtained were not completely satisfactory, due to a low precision induced by a relatively high number of retrieved documents. Improvements have been made into the original prototype that induced a higher precision. Despite the effort, the precision obtained is still very low compared to other methods, namely the space vector model implemented on the SMART system. However, a better effectiveness can be establish using the combinations of both theories if using proper documents description by means of usage of more heterogeneous terms.

Agradecimentos

Gostaria de agradecer:

À minha orientadora, Professora Doutora Cristina Ribeiro. As suas indicações, apoio e permanente incentivos foram essenciais para a conclusão desta dissertação.

Ao meu co-orientador, Professor Doutor Manuel Matos pelas indicações, apontamentos pertinentes. As suas ideias permitiram colocar em perspectiva o meu trabalho.

Ao Professor Doutor Mark Sanderson da Universidade de Sheffield pelos comentários e incentivo.

À minha empresa, PT - Sistemas de Informação S.A., em particular na pessoa da Eng^a. Margarida Almeida, que criou condições internas para a efectiva concretização desta dissertação.

Índice

1	Introdução	1
2	A recuperação de informação	5
2.1	Interesse da abordagem.....	5
2.2	Enquadramento e conceitos	6
2.3	Processos básicos.....	9
3	Modelos na recuperação de informação	13
3.1	Definição formal de modelo	13
3.2	Comparação de modelos.....	14
3.2.1	Modelo booleano	14
3.2.2	Modelo vectorial	15
3.2.3	Modelo probabilístico	18
3.2.4	Modelo baseado na teoria dos conjuntos difusos.....	20
3.2.5	Modelo P -norm.....	22
3.2.6	Modelos alternativos.....	24
4	Fundamentos teóricos das teorias dos conjuntos difusos e aproximativos	27
4.1	Conceitos básicos da teoria dos conjuntos difusos.....	27
4.1.1	Conjunto difuso.....	29
4.1.2	Propriedades e operações.....	30
4.1.3	Relações entre conjuntos difusos e operadores.....	34
4.1.4	Exemplo	38
4.2	Conceitos básicos da teoria dos conjuntos aproximativos.....	40
4.2.1	Relação de indiscernibilidade	41
4.2.2	Aproximação de um conjunto.....	41
4.2.3	Aproximação das classificações	43
4.2.4	Precisão da classificação.....	44

4.2.5	Qualidade da classificação	44
4.2.6	Redução de atributos	45
4.2.7	Exemplo de aplicação da teoria dos conjuntos aproximativos	45
4.3	Combinação das teorias dos conjuntos aproximativos e difusos	50
4.3.1	Conjunto aproximativo difuso (<i>rough fuzzy set</i>)	51
4.3.2	Conjunto difuso aproximativo (<i>fuzzy rough set</i>)	52
4.3.3	Conjunto difuso e espaço difuso aproximativo	53
5	Combinação das teorias dos conjuntos aproximativos e difusos na recuperação de informação	55
5.1	Motivação para a combinação de duas visões distintas	55
5.2	Propostas de combinação	57
5.3	Medida de relevância	62
5.4	RST – Arquitectura do protótipo	67
5.4.1	Representação do documento	68
5.4.2	Representação da interrogação	69
5.4.3	Análise léxica	69
5.4.4	Tesouro	70
5.4.5	Comparação	71
5.4.6	Agrupamento em clusters	71
6	RST1 – Protótipo de recuperação de informação	73
6.1	Análise léxica	73
6.1.1	Ficheiro invertido	76
6.1.2	Ficheiro directo de termos e índice da colecção	78
6.2	Construção do tesouro	79
6.2.1	Matriz de similaridade e classificação difusa de documentos	80
6.3	Comparação	82
6.4	Agrupamento em clusters	84
6.4.1	Métodos de agrupamento em clusters	85
6.4.2	Abordagem implementada	88
6.4.3	Medida e matriz de similaridade	89

6.5	Detalhes de implementação	90
7	Testes e resultados	93
7.1	Introdução	93
7.2	Colecções de teste e definição dos testes.....	94
7.3	Testes iniciais.....	95
7.4	Análise dos componentes do sistema RST1	99
	Análise léxica.....	99
	Co-ocorrência e tesauro	105
	Comparação entre documentos da colecção e interrogação	110
	Agrupamento em clusters	114
7.5	RST2 - Evolução do sistema RST1	117
	Sistema RST2: qualidade da recuperação.....	118
8	Contribuição e conclusões	125
8.1.	Contribuição.....	125
8.2.	Conclusões	126
8.3.	Actividades futuras	128
	Referências e bibliografia.....	131

Índice de figuras

Figura 1	Processo de Recuperação de Informação	10
Figura 2	Representação das aproximações de um conjunto	43
Figura 3	Arquitectura do sistema de recuperação de informação RST	68
Figura 4	Autómato do tipo DFA para identificação de <i>tokens</i>	74
Figura 5	Estrutura interna em memória de um ficheiro invertido	77
Figura 6	Matriz de similaridade simétrica	81
Figura 7	Processo de comparação entre documento e interrogação	84
Figura 8	Números de documentos relevantes para as 20 primeiras interrogações	96
Figura 9	Documentos recuperados por interrogação com diferentes classificações difusas	97
Figura 10	Frequência dos termos da colecção CRAN	100
Figura 11	Frequência dos termos da colecção com radicalização	101
Figura 12	Teste comparativo de recuperação de documentos	102

Figura 13	Precisão entre as duas variantes com alteração da lista de paragem	103
Figura 14	Precisão entre as duas variantes com alteração da lista de paragem	104
Figura 15	Documentos recuperados para diferentes parametrizações	108
Figura 16	Distribuição dos valores de co-ocorrência dos termos	108
Figura 17	Recuperação de Documentos com diferentes tesauros	110
Figura 18	Clusters criados por interrogação para 0.10 como valor de dissimilaridade	115
Figura 19	Clusters criados para a resposta 1	116
Figura 20	Distribuição de documentos num cluster	116
Figura 21	Documentos recuperados por RST1 e RST2	118
Figura 22	Agrupamento em clusters sobre resultados do RST2	120

Índice de tabelas

Tabela 1	Sistema de Informação	46
Tabela 2	Sistema de Informação reduzido	47
Tabela 3	Sistema de Informação reduzido	48
Tabela 4	Taxas de compressão pela eliminação de stopwords	75
Tabela 5	Estrutura do ficheiro invertido	76
Tabela 6	Estrutura do ficheiro directo de termos	78
Tabela 7	Estrutura do ficheiro da classificação difusa de documentos	82
Tabela 8	Parâmetros de Lance-Williams	90
Tabela 9	Precisão Média obtida por diferentes métodos	98
Tabela 10	Valores de co-ocorrência do exemplo	105
Tabela 11	Valores de co-ocorrência definidos com medida BR _x	106
Tabela 12	Valores de co-ocorrência definidos com a medida BR _y	107
Tabela 13	Classificação difusa dos documentos do exemplo	111

Tabela 14	Classificação difusa com medida BRz	112
Tabela 15	Recuperação de documentos para interrogação q_1	112
Tabela 16	Precisão média sobre a colecção da CRAN para diversos métodos	119
Tabela 17	Valores de similaridade para a colecção da CRAN	122

1 Introdução

Esta dissertação investiga o uso de metodologias baseadas nas teorias dos conjuntos difusos e dos conjuntos aproximativos na recuperação de informação. A teoria dos conjuntos difusos e a teoria dos conjuntos aproximativos não são tópicos novos na recuperação de informação. No entanto, a combinação das duas teorias não tem sido objecto de um estudo aprofundado. A principal motivação desta investigação é a exploração desta combinação num sistema experimental de recuperação de informação no qual são aplicadas, em diferentes fases do processo, as teorias dos conjuntos difusos e dos conjuntos aproximativos. Consequentemente, esta dissertação tem como objectivos:

- A análise das potencialidades da teoria dos conjuntos difusos e da teoria dos conjuntos aproximativos na área da recuperação de informação.
- A implementação de um sistema de recuperação de informação com adopção de conceitos teóricos propostos pelas duas teorias.
- Um estudo sobre aplicabilidade e eficiência das teorias referidas em colecções de texto a partir do sistema experimental implementado.

A área da recuperação de informação (*information retrieval*) está historicamente associada à investigação da representação, armazenamento, acesso e recuperação de informação em sistemas automatizados de manipulação de documentos. Contrariamente a aplicações informáticas desenvolvidas com recurso a bases de dados relacionais, um sistema de recuperação de informação não requer uma estrutura complexa e rígida dos dados e permite pesquisas numa linguagem próxima da linguagem natural. Se a génese da recuperação de informação está associada aos primeiros computadores, esta evoluiu consideravelmente nas últimas décadas, pois inovações tecnológicas como a Internet apresentaram novos desafios na recuperação de documentos. Os sistemas iniciais requeriam a intervenção de especialistas ou investigadores que detinham um apurado conhecimento das suas colecções, alcançando os objectivos propostos: recuperação de documentos relevantes para a pesquisa efectuada. Hoje em dia, em vez de investigadores

especializados temos utilizadores sem formação específica, detendo pouco ou nenhum conhecimento sobre as vastas colecções de documentos existentes em diversos suportes (texto, áudio, imagens,...). Por um lado, a tradução de uma necessidade de informação numa interrogação é intrinsecamente imprecisa. Por outro, de uma colecção controlada de documentos evoluímos para algo em constante mutação, difuso, sem classificação prévia por um especialista, em que os documentos são apresentados aos utilizadores como aproximações da resposta à interrogação. Entre os vários modelos matemáticos adoptados na tentativa de responder a estas questões, os modelos associados à teoria dos conjuntos aproximativos e à teoria dos conjuntos difusos têm sido analisados como possíveis soluções. A teoria dos conjuntos aproximativos proposta por Pawlak nos inícios dos anos 80, baseia-se na existência da relação de indiscernibilidade dos objectos, tendo sido aplicada com sucesso em várias aplicações (Slowinsky, 1992). A teoria dos conjuntos difusos, apresentada por Zadeh (1965), fornece um enquadramento mais apropriado do que a teoria dos conjuntos clássicos para captar propriedades que se aplicam aos objectos de forma não rígida.

A motivação para o desenvolvimento de um sistema de recuperação de informação é responder às necessidades de informação do utilizador disponibilizando uma interacção simples e natural com a colecção de documentos de forma a que o utilizador, mesmo sem possuir conhecimento detalhado sobre a colecção, consiga recuperar documentos relevantes. O desenvolvimento recente nos sistemas de recuperação de informação vão no sentido de lidar com utilizadores não familiarizados com as colecções e permitir interrogações que não requeiram expressões complexas, como as de lógica booleana. A indexação automática, por seu lado, recorre a métodos estatísticos baseados em co-ocorrência que estimam a contribuição de termos para os documentos. Estes métodos, porém, não têm em consideração o desconhecimento que o utilizador apresenta face ao vocabulário da colecção e a ambiguidade que os termos da interrogação apresentam, muitas vezes, face à sua real necessidade de informação. A teoria dos conjuntos difusos apresenta-se, neste âmbito, como a resposta natural para o tratamento desta problemática de indexação. De facto, possuindo a capacidade de expandir os conceitos para além da sua concepção binária, esta teoria promove a possível correcção da interrogação do utilizador

expandindo os conceitos introduzidos na fase de comparação entre termos e documentos de um sistema de recuperação. Sabe-se que o utilizador típico de um motor de busca apenas analisa os primeiros 20 ou 30 documentos, ainda que dezenas ou mesmo centenas de milhares de documentos são recuperados. Neste âmbito, a teoria dos conjuntos aproximativos, pela sua capacidade de definição de estruturas, classes de objectos, apresenta-se como uma forma de organizar os documentos nas respostas obtidas num sistema de recuperação. A definição de estruturas de indexação associadas aos termos de uma colecção e a classificação de documentos, são áreas onde as duas teorias têm sido habitualmente aplicadas no processo de recuperação. A construção de tesauros e os métodos de agrupamento em clusters são dois exemplos do seu uso.

No decorrer desta dissertação são investigadas e aplicadas diversas técnicas na área da recuperação de informação para tentar responder a esta motivação. A inexistência de referências da implementação de algumas das técnicas usadas ou da conjugação destas sobre colecções de teste, da obtenção de resultados conclusivos pela investigação efectuada até ao momento, reveste esta investigação de um grau considerável de pioneirismo. Este pioneirismo constituiu um estímulo adicional para abordar e ultrapassar as diferentes dificuldades no decurso desta dissertação sendo igualmente essencial para, a partir do sistema experimental modular obtido, desencadear a realização de múltiplos testes para comparação de diferentes abordagens.

Esta dissertação encontra-se organizada em oito capítulos, no fim dos quais é apresentada a bibliografia. Uma descrição breve dos capítulos é efectuada em seguida.

Capítulo 2: este capítulo apresenta uma introdução sobre a área da recuperação de informação efectuando um enquadramento histórico da evolução desta área. Descreve-se a terminologia, os processos presentes num sistema de recuperação de informação e o interesse da abordagem proposta.

Capítulo 3: este capítulo efectua a caracterização e comparação formal entre os modelos lógicos mais usados na recuperação de informação. Apresenta-se uma definição formal de modelo e a sua importância para a tarefa de recuperação.

Capítulo 4: os fundamentos teóricos das teorias dos conjuntos difusos, dos conjuntos aproximativos e a combinação entre as duas teorias são apresentados neste capítulo.

Efectua-se uma descrição das propriedades, operações e relações existentes nos conjuntos difusos e nos conjuntos aproximativos, separadamente. No fim deste capítulo é descrita a combinação que se pode estabelecer entre as duas teorias, em termos formais.

Capítulo 5: este capítulo apresenta a motivação para a combinação das duas teorias em estudo no âmbito da recuperação de informação. Descrevem-se as propostas de combinação, as medidas de relevância adoptadas neste domínio, sendo introduzido uma arquitectura - RST - para a implementação de um protótipo de recuperação de informação que combina as duas teorias.

Capítulo 6: este capítulo descreve o ambiente para a implementação do protótipo de recuperação de informação, os principais processos, estruturas e algoritmos usados. Os métodos de agrupamento em clusters são analisados, sendo descritos detalhes de implementação como sejam os componentes de hardware e software usados.

Capítulo 7: os testes e resultados obtidos pelo protótipo RST1 são o objecto de estudo neste capítulo. Os resultados são analisados e descrevem-se os principais problemas detectados na recuperação de informação efectuada pelo protótipo. Desta análise surge uma evolução do protótipo RST1, RST2.

Capítulo 8: este capítulo apresenta as conclusões finais e a contribuição efectuada por esta dissertação na área da recuperação de informação. Descrevem-se actividades futuras a serem realizadas no âmbito do protótipo desenvolvido.

2 A recuperação de informação

Este capítulo descreve o interesse suscitado por esta área de investigação, os conceitos e o enquadramento da recuperação de informação e uma descrição dos seus principais conceitos. São descritos os processos básicos presentes num sistema de recuperação, a sua função e importância. A importância da recuperação de informação para a evolução tecnológica é defendida com base no facto que hoje em dia vivemos na era da informação digital.

2.1 Interesse da abordagem

Os últimos três séculos foram dominados pelas revoluções tecnológicas. O século dezoito foi o século onde decorreram os desenvolvimentos mecânicos, como a máquina a vapor que desencadeou a revolução industrial. O século dezanove foi o tempo da electricidade, da difusão do telegráfo e do desenvolvimento da ciência tal como hoje a conhecemos. O século vinte foi, sem dúvida, marcado pelas tecnologias associadas à difusão e processamento de informação. Destas tecnologias, verificou-se a introdução da rádio, do telefone, o nascimento e crescimento exponencial do computador e a difusão das comunicações via satélite. Todas estas áreas foram suportadas, directamente ou não, pela mesma tecnologia: computadores, como suporte de informação digitalizada. Esta crescente integração entre máquina e pensamento continua, progressivamente, a alterar a própria essência da nossa existência. A importância da informação para os indivíduos, as organizações, os países é hoje universalmente aceite e ela é preponderante na evolução económica e social. A evolução exponencial da Web (*World Wide Web*) suportada na Internet fez com que esta se tenha tornado um repositório universal de informação, conhecimento humano e cultura, permitindo uma partilha sem precedentes de ideias e pensamentos. Neste novo universo, digital, um problema clássico ganha nova importância: a recuperação de informação.

Os sistemas tradicionais de armazenamento de informação estavam associados a especialistas ou investigadores cujo conhecimento apurado das colecções, permitia alcançar o objectivo: recuperação de documentos relevantes. Hoje em dia, em vez de investigadores temos utilizadores sem formação específica, detendo pouco conhecimento sobre as vastas colecções de documentos existentes em diversos suportes. A tarefa de traduzir uma necessidade de informação nos termos exactos de uma interrogação tem associada uma imprecisão intrínseca ao uso e riqueza da linguagem natural. Por outro lado, de uma colecção controlada de documentos evoluímos para algo em constante mutação, difuso, sem classificação prévia por um especialista em que documentos não classificados se apresentam aos utilizadores como aproximações aos termos pesquisados. Em resposta a uma pesquisa num motor de busca, não é invulgar um utilizador obter dezenas de milhares de documentos. Assim, os actuais serviços de pesquisa procuram formas de reduzir, de forma efectiva, os problemas de excesso de informação (*information overload*). Esta situação apresenta-se para muitos como um problema de resolução complexa. O principal obstáculo é a ausência de um modelo lógico que suporte a Web sendo frequente encontrar informação e estruturas de suporte de baixa qualidade. Estas dificuldades, a par com a importância e crescimento da Web, fazem com que a área da recuperação de informação tenha ganho maior atenção por parte da investigação científica, colocando as tecnologias aí usadas no centro das atenções.

Entre os vários modelos matemáticos adoptados na tentativa de responder às questões colocadas, os modelos associados à teoria dos conjuntos aproximativos e à teoria dos conjuntos difusos têm sido analisados, sendo os objectos de estudo deste trabalho.

2.2 Enquadramento e conceitos

A recuperação de informação está intimamente associada ao computador, pois a expressão “recuperação de informação” (*information retrieval*) surgiu com a automatização nos anos 50 de pesquisas anteriormente realizadas de forma manual. Nessa altura, a pesquisa era baseada na informação sobre um documento (meta-informação) em

vez da informação existente no documento (conteúdo), devido a restrições de armazenamento de dados.

Hoje em dia, um sistema de recuperação de informação tem associadas duas componentes: *software* e *hardware* e o seu objectivo é a recuperação de **documentos** para o utilizador de forma a satisfazer a sua **necessidade de informação**. Este processo é iniciado pelo utilizador que traduz a sua necessidade de informação pela introdução de **termos**, que constituem a sua **interrogação**, via uma interface entre este e o sistema de recuperação. Esta interrogação funciona como uma pista para a informação que os documentos devem conter para serem considerados na resposta.

A recuperação de informação surge, deste modo, como o processo de representação, armazenamento, organização e acesso a itens de informação (Baeza-Yates e Ribeiro-Neto, 1999). Neste âmbito, um sistema de recuperação de informação gera informação sobre documentos, **documentos recuperados**, em que alguns satisfazem a necessidade de informação do utilizador, **documentos relevantes**. A terminologia “recuperação de dados” é usada para descrever o processo de recuperação de informação realizado sobre dados previamente estruturados, de que é exemplo a resposta a interrogações nas bases de dados relacionais correntes.

Em geral, a informação recuperada não consistirá apenas nos documentos que contêm os termos usados na interrogação, dado que muitas vezes tal não é suficiente para satisfazer a necessidade de informação. O sistema de recuperação deve actuar como um instrumento que interpreta a interrogação para determinar, da forma mais aproximada possível, a necessidade de informação do utilizador, associando a informação existente entre os diversos documentos da **coleção** com o objectivo final de obter:

- uma alta percentagem de documentos recuperados que são relevantes. Esta percentagem designa-se por **precisão**:

$$\text{Precisão} = \frac{|Ra|}{|A|}$$

- uma alta percentagem de documentos relevantes nos documentos recuperados, designada por **recuperação**, tendo em conta o conjunto de documentos relevantes existentes na base de pesquisa:

$$\text{Recuperação} = \frac{|Ra|}{|R|}$$

onde R é o conjunto de documentos relevantes, A é o conjunto de documentos recuperados e Ra é o conjunto de documentos resultante da intersecção dos conjuntos R e A .

Estas medidas básicas têm sido adoptadas para avaliar o desempenho dos sistemas de recuperação, existindo um conjunto de medidas alternativas que têm sido apresentadas ao longo dos anos.

Um sistema ideal de recuperação de informação deve apresentar todos os documentos relevantes existentes na colecção de documentos e só esses. No entanto, em condições reais, este objectivo é difícil de atingir e nos sistemas existentes uma maior precisão corresponde tipicamente uma menor recuperação.

Uma base de dados de documentos para pesquisa sejam estes texto, áudio, imagem ou multimédia, é designada **colecção**. Existem diversas colecções de testes de referência, como sejam as colecções da TREC, CACM, CRAN, estabelecidas para execução de testes em sistemas de recuperação de informação. Uma colecção de teste é assim constituída por uma colecção de documentos, um conjunto de necessidades de informação ou interrogações e uma classificação de relevância dos documentos de acordo com essas necessidades. As colecções de referência são importantes pois permitem estabelecer comparações entre sistemas de recuperação, sendo um instrumento essencial no campo da investigação. As colecções podem ser variadas no seu conteúdo ou muito especializadas dependendo do objectivo proposto pelo sistema. O conjunto de colecções mais usadas no passado recente são as colecções da TREC, que tiveram origem nas conferências *Text REtrieval Conferences*, iniciadas em 1992. Estas permitem, numa metodologia de análise comparativa de sistemas de recuperação, uma medida objectiva do alcance dos objectivos propostos.

A evolução tecnológica ocorrida nos últimos anos no *hardware* tem vindo a permitir que, mesmo quando as colecções não são relativamente pequenas, os sistemas de recuperação possam adoptar uma estratégia de representação em texto integral dos documentos em vez de uma representação por indexação de termos ou texto parcial

(resumo). No entanto, quando esta assume a forma da Web ou de uma biblioteca digital com documentos disseminados por diversas colecções, os documentos são representados por um conjunto de termos o que reduz consideravelmente o espaço necessário. Esta evolução tem desencadeado uma permanente actualização dos requisitos a que um sistema de recuperação deve responder. De colecções de documentos relativamente estáticas e de pequenas dimensões, evoluímos para algo em constante mutação e crescimento, sendo objecto de investigação na área da recuperação de informação o desenvolvimento de sistemas distribuídos, com capacidade de correcção de erros, a pesquisa de informação em diversas línguas.

2.3 Processos básicos

Um sistema de recuperação de informação suporta processos básicos, alguns dos quais dependem directamente do modelo teórico adoptado, existindo três que, em qualquer sistema, estão sempre presentes: a representação do documento; a representação da interrogação (a necessidade de informação do utilizador) e a comparação destas duas representações com vista à resposta.

O gráfico da Figura 1, apresenta estes processos (traduzido de Baeza-Yates e Ribeiro-Neto,1999):

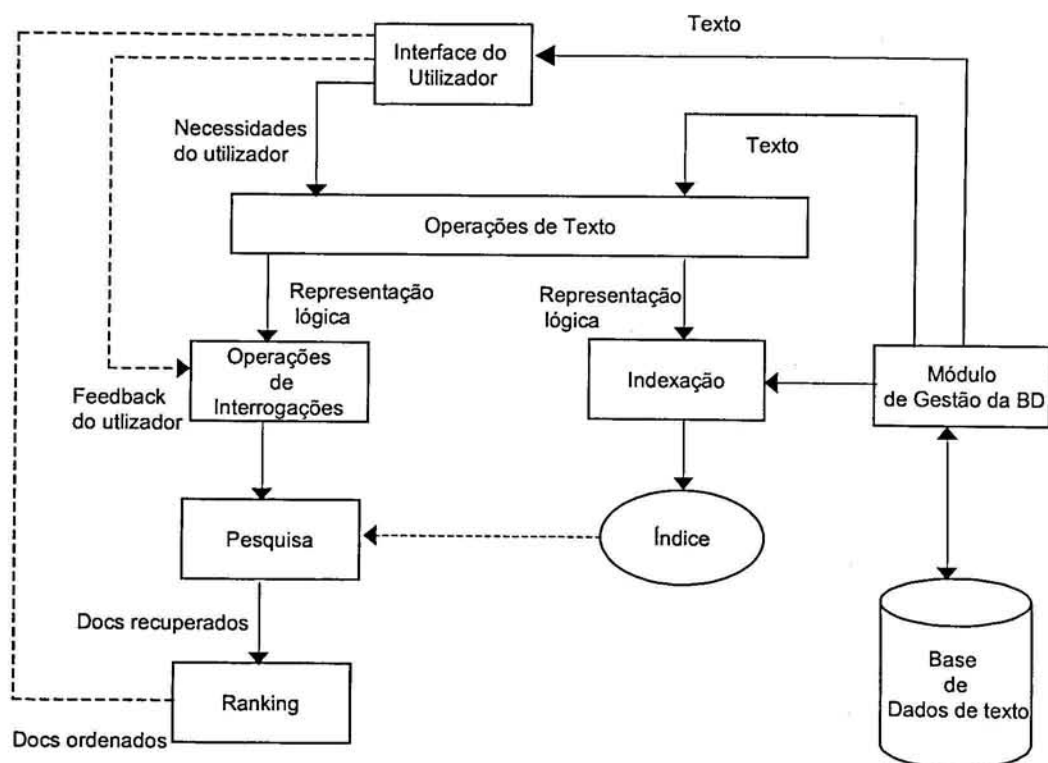


Figura 1 – Processo de Recuperação de Informação

Na base do processo de recuperação de informação está a colecção de documentos. Para especificar um sistema de informação concreto é preciso especificar: os documentos a serem usados; as operações de texto a serem efectuadas; o modelo; a estrutura e os elementos a serem extraídos da base de dados.

As operações de texto processam os documentos e interrogações criando uma representação lógica destes que é interna ao sistema de recuperação de informação. Os termos usados para esta representação interna podem ser obtidos de uma forma automática, criando o próprio sistema uma lista de termos a partir da sua colecção ou usando uma lista de termos relevantes gerada por um especialista. A lista é obtida executando-se um conjunto de operações de modo a eliminar termos de pouco valor para a recuperação. Uma outra estrutura que pode ser usada é o *tesauro*, constituído pelos termos, sua conjugação e conceitos semânticos. Esta serve para denotar conceitos e termos alternativos de modo a obter um melhor desempenho na recuperação de informação. Deste modo, podem usar-se

expressões com valor superior aos dos termos individuais: se um utilizador usou a expressão “guerra fria”, poderá obter uma informação mais adequada de que a que teria com os termos “guerra” e “fria”, separadamente.

A representação do documento e interrogação varia de acordo com o modelo matemático escolhido, sendo os termos não considerados na indexação designados **lista de paragem** (*stopwords*), termos com pouco ou nenhum significado; os casos mais comuns destes são os artigos, os pronomes, os determinantes (exemplo: “de”, “e”, “a”, “que”, “isto”). Os termos com elevada frequência nos documentos de uma dada colecção podem ser considerados lista de paragem. Os modelos matemáticos formais tradicionais (booleano, vectorial, probabilístico) não abordam especificamente a questão das listas de paragem sendo este conceito introduzido posteriormente no próprio sistema de recuperação de informação. Um outro processo de análise léxico é a **radicalização** (*stemming*), que encontra o radical dos termos, como seja “escol” para os termos “escolaridade”, “escolares”, “escolas”.

Para além destas representações, é necessário definir uma estrutura de índice dos documentos, cuja essência é a de atribuir a cada documento um conjunto de termos. Na generalidade, estes termos incluem informação considerada relevante para identificação dos documentos, sobre a localização integral do documento e partes do documento considerados importantes para a decisão do utilizador na avaliação da sua relevância (título, autores, resumo). Muito frequentemente, esta estrutura é baseada no texto integral do documento.

O método de pesquisa condiciona a organização dos ficheiros que contêm a informação de indexação e documentos, sendo usada uma que possa fornecer o melhor desempenho, optimizando as operações de recuperação de informação. Uma destas arquitecturas, ficheiros invertidos, actua como se de um índice de livro se tratasse, com os termos e páginas onde se encontra a informação necessária. Em termos de recuperação de informação, os ficheiros invertidos possibilitam que, para um dado termo, se possa conhecer de imediato todos os documentos associados ao mesmo.

Estas operações são desencadeadas sem a intervenção directa do utilizador.

Uma vez estabelecida a estrutura interna do sistema de recuperação, é possível iniciarem-se os processos de pesquisa e recuperação de informação. O utilizador expressa a sua necessidade de informação que é processada pelo mesmo conjunto de operações de texto aplicadas aos documentos, sendo posteriormente aplicadas operações de interrogação que têm como objectivo a obtenção não somente de uma interrogação mais eficiente mas também uma melhor representação da sua necessidade de informação. Esta interrogação é então processada. Os documentos são usualmente ordenados de acordo com a sua possível relevância para o utilizador antes de serem apresentados. Posteriormente, o utilizador efectua a sua classificação examinando os documentos e, nesta altura, poderá definir um subconjunto de documentos que julga relevantes para iniciar um processo de retorno de informação para reavaliação (*feedback*). A interrogação é refinada pelo sistema de modo a que esta seja uma melhor representação da necessidade de informação do utilizador e que a resposta contenha mais documentos relevantes.

3 Modelos na recuperação de informação

Na área da recuperação de informação tem havido um esforço sistemático no sentido de descrever com rigor o processo de recuperação e caracterizar formalmente a resposta a uma pergunta. Sendo a tarefa de obter respostas a perguntas sobre informação não estruturada intrinsecamente imprecisa, têm surgido múltiplas propostas para este fim.

Este capítulo centra-se nos modelos formais para a investigação e implementação de sistemas de recuperação de informação. Analisam-se, comparativamente, os modelos matemáticos de recuperação de informação com uso mais generalizado: booleano, vectorial, probabilístico, conjuntos difusos e P -norm, bem como alguns modelos alternativos.

3.1 Definição formal de modelo

O modelo adoptado num sistema de recuperação de informação estabelece as representações para documentos e interrogação e os critérios para seleccionar documentos para a resposta. A existência de um modelo explícito facilita a implementação do sistema e o estudo do seu comportamento. Por isso, um modelo deve ser claro, objectivo e preciso. A matemática é assim a linguagem preferencial para a tipificação de modelos de recuperação de informação.

Os modelos propostos centram-se no mecanismo de recuperação e, ao longo dos tempos, várias extensões aos mesmos têm sido apresentados. A extensão dos modelos matemáticos para novos modelos, designados por modelos cognitivos, tem sido sugerida com o propósito de incluir no modelo aspectos como o contexto da interacção entre o utilizador e o sistema operacional.

Para efeito de descrição dos modelos, um sistema de recuperação de informação é definido como um quádruplo $[D, Q, F, R(q_i, d_j)]$ (Baeza-Yates e Ribeiro-Neto, 1999) onde:

D é um conjunto de representações de documentos

Q é um conjunto de representações para as necessidades de informação do utilizador.

F é a “framework” para a modelização da representação de documentos, de interrogações e das suas relações.

$R(q_i, d_j)$ é uma função de ordenação (*ranking*), que associa um número $q_i \in Q$ com uma interrogação e um documento $d_j \in D$. Esta ordenação define uma ordenação entre os documentos associados à interrogação q_i .

Em alguns modelos de recuperação é definido o peso $\omega_{r,i}$ associado ao par de termo e interrogação (k_r, q_i) onde $\omega_{r,i} \geq 0$.

3.2 Comparação de modelos

O conceito essencial no processo de recuperação de informação é o da relevância e a sua medida matemática serve de base ao processo de ordenação, determinada pelo modelo usado.

Existem diversas classificações para os modelos de recuperação de informação e extensões ou variantes destes, tendo sido recentemente apresentado por Baeza-Yates e Ribeiro-Neto (1999) uma taxonomia dos modelos de recuperação de informação. A abordagem para a apresentação e comparação dos modelos não é delinear um sistema matricial, mas estabelecer uma comparação dos fundamentos de cada um dos modelos, apontando diferenças, vantagens e desvantagens em relação a cada uma das metodologias consideradas.

3.2.1 Modelo booleano

O modelo booleano é o mais antigo tendo sido adoptado no passado por muitos dos sistemas comerciais de recuperação de informação. Este modelo tem como princípio matemático a lógica booleana; a forma das interrogações é intuitiva para o utilizador, sendo representadas por expressões booleanas. Seguindo este modelo, a resposta contém um conjunto não ambíguo de documentos da colecção. Apesar desta simplicidade, o

modelo tem sido objecto de muitas críticas. A estratégia de recuperação baseada num modelo de documentos em que cada termo está ou não presente em cada documento, não existindo nenhum grau para esta associação, faz deste modelo mais um modelo de recuperação de dados do que informação (Baeza-Yates e Ribeiro-Neto, 1999).

O modelo booleano não efectua ordenação dos documentos nem usa nenhum tipo de estatística dos termos (Hiemstra, 2001), classificando os documentos como relevantes ou não relevantes e não tendo, contrariamente a outros modelos, como o probabilístico ou difuso, o conceito de relevância parcial.

As expressões booleanas das interrogações podem ser para o utilizador menos experiente algo complexas. Este deve ser capaz de construir expressões com operadores lógicos de AND, OR, NOT para formular a sua interrogação. Em alguns sistemas existe uma extensão destes operadores com o uso de operadores de proximidade e abreviatura.

Apesar destas dificuldades, o modelo booleano é ainda usado, dado a sua simplicidade e precisão formal, sendo binários os pesos associados aos termos, isto é, $\omega_{i,j} \in \{0,1\}$, levando a que poucos ou demasiados documentos sejam recuperados. Tal situação é ainda mais crítica se atentarmos que as colecções são cada vez maiores, como é o caso da Web, sendo indispensável uma ordenação (*ranking*) dos documentos recuperados, uma vez que o utilizador, muito provavelmente, apenas irá analisar os dez ou vinte primeiros documentos recuperados.

Modelos com base na teoria dos conjuntos aproximativos ou difusos têm sido propostos como extensões do modelo booleano.

3.2.2 Modelo vectorial

O modelo vectorial veio dar resposta às limitações do modelo booleano na ordenação dos documentos recuperados.

O modelo vectorial atribui um peso não binário para indexar termos nas interrogações e nos documentos, representando-os como vectores num espaço euclidiano, onde cada termo constitui uma dimensão separada. A relação entre a interrogação e o documento é estabelecida no modelo vectorial pela medida de semelhança, calculada pelo

coseno do ângulo que separa os dois vectores, o do documento $\vec{d}_j$ e o da interrogação $\vec{q}$, obtida a partir de (Baeza-Yates, Ribeiro-Neto, 1999):

$$\begin{aligned} \text{sim}(d_j, q) &= \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^t w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{i=1}^t w_{i,q}^2}} \end{aligned}$$

$|\vec{d}_j|$ e $|\vec{q}|$ são as normas dos vectores do documento e interrogação, respectivamente.

$\omega_{i,q}$, o peso associado ao par de termo e interrogação (k_i, q) , com $\omega_{i,q} \geq 0$; o vector da interrogação $\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$ onde t é o número total de termos usados no sistema.

Um vector para um documento, d_j , é representado de forma análoga por $\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$. A medida de semelhança varia no intervalo $[0,1]$, sendo os documentos recuperados ordenados pela sua maior semelhança relativamente à interrogação introduzida pelo utilizador no sistema. Para fazer funcionar este modelo, é necessário estabelecer a forma de cálculo de pesos para os termos. Este problema, não trivial, abordado inicialmente por Salton (Baeza-Yates, Ribeiro-Neto, 1999) pode ser visto como um problema genérico de agrupamento de objectos em clusters. A metodologia adoptada e testada no sistema SMART (Salton, 1971) revelou que o factor usualmente designado por *tf* - *term frequency* providencia uma referência adequada para a medida de quanto o termo descreve o documento, e que o factor *idf* - *inverse document frequency* informa se o termo é raro no conjunto de documentos. Caso este factor seja alto, será uma indicação de que se trata de um termo adequado para indexar documentos por não ser um

termo comum. As fórmulas para a frequência normalizada de um termo (Baeza-Yates, Ribeiro-Neto, 1999):

$$f_{i,j} = \frac{freq_{i,j}}{\max freq_{d,j}}$$

onde o máximo é calculado pelos termos existentes no documento d_j . A frequência inversa de um termo é obtida por:

$$idf_i = \log \frac{N}{n_i}$$

sendo o esquema de pesagem mais usado para os termos:

$$w_{i,j} = f_{i,j} \times idf_i$$

O modelo vectorial, apesar de ser o mais usado em sistemas actuais dada a sua abordagem para os pesos dos termos, eficiência e ordenação dos documentos recuperados, apresenta limitações. Os termos são assumidos como independentes, não tendo em consideração a construção de frases ou dependência entre termos. Na prática, caso esta dependência seja considerada, podem verificar-se efeitos nefastos em termos de desempenho do sistema dado o grande número de dependências e a sua aplicação indiscriminada em todos os documentos (Baeza-Yates e Ribeiro-Neto, 1999). A implementação deste modelo coloca problemas no cálculo do coseno, indispensável para a medida de semelhança, dado serem necessários todos os valores das componentes do vector, que não estão disponíveis numa arquitectura de ficheiros invertidos. Esta questão leva ao cálculo e armazenamento das normas, ocupando muito mais espaço de que a informação de suporte ao modelo booleano (Witten, Moffat, Bell, 1994 a partir de Hiemstra, 2001).

3.2.3 Modelo probabilístico

O modelo probabilístico introduzido em 1976 por Robertson e Sparck Jones (Baeza-Yates, Ribeiro-Neto, 1999), apresenta uma arquitectura baseada na probabilidade, que se baseia no facto de dada uma interrogação q , um documento d_j , existir uma probabilidade do documento d_j ser considerado relevante pelo utilizador. O modelo admite que a probabilidade de relevância depende exclusivamente da interrogação e do documento, sendo a relevância de um dado documento independente de um outro e dicotómica, os pesos dos termos são binários $w_{i,j} \in \{0,1\}$. Consequentemente, o modelo probabilístico assume a existência de um conjunto perfeito de documentos, inicialmente conhecido ou pré-definido, designado por R , que deve maximizar a probabilidade global de relevância para o utilizador. Os documentos não relevantes da colecção são considerados no conjunto $\bar{R}$, complementar ao R . No modelo probabilístico, a medida de semelhança é obtida através de (Baeza-Yates, Ribeiro-Neto, 1999):

$$sim(d_j, q) = \frac{P(R | d_j)}{P(\bar{R} | d_j)}$$

em que $P(R | d_j)$ é a probabilidade de seleccionar o documento d_j no conjunto R . A medida de semelhança pode ser convertida em:

$$sim(d_j, q) \sim \sum_{i=1}^t w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i | R)}{1 - P(k_i | R)} + \log \frac{1 - P(k_i | \bar{R})}{P(k_i | \bar{R})} \right)$$

Não sendo o conjunto R conhecido inicialmente, é usado um método simplificado para os cálculos iniciais das probabilidades de $P(k_i | R)$ e $P(k_i | \bar{R})$. Esta simplificação é efectuada no início, assumindo que todos os termos têm probabilidade constante,

$$P(k_i | R) = 0.5$$

e que a distribuição dos termos nos documentos não relevantes é semelhante à da colecção toda,

$$P(K_i | \bar{R}) = \frac{n_i}{N}$$

Este método, após a recuperação e ordenação dos primeiros documentos, é alterado de modo a melhorar a ordenação efectuada pelo modelo. Esta alteração consiste na melhoria da estimativa inicialmente efectuada para as probabilidades associadas a R e $\bar{R}$, podendo ser obtida com a aproximação de $P(k_i | R)$ pela distribuição do termo k_i nos documentos recuperados, e aproximar $P(k_i | \bar{R})$ considerando que todos os documentos não recuperados não são relevantes. Consideremos V , o conjunto de documentos inicialmente recuperados e ordenados segundo o modelo. Este conjunto pode ser definido como o conjunto dos primeiro r documentos, onde r é um valor previamente definido. Seja V_i , um subconjunto de V , definido como o conjunto de documentos em V que contém o termo k_i . Definimos as novas probabilidades como:

$$P(K_i | R) = \frac{V_i}{V}$$

$$P(K_i | \bar{R}) = \frac{n_i - V_i}{N - V}$$

Este processo produz uma informação que melhora a aproximação, podendo ser executado repetidamente. Com esta abordagem, é possível usar o modelo sem a informação de relevância do utilizador. Salton e Buckley demonstraram ser expectável uma menor eficiência do modelo probabilístico quando usado em colecções diversas (Baeza-Yates e Ribeiro-Neto, 1999) do que a do modelo vectorial. Este modelo apresenta como vantagem o facto dos documentos serem ordenados pela sua probabilidade de relevância, não necessitando de nenhum algoritmo suplementar de pesos para os termos.

No entanto, no modelo probabilístico é necessário efectuar uma estimativa inicial do conjunto de documentos relevantes e não relevantes da colecção, assumindo uma independência dos termos, não sendo considerada a frequência dos termos dentro de cada documento. Neste contexto, é possível afirmar que o modelo probabilístico sofre parcialmente dos mesmos problemas do que o modelo booleano (Hiemstra, 2001).

3.2.4 Modelo baseado na teoria dos conjuntos difusos

A teoria dos conjuntos difusos tem como conceito fundamental associar aos elementos de uma colecção, uma função que assume valores no intervalo $[0,1]$ para representar o grau de satisfação de um elemento face a uma dada característica variando entre 1, a sua satisfação completa, e 0 a sua não satisfação completa. Entre estes dois valores limites, o elemento apresentará um grau maior ou menor consoante o valor que este elemento apresentar, sendo, por exemplo: 0,95 uma quase satisfação completa. Este grau tem como objectivo representar a imprecisão ou carácter difuso da informação.

Na recuperação de informação, o modelo de conjuntos difusos foi inicialmente apresentado por Tahani (1976). Este associava pesos difusos para indicar o grau de associação entre os documentos (ou interrogações) e os respectivos termos do índice. Pelas suas características, o modelo de conjuntos difusos pode ser usado na construção de um tesouro, na medida em que esta teoria é adequada para o conceito associado ao tesouro (Baeza-Yates, Ribeiro-Neto, 1999). Desde os fins dos anos 70, uma variedade de operadores difusos têm sido apresentados (Lee, 1995, a partir de Hiemstra 2001), não existindo uma explicação clara e concisa sobre as causas para os diferentes desempenhos conseguidos entre eles.

A estrutura do tesouro pode ser construída definindo-se uma matriz C que correlacione os termos. As linhas e colunas desta matriz estão associadas a termos do índice da colecção, sendo os factores de co-ocorrência, $c_{i,l}$, entre dois termos k_i e k_l obtidos por (Baeza-Yates, Ribeiro-Neto, 1999):

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

onde n_i é o número de documentos que contêm o termo k_i , n_i é o número de documentos que contêm o termo k_i , e $n_{i,j}$ é o número de documentos que contêm ambos os termos. Esta matriz de co-ocorrência pode ser usada para a definição de um conjunto difuso para cada termo k_i do índice. Neste conjunto difuso, cada documento d_j tem um grau de pertença $\mu_{i,j}$ definido como (Baeza-Yates, Ribeiro-Neto, 1999):

$$\mu_{i,j} = 1 - \prod_{k_i \in d_j} (1 - c_{i,j})$$

O grau de pertença é calculado no modelo de recuperação de informação como um produto iterado, em vez das funções comuns de máximo e mínimo associadas à teoria dos conjuntos difusos apresentada por Zadeh. Tal método permite a obtenção de valores de pertença que variam numa forma menos abrupta, originando uma certa “continuidade” dos valores. Um documento d_j pertence ao conjunto difuso associado ao termo k_i se os termos presentes no documento estão associados ao termo k_i . Sempre que um termo k_i do documento d_j está fortemente associado ao termo de índice k_i , então o grau de pertença é aproximadamente 1, i.e., $\mu_{i,j} \sim 1$ e o índice k_i é um bom termo para o documento d_j . Caso contrário, $\mu_{i,j} \sim 0$, o índice k_i é um mau índice para o documento d_j indicando que todos os termos do documento d_j apresentam uma associação muito fraca para com o índice k_i .

A interacção do utilizador com o sistema de recuperação de informação é efectuada de um modo booleano, sendo introduzido no sistema uma interrogação que posteriormente é convertida na sua forma normal disjuntiva. Considerando a interrogação $[q = k_a \wedge (k_b \vee \neg k_c)]$, a sua forma normal disjuntiva é $[\bar{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)]$ onde cada componente é um vector binário associado ao énpulo (k_a, k_b, k_c) .

A ordenação dos documentos no modelo de conjuntos difusos é relativa à interrogação do utilizador e é implementada pela co-ocorrência existente entre os termos do documento e do índice. O modelo de conjuntos difusos, pelas suas características, é considerado por muitos como sendo uma extensão do modelo booleano, dado as operações serem do mesmo tipo, realizadas sobre conjuntos difusos e não rígidos.

A aplicação da teoria dos conjuntos difusos tem sido estudada para as várias fases da recuperação de informação, embora os sistemas experimentais implementados mostrarem limitações para a manipulação de colecções grandes. Lee (1995, a partir de Hiemstra, 2001) efectuou uma comparação vasta quer em termos teóricos, quer em termos de eficácia de recuperação de informação. Usando uma das colecções da TREC, foi medida a eficácia das interrogações do tipo booleanas em vários modelos. Os modelos com melhor desempenho foram o modelo de Paice e o modelo P-norm, analisados de seguida. Tal como o modelo vectorial, o modelo baseado nos conjuntos difusos, necessita de uma metodologia suplementar para os pesos dos termos, não providenciando também orientação para os novos problemas de recuperação de informação que estão surgindo.

3.2.5 Modelo P-norm

O modelo P-norm, apresentado por Salton, Fox, Wu (1983) é uma extensão do modelo booleano, onde as operações lógicas são substituídas por operações algébricas de distância (Hiemstra, 2001).

Quando apenas dois termos são considerados, podemos representar os mesmos num espaço euclidiano de duas dimensões. Consideremos uma interrogação do tipo $q = k_i \wedge k_l$, um documento d_j e os pesos $w_{i,j}$, $w_{l,j}$ associados aos pares $[k_i, d_j]$ e $[k_l, d_j]$, respectivamente. Efectuando-se a normalização dos pesos entre 0 e 1 e simplificando nomenclatura dos mesmos apenas para x e y e do ponto $d_j = (x, y)$, podemos definir o ponto (1,1) como sendo o ponto onde ambos os termos ocorrem no documento e (0,0) quando estes estão omissos. Consequentemente, o ponto (1,1) será o mais desejável para uma interrogação com conjunção e o ponto (0,0) o menos desejável para uma interrogação com disjunção. Os documentos serão ordenados por ordem decrescente de distância ao

ponto (0,0) para uma interrogação disjuntiva e por ordem crescente ao ponto (1,1) para uma interrogação conjuntiva. A medida adequada para a semelhança entre o documento e a interrogação com disjunção q_{or} e interrogação com conjunção q_{and} é obtida a partir de:

$$sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

O modelo P-norm generaliza este conceito de distância, incluindo as distâncias p , onde $1 \leq p \leq \infty$. Neste novo contexto, as interrogações são representadas como:

$$q_{or} = k_1 \vee^p k_2 \vee^p \dots \vee^p k_m$$

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \dots \wedge^p k_m$$

As semelhanças são agora obtidas a partir de:

$$sim(q_{or}, d_j) = \sqrt[p]{\left(\frac{x_1^p + x_2^p + \dots + x_m^p}{m} \right)}$$

$$sim(q_{and}, d_j) = 1 - \sqrt[p]{\left(\frac{(1-x_1)^p + (1-x_2)^p + \dots + (1-x_m)^p}{m} \right)}$$

O processamento de interrogações mais genéricas (conjunções e disjunções) é efectuado pelo agrupamento dos operadores numa ordem predefinida, sendo a semelhança obtida através de:

$$sim(q, d) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2} \right)^{\frac{1}{p}} \right)^p + x_3^p}{2} \right)^{\frac{1}{p}}$$

A medida de semelhança para a negação de uma interrogação é obtida

$$sim(q_{not}, d) = 1 - sim(q, d)$$

Baseado em recentes publicações, o modelo P -norm é provavelmente aquele com maior aceitação entre os modelos considerados como extensões do modelo lógico booleano com melhor desempenho (Hiemstra, 2001).

3.2.6 Modelos alternativos

Os modelos apresentados anteriormente podem ser considerados como o núcleo dos modelos para a recuperação de informação, existindo, obviamente, um conjunto de outros modelos que ao longos dos tempos foram sendo apresentados como alternativos e que estão divididos em duas categorias:

- Modelos algébricos, onde se inclui:
 - o Modelo Generalizado Vectorial, aborda a questão da não ortogonalidade dos vectores, para cada par de termos.
 - o Modelo de Indexação Semântico, propõe uma abordagem em que se efectua uma análise semântica sobre a interrogação e sobre os documentos para a extracção de conceitos. Estes são usados como termos para a indexação em contraposição dos termos obtidos por uma análise léxica.

- Modelo de Rede Neuronal, propõe as redes neuronais como instrumentos para a recuperação de informação dada a sua eficácia para a comparação lexical.
- Modelos Probabilísticos, onde estão classificados:
 - Redes de Bayes, que providencia uma framework que permite uma combinação de fontes diversas para a decisão de relevância num dado documento.
 - Modelos de Rede de Inferência, tendo uma visão epistemológica da recuperação de informação associa variáveis aleatórias com os termos, assumindo que a colecção é observada na procura de documentos relevantes.

A teoria dos conjuntos aproximativos não tem recebido atenção suficiente no campo da recuperação de informação, pois para além de alguns artigos publicados, esta abordagem não é considerada suficientemente promissora para ser indicada em obras de referência. Supomos que isto se deve ao facto de que, por si só, a teoria dos conjuntos aproximativos não se constituir como uma alternativa válida face aos modelos existentes.

4 Fundamentos teóricos das teorias dos conjuntos difusos e aproximativos

Este capítulo apresenta os fundamentos matemáticos da teoria dos conjuntos difusos, da teoria dos conjuntos aproximativos e uma combinação entre as duas teorias.

O conceito de objecto difuso é introduzido formalmente no ponto 4.1, sendo definidos os conjuntos difusos, as propriedades, relações e operações existentes sobre os conjuntos difusos. Apresenta-se um exemplo para demonstrar os conceitos associados à teoria dos conjuntos difusos.

O ponto 4.2 apresenta os fundamentos da teoria dos conjuntos aproximativos. Descreve-se a relação de indiscernibilidade, a sua importância para a redução dos atributos que caracterizam um objecto, a necessidade das conseqüentes aproximações estrita e lata de um conjunto de objectos e a determinação da qualidade da classificação. Para uma mais fácil compreensão é introduzido um exemplo de um sistema de informação caracterizado com recurso a esta abordagem.

No ponto 4.3 estabelece-se o ponto de contacto ou convergência das duas teorias, sendo definida a sua ortogonalidade na captação de dois aspectos distintos da imprecisão da informação: indiscernibilidade e incerteza. Definem-se as suas potencialidades no âmbito da recuperação de informação e os conceitos de conjunto aproximativo difuso e conjunto difuso aproximativo.

4.1 Conceitos básicos da teoria dos conjuntos difusos

A teoria dos conjuntos difusos apresentados por Zadeh (1965) emergiu como a teoria para abordar as questões linguísticas, para a forma como as pessoas exprimem sensações, objectos que não possuindo limites bem definidos podem ter diferentes interpretações ou estarem associados a conceitos não rígidos, aos quais não é possível

aplicar uma lógica booleana, de dicotomia entre o falso e o verdadeiro, de pertença ou não pertença a um conjunto.

O conceito base desta teoria é flexibilizar este conceito dicotómico booleano e o de apresentar valores intermédios conferindo uma abordagem mais realista e enriquecedora dos problemas, sendo considerada uma extensão da teoria clássica dos conjuntos.

Um conjunto difuso é uma colecção de objectos que apresentam valores de pertença que variam entre o 0 (total exclusão) e o 1 (inclusão completa). Os valores expressam o grau de aplicabilidade a cada objecto das propriedades que definem o conceito subjacente ao conjunto difuso. Os conjuntos difusos distinguem-se facilmente dos conjuntos rígidos, da teoria clássica dos conjuntos, assim designados por não admitirem valores de pertença. Traduzindo em termos matemáticos, seja x um elemento do universo U , e A um subconjunto de U , $A \subset U$ então $x \notin A$ ou $x \in A$, o que se exprime como:

$$\mu_A(x) = \begin{cases} 1, & \text{se } x \in A \\ 0, & \text{se } x \notin A \end{cases}$$

Um conjunto rígido pode ser considerado como um conjunto degenerado de um conjunto difuso onde os elementos se restringem aos extremos do intervalo $[0,1]$, isto é, ao conjunto $\{0,1\}$.

A teoria dos conjuntos difusos tem sido aplicada na área dos sistemas de recuperação de informação em diferentes momentos do processo de recuperação. A formulação das interrogações que descrevem objectos relevantes por uma representação mais flexível de valores admissíveis para os atributos tem sido uma das formas mais usadas. Do mesmo modo, tem sido aplicado para o cálculo do peso atribuído a um termo em função do grau de satisfação de uma dada propriedade, sendo exemplo a indexação difusa de um termo para um documento.

4.1.1 Conjunto difuso

Seja U um conjunto finito e não vazio designado por universo. Um conjunto difuso A de U é definido pela função de pertença (*membership*) que mapeia os elementos de U no intervalo $[0,1]$.

$$A: U \rightarrow [0,1]$$

É também usada a notação:

$$\mu_A: U \rightarrow [0,1]$$

Deste modo, um conjunto difuso A em U pode ser representado por um conjunto de pares de um elemento $x \in U$ e o seu grau de pertença por $A = \{ A(x)/x \mid x \in U \}$. Existem duas vertentes na teoria dos conjuntos difusos que devem ser acentuadas. A primeira é a que está associada com o aspecto difuso da informação, dado que o valor de pertença representa o grau, a quantificação de quão compatível é x relativamente ao conceito representado por A . O segundo é o da incerteza acerca da informação: $A(x)$ representa o quão possível é que x ocorra caso A esteja presente. Esta vertente é referenciada como uma visão epistemológica de um conjunto difuso, estabelecendo-se o estado de conhecimento acerca de uma variável (Pedrycz e Gomide, 1998).

Os conjuntos difusos podem ser definidos em universos finitos ou não, usando diferentes notações. Se um universo U é discreto e finito, com cardinalidade n , então um conjunto difuso pode ser representado na forma de um vector de dimensão n , cujas coordenadas denotam o grau de pertença dos elementos de U a A . Em princípio, qualquer função da forma $A(x): U \rightarrow [0,1]$ descreve uma função associada a um conjunto difuso que depende não somente do conceito a ser representado mas também do contexto em que é usada.

4.1.2 Propriedades e operações

Um conjunto difuso A é normal, se a sua função atinge o valor 1, isto é,

$$\sup_x A(x) = 1$$

Se o supremo é menor que 1, A é designado por *sub-normal*. O supremo é também designado por *altura de A* .

O suporte de um conjunto difuso A , definido por **Supp**(A), designa todos os elementos de X que pertencem a A para um valor diferente de zero.

$$\text{Supp}(A) = \{ x \in U \mid A(x) > 0 \}$$

O **Núcleo** de um conjunto difuso A , é o conjunto de elementos de U que têm uma pertença 1 em A , isto é,

$$\text{Núcleo}(A) = \{ x \in U \mid A(x) = 1 \}$$

onde $\text{Núcleo}(A) \subseteq \text{Supp}(A)$.

O α -cut ou conjunto de nível α de um conjunto difuso, com notação A_α , é um conjunto de elementos do universo U cujos valores excedem ou igualam o limite α :

$$A_\alpha = \{ x \in U \mid A(x) \geq \alpha \}$$

O conjunto rígido A_α consiste nos elementos de U *identificados com A* num grau de, pelo menos, α . Claramente, quanto menor o α , mais elementos são admitidos no α -cut, isto é,

$$\text{Se } \alpha_1 > \alpha_2 \text{ então } A_{\alpha_1} \subseteq A_{\alpha_2}$$

Pelo α -cut de um conjunto difuso é possível determinar uma família de conjuntos rígidos, permitindo a definição de subconjuntos de elementos que satisfaçam um conjunto de proposições para um dado limite α . Tal propriedade é importante em qualquer modelo baseado na teoria dos conjuntos difusos, pela possibilidade de restringir o estudo a apenas um dado subconjunto.

O teorema da representação (*The Representation Theorem*)

A observação de que qualquer conjunto difuso pode ser decomposto numa série de α -cuts é conhecida como o teorema da representação,

$$A = \bigcup_{\alpha \in [0,1]} (\alpha \mid A_\alpha),$$

em outra forma equivalente,

$$A(x) = \sup_{\alpha \in [0,1]} [\alpha \mid A_\alpha(x)]$$

Como um exemplo, consideremos o conjunto difuso $A = \{0.1/2, 0.3/3, 1.0/4, 0.3/5\}$ no universo $X = \{1, 2, 3, 4, 5\}$. Sejam a família de α -cuts $A_{0.1} = \{2, 3, 4, 5\}$; $A_{0.3} = \{3, 4, 5\}$ e $A_{1.0} = \{4\}$ de A . Apliquemos o teorema da representação para obter o conjunto difuso a partir da família de α -cuts. Calculemos os valores de $A(x)$ para os elementos do universo X :

- O elemento 1 do universo não está presente na família de α -cuts de A , então 1 não é elemento de A .
- Para $x=2$, temos $A(2) = 0.1$ porque $A_{0.1}$ é o único α -cut onde o elemento 2 está presente.
- Para $x=3$, temos $A_{0.1}(3)$ e $A_{0.3}(3)$. Então neste caso o supremo é 0.3, ou seja, $A(3) = 0.3$. Repetindo este processo para os restantes elementos obtemos os diferentes valores de α .

Este teorema implica que qualquer problema formulado no âmbito dos conjuntos difusos pode ser resolvido transformando estes conjuntos nas suas correspondentes famílias de α -cuts e que a partir de uma família de α -cuts de um conjunto difuso é possível obter novamente o conjunto difuso.

Princípio da extensão

O princípio da extensão propõe a transposição de diversos conceitos matemáticos para conceitos difusos. Este princípio tem sido aplicado para operações aritméticas com números difusos, para o cálculo de variáveis induzidas, sendo essencialmente aplicado na transformação de conjuntos difusos através de funções.

Sejam X, Y dois conjuntos e $f : X \rightarrow Y$. Seja A um conjunto difuso em X . O princípio da extensão define que a imagem de A é um conjunto difuso $B = f(A)$ em Y , de modo que para cada $y \in Y$ temos

$$B(y) = \sup_x A(x), \text{ em que } x \in X \text{ e } y = f(x)$$

Dado um conjunto difuso A em um universo finito U , a sua cardinalidade é denotada por $\text{Card}(A)$ e é o número real definido como:

$$\text{Card}(A) = \sum_{x \in U} A(x)$$

Normalização

Seja A um conjunto difuso subnormal, não vazio, a normalização é o processo que converte este conjunto na sua forma normal pela divisão da sua função de pertença original pela altura de A . Para cada elemento tem-se:

$$\text{Norm}_A(x) = \frac{A(x)}{\text{Alt}(A)}$$

Concentração

Os conjuntos difusos designam-se como *concentrados*, quando as suas funções de pertença assumem valores relativamente pequenos. A operação de concentração realiza-se elevando a função de pertença ao expoente p em que $p > 1$, por exemplo.

$$\text{Con_}A(x) = A^p(x), \text{ resultando } \text{Con_}A(x) \leq A(x)$$

Como um exemplo, usando a propriedade associada ao conceito de *bom* investimento numa classificação de aquisição de objectos, a operação de concentração aplicada ao conjunto difuso resultante deve identificar, de forma razoável, os objectos associados ao conceito de *muito bom* investimento.

Dilatação

A operação de dilatação tem o objectivo contrário da concentração e é produzida transformando a função de pertença.

$$\text{Dil_}A(x) = A^r(x), \text{ onde } r \in [0,1]$$

Intensificação de contraste

Na operação de Intensificação de Contraste (*Contrast Intensification*) os valores de pertença menores que 0,5 são diminuídos enquanto que os valores superiores são elevados. Deste modo, intensifica-se o contraste entre os vários elementos, reduzindo-se a incerteza.

$$\text{Int_}A(x) = \begin{cases} 2A^2(x) & \text{se } 0 \leq A(x) \leq 0,5 \\ 1 - 2(1 - A(x))^2, & \text{caso contrário} \end{cases}$$

Difusão

O efeito da difusão (*fuzzification*) é complementar ao da intensificação e é produzido alterando os valores de pertença:

$$\text{Fuzz_}A(x) = \begin{cases} \sqrt{A(x)/2} & \text{se } A(x) \leq 0,5 \\ 1 - \sqrt{(1-A(x))/2}, & \text{caso contrário} \end{cases}$$

4.1.3 Relações entre conjuntos difusos e operadores

Os conjuntos difusos apresentam as mesmas operações básicas da teoria clássica dos conjuntos. Sejam A , B e C conjuntos difusos, definidos no mesmo universo U , estes operadores gozam das seguintes propriedades:

- Comutatividade:
 - o $A \cap B = B \cap A$
 - o $A \cup B = B \cup A$
- Associatividade:
 - o $A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$
 - o $A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C$
- Transitividade:
 - o $A \subset B$ e $B \subset C$ implica que $A \subset C$
- Idempotência:
 - o $A \cup A = A$
 - o $A \cap A = A$
- Condições fronteira:
 - o $A \cap \phi = \phi$
 - o $A \cup \phi = A$

Mas,

- o $A \cap \bar{A} \neq \phi$
- o $A \cup \bar{A} \neq U$

É possível expressar estas operações com base numa função característica. Como exemplo, apresentam-se as operações de intersecção, reunião e negação:

- $(A \cap B)(x) = \min(A(x), B(x))$
- $(A \cup B)(x) = \max(A(x), B(x))$
- $\neg A = 1 - A(x)$

Dois conjuntos difusos A , B são iguais se e só se as suas funções de pertença são idênticas, isto é,

$$A = B \text{ sse } A(x) = B(x) \quad \forall x \in U$$

e A está contido em B :

$$A \subseteq B \text{ sse } A(x) \leq B(x) \quad \forall x \in U$$

As operações de agregação de informação assumem um papel preponderante no desenvolvimento de diferentes sistemas como sejam, como exemplo, os sistemas de suporte à decisão, controladores lógicos e os sistemas com recurso às redes neuronais (Choi e Oh, 2000). A teoria dos conjuntos difusos assume um papel promissor em áreas em que decisões com múltiplos atributos se revelam de complexa resolução. A definição de operadores de agregação de informação como sejam os operadores de intersecção, união e de compensação providenciam modelos genéricos que permitem enquadrar métodos para a resolução destes problemas.

Operadores

As operações binárias t-norm e t-conorm (também conhecida por s-norm), formam classes de operadores genéricos de intersecção e união entre conjuntos difusos, sendo a t-norm associada à operação de intersecção e a t-conorm a de união. A t-norm, operação binária, $t : [0,1] \times [0,1] \rightarrow [0,1]$ satisfaz as seguintes propriedades:

- Comutatividade: $xy = yx$
- Associatividade: $xt(yz) = (xy)tz$

- Monotonicidade: se $x \leq y$ e $w \leq z$, então $x \wedge w \leq y \wedge z$
- Condições fronteira: $0 \wedge x = 0$, $1 \wedge x = x$

A operação t-conorm é idêntica à t-norm excepto na condição fronteira em que:
 $x \vee 0 = x$, $x \vee 1 = 1$.

A intersecção e união são operações que agregam elementos de conjuntos difusos num novo conjunto difuso, sendo que a t-norm e t-conorm são exemplos de operadores de agregação. Estes operadores podem ser genericamente definidos como uma operação

$A: [0,1]^n \rightarrow [0,1]$ que satisfaz:

- Monotonicidade, $A(x_1, \dots, x_n) \geq A(y_1, \dots, y_n)$ se $x_i \geq y_i$, $i = 1, \dots, n$
- Condições fronteira, $A(0, \dots, 0) = 0$ e $A(1, \dots, 1) = 1$

Pode ser provado que o operador de mínimo é um operador t-norm enquanto que o operador de máximo é t-conorm. Dois exemplos de operadores t-norm e t-conorm são os operadores de Yager e Hamacher:

- Yager t-norm

$$Y_p(x, y) = 1 - \min \left\{ 1, \left[(1-x)^p + (1-y)^p \right]^{1/p} \right\}, \text{ em que } p > 0$$

- Yager t-conorm

$$Y_p(x, y) = \min \left\{ 1, \left[x^p + y^p \right]^{1/p} \right\}, \text{ em que } p > 0$$

Quando $p \rightarrow \infty$, o operador Yager t-conorm converge para $\min(x, y)$ enquanto que o operador Yager t-norm converge para $\max(x, y)$.

- Hamacher t-norm

$$H_\gamma(x, y) = \frac{xy}{\gamma + (1-\gamma)(x + y - xy)}, \text{ em que } \gamma \geq 0$$

- Hamacher t-conorm

$$H_{\gamma}(x, y) = \frac{x + y - (2 - \gamma)xy}{1 - (1 - \gamma)xy}, \text{ em que } \gamma \geq 0$$

Em alguns casos, foi detectado que o operador de mínimo não obtém os valores mais adequados, quando usado como modelo para a operação de intersecção, produzindo valores muito baixos. Estes resultados foram obtidos por Zimmermann e Zysno (Pedrycz e Gomide, 1998) que propuseram uma família de operadores designados por operadores de compensação. Estes operadores efectuam uma combinação de operações de intersecção e união, sendo expressa por:

$$(A \odot B)(x) = [(A \cap B)(x)]^{1-\gamma} [(A \cup B)(x)]^{\gamma}$$

onde $\cap$ e $\cup$ são as operações de intersecção e união, γ é o factor de compensação em que $\gamma \in [0,1]$, indicando onde o operador se localiza entre os operadores de intersecção ou união. Existem outras formas de obter este efeito de compensação, como seja:

$$(A \odot B)(x) = (1 - \gamma)[(A \cap B)(x)] + \gamma [(A \cup B)(x)]$$

A família de operadores de cálculo da média agrupam um conjunto de operadores dos mais usuais nos conjuntos difusos. Este tipo de operador satisfaz as propriedades de idempotência, comutatividade e as condições fronteira sendo igualmente monótono. A equação generalizada desta família de operadores, para n argumentos:

$$A(x_1, \dots, x_n) = \sqrt[p]{\frac{1}{n} \sum_{i=1}^n (x_i)^p}, \quad p \in \mathbb{R}, p \neq 0$$

Neste conjunto de operadores encontram-se a média aritmética, geométrica e harmónica, o mínimo e o máximo. Para cada um destes operadores, descreve-se a equação associada, onde p pode ser referenciado como o factor de compensação:

- Média aritmética ($p = 1$), $A(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i)$
- Média geométrica ($p \rightarrow 0$), $A(x_1, \dots, x_n) = \left(\prod_{i=1}^n (x_i) \right)^{1/n}$
- Média harmónica ($p = -1$), $A(x_1, \dots, x_n) = \frac{n}{\sum_{i=1}^n 1/x_i}$
- Mínimo ($p \rightarrow -\infty$), $A(x_1, \dots, x_n) = \min(x_1, \dots, x_n)$
- Máximo ($p \rightarrow +\infty$), $A(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$

A família de operadores ordenados com pesos (OWA: *Order Weighted Averaging Operators*) apresentada por Yager (Pedrycz e Gomide, 1998) situa-se entre os operadores t-norm e t-conorm e tem os seus argumentos ordenados, sendo parametrizado por um vector de n pesos, $w = (w_1, \dots, w_n)$ tal que $\sum_{i=1}^n w_i = 1$ e $w_i \in [0, 1]$. Consideremos a sequência de valores de pertença $\{A(x_i)\}$ ordenada tal que $A(x_1) \leq A(x_2) \leq \dots \leq A(x_n)$, define-se o operador OWA como:

$$OWA(A, w) = \sum_{i=1}^n w_i A(x_i)$$

4.1.4 Exemplo

Consideremos a recuperação de informação associada ao problema “Qual o nível de desenvolvimento das cidades de Portugal?”. Neste exemplo o utilizador necessita de saber o nome da cidade com melhores condições para a sua vida pessoal em Portugal, através da recuperação de documentos relevantes que o informem sobre este facto. Deste modo, ele necessita de estabelecer um conjunto de factores de vida, termos relevantes, para si, a nível das condições que esta cidade deve possuir para ser considerada como candidata.

Enquadremos este problema desenvolvendo um modelo de recuperação de informação recorrendo a conjuntos difusos.

Sejam U o vocabulário da colecção de documentos a ser pesquisada, D o conjunto de documentos da colecção, q a interrogação efectuada e d_i o documento i . A interrogação q definida pelo utilizador foi $q = \{\text{emprego, hospital, educação, desporto, cultura, habitação, espaços_verdes}\}$. Cada documento da colecção encontra-se definido por um conjunto de termos. Suponhamos que o modelo inclui um método de recuperação difuso através do qual se recupera um conjunto difuso de documentos, definido como $B = \{0.7/d_1, 0.8/d_2, 0.65/d_3, 0.6/d_4, 0.68/d_5, 0.59/d_6, 0.71/d_7, 0.70/d_8, 0.63/d_9, 0.56/d_{10}\}$ que identifica as seguintes cidades: Lisboa, Porto, Faro, Braga, Coimbra, Viseu, Maia, Guimarães, Setúbal, Amadora, respectivamente. Cada documento tem um valor de pertença que indica a satisfação dos factores indicados, ou seja, estabelece o grau de desenvolvimento destas cidades face ao conjunto dos factores, em que 0 corresponde a um muito fraco desenvolvimento, enquanto que o valor 1 corresponde a um excelente desenvolvimento. Esta classificação foi previamente estabelecida por um conjunto de especialistas.

A resposta é caracterizada por um conjunto difuso, sendo este sub-normal dado que o supremo é inferior a 1. Dado o número de cidades recuperadas o utilizador considera restringir a resposta. Consideremos $\alpha = 0.7$ e calculemos o α -cut de B , $B_\alpha = \{d_1, d_2, d_7, d_8\}$. Deste modo, reduzimos a nossa área de possíveis localizações para um conjunto mais reduzido de cidades.

4.2 Conceitos básicos da teoria dos conjuntos aproximativos

A teoria dos conjuntos aproximativos proposta por Pawlak no início dos anos 80, baseia-se na existência da relação de indiscernibilidade dos objectos, tendo sido aplicada com sucesso em várias aplicações (Slowinsky, 1992). A relação de indiscernibilidade é uma relação de equivalência (reflexiva, simétrica e transitiva) e parte da observação de que os objectos podem ser indiscerníveis em termos dos dados disponíveis, ou seja, a informação incompleta origina indiscernibilidade entre os objectos (Pawlak, 1991). Dois objectos são indiscerníveis, se e só se tiverem os mesmos valores para um dado conjunto de atributos. A indiscernibilidade do objecto em termos dos seus atributos impede, geralmente, a integração inequívoca desse numa classe definida pela decisão.

Sejam por exemplo dois recursos de informação, x e y , que disponibilizam informação financeira. Existe um conjunto de atributos, $V = \{q_1, \dots, q_n\}$, que discrimina um recurso de informação. Considere-se três dos atributos que os caracterizam. Ambos são de formato electrónico, *pay-per-view* e com uma assinatura base anual similar para acesso aos serviços. Logo, estes dois recursos são indiscerníveis para estes três atributos e num caso em que seja preciso fazer uma escolha, o decisor (gestor de informação) não saberá qual seleccionar para a sua empresa. Atributos como as bases de dados integrantes de cada recurso de informação, o custo de cada documento, disponibilidade do recurso (24x7x365), capacidade de actualização das bases de dados, existência de acções de formação para os utilizadores, podem informar o decisor permitindo efectuar a melhor opção para a sua empresa.

A teoria dos conjuntos aproximativos estabelece que a informação está organizada sob a forma de um sistema de informação, também designado por sistema de representação do conhecimento, podendo ser considerada como uma colecção de objectos descritos pelos seus atributos. O sistema de informação pode ser representado por $S = \langle U, Q, V, f \rangle$ onde

U é um conjunto finito de objectos

Q é um conjunto finito de atributos

$V = \bigcup_{q \in Q} V_q$, V_q é o domínio do atributo q

$f: U \times Q \rightarrow V$ é uma função, designada de função de informação, tal que $f(x,q) \in V_q$ para todo o $q \in Q, x \in U$, denominada função informação.

Qualquer par $(q,v), q \in Q, v \in V_q$ é denominado descritor em S .

4.2.1 Relação de indiscernibilidade

Seja $S = \langle U, Q, V, f \rangle$ um sistema de informação, $P \subseteq Q$ e $x, y \in U$. Diz-se que x e y são indiscerníveis pelo conjunto de atributos P em S se $f(x,q) = f(y,q)$ para todo o $q \in P$. Assim, todo $P \subseteq Q$ gera uma relação binária em U , a qual é denominada de relação de indiscernibilidade, denotada como $IND(P)$. $IND(P)$ é uma relação de equivalência para qualquer P e a família de todas as classes de equivalência da relação $IND(P)$ é notada como $U / IND(P)$ ou, abreviando, por U / P . Esta família é também denominada por conhecimento P -básico de U em S .

As classes de equivalência da relação $IND(P)$ são denominadas conjuntos P -elementares (conceitos ou categorias do conhecimento P -básico) em S . $[X]_{IND(P)}$ é a notação para um conjunto P -elementar incluindo o objecto x . Os conjuntos Q -elementares são denominados átomos. $Des_P(x)$ é a notação para descrição do conjunto P -elementar $X \in U / IND(P)$ em termos dos valores dos atributos de P , i.e.:

$$Des_P(x) = \{ (q,v) : f(q,v) = v, \forall x \in X, \forall q \in P \}$$

A relação de indiscernibilidade é importante para determinar atributos redundantes, fazendo com que estes possam ser eliminados.

4.2.2 Aproximação de um conjunto

Subjacente ao conceito de conjunto aproximativo está o de aproximação de um conjunto por um par de conjuntos, denominados aproximação restrita e lata desse conjunto. Supondo um dado sistema de informação S , a cada conjunto $Y \subseteq U$ e relação de equivalência $IND(P)$ associam-se dois subconjuntos:

- Aproximação P -restrita de Y : $\underline{P}Y = \bigcup \{ X \in U / IND(P) : X \subseteq Y \}$

em que $\underline{P}Y$ é o conjunto de elementos de U , que certamente são classificados como elementos de Y empregando o conjunto de atributos P .

- **Aproximação P-lata de Y:** $\bar{P}Y = \cup \{ X \in U \mid \text{IND}(P): X \cap Y \neq \emptyset \}$

em que $\bar{P}Y$ é o conjunto de elementos de U , que possivelmente são classificados como elementos de Y empregando o conjunto de atributos P .

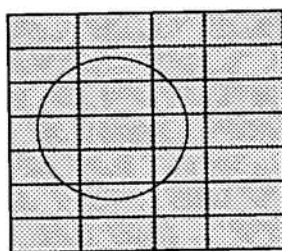
O conjunto **P-fronteira de Y** é definido como:

$$\text{BN}_P(Y) = \bar{P}Y - \underline{P}Y$$

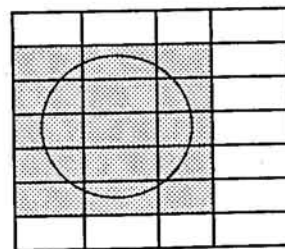
em que $\text{BN}_P(Y)$ é o conjunto de elementos que não se pode garantir que sejam classificados como elementos de Y usando o conjunto de atributos P . Na teoria dos conjuntos aproximativos são igualmente definidas:

- $\text{POS}_P(Y) = \underline{P}Y$, **região P-positiva de Y**, que consitui os elementos que certamente são elementos que se podem classificar como elementos de Y .
- $\text{NEG}_P(Y) = U - \bar{P}Y$, **região P-negativa de Y**, que consitui os elementos que certamente não são elementos classificados como elementos de Y .
- $\text{BN}_P(Y)$, **região fronteira de Y**

Uma representação visual destes conceitos é apresentada na Figura 2 (figura extraída de Chouchoulas e Shen, 2001), em que são representados o universo U como uma matriz não regular de duas dimensões e um conjunto $Y \subseteq U$ como um círculo.



Universo U



Região Negativa (a branco)

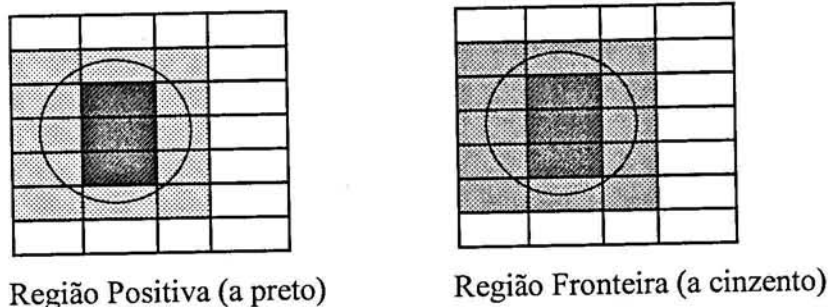


Figura 2 – Representação das aproximações de um conjunto

A **precisão da aproximação** do conjunto $Y \subseteq U$ por P em S , ou, simplesmente, precisão de Y é definida como:

$$\alpha_P(Y) = \frac{\text{card}(\underline{PY})}{\text{card}(\overline{PY})}$$

que tem como objectivo definir o grau do conhecimento que temos sobre o conjunto Y .

4.2.3 Aproximação das classificações

Uma vez que na teoria dos conjuntos aproximativos estamos interessados em classificações de objectos, é interessante calcular a aproximação das classificações. Esta é uma extensão das aproximações de conjuntos. Seja S um sistema de informação, $P \subseteq Q$ e $Y = \{Y_1, Y_2, Y_3, \dots, Y_n\}$ uma classificação (partição) de U em S . A origem desta partição é independente do conjunto de atributos de P . Como exemplo, pode ser obtida a partir da resolução de um problema de ordenação. Os subconjuntos $Y_i, i = 1, \dots, n$ são classes (ou blocos) de Y . Então, definem-se:

- **Aproximação P-restrita (ou região P-positiva)** da classificação Y como:

$$\underline{P}Y = \{\underline{P}Y_1, \underline{P}Y_2, \underline{P}Y_3, \dots, \underline{P}Y_n\}$$

que define todos os objectos que possivelmente são correctamente classificados em cada uma das classes usando o conhecimento estabelecido pelo conjunto de atributos P .

- **Aproximação P-lata** da classificação Y:

$$\bar{P}Y = \{\bar{P}Y_1, \bar{P}Y_2, \bar{P}Y_3, \dots, \bar{P}Y_n\}$$

que define todos os objectos que podem ser correctamente classificados em cada uma das classes usando o conhecimento estabelecido pelo conjunto de atributos P.

- **Aproximação P-fronteira (ou P-duvidosa)** da classificação Y:

$$BN_P(Y) = \bar{P}Y - \underline{P}Y$$

que define todos os objectos que foram incorrectamente classificados em cada uma das classes usando o conhecimento estabelecido pelo conjunto de atributos P.

4.2.4 Precisão da classificação

$$\alpha_P(Y) = \frac{\sum_{j=1}^n \text{card}(\underline{P}Y_j)}{\sum_{j=1}^n \text{card}(\bar{P}Y_j)}$$

e com a notação alternativa de $\mu_P(Y)$ ou $\beta_P(Y)$ expressa a percentagem de possíveis decisões correctas na classificação de objectos recorrendo ao conjunto de atributos P.

4.2.5 Qualidade da classificação

A qualidade da classificação ou qualidade da aproximação de Y por P é definida por:

$$\gamma_P(Y) = \frac{\sum_{j=1}^n \text{card}(PY_j)}{\text{card}(U)}$$

expressa a percentagem de objectos que podem ser correctamente classificados nas classes de Y quando recorrendo ao conjunto de atributos P.

4.2.6 Redução de atributos

O conjunto de atributos $R \subseteq Q$ depende do conjunto de atributos $P \subseteq Q$ em S (notação $P \rightarrow R$) se $\text{IND}(P) \subseteq \text{IND}(R)$. Os valores dos atributos de R são determinados univocamente por valores de atributos de P. A descoberta de relações de dependência entre os atributos é de importância vital para a análise de dados no contexto dos conjuntos aproximativos.

A redução de atributos é, normalmente, efectuada de modo a que o conjunto reduzido de atributos possua a mesma qualidade de classificação do conjunto original de atributos. O subconjunto mínimo $R \subseteq P \subseteq Q$ tal que $\gamma_P(Y) = \gamma_R(Y)$ é denominado Y-reduto (ou simplesmente, reduto, se não existir possibilidade de haver ambiguidade em relação a Y). O Y-reduto de Q é também chamado de conjunto mínimo ou subconjunto em S. De notar que um sistema de informação pode ter mais do que um Y-reduto. $\text{RED}_Y(P)$ é a família de todos os Y-redutos de P. A intersecção de todos os P-redutos de P é chamada de Y-core de P, isto é, $\text{CORE}_Y(P) = \bigcap \text{RED}_Y(P)$. O Y-core de Q é o conjunto dos atributos mais importantes os quais não podem ser eliminados do sistema S sem diminuir a qualidade da aproximação da classificação Y.

4.2.7 Exemplo de aplicação da teoria dos conjuntos aproximativos

Consideremos a situação de um especialista responsável por uma colecção de documentos com informação económica. Este especialista é responsável por estabelecer as diferentes categorias de documentos, analisar novos documentos diariamente determinando se estes devem ou não ser incluídos na colecção e criar, se necessário, novas categorias de

documentos. Este exemplo baseia-se num sistema de informação formada por uma tabela de decisão, sendo esta definida por $TD = \langle U, C \cup D, V, f \rangle$ onde C é o conjunto de atributos de decisão e D é o conjunto de atributos de condição. Designam-se por atributos de condição, os atributos que descrevem um objecto, atributos de entrada de informação. Os atributos de decisão de um objecto são aqueles cujo os seus valores são determinados pelos atributos de condição, isto é, $D = f(C)$. Uma tabela de decisão é determinística se e só se $C \rightarrow D$. Caso contrário, é inconsistente. Uma tabela de decisão determinística descreve univocamente as decisões a serem efectuadas quando algumas das condições são satisfeitas.

Os atributos são, neste caso, obviamente termos do documento. Sejam $R1, R2, R3, R4, R5, R6$ os objectos deste sistema de informação que representam diferentes documentos que devem ser avaliados. Vamos considerar apenas três termos associados a cada documento: $t1, t2, t3$. Para cada termo é atribuído um peso indicativo da sua importância para a representação da informação contida no documento.

Documentos	Termos			Decisão
	t1	t2	t3	Incluir
R1	0.6	0.4	0.6	Sim
R2	0.8	0.4	0.8	Sim
R3	0.9	0.8	0.9	Não
R4	0.6	0.8	0.6	Não
R5	0.8	0.8	0.6	Não
R6	0.9	0.4	0.9	Sim

Tabela 1 – Sistema de Informação

Neste exemplo podemos constatar, se considerarmos o conjunto dos atributos $t1$ e $t3$, que os documentos $R1$ e $R4$ apresentam o mesmo valor (0.6, 0.6). Consequentemente os recursos $R1$ e $R4$ dizem-se indiscerníveis em relação a este conjunto de atributos. Do mesmo modo os documentos $R3$ e $R6$ são também indiscerníveis entre si (0.9, 0.9).

Podemos determinar que os conjuntos elementares, indiscerníveis, formados por estes dois atributos são: $\{R1, R4\}$; $\{R3, R6\}$; $\{R2\}$ e $\{R5\}$. Identificamos dois atributos que definem uma relação de indiscernibilidade nos documentos em análise. A relação de indiscernibilidade permite identificar atributos redundantes, sendo a sua eliminação possível, ou seja, podemos proceder à redução de atributos mantendo a mesma qualidade de classificação. Considerando o conjunto de atributos $\{t1, t2\}$, verificamos que os conjuntos elementares formados por este conjunto são todos singulares, isto é, $\{R1\}$; $\{R2\}$; $\{R3\}$; $\{R4\}$; $\{R5\}$ e $\{R6\}$. Também o são os conjuntos formados pelo conjunto de atributos $\{t1, t2, t3\}$. Se um conjunto de atributos definir as mesmas relações que um conjunto maior de atributos, então o atributo que pertence ao conjunto maior é redundante dado não introduzir nenhuma informação adicional capaz de aumentar a qualidade da decisão. Logo, neste exemplo, o atributo $t3$ é redundante.

O conjunto inicial de atributos $\{t1, t2, t3\}$ foi reduzido para dois $\{t1, t2\}$. Este conjunto é o reduto deste sistema de informação dado ser mínimo e nenhum dos atributos ser redundante. O sistema de informação sobre estes documentos pode ser reformulado tendo em consideração apenas este conjunto de atributos, como apresentado na Tabela 2.

Documentos	Atributos		Decisão
	t1	t2	Incluir
R1	0.6	0.4	Não
R2	0.8	0.4	Sim
R3	0.9	0.8	Sim
R4	0.6	0.8	Não
R5	0.8	0.8	Não
R6	0.9	0.4	Sim

Tabela 2 – Sistema de Informação reduzido

Analisando este sistema de informação verifica-se que o mesmo não é inconsistente, isto é, não apresenta o mesmo conjunto de valores de atributos com uma

decisão diferente. Adicionemos dois novos objectos, documentos, R7 e R8, neste estudo, como apresentado na Tabela 3.

Recursos	Atributos		Decisão
	t1	t2	Incluir
R1	0.6	0.4	Não
R2	0.8	0.4	Sim
R3	0.9	0.8	Sim
R4	0.6	0.8	Não
R5	0.8	0.8	Não
R6	0.9	0.4	Sim
R7	0.8	0.4	Não
R8	0.9	0.4	Não

Tabela 3 – Sistema de Informação reduzido

Os conjuntos elementares definidos pelos atributos são: $\{R1\}$; $\{R2, R7\}$; $\{R3\}$; $\{R4\}$; $\{R5\}$ e $\{R6, R8\}$. Dado que existem decisões diferentes para os documentos R2 e R7 e sendo estes indiscerníveis entre eles, diz-se que o sistema de informação é inconsistente. Tal situação ocorre também para os documentos R6 e R8. Para ultrapassar esta situação e para auxiliar o decisor na sua actividade, são efectuadas aproximações aos conceitos que classificam os objectos. A aproximação restrita é o conjunto de objectos que podem certamente ser classificados como objectos de Y empregando o conjunto de atributos P. Seja R1, R4, R5, R7, R8 os objectos associados à decisão de Incluir=Não, a aproximação restrita é o conjunto $\{R1, R4, R5\}$. Ou seja, estes documentos podem certamente ser classificados como estando associados à decisão de Incluir=Não. A aproximação lata é o conjunto de elementos que podem possivelmente ser classificados como elementos de Y empregando o conjunto de atributos P. Neste caso a aproximação lata é o conjunto $\{R1, R2, R4, R5, R6, R7, R8\}$. A região fronteira é constituída pelos

elementos que não se pode garantir que sejam classificados como elementos de Y, {R2, R6, R7, R8}. A precisão da classificação é:

$$\alpha_p(Y) = \frac{\sum_{j=1}^n \text{card}(\underline{PY}_j)}{\sum_{j=1}^n \text{card}(\overline{PY}_j)} = \frac{3}{7} = 0.43$$

A qualidade da classificação é:

$$\gamma_p(Y) = \frac{\sum_{j=1}^n \text{card}(\underline{PY}_j)}{\text{card}(U)} = \frac{3}{8} = 0.375$$

Para a decisão de Incluir=Sim, a aproximação restrita é o conjunto {R3}, a aproximação lata é {R2, R3, R6, R7, R8} e a região fronteira é o mesmo conjunto anterior {R2, R6, R7, R8}. Para este caso as medidas de classificação apresentam os seguintes valores:

- precisão da classificação:

$$\alpha_p(Y) = \frac{\sum_{j=1}^n \text{card}(\underline{PY}_j)}{\sum_{j=1}^n \text{card}(\overline{PY}_j)} = \frac{1}{5} = 0.2$$

- qualidade da classificação:

$$\gamma_p(Y) = \frac{\sum_{j=1}^n \text{card}(\underline{PY}_j)}{\text{card}(U)} = \frac{1}{8} = 0.125$$

Neste exemplo, o decisor obtém uma maior qualidade sobre a decisão de rejeitar documentos do aquela que obtém para a decisão de incluir documentos na sua colecção sendo, em ambas as decisões, valores relativamente baixos. Consequentemente, o decisor deveria revêr os atributos, termos, que descrevem os documentos em análise para poder aumentar a qualidade e precisão da sua classificação.

4.3 Combinação das teorias dos conjuntos aproximativos e difusos

As teorias de conjuntos aproximativos e difusos são consideradas extensões ou generalizações da teoria clássica de conjuntos. Da descrição efectuada das duas teorias, deve ser claro que estas não são rivais mas captam dois aspectos distintos da imprecisão da informação: indiscernibilidade e incerteza (Dubois e Prade, 1992) que podem estar presentes em simultâneo numa colecção de documentos. Então, é perfeitamente natural efectuar a combinação destas duas teorias. A combinação efectuada baseia-se nos conceitos dos α -cuts desenvolvida por Yao (1997). Partindo do modelo de Pawlak (1982), da teoria dos conjuntos aproximativos, Yao mostra como as *membership functions* (funções de pertença) são usadas para calcular duas aproximações.

Assumindo S um conjunto difuso no universo U , $\mathfrak{R}$ uma relação de equivalência tal que $\mathfrak{R} \subseteq U \times U$ e x, y dois elementos de U , x e y são indiscerníveis se e só se pertencerem à mesma classe de equivalência, isto é, $x\mathfrak{R}y$. Seja μ_S a função de pertença para S , então:

$$\mu_{\underline{apr}_{\mathfrak{R}}(S)}(x) = \inf\{\mu_S(y) \mid y \in U, (x, y) \in \mathfrak{R}\}$$

E se denotar a função de pertença para $\mathfrak{R}$, então:

$$\mu_{\overline{apr}_{\mathfrak{R}}(S)}(x) = \sup\{\mu_S(y) \mid y \in U, (x, y) \in \mathfrak{R}\}$$

Com estas duas expressões combinadas e tendo em conta que $\mu_{\mathfrak{R}}(x, y)$ indica pertença na relação de indiscernibilidade obtemos:

$$\mu_{\underline{apr}_{\mathfrak{R}}(S)}(x) = \inf\{\max[\mu_S(y), 1 - \mu_{\mathfrak{R}}(x, y)] \mid y \in U\}$$

$$\mu_{\overline{apr}_{\mathfrak{R}}(S)}(x) = \sup\{\min[\mu_S(y), 1 - \mu_{\mathfrak{R}}(x, y)] \mid y \in U\}$$

Usando μ_S para definir as aproximações, um elemento x pertence à aproximação restrita de S , com valor de pertença 1, se todos os elementos equivalentes de x pertencerem a S , isto é $\mu_S(y) = 1$. Usando $\mu_{\mathfrak{R}}$ para definir as aproximações, um elemento x pertence à aproximação restrita de S , com valor de pertença 1, se todos os elementos que não pertencem a S não são equivalentes de x , isto é $\mu_{\mathfrak{R}}(x, y) = 0$. Usando ambos $\mu_{\mathfrak{R}}$ e μ_S para definir as aproximações, um elemento x pertence à aproximação restrita de S , com valor de pertença 1, se para todos os elementos do universo U , quando um elemento for equivalente de x , isto é $\mu_{\mathfrak{R}}(x, y) = 1$, então está presente em S , isto é, $\mu_S(y) = 1$. Consequentemente, estas funções μ_S e $\mu_{\mathfrak{R}}$ podem ser usadas para calcular as aproximações no caso em que degeneram, assumindo o valor de 0 ou 1, isto é, quando S e $\mathfrak{R}$ são conjuntos rígidos. Estas funções são igualmente importantes quando S e $\mathfrak{R}$ são conjuntos difusos.

4.3.1 Conjunto aproximativo difuso (*rough fuzzy set*)

Um conjunto aproximativo difuso é obtido pela extensão de um conjunto difuso num espaço de aproximação rígido.

Seja F um conjunto difuso num espaço de aproximação $\underline{apr}_{\mathfrak{R}} = (U, \mathfrak{R})$ com uma relação de equivalência $\mathfrak{R}$.

Para qualquer F_{α} como um conjunto de referência, um conjunto aproximativo $(\underline{apr}_{\mathfrak{R}}(F_{\alpha}), \overline{apr}_{\mathfrak{R}}(F_{\alpha}))$ pode ser definido usando o modelo standard de Pawlak. Considerando a família de α -cuts, obtemos uma família de aproximações $(\underline{apr}_{\mathfrak{R}}(F_{\alpha}))_{\alpha}$ e $(\overline{apr}_{\mathfrak{R}}(F_{\alpha}))_{\alpha}$, $\alpha \in [0, 1]$ de dois conjuntos difusos, dado que F_{α} é construído a partir de um conjunto difuso F . Generalizando, um conjunto pode ser definido como aproximativo difuso dado uma aproximação ser um conjunto difuso, tomando como referência um conjunto difuso. O valor de pertença de x , pertencendo ao conjunto difuso $\underline{apr}_{\mathfrak{R}}(F)$ é o

mínimo dos valores dos elementos na classe de equivalência que contém x , e o valor de x pertencendo a $\overline{apr}_{\mathfrak{R}}(F)$ é o máximo. O par $(\underline{apr}_{\mathfrak{R}}(F), \overline{apr}_{\mathfrak{R}}(F))$ é um conjunto aproximativo difuso tomando F como conjunto de referência difuso. Por outras palavras, um conjunto aproximativo difuso é caracterizado por três conjuntos difusos:

- Conjunto difuso de referência: μ_F ,
- Aproximação restrita: $\mu_{\underline{apr}_{\mathfrak{R}}(F)}(x) = \inf\{\mu_F(y) \mid y \in U, (x, y) \in \mathfrak{R}\}$
- Aproximação lata: $\mu_{\overline{apr}_{\mathfrak{R}}(F)}(x) = \sup\{\mu_F(y) \mid y \in U, (x, y) \in \mathfrak{R}\}$

Um α -cut de um conjunto aproximativo difuso pode ser definido através dos α -cuts de um conjunto difuso F :

$$\begin{aligned} (\underline{apr}_{\mathfrak{R}}(F), \overline{apr}_{\mathfrak{R}}(F))_{\alpha} &= (\underline{apr}_{\mathfrak{R}}(F_{\alpha}), \overline{apr}_{\mathfrak{R}}(F_{\alpha})) \\ &= ((\underline{apr}_{\mathfrak{R}}(F))_{\alpha}, (\overline{apr}_{\mathfrak{R}}(F))_{\alpha}) \end{aligned}$$

4.3.2 Conjunto difuso aproximativo (fuzzy rough set)

Um conjunto difuso aproximativo é derivado da aproximação de um conjunto rígido num espaço de aproximação difuso.

Considerando um espaço de aproximação difuso $apr_{\mathfrak{R}} = (U, \mathfrak{R})$ com $\mathfrak{R}$ representando uma relação de semelhança difusa, é possível definir um β -cut com $\beta \in [0, 1]$, similar ao α -cut de um conjunto difuso, numa relação de semelhança $\mathfrak{R}$ tal que cada β -conjunto de $\mathfrak{R}$ é uma relação de equivalência, que pode ser representada como uma família de relações de equivalência:

$$\mathfrak{R} = (\mathfrak{R}_{\beta})_{\beta} \text{ com } \beta \in [0, 1]$$

que define uma família de espaços de aproximação:

$$apr_{\mathfrak{R}} = (apr_{\mathfrak{R}_\beta} = (U, \mathfrak{R}_\beta))_\beta, \beta \in [0,1]$$

Assim, é possível aplicar-se o modelo de Pawlak para cada β -cut e para um dado subconjunto de S em U podemos derivar um conjunto aproximativo $(\underline{apr}_{\mathfrak{R}_\beta}(S), \overline{apr}_{\mathfrak{R}_\beta}(S))$ em cada uma das relações de equivalência de nível β .

O par $(\underline{apr}_{\mathfrak{R}}(S), \overline{apr}_{\mathfrak{R}}(S))$ é um conjunto difuso aproximativo com S como conjunto de referência. Por outras palavras, um conjunto difuso aproximativo é caracterizado por um conjunto rígido e dois conjuntos difusos:

- Conjunto de referência: $S \subseteq U$,
- Aproximação restrita: $\mu_{\underline{apr}_{\mathfrak{R}}(S)}(x) = \inf\{1 - \mu_{\mathfrak{R}}(x, y) \mid y \notin S\}$
- Aproximação lata: $\mu_{\overline{apr}_{\mathfrak{R}}(S)}(x) = \sup\{\mu_{\mathfrak{R}}(x, y) \mid y \in S\}$

Um β -cut de um conjunto difuso aproximativo é definido através dos β -cuts de um conjunto difuso F :

$$\begin{aligned} (\underline{apr}_{\mathfrak{R}}(F), \overline{apr}_{\mathfrak{R}}(F))_\beta &= (\underline{apr}_{\mathfrak{R}}(F_\beta), \overline{apr}_{\mathfrak{R}}(F_\beta)) \\ &= ((\underline{apr}_{\mathfrak{R}}(F))_\beta, (\overline{apr}_{\mathfrak{R}}(F))_\beta) \end{aligned}$$

4.3.3 Conjunto difuso e espaço difuso aproximativo

Esta abordagem, mais generalizada, permite a aproximação de um conjunto difuso num espaço difuso aproximativo. Assim, podemos usar os α -cuts no conjunto difuso F para obter conjuntos rígidos e β -cuts na relação de semelhança difusa $\mathfrak{R}$ de modo a obter relações de equivalência. Esta operação permite a obtenção de uma família de conjuntos aproximativos $(\underline{apr}_{\mathfrak{R}_\beta}(F_\alpha), \overline{apr}_{\mathfrak{R}_\beta}(F_\alpha))$, $\alpha \in [0,1]$, $\beta \in [0,1]$. A combinação pode ser interpretada de três formas diferentes: como uma família de conjuntos aproximativos, uma

família de conjuntos aproximativos difusos e uma família de conjuntos difusos aproximativos, dependendo da forma de como estes conjuntos estejam agrupados. Independentemente da interpretação, versões generalizadas das expressões anteriormente definidas podem ser obtidas para determinar o valor de pertença nos conjuntos difusos aproximativos.

$$\mu_{\underline{apr}_{\Gamma}(\Delta)}(x) = \inf\{\max[\mu_{\Delta}(y), 1 - \mu_{\Gamma}(x, y)] \mid y \in U\}$$

$$\mu_{\overline{apr}_{\Gamma}(\Delta)}(x) = \sup\{\min[\mu_{\Delta}(y), \mu_{\Gamma}(x, y)] \mid y \in U\}$$

A variável Γ pode definir quer uma relação de equivalência quer uma relação difusa de semelhança. A variável Δ pode definir um subconjunto rígido ou difuso.

5 Combinação das teorias dos conjuntos aproximativos e difusos na recuperação de informação

Este capítulo descreve abordagens da combinação da teoria dos conjuntos aproximativos e da teoria dos conjuntos difusos na recuperação de informação. É apresentada uma arquitectura, RST, para um protótipo de um sistema de recuperação que é usado para experimentar combinações destas teorias.

Nos capítulos anteriores, a teoria dos conjuntos difusos e a teoria dos conjuntos aproximativos foram descritos e analisados em termos formais: a teoria dos conjuntos difusos é adequada para a integração da imprecisão na informação, enquanto que a teoria dos conjuntos aproximativos permite estabelecer classes de equivalência de objectos referindo o mesmo conceito. A terminologia e os modelos lógicos usados na recuperação de informação foram igualmente descritos.

A definição de estruturas de indexação que estão associadas aos termos da colecção e a classificação de documentos são áreas onde estas teorias têm sido aplicadas no processo de recuperação. A construção de tesauros e a geração de clusters são dois exemplos. A combinação das duas teorias não tem sido aplicada em sistemas de recuperação associados às necessidades do mundo real com recurso, por exemplo, às colecções da TREC¹ de modo a comprovar a sua aplicabilidade.

5.1 Motivação para a combinação de duas visões distintas

A motivação para o desenvolvimento de um sistema de recuperação de informação é responder às necessidades de informação do utilizador disponibilizando uma interacção

¹ TREC – Text REtrieval Conference (<http://trec.nist.gov/>)

simples e natural com a colecção de documentos em que este, mesmo sem possuir conhecimento detalhado sobre a colecção, consiga recuperar documentos relevantes.

Os sistemas de recuperação de informação que apresentam uma concepção direccionada para utilizadores não familiarizados com as colecções de documentos apresentam três características principais: ordenação dos resultados obtidos por relevância estimada ou probabilística para o utilizador; o uso da linguagem natural na interrogação, sem a necessidade de introduzir operadores booleanos ou lógicos; *relevance-feedback*, reformulação das interrogações efectuadas pelo utilizador com base na informação previamente recuperada. Estes sistemas efectuem a indexação de termos associados aos documentos, recorrendo a um método de co-ocorrência que estima a contribuição de cada termo para cada documento. No entanto, estes métodos não têm em consideração o desconhecimento que o utilizador apresenta face ao vocabulário da colecção e a ambiguidade que os termos da interrogação muitas vezes apresentam face ao conteúdo dos documentos. A teoria dos conjuntos difusos apresenta-se, neste âmbito, como a resposta natural para o tratamento da indexação. De facto, possuindo a capacidade de expandir os conceitos para além da sua concepção binária, a teoria dos conjuntos difusos, promove a possível correcção da interrogação do utilizador expandindo os conceitos introduzidos na fase de comparação entre termos e documentos.

Uma das condições básicas na adopção de um dado modelo está associado ao mecanismo de ordenação dos documentos (Baeza-Yates e Ribeiro-Neto, 1999). Esta premissa é tanto mais importante se orientarmos o nosso modelo para a recuperação de documentos Web, pois o utilizador típico de um motor de busca apenas irá analisar os primeiros 20 ou 30 documentos, de dezenas ou mesmo centenas de milhares de documentos recuperados. Neste âmbito e para respondermos à nossa permanente motivação na área da recuperação de informação, a teoria dos conjuntos aproximativos, na sua capacidade de categorizar objectos, apresenta-se como a resposta natural a esta dificuldade crescente dos sistemas de recuperação.

Consequentemente, a combinação da teoria dos conjuntos difusos e da teoria dos conjuntos aproximativos apresenta os fundamentos que permitem responder as novas exigências do mundo real.

5.2 Propostas de combinação

A investigação dos fundamentos teóricos para a combinação da teoria dos conjuntos difusos e da teoria dos conjuntos aproximativos tem sido fundamentalmente desenvolvida por Pawlak (1985), Dubois e Prade (1990), Lin (1992), Nakamura (1992), Yao (1997, 1998, 2002) e Yao et al. (1994).

Duas vertentes destacam-se no estudo da combinação das duas teorias na área da recuperação de informação: Miyamoto (1998) apresenta um modelo lógico difuso onde é definido as aproximações restrita e lata para o resultado obtido da comparação da interrogação com a colecção de documentos e Srinivasan (1989), Srinivasan (1991) e Srinivasan et al. (1998) desenvolveram investigações orientadas para a estruturação do vocabulário da colecção de documentos, cujo suporte teórico é baseado no contributo de Yao (1997). Várias abordagens dos conjuntos aproximativos têm sido apresentadas separadamente da dos conjuntos difusos na recuperação de informação. A investigação na sua combinação não é muito extensa, devido possivelmente à ineficiência computacional que resulta do facto de alguns dos princípios teóricos não serem directamente aplicáveis na recuperação de informação, como sejam a definição de relações de equivalência na teoria dos conjuntos aproximativos e as operações entre conjuntos no caso da teoria dos conjuntos difusos.

A abordagem de Miyamoto (1998) é definida na constatação de que as duas teorias devem ser consideradas ortogonais ou complementares, pois captam aspectos diferentes da recuperação de informação. Enquanto que a teoria dos conjuntos difusos está relacionada com a incerteza das linguagens naturais, a teoria dos conjuntos aproximativos está associada à definição de categorias de itens de informação. Sendo assim, devem ser aplicadas nas partes do processo de recuperação onde possuem capacidade para demonstrar toda a sua valência: a teoria dos conjuntos difusos no estabelecimento de estruturas de indexação de documentos, associação de termos, tesouro, captando os problemas de semântica; e a teoria dos conjuntos aproximativos na definição de classes de documentos, categorias da informação criando agrupamentos em clusters.

Miyamoto (1998) apresenta uma metodologia para a recuperação usando a teoria dos conjuntos aproximativos e defendendo que a componente difusa na sua proposta é inerente pelo uso de aproximações que aplica quer numa situação de recuperação difusa quer numa de recuperação rígida. Uma das abordagens da sua metodologia é a definição de uma variante do conceito de relação de equivalência demonstrando que enquanto um conjunto de classes de equivalência estão associadas a uma relação de equivalência, uma classificação hierárquica está associada a uma relação de equivalência difusa. Uma relação de equivalência difusa é uma relação Q que é:

- Reflexiva, $Q(x, x) = 1$ para todo $x \in X$
- Simétrica, $Q(x, y) = Q(y, x)$ para todo $x, y \in X$
- Transitiva, $Q(x, z) \geq \max_{y \in X} \min[Q(x, y), Q(y, z)]$ para todo $x, z \in X$

Assim, embora esta abordagem não tenha a combinação das duas teorias como algo explícito, ela está presente. Consequentemente, é estabelecido o enquadramento teórico para o uso das classificações hierárquicas aglomerativas na recuperação de informação, num contexto da teoria dos conjuntos aproximativos. Nesta abordagem, são referenciadas também outras formas da aplicação da teoria dos conjuntos aproximativos, usando as definições de aproximações estrita e lata de conjuntos difusos e rígidos, conjuntos aproximativos difusos e conjuntos difusos aproximativos, segundo as definições anteriores. O uso de aproximações deverá, no entanto, apresentar dificuldades de implementação quando a informação a recuperar for do tipo multimédia (imagens) dada a complexidade das estruturas de indexação associadas a este tipo de objectos (Miyamoto, 1998). A recuperação de informação nesta área apresenta-se ainda como um campo fértil para a investigação e experimentação.

O contributo teórico de Yao (1997) permite a Srinivasan et al. (1998) estabelecer:

- Descrição difusa de documento, onde um documento é descrito por um conjunto difuso de termos ao qual está associado um valor de pertença que indica o quanto o termo é representativo do documento.

- Noção de interrogação difusa, descrita por um conjunto de termos a que se associa um valor indicativo da importância para a informação que o documento deve conter.
- Relação de semelhança difusa para o espaço de vocabulário dos termos da colecção de documentos.

Na abordagem de Srinivasan et al. (1998), as duas teorias também intervêm em situações distintas cooperando no estabelecimento do vocabulário (classes de equivalência com termos de índice difusos). A teoria dos conjuntos difusos tem a sua intervenção na atribuição de pesos para os termos dos documentos e interrogação e a teoria dos conjuntos aproximativos na definição do vocabulário, na comparação e ordenação dos documentos. A representação lógica dos documentos e interrogações neste modelo é efectuada pelo conjunto de termos que constituem cada um destes objectos (documentos ou interrogações), após terem sido submetidos à fase de processamento de texto, isto é, eliminação de termos pertencentes à lista de paragem, entre outros procedimentos que podem ser considerados. Estabelecidos os valores difusos para cada um dos termos que constituem os documentos da colecção e a interrogação, é efectuada uma comparação, pelo cálculo das aproximações restrita e lata, e posterior ordenação dos documentos. Esta comparação estabelece a semelhança, restrita e lata, entre os documentos e respectiva interrogação, sendo estabelecido matematicamente o modo para comparação de dois conjuntos aproximativos. A investigação nesta nova abordagem continua com o objectivo de estabelecer um só valor para a recuperação combinando as duas novas medidas de similaridade sugeridas. Para melhor compreensão apresenta-se, resumidamente, esta nova proposta.

Sejam dois subconjuntos S_1, S_2 de U num espaço de aproximação (U, R) , as relações de semelhança restrita e lata entre S_1 e S_2 , com S_2 como conjunto de referência e usando as aproximações restrita e lata são calculadas como:

$$\underline{\text{Semelhança}}_R(S_1, S_2) = 1 - \left[\text{card}(B_1) / \text{card}(\underline{\text{apr}}_R(S_2)) \right]$$

$$\overline{\text{Semelhança}}_R(S_1, S_2) = 1 - \left[\text{card}(B_u) / \text{card}(\overline{\text{apr}}_R(S_2)) \right]$$

em que os conjuntos B_u e B_l são obtidos a partir de:

$$B_l = \text{apr}_R(S_2) - |(\text{apr}_R(S_1) \cap \text{apr}_R(S_2))| \text{ e}$$

$$B_u = \overline{\text{apr}}_R(S_2) - |(\overline{\text{apr}}_R(S_1) \cap \overline{\text{apr}}_R(S_2))|$$

onde $|-|$ representa a diferença de conjuntos. Estes conjuntos definem os elementos não comuns entre as aproximações restrita e lata dos subconjuntos S_2 e S_1 . Esta nova medida de similaridade assume o valor 0 para nenhuma semelhança e 1 para uma máxima semelhança entre os dois subconjuntos.

Considere-se que o vocabulário da colecção dos documentos está definido pela relação de equivalência baseada na sinonímia dos termos. Assim, seja $T = \{t_1, t_2, \dots, t_{10}\}$ o conjunto de termos que representa o vocabulário, e R uma relação de sinonímia equivalente. Podemos definir um subconjunto $T/R = \{C_1, C_2, C_3, C_4, C_5\}$ onde C_i é uma classe de equivalência em R . Cada C_i é um subconjunto, tal que $C_1 = \{t_1, t_4, t_6\}$, $C_2 = \{t_3, t_7\}$, $C_3 = \{t_5, t_8, t_{10}\}$, $C_4 = \{t_9\}$, $C_5 = \{t_2\}$.

Assumindo o documento difuso $FD_2 = \{0.9/t_1, 0.7/t_4, 0.5/t_3, 0.8/t_9\}$ e a interrogação $FQ_2 = \{0.5/t_1, 0.2/t_3, 0.3/t_2, 0.5/t_9\}$, a aproximação restrita e lata difusa para FD_2 , aplicando as funções de pertença:

$$\text{apr}_R(FD_2) = \{0.8/t_9\}, \text{ porque só } C_4 \subseteq FD_2$$

$$\overline{\text{apr}}_R(FD_2) = \{0.9/t_1, 0.9/t_4, 0.9/t_6, 0.5/t_3, 0.5/t_7, 0.8/t_9\}, t_4 \text{ e } t_6 \text{ vêm através de } C_1$$

e a operação de $\cup$ mantém o maior grau.

Do mesmo modo para FQ_2 ,

$$\underline{apr}_R(FQ_2) = \{0.3/t_2, 0.5/t_9\}$$

$$\overline{apr}_R(FQ_2) = \{0.5/t_1, 0.5/t_4, 0.5/t_6, 0.2/t_3, 0.2/t_7, 0.3/t_2, 0.5/t_9\}$$

Como descrito anteriormente dois espaços aproximativos S_1, S_2 podem ser comparados no espaço de aproximação (U, R) . Alterando ligeiramente as equações de modo a envolver os cortes de nível α obtemos:

$$\underline{Sim}_R(S_1, S_2)_\alpha = 1 - \left[\text{card}(B_l) / \text{card}(\underline{apr}_R(S_2)) \right]_\alpha$$

$$\overline{Sim}_R(S_1, S_2)_\alpha = 1 - \left[\text{card}(B_u) / \text{card}(\overline{apr}_R(S_2)) \right]_\alpha$$

Substituindo pelos valores obtemos:

$$B_l = \underline{apr}_R(FQ_2) - | \underline{apr}_R(FD_2) \cap \underline{apr}_R(FQ_2) | = \{0.3/t_2\}$$

Logo pela equação

$$\underline{Sim}_R(FD_2, FQ_2)_0 = 1 - 1/2 = 0.5 \text{ e } \underline{Sim}_R(FD_2, FQ_2)_{0.4} = 1.0$$

Também obtemos:

$$B_u = \overline{apr}_R(FQ_2) - | \overline{apr}_R(FD_2) \cap \underline{apr}_R(FQ_2) | = \{1 - 1/4\} = 0.75. \text{ Logo,}$$

$$\overline{Sim}_R(FD_2, FQ_2)_0 = 1 - 1/7 = 0.86 \text{ e}$$

$$\overline{Sim}_R(FD_2, FQ_2)_{0.4} = 1 - 1/4 = 0.75$$

Este modelo é diferente, devendo ser mais capaz de responder às questões postas pela recuperação de informação, estando baseado na construção de um vocabulário. No entanto, a proposta de Srinivasan et al. (1998) sofre de um problema que é a necessidade de definir uma relação de equivalência. A definição de uma tal relação na área da recuperação de informação é demasiado restritiva, uma vez que nem sempre a propriedade de transitividade se confirma (Ho, Funakoshi, 1998) (Miyamoto, 1998). Do mesmo modo, os testes efectuados restringiram-se apenas a um domínio, *Unified Medical Language System* (UMLS), tendo sido detectadas várias relações binárias que não possuíam algumas das propriedades necessárias para ser considerada uma relação de equivalência. Esta nova abordagem desenvolvida num âmbito estrito na área da recuperação carece, assim, de uma experimentação mais profunda. O desafio que se coloca é na forma de construção deste vocabulário e na performance do sistema quando aplicado a uma colecção do mundo real. Foi esta dificuldade da teoria dos conjuntos aproximativos que levou Ho e Funakoshi (1998) a apresentar uma inovadora abordagem na área da recuperação de informação desenvolvendo um novo modelo designado por *Tolerance Rough Set Model* (TRSM) que expande o conceito do modelo de Pawlak (1982), aqui designado por *Equivalence Rough Set Model* (ERSM), cujo fundamento teórico são classes de tolerância, classes não disjuntas, construídas a partir da co-ocorrência de termos e implementando um sistema de recuperação usando um novo algoritmo recorrendo a aproximações. Este novo modelo do TRSM foi aplicado com sucesso a uma pequena colecção de documentos da JSAI (*Journal of the Japanese Society for Artificial Intelligence*) com 802 documentos.

5.3 Medida de relevância

Na avaliação de um sistema de recuperação, um critério chave de performance do sistema é o da resposta apresentada pelo sistema ser idêntica à recuperada pelo decisor ou especialista que efectuou uma avaliação prévia de todos os objectos da colecção, o que nem sempre é praticável.

A recuperação de informação desenvolvida nos fundamentos da teoria dos conjuntos difusos, pela sua propriedade de extensão de conceitos semânticos, pode ter

como efeito a recuperação de um número elevado de documentos, em que cada documento na resposta obtida se apresenta como uma aproximação à interrogação definida pelo utilizador. Consequentemente, é expectável que a precisão da resposta seja relativamente baixa enquanto que a recuperação seja alta. Tal confirma o comportamento que uma maior recuperação corresponde uma menor precisão. No âmbito desta particular abordagem, foram desenvolvidas várias medidas de relevância associada à teoria dos conjuntos difusos, quase todos baseados na cardinalidade relativa entre dois conjuntos difusos, desenvolvida por Zadeh (1965):

$$Card(A/B) = \frac{\sum -count(A \cap B)}{\sum -count(B)}$$

onde $\sum -count(F) = \sum_{x \in X} F(x)$ é a medida, introduzida por De Luca e Termini (1972), da cardinalidade de um conjunto difuso F definido num universo finito X . Exemplos de medidas difusas baseadas na cardinalidade de Zadeh são as propostas por Buell e Kraft (1981), que apresentaram uma das primeiras extensões difusas ao modelo de recuperação booleano, para a recuperação $\tilde{\rho}$ e precisão $\tilde{\psi}$:

$$\tilde{\rho} = \frac{\sum_{i=1}^n (\beta_i \wedge \alpha_i)}{\sum_{i=1}^n \beta_i} \quad \tilde{\psi} = \frac{\sum_{i=1}^n (\beta_i \wedge \alpha_i)}{\sum_{i=1}^n \alpha_i}$$

em que:

- Q denota a interrogação do utilizador ao sistema que é definida por um conjunto difuso $Q = \beta_1 / \omega_1 + \beta_2 / \omega_2 + \dots + \beta_n / \omega_n$, em que ω_i indica o documento i da colecção.
- β_i é o grau, definido pelo especialista, pelo qual o documento ω_i satisfaz a interrogação Q .

- α_i é o grau, definido pelo sistema, pelo qual o documento ω_i satisfaz a interrogação Q .
- $\wedge$ é a operação de mínimo sobre os valores de pertença entre os termos comuns de dois conjuntos difusos.

e a medida difusa proposta por Sánchez e Pierre (1994), onde um limiar θ é fixado para os documentos serem incluídos na resposta e a cardinalidade do denominador para a precisão é calculada com base na cardinalidade do conjunto dos documentos recuperados. Recentemente, Martín-Bautista et al. (2000) apresentaram uma medida difusa alternativa, não baseada na cardinalidade relativa de Zadeh, que segundo os autores é mais eficiente em termos computacionais, $O(n)$, e mais intuitiva. Esta medida é baseada num novo método introduzido por Delgado et al. (1999) de cálculo da cardinalidade relativa de conjuntos difusos. A recuperação, ρ , designada por $SVD(S/E)$ é definida como:

$$\rho = SVD(S/E) = \sum_{\alpha_i \in \Delta(S/E)} \left((\alpha_i - \alpha_{i+1}) \frac{|(S \cap E)^{(\alpha_i)}|}{|E^{(\alpha_i)}|} \right)$$

onde:

- $\Delta(S/E) = \Delta(S \cap E) \cup \Delta(E) = \{\alpha_1, \dots, \alpha_q\}$ em que $S \cap E$ designa a intersecção de dois conjuntos difusos, $\cup$ a operação de união e os elementos do conjunto estão em ordem decrescente, isto é, $1 = \alpha_1 > \alpha_2 > \alpha_3 > \dots > \alpha_q > \alpha_{q+1} = 0$.
- $\Lambda(F) = \{F(x) \mid x \in X, F(x) > 0\}$ é o conjunto de nível de F .
- $F^{(\alpha)}$ é o α -cut de F e $|\cdot|$ é a cardinalidade de um conjunto.

É assumido que o conjunto E é normalizado. Caso não esteja, este é normalizado previamente ao cálculo da precisão sendo o factor de normalização aplicado a $S \cap E$. A precisão ψ é definida como:

$$\psi = SVD(E/S) = \sum_{\alpha_i \in \Delta(E/S)} \left((\alpha_i - \alpha_{i+1}) \frac{|(S \cap E)^{(\alpha_i)}|}{|S^{(\alpha_i)}|} \right)$$

sendo igualmente assumido que o conjunto S se encontre normalizado.

Como exemplo consideremos dois conjuntos difusos A e B , definidos sobre um universo finito $X = \{x_1, x_2, \dots, x_n\}$ em que $n = 1000$.

$$A(x) = \begin{cases} 1 & x = x_1 \\ 0.01 & x \neq x_1 \end{cases}$$

$$B(x) = \begin{cases} 1 & x = x_2 \\ 0.01 & x \neq x_2 \end{cases}$$

Neste exemplo obtemos $(A \cap B)(x) = 0.01$ e a cardinalidade relativa de A em relação a B obtida pela definição de Zadeh tem como resultado:

$$\begin{aligned} Card(A/B) &= \frac{1000 * 0.01}{1 + 999 * 0.01} \\ &= 0.91 \end{aligned}$$

Pela cardinalidade de Zadeh foi obtido um valor muito elevado para dois conjuntos que não apresentam elementos comuns com elevados valores de pertença. Este valor deveria ser baixo pois intuitivamente apenas o elemento x_1 está claramente no conjunto A .

Consideremos este exemplo aplicado à nova medida de recuperação proposta por Martín-Bautista et al. (2000) em que $S=A$ e $E=B$. Aplicando os conceitos obtemos $\Delta(A/B) = \{1, 0.01\}$. De imediato $q = 2$ e $\alpha_1 = 1$, $\alpha_2 = 0.01$, $\alpha_3 = 0$. Aplicando a medida proposta para a recuperação obtemos:

$$\begin{aligned} \rho = SVD(A/B) &= \sum_{\alpha_i \in \Delta(A/B)} \left((\alpha_i - \alpha_{i+1}) \frac{|(A \cap B)^{(\alpha_i)}|}{|B^{(\alpha_i)}|} \right) \\ &= (\alpha_1 - \alpha_2) \frac{|(A \cap B)^{(\alpha_1)}|}{|B^{(\alpha_1)}|} + (\alpha_2 - \alpha_3) \frac{|(A \cap B)^{(\alpha_2)}|}{|B^{(\alpha_2)}|} \\ &= (1 - 0.01) \frac{|(A \cap B)^{(1)}|}{|B^{(1)}|} + (0.01 - 0) \frac{|(A \cap B)^{(2)}|}{|B^{(2)}|} \end{aligned}$$

$$= (1 - 0.01) \frac{0}{1} + (0.01 - 0) \frac{1000}{1000}$$

$$= 0.01$$

Este valor para a recuperação é mais intuitivo do que o anteriormente obtido, uma vez que só existe, de facto, um documento relevante não sendo considerada esta informação com a medida de Zadeh, que indicava uma recuperação elevada.

A teoria dos conjuntos aproximativos tem sido fundamentalmente adoptada como o método de ordenação e definição de categorias dos resultados ao utilizador (Miyamoto, 1998). Para tal, são usados múltiplos métodos para geração de clusters, hierárquicos ou não, para agrupar os documentos recuperados pelo sistema. Nestes métodos não são usadas medidas de recuperação ou precisão associadas aos clusters dado os documentos serem inseridos num determinado cluster com base numa medida de similaridade entre documentos. Esta é usada também para ordenar os clusters entre si, existindo uma variedade de medidas desenvolvidas ao longo do tempo, dado este tópico na área de recuperação estar continuamente sob o olhar atento dos investigadores.

Como podemos constatar cada abordagem apresenta as suas próprias medidas para indicar a qualidade da recuperação de informação apresentada ao utilizador. Os modelos baseados na teoria dos conjuntos difusos descrevem, inclusivamente, medidas próprias. O utilizador, quando visualiza os resultados obtidos, deve ser informado de forma clara e concisa sobre a qualidade da resposta obtida, conforme a sua apresentação. Tratando-se de uma abordagem que efectua a combinação de duas teorias em fases distintas do processo de recuperação, não existe uma medida de relevância desenvolvida que contemple a combinação das duas teorias. Neste âmbito, considera-se que a melhor proposta para uma correcta informação ao utilizador, consiste numa medida de relevância sobre o conjunto de documentos obtidos antes do agrupamento em clusters. Nesta situação, as medidas de recuperação e precisão usualmente usadas servem este objectivo podendo, eventualmente, ser adoptada uma medida difusa. Posteriormente, e para cada cluster gerado, uma medida de similaridade dos documentos do cluster para com a interrogação é adoptada. Esta abordagem permitirá ao utilizador ter uma perspectiva global da resposta, identificar e

estabelecer relações entre clusters de modo a permitir uma redefinição da sua interrogação face aos documentos definidos em cada cluster.

5.4 RST – Arquitectura do protótipo

A definição de uma arquitectura que combine as duas teorias em estudo para a implementação de um protótipo de um sistema de recuperação de informação, teve como motivação determinar a aplicabilidade prática e a eficiência da proposta teórica. Em múltiplas situações, novas abordagens só conseguem obter aplicabilidade numa área ou domínio muito restrito, dependendo, inclusivamente, do tipo de recuperação (texto, imagens, etc.). Consequentemente, uma nova abordagem deve ser apresentada com o devido suporte prático que esta área requer. A proposta consiste no desenvolvimento de um protótipo de um sistema de recuperação, designado por RST1, cujo objectivo é o da recuperação de informação em colecções de documentos de texto. A arquitectura do sistema designa-se por RST.

Um sistema de recuperação de informação tem como objectivo recuperar documentos que possam satisfazer as necessidades de informação do utilizador. Um sistema de informação permite informar o utilizador da existência e localização dos documentos, sendo possível visualizar os documentos em parte ou no seu todo. Neste caso, obviamente, também foi este o objectivo. Para além deste, e dado tratar-se de um protótipo, a arquitectura foi desenhada no sentido de oferecer a maior flexibilidade, modularidade ao sistema com a motivação de algum componente ser facilmente alterado sem interferir com os restantes. A arquitectura do RST é apresentada na Figura 3, denotando uma separação clara entre os diversos processos básicos existentes num sistema de recuperação de informação: processamento da interrogação, indexação de documentos, processamento da comparação entre a interrogação e a colecção de documentos de modo a obter a resposta a ser apresentada ao utilizador. Cada módulo aplicacional tem como dados de entrada os de saída do processo ou módulo anterior, estabelecendo-se a troca de informação por meio de ficheiros sequenciais associados a estruturas internas.

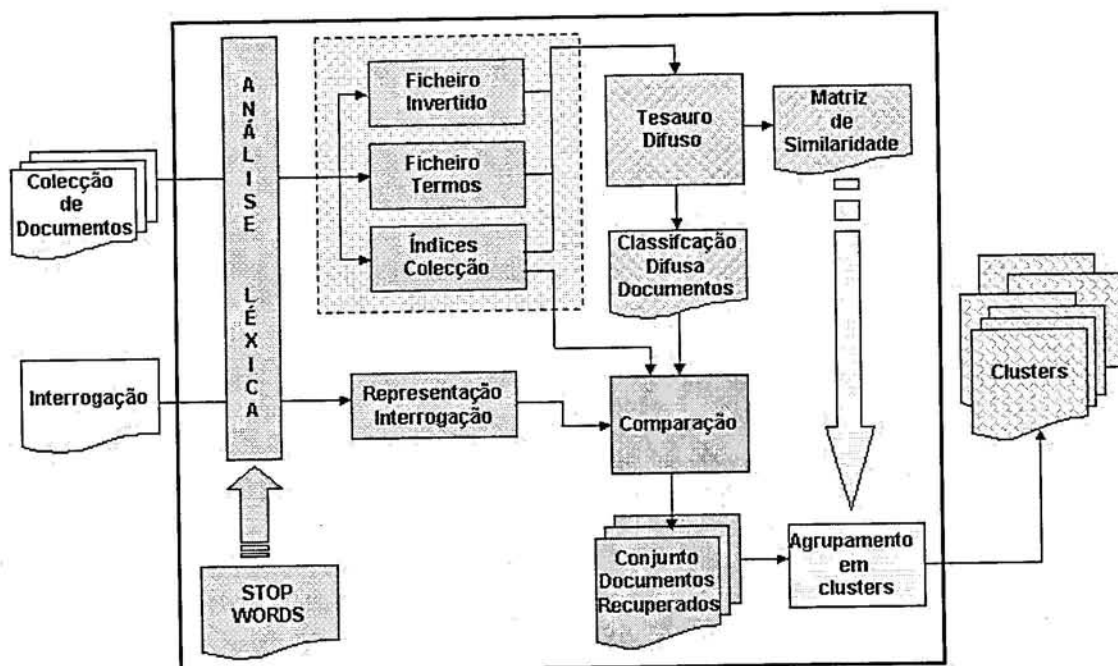


Figura 3 – Arquitectura do sistema de recuperação de informação RST

Em seguida são analisados cada um dos processos do sistema de recuperação de informação.

5.4.1 Representação do documento

A representação interna de um documento está intimamente associada ao modelo lógico de recuperação adoptado. No modelo adoptado, um documento D é representado por um conjunto difuso $D = \{\mu_1 T_1, \mu_2 T_2, \mu_3 T_3, \dots, \mu_n T_n\}$. Cada termo de indexação do documento, T_i é um termo presente no documento excluindo termos com baixo valor para a representação da informação contida no documento – lista de paragem. Este processo de eliminação ocorre durante a análise léxica do RST. Um valor de inclusão, μ_i , é atribuído a cada um dos termos que indica o nível de coesão do termo relativamente à informação

descrita pelo documento. Em outras palavras, μ_i indica a compatibilidade do termo para com os conceitos presentes na informação descrita pelo documento. A representação de um documento é obtida no fim do processo da classificação difusa de documentos, sendo usada no processo de comparação.

5.4.2 Representação da interrogação

A interrogação, q , é representada por um conjunto rígido de termos $q = \{t_1, t_2, t_3, \dots, t_n\}$, onde um termo pode estar ou não presente no conjunto de termos usados para a indexação dos documentos, dado ser permitido introduzir uma interrogação em linguagem natural no sistema RST. A interrogação, tal como o documento, é sujeita a um processo de análise léxica no sentido de serem removidos termos com pouco ou nenhum valor para posterior comparação. Consideremos uma interrogação, em inglês, retirada da colecção de teste da CRAN e definida como: *"what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft"*. A análise léxica é aplicada e cria o conjunto: {aircraft, speed, high, heated, models, aeroelastic, constructing, obeyed, laws, similarity}, sendo definida uma importância idêntica para todos os termos introduzidos na interrogação. A pressuposição de que todos os termos na interrogação contribuem no mesmo grau pode ser considerada redutora, uma vez que não é permitida a definição pelo utilizador do valor μ_i de cada termo t_i . No entanto, esta pressuposição é efectuada tendo em consideração que o sistema é implementado de modo a permitir a introdução de interrogações em linguagem natural e que a possível imprecisão intrínseca em cada termo é posteriormente corrigida no processo da construção do tesauro. Consequentemente, esta representação é considerada adequada para o contexto do sistema RST.

5.4.3 Análise léxica

A análise léxica é uma operação básica num sistema de recuperação de informação (Fox, 1992). Este processo possibilita a identificação de sequências de caracteres válidos, *tokens*, sendo removidos todos aqueles que apresentam pouco valor para a descrição do documento ou interrogação incluídos na lista de paragem. Normalmente, associada a cada

colecção de documentos de teste, é disponibilizada uma lista de paragem a considerar no âmbito daquela colecção. Este foi o caso do protótipo RST1 que usou a lista de paragem disponibilizada pelas colecções de teste da CRAN e da LISA. Os casos mais comuns de *tokens* incluídos numa lista de paragem são os artigos definidos ou indefinidos; os pronomes; os determinantes, como por exemplo: “de”, “e”, “a”, “que”, “isto”. Podem ser considerados lista de paragem os termos com elevada frequência nos documentos de uma dada colecção. Caso não fossem removidos, uma pesquisa usando termos contidos na lista de paragem provavelmente poderia recuperar todos os documentos da colecção.

5.4.4 Tesouro

A definição de um tesouro no contexto do sistema RST tem como objectivo a correcção da imprecisão dos termos usados na interrogação, recorrendo à teoria dos conjuntos difusos. Esta correcção consiste na expansão dos termos para um conjunto de documentos associados, com um valor de pertença, que identifica “o quão adequado é o documento para o termo”. Existem vários tipos de tesouro, todos providenciando um vocabulário controlado e pré-definido. A diversidade dos tesouros decorre de aspectos como a interacção com o utilizador, o tipo de informação que pode ser indexada, existindo tesouros especializados numa determinada área de informação, como por exemplo, a informação biomédica e pela forma como é construído – manual ou automática. Alguns tesouros não somente indexam termos, isoladamente, mas também expressões, dado que a conjunção de termos revela mais a sua semântica. Neste aspecto, os tesouros podem ser classificados como *precoordinated* ou *postcoordinated* consoante o tesouro permita a indexação de frases ou não, respectivamente. Consequentemente, a construção de um tesouro não é algo trivial e existe uma vasta literatura sobre os princípios e problemas envolvidos na sua construção. No entanto, apenas uma pequena porção se dedica à construção automática do tesouro, sendo a maioria construída manualmente uma vez que existe algum cepticismo sobre a possibilidade dessa construção (Srinivasan, 1992). Os resultados obtidos na recuperação de informação com recurso a tesouros construídos automaticamente permanecem inconclusivos existindo divergências sobre a sua eficiência. No mesmo sentido e no âmbito do projecto FACET (*Faceted Access to Cultural hEritage*

Terminology), foram descritos alguns problemas na recuperação de informação com recurso a um tesauro, permanecendo o seu potencial ainda por confirmar (Blocks et al., 2002). Apesar destas dificuldades, foi considerada, no âmbito das teorias em estudo, como mais adequada a opção da aplicação do tesauro no sistema RST.

A construção do tesauro é efectuada automaticamente com base nos documentos da colecção, após ter sido executada a análise léxica, de modo a obter um vocabulário mais adequado. Dado se tratar de uma construção automática, o nível de coordenação seleccionado foi o de *postcoordination* dado que a construção automática de frases consegue ser efectiva apenas quando a colecção de documentos é suficientemente grande, o que não é o caso (Choueka, 1988 a partir de Srinivasan, 1992).

5.4.5 Comparação

O processo de comparação no sistema RST para efectuar a recuperação dos documentos define um conjunto rígido de documentos, que possuem um valor de satisfação igual ou superior a um valor limite estabelecido no momento da sua execução. Este processo é implementado num módulo para permitir a experimentação de diversas abordagens e valores limites. A comparação entre a interrogação e o documento é baseada na estrutura de classificação difusa de documentos definida no processo de construção do tesauro e pela lista de termos da colecção criada na análise léxica.

5.4.6 Agrupamento em clusters

O uso do agrupamento em clusters no âmbito do sistema RST é baseado na hipótese de que “documentos relevantes muito próximos têm a tendência de serem relevantes para o mesmo pedido” (van Rijsbergen, 1979), mas com um contraste com a estratégia dos testes inicialmente desenvolvidos por Salton (1971). Contrariamente aos testes iniciais, em que todos os documentos da colecção eram agrupados em clusters para permitir uma redução do processamento na comparação uma vez que esta passava a ocorrer entre interrogação e representante do cluster (centróide), no âmbito do sistema RST1, o agrupamento em clusters tem como objectivo agrupar em clusters os documentos recuperados pelo processo de comparação, promovendo uma mais fácil pesquisa dos

documentos relevantes pelo utilizador. Neste sentido, o agrupamento em clusters no âmbito do sistema RST tem como objectivo efectuar a separação dos documentos relevantes dos documentos não relevantes.

O agrupamento em clusters é uma técnica usada para análise multi-diferenciada que possui aplicações em áreas muito diversas (Rasmussen, 1992). Esta, não deve ser referenciada como uma técnica de classificação automática dado não se saber à partida as classes ou grupos gerados. Os documentos dentro de um grupo ou classe, devem possuir um elevado grau de associação entre si e um baixo grau de associação com outros documentos de outros grupos ou classes, pelo que se pode afirmar que o agrupamento em clusters tem como objectivo agrupar objectos que são similares numa determinada perspectiva. Esta técnica, estudada intensivamente na década de 70, teve um novo ressurgimento em anos recentes devido à explosão de documentos na Web, sendo muito popular em áreas como a biologia (Fasulo, 1999). Tal facto não é de estranhar uma vez que com esta explosão de documentos é urgente introduzir uma estrutura, uma classificação dos documentos. Consequentemente, é possível afirmar que consciente ou inconscientemente o conceito base do agrupamento em clusters é usado desde há centenas de anos, tendo o processo sofrido uma redefinição graças às novas metodologias desenvolvidas na sequência da evolução tecnológica ocorrida nas últimas décadas (Willet, 1988).

No campo da recuperação de informação, o agrupamento em clusters foi inicialmente desenvolvido com o objectivo de melhorar a eficiência na recuperação e tem sido aplicada a nível de documentos e termos. Na área da recuperação de informação, esta técnica tem sido usada em domínios como a reformulação da interrogação (*relevance-feedback*), visualização gráfica de resultados, ordenação de resultados (*ranking*), reformulação de interrogação (*query expansion*), etc., existindo técnicas e diversas propostas de algoritmos para a sua implementação.

6 RST1 – Protótipo de recuperação de informação

Este capítulo apresenta a implementação do protótipo RST1 cuja arquitectura foi apresentada no capítulo anterior. Este protótipo usa uma combinação das duas teorias em estudo: teoria dos conjuntos difusos e teoria dos conjuntos aproximativos. Esta combinação é introduzida da forma mais natural no âmbito da recuperação de informação através da definição de uma componente difusa que estabelece associação entre documentos e termos e pela definição de agrupamentos em clusters dos resultados obtidos após comparação entre interrogação e documentos da colecção. Descrevem-se as estruturas usadas para a implementação do protótipo, os detalhes do ambiente de implementação e as colecções de teste usadas.

6.1 Análise léxica

O desenho de um analisador léxico levanta um conjunto de questões técnicas importantes. Em primeiro, o que considerar como uma sequência válida de caracteres? Sendo identificadas estas sequências, deverão ser usadas como índices para indexar documentos? Estas questões são pertinentes e podem ter impacto significativo na qualidade da recuperação da informação. Tendo em mente estas questões, foi considerado como mais apropriado usar como base para o analisador léxico o autómato DFA desenvolvido por Fox (1992) para o reconhecimento de *tokens*, dado o mesmo ser implementado com um algoritmo eficaz, simples e previamente testado. Nesta implementação, qualquer sequência alfanumérica de caracteres, não iniciada por um dígito é considerada como um *token* válido. Todos os caracteres de espaço, pontuação e controlo são considerados delimitadores de *tokens*, sendo definidas dez classes diferentes de caracteres (letras, dígitos, espaço, fim de *string* – *fds*,...). Todos os caracteres são convertidos de maiúsculas para minúsculas. Esta definição foi considerada adequada para as colecções de teste usadas. No entanto, para colecções associadas a outros âmbitos como sejam o do direito judicial, por exemplo, sequências de caracteres iniciadas por dígitos e contendo caracteres

específicos como sejam o hífen deveriam ser eventualmente considerados como *tokens*. Na Figura 4 é apresentado o diagrama de transição do autómato finito discreto.

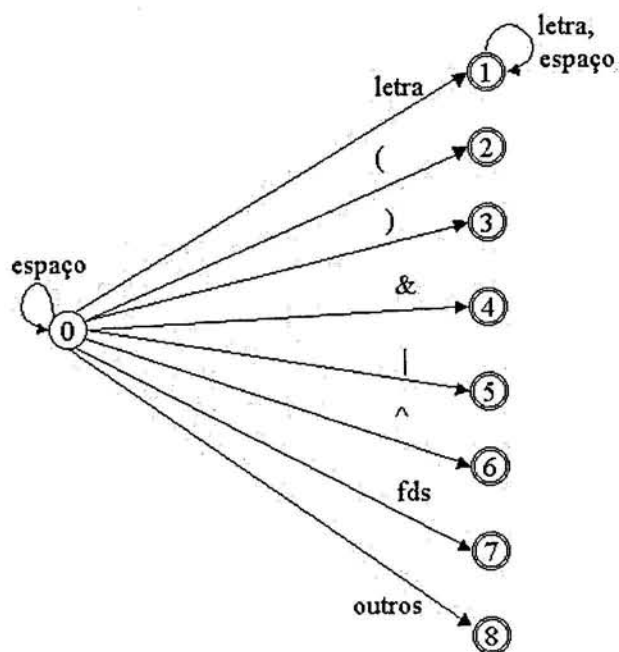


Figura 4 - Autómato do tipo DFA para identificação de *tokens*²

Como referenciado anteriormente, foram usadas duas colecções distintas, cujos documentos apresentam estruturas diferentes. Tal situação teve como efeito o desenvolvimento de funções de entrada de dados adaptados a esta realidade. A análise léxica dos documentos e das interrogações usou a mesma lista de paragem. A análise léxica identifica sequências de caracteres válidos eliminando aquelas que pertencem à lista de paragem. Esta eliminação das *stopwords* também tem benefícios óbvios no espaço necessário para a definição de estruturas de indexação, conseguindo-se, em alguns casos, taxas de compressão superiores a 40% (Baeza-Yates, Ribeiro-Neto, 1999).

² Adaptado do diagrama construído por Fox (1992)

No caso das colecções em análise e usando a lista de paragem disponibilizada por cada colecção, a CRAN foi a colecção que apresentou uma taxa de redução de termos e estruturas de indexação mais significativa, tal como indicado na Tabela 4.

	CRAN		
	Ficheiro Invertido (Mb)	Ficheiro Directo (Mb)	Nº. Termos Colecção
Sem eliminação stopwords	2.074	1.817	8372
Com eliminação stopwords	1.479	1.310	8019
Taxa de redução	29.7%	27.9%	4.21%

Tabela 4 – Taxas de compressão pela eliminação de *stopwords*

O analisador léxico implementado no RST1, como se pode ver na Figura 3, está acoplado a um módulo que efectua a construção de estruturas de dados, criando representações internas dos documentos e das interrogações. Estas estruturas são usadas em fases posteriores do processo de recuperação, nomeadamente na criação do tesauro e na comparação, promovendo uma maior eficiência. Por questões de optimização de acessos, uma vez que é necessário efectuar uma leitura a todos os documentos, foi construído um ficheiro invertido (analisado posteriormente em detalhe) e um ficheiro de acesso directo ao mesmo tempo.

Um outro processo de análise léxica usualmente adoptado é a radicalização, que encontra o radical dos termos, como seja “escol” para os termos “escolaridade”, “escolares”, “escolas”. Na versão final do RST1 não foi implementado nenhum módulo de radicalização dado ter sido considerado desadequado no âmbito de uma recuperação difusa, onde o valor semântico dos termos é importante. O algoritmo de radicalização inicialmente usado foi implementado por Stuart J. Barr e baseia-se no algoritmo de Porter.

Este encontra-se disponível nas páginas dos recursos de recuperação de informação da Universidade de Glasgow em http://www.dcs.gla.ac.uk/ir_resources/linguistic_utils/porter.c

6.1.1 Ficheiro invertido

Um dos produtos resultantes do processo da análise léxica é o ficheiro invertido da colecção de documentos. Este ficheiro é uma das três estruturas de ficheiros mais comuns na área da recuperação de informação (Harman et al., 1992), sendo agrupados na categoria dos índices lexicográficos. Para além desta, as estruturas do tipo cluster, analisados em detalhe posteriormente, e os índices baseados em algoritmos de dispersão (algoritmos de compressão e indexação de dados) são também usados na organização de ficheiros para a recuperação de informação.

Os ficheiros invertidos possuem uma estrutura física indicada na Tabela 5 e permitem o eficiente acesso aos documentos associados a um determinado termo, indicando o número de ocorrências do termo num dado documento.

Termo	Nº documento	Frequência
zone	14	2
zone	126	2
zone	167	1
...	...	...
young	145	1
...	...	...
work	415	2
work	540	1
work	540	1
...	...	...

Tabela 5 – Estrutura do ficheiro invertido

No caso do RST1, este ficheiro invertido é criado internamente recorrendo a uma árvore binária, em que cada nó está associado a uma lista de apontadores de documentos (*linked*

list), como indica a Figura 5. Cada nó da árvore binária possui informação sobre o termo e apontadores para os termos à sua direita e esquerda que permitem ligá-los aos elementos seguintes na ordem lexicográfica. Um elemento da lista possui informação sobre o documento, frequência do termo no documento e apontador para o documento seguinte onde o termo está presente. Sempre que um termo é processado é criado um novo nó na árvore ou actualizada informação na lista associada ao nó, caso este já esteja presente. A actualização na lista pode traduzir-se na actualização simples do campo de frequência associado a um dado documento ou pela criação de uma nova entrada caso este termo esteja associado a um documento ainda não presente na lista. Esta técnica é uma variante da proposta por Harman e Candela (1990) e permite criar dinamicamente o ficheiro invertido, conseguindo obter a informação numa só passagem e de forma eficiente, sem recurso a ordenações intermediárias.

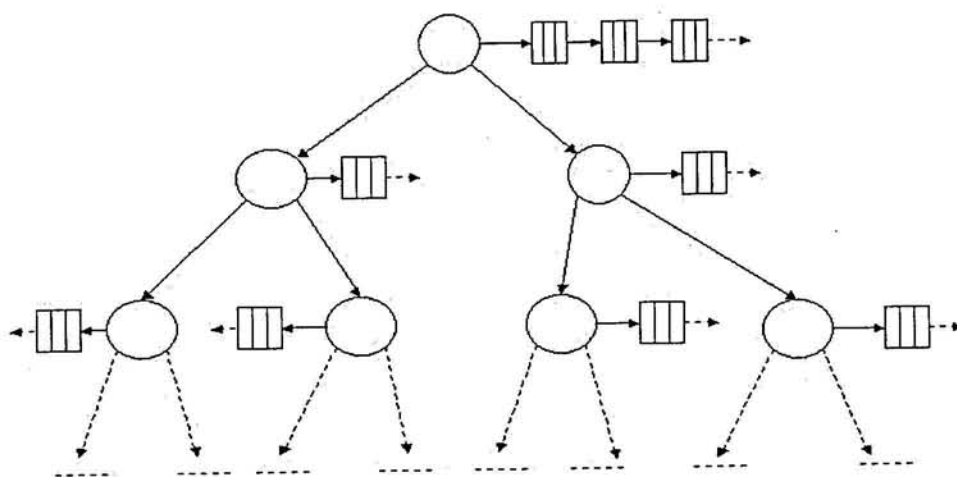


Figura 5 – Estrutura interna em memória de um ficheiro invertido

6.1.2 Ficheiro directo de termos e índice da colecção

O ficheiro directo de termos, *direct term file*, é uma estrutura que em termos de organização apresenta informação sobre os documentos, termos e frequência deste no documento, em que a chave de pesquisa é o documento e não o termo, como apresentado na Tabela 6. Este tipo de ficheiro é útil no sentido de identificar eficientemente todos os termos associados a um dado documento. Esta estrutura é usada no processo do tesauro.

Nº documento	Termo	Frequência
1	zone	2
1	young	2
1	work	1
...	...	...
145	airplane	1
145	mountain	2
145	work	2
...	...	...
540	work	1
540	manufactury	6
...	...	...

Tabela 6 – Estrutura do ficheiro directo de termos

O ficheiro de índice da colecção identifica todos os termos que são usados para indexação de documentos. Este ficheiro é usado no processo do tesauro e na comparação entre interrogação e documentos de modo a estabelecer se os termos usados na interrogação são ou não usados na indexação de documentos da colecção. Caso não o sejam, estes são ignorados no processo de comparação.

6.2 Construção do tesauro

A fase de geração do tesauro é a mais complexa em termos computacionais, dado que é necessário gerar informação sobre todos os documentos da colecção que partilham pelo menos um termo. A base para a sua construção são os valores de co-ocorrência entre os termos de indexação, k_i , k_l . Estes valores são usados para definir uma matriz de co-ocorrência que é usada para estabelecer os valores de pertença de um documento relativamente a um termo do índice. As fórmulas usadas para o seu cálculo são as sugeridas por Ribeiro-Neto e Baeza-Yates (1999). Para a geração de valores de co-ocorrência no âmbito da geração automática é necessário definir uma medida estatística que consiga traduzir a importância de cada termo na descrição do documento. A medida, normalizada, adoptada é:

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

onde n_i é o número de documentos que contém o termo k_i , n_l é o número de documentos que contém o termo k_l e $n_{i,l}$ o número de documentos que contém ambos os termos de indexação. A estrutura do tesauro é definida em memória e usada para estabelecer uma classificação difusa de documentos. Os produtos finais do processo de construção do tesauro são uma matriz de similaridade, usada no processo de agrupamento em clusters e um ficheiro de classificação difusa de documentos usado no processo de comparação.

O cálculo do valor de pertença difuso associado a cada documento d_j , foi calculado como um produto iterado sobre os termos do documento d_j , como proposta por Ribeiro-Neto e Baeza-Yates (1999):

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l})$$

Este produto iterado promove uma transição mais suave entre os diversos valores de pertença. Com base nesta informação, podemos concluir que um documento d_j pertence ao conjunto difuso associado ao termo de indexação k_i , se os próprios termos do documento estão associados ao índice k_i . Sempre que exista, pelo menos, um termo de indexação k_j de d_j que está fortemente associado ao termo de indexação k_i , então $\mu_{i,j} \sim 1$ e o índice k_i é um bom índice difuso para o documento d_j . Na situação em que todos os termos de indexação de d_j tenham um grau de associação baixo relativamente ao índice k_i , então o índice k_i não é um bom índice difuso para o documento d_j e $\mu_{i,j} \sim 0$ (Ribeiro-Neto, Baeza-Yates, 1999). Um valor mínimo (limiar) foi definido de modo a estabelecer o valor a partir do qual um índice de indexação pode ser considerado como bom. Neste caso, usa-se um α -cut do conjunto difuso, uma vez que para um documento d_j , obtemos os índices k_j com um limiar maior do que α .

A definição em memória do tesauro foi orientada pelos mesmos princípios usados na análise léxica. É usada uma árvore binária em que cada nó representa um termo de indexação ao qual está associado uma lista com informação sobre documentos que contém o termo, como indicado na Figura 5.

6.2.1 Matriz de similaridade e classificação difusa de documentos

Dois dos resultados da construção do tesauro são usados no processo de recuperação.

A matriz de similaridade usada no agrupamento em clusters é construída nesta fase de forma mais eficiente uma vez que sendo necessário efectuar a construção de dados estatísticos sobre os documentos, nomeadamente a co-ocorrência de termos, para a construção do tesauro é usada a mesma informação base. A matriz de similaridade pode ser calculada de diversas formas, existindo um conjunto de medidas ou coeficientes diversos usualmente usados na área da recuperação de informação. Assim, os coeficientes ou medidas mais usados são o Dice, Jaccard, Coseno, dado serem simples e normalizados

(Rasmussen, 1992), isto é, terem em consideração o número de termos que ocorrem no documento. O coeficiente de Dice é obtido da seguinte forma:

$$S_{D_i, D_j} = \frac{2 \sum_{k=1}^L (w_{ik} w_{jk})}{\sum_{k=1}^L w_{ik}^2 + \sum_{k=1}^L w_{jk}^2}$$

onde w_{ik} é o peso atribuído ao termo k no documento D_i . No contexto do sistema RST1 foi adoptada a versão binária do mesmo, uma vez que não está de todo aceite que o uso de coeficientes com pesos possa induzir uma melhoria (antes pelo contrário), como argumentam Sneath e Sokal (1973, a partir de Rasmussen, 1992):

$$S_{D_i, D_j} = \frac{2C}{A + B}$$

onde C é o número de termos que D_i e D_j têm em comum, A e B são os números de termos nos documentos D_i e D_j , respectivamente. Neste caso, uma vez que a medida é simétrica, isto é $S_{i,j} = S_{j,i}$, a matriz de similaridade associada também é simétrica, pelo que apenas será necessário calcular parte dos valores, como indica a Figura 6.

$$S = \begin{bmatrix} 1 & & & & \\ S_{21} & 1 & & & \\ S_{31} & S_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ S_{N1} & S_{N2} & \cdots & S_{N(N-1)} & 1 \end{bmatrix}$$

Figura 6 – Matriz de similaridade simétrica

A estrutura difusa de classificação de documentos é constituída por uma representação interna do índice que associa todos os documentos e ao respectivo valor difuso de pertença, como indicado na Tabela 7.

Termo (índice interno- k_i)	Nº documento (d_j)	Valor Difuso ($\mu_{i,j}$)
1	10	1
1	23	0.9
2	35	1
...	...	...
40	1	0.43
40	2	0.65
40	4	0.77
...	...	...
100	36	0.8
100	44	0.6
..	...	...

Tabela 7 – Estrutura do ficheiro da classificação difusa de documentos

6.3 Comparação

O processo de comparação é o segundo mais oneroso em termos dos recursos necessários para a execução no sistema RST. A estrutura de classificação difusa de documentos armazenada num ficheiro sequencial é nesta fase acedida sendo criada uma cópia em memória para promover um acesso mais eficiente. Os termos presentes na interrogação são convertidos numa representação interna desde que sejam termos de indexação. Esta conversão tem como motivação uma melhoria da pesquisa nas estruturas, dado que, deste modo, serão efectuadas comparações directas entre números, que são substancialmente mais eficazes do que comparações entre *strings* (cadeias de caracteres). A extracção dos documentos é efectuada por comparação entre os termos que constituem a interrogação e

os termos de indexação dos documentos, numa primeira fase. É óbvio que nenhuma informação sobre termos introduzidos na interrogação que não estejam definidos na colecção de documentos foi gerada pelo tesouro. Após esta validação é iniciada uma pesquisa. A estratégia adoptada é a de uma pesquisa extensiva a todos os termos do índice para obter os documentos que partilham, pelo menos, um termo com a interrogação. Para cada termo de indexação i , é criado um conjunto difuso de documentos sendo que cada documento possui um valor de pertença igual ou superior a α (limiar). Para cada documento j seleccionado é calculado um valor difuso de pertença de modo a estabelecer o seu grau de satisfação relativamente à interrogação q . A fórmula base usada para o seu cálculo foi a proposta por Ribeiro-Neto e Baeza-Yates (1999):

$$\mu_{q,j} = 1 - \prod_{i=1}^N (1 - \mu_{i,j})$$

onde N é o número de termos da interrogação. A resposta final do sistema RST1 neste processo é definido como um α -cut de um conjunto difuso, onde α foi previamente definido. Podemos descrever genericamente o algoritmo do processo de comparação como sendo:

1- *Selecciona interrogação Q*

Para cada termo q_i em Q

Verifica se termo q_i é termo de indexação da colecção

Para cada q_i selecciona informação sobre documentos que possuem o termo q_i na estrutura difusa de classificação F , criando conjunto H

2- *Para cada documento em H*

Calcula o valor difuso de pertença relativamente à interrogação Q

3- *Se valor difuso de pertença do documento é maior ou igual a α (limiar)*

Insere documento no conjunto final de resposta do processo

A Figura 7 apresenta, esquematicamente, o processo de comparação entre interrogação e documentos da colecção.

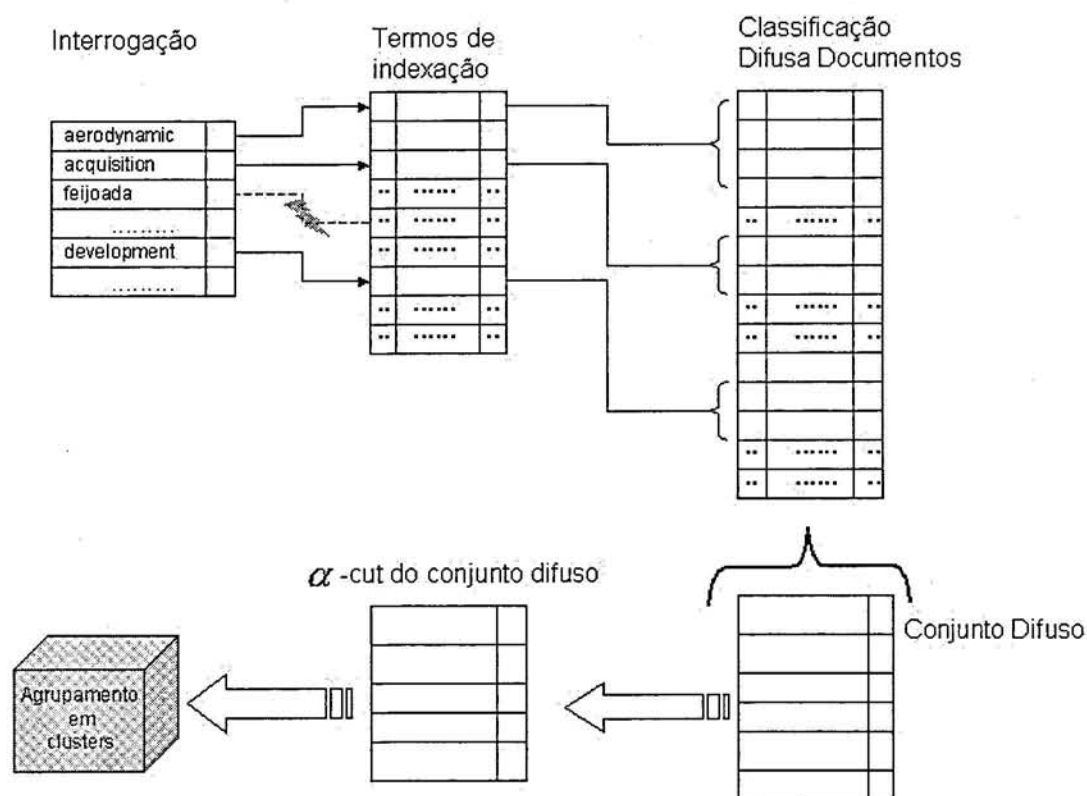


Figura 7 – Processo de comparação entre documento e interrogação

Neste exemplo, o termo feijoada, introduzido na interrogação, não está presente nos termos de indexação da colecção de documentos.

6.4 Agrupamento em clusters

O agrupamento em clusters tem gerado um entusiasmo assinalável na área da recuperação como um método que apresenta o potencial para ultrapassar os problemas de grandes volumes de documentos na resposta. No entanto, apenas um número surpreendentemente reduzido de sistemas de recuperação têm aplicado estas técnicas para a organização dos resultados, não existindo, virtualmente, nenhuma literatura onde sejam

efectuadas análises detalhadas do seu uso (Leuski e Croft, 1996). Tal situação pode dever-se ao facto que, embora exista uma base teórica forte baseada em relações estatísticas, o agrupamento em clusters ainda não conseguiu demonstrar plena capacidade em melhorar a recuperação de informação em ambientes efectivos de recuperação (Yu et al., 1983; Smeaton e van Rijsbergen, 1983 a partir de Willet, 1988). Apesar destes factos, a hipótese cluster foi estudada por vários investigadores, tendo sido dada como provada por Croft (1978), Hearst e Pederson (1996), Leuski e Allan (2000) usando colecções da TREC e combinando os clusters com as listas ordenadas que indicaram uma clara melhoria de eficiência face ao método isolado e tradicional da lista ordenada.

Mais recentemente, Tombros (2002), confirma que a investigação desenvolvida até ao momento foi inconclusiva na demonstração da capacidade desta técnica produzir ou não uma melhoria na recuperação de informação defendendo, no entanto, na sua tese essa mesma demonstração. A investigação de Tombros (2002) consiste na incorporação da informação da interrogação para o agrupamento em clusters dos documentos por dois métodos: o *post-retrieval clustering*, ou seja, o agrupamento em clusters dos documentos recuperados por um sistema de recuperação e o agrupamento em clusters dos documentos com recurso a medidas de similaridade definidas com base na interrogação.

Assim, tendo plena consciência das possíveis dificuldades operacionais, a motivação para o uso do agrupamento em clusters no sistema RST1 é permitir uma análise, uma visualização mais fácil dos resultados, induzindo o utilizador a melhorar, eventualmente, a sua interrogação ao sistema RST1 pela descoberta de novas relações entre os documentos existentes nos diversos clusters. Esta abordagem, no domínio do *post-retrieval clustering*, agrupamento em clusters após recuperação de documentos, é particularmente interessante nesta aplicação dado que o número de documentos recuperados pela abordagem difusa é demasiado elevado para ser analisado pelo utilizador.

6.4.1 Métodos de agrupamento em clusters

Dos métodos de agrupamento em clusters existentes, destacam-se dois tipos na área da recuperação de informação: não-hierárquico e hierárquico. O método não-hierárquico estabelece um conjunto pré-definido de condições sobre o número de clusters, o seu

tamanho, os critérios de pertença, sendo implementado por um algoritmo recursivo que tenta divisões, partições sucessivas de N objectos em M clusters. Deste modo, o método não-hierárquico é heurístico por natureza, não existindo uma hierarquia entre os diversos clusters e sendo os resultados não determinísticos dado que estes dependem da ordem com que os documentos são analisados (Willet, 1988). Estes métodos apresentam vantagens computacionais dado que os seus requisitos são baixos, usualmente da ordem de $O(n)$ até $O(N \log N)$ (Willet, 1988), tendo sido fundamentalmente usados no desenvolvimento inicial do agrupamento em clusters de documentos (Rasmussen, 1992). Estes métodos foram considerados como inadequados para a recuperação de informação, dado resultarem numa diminuição da eficiência comparativamente com pesquisas efectuadas directamente sobre o ficheiro não cluster (Salton, 1971, Fritsche, 1974, a partir de Willet, 1988).

O método hierárquico, computacionalmente mais complexo do que o anterior, pode ser aglomerativo (*agglomerative*) ou divisivo (*divisive*), dependendo da construção dos clusters ser iniciada com um conjunto de documentos não agrupados em cluster, caso em que é efectuada uma união de itens ($N-1$ para um conjunto de N documentos) em cada iteração do algoritmo, ou se é iniciado com um único cluster e progride através de divisões do mesmo para obter clusters mais pequenos. O método hierárquico, desenvolvido inicialmente no âmbito da biologia, na taxonomia, é o método mais adoptado na recuperação de informação. Os métodos hierárquicos aglomerativos são os mais usados dado permitirem que os documentos presentes no cluster tenham alguns ou muitos termos em comum não existindo a obrigatoriedade da presença de termos específicos tal como acontece nos métodos hierárquicos divisivos (van Rijsbergen, 1979 a partir de Willet, 1988). Em cada iteração a classificação hierárquica define um número de clusters que são disjuntos e tal como demonstrado por Miyamoto (1998) o agrupamento em cluster hierárquico aglomerativo está intimamente associado à teoria dos conjuntos aproximativos tomando uma relação difusa como uma relação de equivalência. Deste modo, o método cluster hierárquico aglomerativo (*Hierarchical Agglomerative Cluster Method – HACM*) providencia a base teórica necessária no âmbito da teoria dos conjuntos aproximativos.

Os algoritmos do tipo HACM mais usados são o *Single Link*, *Complete Link*, *Group Average Link* e o método de *Ward* (Rasmussen, 1992).

O algoritmo *Single Link* ou *Nearest Neighbor* efectua, em cada iteração, a união do par de objectos mais similar em que um dos objectos é um cluster. Dada a sua simplicidade, este método é um dos mais usados tendo sido um dos primeiros a ser aplicado com sucesso a uma colecção de documentos por Croft (1977). Existem vários algoritmos que implementam este método sendo o SLINK (Sibson, 1973) ou Prim-Dijkstra (Dijkstra, 1976) os mais usados para recuperação em colecções grandes tendo apresentado bons resultados (Rasmussen, 1992). No entanto, possui alguns problemas uma vez que tem a tendência de criar clusters com baixa coesão, problema designado por *chaining*.

O método *Complete Link* é o oposto do *Single Link* dado que efectua, em cada iteração, a união dos pares com menor similaridade. Este método efectua a definição de cluster pequenos, com forte coesão entre si tendo apresentado bons resultados na recuperação de informação efectuada por Voorhees (1986, a partir de Rasmussen, 1992), sendo de difícil aplicação para colecções de grande dimensão (Rasmussen, 1992). Os algoritmos mais usados na implementação do *Complete Link* são o CLINK (Defays, 1977) e o algoritmo de Voorhees (1986, a partir de Rasmussen, 1992).

O método *Group Average* efectua a união de dois clusters baseado no valor de similaridade que é o valor médio de todos os membros dos dois clusters a serem unidos. Este método apresenta problemas no seu algoritmo genérico no que diz respeito a espaço e processamento sendo de $O(N^2)$ e $O(N^3)$, respectivamente (Rasmussen, 1992), sendo um compromisso entre o *Single Link* e o *Complete Link*. O método de *Group Average* é criticado por criar cluster pequenos que são completamente diferentes dos restantes clusters (Willet, 1988). Um membro de um cluster criado por este método apresenta uma maior similaridade média para com os outros membros do cluster do que com qualquer outro membro de qualquer outro cluster.

O método de *Ward* efectua a união de dois clusters que resulta no menor incremento da distância de cada documento para com o centróide (representante do cluster) do seu cluster. Este método é aplicado correctamente quando a medida de similaridade adoptada é euclidiana, tendo sido criticado por criar clusters esféricos uma vez que estes podem não reproduzir correctamente as formas dos clusters presentes nos objectos em análise (Willet, 1988).

6.4.2 Abordagem implementada

No sistema RST1, o agrupamento em clusters é um processo dinâmico que não efectua o agrupamento em clusters estático de todos os documentos da colecção, tal como ocorria nos testes iniciais desenvolvidos na área da recuperação de informação (Willet, 1988), (Hearst e Pederson, 1996). Nesse contexto inicial, era efectuado um agrupamento em clusters de toda a colecção e calculados os centróides ou membros representativos de todos os clusters sendo estes, posteriormente, comparados com o objectivo de determinar o cluster mais similar relativamente à interrogação. Estes testes foram desenvolvidos com o intuito de evitar a análise e comparação de todos os documentos da colecção e de categorizar, classificar os documentos de uma colecção. O sistema RST1 adoptou uma abordagem diferente dos testes iniciais e um princípio algo similar a do sistema *Scatter/Gather* desenvolvido por Hearst e Pederson (1996). Assim, é admitido que dois documentos que estejam presentes no mesmo cluster são similares um para o outro e que documentos relevantes devem ter a tendência de estarem presentes no mesmo cluster enquanto que os documentos não-relevantes devem ter a tendência de estarem em outros clusters. A diferença ocorre dado que não é assumido que no caso de dois documentos d_1 e d_2 serem relevantes para uma interrogação q_i , também o sejam para a interrogação q_1 .

Os algoritmos implementados no âmbito do sistema RST1 foram o *Single Link*, *Group Average* e o *Complete Link*. O primeiro documento de cada cluster é aquele que apresenta a maior similaridade relativamente ao grupo de documentos apresentados ao utilizador. O agrupamento em clusters aplicado sobre o conjunto de resultados, *post-clustering*, apresenta vantagens sobre o agrupamento em clusters aplicado em todos os documentos da colecção. Este método apresenta exigências menores em termos dos recursos computacionais usados, uma vez que o processo se efectua em apenas parte dos documentos. A actualização dos clusters é inexistente uma vez que não sendo estática, não é necessário implementar processos de actualização dos grupos de clusters.

Esta abordagem, no âmbito do agrupamento em clusters, é ainda relativamente recente e não existe ainda muitos desenvolvimentos nesta área, tendo sido Willet (1985) um dos primeiros a estudá-la (Tombros, 2002). Deste modo, o agrupamento em clusters é

introduzido de modo a ser alcançada uma maior separação entre documentos relevantes e não relevantes, o que induz uma óbvia melhoria na interacção do utilizador com o sistema e naturalmente uma recuperação mais eficaz.

6.4.3 Medida e matriz de similaridade

Os algoritmos HACM distinguem-se pela forma como são definidos os pares mais similares ou dissimilares em cada iteração do algoritmo. Uma componente importante destes algoritmos é o cálculo da similaridade entre documentos ou clusters, sendo criada uma matriz de similaridade que contém os valores desta para todos os pares. Esta matriz é recalculada em cada iteração, após a união de dois pares, podendo ser simétrica se a medida seleccionada o for. Da diversidade de medidas existentes e descritas na literatura, Salton (1989; a partir Frakes e Baeza-Yates, 1992) analisa as mais adequadas para a comparação de documentos, existindo uma discussão sobre se devem ou não serem usadas medidas com peso. Neste domínio, Sneath e Sokal (1973, a partir de Willet, 1988) argumentam a favor da sua não utilização, tendo Willet (1988) desenvolvido testes que demonstram que pouca ou nenhuma diferença existe no resultado obtido com o uso de tais medidas. Para o cálculo da matriz de similaridade, construída na fase do processo do tesouro foi seleccionado o coeficiente de DICE na versão binária. A fórmula proposta por Lance-Williams (1967, a partir de Rasmussen, 1992), foi usada dado permitir o cálculo das dissimilaridades entre os novos clusters e os existentes. Esta usa os valores existentes anteriores à criação do novo cluster, sendo assim evitado o refazer do cálculo da matriz de similaridade. A dissimilaridade d entre o novo cluster, $C_{i,j}$, resultado da união entre os clusters C_i e C_j , e qualquer outro cluster C_k existente é dada por:

$$d_{C_{i,j}C_k} = \alpha_i d_{C_iC_k} + \alpha_j d_{C_jC_k} + \beta d_{C_iC_j} + \gamma |d_{C_iC_k} - d_{C_jC_k}|$$

onde os parâmetros β , α , γ estão dependentes dos diferentes algoritmos seleccionados para o agrupamento em clusters. Os valores para os parâmetros dos algoritmos implementados são apresentados na Tabela 8.

HACM	Parâmetros Lance-Williams
<i>Single Link</i>	$\alpha_i = \frac{1}{2}$
	$\beta = 0$
	$\gamma = -\frac{1}{2}$
<i>Complete Link</i>	$\alpha_i = \frac{1}{2}$
	$\beta = 0$
	$\gamma = \frac{1}{2}$
<i>Group Average</i>	$\alpha_i = \frac{m_i}{m_i + m_j},$ onde m_i é o número de itens em C_i
	$\beta = 0$
	$\gamma = 0$

Tabela 8 – Parâmetros de Lance-Williams

6.5 Detalhes de implementação

O software foi desenvolvido usando a linguagem C e o compilador do Visual C++. Para o armazenamento dos dados foram usados ficheiros sequenciais e nomeados de acordo com objectivo e conteúdo.

Em cada módulo do sistema foram desenvolvidas funções responsáveis pela leitura dos dados de entrada criados em processos anteriores. Durante esta operação, são criadas estruturas internas de memória para referência rápida dos dados. Estas estruturas são de

vários tipos desde a simples variável, vector ou matriz até à combinação de árvores binárias associadas a listas, criadas dinamicamente.

O sistema RST1 foi dividido em vários programas:

- lexDoc.c e lexQuery.c, efectuam a análise léxica dos documentos da colecção e das interrogações.
- tesauro.c, efectua a construção do tesauro, classificação difusa dos documentos e geração da matriz de similaridade.
- extract.c, efectua a comparação entre interrogação e classificação difusa de documentos.
- cluster.c, efectua o agrupamento em clusters dos resultados para os três métodos testados: *Single Link*, *Complete Link* e *Group Average Link*.
- quality.c, efectua a extracção de informação qualitativa sobre resultados obtidos, gerando informação sobre número de clusters gerados para determinada interrogação e a sua qualidade.

Os documentos das colecções de teste da CRAN e da LISA possuem estruturas diferentes. Esta situação teve como consequência o desenvolvimento de algumas funções específicas de formatação de dados de entrada e saída. Todo o software foi desenvolvido no âmbito da construção do sistema RST1, com excepção do módulo DFA de Christopher Fox (1992) que foi incorporado, com pequenas alterações, directamente nas funções de análise léxica e devidamente identificado. Para o desenho do algoritmo de agrupamento em clusters, foi inicialmente usado o algoritmo genérico dos HACM apresentado por Edie Rasmussen (1992) e que consiste em:

- 1 – Identificar os dois pontos mais próximos e combinar os mesmos num cluster.
- 2 – Identificar e combinar os próximos dois pontos mais similares, tratando os clusters existentes como pontos.
- 3 – Se mais de um cluster permanecer, voltar ao ponto 1.

Nos casos de testes como iremos analisar de seguida, foi estabelecido um ponto de *cut-off*, corte da árvore binária da classificação hierárquica.

7 Testes e resultados

Este capítulo apresenta as colecções de teste usadas, os testes executados e os respectivos resultados. Os testes iniciais do protótipo RST1 indicam uma recuperação com uma precisão muito baixa devido ao elevado número de documentos recuperados. Esta precisão desencadeou uma análise detalhada às possíveis causas e a execução de inúmeros testes procurando detectar pontos fracos do modelo adoptado e melhorar os resultados obtidos. Desta acção emergiu uma evolução do protótipo RST1, RST2, que obteve resultados mais satisfatórios mas ainda muito aquém dos obtidos com outros métodos. O processo de agrupamento em clusters dos documentos recuperados obteve resultados considerados satisfatórios, não tendo sido obtido no mesmo cluster a totalidade dos documentos classificados como relevantes pelo especialista da colecção de teste para uma dada interrogação.

Os testes efectuados sobre o sistema RST1 são diversificados, sendo implementadas variantes nas diferentes fases do processo de recuperação. Deste modo, durante a fase da análise léxica é avaliado o uso do algoritmo de radicalização de Porter, o refinamento do vocabulário da colecção aumentando a lista de paragem, abrangendo termos considerados com uma alta frequência no âmbito da colecção de teste. Na fase de comparação são estabelecidos diferentes fórmulas e diversos limiares para analisar o seu impacto na qualidade da recuperação. Na fase de agrupamento em clusters são também analisados diversos valores de corte da construção dos clusters de modo a obter-se o maior número de documentos relevantes agrupados no mesmo cluster. Os clusters criados devem igualmente possuir documentos similares entre si de modo a revelar ao utilizador outros documentos de potencial interesse.

7.1 Introdução

Os testes foram efectuados com recurso a duas colecções: LISA e CRAN. Foram seleccionadas um conjunto de interrogações em cada colecção. O sistema RST1 foi

submetido a diversas parametrizações ou configurações no sentido de avaliar as possíveis melhorias que estas produzem nos resultados obtidos. Os resultados são apresentados com informação das alterações ocorridas entre os testes. Como esperado, a matriz de similaridade usada no agrupamento em clusters não é afectada pelo valor de limiar definido na construção do tesauro dado que é baseada no total dos termos da colecção de documentos. Os testes iniciais foram efectuados na colecção da LISA, sendo posteriormente usada, em exclusivo, a colecção da CRAN devido à existência de alguns estudos publicados e disponíveis em formato digital para comparação com os resultados obtidos pelo sistema RST1 e suas variantes.

7.2 Colecções de teste e definição dos testes

A arquitectura do RST1, foi testada em duas colecções de documentos de texto, a LISA (*Library and Information Science Abstracts Collection*) e a CRAN (*Cranfield Collection*).

A colecção da CRAN é uma das colecções mais antigas usadas na investigação no âmbito da recuperação de informação. É constituída por 1398 documentos e o seu domínio é a aeronáutica. A versão usada desta colecção foi seleccionada a partir das páginas de recursos de recuperação de informação da Universidade de Glasgow (http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cran). Associada a esta colecção estão 225 interrogações e uma avaliação dos documentos relevantes associados efectuada por um especialista, sendo disponibilizada uma lista de paragem.

A colecção da LISA foi também obtida a partir das páginas de recursos de recuperação em http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/lisa. Esta colecção é composta por 6005 documentos, 35 interrogações com documentos relevantes associados e disponibiliza também uma lista de paragem. Esta foi a colecção base para o desenvolvimento do sistema RST1.

A motivação para o uso destas duas colecções em particular e não as colecções da TREC, teve como factor essencial a sua dimensão. Dado estas colecções serem muito pequenas, comparativamente com as da TREC, é possível obter resultados rapidamente, o que é

indispensável no ambiente de desenvolvimento de um protótipo. O espaço de armazenamento dos dados associados e das estruturas internas criadas também não seriam compatíveis com as necessidades das colecções da TREC sendo o processamento, muito provavelmente, uma questão inultrapassável dado as limitações do *hardware* usado.

7.3 Testes iniciais

Os testes foram inicialmente efectuados na colecção da LISA, que serviu de suporte para o desenvolvimento do sistema RST1. Posteriormente, e face aos fracos resultados obtidos, sendo necessária a comparação com outros estudos, foi aplicado o sistema RST1 à colecção da CRAN. Esta colecção é muito usada na área da recuperação de informação, tendo sido possível a comparação de resultados. Deste modo, a colecção da CRAN tornou-se a colecção de teste de referência do sistema RST1. Os resultados apresentados a seguir referem-se, assim, exclusivamente à colecção da CRAN. Esta colecção inclui 225 interrogações sendo as 20 primeiras interrogações da colecção definidas como teste para este estudo. A CRAN apresenta um ficheiro com a classificação dos documentos que devem ser considerados relevantes para cada interrogação, numa escala de 1 a 5, 1 indicando um documento que constitui uma resposta completa à interrogação e 5 um documento sem nenhum interesse para a interrogação. Como seria de esperar, neste ficheiro não estão definidos explicitamente os documentos classificados com nível 5 para cada interrogação. Com base nesta classificação, definiram-se no âmbito destes testes como documentos relevantes todos os documentos classificados com nível 1 ao nível 4, não se efectuando nenhuma distinção na resposta ao utilizador, nesta fase do protótipo. A colecção da CRAN apresenta, em média, 8.3 documentos relevantes por interrogação. A Figura 8 apresenta o número de documentos relevantes associado a cada uma das interrogações definida pelo especialista da colecção da CRAN.

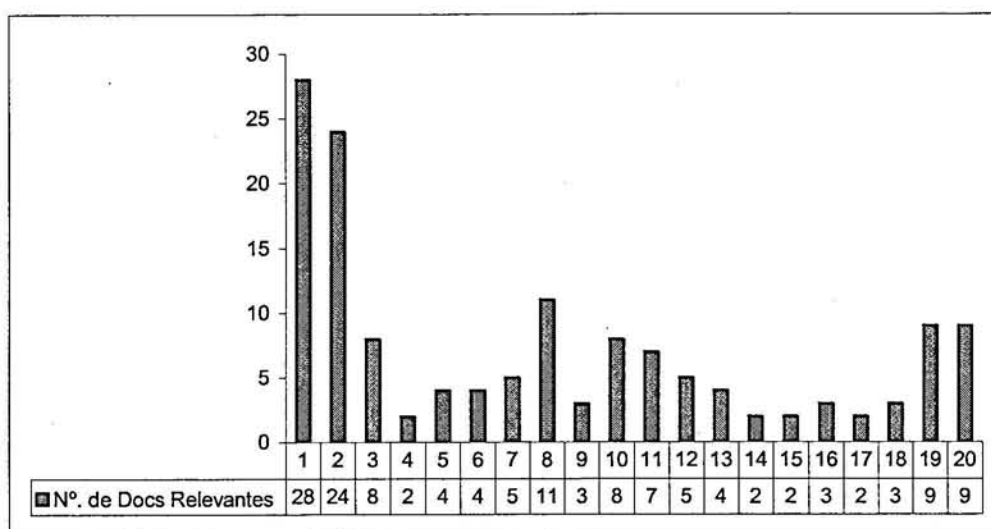


Figura 8 – Números de documentos relevantes para as 20 primeiras interrogações

Os testes com a configuração inicial do sistema RST1 foram efectuados para diferentes valores de limiar no processo de recuperação dos documentos. O valor inicial de limiar foi de 0.6 para a comparação entre os documentos e a interrogação, tendo sido usada uma classificação difusa obtida a partir da construção do tesauro com um grau de inclusão de 0.6. Estes resultados são apresentados, comparativamente com outros testes, na Figura 9 que indica o número de documentos recuperados no conjunto difuso, por interrogação e para diferentes classificações difusas de documentos obtidas a partir da construção de tesauros com diferentes valores de inclusão. O primeiro valor indicado na legenda da Figura 9 corresponde ao valor do limiar com que a recuperação de informação foi executada, enquanto que o segundo indica o valor com que a classificação difusa dos documentos foi obtida (*T grau de pertença*). Como exemplo, o segundo caso de teste apresentado como 0.60/T 0.95 indica que a comparação foi executada para recuperar todos os documentos com valor de satisfação da interrogação de pelo menos 0.60 tendo sido usada uma classificação difusa de documentos definida com um limiar de 0.95.

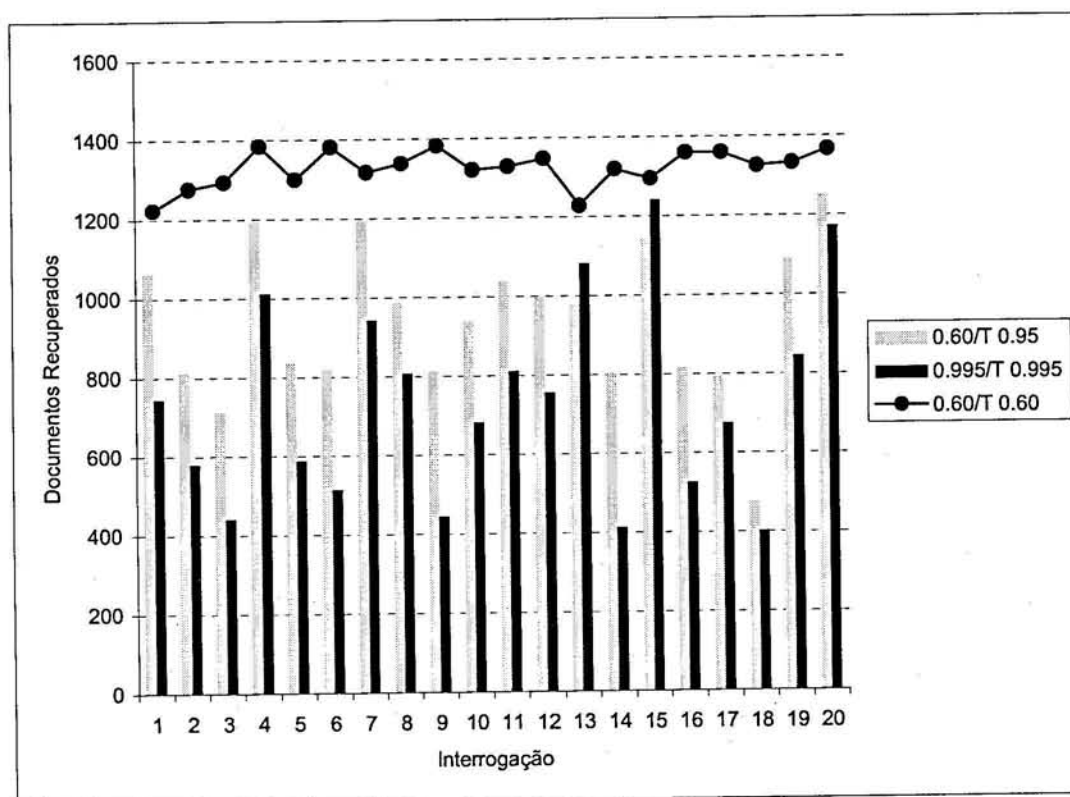


Figura 9 – Documentos recuperados por interrogação com diferentes classificações difusas

O número de documentos recuperados diminui, em média, à medida que o valor de limiar é aumentado, o que seria de esperar. No entanto, o número de documentos recuperados é demasiado elevado, apesar das sucessivas alterações efectuadas. O valor médio obtido pelo sistema RST1 na precisão da recuperação, 0.017, é substancialmente inferior à obtida por outros modelos de recuperação. A precisão obtida por um modelo de recuperação de informação baseado numa extensão do modelo vectorial e descrito por Voorhees (1999) é substancialmente superior à obtida pelo RST1 (precisão entre 0.25 e 0.35). Este modelo implementa uma indexação conceptual recorrendo à informação contida num sistema designado por WordNet. O WordNet é um sistema léxico, construído manualmente na Universidade de Princeton e que contém informação conceptual sobre termos com o objectivo de induzir uma melhoria na recuperação de informação. Nesta abordagem foram efectuados diversos testes e comparada a precisão obtida com os valores

de referência obtidos pela sistema SMART quando aplicado sobre a colecção da CRAN. Nos testes descritos por Voorhees (1993), a precisão foi sempre inferior à obtida no sistema SMART. Estes resultados foram atribuídos à dificuldade de efectuar uma análise conceptual sobre os termos das interrogações da CRAN que foram consideradas de dimensão reduzida. No caso do sistema RST1, a precisão média obtida é muito baixa: 0.017, enquanto a recuperação média é mais elevada, 0.90, facto que deriva da recuperação da quase totalidade dos documentos da colecção. A Tabela 9 apresenta os valores da precisão média obtida por diferentes métodos de recuperação de informação aplicados à colecção da CRAN e calculada sobre todas as interrogações da colecção. Estes valores foram extraídos dos trabalhos de Hust et al. (2002) que desenvolveram novos métodos para a recuperação de informação direccionados para a Web. Os métodos são: modelo vectorial - *Vector Space Model* (VSM), *Pseudo Relevance Feedback* (PRF) que usa uma conjunção da similaridade entre documentos e interrogação, ordenação dos documentos relevantes e dos seus termos para expansão da interrogação e nova ordenação dos documentos obtidos após a expansão, *Query Similarities and Relevant Documents* (QSD) que usa *relevance feedback* e informação da execução das interrogações anteriores para comparar a similaridade entre a nova e anteriores interrogações para determinar possíveis documentos relevantes (documento anteriormente classificado como relevante para uma dada interrogação tem uma alta probabilidade de ser considerado relevante para uma interrogação similar), QSDPRF e PRFQSD que são duas combinações dos dois métodos anteriores.

Métodos	Precisão Média CRAN
VSM	0.384
PRF	0.485
QSD	0.428
QSDPRF	0.451
PRFQSD	0.463
RST1	0.017

Tabela 9 – Precisão média obtida por diferentes métodos

Os resultados obtidos pelo sistema RST1 são muito inferiores aos resultados obtidos por qualquer um dos outros métodos. Este desempenho do RST1 deve estar associado a uma ou mais das hipóteses colocadas em seguida:

- A análise léxica cria um vocabulário com termos pouco restritivos.
- Os valores de co-ocorrência definidos induzem uma classificação difusa incorrecta.
- A classificação difusa de documentos é efectuada com fórmula excessivamente disjuntiva provocando a recuperação de um número muito elevado de documentos.

A introdução destas hipóteses teve como consequência o desenvolvimento de um conjunto de testes específicos para análise dos diferentes componentes do sistema RST1.

7.4 Análise dos componentes do sistema RST1

A definição dos testes iniciais no sistema RST1 teve como objectivo primordial comparar a precisão obtida pelo sistema com os valores de precisão média obtida por outros sistemas com recurso à colecção da CRAN. Tendo em consideração que estes foram substancialmente inferiores, executaram-se uma multiplicidade de testes para analisar a variação e importância de cada componente no comportamento do RST1.

Análise léxica

O estudo do impacto da análise léxica no processo de recuperação de informação do sistema RST1 inicia-se com a obtenção dos gráficos de distribuição dos termos definidos na colecção, apresentado na Figura 10.

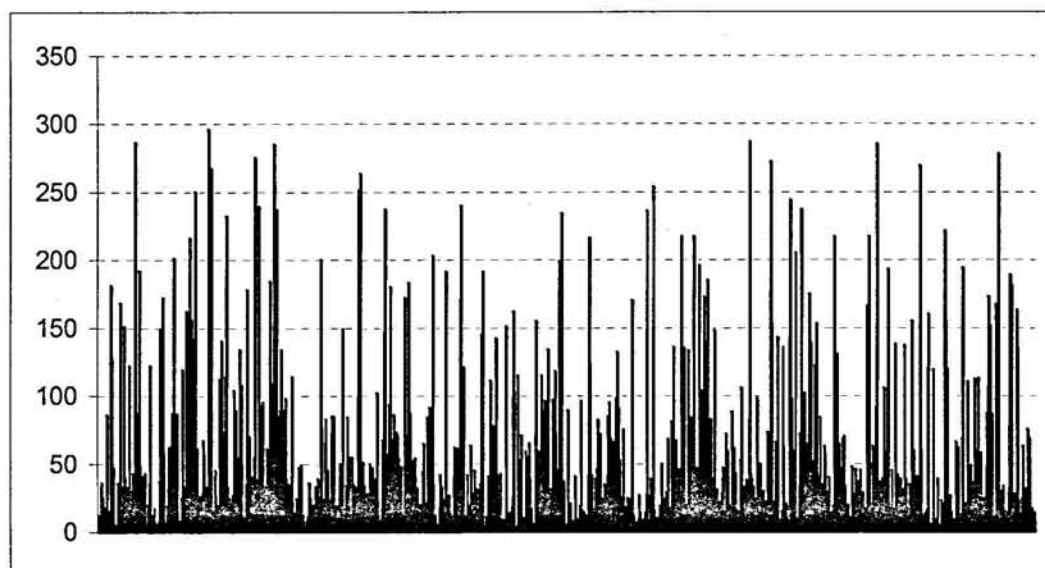


Figura 10 - Frequência dos termos da colecção CRAN

Considerando a dimensão da colecção e o tópico altamente especializado desta, os termos com uma alta frequência devem ser relativamente comuns e, consequentemente, a definição da lista de paragem deve ter esse facto em consideração. No entanto, após análise da mesma, esta continha os termos mais comuns da língua inglesa, identificados normalmente em qualquer obra de referência da área da recuperação de informação. A lista de paragem foi inicialmente definida com os termos com frequência superior a 200 na colecção, sendo este valor progressivamente reduzido até ao valor de 100. Esta alteração pode produzir uma dificuldade adicional na recuperação de informação através da remoção de termos da interrogação uma vez que a lista de paragem também é aplicada nas interrogações. A remoção de termos pode alterar a semântica da interrogação e impossibilitar a recuperação de determinados documentos relevantes.

Os testes iniciais de alteração do vocabulário foram executados com valor de 200 e desde logo indicaram uma melhoria significativa na redução dos documentos recuperados, atingindo mais de 50% para algumas interrogações (ver Figura 12), permanecendo o seu número ainda muito elevado. A alteração efectuada na definição dos termos classificados

na lista de paragem produziu uma melhoria clara, recuperando um número substancialmente menor de documentos. Esta alteração não induziu uma menor recuperação de documentos relevantes dado o número ainda elevado de documentos recuperados. Estes resultados levantam a questão se o algoritmo de geração de radicais de Porter pode introduzir para a análise léxica alguma melhoria. Consequentemente, foram desenvolvidos testes e obtido o vocabulário da colecção da CRAN segundo esta configuração do sistema. Recorrendo à metodologia anterior foi analisado o vocabulário obtido de radicais. A Figura 11 indica a frequência dos radicais.

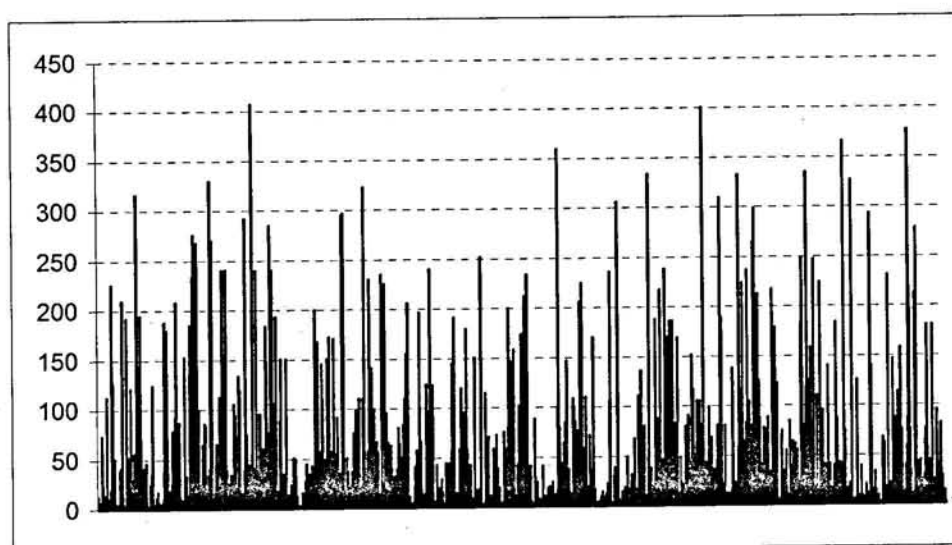


Figura 11 - Frequência dos termos da colecção com radicalização

A implementação do algoritmo de Porter criou alguns radicais com uma taxa de frequência considerada elevada. Os radicais com uma frequência superior a 200 foram inicialmente considerados lista de paragem sendo removidos do vocabulário da colecção da CRAN. Efectuada esta redefinição, foram executados 10 testes e comparados com outras abordagens ou variantes do sistema RST1 com o intuito de avaliar o impacto em termos de performance global do sistema, com excepção das vantagens óbvias de redução de espaço das estruturas de indexação e da consequente eficiência induzida no tempo de resposta do sistema ao utilizador.

A Figura 12 apresenta 6 testes, em que todos usaram o mesmo valor de recuperação e mesma classificação difusa de documentos:

- 0.995/ T 0.995, recuperação sem alteração da lista de paragem.
- 0.995/ T 0.995 Voc., recuperação usando lista de paragem alterada com introdução dos termos com frequência superior a 200.
- 0.995/ T 0.995 St. Voc. 1k., recuperação recorrendo ao algoritmo de Porter na análise léxica com alteração da lista de radicais redundantes com introdução dos radicais com frequência superior a 100.
- 0.995/ T 0.995 St., recuperação recorrendo ao algoritmo de Porter na análise léxica sem alteração da lista de radicais redundantes.
- 0.995/ T 0.995 St. Voc., recuperação recorrendo ao algoritmo de Porter na análise léxica com alteração da lista de radicais redundantes com introdução dos radicais com frequência superior a 200.
- 0.995/ T 0.995 Voc 1k., recuperação usando lista de paragem alterada com introdução dos termos com frequência superior a 100.

Os valores usados para os testes iniciais de recuperação são muito elevados porque as funções usadas geram valores muito próximas de 1.

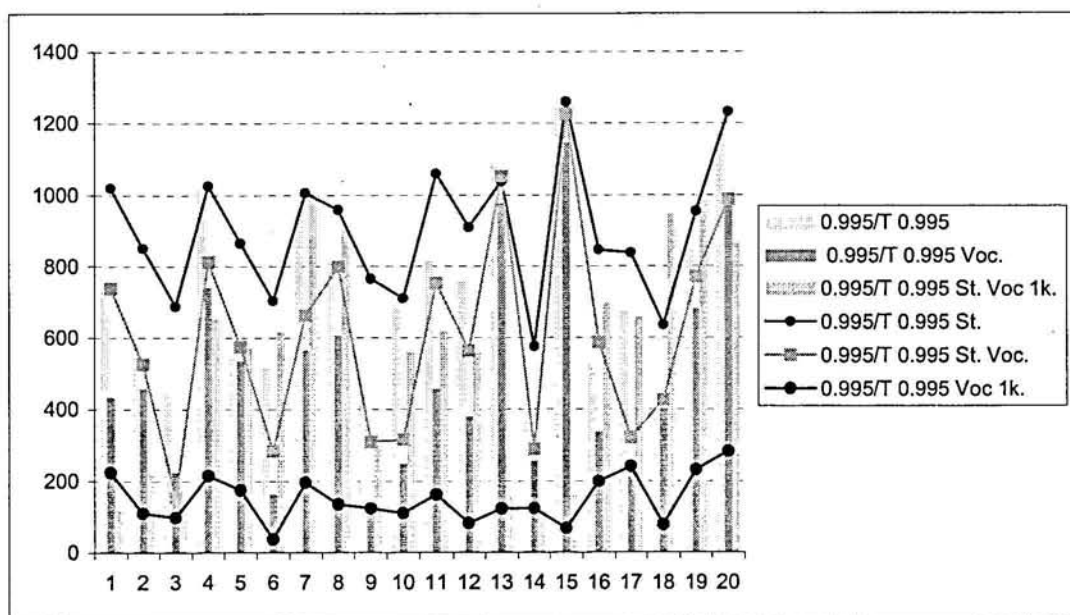


Figura 12 – Teste comparativo de recuperação de documentos

Os testes efectuados indicam que o resultado mais promissor para a precisão está associado ao último caso de teste na medida em que foi obtido o menor número de documentos recuperados. Este teste indica uma recuperação muito menor de documentos relativamente aos testes executados em que foram incluídos os termos com frequência superior a 200 apresentando, no entanto, uma descida do número de documentos relevantes recuperados. As duas figuras seguintes apresentam os resultados comparativos, para as primeiras 10 interrogações, entre os dois testes mais promissores com alteração da lista de paragem (inclusão de termos com frequência superior a 100 e 200). A Figura 13 apresenta a precisão obtida entre os dois testes enquanto que a Figura 14 apresenta os resultados para a recuperação. A variante da lista de paragem com inclusão de termos com frequência superior a 100 é indicada pela sigla Voc 1k.

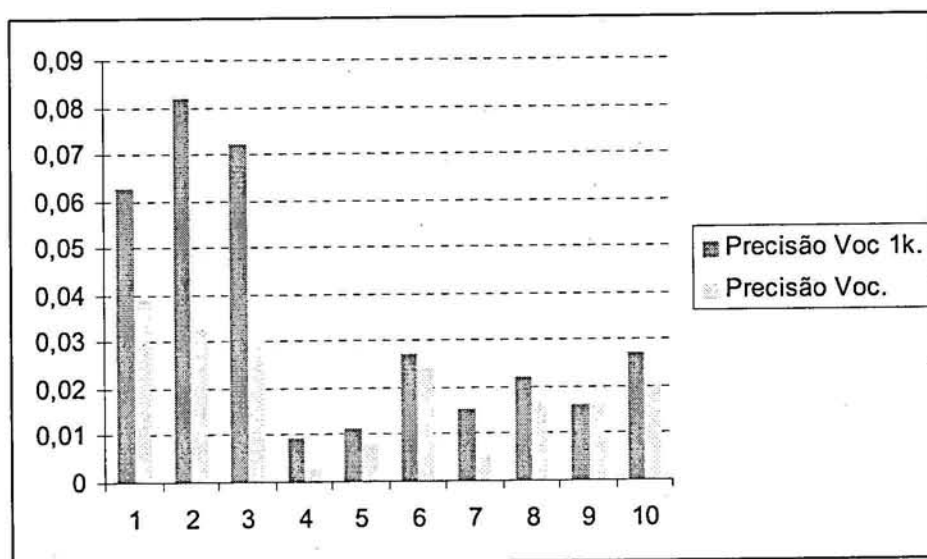


Figura 13 – Precisão entre as duas variantes com alteração da lista de paragem

A alteração da lista de paragem com inclusão de termos com frequência superior a 100, produziu resultados satisfatórios com o aumento significativo da precisão.

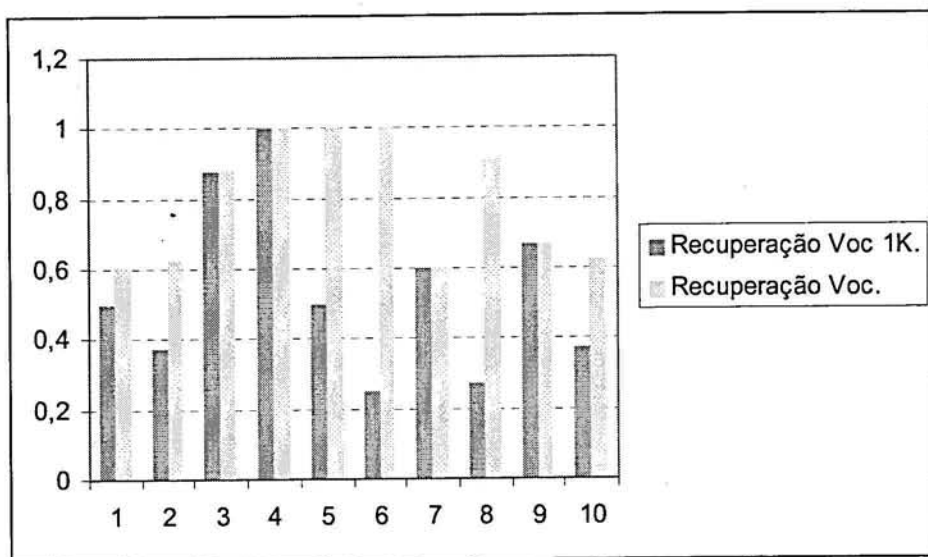


Figura 14 – Recuperação entre as duas variantes com alteração da lista de paragem

A recuperação obtida no teste com a variante Voc 1K. foi, em média, inferior a obtida pela variante com inclusão dos termos com frequência superior a 200 (Voc.) uma vez que esta última variante recupera um número ainda muito elevado de documentos. Em média, são recuperados 3.8 documentos por interrogação pela variante Voc 1K. enquanto que a variante Voc. recupera 5.5 documentos relevantes por interrogação. Os documentos da colecção da CRAN são constituídos por um número elevado de termos, tendo em consideração a dimensão de cada documento (Losada e Barreiro, 2001). No estudo efectuado por Losada e Barreiro (2001) a dimensão média para um documento é de 86.1 termos enquanto no sistema RST1 é de 37.9, após aplicação da análise léxica com lista de paragem alterada com inclusão dos termos com frequência superior a 100. Esta diferença da dimensão sobre a mesma colecção está directamente associada às diferentes abordagens implementadas para a análise léxica. Losada e Barreiro (2001) indicam no seu estudo que esta colecção contém inúmeros termos que não são importantes para determinar o conteúdo de um documento. Este facto pode estar na origem da menor recuperação de documentos relevantes no caso do sistema RST1 dada a importância da correcta determinação da semântica do termo nesta abordagem. Apesar desta descida, este resultado indica que os termos com frequência superior a 100 devem ser considerados lista de paragem da colecção da CRAN, dada a melhoria significativa da precisão.

O teste de recuperação executado com radicalização para além da alteração do vocabulário com a eliminação dos termos muito frequentes revelaram resultados inferiores aos obtidos pela método sem recurso a esta técnica.

Co-ocorrência e tesouro

A co-ocorrência e a definição do tesouro são os próximos objectos de estudo no impacto que introduzem na recuperação de informação do sistema RST1.

Consideremos a medida adoptada no sistema RST1, apresentada por Ribeiro-Neto e Baeza-Yates (1999) e doravante designada por BR:

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

onde n_i é o número de documentos que contém o termo k_i , n_l é o número de documentos que contém o termo k_l e $n_{i,l}$ o número de documentos que contém ambos os termos de indexação. O exemplo seguinte pretende efectuar um estudo das possíveis alterações a introduzir neste processo. Sejam $d_1 = \{a, b, c\}$, $d_2 = \{a, d, e, f\}$, $d_3 = \{g\}$, $d_4 = \{a, b, e, f\}$ os documentos de uma colecção exemplo C em que a, b, c, d, e, f, g são os termos que ocorrem nos documentos. O sistema RST1 foi concebido para efectuar o cálculo definindo uma matriz simétrica da co-ocorrência dos termos. A Tabela 10 indica os valores obtidos de co-ocorrência para este exemplo.

	a	b	c	d	e	f	g
a	1,00	0,67	0,33	0,33	0,67	0,67	0,00
b	0,67	1,00	0,50	0,00	0,33	0,33	0,00
c	0,33	0,50	1,00	0,00	0,00	0,00	0,00
d	0,33	0,00	0,00	1,00	0,50	0,50	0,00
e	0,67	0,33	0,00	0,50	1,00	1,00	0,00
f	0,67	0,33	0,00	0,50	1,00	1,00	0,00
g	0,00	0,00	0,00	0,00	0,00	0,00	1,00

Tabela 10 – Valores de co-ocorrência do exemplo

Este exemplo indica, neste caso, valores de co-ocorrência relativamente altos mas fundamentalmente que o estabelecimento de uma matriz simétrica para esta medida deve ser reavaliado uma vez que a atribuição de alguns valores não é representativa da real informação contida na colecção. Um caso desta situação é o valor de co-ocorrência entre os termos 'e' e 'd'. O termo 'd' surge sempre que o termo 'e' ocorre mas o termo 'e' ocorre apenas em metade das ocorrências do termo 'd'. O valor de 0.50 em ambos os casos não reflecte esta situação. Coloca-se a questão desta falta de interpretação poder ter influência sensível nos resultados da recuperação de informação. Para ser avaliada esta hipótese foi definido uma variante da medida BR designada por BRx, definida como:

$$c_{i,l} = \left(\frac{n_{i,l}}{n_i + n_l - n_{i,l}} \right) \times \left(\frac{n_l}{n_i} \right)$$

em que se pretende captar uma co-ocorrência na eventual assimetria da ocorrência dos termos. Aplicando esta nova medida ao exemplo anterior, obtemos os dados apresentados na Tabela 11.

	a	b	c	d	e	f	g
a	1,00	0,44	0,11	0,11	0,44	0,44	0,00
b	1,00	1,00	0,25	0,00	0,33	0,33	0,00
c	1,00	1,00	1,00	0,00	0,00	0,00	0,00
d	1,00	0,00	0,00	1,00	1,00	1,00	0,00
e	1,00	0,33	0,00	0,25	1,00	1,00	0,00
f	1,00	0,33	0,00	0,25	1,00	1,00	0,00
g	0,00	0,00	0,00	0,00	0,00	0,00	1,00

Tabela 11 – Valores de co-ocorrência definidos com medida BRx

Os valores indicam uma melhoria na captação da informação sobre a co-ocorrência entre os termos do exemplo. A medida BRx origina uma muito ligeira alteração na distribuição dos valores de co-ocorrência dos termos para o caso da colecção da CRAN. Apesar das primeiras indicações positivas obtidas no exemplo anterior, a medida BRx não

produz melhores resultados a nível da precisão média, como se pode comprovar na Figura 15. Com base nesta informação dois testes foram executados com recurso à uma medida designada por BRy, não simétrica, definida como

$$c_{i,l} = \frac{n_{i,l}}{n_l}$$

em que se pretende captar na co-ocorrência dos termos o efeito da assimetria e da frequência dos termos na colecção, criando valores mais baixos de co-ocorrência para termos mais frequentes. A Tabela 12 indica os valores obtidos de co-ocorrência para o exemplo anteriormente enunciado.

	a	b	c	d	e	f	g
a	1,00	1,00	1,00	1,00	1,00	1,00	0,00
b	1,00	1,00	1,00	0,00	0,50	0,50	0,00
c	0,30	0,50	1,00	0,00	0,00	0,00	0,00
d	0,30	0,00	0,00	1,00	0,50	0,50	0,00
e	0,67	0,50	0,00	1,00	1,00	1,00	0,00
f	0,67	0,50	0,00	1,00	1,00	1,00	0,00
g	0,00	0,00	0,00	0,00	0,00	0,00	1,00

Tabela 12 – Valores de co-ocorrência definidos com a medida BRy

Os resultados obtidos com a medida BRy foram melhores que os obtidos com a medida BR no número de documentos recuperados.

A Figura 15 apresenta, para cada interrogação, o número de documentos recuperados para diferentes parametrizações do sistema RST1. Na legenda são indicados os valores de limiar usados e as medidas usadas: 0.995 BRx, 0.995 BR e 0.995 BRy usando as respectivas medidas associadas. Estes testes foram efectuados sem alteração do vocabulário da colecção.

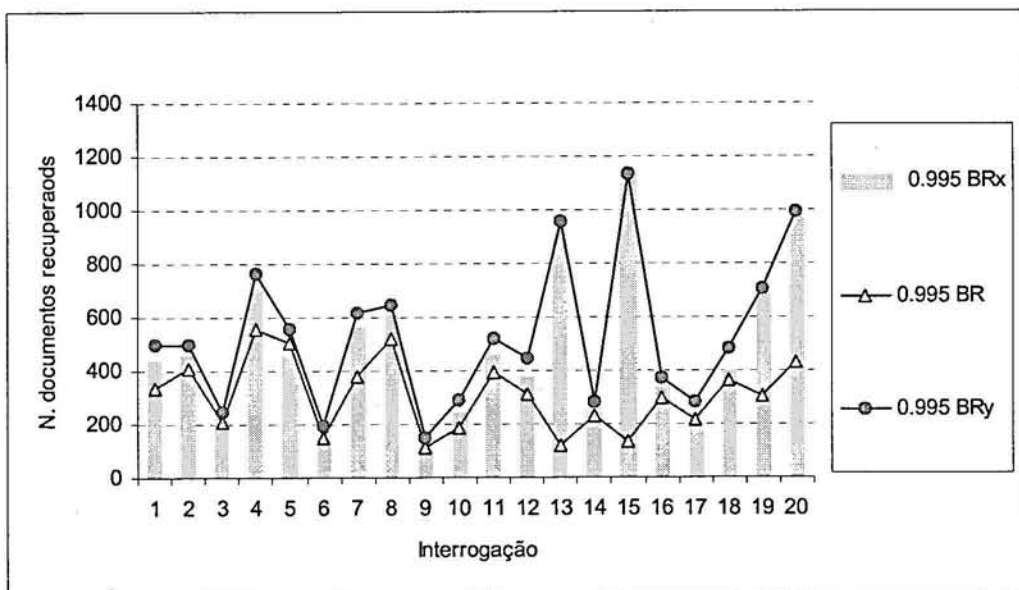


Figura 15 – Documentos recuperados para diferentes parametrizações

O resultado mais promissor de todos os testes efectuados é obtido quando a medida de BR é usada, uma vez que o número de documentos é, em média, consideravelmente menor. No entanto, é necessário avaliar a recuperação de documentos relevantes. Esta informação é apresentada na Figura 16.

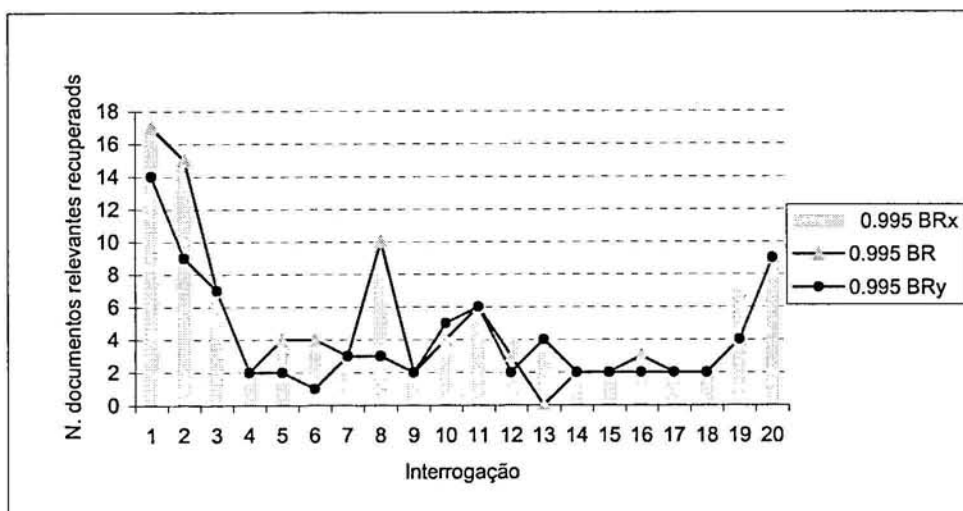


Figura 16 – Documentos relevantes recuperados para diferentes parametrizações

Os resultados apresentados nas duas figuras anteriores indicam que a configuração mais adequada para a recuperação de informação deve estar associada à configuração do sistema RST1 obtida com recurso à medida BR, de Baeza-Yates e Ribeiro-Neto. A precisão média obtida para a medida BR é de 0.017, valor muito inferior ao obtido por outros métodos. Deste modo, estabelece-se que a utilização de uma medida não simétrica não aparenta ter influência na qualidade da recuperação.

No decorrer da execução dos testes foi detectada uma diferença significativa do número de documentos recuperados entre dois testes particulares. Estes testes foram executados com recurso a tesauros definidos com limiares diferentes em que não foi alterada a lista de paragem da colecção da CRAN com inclusão de termos frequentes (termos com frequência superior a 100). O teste com tesouro definido com o limiar menor, 0.60, recupera um número menor de documentos do que o teste definido com limiar 0.95. Estes resultados parecem indicar que a construção de um tesouro que permite a inclusão de um maior número de termos de indexação pode permitir efectuar uma classificação difusa de documentos mais consistente. Com base nesta observação foi efectuada a construção de um tesouro com um limiar de 0.1 não se tendo registado, no entanto, melhorias na classificação difusa. Este facto indica a existência de um limiar limite a partir do qual a construção do tesouro deixe de ter impacto na recuperação.

Posteriormente, foram executados os mesmos dois testes para os mesmos valores, mas agora com o vocabulário alterado com os termos considerados com alta frequência. Estes testes estão identificados com a sigla Voc. 1k na Figura 17.

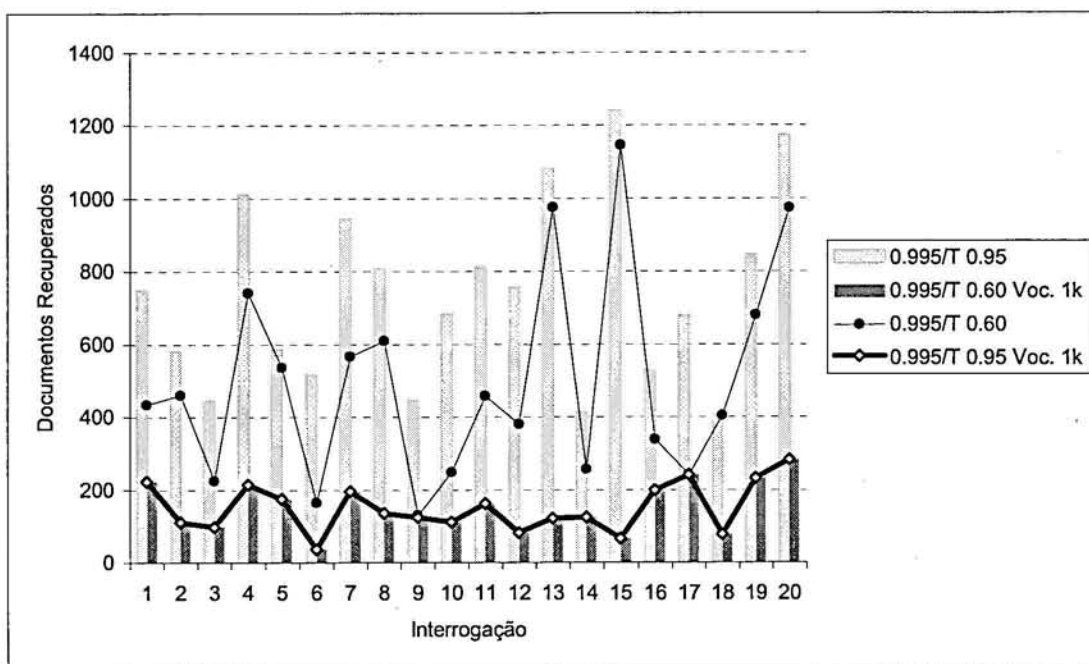


Figura 17 – Recuperação de Documentos com diferentes tesauros

Os resultados da figura anterior indicam que a definição do limiar de construção do tesauro é determinante na obtenção da melhor qualidade da recuperação quando o vocabulário da colecção apresenta termos com elevada frequência. Nesta situação, o valor deve ser relativamente mais baixo. Na situação em que o vocabulário não inclui termos com elevada frequência na colecção, o limiar para a construção do tesauro não é determinante.

Comparação entre documentos da colecção e interrogação

O estabelecimento do grau de inclusão de um termo k_i relativamente a um documento d_j é estabelecida pela fórmula

$$\mu_{i,j} = 1 - \prod_{k_i \in d_j} (1 - c_{i,l})$$

que produz valores de inclusão máximo para os termos que ocorrem no documento, não tendo presente a informação contida no documento e na colecção na sua globalidade. Consideremos o exemplo anteriormente apresentado no estudo da co-ocorrência do sistema RST1. Na Tabela 13 apresentam-se os valores de inclusão obtidos para os diferentes termos desta colecção.

	a	b	c	d	e	f	g
d_1	1	1	1	0,33	0,78	0,78	0
d_2	1	0,85	0,33	1	1	1	0
d_3	0	0	0	0	0	0	1
d_4	1	1	0,67	0,83	1	1	0

Tabela 13 – Classificação difusa dos documentos do exemplo

O grau de inclusão obtido para um termo para um dado documento não reflecte de forma mais correcta a informação existente no documento uma vez que não deve ser suficiente a ocorrência de um termo num documento para este ser, de imediato, considerado com o grau de inclusão máximo, independentemente da sua importância na definição semântica do documento e na própria colecção. Tendo em conta estes factores, foi desenvolvido uma variante do sistema RST1 em que uma nova medida de inclusão, BRz, foi adoptada. Esta medida é normalizada pelo conjunto de documentos que possuem o índice, ndc , para o qual está a ser definido o valor de inclusão:

$$\mu_{i,j} = \frac{1 - \prod_{k_l \in d_j} (1 - c_{i,l})}{ndc}$$

Esta medida, aplicada ao exemplo anterior, produz os valores descritos na Tabela 14.

	a	b	c	d	e	f	g
d_1	0,33	0,33	0,33	0,11	0,26	0,26	0
d_2	0,25	0,21	0,08	0,25	0,25	0,25	0
d_3	0	0	0	0	0	0	1
d_4	0,33	0,33	0,22	0,28	0,33	0,33	0

Tabela 14 – Classificação difusa com medida BRz

Os resultados obtidos com esta medida, apesar das potencialidades apresentadas no exemplo anterior, são muito fracos traduzindo-se numa muito baixa recuperação e precisão média, inferior a 0.01 para ambos os casos, sem e com o vocabulário da colecção da CRAN alterado. Esta variante da medida usada não permitiu estabelecer uma melhoria na qualidade da recuperação. A fórmula usada para a comparação entre a interrogação e os documentos da colecção, definida por

$$\mu_{q,j} = 1 - \prod_{i=1}^N (1 - \mu_{i,j})$$

é uma fórmula com características disjuntivas podendo recuperar demasiados documentos, tendo em consideração que os valores da classificação difusa dos documentos são elevados. A Tabela 15 apresenta os valores de satisfação dos documentos do exemplo apresentado anteriormente para a interrogação $q_1 = \{a, b, c\}$.

q_1	a	b	c	d	e	f	g	Satisfação Documento ($\mu_{q_1,j}$)
d_1	0	0	0	1	1	1	1	1
d_2	0	0,15	0,67	1	1	1	1	1
d_3	1	1	1	1	1	1	1	0
d_4	0	0	0,33	1	1	1	1	1

Tabela 15 – Recuperação de documentos para interrogação q_1

A Tabela 15 indica que apenas um dos documentos da colecção será rejeitado pela interrogação e todos os outros serão incluídos na resposta do sistema com indicação de total satisfação. Esta situação pode não traduzir correctamente a informação contida na interrogação, devendo ser questionada se apenas a presença de um dos termos na interrogação é suficiente para que todos os documentos que possuam este termo sejam considerados na resposta. Neste sentido, foram averiguadas fórmulas alternativas:

- $\mu_{q,j} = \prod_{i=1}^N (\mu_{i,j})$, definida pelo produto dos valores de inclusão que corresponde ao complemento da fórmula definida anteriormente para a comparação.

- $\mu_{q,j} = \min_{i=1,\dots,N} \mu_{i,j}$, definida pelo mínimo dos valores de inclusão.

- $\mu_{q,j} = \frac{1 - \prod_{i=1}^N (1 - \mu_{i,j})}{ndc}$, definida por uma variante da medida proposta por Baeza-Yates e Ribeiro-Neto em que ndc é o número de casos em que $\mu_{i,j} \neq 0$ e aqui designada por BRv.

A recuperação de documentos com recurso ao complemento da fórmula inicial ou ao valor mínimo obtiveram resultados muito fracos tendo sido recuperados um número muito reduzido de documentos relevantes, obtendo uma precisão e recuperação muito baixas (<0.1). Estes testes foram executados para diferentes valores de recuperação tendo em consideração as diferentes abordagens, isto é, os valores de recuperação de documentos foram sendo sucessivamente alterados com o intuito de tentar obter um valor de referência que obtivesse os melhores resultados possíveis para todas as interrogações. A fórmula BRv foi a que obteve resultados positivos uma vez que foi obtida uma precisão média superior aos testes efectuados anteriormente.

Agrupamento em clusters

O agrupamento em clusters tem como objectivo efectuar a definição de subconjuntos de documentos similares a partir da resposta obtida na fase de comparação do sistema RST1. Estes subconjuntos devem permitir ao utilizador efectuar a fácil identificação dos documentos relevantes e obter ao mesmo tempo documentos similares de potencial interesse. O agrupamento em clusters é particularmente útil nos casos em que é recuperado um número elevado de documentos. Na fase inicial dos testes no sistema RST1 foram usados os três algoritmos implementados: *Single Link*, *Complete Link* e o *Group Average Link*, tendo o último apresentado os melhores resultados. Deste modo, foi decidido restringir os testes para o método *Group Average Link*, não sendo apresentados os resultados preliminares dos outros dois métodos dado os mesmos serem inferiores. O teste de agrupamento para o método do *Single Link* foi implementado sendo definido como valor de corte $(N-1)/3$ iterações, onde N é o número de documentos. A Figura 18 apresenta o número de clusters criados, por interrogação, em que todos os conjuntos de documentos foram recuperados com o mesmo valor de inclusão e recorrendo a tesauros definidos com diferentes limiares (0.99, 0.95, 0.995, 0.90 e 0.85). O método usado para o agrupamento em clusters foi o *Group Average Link* parametrizado com um valor limite de dissimilaridade de 0.10.

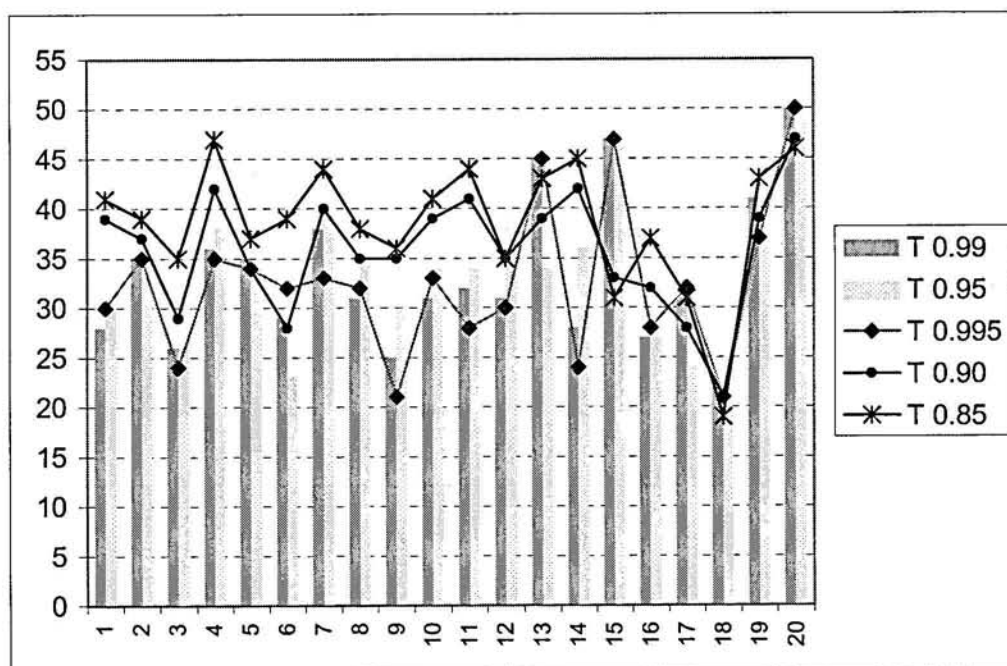


Figura 18 – Clusters criados por interrogação para 0.10 como valor de dissimilaridade

O número de clusters criados varia conforme os documentos recuperados, não sendo possível estabelecer nenhuma ligação entre o agrupamento em clusters e a recuperação difusa de documentos face à diversidade de resultados obtidos. Para todos os testes efectuados com dissimilaridade de 0.10, o número de documentos agrupados no mesmo cluster foi considerado muito baixo. Face a estes resultados, foram efectuados testes e análises no sentido de determinar e corrigir as possíveis causas. Os testes de agrupamento em clusters foram efectuados com o objectivo de averiguar a criação de clusters que podiam constituir-se como clusters interessantes para o utilizador. Neste domínio, consideram-se clusters interessantes aqueles que apresentam uma percentagem elevada de documentos relevantes no mesmo cluster e, em simultâneo, possam desvendar novas relações de modo a informar o utilizador para este melhorar a sua interrogação futura ao sistema. A Figura 19 apresenta o número de clusters criados para o conjunto difuso associado à resposta da primeira interrogação e para diferentes valores de dissimilaridade.

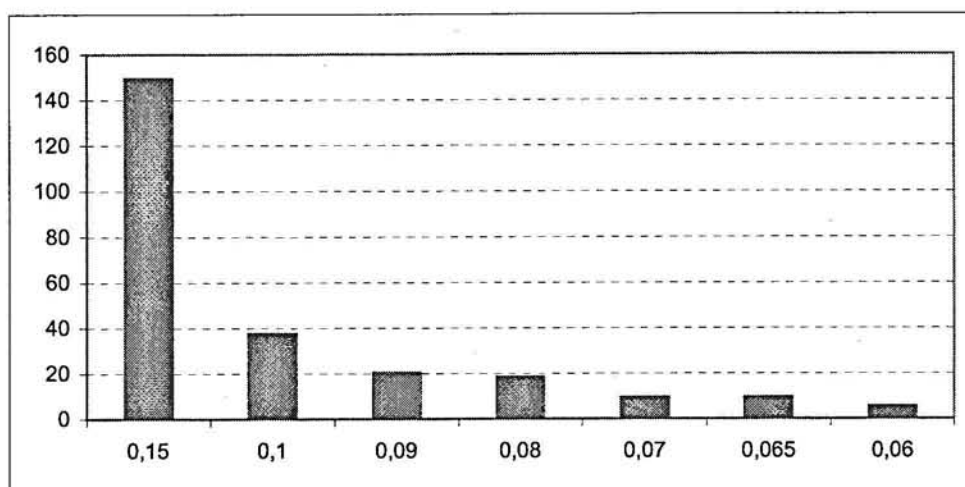


Figura 19 - Clusters criados para a resposta 1

O número de clusters criados decresce de forma acentuada à medida que o valor de dissimilaridade diminui, aumentando obviamente o número de documentos relevantes agrupados no mesmo cluster. No entanto, para valores inferiores a 0.06 o número de clusters é próximo de 1. Na Figura 20 é apresentada a análise tipo sobre a informação criada para definição do melhor cluster para diversos valores de dissimilaridade.

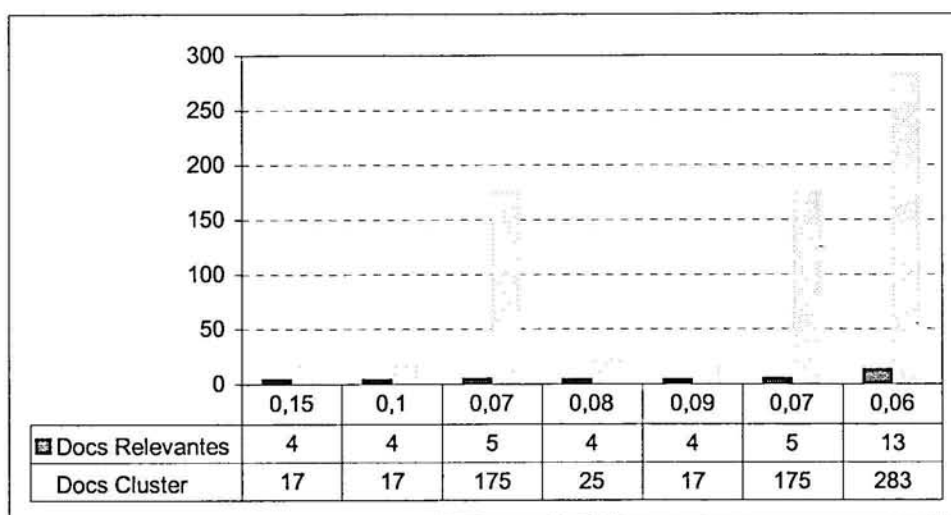


Figura 20 – Distribuição de documentos num cluster

A análise indicada na Figura 20 foi efectuada para as interrogações do conjunto de teste e para diferentes parametrizações. Deste modo, após a realização de cada um dos testes, foi efectuada uma análise exaustiva para cada um dos cluster de modo a poder ser obtido um valor de referência para o agrupamento em clusters para a colecção da CRAN. Esta análise teve como conclusão que o valor mais adequado para a dissimilaridade é o 0.065, valor onde se obtém um número adequado de documentos agrupados no mesmo cluster em que a semelhança dos documentos é significativa.

7.5 RST2 - Evolução do sistema RST1

Os testes iniciais do sistema RST1 indicaram a recuperação de um número demasiado elevado de documentos no conjunto difuso definido pelo processo de comparação, um número próximo do número total de documentos da colecção. Esta situação desencadeou uma análise às diferentes componentes do sistema RST1 que permitiu retirar ilações para introduzir alterações na sua configuração inicial, uma vez que os resultados indicavam diversos problemas, localizados em alguns componentes do sistema. Para alguns destes problemas e após múltiplos testes, foi possível diminuir significativamente o número de documentos recuperados. Esta alterações conduziram a uma evolução do RST1, designado por RST2, permitindo alcançar melhores resultados comparativamente aos iniciais. As alterações introduzidas consistem:

- redefinição da lista de paragem do sistema com inclusão dos termos com frequência superior a 100.
- uso de fórmula alterada para comparação, denominada anteriormente por BRv.
- definição de estrutura de tesauro com valor de 0.7

Sistema RST2: qualidade da recuperação

A multiplicidade de testes executados permitiu desenvolver uma melhoria nos resultados obtidos face aos resultados iniciais. A Figura 21 apresenta comparativamente os documentos recuperados entre as duas versões do protótipo, RST1 e RST2.

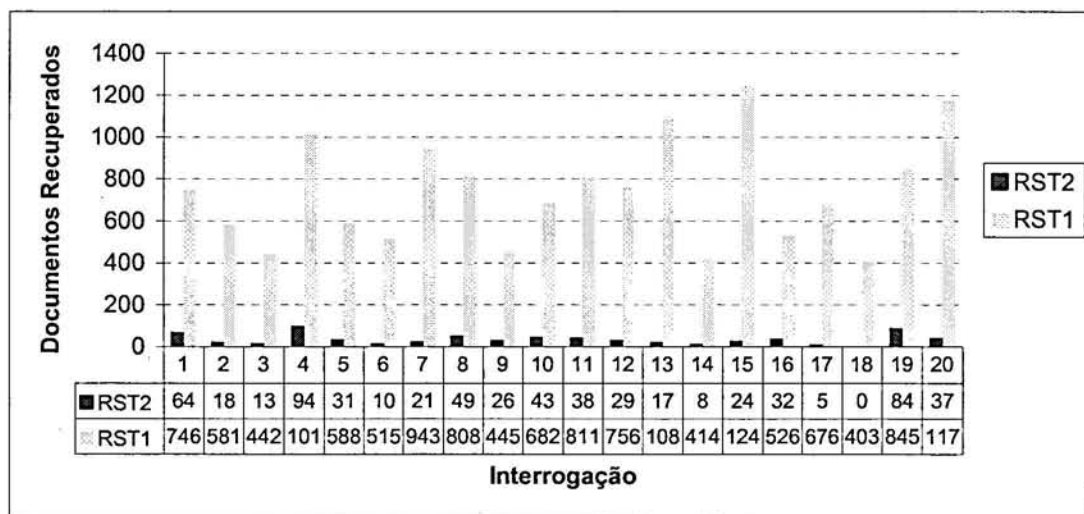


Figura 21 – Documentos recuperados por RST1 e RST2

No entanto, estes continuam ainda muito aquém em termos de precisão obtida por outros métodos. A precisão média obtida do RST2 é apresentada na Tabela 16 em parceria com os valores dos outros métodos. Esta precisão média do RST2 corresponde a uma recuperação média de 0.386, enquanto que os outros valores correspondem a uma precisão média calculada a partir de diferentes valores de recuperação: 0,25; 0.50 e 0,75.

Métodos	Precisão Média CRAN
VSM	0.384
PRF	0.485
QSD	0.428
QSDPRF	0.451
PRFQSD	0.463
RST1	0.017
RST2	0.100

Tabela 16 – Precisão média sobre a colecção da CRAN para diversos métodos

Como podemos constatar o problema fulcral continua a localizar-se na baixa precisão do sistema. Na investigação efectuada durante esta dissertação não foi detectada nenhuma abordagem similar com implementação de um protótipo, pelo que a comparação desta nova metodologia foi estabelecida com sistemas com fundamentos teóricos totalmente diferentes. De seguida são abordadas algumas das possíveis causas para os problemas persistentes desta abordagem.

Este estudo foi desenvolvido em detalhe sobre a colecção da CRAN, tendo sido inicialmente usada a colecção da LISA. Esta abordagem de concentrar o estudo apenas sobre a colecção da CRAN teve como intuito, como referenciado anteriormente, possibilitar a comparação de resultados e execução rápida de testes. Deste modo, este sistema pode apresentar comportamento substancialmente diferente em colecções onde a distribuição de termos possibilite a correcta definição da semântica e da co-ocorrência do termo para com o documento.

O agrupamento em clusters demonstrou resultados positivos. A Figura 22 apresenta detalhadamente os resultados do agrupamento em clusters para as primeiras 10 interrogações da colecção da CRAN. Nesta Figura é indicado o número de documentos recuperados para a interrogação como “Recuperados”, o número total de documentos dentro do cluster como “Cluster”, o número relevante de documentos para a interrogação como “Relevantes”, o número de documentos relevantes agrupados dentro do cluster como “Relevantes Cluster”, o número de documentos relevantes recuperados como “Relevantes Recuperados” e a melhor precisão obtida nos clusters criados face aos documentos

relevantes presentes nas respostas para cada uma das interrogações como “Precisão Cluster”.

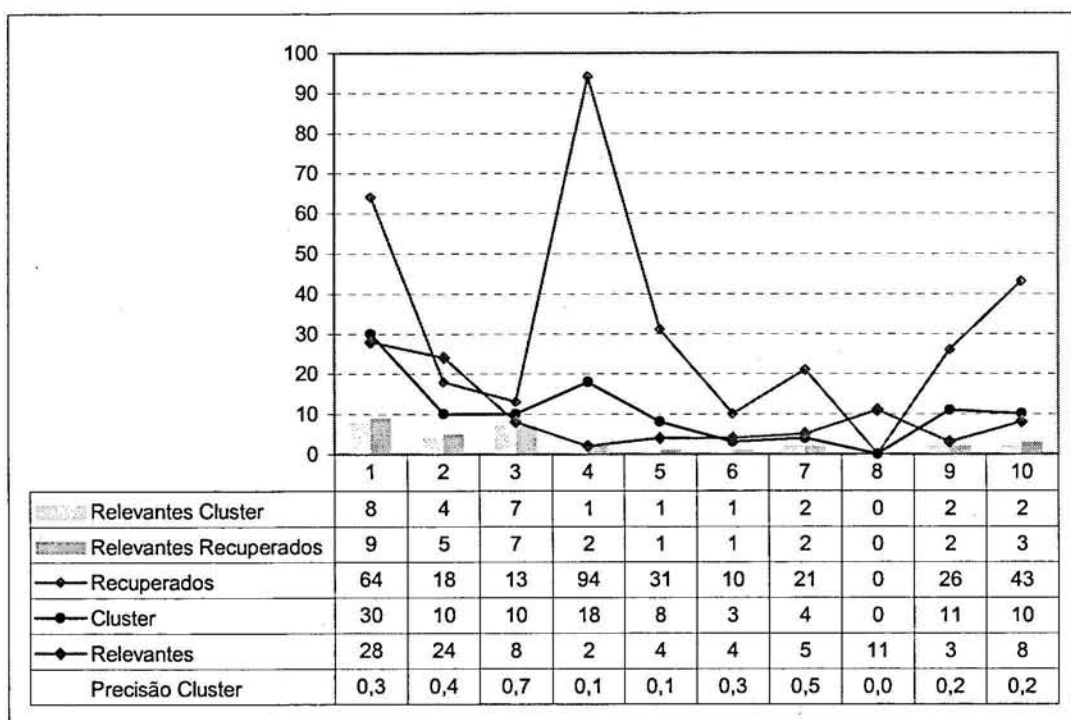


Figura 22 – Agrupamento em clusters sobre resultados do RST2

Em alguns dos clusters foi possível obter um agrupamento de documentos relevantes assinalável (>60%), tendo sido obtido alguns clusters com um número reduzido de documentos não relevantes. Estes valores podem ser superiores se o corte da construção do dendograma ocorrer mais tarde. O agrupamento em clusters para um conjunto de resultados de um sistema de recuperação de informação (*post-retrieval clustering*) deve criar clusters com o maior possível de documentos relevantes. Esta expectativa parte do princípio que documentos muito similares tendem a ser relevantes para a mesma interrogação.

A colecção da CRAN foi testada no contexto do agrupamento em clusters por alguns investigadores com o objectivo de averiguar se este processo podia ou não produzir uma maior eficiência na recuperação de informação. Um destes investigadores, Croft (1980, a

partir de Hearst e Pederson, 1996) descreve um método de agrupamento em clusters que usa a medida de co-ocorrência do coseno. Com recurso a este método, são obtidos resultados melhores do que uma simples recuperação ordenada de resultados. No entanto, estes resultados só conseguem ser obtidos desde que o cluster óptimo seja identificado sendo posteriormente efectuado sucessivas optimizações no sentido de melhorar o resultado obtido (Hearst e Pederson, 1996). No sistema *Scatter/Gather* (Hearst e Pederson, 1996), os resultados obtidos foram sendo ajustados ao número de documentos agrupados em clusters, isto é, efectuaram-se testes com os primeiros 100, 250, 500, 1000 documentos recuperados. Perante os dados obtidos Hearst e Pederson (1996) conjecturam que a hipótese cluster se confirma, sendo que os resultados apontam para uma situação onde o melhor cluster obtinha resultados pouco acima dos 50% de documentos relevantes agrupados no mesmo cluster. No entanto, Hearst e Pederson (1996) consideram sempre cinco partições, independentemente do número de documentos agrupados em clusters. Ora, é óbvio, que nestas circunstâncias, a percentagem de documentos relevantes venha a aumentar à medida que se aumente o número de documentos agrupados em clusters.

O agrupamento em clusters no sistema RST2 obteve resultados positivos, comparativamente ao sistema *Scatter/Gather*, tendo sido possível efectuado o agrupamento de um número significativo de documentos similares e relevantes no mesmo cluster.

O sistema *Scatter/Gather*, implementado por Hearst e Pederson (1996) usa uma medida de similaridade diferente, um método não-hierárquico, determinístico, designado por *Fractionation* que efectua a divisão de n documentos em k grupos e uma colecção cujos os documentos não são simples resumos tal como acontece na colecção da CRAN. A medida usada é baseada no coseno, com pesos atribuídos aos termos com base na frequência inversa dos termos nos documentos.

A medida usada no RST2, DICE, é normalizada, logo tem em consideração o número de termos em cada um dos documentos. Esta medida, binária e normalizada, não deve ser a responsável por alguns dos resultados obtidos menos positivos uma vez que não existe nenhuma indicação de que medidas binárias tenham influência sensível na definição dos clusters (Willet, 1983, a partir de Tombros, 2002), (Can e Ozkaran, 1990, a partir de Tombros, 2002), não existindo, consequentemente, nenhuma evidência que confirma ou

contraria o uso de medidas binárias no agrupamento em clusters de documentos. Foi observado que as similaridades obtidas entre documentos são todas muito baixas, como apresentadas na Tabela 17.

Intervalo	Valores Similaridade (%)
[0-0,1[	64,8
]0,1-0,2]	32
]0,2-0,9]	3
]0,9-1]	0,2

Tabela 17 – Valores de similaridade para a colecção da CRAN

O facto da colecção usada ser muito pequena relativamente a outras colecções mais recentes e usadas em outros sistemas pode ter tido uma influência determinante para alguns resultados menos positivos.

Os métodos usados *Complete Link* e *Group Average Link* são considerados como os mais adequados para a recuperação de documentos, sendo que o *Single Link* foi implementado após obtenção e análise dos resultados dos dois primeiros métodos, para comparação de resultados. Existe referência que Murtagh (1984, a partir de Tombros 2002) efectuou um estudo comparativo de métodos hierárquicos, sendo o objectivo a análise do tipo de estruturas definidas pelos diversos métodos e quão equilibradas eram estas. Neste estudo, Murtagh (1984) indica que não deveria ser dada preferência ao *Single Link*, indicando os outros três métodos: *Complete Link*, *Group Average Link* e *Ward* como os mais adequados (Tombros, 2002).

Tendo conhecimento das diversas abordagens que efectuem a indexação de valores difusos aos termos da interrogação, por exemplo Buell e Kraft (1981), tal não foi considerado no âmbito do sistema desta dissertação dado que estas requerem que o utilizador especifique qual o peso a atribuir a cada termo da interrogação. Shaw et al.

(1997) referencia que a eficiência do agrupamento hierárquico de documentos relevantes em clusters não é significativamente diferente da obtida por procedimentos aleatórios, após este ter obtido resultados negativos na tentativa de demonstrar a eficiência deste método (Tombros, 2002).

No sistema RST (RST1 ou RST2) foi experimentado um método hierárquico, sem determinação de medidas heurísticas para efectuar, com um mínimo de intervenção, o agrupamento em clusters de documentos recuperados. Alguns dos resultados menos positivos podem ser o reflexo dos seguintes factores conjugados, num menor ou maior grau:

- Uso de colecções de documentos relativamente pequenas, constituídas por documentos de pequenas dimensão, que não conseguem produzir uma medida de discriminação totalmente correcta de similaridade entre os diversos documentos.
- Uso de métodos hierárquicos, não heurísticos, em que nenhuma predefinição de parâmetros sobre tamanho, número de clusters é efectuada, para respeitar pressupostos teóricos da teoria dos conjuntos aproximativos.

No entanto, e apesar de todo estes condicionalismos os resultados são considerados positivos e nesta situação em particular, a afirmação de Cormack (1971, a partir de Tombros, 2002): “Some – I am tempted to say most – data are just not classifiable” não se aplica de todo, podendo mesmo afirmar-se, à luz dos dados obtidos por Hearst e Pederson (1996) que o sistema RST conseguiu efectuar o agrupamento em clusters com um grau bastante satisfatório.

8 Contribuição e conclusões

Este capítulo identifica a contribuição efectuada na área da gestão de informação, as principais conclusões obtidas pela pesquisa desenvolvida no âmbito desta dissertação e possíveis actividades futuras.

8.1. Contribuição

Esta dissertação estuda o uso de metodologias baseadas em conjuntos difusos (*fuzzy sets*) e conjuntos aproximativos (*rough sets*) na recuperação de informação. No decorrer desta dissertação são investigadas e aplicadas diversas técnicas na área da recuperação de informação, tendo sido desenvolvido, quase integralmente pelo autor, um sistema de recuperação de informação que combina de forma inovadora os fundamentos teóricos das duas teorias em estudo. Este sistema experimental de recuperação de informação, designado por RST, é definido e implementado sendo aplicadas, em diferentes fases do processo, a teoria dos conjuntos difusos e a teoria dos conjuntos aproximativos. Uma estratégia de indexação difusa é usada para a descrição de documentos, baseada num tesouro criado a partir da colecção de documentos com recurso a uma medida estatística. Os documentos recuperados são agrupados em clusters usando uma abordagem no domínio da teoria dos conjuntos aproximativos. Os algoritmos são testados com recurso a duas colecções de teste: LISA e CRAN. São implementados três algoritmos distintos de agrupamento em clusters: *Single Link*, *Complete Link* e *Group Average Link*.

O sistema RST apresenta algumas potencialidades e foram identificados alguns dos seus problemas quando aplicado a colecção da CRAN. Alguns dos problemas detectados estão directamente associado a própria colecção e outros são inerentes aos processos implementados e sofreram diversas alterações e melhorias. Estas alterações, limitadas para não comprometer a base teórica inerente à proposta desta dissertação, conduziram a uma evolução do protótipo inicial. A experimentação desenvolvida foi possível dada a modularidade alcançada na implementação deste sistema que permitiu a realização de

múltiplos testes para comparação de diferentes abordagens constituindo-se um instrumento nuclear da actividade desenvolvida e um dos seus principais produtos.

Nos capítulos iniciais desta dissertação são analisadas duas teorias que captam aspectos distintos da imprecisão da informação: indiscernibilidade e incerteza. A teoria dos conjuntos aproximativos e a teoria dos conjuntos difusos são consideradas, nesse sentido, ortogonais. As suas potencialidades no âmbito da recuperação de informação são definidas, tendo sido descritas abordagens para uma possível combinação das duas teorias para uma melhoria da recuperação de informação. A combinação das duas teorias apresenta uma base teórica sólida, tendo esta dissertação, através da apresentação do sistema RST, contribuído para fortalecer uma área de investigação com capacidade de contribuir para o desenvolvimento da recuperação de informação.

8.2. Conclusões

A gestão de informação possui uma área vasta de problemas em discussão. A recuperação de informação, sendo apenas um dos problemas, é deveras estimulante e muito vasta em que uma dissertação pode apenas analisar uma questão particular. Embora devidamente focalizada, esta investigação abrangeu um domínio de conhecimento e de técnicas muito vastos dada a riqueza da área.

A implementação do sistema RST permitiu delinear a importância dos vários processos e dos conceitos teóricos subjacentes a um sistema de recuperação de informação e comparar experimentalmente diferentes abordagens. Os resultados obtidos não demonstram, nesta implementação, uma melhoria na recuperação de informação relativamente a outros sistemas existentes, nomeadamente baseados no modelo vectorial, como seja o sistema SMART. No sistema RST foram aplicadas algumas técnicas usadas anteriormente por outros sistemas de recuperação em que os resultados obtidos foram inconclusivos relativamente à sua eficácia. Nesta situação, a construção automática de um tesouro para actuar sobre a interrogação do utilizador pode ter contribuído para que alguns documentos relevantes não fossem recuperados.

Apesar disto e fundamentado nos resultados obtidos, pode ser afirmado que a teoria dos conjuntos difusos e a teoria dos conjuntos aproximativos têm potencial para a recuperação de informação. As duas teorias possuem os fundamentos para promover uma melhoria na recuperação de informação.

A abordagem para a implementação do protótipo RST1 e sua evolução RST2 foi seleccionada tendo em consideração os fundamentos teóricos associados a ambas as teorias e os recursos disponíveis para a concretização dos testes. Os recursos disponíveis, em termos de *hardware*, constituíram-se como uma preocupação premente tendo sido um dos factores decisivos na selecção das colecções de teste colocando-se a maior ênfase na dimensão para permitir um ciclo temporal reduzido entre a alteração e o respectivo resultado.

As colecções da CRAN e LISA são muito homogéneas, especializadas num dado domínio. Os métodos aplicados a estas colecções não apresentam bons resultados em termos de precisão. As variantes do protótipo inicial definidas pela actuação sobre a lista de paragem, incluindo termos frequentes na colecção, pela alteração do método de cálculo da co-ocorrência e do método de agrupamento em clusters produziram uma melhoria significativa de resultados. Esta alteração sobre a lista de paragem permite concluir que as fórmulas usadas durante a análise léxica não conseguiram inibir o efeito produzido por esses termos inicialmente não identificados na lista de paragem. A selecção de colecções de teste menos homogéneas e de maior dimensão como sejam as colecções da TREC podem permitir a obtenção de resultados mais satisfatórios. Do mesmo modo, interrogações de maior dimensão podem beneficiar este tipo de abordagem em que a determinação da semântica dos termos é essencial.

A combinação das duas teorias apresenta alguns constrangimentos que são inerentes às próprias teorias como sejam a definição de relações de equivalência na teoria dos conjuntos aproximativos e as operações entre conjuntos no caso da teoria dos conjuntos difusos. Esta situação constitui-se como uma das possíveis causas para um estudo que não tem sido muito aprofundado na recuperação de informação. A descrição formal de um novo modelo é vital para permitir a partilha de informação, implementação de protótipos e desenvolvimento de variantes. No entanto, os modelos teóricos, por melhor

que possam ser definidos em termos matemáticos, requerem sempre uma experimentação e comparação com outros sistemas existentes, sem a qual não podem ser validados.

8.3. Actividades futuras

O sistema RST apresentado no decurso desta investigação apresenta algumas potencialidades que podem ser desenvolvidas uma vez que esta dissertação não esgotou, de todo, a problemática sobre a combinação das duas teorias na recuperação de informação.

Esta nova abordagem aplicada no contexto de recuperação de documentos de texto apresenta potencialidades que devem ser analisadas com testes sobre colecções da TREC, definindo-se algoritmos que possam captar com maior rigor a semântica dos termos definidos nos documentos. A aplicação de um operador do tipo Yager ou Hamacher na comparação entre a interrogação e o documento pode introduzir uma definição mais adequada da relevância do documento, uma vez que estes operadores apresentam uma maior flexibilidade relativamente às operações implementadas no sistema RST permitindo captar uma combinação entre a *t-norm* e a *t-conorm*. Um operador deste tipo permitirá, eventualmente, reduzir o efeito disjuntivo da fórmula usada no sistema RST. Estes desenvolvimentos podem levar a uma melhoria dos resultados obtidos.

A implementação de um mecanismo de reformulação da interrogação pelo utilizador (*relevance-feedback*) tem demonstrado em outros sistemas de recuperação de informação uma melhoria significativa dos resultados. Esta técnica deverá, do mesmo modo, produzir resultados positivos no sistema RST.

O agrupamento em clusters de documentos contidos na resposta de um sistema de recuperação, *post-retrieval clustering*, apresenta um campo fértil para a investigação. Este método aborda o problema do número cada vez maior de documentos recuperados pelos sistemas de recuperação de informação, sendo possível a definição de estruturas para permitir uma análise mais efectiva do utilizador sobre os documentos recuperados. As questões de eficiência na definição dos clusters também não são abordadas nesta dissertação, dado se tratar de um sistema experimental e as colecções usadas serem pequenas. Num contexto diferente, esta preocupação deva ser colocada de forma mais

premente. Do mesmo modo, a interface do sistema RST foi desenvolvida apenas o suficiente para permitir efectuar a análise dos resultados obtidos podendo ser objecto de redefinição em futuras actividades.

O armazenamento da informação das diversas estruturas criadas durante a implementação do RST como sejam, como exemplo, o tesauro, a matriz de co-ocorrência foi efectuado em ficheiros sequenciais. Deste modo, para permitir o uso das colecções da TREC deverá ser efectuada a definição de estruturas mais optimizadas e de algoritmos capazes de tratar o volume da informação, como seja o uso de técnicas de *hashing*. Caso contrário, o tempo de resposta do sistema ao utilizador será inevitavelmente muito longo.

A implementação de um sistema baseado na combinação destas duas teorias para a recuperação de documentos multimédia deverá ser enriquecedora uma vez que ambas as teorias possuem potencialidades neste domínio e o número deste tipo de documento na Web tem aumentado significativamente nos últimos anos. A dificuldade de classificação de documentos multimédia tem origem nos múltiplos e diferentes elementos que os constituem. A teoria dos conjuntos aproximativos apresenta os fundamentos teóricos para ultrapassar esta dificuldade podendo ser aplicada para efectuar a separação e classificação desses elementos. A pesquisa de documentos multimédia é efectuada inicialmente, na maioria dos sistemas, por meta-informação indicando o utilizador características do documento pretendido. A teoria dos conjuntos difusos pode ultrapassar esta dificuldade permitindo ao utilizador, por exemplo, a inserção de uma imagem como termo de pesquisa.

Referências e bibliografia

- Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999) - Modern Information Retrieval. ACM Press. Addison-Wesley, New York.
- Blocks, D.; Binding, C.; Cunliffe, D.; Tudhope, D. (2002) – Qualitative Evaluation of Thesaurus-Based Retrieval. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002). Lectures Notes in Computer Science, Springer-Verlag, 2458, 346-361.
- Buell, D. A.; Kraft, D. H. (1981) – A model for a weighted retrieval systems. Journal of the American Society for information Science, 32(3), 211-216.
- Can, F.; Ozkarahan, E. A. (1990) – Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. ACM transactions on Databases Systems, 15(4), 483-517.
- Choi, Dae-Young; Oh, Kwung-Yang (2000) – Case Study – ASA and its application to multi-criteria decision making. Fuzzy Sets and Systems, 114, 89-112.
- Chouchoulas, Alexios; Shen, Qiang (2001) – Rough Set-Aided Keyword Reduction for Text Categorisation. Informatics Research Report EDI-INF-RR-0121. Division of Informatics – University of Edinburgh, Centre for Intelligent Systems and their Applications. Disponível em <http://www.informatics.ed.ac.uk>.
- Choueka, Y. (1988) - Looking for Needles in a Haystack OR Locating Interesting Collocation Expressions in Large Textual Databases. Paper presented at the

Conference on User-Oriented Content-Based Text and Image Handling, MIT, Cambridge, Mass. 609-623.

- Cormack, R. M. (1971) – A review of classification. *Journal of the Royal Statistical Society, series A*, 134, 321-353.
- Croft, W. B. (1980) – A model of cluster searching based on classification. *Information Systems*, 5, 189-195.
- De Luca, A.; Termini, S. (1972) – A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, 20, 301-312.
- Dubois, D.; Prade, H. (1992) - Putting Rough Sets and Fuzzy Sets together. In *Intelligent Decision Support, Handbook of Advances of the Rough Set Theory* by R. Slowinski. Kluwer Academic Publishers, Dordrecht, Netherlands.
- Fasulo, Daniel (1999) – An analysis of recent work on clustering algorithms. Technical Report # 01-03-02, Department of Computer Science & Engineering. University of Washington. Seattle.
- Frakes William B.; Baeza-Yates, Ricardo (1992) - *Information Retrieval: Data Structures and Algorithms*. Prentice Hall.
- Frants Valery I.; Shapiro, Jacob; Voiskunskii, Vladimir G. (1997) - *Automated Information Retrieval: Theory and Methods*. Academic Press, London.
- Fritsche, M. (1974) - *Automatic Clustering Techniques In Information Retrieval*. Luxembourg: Commission of the European Communities.

- Fuhr, Norbert (1992) - Probabilistic Models in Information Retrieval. *The Computer Journal*, 35(3), 243-255.
- Fox, Cristopher (1992) - Lexical Analysis and Stoplists. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, 102-130, Prentice Hall.
- Harman, D.; Candela, G. (1990) - Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking. *Journal of the American Society for Information Science*, 41(8), 581-589.
- Harman, Donna; Fox, Edward; Baeza-Yates, R.; Lee, W (1992) - Inverted Files. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, 28-43, Prentice Hall.
- Hearst, M. A.; Pederson, J. O. (1996) – Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of ACM SIGIR*, 254-259.
- Hiemstra, Djoerd (2001) - Using language models for information retrieval Ph. D. thesis. University of Twente, Netherlands. ISBN 90-75296-05-3.
- Hiemstra, Djoerd; De Jong, Franciska (2001) - Statistical Language Models and Information Retrieval: natural language processing really meets retrieval. *Glott International*, 5(8), 288-294.
- Ho, Tu Bao; Funakoshi, Kaname (1998) - Information Retrieval Using Rough Sets. *Journal of Japanese Society for Artificial Intelligence*, 13(4), July.
- Hust, Armin; Klink, Stefan; Junker, Markus; Dengel, Andreas (2002) – Query Expansion for Web Information Retrieval. *Proceedings of Web Information*

Retrieval Workshop, 32nd Annual Conference of the German Informatics Society, 176-180. German Informatics Society. ISBN 3-88579-348-2.

- Jardine, N.; Van Rijsbergen, C. J. (1971) - The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7, 217-240.
- Lee, J. H. (1995) - Analyzing the effectiveness of extended boolean models in information retrieval. Technical Report TR95-1501, Cornell University. <http://cs-tr.cs.cornell.edu/>.
- Leuski, A., Allan, J. (2000) - Improving Interactive Retrieval by combining Ranked Lists and Clustering. *Proceedings of RYAO'2000*, 665-681.
- Leusky A., Croft B. (1996) - An Evaluation of Techniques for Clustering Search results. Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst.
- Losada, David E; Barreiro, Alvaro (2001) – Rating the Impact of Logical Representations on Retrieval Performance. *Proceedings of the 12th International Workshop on Database and Expert Systems Applications*, 247-253.
- Lowen, R. (1996) - Fuzzy Set Theory. Basic concepts, Techniques and Bibliography. Kluwer Academic Publishers, Dordrecht, Netherlands.
- Martín-Bautista, María J.; Sánchez, Daniel; Vila, María-Amparo; Larsen, Henrik L. (2000) - Measuring Effectiveness in Fuzzy Information Retrieval. *Flexible Query Answering Systems, Recent Advances, Proceedings of the Fourth International Conference on Flexible Query Answering Systems, FQAS 2000, October 25 - 28, 2000, Warsaw, Poland*, 396-402.

- Mateus, João Guilherme Conde Magalhães (1997) - A teoria dos conjuntos aproximativos para apoio à decisão – Aplicação a um problema de recursos humanos. Dissertação para obtenção do grau de Mestre em Investigação Operacional e Engenharia de Sistemas. Universidade Técnica de Lisboa, Instituto Superior Técnico.
- Miyamoto, S. (1998) - Application of rough sets to Information Retrieval. *Journal of the American Society for Information Retrieval*, 49, 195-205.
- Murtagh, F. (1984) – Structure of hierarchic clustering: implications for information retrieval and for multivariate data analysis. *Information Processing & Management*, 20(5/6), 611-617.
- Pawlak, Z. (1982) - Rough Sets. *International Journal of Computer and Information Sciences*, 11, 341-356.
- Pawlak, Z. (1991) - Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, Netherlands.
- Pedrycz, Witold; Gomide, Fernando (1998) - An introduction to Fuzzy Sets: Analysis and Design. A Bradford Book. MIT Press, London.
- Kise, Koichi; Junker, Markus; Dengel, Andreas; Matsumoto, Keinosuke (2001) – Experimental Evaluation of Passag-eBased Document Retrieval. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*.
- Rasmussen, Edie (1992) - Clustering Algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, 419-442. Prentice Hall.

- Salton, G. (1971) – The SMART Retrieval system. Englewood Cliffs, NJ: Prentice Hall.
- Salton, G.; Fox, E. A.; Wu, H. (1983) - Extended boolean information retrieval. *Communications of the ACM* 26 (11), 1022–1036.
- Sánchez, E.; Pierre, P. (1994) – Fuzzy logic and Genetic Algorithms in Information Retrieval. In *Proceedings of the Third Conference on Fuzzy Logic, Neural Nets and Soft Computing*, 29-35, Iizuka, Japan.
- Shaw, W. M.; Burgin, R.; Howell, P. (1997) – Performance standards and evaluations in IR test collections: Cluster-Based retrieval models. *Information Processing & Management*, 33(1), 1-14.
- Skowron, A.; Stepaniuk, J. (1994) - Generalized Approximation Spaces, *The 3rd International Workshop on Rough Sets and Soft Computing*, 156-163.
- Slowinski, Roman (1992) - *Intelligent Decision Support, Handbook of applications and Advances of the Rough Set Theory*. Kluwer Academic Publishers, Dordrecht, Netherlands.
- Sneath, P. H. A.; Sokal, R. R. (1973) - *Numerical Taxonomy*. San Francisco: Freeman.
- Srinivasan, Padmini (1989) - Intelligent Information Retrieval using Rough Set Approximations. *Information Processing & Management*, 25(4), 347-361.
- Srinivasan, Padmini (1991) - The importance of rough approximations for information retrieval. *International Journal of Man-Machines Studies*, 34, 657-671.

- Srinivasan, Padmini (1992). Tesauro Construction. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, 161-218, Prentice Hall.
- Srinivasan, Padmini; Ruiz, Miguel; E. Kraft, Donald H., Chen, Jianhua, Kundu, Sukhamay (2001) - Vocabulary Mining for Information Retrieval: Rough Sets and Fuzzy Sets. *Information Processing and Management*, 37, 15-38.
- Tahani, V. (1976) - A fuzzy model of document retrieval systems. *Information Processing and Management*, 12, 177-187.
- Takagi, Tomohiro; Kawase, Kazushi (2001) - A Trial for Data Retrieval Using Conceptual Fuzzy Sets. *IEEE Transactions on Fuzzy Systems*, 9(4), August.
- Tombros, A. (2002) - The effectiveness of query-based hierarchic clustering of documents for information retrieval. PhD Thesis. University of Glasgow, Department of Computing Science, Faculty of Computing Science, Mathematics and Statistics.
- Turtle, H. R.; Croft, W. B. (1992) - A comparison of Text Retrieval Models. *The Computer Journal*, 35(3), 279-290.
- Voorhees, Ellen (1986), *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval*. Ph. D. thesis, Cornell University.
- Voorhees, Ellen (1993) – On expanding query vectors with lexically related words. *Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*.

- Voorhees, Ellen (1999) – Natural Language Processing and Information Retrieval. In M.T. Pazienza (ed.), *Information Extraction: Towards Scalable, Adaptable Systems*. New-York, Springer-Verlag. 32-48.
- Willett, P. (1983) – Similarity coefficients and weighting functions for automatic document classification: an empirical comparison. *International Classification*, 3, 138-142.
- Willett, P. (1988) - Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5), 577–597.
- Witten, I. H.; Moffat, A.; Bell, T. C. (1994) - *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold.
- Yao, Y. Y. (1997) - Combination of Rough and Fuzzy Sets based on α -level sets, in T.Y. Lin and N. Cercone (Eds), *Rough Sets and Data Mining: Analysis for Imprecise Data*, Kluwer Academic Publishers, Boston, 301-321.
- Yao, Y.Y. (1998) - A Comparative Study of Fuzzy Sets and Rough Sets. *Information Sciences*, 109(1-4), 227-242.
- Yao, Y.Y. (2002) - Information Retrieval Support Systems. FUZZ-IEEE'02 in The 2002 IEEE World Congress on Computational Intelligence, Honolulu, Hawaii, USA, May 12-17, 773-778.
- Yao, Y.Y.; Li, X.; Lin, T.T.; Liu, Q. (1994) - Representation and Classification of Rough Set Models, *The 3rd International Workshop on Rough Sets and Soft Computing*, 630-637.
- Zadeh, L.A. (1965) - Fuzzy Sets. *Information and Control*, 8, 338-353.



FACULDADE DE ENGENHARIA

UNIVERSIDADE DO PORTO

BIBLIOTECA



0000078314