

---

## Organização flexível de documentos

---

*Tatiane Nogueira Rios*

---



SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: \_\_\_\_\_

# Organização flexível de documentos

**Tatiane Nogueira Rios**

***Orientadora: Profa. Dra. Solange Oliveira Rezende***

***Coorientadora: Profa. Dra. Heloisa de Arruda Camargo***

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA.*

**USP – São Carlos**  
**Abril 2013**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi  
e Seção Técnica de Informática, ICMC/USP,  
com os dados fornecidos pelo(a) autor(a)

R586o Rios, Tatiane Nogueira  
Organização flexível de documentos / Tatiane  
Nogueira Rios; orientadora Solange Oliveira Rezende;  
co-orientadora Heloisa de Arruda Camargo. -- São  
Carlos, 2013.  
128 p.

Tese (Doutorado - Programa de Pós-Graduação em  
Ciências de Computação e Matemática Computacional) --  
Instituto de Ciências Matemáticas e de Computação,  
Universidade de São Paulo, 2013.

1. Mineração de textos. 2. Agrupamento fuzzy. 3.  
Agrupamento de documentos. 4. Organização de  
documentos. I. Rezende, Solange Oliveira, orient.  
II. Camargo, Heloisa de Arruda, co-orient. III.  
Título.

# Agradecimentos

---

Após quatro anos de muito estudo e trabalho, chegou o momento da defesa de doutorado. Em mais um momento importante da minha vida, eu não posso deixar de agradecer à todos aqueles que de alguma forma participaram desta jornada, pois nada disso teria sido possível sem o apoio advindo do carinho da família, dos amigos e dos colegas de profissão.

Em primeiro lugar, agradeço a Deus, porque creio que Ele sempre está no controle de tudo.

Ao meu amado esposo, Ricardo, que sempre me estimula a crescer científica e pessoalmente. Agradeço à ele, por ser o pilar de sustentação da minha vida pessoal e profissional com seu inestimável apoio que preenche as minhas diversas falhas. Sua paciência e compreensão foram tão importantes quanto as pesquisas que resultaram diretamente nesta tese.

À minha querida filha, Marina, minha grande companheira na finalização deste trabalho, porque ela já me faz uma pessoa mais otimista e feliz.

Aos meus pais e irmãos, Valdeci, Dilzete, Tiago e Rodrigo, pelo incansável amor, apoio e carinho que ultrapassam distâncias e me fazem mais forte a cada dia.

À família que Deus me deu de presente, Robério, Josinha, Gustavo, Bel, Joyce e Eva, por sempre confiarem no meu potencial. À eles o meu eterno carinho.

À Solange, por sua orientação e amizade. Sem a sua colaboração e persistência, não teria sido possível concluir esta tese.

À Heloisa, por sua orientação segura e sua serenidade que sempre me servirão como exemplo.

Ao Dr. Enrique Herrera-Viedma, por sua orientação e apoio no período em que realizei pesquisas na Universidade de Granada, em Granada - Espanha.

Ao Dr. Pierre Pluye, por sua orientação e apoio no período em que realizei pesquisas na Universidade McGill, em Montreal-Canadá.

Aos colegas e amigos do Labic, agradeço pela amizade, disponibilidade e pelas trocas vivenciadas.

Aos amigos do CIG, agradeço pela amizade, pelo companheirismo e pelas divertidas reuniões que me encheram de otimismo.

Aos amigos espalhados pelo mundo que o doutorado me deu a oportunidade de conhecer. Aos amigos de Montreal-Canadá, Pedro, Tay, Aninha, Leo, Naty, Lucas, os pequenos Théo e Sophie, Janique, David, Dr. Grad, Mme Marrie, Mme Yanick, Mme Nacera, Afshin, Roza, Ali, Maria e todos os outros colegas que animaram as aulas de francês. Aos amigos de Granada-Espanha, Prof. Francisco Herrera, Michela, María, Manolo, Luisa, Jorge, Mati, Joaquim, Victoria, Alvaro e Isaac. Aos amigos de Hong Kong-China, Rouxi, Arisa, Asako, Naoko e Chano, que partiu mas deixou grandes recordações.

Aos professores de inglês, Marcos, Anand, Fernanda, Marika e Peter, que me mostraram que a língua ultrapassa barreiras e me auxiliaram para que este trabalho fosse ainda mais longe.

Aos amigos do Brasil, os que fiz durante o doutorado e aqueles que a amizade foi fortalecida nesse período.

Ao Santander pelo suporte financeiro na Espanha.

À CAPES pelo suporte financeiro no Brasil e no exterior.

E, por fim, mas não menos importante, à todos aqueles que não tem os seus nomes citados aqui, mas que de alguma forma estiveram comigo durante este período.

# Resumo

---

---

Diversos métodos têm sido desenvolvidos para a organização da crescente quantidade de documentos textuais. Esses métodos frequentemente fazem uso de algoritmos de agrupamento para organizar documentos que referem-se a um mesmo assunto em um mesmo grupo, supondo que conteúdos de documentos de um mesmo grupo são similares. Porém, existe a possibilidade de que documentos pertencentes a grupos distintos também apresentem características semelhantes. Considerando esta situação, há a necessidade de desenvolver métodos que possibilitem a organização flexível de documentos, ou seja, métodos que possibilitem que documentos sejam organizados em diferentes grupos com diferentes graus de compatibilidade. O agrupamento fuzzy de documentos textuais apresenta-se como uma técnica adequada para este tipo de organização, uma vez que algoritmos de agrupamento fuzzy consideram que um mesmo documento pode ser compatível com mais de um grupo. Embora tem-se desenvolvido algoritmos de agrupamento fuzzy que possibilitem a organização flexível de documentos, tal organização é avaliada em termos do desempenho do agrupamento de documentos. No entanto, considerando que grupos de documentos devem possuir descritores que identifiquem adequadamente os tópicos representados pelos mesmos, de maneira geral os descritores de grupos tem sido extraídos utilizando alguma heurística sobre um conjunto pequeno de documentos, realizando assim, uma avaliação simples sobre o significado dos grupos extraídos. No entanto, uma apropriada extração e avaliação de descritores de grupos é importante porque os mesmos são termos representantes da coleção que identificam os tópicos abordados nos documentos. Portanto, em aplicações em que o agrupamento fuzzy é utilizado para a organização flexível de documentos, uma descrição apropriada dos grupos obtidos é tão importante quanto um bom agrupamento, uma vez que, neste tipo de agrupamento, um mesmo descritor pode indicar o conteúdo de mais de um grupo. Essa necessidade motivou esta tese, cujo objetivo foi investigar e desenvolver métodos para a extração de descritores de grupos fuzzy para a organização flexível de documentos. Para cumprir esse objetivo desenvolveu-se: i) o método SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*), pelo qual descritores de grupos fuzzy *flat* são extraídos após o processo de agrupamento fuzzy, visando identificar tópicos da organização flexível de documentos independentemente do

algoritmo de agrupamento fuzzy utilizado; ii) o método SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*), pelo qual descritores de grupos fuzzy *flat* também são extraídos após o processo de agrupamento fuzzy utilizando o grau de pertinência dos documentos em cada grupo, obtidos do agrupamento fuzzy, como fator de ponderação dos termos candidatos a descritores; iii) o método HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*), pelo qual descritores de grupos fuzzy hierárquicos são extraídos após o processo de agrupamento hierárquico fuzzy, identificando tópicos da organização hierárquica flexível de documentos. Adicionalmente, apresenta-se nesta tese uma aplicação do método SoftO-FDCL no contexto do programa de educação médica continuada canadense, reforçando a utilidade e aplicabilidade da organização flexível de documentos.

# Abstract

---

---

Several methods have been developed to organize the growing number of textual documents. Such methods frequently use clustering algorithms to organize documents with similar topics into clusters. However, there are situations when documents of different clusters can also have similar characteristics. In order to overcome this drawback, it is necessary to develop methods that permit a soft document organization, i.e., clustering documents into different clusters according to different compatibility degrees. Among the techniques that we can use to develop methods in this sense, we highlight fuzzy clustering algorithms (FCA). By using FCA, one of the most important steps is the evaluation of the yield organization, which is performed considering that all analyzed topics are adequately identified by cluster descriptors. In general, cluster descriptors are extracted using some heuristic over a small number of documents. The adequate extraction and evaluation of cluster descriptors is important because they are terms that represent the collection and identify the topics of the documents. Therefore, an adequate description of the obtained clusters is as important as a good clustering, since the same descriptor might identify one or more clusters. Hence, the development of methods to extract descriptors from fuzzy clusters obtained for soft organization of documents motivated this thesis. Aiming at investigating such methods, we developed: i) the SoftO-FDCL (Soft Organization - Fuzzy Description Comes Last) method, in which descriptors of fuzzy clusters are extracted after clustering documents, identifying topics regardless the adopted fuzzy clustering algorithm; ii) the SoftO-wFDCL (Soft Organization - weighted Fuzzy Description Comes Last) method, in which descriptors of fuzzy clusters are also extracted after the fuzzy clustering process using the membership degrees of the documents as a weighted factor for the candidate descriptors; iii) the HSoftO-FDCL (Hierarchical Soft Organization - Fuzzy Description Comes Last) method, in which descriptors of hierarchical fuzzy clusters are extracted after the hierarchical fuzzy clustering process, identifying topics by means of a soft hierarchical organization of documents. Besides presenting these new methods, this thesis also discusses the application of the SoftO-FDCL method on documents produced by the Canadian continuing medical education program, presenting the utility and applicability of the soft organization of documents in real-world scenario.



# Sumário

---

---

|  |           |
|--|-----------|
| Lista de Figuras . . . . .   | ix        |
| Lista de Tabelas . . . . .   | xi        |
| Notação Matemática . . . . .   | xv        |
| <b>1 Introdução</b>  | <b>1</b>  |
| 1.1 Motivação, hipótese e objetivo . . . . .                               | 3         |
| 1.2 Contribuições . . . . .  | 5         |
| 1.3 Organização da tese . . . . .  | 6         |
| <b>2 Fundamentos da Organização de Documentos</b>                          | <b>7</b>  |
| 2.1 Considerações iniciais . . . . .                                       | 7         |
| 2.2 Agrupamento de documentos . . . . .                                    | 7         |
| 2.2.1 Agrupamento fuzzy de documentos . . . . .                            | 10        |
| 2.2.2 Algoritmo de agrupamento Fuzzy C-Means . . . . .                     | 10        |
| 2.2.3 Algoritmo de agrupamento Possibilístico C-Means . . . . .            | 13        |
| 2.2.4 Algoritmo de agrupamento Hierárquico Fuzzy C-Means . . . . .         | 14        |
| 2.2.5 Validação de agrupamento fuzzy . . . . .                             | 17        |
| 2.3 Extração de descritores de grupos . . . . .                            | 19        |
| 2.4 Considerações finais . . . . .   | 24        |
| <b>3 Abordagem para Organização Flexível de Documentos</b>                 | <b>27</b> |
| 3.1 Considerações iniciais . . . . .                                       | 27        |
| 3.2 Contextualização do problema . . . . .                                 | 27        |
| 3.3 Explorações preliminares . . . . .                                     | 30        |
| 3.3.1 Geração de regras fuzzy para classificação de documentos . . . . .   | 31        |
| 3.3.2 Recuperação de informação por meio de regras fuzzy . . . . .         | 35        |
| 3.4 Uma Abordagem Proposta para Organização flexível de documentos . . . . | 37        |
| 3.4.1 Trabalhos relacionados à organização flexível de documentos . . . .  | 38        |
| 3.4.2 Métodos propostos para extração de descritores de grupos fuzzy . .   | 40        |
| 3.5 Considerações finais . . . . .   | 42        |

|   |            |
|---|------------|
| <b>4 Métodos Propostos para Extração de Descritores de Grupos na Organização Flexível de Documentos</b>   | <b>43</b>  |
| 4.1 Considerações iniciais . . . . .  | 43         |
| 4.2 O método SoftO-FDCL . . . . .   | 44         |
| 4.3 O método SoftO-wFDCL . . . . .  | 46         |
| 4.4 O método HSoftO-FDCL . . . . .  | 48         |
| 4.5 Avaliação dos métodos propostos . . . . .   | 52         |
| 4.5.1 Coleções de documentos utilizados na avaliação dos métodos propostos  | 53         |
| 4.5.2 Pré-processamento dos documentos utilizados na avaliação dos métodos propostos . . . . .  | 56         |
| 4.5.3 Algoritmos de classificação utilizados na avaliação dos métodos propostos . . . . .   | 56         |
| 4.5.4 Avaliação do método SoftO-FDCL . . . . .  | 57         |
| 4.5.5 Avaliação do método SoftO-wFDCL . . . . .   | 71         |
| 4.5.6 Avaliação do método HSoftO-FDCL . . . . .   | 74         |
| 4.6 Considerações finais . . . . .  | 78         |
| <b>5 Aplicação do Método SoftO-FDCL: organização flexível de comentários de médicos de família sobre um processo de avaliação da educação médica continuada canadense</b> | <b>81</b>  |
| 5.1 Considerações iniciais . . . . .  | 81         |
| 5.2 Seleção manual de comentários construtivos . . . . .  | 85         |
| 5.3 Identificação automática de comentários . . . . .   | 86         |
| 5.4 Resultados obtidos . . . . .  | 87         |
| 5.5 Considerações finais . . . . .  | 92         |
| <b>6 Conclusões</b>   | <b>95</b>  |
| 6.1 Resumo das contribuições . . . . .  | 96         |
| 6.2 Publicações provenientes deste doutorado . . . . .  | 98         |
| 6.3 Parcerias . . . . .   | 100        |
| 6.4 Limitações . . . . .  | 102        |
| 6.5 Trabalhos futuros . . . . .   | 103        |
| <b>Referências Bibliográficas</b>   | <b>115</b> |
| <b>A Classificação de Documentos Utilizando Regras Fuzzy</b>  | <b>117</b> |
| <b>B Estratégia de <i>matching</i> para Recuperação Flexível de Documentos</b>  | <b>125</b> |

# Lista de Figuras

---

---

|     |  |    |
|-----|--|----|
| 2.1 | Posição dos documentos $d_1$ e $d_1$ com relação aos grupos $A_1$ e $A_2$ (Adaptado de Oliveira e Pedrycz (2007)) . . . . .  | 13 |
| 2.2 | Agrupamento hierárquico fuzzy obtido da execução do algoritmo HFCM (Adaptado de Pedrycz (1996)) . . . . .  | 14 |
| 2.3 | Abordagem DCF ( <i>Description Comes First</i> ) . . . . .   | 21 |
| 2.4 | Abordagem DCL ( <i>Description Comes Last</i> ) . . . . .  | 21 |
| 3.1 | Níveis de um Sistema de Recuperação de Informação . . . . .  | 29 |
| 3.2 | Exemplo de Organização Flexível com três grupos . . . . .  | 30 |
| 3.3 | Nível de representação de documentos em um SRI . . . . .   | 31 |
| 3.4 | Nível de organização de documentos em um SRI . . . . .   | 34 |
| 3.5 | Nível de recuperação de documentos em um SRI . . . . .   | 36 |
| 3.6 | Contextualização da abordagem proposta para organização flexível no nível da organização de um SRI . . . . .   | 41 |
| 4.1 | Agrupamento fuzzy hierárquico . . . . .  | 49 |
| 4.2 | Frequência de 50 descritores nos sumários dos documentos que possuem grau de pertinência no grupo representado por um gráfico maior ou igual ao limiar $\delta=0,25$ . . . . .                 | 63 |
| 4.3 | Organização flexível hierárquica de documentos da coleção Opinosis . . . .   | 75 |
| 4.4 | Visão Parcial da Organização flexível hierárquica de documentos da coleção Opinosis - observação da especialização/generalização dos tópicos identificados por descritores de grupos . . . . . | 76 |
| 4.5 | Exemplo de hierarquia fuzzy com cinco documentos . . . . .   | 76 |
| 5.1 | Exemplo de email enviado para um médico de família canadense sobre um <i>e-Therapeutics+ Highlight</i> . . . . .   | 82 |
| 5.2 | Exemplo de um <i>e-Therapeutics+ Highlight</i> . O trecho destacado em verde corresponde a um <i>Highlight</i> . . . . .   | 83 |
| 5.3 | Questionário IAM . . . . .   | 84 |

|     |  |     |
|-----|--|-----|
| 5.4 | Frequência dos descritores obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>all-2011</i> nas coleções <i>cfb-2011</i> e <i>ncfb-2011</i> . . . . . | 89  |
| 5.5 | Frequência dos descritores obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>all-2012</i> nas coleções <i>cfb-2012</i> e <i>ncfb-2012</i> . . . . . | 91  |
| A.1 | Método fuzzy para classificação de documentos . . . . .  | 117 |
| A.2 | Variável linguística $G_1$ . . . . .   | 119 |
| A.3 | Influência da quantidade de termos no desempenho da classificação obtida pelo método proposto . . . . .  | 123 |
| A.4 | Influência da quantidade de termos no desempenho da classificação pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR . . . . .                  | 123 |
| A.5 | Resultados obtidos pela mudança na frequência mínima . . . . .   | 124 |
| B.1 | Função de pertinência da palavra-chave $\iota_1$ . . . . .   | 127 |

# Lista de Tabelas

---

---

|      |  |    |
|------|--|----|
| 2.1  | Exemplo de matriz documentos-termos de uma coleção de 3 documentos . . . . .   | 8  |
| 2.2  | Exemplo de matriz documentos-grupos de uma coleção de 3 documentos . . . . .   | 11 |
| 2.3  | Métodos de extração de descritores de grupos . . . . .   | 22 |
| 4.1  | Matriz de contingência do termo $t_j$ para o grupo $g_l$ para as medidas de Recuperação de Informação utilizadas pelo método SoftO-FDCL . . . . .      | 45 |
| 4.2  | Matriz de contingência para as medidas de Recuperação de Informação utilizadas pelo método SoftO-wFDCL . . . . .                                       | 47 |
| 4.3  | Matriz de contingência do termo $t_j$ para o grupo $g_{l_u}$ para as medidas de Recuperação de Informação utilizadas pelo método HSoftO-FDCL . . . . . | 50 |
| 4.4  | Coleções de documentos utilizadas nos experimentos . . . . .   | 53 |
| 4.5  | Sumários escritos por humanos sobre o documento “ <i>battery life of the amazon kindle</i> ” da coleção opinosis . . . . .                             | 54 |
| 4.6  | Vinte descritores com maior valor de $f_1$ obtidos pelo método SoftO-FDCL para cada grupo da coleção Opinosis . . . . .                                | 59 |
| 4.7  | Taxas de acerto obtidas pelos algoritmos de classificação utilizando os descritores extraídos pelos métodos Centroide e SoftO-FDCL . . . . .           | 59 |
| 4.8  | Organização flexível da coleção Opinosis representada no formato documentos-grupos obtida pelo método SoftO-FDCL . . . . .                             | 61 |
| 4.9  | Exemplo de matriz documentos-termos . . . . .  | 63 |
| 4.10 | Matriz documentos-grupos obtida do agrupamento FCM para o exemplo da Tabela 4.9 . . . . .  | 64 |
| 4.11 | Matriz documentos-grupos obtida do agrupamento PCM para o exemplo da Tabela 4.9 . . . . .  | 64 |
| 4.12 | Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção Opinosis) . . . . .              | 65 |
| 4.13 | Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção 20Newsgroups) . . . . .          | 65 |

|   |    |
|---|----|
| 4.14 Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção Reuters-21578) . . . . . | 65 |
| 4.15 Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção WAP) . . . . .           | 66 |
| 4.16 Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção Hitech) . . . . .        | 66 |
| 4.17 Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção NSF) . . . . .           | 66 |
| 4.18 Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e $\chi^2$ (Coleção Opinosis) . . . . .                           | 69 |
| 4.19 Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e $\chi^2$ (Coleção 20NewsGroups) . . . . .                       | 70 |
| 4.20 Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e $\chi^2$ (Coleção Reuters) . . . . .                            | 70 |
| 4.21 Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e $\chi^2$ (Coleção WAP) . . . . .                                | 70 |
| 4.22 Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção Opinosis . . . . .  | 71 |
| 4.23 Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção 20NewsGroups . . . . .                                    | 72 |
| 4.24 Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção Reuters . . . . .   | 72 |
| 4.25 Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção WAP . . . . .   | 72 |
| 4.26 Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e $w\chi^2$ (Coleção Opinosis) . . . . .             | 73 |
| 4.27 Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e $w\chi^2$ (Coleção 20NewsGroups) . . . . .         | 73 |
| 4.28 Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e $w\chi^2$ (Coleção Reuters) . . . . .              | 73 |
| 4.29 Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e $w\chi^2$ (Coleção WAP) . . . . .                  | 73 |
| 4.30 Matriz atributo-valor obtida do corte no nível 3 da hierarquia apresentada na Figura 4.5 . . . . .   | 77 |
| 4.31 Comparação entre o método SoftO-FDCL e o método HSoftO-FDCL (Coleção Opinosis) . . . . .   | 77 |
| 4.32 Comparação entre o método SoftO-FDCL e o método HSoftO-FDCL (Coleção Hitech) . . . . .   | 77 |
| 4.33 Comparação entre o método SoftO-FDCL e o método HSoftO-FDCL (Coleção Reuters) . . . . .  | 78 |

|      |   |     |
|------|---|-----|
| 5.1  | CFB e non-CFB para o <i>Highlight</i> - Exemplo . . . . .   | 85  |
| 5.2  | Coleções utilizadas na aplicação do método SoftO-FDCL para organização flexível dos comentários de médicos de família canadenses. As coleções são identificadas pela coluna “ID” e a quantidade de comentários que compõe cada coleção é identificada pela coluna “# comentários”. A porcentagem de CFBs e non-CFBs obtida a partir das coleções de 2011 e 2012 são também apresentadas na coluna “# comentários” . . . . . | 87  |
| 5.3  | Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>all-2011</i> . . . . .   | 88  |
| 5.4  | Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>cfb-2011</i> . . . . .   | 88  |
| 5.5  | Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>ncfb-2011</i> . . . . .  | 88  |
| 5.6  | Exemplo de comentários em que o descritor “ <i>good</i> ” ocorre . . . . .  | 89  |
| 5.7  | Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>all-2012</i> . . . . .   | 90  |
| 5.8  | Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>cfb-2012</i> . . . . .   | 90  |
| 5.9  | Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção <i>ncfb-2012</i> . . . . .  | 90  |
| 5.10 | Graus de pertinência de três comentários em dois grupos . . . . .   | 91  |
| A.1  | Variação de frequência para seleção de termos . . . . .   | 121 |
| A.2  | Coleções utilizadas nos experimentos e respectivas quantidades de documentos e termos . . . . .   | 121 |
| A.3  | Teste 1 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR . . . . .   | 121 |
| A.4  | Teste 2 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR . . . . .   | 122 |
| A.5  | Teste 3 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR . . . . .   | 122 |
| A.6  | Teste 4 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR . . . . .   | 122 |
| A.7  | Teste 5 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR . . . . .   | 122 |
| A.8  | Configuração das frequências dos Testes 1 a 5 organizados em ordem crescente da frequência mínima e renomeados como testes de A a E . . . . .   | 124 |
| B.1  | Base de regras geradas a partir da matriz documentos-grupos . . . . .   | 126 |
| B.2  | Critérios de relevância definidos pelo usuário . . . . .  | 127 |



# Notação Matemática

---



---

| Notação                        | Significado   |
|--------------------------------|---|
| $D$                            | coleção de documentos   |
| $n$                            | quantidade de documentos  |
| $d_i$                          | um documento da coleção, com $1 \leq i \leq n$  |
| $k$                            | quantidade de termos  |
| $t_j$                          | um termo, com $1 \leq j \leq k$   |
| $\sigma(t_j, d_i)$             | frequência do termo $t_j$ no documento $d_i$  |
| $c$                            | quantidade de grupos  |
| $g_l$                          | um grupo, com $1 \leq l \leq c$   |
| $P = \{g_1, g_2, \dots, g_c\}$ | pseudo partição fuzzy   |
| $\mu(d_i, g_l)$                | grau de pertinência do documento $d_i$ no grupo $g_l$   |
| $\mu_1(d_i)$                   | primeiro maior grau de pertinência do documento $d_i$   |
| $\mu_2(d_i)$                   | segundo maior grau de pertinência do documento $d_i$  |
| $v_l$                          | protótipo do grupo $g_l$  |
| $dist(d_i, v_l)$               | distância entre o documento $d_i$ e o protótipo $v_l$ do grupo $g_l$  |
| $\alpha(d_i, g_l)$             | distância média entre o documento $d_i$ e todos os outros documentos pertencentes ao grupo $g_l$                  |
| $\beta(d_i, g_l)$              | distância média entre o documento $d_i$ e todos os documentos pertencentes à todos os grupos diferentes de $g_l$  |
| $m$                            | fator de fuzificação  |
| $\epsilon$                     | critério de parada do algoritmo Fuzzy C-Means   |
| $J(P)$                         | função objetivo $J$ do algoritmo Fuzzy C-Means sobre a pseudo partição $P$  |
| $\varphi(d_i, g_l)$            | tipicidade do documento $d_i$ com relação ao grupo $g_l$  |
| $K(Q)$                         | função objetivo $K$ do algoritmo Possibilístico C-Means sobre a pseudo partição $Q$                               |
| $y$                            | quantidade de níveis da hierarquia fuzzy de documentos  |
| $S(d_i)$                       | silhueta do documento $d_i$   |
| $M_{n \times k}$               | matriz documentos-termos  |
| $W_{n \times c}$               | matriz documentos-grupos  |
| $z$                            | quantidade de regras  |
| $R_s$                          | uma regra, com $1 \leq s \leq z$  |
| $Class$                        | classe de um documento  |
| $G_l$                          | variável linguística representativa do grupo $g_l$  |
| $x_i$                          | variável auxiliar do documento $d_i$ no agrupamento CFCM  |
| $Compat(R_s, d_i)$             | grau de compatibilidade do documento $d_i$ com a regra $R_s$  |
| $A = \{a_1, a_2, \dots, a_o\}$ | conjunto de termos linguísticos   |
| $o$                            | quantidade de termos linguísticos   |
| $a_q$                          | termo linguístico, no qual $1 \leq q \leq  A $  |
| $A_q(\mu(d_i, g_l))$           | grau de pertinência do grau de pertinência do documento $d_i$ no grupo $g_l$ no conjunto fuzzy $a_q$              |
| $\delta$                       | limiar que define se um documento pertence ou não a um determinado grupo no método SoftO-FDCL                     |
| $\zeta$                        | limiar que define se um documento pertence ou não a um determinado grupo no método HSoftO-FDCL                    |
| $I$                            | índice de desempenho de um grupo, o qual indica a qualidade de estrutura hierárquica a ser formada a partir deste |



## CAPÍTULO

### 1

# Introdução

meio da Mineração de Dados (MD) (Fayyad et al., 1996), a qual atua como um intermediário entre os dados registrados em um determinado conjunto de dados e o conhecimento que pode ser extraído por meio da identificação de padrões e regularidades presentes nos mesmos. Uma das principais características da MD é a forma como os dados são estruturados e organizados, que pode ser, por exemplo, por meio de tabelas de bancos de dados. Entretanto, nem todos os dados como, por exemplo, os documentos textuais, podem ser estruturados e organizados em um formato bem definido.

O avanço e a popularização da tecnologia vivenciados ao longo dos anos tornaram comum o uso de sistemas de coleta e armazenamento digital de dados por parte das mais diversas organizações, gerando bases de dados que crescem rapidamente, atingindo quantidades de dados que extrapolam a capacidade humana de, manualmente, analisá-las e compreendê-las por completo.

A extração automática de conhecimento a partir dessa crescente quantidade de dados armazenada digitalmente tornou-se uma tarefa de grande importância para as corporações, uma vez que, dessa maneira, é possível obter conhecimento por meio de informações novas e potencialmente úteis. Essa extração pode ser obtida por meio da Mineração de Dados (MD) (Tan et al., 2005), o qual possui como principal característica a exploração dos dados, de forma a estruturar e evidenciar padrões nestes dados, auxiliando na descoberta de conhecimento. Entretanto, nem todos os dados como, por exemplo, os documentos textuais, podem ser estruturados e organizados em um formato bem definido.

Quando o conjunto de dados consiste de documentos textuais, isto é, dados não-estruturados, utiliza-se uma especialização do processo de MD, denominada Mineração de Textos (MT) (Berry e Kogan, 2010; Aggarwal e Zhai, 2012). A MT diferencia-se da MD pela incorporação de atividades que visam a estruturação dos documentos em um formato apropriado para a obtenção automática de conhecimento, sem que haja perda de

---

informações relevantes em relação ao formato originalmente não-estruturado. Uma vez estruturados os documentos, algoritmos de MD convencionais podem ser aplicados para extrair conhecimento e informação desses dados por meio de padrões detectados em toda a coleção de documentos.

Uma grande variedade de métodos tem sido desenvolvida para gerenciar e organizar a crescente quantidade de documentos textuais, além de extrair automaticamente o conhecimento embutido nesses documentos (Bordogna e Pasi, 2001, 2004; Zadrozny e Nowacka, 2009). Entre esses métodos, pode-se destacar o agrupamento de documentos, o qual é muito utilizado para se obter conhecimento útil sobre os documentos.

Nesse contexto, Sistemas de Recuperação de Informação (SRIs) têm sido desenvolvidos para que os usuários deste tipo de sistema tenham acesso ao conhecimento obtido de documentos textuais, de maneira mais natural e intuitiva. Para tanto, faz-se necessário organizar os documentos a fim de que sejam compreendidos. Sendo assim, o tratamento de imprecisão e incerteza na organização de documentos é um problema de pesquisa importante devido ao grande volume de documentos que os SRIs têm gerenciado. Em geral, a imprecisão e incerteza estão presentes em documentos, porque diferentes usuários dos SRIs veem o documento sob diferentes perspectivas. Por exemplo, um determinado usuário pode organizar um conjunto de documentos com base em algum critério, como por exemplo pelos assuntos abordados nos documentos, enquanto outro usuário pode organizar o mesmo conjunto de documentos com base em algum outro critério, como por exemplo pela importância de cada documento com relação a um assunto específico. Do mesmo modo, ao realizar uma consulta em uma máquina de busca, os documentos obtidos como resultado desta consulta podem ter graus de importância diferentes para diferentes usuários.

No entanto, geralmente, SRIs são intolerantes em termos de incorporar imprecisão e incerteza. Para solucionar este problema e organizar os documentos de forma flexível, existe a necessidade de desenvolver SRIs flexíveis, os quais são capazes de gerenciar informações imperfeitas, isto é, informações imprecisas e/ou incertas (Kraft et al., 2006). Para tanto, a Computação Flexível (*Soft Computing*) (CF) (Zadeh, 1997) tem sido amplamente experimentada (Crestani e Pasi, 2000; Herrera-Viedma et al., 2006). A CF é um paradigma abrangente que agrega, tradicionalmente, as metodologias de Redes Neurais Artificiais, da Teoria de Conjuntos Fuzzy e da Computação Evolutiva, oferecendo uma nova visão para a solução de problemas complexos, os quais muitas vezes não dispõem de modelos matemáticos específicos (Pedrycz, 1998). A aplicação da CF em problemas complexos pode introduzir o conhecimento humano como, por exemplo, a cognição, o reconhecimento, a compreensão e a aprendizagem.

Considerando que não há a definição de um modelo matemático específico para o tratamento da imprecisão e incerteza presentes em documentos, a Teoria de Conjuntos Fuzzy (Zadeh, 1965) (TCF) tem sido bastante aplicada a fim de obter SRIs flexíveis (Zadrozny e Nowacka, 2008; Lopez-Herrera et al., 2009; Akinribido et al., 2011; Yan et al.,

2012).

Portanto, o grande volume de documentos disponíveis digitalmente, a imprecisão e incerteza inerente aos documentos e a exigência de interpretação dos grupos de documentos, resultaram em novos requisitos, com seus respectivos desafios científicos, os quais são objetivos de pesquisa deste doutorado.

Partindo-se desse contexto, o cenário motivacional, a hipótese e o objetivo que guiam este doutorado são apresentados a seguir.

## 1.1 Motivação, hipótese e objetivo

A TCF possibilita a definição de técnicas de recuperação capazes de modelar, até certo ponto, a subjetividade humana em termos de estimativa da relevância parcial de documentos que atendem às necessidades dos usuários como, por exemplo, documentos que resultam de uma consulta na web (Kraft et al., 2006). A TCF também permite a representação de conceitos vagos expressos por meio de termos linguísticos como, por exemplo, temperatura alta, custo baixo ou clima frio, oferecendo mecanismos mais poderosos para representação do conhecimento (Klir e Yuan, 1995). Entre estes mecanismos, o agrupamento fuzzy de documentos destaca-se como mecanismo associativo para capturar a inerente imprecisão e incerteza dos documentos dentro de uma coleção.

Agrupamento de documentos, de maneira geral, é aplicado no contexto de RI porque se existe um documento relevante para a busca requisitada que é pertencente a um determinado grupo, então é provável que outro documento deste mesmo grupo também seja relevante (Bordogna e Pasi, 2011). A suposição de que conteúdos de documentos de um mesmo grupo são similares ocorre devido ao processo de agrupamento, o qual busca organizar documentos que referem-se a um mesmo assunto em um mesmo grupo.

Por outro lado, o agrupamento fuzzy de documentos propicia a alocação de documentos em mais de um grupo ao mesmo tempo, já que existe a possibilidade de que documentos pertencentes a grupos distintos também apresentem características semelhantes. Para observar como esse mecanismo auxilia no tratamento de imprecisão e incerteza em documentos, considere como exemplo o documento cujo título é “*Extração de Regras de Redes Neurais via Algoritmos Genéticos*” (Santos et al., 1999), o qual aborda os tópicos: *Redes Neurais* e *Algoritmos Genéticos*. Considerando que cada tópico é identificado por descritores de grupos de documentos encontrados na coleção de documentos, da qual esse documento faz parte, o mesmo pode estar em dois grupos distintos: o grupo que representa o tópico *Redes Neurais* ou o grupo que representa o tópico *Algoritmos Genéticos*. No entanto, o documento citado como exemplo aborda os dois tópicos simultaneamente, dificultando a alocação exata desse documento em somente um dos grupos. Logo, cada grupo deve possuir uma descrição sucinta, ou seja, descritores que identifiquem tópicos da coleção de documentos e que permitam auxiliar o usuário na sua busca pela informação contida nos documentos. Essa organização em tópicos facilita a busca pela informação de

interesse, obtendo-se uma visão complementar ao modelo baseado em uma simples lista ordenada de acordo com a relevância dos documentos. No entanto, o agrupamento de documentos tem desafios e requisitos específicos. Portanto, tanto o agrupamento fuzzy de documentos, quanto a extração de descritores de grupos que identifiquem adequadamente os tópicos representados por cada grupo de documentos, são tarefas chave para a obtenção de flexibilidade na organização de documentos.

Existem na literatura várias abordagens para organização de documentos por meio de agrupamento fuzzy (Lee, 2001; Horng et al., 2005; Rodrigues e Sacks, 2005; Bordogna et al., 2006; Kozielski, 2007; Saraçoglu et al., 2007, 2008; Chowdhury e Bhuyan, 2010; Hüllermeier, 2011). Essas abordagens apresentam boas estratégias, as quais podem ser utilizadas para obter flexibilidade na organização de documentos pelo tratamento de imprecisão e incerteza inerentes aos documentos. No entanto, os autores possuem como foco principal de suas abordagens o agrupamento fuzzy de documentos e destacam a descrição de grupos como um problema em aberto, indicando a necessidade de extração de descritores para que seja possível compreender os grupos de documentos encontrados.

Geralmente, as abordagens apresentadas na literatura para a organização flexível de documentos são avaliadas em termos do desempenho do agrupamento de documentos, enquanto os descritores de grupos são extraídos utilizando alguma heurística sobre um conjunto pequeno de documentos, realizando assim, uma avaliação simples sobre o significado dos grupos extraídos. No entanto, uma apropriada extração e avaliação de descritores de grupos que considere toda a coleção e seja fundamentada em critérios bem definidos é importante porque os descritores são termos representantes da coleção que identificam os tópicos abordados nos documentos.

A extração de bons descritores é um problema desafiador, pois coleções de documentos são geralmente representadas por uma grande quantidade de termos, isto é, por um espaço de características de alta dimensionalidade. Além disso, em aplicações em que o agrupamento fuzzy é utilizado para a organização flexível de documentos, uma descrição apropriada dos grupos obtidos é tão importante quanto um bom agrupamento, pois, neste tipo de agrupamento, um mesmo descritor pode indicar o conteúdo de mais de um grupo, uma vez que no agrupamento fuzzy um mesmo documento pode ser compatível com mais de um grupo (Feldman e Sanger, 2007).

Sendo assim, dado que a falta de uma descrição apropriada para os grupos obtidos dificulta a interpretação desses grupos (Anaya-Sánchez et al., 2008), este doutorado beneficia-se deste cenário motivacional e é guiado pela seguinte hipótese:

*A extração de descritores de grupos fuzzy de documentos viabiliza a organização flexível de documentos, a qual permite que usuários de sistemas de recuperação de informação accessem o conteúdo dos documentos organizados considerando a imprecisão e incerteza típicas de situações reais.*

Partindo-se dessa hipótese, este doutorado tem como principal objetivo:

*Investigar e desenvolver métodos para a extração de descritores de grupos fuzzy que viabilizem a organização flexível de documentos.*

Dados esta hipótese e objetivo, este doutorado apresenta como contribuição principal a organização flexível de documentos pela proposta de investigação e desenvolvimento de métodos para a extração de descritores de grupos fuzzy. Por meio dessa organização, é possível alocar documentos a múltiplos grupos simultaneamente, respeitando as relações entre seus tópicos abordados e considerando a imprecisão e incerteza inerentes aos documentos. Partindo-se dessa contribuição, outras contribuições específicas da extração de descritores de grupos fuzzy são apresentadas a seguir.

## 1.2 Contribuições

De maneira geral, a inovação da pesquisa desenvolvida neste doutorado consiste no estudo, proposta, desenvolvimento, avaliação e aplicação de métodos para extração de descritores de grupos fuzzy que viabilizam a organização flexível de documentos. Neste contexto, quatro principais contribuições foram obtidas neste doutorado.

A primeira contribuição consiste da proposta e desenvolvimento do método SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*). Por meio desse método, descritores de grupos fuzzy *flat* são extraídos após o processo de agrupamento fuzzy independentemente do algoritmo de agrupamento fuzzy utilizado, visando identificar tópicos da organização flexível de documentos.

A segunda contribuição consiste da proposta e desenvolvimento do método SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*). Por meio desse método, descritores de grupos fuzzy *flat* também são extraídos após o processo de agrupamento fuzzy. Porém, o grau de pertinência dos documentos em cada grupo, obtidos do agrupamento fuzzy, é utilizado diretamente como fator de ponderação dos termos candidatos a descritores.

A terceira contribuição consiste da proposta e desenvolvimento do método HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*). Por meio desse método, descritores de grupos fuzzy hierárquicos são extraídos após o processo de agrupamento hierárquico fuzzy, identificando tópicos da organização hierárquica flexível de documentos.

Por fim, uma quarta contribuição foi obtida neste doutorado pela aplicação do método SoftO-FDCL no contexto do programa de educação médica continuada canadense, reforçando a utilidade e aplicabilidade da organização flexível de documentos.

Além dessas quatro principais contribuições, foram obtidas duas outras contribuições que não fazem parte da proposta principal deste doutorado, mas que foram importantes para o entendimento do problema abordado no mesmo: a primeira contribuição consiste na representação de documentos por meio de agrupamento fuzzy e a geração de regras

fuzzy a partir desta; e a segunda contribuição está relacionada à aplicação de regras fuzzy no nível da consulta de um Sistema de Recuperação de Informação (SRI).

Cada uma destas contribuições são apresentadas e discutidas ao longo desta tese, a qual é organizada conforme apresentado na próxima seção.

### 1.3 Organização da tese

Para mostrar a validação da hipótese, cumprindo com o objetivo deste doutorado, esta tese está dividida em 6 capítulos:

**Capítulo 2.** Neste capítulo, são apresentados os fundamentos da organização de documentos por meio do agrupamento de documentos, bem como a descrição de algoritmos de agrupamento fuzzy bem conhecidos na literatura. Uma revisão da literatura acerca da extração de descritores de grupos é também apresentada neste capítulo.

**Capítulo 3.** Neste capítulo, é apresentada uma abordagem proposta para organização flexível de documentos. Para tanto, a contextualização do problema e os experimentos preliminares que levaram à abordagem proposta também são apresentados.

**Capítulo 4.** Neste capítulo, são apresentados os métodos propostos para extração de descritores de grupos importantes para a organização flexível de documentos, bem como os resultados experimentais obtidos a partir da avaliação desses métodos.

**Capítulo 5.** Neste capítulo, é apresentada uma aplicação do método SoftO-FDCL no contexto do programa de educação médica continuada canadense.

**Capítulo 6.** Por fim, neste capítulo, são apresentados um resumo das contribuições deste doutorado, as parcerias com grupos de pesquisa de outras instituições, as quais enriqueceram a pesquisa desenvolvida neste doutorado, as limitações encontradas e os trabalhos futuros.

## CAPÍTULO

## 2

# Fundamentos da Organização de Documentos

---

## 2.1 Considerações iniciais

A organização de documentos é de bastante importância para a recuperação de informação textual. Tradicionalmente, métodos de agrupamento são utilizados para auxiliar a organização de documentos em grupos, cujos documentos similares e relacionados a um mesmo tema são alocados juntos em um mesmo grupo.

Nesse contexto, os fundamentos da organização de documentos por meio de agrupamento, bem como da extração de descritores de grupos, são apresentados neste capítulo.

## 2.2 Agrupamento de documentos

O agrupamento de documentos é muito utilizado para se obter conhecimento útil sobre os documentos (Feldman e Sanger, 2007; Manning et al., 2008; Baeza-Yates e Ribeiro-Neto, 2011). Para tanto, em uma etapa anterior ao agrupamento de documentos, no pré-processamento de documentos, os mesmos devem ser estruturados de maneira a torná-los processáveis pelos algoritmos de agrupamento.

A estruturação mais comum para documentos é a representação destes em um espaço vetorial no formato de tabelas atributo-valor. Considerando que cada linha dessas tabelas corresponde a um documento da coleção e cada coluna corresponde a um termo presente em toda a coleção de documentos, estas tabelas são usualmente nomeadas matriz documentos-termos. Assim, a cada célula da matriz é associada uma medida, como a medida binária, que indica a presença ou não de um termo em um documento; a frequência de um termo em um documento; e a frequência ponderada de um termo em um documento em função de sua distribuição ao longo da coleção.

Geralmente, os termos presentes nas matrizes documentos-termos são previamente

analisados e tratados. Em um esforço inicial, busca-se desconsiderar dos documentos termos que não representem conhecimento útil, via eliminação de *stopwords*, as quais são palavras não relevantes na análise dos textos, sendo geralmente constituídas por preposições, pronomes, artigos, interjeições, dentre outras. Posteriormente, busca-se identificar similaridades de significados entre palavras, como em casos de variações morfológicas ou de palavras sinônimas. Para tal, pode-se reduzir uma palavra à sua raiz por meio de processos de *stemming* ou reduzir as palavras ao seu lema (lematização). O efeito do uso de diferentes formas de extração de termos, das quais destacam-se as duas anteriormente descritas, *stemming* e lematização, bem como a substantivação, a qual visa reduzir o termo à sua forma semelhante àquela própria de um nome, foi explorado por Conrado (2009).

A matriz documentos-termos referente à coleção de documentos é inherentemente esparsa e de alta dimensionalidade, o que, por vezes, pode tornar o processo de análise computacionalmente muito custoso ou mesmo inviável, além de afetar negativamente o resultado de alguns algoritmos de extração de conhecimento. É vital para o processo de análise, portanto, selecionar os termos mais relevantes da coleção de documentos, tornando o conjunto de termos com o qual se trabalha mais conciso, porém não menos representativo em relação ao conjunto original.

Sendo assim, considere que uma coleção de documentos  $D$  é formada por  $n$  documentos, no qual  $d_i$  é um documento da coleção, com  $1 \leq i \leq n$ . Um documento  $d_i$  é representado por um vetor de valores relacionados aos  $k$  termos representativos da coleção, tal que  $d_i = [\sigma(t_1, d_i), \sigma(t_2, d_i), \dots, \sigma(t_k, d_i)]$ , no qual  $t_j$  é um termo, com  $1 \leq j \leq k$ , e  $\sigma(t_j, d_i)$  é a medida de associação entre um termo  $t_j$  e o documento  $d_i$ , a qual pode ser uma medida binária, que indica a presença ou não do termo  $t_j$  no documento  $d_i$ ; a frequência do termo  $t_j$  no documento  $d_i$ ; ou a frequência ponderada do termo  $t_j$  no documento  $d_i$  em função de sua distribuição ao longo da coleção  $D$ .

Como exemplo de representação de documentos, observa-se na Tabela 2.1 uma coleção  $D$ , a qual possui 3 documentos,  $\{d_1, d_2, d_3\}$ , e 5 termos representativos da coleção,  $\{\text{saúde}, \text{política}, \text{educação}, \text{esporte}, \text{segurança}\}$ , ou seja  $n = 3$  e  $k = 5$ . As células da tabela são compostas por valores de frequência dos termos nos documentos. Por exemplo, o termo “saúde” ocorre 2 vezes no documento  $d_1$ .

Tabela 2.1: Exemplo de matriz documentos-termos de uma coleção de 3 documentos

| Documentos | Termos representativos da coleção |          |          |         |           |
|------------|-----------------------------------|----------|----------|---------|-----------|
|            | saúde                             | política | educação | esporte | segurança |
| $d_1$      | 2                                 | 3        | 0        | 2       | 1         |
| $d_2$      | 1                                 | 2        | 4        | 0       | 0         |
| $d_3$      | 3                                 | 0        | 1        | 3       | 0         |

Uma vez representados os documentos por meio da matriz documentos-termos, o agrupamento de documentos visa organizar em um mesmo grupo documentos similares. No agrupamento, a distribuição dos documentos em grupos é realizada de acordo com as

características próprias da coleção de documentos. Sendo assim, o parâmetro chave de um algoritmo de agrupamento é a medida de similaridade entre dois documentos. Essa medida influencia diretamente o processo de agrupamento. Diferentes medidas, levam a diferentes agrupamentos.

Quando os objetos a serem agrupados são representados por atributos contínuos, a medida de distância mais comum utilizada para medir a similaridade entre esses objetos é a distância Euclidiana. No entanto, quando os objetos a serem agrupados referem-se a documentos, a matriz documentos-termos que representa os documentos é naturalmente esparsa, *i.e.*, os vetores que representam os documentos possuem muitos termos cujos valores de frequência são iguais a zero. Portanto, a medida de similaridade mais comum para este tipo de objeto é o coeficiente de similaridade de cosseno, a qual não considera os termos que não ocorrem nos documentos e, portanto, desconsidera os valores iguais a zero dos vetores que representam os documentos, considerando apenas o ângulo formado entre eles. O coeficiente de similaridade de cosseno é apresentado na Equação (2.1), com a qual mede-se a similaridade  $sim(\mathbf{d}_1, \mathbf{d}_2)$  entre os documentos  $\mathbf{d}_1$  e  $\mathbf{d}_2$ . A similaridade entre dois documentos pode variar entre 0 e 1, no qual 1 indica que o ângulo entre dois documentos é de zero grau e que, portanto, eles são muito similares.

$$sim(\mathbf{d}_1, \mathbf{d}_2) = \cos\theta = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{|\mathbf{d}_1| |\mathbf{d}_2|} \in [0, 1] \quad (2.1)$$

De maneira geral, o processo de agrupamento tem por objetivo organizar os documentos mais similares entre si dentro de um mesmo grupo e documentos que apresentam máxima dissimilaridade dentro de grupos diferentes. Como resultado desse processo, pode-se obter grupos *flat*, nos quais os grupos são isolados, ou hierárquicos, nos quais os grupos apresentam alguma estrutura hierárquica. Além disto, pode-se obter grupos *hard*, cujos documentos são alocados em um único grupo, ou *soft*, cujos documentos podem pertencer a mais de um grupo com diferentes graus de pertinência (Manning et al., 2008; Baeza-Yates e Ribeiro-Neto, 2011).

O agrupamento *hard* é baseado na tradicional lógica booleana, pela qual um documento é alocado em um único grupo, sendo intolerante em termos de incorporar imprecisão e incerteza. Por outro lado, o agrupamento *soft* é baseado em técnicas que acrescentam melhorias à abordagem booleana, possibilitando que um documento seja alocado em mais de um grupo. Entre essas técnicas, destacam-se aquelas que são baseadas na Teoria de Probabilidades e na Lógica Fuzzy (Bordogna e Pasi, 2011). Técnicas baseadas na Teoria de Probabilidades representam e processam a incerteza considerando que a mesma é ocasionada pela não ocorrência de um determinado evento. Por outro lado, técnicas baseadas na Lógica Fuzzy consideram que a incerteza é ocasionada pela falta de significado das palavras, já que as mesmas são naturalmente imprecisas e dependentes do contexto no qual são utilizadas.

Considerando que descritores de grupos de documentos identificam tópicos da coleção de documentos, a abordagem proposta neste doutorado está relacionada à grupos *soft*,

uma vez que documentos podem referir-se a mais de um tópico e, portanto, apresentarem alguma similaridade com documentos de outros grupos. Além disso, considerando que os descritores de grupos são extraídos a partir de palavras representativas da coleção de documentos, o agrupamento fuzzy é uma técnica de agrupamento *soft* baseada na Lógica Fuzzy apropriada para o tratamento de imprecisão e incerteza em documentos.

### 2.2.1 Agrupamento fuzzy de documentos

Os algoritmos de agrupamento fuzzy mais comuns são os algoritmos Fuzzy C-Means (FCM) (Bezdek, 1981), Guztafson-Kessel (GK) (Guztafson e Kessel, 1979) e Gath-Geva (GG) (Gath e Geva, 1989). Com base nestes algoritmos, em especial o FCM, foi desenvolvida a maioria dos algoritmos de agrupamento fuzzy hierárquicos como, por exemplo, os algoritmos H<sup>2</sup>-FCM (*Hierarchical Hyper-spherical c-Means Algorithm*) (Rodrigues e Sacks, 2005), HFCM (*Hierarchical FCM in a stepwise discovery of structure in data*) (Pedrycz e Reformat, 2006) e *A Dynamic Hierarchical Fuzzy Clustering Algorithm for Information Filtering* (Bordogna et al., 2006).

Os algoritmos de agrupamento fuzzy são, de maneira geral, baseados em protótipo, ou seja, eles otimizam um conjunto de protótipos, um para cada grupo, o qual consiste de parâmetros de localização, tamanho ou formato do grupo. Cada protótipo, por sua vez, capta a distribuição de um grupo de objetos com base na semelhança entre os objetos e o protótipo de cada grupo ou a aproximação de sua localização (distância). Assim, diferentes algoritmos de agrupamento são distinguidos pelo protótipo e pela medida de distância entre os objetos e o grupo.

Dentre os algoritmos de agrupamento fuzzy mais conhecidos, os algoritmos FCM, Possibilístico C-Means (PCM) (Pal et al., 2005), o qual consiste de uma melhoria do algoritmo FCM, e HFCM são descritos detalhadamente a seguir.

### 2.2.2 Algoritmo de agrupamento Fuzzy C-Means

O algoritmo FCM é o mais comum dos algoritmos de agrupamento fuzzy, o qual consiste de uma generalização do algoritmo de agrupamento convencional K-Means (Kaufman e Rousseeuw, 1990).

O algoritmo FCM consiste de um processo iterativo que atualiza os protótipos dos grupos, definidos inicialmente a partir de uma pseudo partição fuzzy. No agrupamento *hard* utiliza-se o termo *partição* para a definição de cada grupo, uma vez que neste tipo de agrupamento os objetos são alocados exatamente em um único grupo. Já no agrupamento fuzzy utiliza-se o termo *pseudo participação*, pois não há uma divisão bem definida dos grupos, estando os objetos em mais de um grupo simultaneamente. A pseudo participação fuzzy inicial, geralmente, é composta pela distribuição aleatória de graus de pertinência dos objetos nos grupos definida previamente.

Assim, uma pseudo participação fuzzy é uma família de grupos fuzzy  $P = \{g_1, g_2, \dots, g_c\}$ ,

sendo  $c$  a quantidade de grupos, obtidas da coleção de documentos  $D = \{d_1, d_2, \dots, d_n\}$ , sendo  $n$  a quantidade de documentos, que satisfaz as Equações (2.2) e (2.3), nas quais  $\mu(d_i, g_l)$  é o grau de pertinência do documento  $d_i$  no grupo  $g_l$ , para  $1 \leq l \leq c$  (Klir e Yuan, 1995).

$$\sum_{l=1}^c \mu(d_i, g_l) = 1 \quad (2.2)$$

$$0 < \sum_{i=1}^n \mu(d_i, g_l) < n \quad (2.3)$$

Para exemplificar este processo, observa-se na Tabela 2.2 o resultado do agrupamento fuzzy da coleção de documentos apresentada na Tabela 2.1, no formato matriz documentos-grupos, a qual possui agora os mesmos  $n = 3$  documentos agrupados, na qual as células contém os graus de pertinência dos documentos nos  $c = 3$  grupos. Observa-se ainda que a soma dos graus de pertinência de um documento em todos os grupos é igual a 1, satisfazendo a restrição da Equação 2.2. Este exemplo também satisfaz a restrição apresentada na Equação 2.3, pela qual evita-se a construção de grupos vazios.

Tabela 2.2: Exemplo de matriz documentos-grupos de uma coleção de 3 documentos

| Documentos | Grupos |       |       |
|------------|--------|-------|-------|
|            | $g_1$  | $g_2$ | $g_3$ |
| $d_1$      | 0,2    | 0,5   | 0,3   |
| $d_2$      | 0,4    | 0,1   | 0,5   |
| $d_3$      | 0,6    | 0,2   | 0,2   |

Os  $c$  vetores de protótipos dos grupos,  $v_1, v_2, \dots, v_c$ , são calculados pela Equação (2.4), sendo  $m > 1$  um número real chamado fator de fuzificação (*fuzzifier*) ou expoente de ponderação. A exponenciação  $m$  das pertinências dos documentos a serem agrupados pode ser vista como uma função sobre os graus de pertinência que leva a uma generalização da função de erro mínimo quadrado aplicado no agrupamento *hard* (agrupamento que não permite sobreposição de grupos). Assim, o valor de  $m$  controla a fuzificação (*fuzziness*) do agrupamento fuzzy. Quando  $m \rightarrow 1$ , o FCM converge para o clássico k-means. Já quando  $m \rightarrow \infty$ , as fronteiras entre os grupos são mais “suaves”. Finalmente, a definição do valor de  $m$  é feita de acordo com o problema considerado, embora  $m = 2$  seja usualmente escolhido (Pedrycz e Gomide, 2007; Klir e Yuan, 1995).

$$v_l = \frac{\sum_{i=1}^n [\mu(d_i, g_l)]^m d_i}{\sum_{i=1}^n [\mu(d_i, g_l)]^m}, \forall l \leq c \quad (2.4)$$

O processo iterativo do FCM dá-se por meio da atualização da pseudo partição definida inicialmente. Esta atualização ocorre na tentativa de minimizar a distância,  $|d_i - v_l|$ , entre um determinado documento  $d_i$  e um protótipo de grupo  $v_l$ .

No agrupamento de documentos, essa distância é calculada a partir do coeficiente de similaridade de cosseno, Equação (2.5) e Equação (2.6), ao invés da distância euclidiana utilizada no algoritmo FCM original, uma vez que a matriz documentos-termos que representa os documentos é naturalmente esparsa e de alta dimensionalidade (Deng et al., 2010).

$$sim(\mathbf{d}_i, \mathbf{v}_l) = cos\theta = \frac{\mathbf{d}_i \cdot \mathbf{v}_l}{|\mathbf{d}_i| |\mathbf{v}_l|} \in [0, 1] \quad (2.5)$$

$$dist(\mathbf{d}_i, \mathbf{v}_l) = 1 - sim(\mathbf{d}_i, \mathbf{v}_l) \in [1, 0] \quad (2.6)$$

Sendo assim, os graus de pertinência de cada documento em cada grupo, ou seja, a pseudo partição inicial, são redefinidos pela Equação (2.7).

$$\mu(\mathbf{d}_i, g_l) = \frac{1}{\sum_{h=1}^c \left( \frac{\|\mathbf{d}_i - \mathbf{v}_h\|}{\|\mathbf{d}_i - \mathbf{v}_l\|} \right)^{\frac{1}{m-1}}} \quad (2.7)$$

O objetivo do FCM é minimizar a função objetivo  $J(P)$  sobre a pseudo partição  $P$ , conforme Equação (2.8), ou seja, minimizar a distância entre os documentos e os protótipos dos grupos.

$$J(P) = \sum_{i=1}^n \sum_{j=1}^c [\mu(\mathbf{d}_i, g_l)]^m \|\mathbf{d}_i - \mathbf{v}_l\| \quad (2.8)$$

O grau de pertinência  $\mu(\mathbf{d}_i, g_l)$  que o FCM atribui a um documento  $\mathbf{d}_i$  está relacionado à distância relativa do documento  $\mathbf{d}_i$  ao protótipo de grupo  $\mathbf{v}_l$ ,  $\forall l \leq c$ . Se um documento  $\mathbf{d}_i$  é igualmente distante de dois protótipos,  $\mathbf{v}_1$  e  $\mathbf{v}_2$ , o grau de pertinência do documento  $\mathbf{d}_i$  em cada grupo será o mesmo:  $\mu(\mathbf{d}_i, g_1) = 0,5$  e  $\mu(\mathbf{d}_i, g_2) = 0,5$ .

Nesse contexto, considere um dado ruidoso como um documento que está distante, porém igualmente distante, dos protótipos de dois grupos. Por meio do FCM, pode-se atribuir a este dado o mesmo grau de pertinência de um documento que está mais próximo do protótipo de um grupo. Esta situação é ilustrada na Figura 2.1, na qual os documentos  $\mathbf{d}_1$  e  $\mathbf{d}_2$  tem ambos os mesmos graus de pertinência, 0,5 nos grupos, embora o documento  $\mathbf{d}_1$  esteja mais próximo dos grupos do que o documento  $\mathbf{d}_2$ .

De acordo com Pal et al. (2005), esta situação decorre da noção básica de partição probabilística do conjunto de dados do FCM, o qual possui a restrição  $\sum_{l=1}^c \mu(\mathbf{d}_i, g_l) = 1$ , i.e., a soma dos graus de pertinência de um documento em todos os grupos devem ser igual a 1.

Para solucionar este problema, foi desenvolvido o algoritmo Possibilístico C-Means (PCM) (Krishnapuram e Keller, 1993), o qual relaxa a restrição do FCM considerando o valor absoluto da distância de um documento  $\mathbf{d}_i$  aos protótipos de grupos. Neste sentido, o grau de pertinência do documento  $\mathbf{d}_i$  em um grupo  $g_l$  obtido por meio do PCM deve ser interpretado como a tipicidade do documento  $\mathbf{d}_i$  com relação ao grupo  $g_l$ .

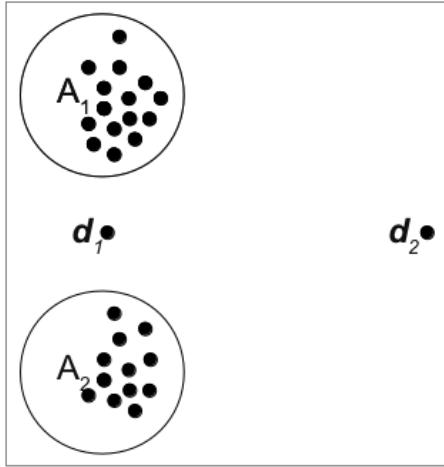


Figura 2.1: Posição dos documentos  $d_1$  e  $d_2$  com relação aos grupos  $A_1$  e  $A_2$  (Adaptado de Oliveira e Pedrycz (2007))

### 2.2.3 Algoritmo de agrupamento Possibilístico C-Means

De maneira similar ao FCM, o PCM é um processo iterativo que atualiza os protótipos de grupos definidos inicialmente a partir de uma pseudo partição. Essa atualização tenta minimizar a distância entre um documento e os protótipos de grupos. A atualização de protótipos de grupos do PCM é idêntica à atualização de protótipos do FCM, conforme apresentado na Equação (2.4). Assim, utilizando a medida de distância apresentada na Equação (2.6), o PCM realiza uma série de atualizações na pseudo partição definida inicialmente de acordo com a Equação (2.9) (Pal et al., 2005), no qual  $\varphi(\mathbf{d}_i, g_l)$  é tipicidade do documento  $\mathbf{d}_i$  com relação ao grupo  $g_l$ .

$$\varphi(\mathbf{d}_i, g_l) = \frac{1}{1 + \left( \frac{\|\mathbf{d}_i - \mathbf{v}_l\|}{\gamma_i} \right)^{\frac{1}{m-1}}} \quad (2.9)$$

A constante  $\gamma_i > 0$  é definida pelo usuário para minimizar o problema de singularidade do FCM e, assim, a distância  $\|\mathbf{d}_i - \mathbf{v}_l\|$  pode ser zero, relaxando a restrição do FCM (Equação 2.2). Mais detalhes sobre a definição de  $\gamma_i$  foram apresentados por Krishnapuram e Keller (1993).

O objetivo do PCM é minimizar a função objetivo  $K(Q)$  sobre a pseudo partição  $Q$ , de acordo com a Equação (2.10).

$$K(Q) = \sum_{i=1}^n \sum_{l=1}^c \varphi(\mathbf{d}_i, g_l)^m \|\mathbf{d}_i - \mathbf{v}_l\| + \sum_{l=1}^c \gamma_l \sum_{i=1}^n (1 - \varphi(\mathbf{d}_i, g_l))^m \quad (2.10)$$

A primeira parte,  $\sum_{i=1}^n \sum_{l=1}^c \varphi(\mathbf{d}_i, g_l)^m \|\mathbf{d}_i - \mathbf{v}_l\|$ , da função objetivo  $K(Q)$  do PCM corresponde à função objetivo  $J(P)$  do FCM, portanto, na ausência da segunda parte,  $\sum_{l=1}^c \gamma_l \sum_{i=1}^n (1 - \varphi(\mathbf{d}_i, g_l))^m$ , a otimização sem restrições leva a uma solução trivial  $\varphi(\mathbf{d}_i, g_l) = 0$ . Assim, a segunda parte de  $P(Q)$  atua como uma penalidade que tenta aproximar

$\varphi(\mathbf{d}_i, g_l)$  do valor 1 (Pal et al., 2005).

Além dos algoritmos de agrupamento *flat*, FCM e PCM, apresentados anteriormente, é possível obter uma hierarquia de documentos por meio do algoritmo de agrupamento Hierárquico Fuzzy C-Means (*Hierarchical Fuzzy C-Means*) (HFCM) (Pedrycz e Reformat, 2006), o qual é apresentado a seguir.

#### 2.2.4 Algoritmo de agrupamento Hierárquico Fuzzy C-Means

O algoritmo de agrupamento Hierárquico Fuzzy C-Means (*Hierarchical Fuzzy C-Means*) (HFCM) executa um modelo de agrupamento hierárquico divisivo cujos grupos são formados e refinados pela divisão dos nós da hierarquia (grupos individuais) por meio do algoritmo de agrupamento FCM.

Em contraste com conceitos e algoritmos existentes para agrupamento, os autores combinam em sua abordagem a ideia de agrupamento hierárquico com agrupamento baseado em objetivo fuzzy. Para tanto, o nível mais alto da hierarquia é considerado o contexto no qual o agrupamento é realizado. Essa restrição baseada em contexto é obtida por meio do algoritmo de agrupamento Fuzzy C-Means Condisional (*Conditional Fuzzy C-Means*) (CFCM) proposto por Pedrycz (1996). O algoritmo CFCM executa um agrupamento sensível a contexto, pelo qual o agrupamento é direcionado.

A Figura 2.2 ilustra como os algoritmos FCM e CFCM são aplicados no algoritmo de agrupamento fuzzy hierárquico proposto por Pedrycz e Reformat (2006).

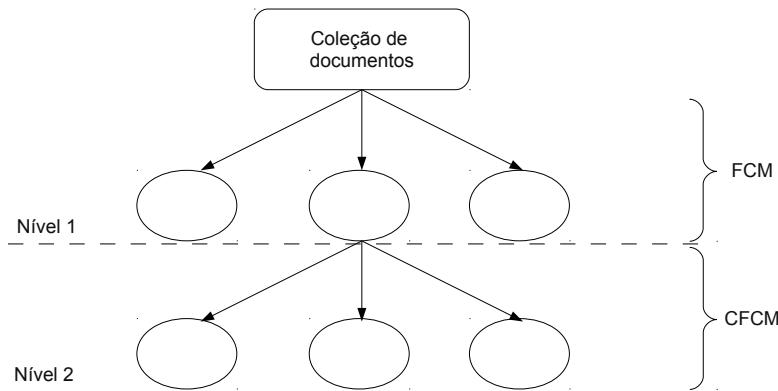


Figura 2.2: Agrupamento hierárquico fuzzy obtido da execução do algoritmo HFCM (Adaptado de Pedrycz (1996))

O algoritmo HFCM é definido como segue. Considere a coleção  $D = \{\mathbf{d}_1, \mathbf{d}_2 \dots \mathbf{d}_n\}$  composta por  $n$  documentos. O algoritmo HFCM tem início com o agrupamento de todos

os documentos em  $c$  grupos fuzzy por meio do algoritmo FCM. Desse agrupamento obtém-se uma pseudo partição fuzzy  $P = \{g_1, g_2, \dots, g_c\}$  e um conjunto de protótipos de grupo  $\mathbf{v}_1[1], \mathbf{v}_2[1], \dots, \mathbf{v}_c[1]$ , cujos valores entre colchetes correspondem ao nível da hierarquia no qual esses grupos estão. Sendo assim, o algoritmo FCM é utilizado pelo algoritmo HFCM para obter o primeiro nível da hierarquia, cuja largura no primeiro nível corresponde à quantidade de grupos  $c$ .

Para continuar gerando a hierarquia, o algoritmo HFCM mede o desempenho do agrupamento em um determinado nível por meio do índice  $I$ , o qual indica a qualidade da estrutura hierárquica a ser formada. Esse índice considera que quanto menores os valores dos protótipos de grupos, melhores são os grupos. Assim, quando um grupo possui valor  $I$  próximo de zero, diz-se que ele apresenta o mapeamento ideal e não precisa ser refinado. Por outro lado, o grupo que apresenta maior valor de  $I$  necessita ser expandido. Por exemplo, considere que o grupo  $g_l$  no nível 1 possui o maior valor de  $I$ , uma vez identificados os documentos que apresentam maior grau de pertinência no grupo  $g_l$ , considere  $D_l$  a coleção de documentos do grupo  $g_l$ . Para que esse grupo seja expandido, o algoritmo CFCM é aplicado sobre  $D_l$  e  $c$  novos grupos são obtidos, cujos protótipos são  $\mathbf{v}_{1l}[2], \mathbf{v}_{2l}[2], \dots, \mathbf{v}_{cl}[2]$ , essa notação refere-se aos  $c$  protótipos dos novos grupos expandidos do grupo  $g_l$  e o número entre colchetes indica o nível em que os mesmos encontram-se. O processo de avaliação dos grupos por meio do índice  $I$  é repetido considerando o critério de parada pré-definido pelo usuário, como por exemplo um valor mínimo de  $I$ .

Formalmente, o processo iterativo descrito anteriormente ocorre como segue. Primeiramente, o algoritmo FCM é executado para agrupar os documentos da coleção  $D = \{\mathbf{d}_1, \mathbf{d}_2 \dots \mathbf{d}_n\}$ , formando  $c$  grupos e obtendo a matriz de partição  $U[1] = [\mu(\mathbf{d}_i, g_l)[1]]$ ,  $l = 1, 2, \dots, c$ ;  $i = 1, 2 \dots, n$ , cujas células correspondem ao grau de pertinência de cada documento em cada grupo e o valor entre colchetes indica o nível no qual esses grupos se encontram. Esse agrupamento inicial guiará a formação do restante da hierarquia. Para tanto, o algoritmo de agrupamento CFCM é executado utilizando os documentos do grupo que apresenta maior valor de  $I$ . Antes de verificar como o valor de  $I$  é computado, observe como um determinado grupo é expandido utilizando o algoritmo CFCM.

O algoritmo HFCM utiliza como fator condicionante do algoritmo CFCM a restrição de que os grupos em um nível mais abaixo na hierarquia são influenciados pelo grupo que os originaram em um nível mais acima. Assim, a restrição do algoritmo de agrupamento FCM, a qual afirma que a soma dos graus de pertinência em todos os grupos deve ser igual a 1, conforme Equação (2.11), é modificada no algoritmo CFCM conforme Equação (2.12).

$$\sum_{l=1}^c \mu(\mathbf{d}_i, g_l) = 1 \quad (2.11)$$

$$\sum_{h=1}^c \mu(\mathbf{d}_i, g_h)[l, 2] = \mu(\mathbf{d}_i, g_l)[1], \quad \mathbf{d}_i \in D_l \quad (2.12)$$

De acordo com a Equação (2.12), tem-se que o grupo  $g_l$  é expandido nos grupos  $g_h$ ,  $h = 1, \dots, c$ . Assim, considere  $g_l$  o grupo a ser expandido e  $D_l$  a coleção de documentos composta pelos documentos que apresentam maior valor de pertinência no grupo  $g_l$ . Esse grupo será expandido em  $c$  novos grupos visando obter a matriz de partição  $U[2]$ , i.e., a matriz de partição obtida no nível 2, considerando a restrição de que a soma dos graus de pertinência de um documento  $\mathbf{d}_i \in D_l$  nos grupos  $g_h[2]$ ,  $h = 1, \dots, c$ , expandidos a partir do grupo  $g_l[1]$ , é equivalente ao grau de pertinência desse documento no grupo  $g_l[1]$ .

A execução do algoritmo CFCM realiza a atualização dos protótipos de grupos da mesma maneira que o algoritmo FCM. Porém, de forma a satisfazer a restrição apresentada na Equação (2.12), o algoritmo CFCM calcula a matriz de partição  $U[2]$  modificando a Equação (2.13), do algoritmo FCM, conforme Equação (2.14).

$$\mu(\mathbf{d}_i, g_l) = \frac{1}{\sum_{h=1}^c \left( \frac{\|\mathbf{d}_i - \mathbf{v}_h\|}{\|\mathbf{d}_i - \mathbf{v}_e\|} \right)^{\frac{1}{m-1}}} \quad (2.13)$$

$$\mu(\mathbf{d}_i, g_h)[l, 2] = \frac{\mu(\mathbf{d}_i, g_l)[1]}{\sum_{e=1}^c \left( \frac{\|\mathbf{d}_i - \mathbf{v}_h[2]\|}{\|\mathbf{d}_i - \mathbf{v}_e[2]\|} \right)^{\frac{1}{m-1}}}, \quad h = 1, \dots, c, \quad \mathbf{d}_i \in D_l \quad (2.14)$$

Conforme mencionado anteriormente, o algoritmo CFCM é executado visando expandir os grupos que apresentam menor desempenho  $I$ . Essa medida de desempenho considera que a capacidade de mapeamento de um grupo está relacionada aos protótipos dos grupos. Assim, assumindo que os protótipos de grupos correspondem a uma versão condensada dos documentos agrupados, cada documento  $\mathbf{d}_i$  pode ser representado pela combinação linear entre os protótipos de grupo e os graus de pertinência desses documentos nos grupos, conforme Equação (2.15) para os documentos alocados nos grupos do nível 1, e conforme Equação (2.16) para os documentos alocados no nível 2.

$$\hat{\mathbf{d}}_i = \sum_{l=1}^c \mu(\mathbf{d}_i, g_l)[1] \mathbf{v}_l[1] \quad (2.15)$$

$$\hat{\mathbf{d}}_i = \sum_{h=1}^c \tilde{\mu}(\mathbf{d}_i, g_h)[l, 2] \mathbf{v}_h[2], \quad \mathbf{d}_i \in D_l \quad (2.16)$$

no qual os graus de pertinência  $\tilde{\mu}(\mathbf{d}_i, g_h)[l, 2]$  são computados de forma a somarem 1, conforme restrição do algoritmo FCM.

Para documentos em um nível mais abaixo da hierarquia, a Equação (2.16) é também aplicada considerando os protótipos correspondentes. A representação  $\hat{\mathbf{d}}_i$ , deve ser o mais próximo possível do documento original  $\mathbf{d}_i$ . Assim, a soma das distâncias  $\|\hat{\mathbf{d}}_i - \mathbf{d}_i\|^2$  entre  $\hat{\mathbf{d}}_i$  e  $\mathbf{d}_i$  captura o mapeamento do grupo. Quanto menor o valor dessa soma, melhor a capacidade de aproximação dos grupos e, portanto, melhor a qualidade do grupo.

Sendo assim, o desempenho  $I$  de cada grupo é medido conforme Equação (2.17), para os grupos do nível 1, e conforme Equação (2.18), para os grupos do nível 2.

$$I_l[1] = \sum_{i:d_i \in D_l} \|\hat{d}_i - d_i\|^2 \quad (2.17)$$

$$I_h[2] = \sum_{i:d_i \in D_h} \|\hat{d}_i - d_i\|^2 \quad (2.18)$$

Assim, é computado o desempenho  $I$  de cada grupo em um determinado nível da hierarquia e será expandido aquele que possuir maior valor de  $I$ . A composição da hierarquia é, portanto, realizada de maneira sucessiva pela determinação do desempenho de todos os grupos em todos os níveis, expandindo aqueles que possuírem desempenho menor em grupos mais especializados. O critério de parada do algoritmo HFCM é estabelecido pelo usuário e depende do problema abordado. No entanto, os autores sugerem dois critérios: a determinação de um limiar relacionado ao valor de  $I$ , pelo qual a hierarquia é expandida até que se obtenha este valor terminal de  $I$ ; ou pela observação da estrutura de agrupamento obtida, já que se os protótipos de grupos tendem a ser alocados muito próximos uns dos outros significa que nenhuma nova estrutura está sendo revelada. Logo, não há necessidade de expansão do grupo. Assim, de acordo com os autores, o algoritmo HFCM é centrado no usuário (*user-centric*) significando que é sempre prudente que o usuário observe o passo-a-passo da estrutura revelada e decida se a hierarquia deve ou não continuar sendo expandida.

No processo de agrupamento fuzzy, a validação do agrupamento obtido é muito importante, pois o mesmo determina a estrutura de agrupamento mais adequada de acordo com os dados e o problema Gomez-Skarmeta et al. (1999). Não há um método geral que valide os métodos de agrupamento fuzzy, mas existem várias abordagens de validação que sugerem diferentes soluções para diferentes problemas (Bezdek, 1981; Bezdek e Pal, 1992). Uma medida de validação conhecida é a medida de validação Silhueta Fuzzy (*Fuzzy Silhouette - FS*) (Campello e Hruschka, 2006) para escolha do número adequado de grupos de documentos, uma vez que a mesma apresenta melhor equilíbrio entre eficácia e custo computacional, além de utilizar os graus de pertinência e os valores dos dados em sua função. Essa medida em geral é utilizada para escolha da quantidade de grupos porque o processo de partição de um conjunto de dados em um número apropriado de subconjuntos é uma tarefa difícil, uma vez que a divisão deste em muitos grupos pode levar à uma organização de difícil interpretação e/ou análise, enquanto a divisão em poucos grupos pode causar a perda de informação. Essa medida é apresentada detalhadamente a seguir.

### 2.2.5 Validação de agrupamento fuzzy

O método FS é uma extensão da versão simplificada do critério de largura média da silhueta (*Average Silhouette Width Criterion - ASWC*) (Kaufman e Rousseeuw, 1990), o qual foi originalmente desenvolvido para validação de agrupamento crisp. A ASWC é definida como segue (Adaptado de Campello et al. (2009)): considere um documento  $d_i \in \{d_1, d_2, \dots, d_n\}$ , pertencente a um grupo  $g_l \in \{g_1, g_2, \dots, g_c\}$ , para  $c$  igual a quantidade

de grupos. Em um agrupamento crisp, isto significa que o documento  $\mathbf{d}_i$  é mais próximo do protótipo do grupo  $g_l$  do que dos outros protótipos. Seja  $\alpha(\mathbf{d}_i, g_l)$  a distância média entre o documento  $\mathbf{d}_i$  e todos os outros documentos pertencentes ao grupo  $g_l$ . Seja também  $\lambda(\mathbf{d}_i, g_p)$  a distância média entre o documento  $\mathbf{d}_i$  e todos os documentos pertencentes à outro grupo  $g_p$ ,  $p \neq l$ . Considere o menor valor de  $\lambda(\mathbf{d}_i, g_p)$  computado sobre  $p = 1, \dots, c$ ,  $p \neq l$ , como a dissimilaridade  $\beta(\mathbf{d}_i, g_l)$  do documento  $\mathbf{d}_i$  ao grupo vizinho mais próximo de  $g_l$ . Desta maneira, a silhueta do documento  $\mathbf{d}_i$  é definida pela Equação (2.19)

$$S(\mathbf{d}_i) = \frac{\beta(\mathbf{d}_i, g_l) - \alpha(\mathbf{d}_i, g_l)}{\max\{\alpha(\mathbf{d}_i, g_l), \beta(\mathbf{d}_i, g_l)\}}, \quad (2.19)$$

na qual o denominador é utilizado apenas como um fator de normalização. Quanto maior o valor de  $S(\mathbf{d}_i)$  mais o documento  $\mathbf{d}_i$  é considerado pertencente ao grupo  $g_l$ . É importante ressaltar que se um grupo contém apenas um único documento  $S(\mathbf{d}_i)$ , então a silhueta deste objeto é  $S(\mathbf{d}_i) = 0$ . Esta restrição evita que, por meio da medida de silhueta, um agrupamento encontre um grupo para cada documento. Sendo assim, a silhueta média de todos os documentos é definida pela equação (2.20).

$$ASWC = \frac{1}{n} \sum_{i=1}^n S(\mathbf{d}_i) \quad (2.20)$$

O melhor agrupamento encontrado é aquele com maior ASWC, ou seja, com a menor distância intra-grupos  $\alpha(\mathbf{d}_i, g_l)$  e a maior distância inter-grupos  $\beta(\mathbf{d}_i, g_l)$ .

Para validação de um agrupamento fuzzy, deve-se considerar que, quando um documento  $\mathbf{d}_i$  é dito pertencente ao grupo  $g_l$ , o documento  $\mathbf{d}_i$  tem grau de pertinência maior no grupo  $g_l$  do que nos outros grupos. Mesmo assumindo este critério de definição exata da pertinência de um documento a um grupo, não há explicitamente a utilização da pseudo participação (graus de pertinência do objeto em todos os grupos) quando da utilização da medida ASWC para validação do agrupamento fuzzy.

Diante disto, o método Silhueta Fuzzy considera não apenas o grupo no qual um determinado documento possui maior grau de pertinência, mas também o grupo no qual ele possui o segundo maior grau de pertinência explicitando a importância do documento no grupo vizinho mais próximo. Sendo assim, a definição da silhueta do agrupamento fuzzy dá-se como segue na Equação (2.21)

$$FS = \frac{\sum_{i=1}^n (\mu_1(\mathbf{d}_i) - \mu_2(\mathbf{d}_i)) S(\mathbf{d}_i)}{\sum_{i=1}^n (\mu_1(\mathbf{d}_i) - \mu_2(\mathbf{d}_i))}, \quad (2.21)$$

na qual  $\mu_1(\mathbf{d}_i)$  e  $\mu_2(\mathbf{d}_i)$  são, respectivamente, o primeiro e segundo maiores graus de pertinência do documento  $\mathbf{d}_i$  nos grupos.

Sendo assim, a escolha da quantidade de grupos por meio dessa medida de avaliação de grupos fuzzy dá-se conforme Algoritmo 1, pela execução repetida do algoritmo FCM para diferentes valores de  $c$ , ou seja, diferentes quantidades de grupos. Calcula-se a medida FS

a cada execução e aquele que obtiver maior valor é escolhido como um número adequado de grupos. Para fins de otimização do custo computacional, conforme sugerido pelos autores do método, considerou-se apenas a similaridade dos documentos aos protótipos dos grupos vizinhos ao invés da similaridade entre um documento e todos os outros da coleção (Hruschka et al., 2004).

---

**Algoritmo 1:** Validação de agrupamento fuzzy de documentos (Adaptado de Campanello e Hruschka (2006))

---

**Entrada:** Coleção de documentos  $D = \{d_1, d_2 \dots d_n\}$ ;

Quantidade mínima de grupos  $c_{min}$ ;

Quantidade máxima de grupos  $c_{max}$ ;

Quantidade de grupos  $c \in [c_{min}, c_{max}]$ ;

Silhueta Fuzzy atual  $FS_{atual} = 0$ ;

Silhueta Fuzzy final  $FS_{final} = 0$ ;

**Saída:** Pseudo participação final  $P_{final}$ , a qual corresponde a melhor pseudo participação fuzzy  $P_c$ ;

**início**

$c = c_{min}$ ;

**repita**

Obter uma pseudo participação fuzzy  $P_c$  a partir de um algoritmo de agrupamento fuzzy sobre a coleção  $D$  com  $c$  grupos;

Calcular o  $FS_{atual}$  da pseudo participação  $P_c$  pela Equação (2.21);

**se**  $FS_{atual} > FS_{final}$  **então**

$FS_{final} = FS_{atual}$

$P_{final} = P_c$

**fim**

$c = c + 1$ ;

**até**  $c = c_{max}$ ;

**retorna**  $P_{final}$

**fim**

---

Além da validação dos grupos obtidos a partir de um processo de agrupamento, a extração de descritores de grupos é uma tarefa de muita importância quando da utilização de agrupamento para organizar documentos e extrair automaticamente o conhecimento embutido nos mesmos. Alguns dos principais métodos existentes na literatura para a realização da tarefa de extração de descritores de grupos, em especial, grupos fuzzy, são apresentados na seção a seguir.

## 2.3 Extração de descritores de grupos

Existem na literatura vários trabalhos que abordam a extração de descritores de grupos de documentos, os quais podem ser divididos em: abordagens baseadas em conhecimento interno e abordagens baseadas em conhecimento externo. Entende-se por conhecimento interno o conhecimento que pode ser adquirido diretamente dos documentos como, por exemplo, a medição de importância de um determinado termo pela frequência do mesmo na coleção dos documentos. Por outro lado, o conhecimento externo é adquirido por meio

de ferramentas ou métodos que auxiliam os desenvolvedores na escolha de termos mais compreensíveis aos usuários, ainda que o mesmo não ocorra na coleção de documentos. Um exemplo de ferramenta utilizada nesse contexto é a *Wikipedia*<sup>1</sup>, uma enciclopédia livre disponível na web, a qual pode ser utilizada para interpretação dos termos escolhidos como descritores de grupos.

As abordagens baseadas em conhecimento interno são, geralmente, abordagens que utilizam alguma medida estatística para escolha dos descritores de grupos, buscando pelos termos mais “importantes” na coleção de documentos e escolhendo-os como os termos representativos da coleção. Neste sentido, Geraci et al. (2006) propôs o uso de uma versão modificada da medida de ganho de informação (*information gain*) para identificar termos que melhor representam o conteúdo de um determinado grupo de documentos e que são menos representativos de outros grupos. Nos trabalhos apresentados por Osinski e Weiss (2005); Treeratpituk e Callan (2006), não apenas a frequência de termos é considerada, mas também a frequência de frases nos documentos, as quais correspondem à uma sequência de termos. Toda e Kataoka (2005) utilizam entidades nomeadas extraídas dos documentos como descritores de grupos de documentos. A abordagem *Scatter/Gather* (Cutting et al., 1992) obtém descritores considerando a frequência dos termos no título dos documentos.

Por outro lado, as abordagens baseadas em conhecimento externo fazem uso de algum conhecimento a priori para melhorar a geração dos descritores de grupos como, por exemplo, a abordagem proposta por Chin et al. (2006), no qual os autores utilizaram o banco de dados léxico *WordNet*<sup>2</sup>(Miller, 1995) para encontrar o significado dos termos escolhidos como descritores, bem como determinar relações semânticas entre eles. Neste mesmo contexto, Hotho et al. (2003) propõem a utilização do *WordNet* para reduzir a variância de documentos dentro de um mesmo grupo, uma vez que termos diferentes, mas semanticamente similares presentes em dois documentos podem contribuir para melhorar a taxa de similaridade entre estes documentos. Porém, segundo Hu et al. (2008), o uso do *WordNet* para extração de conhecimento externo pode não ser a melhor abordagem, dada sua capacidade limitada para desambiguação de palavras e simplicidade nas estratégias de enriquecimento da representação de texto, a qual é feita pela substituição dos termos do documento pelos seus sinônimos. Sendo assim, os autores propuseram a construção de um *thesaurus* com base em relações semânticas extraídas da *Wikipedia* e desenvolveram um *framework* que utiliza estas relações semânticas para reforçar a medida de similaridade tradicional no agrupamento de texto. Segundo Carmel et al. (2009), a *Wikipedia* é um recurso bem sucedido quando da sua utilização para a geração de descritores de grupos de documentos, embora o conhecimento interno dos documentos deva ser considerado para os casos em que a *Wikipedia* não abrange o conteúdo dos grupos. Assim, os autores propõem, diferentemente dos métodos que utilizam a *Wikipedia* para a identificação de categorias de documentos (Gabrilovich e Markovitch, 2007), um método que extrai dos

---

<sup>1</sup><http://www.wikipedia.org/>

<sup>2</sup><http://wordnet.princeton.edu/>

documentos um conjunto de termos considerados importantes para representação de cada grupo e, a partir destes são identificados os descritores finais dos grupos de documentos por meio da *Wikipedia*.

A tarefa de extrair descritores de grupos de documentos pode ser dividida ainda em duas possibilidades: *Description Comes First* (DCF), ilustrado na Figura 2.3, e *Description Comes Last* (DCL) (Zhang, 2009), ilustrado na Figura 2.4. Por meio de métodos do tipo DCF, também conhecido como baseado em rótulo (*label-based*), os descritores são extraídos na etapa de pré-processamento dos documentos antes, ou ao mesmo tempo, do agrupamento dos documentos. Por meio de métodos do tipo DCL, também conhecido como baseado em documentos (*document-based*), os descritores são extraídos após o agrupamento dos documentos.

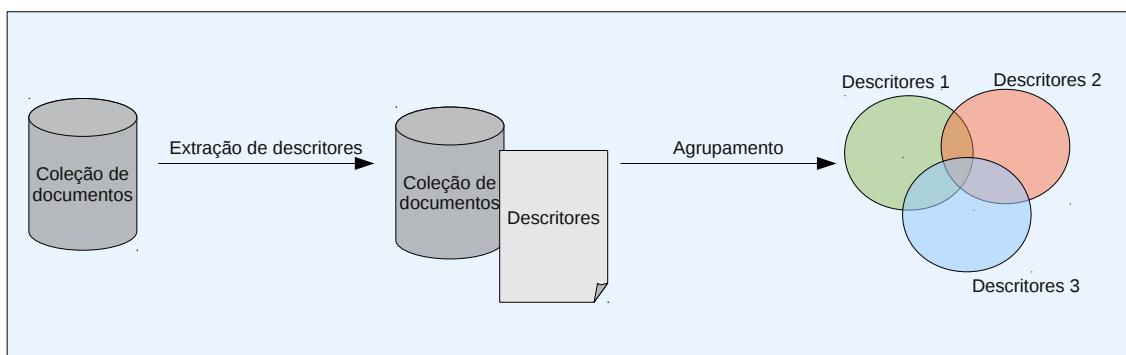


Figura 2.3: Abordagem DCF (*Description Comes First*)

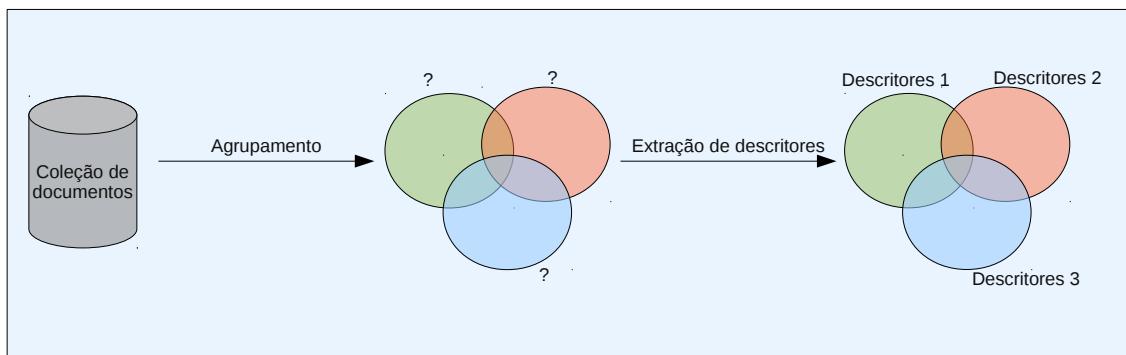


Figura 2.4: Abordagem DCL (*Description Comes Last*)

Existem na literatura vários métodos de extração de descritores de grupos, cujas características variam de acordo com o resultado do agrupamento: agrupamento *flat*, agrupamento hierárquico, grupos *hard*, grupos *soft*. Além disso, métodos do tipo DCF, geralmente, são compostos por algoritmos de agrupamento que tem a extração de descritores embutida em seu processo de agrupamento. Por outro lado, métodos DCL independem

do algoritmo de agrupamento utilizado. Alguns métodos e algoritmos mais conhecidos da literatura são apresentados na Tabela 2.3 de acordo com suas características: agrupamento obtido, grupos obtidos e tipo de método de extração de descritores de grupos. Por serem os mais citados na literatura, os mesmos citam uns aos outros em suas avaliações. A seguir, esses métodos são apresentados do mais antigo para o mais recente.

Tabela 2.3: Métodos de extração de descritores de grupos

| Autores                 | Nome      | Agrupamento      | Grupos | Tipo |
|-------------------------|-----------|------------------|--------|------|
| Zamir e Etzioni (1998)  | STC       | flat             | soft   | DCF  |
| Fung et al. (2003)      | FIHC      | hierárquico      | hard   | DCF  |
| Osinski e Weiss (2005)  | Lingo     | flat             | soft   | DCF  |
| Chen et al. (2010a)     | FMDC      | flat/hierárquico | hard   | DCF  |
| Chen et al. (2010b)     | $F^2$ IHC | hierárquico      | hard   | DCF  |
| Matsumoto e Hung (2012) | FTCA      | flat/hierárquico | soft   | DCL  |

Zamir e Etzioni (1998) propuseram o algoritmo de agrupamento STC, o qual é um algoritmo de agrupamento incremental cujos grupos obtidos são baseados em frases compartilhadas por documentos extraídos da web. Os autores realizaram avaliações comparativas com outros algoritmos de agrupamento nas quais foram observadas somente o desempenho do processo de agrupamento, não considerando a qualidade dos descritores de grupo obtidos.

Fung et al. (2003) propuseram um método chamado *Frequent Itemset-based Hierarchical Clustering* (FIHC), o qual produz uma hierarquia de tópicos para agrupar documentos. Esse método oferece bons resultados relacionados à redução da dimensionalidade, quantidade de grupos e fácil exploração com descritores de grupos obtidos da mineração de regras de associação cujos *itemsets* frequentes são os termos chave candidatos a descritores de grupos. Um *itemset* frequente é definido pelos autores como um conjunto de palavras que ocorrem juntas em uma quantidade mínima de documentos em um grupo. Uma vez obtidos os descritores candidatos, uma hierarquia de grupos é construída. Os autores realizaram avaliações comparativas com outros algoritmos de agrupamento hierárquico, mas não realizaram uma avaliação dos descritores de grupo.

Osinski e Weiss (2005) propuseram o método Lingo, que é um método de extração de descritores do tipo DCF. O autores sugerem o uso de um algoritmo de agrupamento baseado em rótulo (*label-based*) que identifica os conceitos abstratos que melhor descrevem uma amostra da coleção de documentos a serem agrupados. A representação desses conceitos é obtida a partir das frases mais frequentes dos documentos. Os conceitos, por sua vez, produzem um conjunto de descritores que determinam o conteúdo dos grupos e o algoritmo de agrupamento utilizado é orientado por esses descritores. Os autores realizaram apenas avaliações empíricas com um conjunto reduzido de grupos e descritores, cujos resultados foram comparados com os resultados obtidos pela utilização do método STC.

Chen et al. (2010a) propuseram o algoritmo FMDC (*Fuzzy-based Multi-label Document*

*Clustering*) que integra a mineração de regras de associação com a ontologia da WordNet de forma a explorar as relações semânticas fuzzy entre os termos que ocorrem nos documentos. Em sua abordagem, os autores extraem um conjunto de termos chave para representação inicial dos documentos os quais são enriquecidos pelo uso da WordNet. A partir dessa seleção de termos, um algoritmo de mineração de regras de associação fuzzy é executado para extrair um conjunto de *itemsets* frequentes altamente relacionados compostos por termos chave que serão considerados como candidatos a descritores de grupos. Uma vez selecionados os descritores os documentos são organizados em grupos definidos por esses descritores.

Segundo Chen et al. (2010a), a utilização de mineração de regras de associação como parte do processo de agrupamento é um caso especial de agrupamento, com o qual pode-se obter tanto agrupamento *flat* quanto hierárquico. Assim como o  $F^2$ IHC, a extração de descritores do algoritmo FMDC considera a possibilidade de descritores serem significativos para mais de um documento, *i.e.*, fuzzy. Porém, o algoritmo proposto obtém grupos *hard*, ou seja, documentos são alocados em um único grupo.

Chen et al. (2010b) propuseram uma extensão do algoritmo FIHC chamado *Fuzzy Frequent Itemset-Based Hierarchical Clustering* ( $F^2$ IHC) no qual um algoritmo de mineração de regras de associação é aplicado para descobrir um conjunto de *itemsets* frequentes que contém os termos chave candidatos a descritores de grupos. Uma vez extraídos os descritores candidatos, os documentos são organizados em uma estrutura hierárquica baseada nos descritores candidatos. A hierarquia de grupos é construída em um modelo *top-down*, o qual seleciona recursivamente os grupos pais no nível  $e - 1$  distribuindo os documentos dentro de seus grupos filhos no nível  $e$ . Embora a extração de descritores considere a possibilidade de descritores serem significativos para mais de um documento, o algoritmo proposto obtém grupos *hard*, ou seja, documentos são alocados em um único grupo.

Chen et al. (2010b) realizaram avaliações comparativas com outros algoritmos de agrupamento hierárquico, mas não realizaram uma avaliação dos descritores de grupo. Inclusive, os autores realizaram avaliações comparativas com o algoritmo FIHC, o qual também sugere a extração de descritores em uma etapa anterior ao agrupamento. No entanto, os descritores extraídos por cada um dos métodos não foram avaliados. Considerou-se apenas o desempenho do processo de agrupamento.

Matsumoto e Hung (2012) propuseram o algoritmo de agrupamento FTCA (*Fuzzy Transduction-based Clustering Algorithm*) para agrupar documentos extraídos da Web. Para tanto, os autores aplicam um modelo de relevância (*transduction-based relevance model* - TRM) que considera a relação local entre documentos. Resultados experimentais sobre o FTCA utilizando coleções de documentos reais e sintéticas mostraram que os resultados são favoráveis quando comparados com dois algoritmos bastante comuns: STC e Lingo. No entanto, embora os autores propuseram a extração de descritores de grupos após a execução do algoritmo FTCA, os autores não realizaram uma avaliação quantitativa dos descritores. Além disso, STC e Lingo são métodos DCF de extração de descritores de

grupos, não correspondendo ao algoritmo FTCA que sugere a utilização de um método DCL para extração de descritores.

Considerando os trabalhos citados anteriormente e outros trabalhos disponíveis na literatura, observa-se que o estado-da-arte aponta a existência de uma grande quantidade de algoritmos de agrupamento cuja extração de descritores é embutida no processo de agrupamento, *i.e.*, a maioria dos autores propõem métodos DCF para extração de descritores. Por esta razão, a qualidade dos descritores de grupos é medida pelo desempenho do algoritmo de agrupamento. Entretanto, apenas o agrupamento de documentos não descreve a organização de documentos, já que os grupos resultantes não possuem significado sem quem sejam extraídos descritores dos mesmos. Logo, quando a organização de documentos é obtida usando grupos de documentos, é preciso extrair descritores apropriados dos grupos e a qualidade dos mesmos deve ser avaliada considerando sua concisão, que significa que eles devem ser o mais curtos possíveis, mas suficientes para abordar o tópico do grupo; sua comprehensibilidade, também conhecida como transparência, que significa que eles devem mapear o conteúdo dos grupos; acurácia, que significa que eles devem refletir o tópico que corresponde ao grupo; e, distinção, que significa que eles devem ser mais frequentes em um grupo do que em outros (Zhang et al., 2009). Cada um destes pontos de avaliação dos descritores possuem desafios científicos específicos e são considerados conforme o domínio dos documentos a serem organizados.

Além disso, segundo Zhang (2009), métodos DCF apresentam um intervalo semântico entre a extração de descritores e os protótipos de grupos, o qual contradiz que “Primeiro agrupa-se, segundo descreve-se” (“*First clustering, second description*”), e diminui a habilidade explicativa dos descritores de grupo. Esse intervalo significa que não é possível identificar se os descritores influenciam no agrupamento ou o contrário. Métodos DCL são tipicamente menos complexos e capazes de obter tanto bom desempenho no agrupamento quanto descritores significativos. Além disto, separando o algoritmo de agrupamento da extração de descritores, diferentes algoritmos podem ser testados e usados.

## 2.4 Considerações finais

No contexto da pesquisa de doutorado apresentada nesta tese, considera-se o tratamento de imprecisão e incerteza de documentos por meio da organização flexível de documentos. Assumindo que a organização de documentos pode ser obtida por meio do agrupamento de documentos, neste capítulo foram apresentados os fundamentos básicos relacionados ao processo de agrupamento de documentos. Além disso, a fim de obter flexibilidade na organização de documentos, considera-se que um documento pode ser alocado em diferentes grupos. Assim, o agrupamento fuzzy de documentos é indicado como principal técnica utilizada para obter a pertinência de documentos em mais de um grupo. Ainda neste capítulo foram apresentados os principais algoritmos de agrupamento fuzzy, uma medida de validação de agrupamento fuzzy e uma revisão de trabalhos sobre

extração de descritores de grupos.

Especificamente neste doutorado, será abordada a extração de descritores de grupos a partir de conhecimento interno, uma vez que os algoritmos de agrupamento fuzzy utilizam a representação documentos-termos em seu processo e que, portanto, é natural que os descritores de grupos sejam obtidos desta mesma representação. Para tanto, neste doutorado é abordado o desenvolvimento de métodos DCL (*Description Comes Last*) para extração de descritores de grupos fuzzy.

No capítulo a seguir é apresentada a contextualização da pesquisa desenvolvida neste doutorado, os experimentos iniciais realizados e a abordagem proposta para organização flexível de documentos.



# Abordagem para Organização Flexível de Documentos

## 3.1 Considerações iniciais

O tratamento de imprecisão e incerteza na representação, organização e recuperação de documentos é um problema de pesquisa importante especialmente para a área de Recuperação de Informação (RI). Em geral, a imprecisão e a incerteza estão presentes em qualquer documento, pois diferentes leitores veem o documento sob diferentes perspectivas. Assim sendo, um determinado leitor pode organizar um conjunto de documentos com base em algum critério por ele definido, como por exemplo, pelos assuntos que ele considera mais importantes, enquanto outro leitor pode organizar o mesmo conjunto de documentos com base em algum outro critério, como por exemplo, pelos documentos que ele lê com mais frequência. Do mesmo modo, ao realizar uma consulta em uma máquina de busca, os documentos obtidos como resultado desta consulta podem ter graus de importância diferentes para diferentes leitores.

Nesse contexto, é abordado o tratamento de imprecisão e incerteza típicas de situações reais por meio de uma organização flexível de documentos. Nessa organização considera-se que um conjunto de documentos é organizado por tópicos e que um mesmo documento pode referir-se a diferentes tópicos. Uma proposta de abordagem para tal organização é apresentada neste capítulo.

## 3.2 Contextualização do problema

Geralmente, Sistemas de Recuperação de Informação (SRIs) são baseados no modelo booleano, apresentando limitações quanto à flexibilidade, uma vez que tal modelo é into-

lerante em termos de incorporar imprecisão e incerteza. Para superar essa limitação, dois outros modelos são comumente utilizados: modelo probabilístico e modelo flexível.

Por meio do modelo probabilístico de RI, os documentos a serem recuperados são listados em ordem decrescente de suas avaliações probabilísticas de relevância à informação que o usuário do SRI necessita. Muitas pesquisas tem sido feitas com o uso da teoria formal de probabilidade e da estatística a fim de avaliar, ou até mesmo estimar, a probabilidade de relevância dos documentos (Crestani et al., 1998). O problema em se estimar probabilisticamente a relevância de um determinado documento em uma coleção está na grande quantidade de variáveis envolvidas na representação dos documentos em comparação com a pequena quantidade de informação disponível acerca da relevância dos documentos. Assim, os modelos probabilísticos diferem principalmente pela forma como são estimadas estas informações relacionadas à probabilidade de relevância. Para tanto, os modelos de inferência probabilística aplicam conceitos e técnicas de lógica e inteligência artificial.

Por outro lado, o modelo flexível de RI provê melhorias nos SRIs por meio da aplicação de técnicas baseadas em Redes Neurais e Conjuntos Fuzzy. Tais modelos são referidos por modelos de recuperação de informação flexível (*Soft Information Retrieval*), em analogia à área de computação flexível (*Soft Computing*) (Crestani e Pasi, 1999; Kraft et al., 2006). Tem-se utilizado conjuntos fuzzy a fim de permitir a caracterização dos elementos de RI por meio do conceito de gradualidade. Com isto, os principais níveis de aplicação da teoria de conjuntos fuzzy para a RI estão na definição de extensões do modelo booleano, tanto em relação à representação de documentos quanto à consulta realizada por um usuário de um SRI, e à definição de mecanismos associativos, como por exemplo agrupamento fuzzy, o qual captura a inerente imprecisão e incerteza dos documentos dentro da coleção. Já as Redes Neurais ou modelos conexionistas de RI podem ser aplicadas tanto como um procedimento de aprendizado supervisionado quanto não supervisionado. No aprendizado supervisionado, durante a fase de aprendizado, a rede neural adapta os valores dos pesos nas conexões a fim de obter a saída desejada. Assim, no modelo neural cada documento é representado por uma unidade, cujo nível de ativação indica a relevância do documento. Porém, uma vez que os pesos entre documentos são identificados apenas por  $-1$  ou  $1$ , isso não reflete a importância da representação do conteúdo dos documentos. Já no aprendizado não-supervisionado, as redes neurais são utilizadas nos SRIs principalmente para agrupamento ou classificação de documentos ou termos, uma vez que não há *feedback* no processo de aprendizado e a rede neural é executada sobre informações locais e controles internos para capturar regularidades nos padrões de entrada.

definições a eventos; simplificação é sempre necessária. Finalmente, pode-se investigar probabilidade como uma descrição da crença, porém o tratamento é de abstração matemática (admitindo “eventos bem definidos”), que tem relação não muito clara com a realidade.

Neste doutorado, foi investigada a utilização de um modelo flexível para o trata-

mento de imprecisão e incerteza dos SRIs. Mais especificamente, foi considerado o uso de agrupamento fuzzy de documentos como mecanismo associativo para capturar a inerente imprecisão e incerteza dos documentos dentro de uma coleção. A fuzzificação é uma ambiguidade ocasionada por palavras, ou a falta de informação em significados, uma vez que termos requerem definições, e definições são impossíveis de obter, já que o uso de palavras para expressar a realidade é impreciso. O grau de imprecisão é medido como graus de fuzzificação. Assim, por meio de tal modelo, a flexibilidade dos SRIs pode ser considerada em seus três níveis: representação, organização e recuperação, conforme apresentado na Figura 3.1.

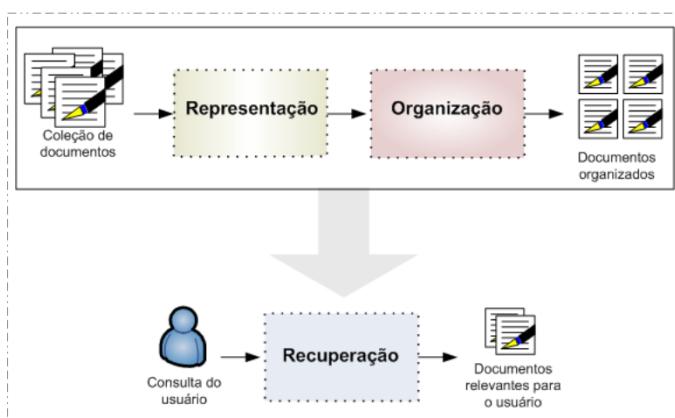


Figura 3.1: Níveis de um Sistema de Recuperação de Informação

O agrupamento de documentos, entretanto, não é uma tarefa simples porque um determinado documento pode, eventualmente, abordar diferentes tópicos, sendo necessário atribuí-lo a mais de um grupo. Assim, visando solucionar esse problema, técnicas de agrupamento fuzzy são aplicadas sobre documentos permitindo realizar uma organização flexível, alocando documentos à múltiplos grupos simultaneamente e respeitando as relações entre seus assuntos abordados (Saraçoglu et al., 2008; Hüllermeier, 2011). Os graus de pertinência obtidos no agrupamento fuzzy podem ser utilizados como medição da compatibilidade dos documentos com os grupos. Tal compatibilidade pode ainda ser representada por meio de termos linguísticos, aproximando-se da indicação de importância dada pelos seres humanos. Por exemplo, um documento pode ser “muito” ou “pouco” compatível com um determinado grupo.

Como exemplo de organização flexível de documentos por meio de agrupamento fuzzy, considere a Figura 3.2. Nesta figura, o documento, em formato de notícia, (a) “*Educação é deixada de lado, e povo é que paga*” pertence tanto ao grupo de notícias sobre *Educação* quanto ao grupo de notícias sobre *Política*. Neste mesmo exemplo, observa-se que a notícia (b) “*Vereadores retomam debate sobre coleta de lixo*” pertence aos três grupos de notícias: *Educação*, *Política* e *Saúde*. Já a notícia (c) “*Saúde abandonada*” encontra-se no grupo sobre *Política* e *Saúde*.

Pode-se observar ainda na Figura 3.2 que cada notícia tem compatibilidade maior com um grupo do que com outro(s). Por exemplo, a notícia “*Educação é deixada de lado, e povo*

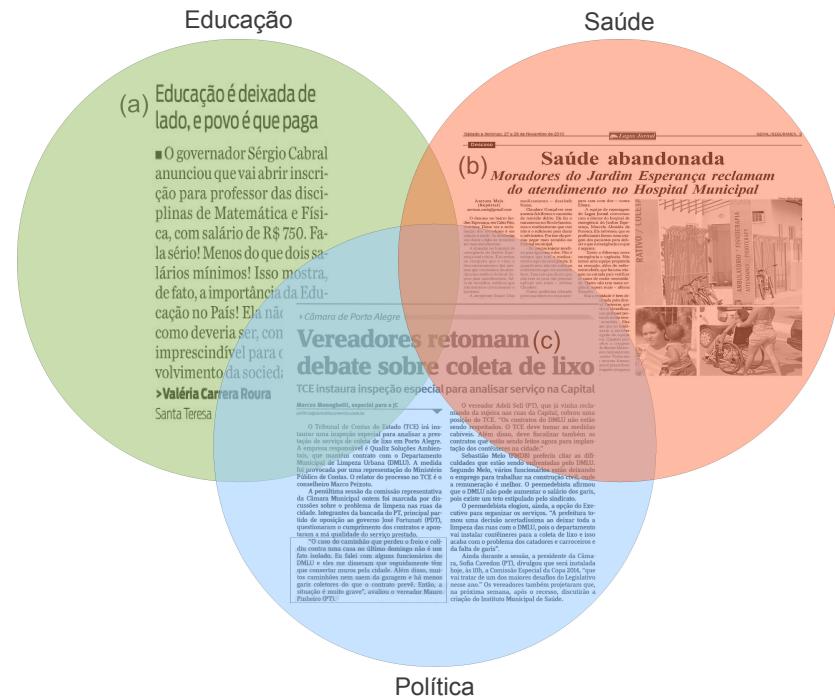


Figura 3.2: Exemplo de Organização Flexível com três grupos

é *que paga*” aborda com mais intensidade o assunto *Educação* do que o assunto *Política*, e não aborda o assunto *Saúde*. Esta intensidade pode ser o grau de compatibilidade de documentos com grupos, o qual é obtido por meio do grau de pertinência, quando executado o agrupamento fuzzy.

Nesse contexto, deu-se o desenvolvimento deste doutorado. Tendo como objetivo o tratamento de imprecisão e incerteza de documentos, foram realizadas explorações preliminares, descritas na Seção 3.3, que, juntamente com a avaliação do estado da arte, apresentado no Capítulo 2 e na Seção 3.4.1, apontaram a organização flexível como um problema em aberto, o qual foi o foco principal deste doutorado.

### 3.3 Explorações preliminares

O tratamento de imprecisão e incerteza é um problema amplo e em aberto. Sendo assim, foram feitas explorações iniciais de maneira a explorar o estado da arte do problema e propor uma abordagem adequada para o desenvolvimento de um SRI flexível.

A primeira exploração deu-se sobre a geração de regras fuzzy para classificação de documentos, uma vez que o processo de classificação de documentos é uma das principais tarefas dos SRIs. Além disso, as regras fuzzy tornam a modelagem do problema mais fiel e adequada ao mundo real. Essa primeira exploração, por sua vez, levou à exploração de um SRI que faz uso das regras fuzzy propostas em seu processo de recuperação.

Conforme apresentado na Seção 2.2 do Capítulo 2, a representação de documentos é usualmente realizada no formato de uma matriz documentos-termos obtida do pré-

processamento da coleção de documentos. Cada linha desta matriz representa um documento  $d_i$ , com  $1 \leq i \leq n$ , e cada coluna representa um termo  $t_j$ , com  $1 \leq j \leq k$ . Cada célula desta matriz é composta pela frequência  $\sigma(t_j, d_i)$  do termo  $t_j$  no documento  $d_i$ .

Neste projeto de doutorado, especificamente, a matriz documentos-termos é utilizada para extração de padrões por meio do agrupamento fuzzy de documentos. Este agrupamento reduz a dimensionalidade da matriz documentos-termos e modifica a representação dos documentos para a forma documentos-grupos. Cada célula da matriz documentos-grupos é composta pelo grau de pertinência  $\mu(d_i, g_l)$  do documento  $d_i$  no grupo  $g_l$ ,  $1 \leq l \leq c$ . Uma ilustração desse formato de representação de documentos de um SRI é apresentada na Figura 3.3.

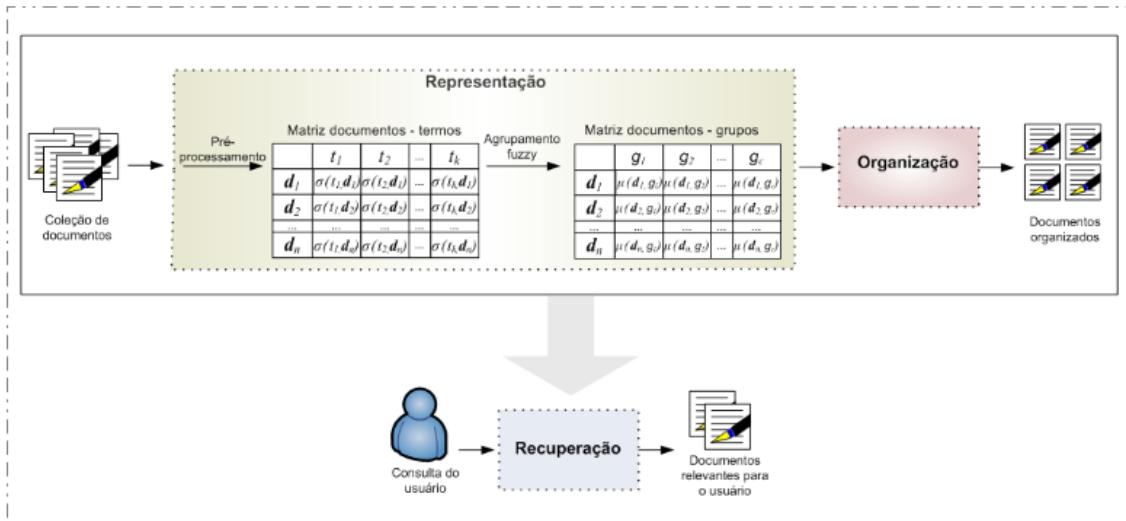


Figura 3.3: Nível de representação de documentos em um SRI

Portanto, considerando que os tópicos abordados pelos diferentes documentos da coleção podem ser representados por grupos, dada essa representação é possível explorar o nível de organização e recuperação em um SRI de modo a torná-lo flexível.

### 3.3.1 Geração de regras fuzzy para classificação de documentos

As regras fuzzy permitem a representação de conhecimento impreciso e possuem o seguinte formato:

*SE antecedente ENTÃO consequente*

Este formato tem a finalidade de estabelecer relações entre as variáveis que aparecem no antecedente, também chamado de condição ou premissa, e as que aparecem no consequente, também chamado de conclusão ou ação. Por exemplo, a regra:

*SE um documento é importante ENTÃO a possibilidade\_de\_recuperá-lo é alta.*

Esta regra estabelece uma relação entre as variáveis linguísticas *documento* e *possibilidade\_de\_recuperá-lo* usando os termos linguísticos *importante* e *alta*. Com o conceito de variáveis linguísticas, problemas naturalmente imprecisos e complexos passam a ser manipuláveis por computadores. A interpretação de um determinado conhecimento expresso na forma linguística torna-se passível de uma representação matemática por meio destas variáveis.

Sendo assim, variáveis linguísticas são variáveis cujos valores são palavras ou sentenças em linguagem natural em vez de números (Zimmermann, 1991). Elas são definidas sobre um determinado domínio, o qual é granularizado em termos linguísticos definidos por conjuntos fuzzy. O processo de granularização de um domínio de uma variável em conjuntos fuzzy define a chamada partição fuzzy.

Os conjuntos fuzzy são usados para modelar informação imprecisa, os quais podem ser a de maneira geral como, por exemplo, imagens e vídeos bordados como uma generalização da noção clássica de conjuntos. Nos conjuntos fuzzy, os elementos pertencem ao conjunto com um certo grau, que usualmente é um valor entre 0 e 1. Quando os graus de pertinência de um elemento assumem os valores 0 ou 1, tem-se o caso clássico de não pertinência total ou pertinência total do elemento ao conjunto, respectivamente. Assim, a definição de um conjunto fuzzy é obtida ampliando-se o contradomínio da função característica  $\{0, 1\}$ , da definição clássica de conjuntos, para o intervalo  $[0, 1]$ , de forma a atribuir o grau com que um elemento pertence a um conjunto fuzzy. A generalização da função característica passa a ser chamada de função de pertinência, a qual definirá o conjunto fuzzy, associando elementos de um dado conjunto universo  $U$  a números reais do intervalo  $[0, 1]$ .

Logo, um conjunto fuzzy  $a_q$  no conjunto universo  $U$  é caracterizado pela função

$$a_q : U \longrightarrow [0, 1]$$

e  $\mu(x, a_q)$  define o grau com que o elemento  $x$  pertence ao conjunto fuzzy  $a_q$  (Klir e Yuan, 1995).

Assim, uma proposição fuzzy é a parcela de informação básica que pode aparecer em uma regra e seu formato mais simples é:

$$X \text{ é } a_q$$

na qual  $X$  é uma variável linguística e  $a_q$  é um termo linguístico do conjunto de termos linguísticos  $A = \{a_1, a_2, \dots, a_o\}$ ,  $1 \leq q \leq |A|$ , da variável  $X$  que representa um conjunto fuzzy, o qual é caracterizado por uma função de pertinência.

Por exemplo, se  $X$  for a variável *documento* e  $a_q$  o termo *importante*, a proposição  $X \text{ é } a_q$  representa formalmente a parcela de informação do exemplo anterior que diz “*um documento é importante*”.

Utilizando proposições fuzzy, regras fuzzy são geradas de forma a serem utilizadas nos chamados Sistemas Fuzzy Baseado em Regras (SFBR), os quais são compostos por dois componentes principais: a Base de Conhecimento (BC) e o Mecanismo de Inferência

(MI) (Klir e Yuan, 1995). A BC é composta pela Base de Dados (BD), a qual contém as definições dos conjuntos fuzzy relacionados aos termos linguísticos usados nas regras fuzzy e pela Base de Regras (BR), que armazena o conjunto de regras que modelam um determinado problema. O MI é responsável pelo processamento das regras, o qual é realizado por algum método de raciocínio. Este consiste da aplicação de um procedimento de inferência para derivar conclusões a partir das regras e de fatos conhecidos.

O modo de operação de um SFBR, geralmente, consiste das seguintes etapas:

1. Transformação dos valores de entrada em conjuntos fuzzy, ou fuzificação. Para entradas numéricas este passo consiste no cálculo do grau de pertinência de cada valor de entrada no conjunto correspondente, e indica a compatibilidade da entrada com o antecedente de cada regra;
2. Agregação de antecedentes de cada regra por meio de operadores para o cálculo de conjunção fuzzy. O resultado obtido é chamado de Grau de Disparo da regra;
3. Aplicação da inferência com a derivação de resultados individuais para cada regra;
4. Combinação de possíveis saídas fuzzy;
5. Transformação do resultado fuzzy em um resultado preciso, processo conhecido como defuzificação. Esta etapa nem sempre ocorre, pois depende do tipo de problema em que será aplicado o Sistema Fuzzy.

Quando os SFBRs são desenvolvidos com o objetivo específico de executar a tarefa de classificação, tem-se os chamados Sistemas de Classificação Fuzzy (SCF), os quais utilizam métodos de raciocínio próprios para essa tarefa.

O formato usual de regras fuzzy para classificação é:

$$\text{SE } X \text{ é } a_q \text{ ENTÃO } Class$$

na qual  $X$  é uma variável linguística sobre o domínios  $\mathbf{X}$ , e  $a_q$  é um termo linguístico, que rotula um conjunto fuzzy definido sobre o domínio  $\mathbf{X}$ .

Os SCF utilizam métodos de raciocínio próprios para a tarefa de classificação utilizando regras fuzzy, cujos mais utilizados são: Método de Raciocínio Fuzzy Clássico (Chi et al., 1996) e Método de Raciocínio Fuzzy Geral (Ishibuchi et al., 1999).

No contexto deste doutorado, apresentado na Seção 3.2, assume-se que a representação dos documentos (Figura 3.3) propiciam o tratamento de imprecisão e incerteza de documentos. Assim, a geração de regras fuzzy a partir do agrupamento fuzzy de documentos é uma abordagem interessante para a classificação de documentos, uma vez que o agrupamento fuzzy de documentos reduz a dimensionalidade da matriz documentos-termos. Geralmente, a alta dimensionalidade da matriz documentos-termos dificulta a interpretabilidade de regras geradas a partir desta. Se fossem geradas regras a partir da matriz documentos-termos, as variáveis linguísticas que comporiam as regras corresponderiam a

cada um dos termos que representam toda a coleção de documentos. Gerando as regras a partir da matriz documentos-grupos, esta dificuldade é superada.

Uma ilustração do nível de organização de documentos em um SRI é apresentada na Figura 3.4. Nessa organização, regras fuzzy são geradas a partir do agrupamento fuzzy de documentos possibilitando o tratamento de imprecisão e incerteza para classificação de documentos.

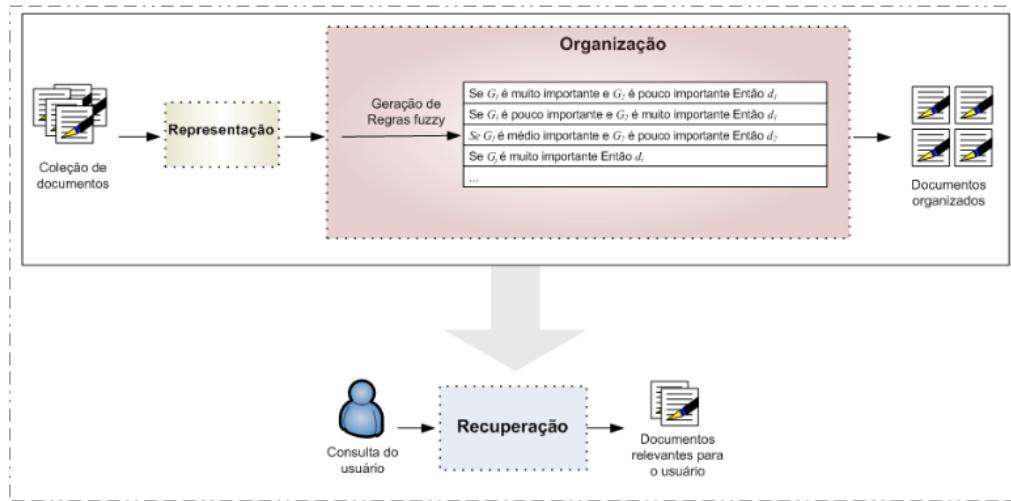


Figura 3.4: Nível de organização de documentos em um SRI

Nessa organização proposta, as regras fuzzy são geradas por meio do algoritmo Wang & Mendell (Wang e Mendel, 1992) após o agrupamento fuzzy de documentos obtido pela aplicação do algoritmo Fuzzy C-Means (FCM) (Bezdek, 1981), obtendo-se um SCF cujas regras assumem o seguinte formato:

$$\text{SE } G_1 \text{ é } a_1 \text{ E } G_2 \text{ é } a_2 \text{ E } \dots \text{ E } G_c \text{ é } a_c \text{ ENTÃO } Class$$

Nesse formato de regra,  $G_1, G_2 \dots G_c$  são variáveis linguísticas que representam os  $c$  grupos formados pelo agrupamento fuzzy de documentos, as quais foram granularizadas nos termos linguísticos  $A = \{a_1, a_2, a_3\}$ . Por exemplo, o grupo  $g_1$  é representado como uma variável linguística  $G_1$  granularizada nos termos linguísticos  $a_1 = Baixo$ ,  $a_2 = Médio$ ,  $a_3 = Alto$ , os quais são caracterizados por uma função de pertinência. Os termos *Baixo*, *Médio* e *Alto* referem-se aos graus de pertinência dos documentos nos grupos.

Os resultados obtidos com esta exploração preliminar mostraram que a classificação de documentos utilizando regras fuzzy apresenta um bom desempenho quando comparada com métodos para classificação bastante conhecidos na literatura: KNN, J48, Naive Bayes, OneR e SVM. Os resultados obtidos com o mecanismo proposto foram apresentados por Nogueira et al. (2010) e Nogueira et al. (2011b).

Por outro lado, a interpretabilidade das regras geradas a partir do agrupamento fuzzy de documentos é comprometida pela ausência de significado nos grupos. Entende-se por significado, os descritores de grupos que referem-se aos tópicos abordados pelos documentos agrupados. Por exemplo, na regra

SE  $G_1$  é *Alto* ENTÃO *Class* é *Redes Computadores*,

o grupo  $G_1$  não tem significado. Logo, não é possível interpretar a proposição  $G_1$  é *Alto*. Por outro lado, se o grupo  $G_1$  fosse composto pelos descritores {*rede de sensores*, *redes wireless*, *grades computacionais* ...}, seria possível a interpretação de que documentos do grupo  $G_1$  são compostos por palavras que correspondem aos descritores deste grupo. Além disso, se a ocorrência de tais palavras é *Alto*, então este documento corresponde ao tópico *Redes Computadores*. Sendo assim, a interpretabilidade de regras é importante porque possibilita a expressão do conhecimento, além de propiciar transparência e comprehensibilidade ao sistema (Luger, 2004).

Dessa exploração, portanto, observou-se a necessidade de extração de descritores de grupos de forma a melhorar a interpretabilidade dos grupos e, consequentemente, das regras fuzzy a partir do agrupamento fuzzy de documentos.

Um detalhamento da abordagem proposta para geração de regras fuzzy para classificação de documentos, bem como os resultados obtidos dessa exploração preliminar, são apresentados no Apêndice A.

Outra exploração para o tratamento de imprecisão e incerteza de documentos deu-se no nível da recuperação. Essa exploração faz uso da proposta de geração de regras fuzzy descrita anteriormente e é apresentada na próxima seção.

### 3.3.2 Recuperação de informação por meio de regras fuzzy

Considerando que a representação dos documentos por meio da matriz documentos-grupos obtidas do processo de agrupamento fuzzy reduz a dimensionalidade da matriz documentos-termos, optou-se por explorar o tratamento de imprecisão e incerteza no nível da recuperação de um SRI quando da análise da consulta realizada por um usuário.

A noção de medição da importância de uma palavra-chave na consulta a um SRI é amplamente defendida. Por outro lado, não é óbvio como isto deve ser medido e expressado. Usualmente, o intervalo [0,1] é adotado como faixa do grau de importância. Entretanto, é pouco realista esperar que esta medida possa ser precisamente expressa por um número. Logo, tem-se utilizado a Recuperação de Informação Fuzzy (RIF), a qual é uma extensão do modelo booleano e representa um documento como um conjunto de termos fuzzy tornando a descrição da informação contida nos documentos mais precisa (Radecki, 1979). A lógica fuzzy tem sido reconhecida como uma forma conveniente de modelagem do processo de RI, ou seja, interpretação dos conceitos de relevância de um documento com relação à consulta ou à importância de uma palavra-chave para a representação de um documento e/ou consulta.

Nesse contexto, a flexibilidade de um SRI pode ser obtida usando dois tipos de técnicas para refinamento de consultas (Manning et al., 2008; Baeza-Yates e Ribeiro-Neto, 2011): Métodos Globais e Métodos Locais. O primeiro tipo inclui técnicas para expandir ou reformular os termos da consulta (Chli e Wilde, 2006; Dae-Young e Choi, 2003). Esta

reformulação modifica a consulta para obter uma nova consulta que corresponda a outros termos semanticamente similares. Este tipo de técnica usualmente provê uma estratégia de *matching*, ou seja, uma estratégia que possibilite a correspondência entre os termos utilizados pelo usuário no momento da consulta e os termos que caracterizam a informação a ser recuperada. O segundo tipo, por sua vez, inclui técnicas que ajustam uma consulta de acordo com os documentos que inicialmente aparecem como resultado da consulta.

Para proporcionar flexibilidade a um SRI no nível da consulta, conforme ilustrado na Figura 3.5, foi desenvolvida em uma exploração preliminar uma estratégia de *matching* para recuperação de documentos com base em um critério de relevância definido pelo usuário.

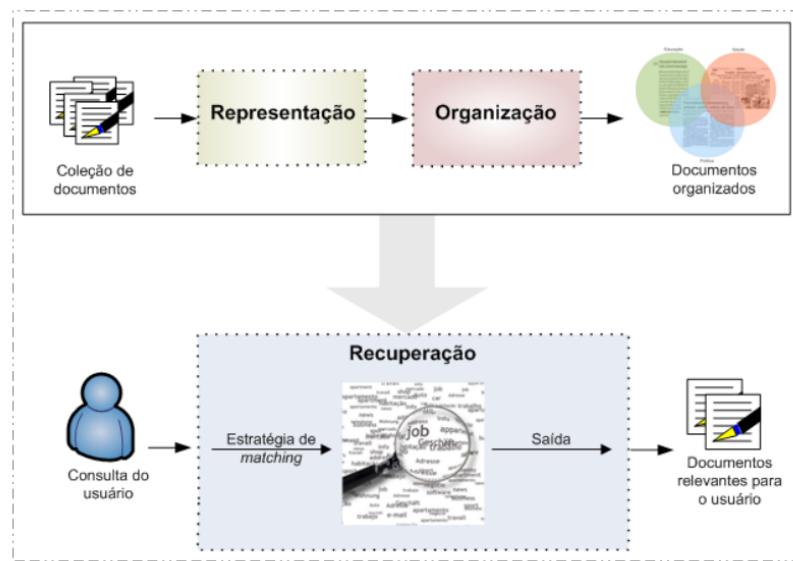


Figura 3.5: Nível de recuperação de documentos em um SRI

Para tanto, considere uma coleção de documentos e um conjunto de regras geradas a partir do agrupamento destes documentos. Considere ainda um conjunto de termos representativos da coleção e um conjunto de classes dos documentos definidos manualmente. Ao realizar a consulta, o usuário deve definir a relevância de cada um dos termos, bem como das classes dos documentos. Com isto, a estratégia de *matching* deve realizar as composições necessárias sobre as regras fuzzy para garantir que os documentos recuperados satisfaçam os critérios de relevância definidos pelo usuário.

Desta exploração observou-se dois problemas: a definição de termos a serem utilizados pelo usuário no momento da consulta e a utilização de coleções de documentos manualmente classificados.

O primeiro aspecto limita o espaço de busca do SRI, pois somente termos previamente selecionados podem ser utilizados na consulta. Além disto, no momento desta exploração, os termos utilizados na consulta eram selecionados por especialistas do domínio a partir de todos os termos obtidos do pré-processamento dos documentos. Esta seleção é custosa e inviável dado que, no geral, SRIs possuem coleções de documentos que crescem ao longo do tempo. O segundo aspecto limita a flexibilidade de um SRI, uma vez que, para a

estratégia de *matching* proposta, deve-se considerar documentos previamente classificados, o que não corresponde ao conceito de agrupamento fuzzy de documentos, cujo aprendizado é não-supervisionado e um documento pode ser compatível com mais de um grupo.

O funcionamento da estratégia de *matching* desenvolvida nesta exploração pode ser conferida detalhadamente no Apêndice B.

Já existem na literatura métodos que gerenciam a imprecisão e incerteza de um SRI no nível da consulta. Porém, elas dependem dos termos utilizados na consulta e da participação do usuário, conforme observado na estratégia proposta no experimento preliminar e em algumas abordagens existentes. Lopez-Herrera et al. (2009); Zadrozy e Nowacka (2008) desenvolveram modelos de RI utilizando variáveis linguísticas fuzzy para caracterizar a subjetividade da iteração com o usuário, assinalando valores qualitativos para os termos da consulta. Pasi (2002) apresenta algumas pesquisas relacionadas à consultas que tornam um SRI flexível, tais como o controle personalizado do processo de indexação pela inserção de restrições sobre a estrutura do documento, ou a introdução de níveis de importância e quantificadores fuzzy nas consultas. Lynn e Ng (2008) propuseram um modelo de IR baseado em conjuntos fuzzy que classifica documentos recuperados para qualquer consulta imprecisa usando uma “pontuação de imprecisão” dos documentos com base no significado das palavras. Por meio desta pontuação, os documentos que melhor correspondem aos diferentes significados dos termos usados na consulta são recuperados. Akinribido et al. (2011) propuseram um SRI baseado em ontologia fuzzy que determina a equivalência semântica entre termos da consulta e termos nos documentos pela relação de sinônimos dos termos das consultas com os termos dos documentos. Segundo Zadrozy e Nowacka (2009), o primeiro passo para a aplicação da lógica fuzzy em sistemas de RI é aplicar a lógica multivalorada ao invés da clássica, ou seja, binária. Logo, um documento é tratado como um conjunto fuzzy de palavras-chave, as quais são palavras com capacidade de descrever semanticamente o conteúdo de um documento. Essas palavras são utilizadas pelos SRIs para realizar pesquisas relacionadas a um assunto específico e a pertinência das palavras-chave reflete sua importância em representar o significado do documento.

Portanto, a partir das explorações preliminares realizadas e da revisão da literatura, concluiu-se que para que um documento seja satisfatoriamente recuperado por um SRI, considerando que a imprecisão e incerteza são típicas de documentos textuais, a coleção da qual ele faz parte deve ser organizada de maneira flexível, pois, em concordância com Chowdhury e Bhuyan (2010), a flexibilidade abordada somente no nível da consulta, limita um SRI.

### 3.4 Uma Abordagem Proposta para Organização flexível de documentos

Entende-se por organização flexível aquela em que documentos sobre assuntos diferentes podem apresentar características similares. Neste sentido, neste doutorado, deu-se

início à investigação da flexibilidade de um SRI no nível da organização de documentos por meio de agrupamento fuzzy e da descrição dos grupos obtidos.

Alguns trabalhos tem sido desenvolvidos de forma a obter flexibilidade na organização de documentos. Tais trabalhos fazem uso de agrupamento fuzzy como principal meio para o tratamento de imprecisão e incerteza inerentes à documentos. Alguns deles são brevemente apresentados na seção a seguir.

### 3.4.1 Trabalhos relacionados à organização flexível de documentos

Torra (2005) apresenta uma proposta de algoritmo de agrupamento fuzzy hierárquico, no qual alguns grupos são previamente definidos por meio do algoritmo de agrupamento Fuzzy C-Means (FCM) (Bezdek, 1981). Com isso, um processo iterativo é aplicado para a construção da hierarquia seguindo a estratégia *top-down*, na qual os grupos definidos anteriormente são particionados utilizando um agrupamento hierárquico divisivo.

Segundo Rodrigues e Sacks (2005), tópicos que caracterizam um dado domínio de conhecimento são algumas vezes associados uns aos outros e podem, também, ser relacionados à tópicos de outros domínios. Logo, documentos podem conter informações relevantes para diferenciar domínios em algum grau e descobrir relacionamento útil entre os domínios. Nessa linha, Rodrigues e Sacks (2004) propuseram uma modificação do algoritmo FCM para agrupamento de documentos que utiliza o coeficiente de similaridade de cosseno ao invés da distância euclidiana. Esta modificação, por sua vez, foi considerada por Rodrigues e Sacks (2005) para o desenvolvimento de um algoritmo de agrupamento fuzzy hierárquico chamado *Hierarchical Hyper-spherical c-Means Algorithm* ( $H^2$ -FCM) para construção de uma taxonomia de tópicos que explora a noção de similaridade assimétrica para organizar grupos fuzzy hierarquicamente formando uma hierarquia de tópicos significante baseada no centróide dos grupos.

De acordo com Bordogna et al. (2006), usuários de um SRI podem buscar por documentos considerando interesses específicos ou gerais de acordo com seu perfil. Ou seja, um determinado usuário pode realizar uma busca por um conjunto de documentos que abordam um determinado assunto de maneira geral, ou que abordam um determinado assunto de maneira específica. Portanto, um SRI flexível deve prover documentos organizados em categorias de interesse que correspondam tanto a um tópico geral, como por exemplo, *esportes*, quanto a um tópico específico, como por exemplo, *futebol*. Nesse contexto, Bordogna et al. (2006) propuseram um algoritmo de agrupamento fuzzy hierárquico dinâmico, o qual obtém uma estrutura de grupos hierárquicos de documentos, cujos grupos são identificados automaticamente. Nessa estrutura, cada nível da hierarquia corresponde a um nível distinto de sobreposição de grupos, no qual no nível mais alto da hierarquia o valor de sobreposição aumenta, uma vez que os tópicos representados nestes níveis são mais gerais.

Saraçoglu et al. (2007) propuseram, e posteriormente Saraçoglu et al. (2008) melhoraram, uma abordagem com o uso da lógica fuzzy também para busca de similaridade entre

documentos na tentativa de solucionar o problema de multi categorias. Segundo Saraçoglu et al. (2007), o maior problema dos atuais sistemas de busca é o resultado da busca, os quais disponibilizam documentos não relacionados ou diminuem ao máximo o número de documentos não relacionados como resultado da busca. Geralmente, nestes sistemas, os documentos pertencem apenas a uma categoria. Assim, os autores propuseram a indexação da saída dos documentos pertencentes a mais de uma categoria e determinada as categorias as quais pertencem.

Tjhi et al. (2009) propuseram um algoritmo de agrupamento de documentos que capta simultaneamente as vantagens do agrupamento fuzzy, do agrupamento possibilístico e do co-agrupamento. O algoritmo proposto é chamado DFPC (*Dual Fuzzy-Possibilistic Co-clustering*). Esse algoritmo identifica e representa grupos de documentos mais realistas, além de ser mais robusto com relação à ruídos e estável com relação à grupos coincidentes. A formulação de co-agrupamento, por sua vez, possibilita a organização de documentos e habilita o algoritmo para gerar a pertinência flexível das palavras nos documentos, beneficiando a interpretabilidade dos grupos de documentos.

Chowdhury e Bhuyan (2010) propuseram um método que faz uso do algoritmo de agrupamento FCM para verificar similaridade entre documentos de grupos diferentes: grupos de documentos recuperados e grupos de documentos não-recuperados pelos SRIs. Segundo os autores, de maneira geral, um SRI recupera documentos que são relevantes para um usuário considerando a similaridade entre os termos utilizados pelo usuário no momento da consulta e os termos que ocorrem nos documentos. No entanto, documentos que não são recuperados a partir dessa estratégia podem ter similaridade com os documentos recuperados.

Yan et al. (2012) propuseram um novo algoritmo denominado SS-HFCR (*Heuristic Semi-Supervised Fuzzy Co-clustering*) para organização de grandes volumes de documentos provenientes da Web. No processo de agrupamento realizado pelo algoritmo proposto, os autores incluem um conhecimento obtido previamente na forma de restrições fornecidas por usuários de SRIs. Cada restrição especifica se um par de documentos deve ou não ser agrupado junto. Por meio desse algoritmo, não somente os graus de pertinência dos documentos nos grupos são obtidos, mas também os graus de pertinência das palavras nos documentos. Ao desenvolver este algoritmo, os autores tinham como principal objetivo melhorar a acurácia do agrupamento e reduzir a sensibilidade dos parâmetros de fuzzificação pela inserção do conhecimento prévio.

Dante dos trabalhos citados anteriormente e de outras abordagens para organização de documentos por meio de agrupamento fuzzy, as quais podem ser conferidas em (Lee, 2001; Horng et al., 2005; Bordogna et al., 2006; Kozielski, 2007; Song et al., 2011; Bordogna e Pasi, 2011, 2012), observa-se que, de maneira geral, todos eles apresentam boas estratégias para organização flexível de documentos, considerando como principal foco de suas abordagens o processo de agrupamento.

No entanto, para uma organização flexível completa, além do agrupamento, é impor-

tante a extração de bons representantes de grupos, como pode ser observado no exemplo apresentado na Figura 3.2 na Seção 3.2. Neste exemplo, os representantes de grupos *Educação*, *Política* e *Saúde* foram escolhidos manualmente para indicar o tópico abordado por cada grupo de documentos naquela organização de notícias. No entanto, a escolha de representantes de grupos não é uma tarefa trivial, uma vez que os mesmos devem ser capazes de representar, da melhor maneira possível, o conteúdo de todos os documentos de um determinado grupo. As técnicas tradicionais de agrupamento não proveem uma descrição apropriada para os grupos obtidos, dificultando a interpretação dos mesmos (Anaya-Sánchez et al., 2008). Esta dificuldade é ainda maior quando deseja-se obter representantes de grupos fuzzy. Neste tipo de agrupamento um mesmo representante pode indicar o conteúdo de mais de um grupo, uma vez que no agrupamento fuzzy um mesmo documento pode ser compatível com mais de um grupo.

Em agrupamento de documentos a escolha de representantes de grupos é realizada por meio da identificação de descritores, que são termos significantes dos tópicos abordados nos documentos. Contudo, documentos são representados por uma grande quantidade de termos, isto é, por um espaço de características de alta dimensionalidade. Logo, a extração de bons descritores é um problema desafiador e em aberto. Além disso, em aplicações em que o agrupamento é utilizado para a RI, a extração do significado dos grupos é tão importante quanto um bom agrupamento (Feldman e Sanger, 2007). Neste contexto, a principal contribuição deste doutorado deu-se na organização flexível de documentos pela proposta de investigação e desenvolvimento de métodos para a extração de descritores de grupos fuzzy.

### 3.4.2 Métodos propostos para extração de descritores de grupos fuzzy

Conforme apresentado na Seção 2.3 do Capítulo 2, existem na literatura duas possibilidades de extração de descritores: DCF (*Description Comes First*) e DCL (*Description Comes Last*). Os métodos desenvolvidos neste doutorado são do tipo DCL, o qual permite que diferentes algoritmos de agrupamento sejam utilizados, dependendo do que deseja-se obter a partir da organização de documentos, tornando a proposta de organização flexível mais abrangente.

Na Figura 3.6 é apresentada a contextualização da abordagem proposta para organização flexível no nível da organização de um SRI. Neste contexto, a abordagem proposta para organização flexível tem início com a representação dos documentos, a qual é composta pelo pré-processamento e agrupamento dos documentos. Especificamente neste doutorado, foram realizadas investigações utilizando o algoritmo de agrupamento Fuzzy C-Means (FCM) (Bezdek, 1981), o algoritmo de agrupamento Possibilístico C-Means (PCM) (Pal et al., 2005) e o algoritmo de agrupamento Hierárquico Fuzzy C-Means (HFCM) (Pedrycz e Reformat, 2006), apresentados na Seção 2.2.1. Uma vez agrupados os documentos, os grupos não possuem significados. Para tanto, foram desenvolvidos métodos para extração de descritores de grupos, considerando o processo de agrupamento fuzzy, pelo qual

documentos podem pertencer a diferentes grupos com diferentes graus de pertinência. A partir do tipo de agrupamento utilizado para organizar os documentos, foram propostos métodos para extração de descritores que utilizam esta informação de pertinência.

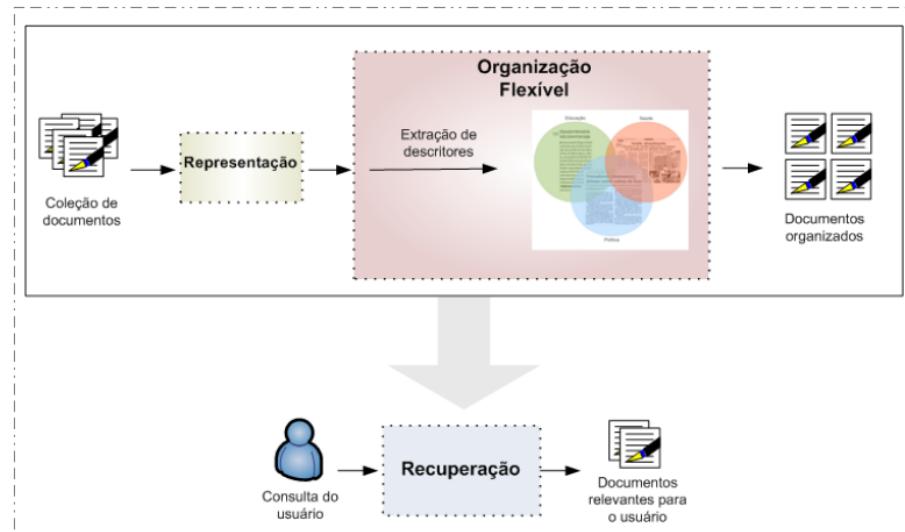


Figura 3.6: Contextualização da abordagem proposta para organização flexível no nível da organização de um SRI

O ponto de inovação deste doutorado concentra-se no desenvolvimento de métodos para extração de descritores de grupos fuzzy que possibilitam organizar, de maneira flexível, os documentos de acordo com os principais tópicos da coleção, considerando que cada grupo refere-se a um tópico. Esses métodos foram desenvolvidos com base em medidas conhecidas da área de RI (precisão, revocação e medida-F ( $F_1$ )) e, a partir do tipo de agrupamento utilizado para organizar os documentos (*flat* ou hierárquico), utilizam a informação de pertinência para extração de descritores de grupos fuzzy. As investigações realizadas sobre cada um dos métodos propostos serão apresentadas detalhadamente no próximo capítulo divididos em dois tópicos:

- **Organização flexível de documentos usando agrupamento fuzzy *flat*.** Desta investigação, foram propostos dois métodos: SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*) e SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*).
- **Organização flexível de documentos usando agrupamento fuzzy hierárquico.** Nesta investigação, na qual foi utilizado o algoritmo HFCM, foi proposto um método para extração de descritores de grupos hierárquicos fuzzy com base na medida  $F_1$  denominado HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*).

## 3.5 Considerações finais

Neste capítulo foi apresentado o contexto em que este doutorado se insere, o qual consiste do tratamento de imprecisão e incerteza de documentos de modo a proporcionar flexibilidade aos Sistemas de Recuperação de Informação.

Foram apresentadas também explorações preliminares realizadas, as quais apoiaram a especificação do foco principal da proposta de doutorado. As explorações realizadas são relacionadas à geração de regras fuzzy para classificação de documentos e à recuperação de informação a partir de tais regras. Assim, com os experimentos iniciais e a revisão do estado-da-arte, concluiu-se que a organização flexível de documentos é de grande importância para o tratamento de imprecisão e incerteza em SRI. Além disso, observou-se que a organização flexível de documentos pode ser obtida por meio de agrupamento fuzzy, sendo a extração de descritores de grupos a principal tarefa a ser realizada neste doutorado.

Sendo assim, ainda neste capítulo, foi apresentada a abordagem proposta para a organização flexível de documentos, cujos detalhes relacionados aos métodos desenvolvidos para extração de descritores de grupos fuzzy, experimentos realizados e resultados obtidos são apresentados no próximo capítulo.

# Métodos Propostos para Extração de Descritores de Grupos na Organização Flexível de Documentos

## 4.1 Considerações iniciais

Conforme apresentado no capítulo anterior, o tratamento de imprecisão e incerteza de documentos pode ser abordado nos níveis de representação, organização e recuperação de um Sistema de Recuperação de Informação (SRI). Nesse contexto, considerando os experimentos preliminares realizados e a revisão da literatura, observou-se a necessidade da organização flexível de documentos. Esse tipo de organização é abordado neste doutorado por meio do agrupamento fuzzy de documentos, para o qual a extração de descritores dos grupos obtidos é de grande importância, uma vez que por meio dos descritores de grupos é possível interpretar o significado dos grupos.

Sendo assim, neste capítulo são apresentados três novos métodos para extração de descritores de grupos fuzzy do tipo *Description Comes Last* (DCL) (Zhang, 2009), pelo qual o processo de agrupamento é separado da extração de descritores. Os métodos propostos viabilizam a organização flexível de documentos pela extração de descritores de grupos após um processo de agrupamento fuzzy de documentos e possuem como base medidas clássicas da área de Recuperação de Informação (RI): precisão (*precision*), revocação (*recall*) e medida *f1*. O primeiro método proposto é denominado SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*), o qual extrai descritores de grupos fuzzy *flat* de documentos. O segundo método proposto é denominado SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*), o qual também extrai descritores de grupos fuzzy *flat* de documentos, porém incluindo os graus de pertinência dos documentos

nos grupos no cálculo da precisão, revocação e medida  $f1$  dos candidatos a descritores de grupos. Por fim, o terceiro método proposto é denominado HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*), o qual extrai descritores de grupos fuzzy hierárquicos de documentos.

Após a descrição dos métodos propostos, a avaliação realizada e os resultados obtidos sobre os mesmos também são apresentados neste capítulo.

## 4.2 O método SoftO-FDCL

O método SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*) foi desenvolvido com o propósito de fornecer flexibilidade a um SRI no nível da sua organização por meio da extração de descritores de grupos fuzzy *flat* de documentos.

Quando a organização de documentos é obtida por meio do agrupamento *flat* convencional, ou *hard*, tem-se grupos cujos documentos não apresentam relação nenhuma com documentos de outros grupos que compõem a organização. No entanto, no agrupamento fuzzy *flat*, documentos de grupos distintos possuem uma relação de semelhança definida pelo conceito de grau de pertinência dos documentos nos grupos. O grau de pertinência de um documento em um grupo é obtido pelo processo de agrupamento fuzzy, pelo qual é medido o quanto um documento pertence a um determinado grupo. Dessa maneira, documentos podem pertencer a mais de um grupo e se assemelhar com documentos de grupos distintos.

A semelhança entre documentos de grupos distintos é de extrema importância para a obtenção da organização flexível de documentos, pois se assumirmos que grupos de documentos representam tópicos abordados pelos documentos da coleção organizada, dois documentos que abordam tópicos diferentes, ou seja, que estão alocados em dois grupos diferentes, podem também abordar um tópico em comum, *i.e.*, podem também ambos estarem alocados em um terceiro grupo.

Além disso, por meio do agrupamento fuzzy de documentos, dois documentos que abordam um mesmo tópico, mas com diferente intensidade, são alocados em um mesmo grupo, pois o grau de pertinência mede o quanto um determinado documento pertence a um determinado grupo, *i.e.*, com qual intensidade um determinado documento aborda um determinado tópico.

Sendo assim, o método SoftO-FDCL extrai descritores de grupos fuzzy de forma a identificar tópicos para a organização flexível de documentos. Para tanto, a informação de pertinência obtida do processo de agrupamento fuzzy da coleção de documentos a ser organizada é fundamental. Essa informação é considerada pelo método SoftO-FDCL quando todos os termos representativos da coleção de documentos são avaliados como candidatos a descritores de grupos. Tal avaliação é realizada utilizando as medidas clássicas para avaliação quantitativa da efetividade da Recuperação de Informação (RI): precisão (*precision*), revocação (*recall*) e medida  $f1$  (*F1-measure*) (Salton e McGill, 1983). Uma

medida de precisão verifica a proporção de documentos relevantes entre os documentos recuperados. A revocação indica a proporção de documentos relevantes recuperados entre todos os documentos que são conhecidamente relevantes para uma dada consulta em um SRI. A medida  $f1$  é a média harmônica entre precisão e revocação.

No método SoftO-FDCL, um documento  $\mathbf{d}_i$ , com  $1 \leq i \leq n$ , para  $n$  igual a quantidade de documentos da coleção, pertence a um grupo  $g_l$ , com  $1 \leq l \leq c$ , para  $c$  igual a quantidade de grupos, se o mesmo possui grau de pertinência no grupo maior ou igual a  $\delta$ , i.e.,  $\mu(\mathbf{d}_i, g_l) \geq \delta$ , no qual  $\delta$  é um limiar definido por  $\delta = \frac{1}{c}$ . Esse limiar é considerado por duas razões. Primeiro, ele permite que os candidatos a descritores de grupos representem documentos que pertençam à mais de um grupo com diferentes graus, ao invés de considerar somente documentos com o maior grau de pertinência em um grupo. Segundo, por meio desse limiar é possível penalizar os candidatos a descritores que ocorrem em documentos com baixo grau de pertinência em um grupo. Assim, todos os termos  $t_j$ , com  $1 \leq j \leq k$ , para  $k$  igual a quantidade de termos representativos da coleção, são avaliados como candidatos a descritores de um grupo  $g_l$ , considerando a matriz de contingência apresentada na Tabela 4.1.

Tabela 4.1: Matriz de contingência do termo  $t_j$  para o grupo  $g_l$  para as medidas de Recuperação de Informação utilizadas pelo método SoftO-FDCL

|  | Documentos que pertencem ao grupo $g_l$ | Documentos que não pertencem ao grupo $g_l$ |
|--|---|---|
| Documentos que possuem o descritor candidato $t_j$     | $ganhos(t_j, g_l)$                      | $ruidos(t_j, g_l)$                          |
| Documentos que não possuem o descritor candidato $t_j$ | $perdas(t_j, g_l)$                      | $rejeitos(t_j, g_l)$                        |

Sendo  $\mu(\mathbf{d}_i, g_l)$  o grau de pertinência do documento  $\mathbf{d}_i$  no grupo  $g_l$  e  $\sigma(t_j, \mathbf{d}_i)$  a frequência do termo  $t_j$  no documento  $\mathbf{d}_i$ , a formalização da medição dos  $ganhos(t_j, g_l)$ ,  $perdas(t_j, g_l)$ ,  $ruidos(t_j, g_l)$  e  $rejeitos(t_j, g_l)$ , é definida pelas Equações (4.1), (4.2), (4.3) e (4.4), respectivamente, considerando as funções de grau apresentadas nas Equações (4.5) e (4.6).

$$ganhos(t_j, g_l) = \sum_{i=1}^n \phi(t_j, \mathbf{d}_i) \cdot \psi(\mathbf{d}_i, g_l) \quad (4.1)$$

$$perdas(t_j, g_l) = \sum_{i=1}^n (1 - \phi(t_j, \mathbf{d}_i)) \cdot \psi(\mathbf{d}_i, g_l) \quad (4.2)$$

$$ruidos(t_j, g_l) = \sum_{i=1}^n \phi(t_j, \mathbf{d}_i) \cdot (1 - \psi(\mathbf{d}_i, g_l)) \quad (4.3)$$

$$rejeitos(t_j, g_l) = \sum_{i=1}^n 1 - (\phi(t_j, \mathbf{d}_i) \cdot \psi(\mathbf{d}_i, g_l)) \quad (4.4)$$

$$\phi(t_j, \mathbf{d}_i) = \begin{cases} 1, & \sigma(t_j, \mathbf{d}_i) > 0 \\ 0, & \sigma(t_j, \mathbf{d}_i) = 0 \end{cases} \quad (4.5)$$

$$\psi(\mathbf{d}_i, g_l) = \begin{cases} 1, & \mu(\mathbf{d}_i, g_l) \geq \delta \\ 0, & \mu(\mathbf{d}_i, g_l) < \delta \end{cases} \quad (4.6)$$

Sendo assim, a extração de descritores de um determinado grupo  $g_l$  tem início com o cálculo do  $f1$  de cada descritor candidato  $t_j$ . Ao final, tem-se um *ranking* de termos candidatos a descritores de cada grupo. Considerando a matriz de contingência apresentada na Tabela 4.1, o cálculo do  $f1$  de cada descritor candidato  $t_j$  é obtido como segue.

- i. Calcular a precisão  $p(t_j, g_l)$  do termo  $t_j$  candidato a descritor do grupo  $g_l$ :

$$p(t_j, g_l) = \frac{ganhos(t_j, g_l)}{ganhos(t_j, g_l) + ruidos(t_j, g_l)} \quad (4.7)$$

- ii. Calcular a revocação  $r(t_j, g_l)$  do termo  $t_j$  candidato a descritor do grupo  $g_l$ :

$$r(t_j, g_l) = \frac{ganhos(t_j, g_l)}{ganhos(t_j, g_l) + perdas(t_j, g_l)} \quad (4.8)$$

- iii. Calcular a medida  $f1(t_j, g_l)$  do termo  $t_j$  candidato a descritor do grupo  $g_l$ :

$$f1(t_j, g_l) = \frac{2 \cdot p(t_j, g_l) \cdot r(t_j, g_l)}{p(t_j, g_l) + r(t_j, g_l)} \quad (4.9)$$

Assim, a quantidade de descritores de cada grupo é selecionada empiricamente a partir dos candidatos a descritores que possuem maior  $f1$ . A medida  $f1$  mede quão representativo um descritor é para um grupo e os descritores identificam tópicos da organização flexível. Essa flexibilidade é alcançada porque os graus de pertinência indicam a compatibilidade entre documentos e grupos. Além disso, grupos distintos podem ter os mesmos descritores.

## 4.3 O método SoftO-wFDCL

O método SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*) é uma extensão do método SoftO-FDCL. O método SoftO-wFDCL avalia a eficiência de cada descritor candidato em identificar os documentos em um grupo incluindo o grau de pertinência dos documentos em cada grupo nas medidas de precisão, revocação e  $f1$ . Essa nova forma de avaliação considera que os graus de pertinência carregam uma informação

adicional sobre a representatividade dos termos, a qual pode contribuir para uma avaliação mais precisa acerca da importância de um termo candidato a descritor de grupo. O uso do grau de pertinência na avaliação de um descritor candidato é útil, pois, essa medida garante que os descritores extraídos representam a informação de que o documento pode pertencer a mais de um grupo com diferentes graus de compatibilidade.

Assim como no método SoftO-FDCL, no método SoftO-wFDCL um documento  $d_i$ , com  $1 \leq i \leq n$ , pertence a um grupo  $g_l$ , com  $1 \leq l \leq c$ , para  $c$  igual a quantidade de grupos, se o mesmo possui grau de pertinência  $\mu(d_i, g_l) \geq \delta$ , no qual  $\delta$  é um limiar definido por  $\delta = \frac{1}{c}$ .

A extração de descritores de um determinado grupo pelo método SoftO-wFDCL tem início com o cálculo da medida  $f1$  de cada descritor candidato. Um *ranking* de termos ponderados pela sua medida  $f1$  é obtido para cada grupo, considerando a matriz de contingência apresentada na Tabela 4.2<sup>1</sup>. Nas equações (4.10), (4.11), (4.12) e (4.13), tem-se a formalização da inserção dos graus de pertinência na medição dos *ganhos*( $t_j, g_l$ ), *perdas*( $t_j, g_l$ ), *ruidos*( $t_j, g_l$ ) e *rejeitos*( $t_j, g_l$ ), respectivamente, considerando funções de grau apresentadas nas Equações (4.15) e (4.16). A definição de *ganhos*( $t_j, g_l$ ), *perdas*( $t_j, g_l$ ) e *rejeitos*( $t_j, g_l$ ) inclui o maior grau de pertinência dos documentos no grupo para o qual estão sendo avaliados os termos candidatos a descritores. A definição de *ruidos*( $t_j, g_l$ ) para o método SoftO-wFDCL, por sua vez, tem uma particularidade. Considerando que se um termo  $t_j$  ocorre em um documento que não pertence ao grupo  $g_l$ , então este documento pertence a um outro grupo com um grau de pertinência maior do que o grau de pertinência em  $g_l$ , ou seja, o grupo para o qual estão sendo avaliados os termos candidatos a descritores. Sendo assim, a informação de pertinência a ser inserida na medida de *ruidos* utilizada pelo método SoftO-wFDCL é o maior grau de pertinência que o documento que contém o termo  $t_j$  possui em um grupo diferente de  $g_l$ . Essa definição é apresentada na Equação (4.14).

Tabela 4.2: Matriz de contingência para as medidas de Recuperação de Informação utilizadas pelo método SoftO-wFDCL

|  | Documentos que pertencem ao grupo $g_l$ | Documentos que não pertencem ao grupo $g_l$ |
|--|---|---|
| Documentos que possuem o descritor candidato $t_j$     | $ganhos(t_j, g_l)$                      | $ruidos(t_j, g_l)$                          |
| Documentos que não possuem o descritor candidato $t_j$ | $perdas(t_j, g_l)$                      | $rejeitos(t_j, g_l)$                        |

<sup>1</sup>Esta tabela é igual a Tabela 4.1, porém o cálculo de ganhos, ruidos, perdas e rejeitos é realizado diferentemente do método SoftO-FDCL

$$ganhos(t_j, g_l) = \sum_{i=1}^n \phi(t_j, \mathbf{d}_i) \cdot (1 + \mu(\mathbf{d}_i, g_l)) \cdot \psi(\mathbf{d}_i, g_l) \quad (4.10)$$

$$perdas(t_j, g_l) = \sum_{i=1}^n (1 - \phi(t_j, \mathbf{d}_i)) \cdot (1 + \mu(\mathbf{d}_i, g_l)) \cdot \psi(\mathbf{d}_i, g_l) \quad (4.11)$$

$$ruidos(t_j, g_l) = \sum_{i=1}^n \phi(t_j, \mathbf{d}_i) \cdot (1 + \tau(\mathbf{d}_i, \bar{g}_l)) \cdot (1 - \psi(\mathbf{d}_i, g_l)) \quad (4.12)$$

$$rejeitos(t_j, g_l) = \sum_{i=1}^n (1 - \phi(t_j, \mathbf{d}_i)) \cdot (1 + \mu(\mathbf{d}_i, g_l)) \cdot (1 - \psi(\mathbf{d}_i, g_l)) \quad (4.13)$$

$$\tau(\mathbf{d}_i, \bar{g}_l) = \max_{x=1}^c \mu(\mathbf{d}_i, g_x), \quad \forall g_x \neq g_l \quad (4.14)$$

$$\phi(t_j, \mathbf{d}_i) = \begin{cases} 1, & \sigma(t_j, \mathbf{d}_i) > 0 \\ 0, & \sigma(t_j, \mathbf{d}_i) \leq 0 \end{cases} \quad (4.15)$$

$$\psi(\mathbf{d}_i, g_l) = \begin{cases} 1, & \mu(\mathbf{d}_i, g_l) \geq \delta \\ 0, & \mu(\mathbf{d}_i, g_l) < \delta \end{cases} \quad (4.16)$$

O cálculo de  $f1$  do candidato a descritor  $t_j$  para o grupo  $g_l$  é feito como segue.

i. Calcular a precisão do termo  $t_j$  candidato a descritor do grupo  $g_l$ :

$$p(t_j, g_l) = \frac{ganhos(t_j, g_l)}{ganhos(t_j, g_l) + ruidos(t_j, g_l)} \quad (4.17)$$

ii. Calcular a revocação do termo  $t_j$  candidato a descritor do grupo  $g_l$ :

$$r(t_j, g_l) = \frac{ganhos(t_j, g_l)}{ganhos(t_j, g_l) + perdas(t_j, g_l)} \quad (4.18)$$

iii. Calcular a medida  $f1$  do termo  $t_j$  candidato a descritor do grupo  $g_l$ :

$$f1(t_j, g_l) = \frac{2 \cdot p(t_j, g_l) \cdot r(t_j, g_l)}{p(t_j, g_l) + r(t_j, g_l)} \quad (4.19)$$

A quantidade de descritores de cada grupo é selecionada empiricamente a partir dos candidatos a descritores que possuem maior  $f1$ .

## 4.4 O método HSoftO-FDCL

O método HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*) foi desenvolvido como uma extensão do método SoftO-FDCL e tem como propó-

sito fornecer flexibilidade a um SRI no nível da sua organização por meio da extração de descritores de grupos fuzzy hierárquicos de documentos. Como apresentado, o agrupamento hierárquico fornece uma visão dos documentos agrupados em diferentes níveis de abstração. Além disso, a organização hierárquica de documentos permite que a coleção de documentos seja visualizada e explorada iterativamente, já que por meio desta organização dois tópicos podem ser a especialização ou generalização um do outro. Ou seja, documentos podem abordar um tópico, como por exemplo *esporte*, representado por um grupo em um nível mais alto da hierarquia, ou um sub-tópico, como por exemplo *futebol*, representado por um grupo em um nível abaixo.

Assim como o método SoftO-FDCL, o método HSoftO-FDCL também considera a informação de pertinência obtida do processo de agrupamento fuzzy, já que os documentos também podem estar alocados em mais de um grupo em um nível da hierarquia. No entanto, para o método HSoftO-FDCL, um documento  $d_i$ , com  $1 \leq i \leq n$ , para  $n$  igual a quantidade de documentos da coleção organizada, pertence a um grupo  $g_{l_u}$ , com  $1 \leq l \leq c$ , para  $c$  igual a quantidade de grupos, e  $1 \leq u \leq y$ , para  $y$  igual a quantidade de níveis da hierarquia, se o grau de pertinência do documento  $d_i$  no grupo  $g_{l_u}$  é maior ou igual a  $\zeta$ , i.e.,  $\mu(d_i, g_{l_u}) \geq \zeta$ , no qual  $\zeta$  é um limiar definido por  $\zeta = \frac{\mu(d_i, g_{l_{u-1}})}{c}$ , para  $\mu(d_i, g_{l_{u-1}})$  o grau de pertinência do documento  $d_i$  no grupo  $g_{l_{u-1}}$ , i.e., o grau de pertinência do documento  $textbf{d}_i$  no grupo que deu origem a  $g_{l_u}$ .

Sendo assim, o método HSoftO-FDCL considera que a informação de pertinência de um documento em um grupo em um determinado nível da hierarquia está condicionada à informação de pertinência desse documento em um nível acima. Por exemplo, observe a hierarquia ilustrada na Figura 4.1, na qual tem-se a distribuição dos graus de pertinência do documento  $d_1$ .

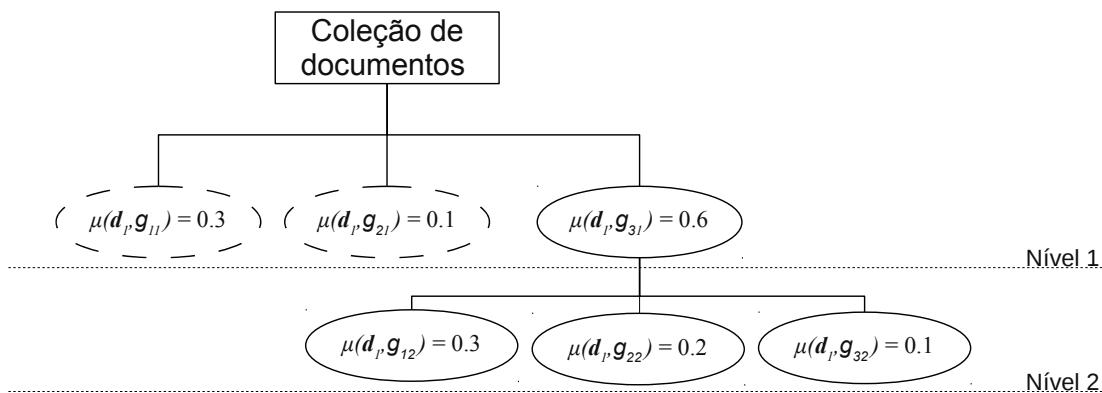


Figura 4.1: Agrupamento fuzzy hierárquico

Observa-se na Figura 4.1 que a soma dos graus de pertinência do documento  $d_1$  nos grupos do nível 2 da hierarquia é igual ao grau de pertinência do documento  $d_1$  no grupo que deu origem aos grupos nos quais ele está alocado no nível 2.

Sendo assim, o método HSoftO-FDCL extrai descritores dos grupos do nível  $u$  da hie-

rarquia fuzzy de documentos considerando o limiar de pertinência  $\zeta$  apresentado anteriormente e a matriz de contingência apresentada na Tabela 4.3. Nas equações (4.20), (4.21), (4.22) e (4.23), tem-se a formalização da medição dos  $ganhos(t_j, g_{l_u})$ ,  $perdas(t_j, g_{l_u})$ ,  $ruidos(t_j, g_{l_u})$  e  $rejeitos(t_j, g_{l_u})$ , respectivamente, considerando funções de grau apresentadas nas Equações (4.24) e (4.25). Considerando  $\mu(\mathbf{d}_i, g_l)$  o grau de pertinência do documento  $\mathbf{d}_i$  no grupo  $g_l$  e  $\sigma(t_j, \mathbf{d}_i)$  a frequência do termo  $t_j$  no documento  $\mathbf{d}_i$ .

Tabela 4.3: Matriz de contingência do termo  $t_j$  para o grupo  $g_{l_u}$  para as medidas de Recuperação de Informação utilizadas pelo método HSoftO-FDCL

|  | Documentos que pertencem ao grupo $g_{l_u}$ | Documentos que não pertencem ao grupo $g_{l_u}$ |
|--|---|---|
| Documentos que possuem o descritor candidato $t_j$     | $ganhos(t_j, g_{l_u})$                      | $ruidos(t_j, g_{l_u})$                          |
| Documentos que não possuem o descritor candidato $t_j$ | $perdas(t_j, g_{l_u})$                      | $rejeitos(t_j, g_{l_u})$                        |

$$ganhos(t_j, g_{l_u}) = \sum_{i=1}^n \phi(t_j, \mathbf{d}_i) \cdot \psi(\mathbf{d}_i, g_{l_u}) \quad (4.20)$$

$$perdas(t_j, g_{l_u}) = \sum_{i=1}^n (1 - \phi(t_j, \mathbf{d}_i)) \cdot \psi(\mathbf{d}_i, g_{l_u}) \quad (4.21)$$

$$ruidos(t_j, g_{l_u}) = \sum_{i=1}^n \phi(t_j, \mathbf{d}_i) \cdot (1 - \psi(\mathbf{d}_i, g_{l_u})) \quad (4.22)$$

$$rejeitos(t_j, g_{l_u}) = \sum_{i=1}^n 1 - (\phi(t_j, \mathbf{d}_i) \cdot \psi(\mathbf{d}_i, g_{l_u})) \quad (4.23)$$

$$\phi(t_j, \mathbf{d}_i) = \begin{cases} 1, & \sigma(t_j, \mathbf{d}_i) > 0 \\ 0, & \sigma(t_j, \mathbf{d}_i) = 0 \end{cases} \quad (4.24)$$

$$\psi(\mathbf{d}_i, g_{l_u}) = \begin{cases} 1, & \mu(\mathbf{d}_i, g_{l_u}) \geq \zeta \\ 0, & \mu(\mathbf{d}_i, g_{l_u}) < \zeta \end{cases} \quad (4.25)$$

Assim como o método SoftO-FDCL, o método HSoftO-FDCL avalia todos os termos representativos da coleção, os quais são considerados candidatos a descritores dos grupos, utilizando as medidas de precisão, revocação e  $f1$ . Tal avaliação visa obter um *ranking* de termos candidatos a descritores de cada grupo em cada nível, considerando a matriz de contingência apresentada na Tabela 4.3. O método HSoftO-FDCL tem início com

a aplicação do método SoftO-FDCL para extraer os descritores dos grupos do primeiro nível da hierarquia, uma vez que nesse nível os grupos são isolados, ou seja, os grupos não são originados de outros grupos. Além disso, pelo método HSoftO-FDCL, os termos que tenham sido escolhidos como descritores do grupo  $g_{l_{u-1}}$  não são considerados termos candidatos a descritores dos grupos  $g_{l_u}$ . Sendo assim, os descritores de grupos fuzzy hierárquicos são extraídos pelo método HSoftO-FDCL como segue.

1. Extrair descritores dos grupos do primeiro nível,  $u = 1$ , utilizando o método SoftO-FDCL.
2. Para cada nível  $u$ ,  $2 \leq u \leq y$ , faça:
  - (a) Se o termo  $t_j$  tiver sido escolhido como descritor do grupo  $g_{l_{u-1}}$ , desconsidere-lo.
  - (b) Para cada grupo  $g_l$ ,  $1 \leq l \leq c$ , do nível  $u$ , faça:
    - i. Calcular a precisão  $p(t_j, g_{l_u})$  do termo  $t_j$  candidato a descritor do grupo  $g_{l_u}$ :
$$p(t_j, g_{l_u}) = \frac{ganhos(t_j, g_{l_u})}{ganhos(t_j, g_{l_u}) + ruidos(t_j, g_{l_u})} \quad (4.26)$$
    - ii. Calcular a revocação  $r(t_j, g_{l_u})$  do termo  $t_j$  candidato a descritor do grupo  $g_{l_u}$ :
$$r(t_j, g_{l_u}) = \frac{ganhos(t_j, g_{l_u})}{ganhos(t_j, g_{l_u}) + perdas(t_j, g_{l_u})} \quad (4.27)$$
    - iii. Calcular a medida  $f1(t_j, g_{l_u})$  do termo  $t_j$  candidato a descritor do grupo  $g_{l_u}$ :
$$f1(t_j, g_{l_u}) = \frac{2 \cdot p(t_j, g_{l_u}) \cdot r(t_j, g_{l_u})}{p(t_j, g_{l_u}) + r(t_j, g_{l_u})} \quad (4.28)$$
3. A quantidade de descritores de cada grupo é selecionada empiricamente a partir dos candidatos a descritores que possuem maior  $f1$ .

Ao final desse procedimento, tem-se a organização flexível de documentos apresentada em um estrutura hierárquica obtida pelo agrupamento fuzzy hierárquico de documentos e pela extração dos descritores de grupos. Nessa organização, a flexibilidade é alcançada porque os documentos pertencem a mais de um grupo, em um mesmo nível da hierarquia, e os graus de pertinência indicam a compatibilidade entre documentos e grupos. Além disso, os descritores extraídos identificam tópicos da organização, os quais podem ser a especialização ou generalização uns dos outros.

Sendo assim, a principal novidade dos métodos propostos consiste em possibilitar a organização flexível de documentos. Por meio dos métodos apresentados, descritores de grupos são extraídos para representar o conteúdo dos documentos de um determinado grupo considerando a imprecisão e a incerteza inerentes aos documentos. Além disso, os métodos propostos são independentes do algoritmo de agrupamento utilizado. Foram

propostos três métodos que extraem descritores de grupos fuzzy sobre as seguintes perspectivas: o método SoftO-FDCL extrai descritores de grupos fuzzy *flat*, o qual é suficiente para proporcionar a organização flexível de documentos; o método SoftO-wFDCL é uma extensão do método SoftO-FDCL, o qual também extraí descritores de grupos fuzzy *flat*, mas acrescenta o grau de pertinência obtido do agrupamento fuzzy, como uma informação adicional para a extração de descritores; e o método HSoftO-wFDCL, o qual também é uma extensão do método SoftO-FDCL para extração de descritores de grupos hierárquicos, proporcionando uma visão dos documentos em diferentes níveis de abstração. A avaliação de cada um desses métodos propostos é apresentada a seguir.

## 4.5 Avaliação dos métodos propostos

Conforme apresentado na Seção 2.3 do Capítulo 2, é muito comum que os descritores de grupos obtidos pelos métodos disponíveis na literatura sejam avaliados de acordo com o desempenho do processo de agrupamento. No entanto, de acordo com Zhang et al. (2009), os descritores de grupos devem ser avaliados com relação à sua concisão, à sua comprehensibilidade, à sua acurácia e/ou distinção.

Nos experimentos realizados para a avaliação dos métodos propostos neste doutorado, avaliou-se a acurácia dos descritores obtidos, considerando que, em qualquer organização de documentos, é importante que um tópico seja o mais representativo possível para o conjunto de documentos. Quando a organização de documentos é alcançada utilizando grupos de documentos, os descritores de grupos devem ser representativos para os documentos pertencentes a um determinado grupo. Com o objetivo de avaliar quão representativos os descritores são, foi avaliado o poder preditivo dos descritores considerando um grupo como uma classe e os descritores dos grupos como atributos dos documentos. Uma vez que no agrupamento fuzzy os documentos podem pertencer a mais de um grupo, a classe do documento é considerado o grupo no qual ele possui maior grau de pertinência.

Depois de rotular cada documento na coleção com o grupo correspondente, foi criada uma matriz atributo-valor com cada descritor sendo um atributo. As células da matriz contém a frequência ponderada de um termo em um documento em função de sua distribuição ao longo da coleção, denominada *tf-idf* (*Term Frequency-Inverse Document Frequency*). A partir dessa medida, a célula  $d_{ij} = tfidf(t_j, d_i) = freq(t_j, d_i) \times idf(t_j)$ ,  $1 \leq j \leq k$  e  $1 \leq i \leq n$ . A frequência do termo  $t_j$  no documento  $d_i$  é  $freq(t_j, d_i)$ , e o inverso da frequência do termo  $t_j$  no documento  $d_i$  é  $idf(t_j) = \log \frac{K}{d(t_j)}$ , com  $d(t_j)$  igual a quantidade de documentos em que  $t_j$  ocorre. Por meio dessa medida, a importância dos termos no documento é ponderada, de forma que os termos presentes em muitos documentos tem um peso menor do que os termos que ocorrem mais raramente na coleção. Termos que ocorrem em muitos documentos não distinguem tópicos. Por outro lado, termos que ocorrem em poucos documentos podem distinguir tópicos.

Utilizando essa matriz, foram realizados experimentos de forma a analisar as taxas

de acerto obtidas da utilização de algoritmos de classificação bastante conhecidos na literatura. Para tais experimentos, foram utilizadas coleções de documentos conhecidas na literatura e considerou-se o pré-processamento dos documentos como a primeira etapa de todas as análises, uma vez que é muito importante que os documentos estejam em um formato adequado para que seja realizado o processo de agrupamento.

As coleções de documentos, a forma de pré-processamento dos documentos e os algoritmos de classificação utilizados na avaliação dos métodos propostos são descritos a seguir.

#### 4.5.1 Coleções de documentos utilizados na avaliação dos métodos propostos

Para avaliar os métodos propostos foram utilizadas as coleções de documentos cujas características, quantidade de classes e quantidade de documentos, são apresentadas na Tabela 4.4.

Tabela 4.4: Coleções de documentos utilizadas nos experimentos

| Coleção       | # classes | # documentos |
|---------------|-----------|--------------|
| Opinosis      | 3         | 51           |
| Reuters-21578 | 43        | 1052         |
| WAP           | 20        | 1560         |
| 20Newsgroups  | 4         | 2000         |
| NSF           | 16        | 1600         |
| Hitech        | 6         | 600          |

Nos experimentos, foram utilizados subconjuntos das coleções visando a otimização do custo computacional dos algoritmos de classificação, bem como uma observação manual dos descritores obtidos pelos métodos propostos. Os subconjuntos foram obtidos pela seleção aleatória de documentos de cada classe das coleções. A descrição de cada coleção é apresentada a seguir.

##### *Opinosis*

A coleção de documentos Opinosis é composta por documentos que possuem como conteúdo revisões feitas por consumidores sobre as características de algum produto. As revisões dos consumidores foram obtidas dos seguintes sítios da web: *Tripadvisor.com*, pelo qual consumidores podem fazer revisões sobre hotéis; *Amazon.com*, pelo qual consumidores podem fazer revisões sobre carros; e *Edmunds.com*, pelo qual consumidores podem fazer revisões sobre produtos eletrônicos. Essa coleção foi adquirida do repositório *UCI Machine Learning Repository* (Frank e Asuncion, 2010).

Essa coleção foi bastante utilizada nos experimentos, pois ela apresenta características que remetem bastante ao problema abordado neste doutorado. As sentenças dos documentos dessa coleção são altamente subjetivas, imprecisas e incertas, uma vez que diferentes documentos com diferentes revisões e sobre diferentes características de produtos podem

compartilhar sentenças semelhantes. O objetivo de organizar esses documentos por meio de agrupamento fuzzy é encontrar grupos de documentos que apresentam alguma similaridade relacionada aos tópicos das revisões realizadas pelos consumidores. Por exemplo, considere as revisões de dois produtos: carro e computador portátil, respectivamente: “O limite de velocidade deste carro é bom” e “Este computador portátil tem boa velocidade de desempenho”. Nesse contexto, o tópico “velocidade” é um tópico em comum entre os dois produtos. Na organização flexível de documentos proposta, diferentes produtos como esses podem ser alocados em um mesmo grupo que representa o tópico “velocidade”.

Além disso, essa coleção possui um conjunto de sumários dos documentos, com os quais é possível realizar uma avaliação qualitativa acerca dos descritores de grupos obtidos pelos métodos propostos.

Os sumários dos documentos da coleção Opinosis foram escritos por humanos por meio do *Amazon's Online Workforce*<sup>2</sup> pelo qual 5 diferentes pessoas puderam escrever livremente um sumário para cada documento. Os sumários foram previamente utilizados pelos autores da coleção para avaliar um método de sumarização automática proposto pelos mesmos (Ganesan et al., 2010). Segundo os autores, a comparação dos sumários escritos por humanos com os sumários gerados automaticamente tem melhor relação com o julgamento humano de qualidade do método proposto, pois os sumários são compostos por sentenças curtas compostas por palavras que representam a informação essencial dos documentos.

Um exemplo do conjunto de sumários do documento que contém revisões sobre a vida útil da bateria do equipamento *amazon kindle* (“battery life of the amazon kindle”) é apresentado na Tabela 4.5.

Tabela 4.5: Sumários escritos por humanos sobre o documento “battery life of the amazon kindle” da coleção opinosis

|   |
|---|
| <i>The Kindle can run for days without a need for recharging.</i>   |
| <i>Battery life is exceptional.</i>   |
| <i>It can be a very difficult process when trying to replace the battery. The battery seems to lose charge quickly. The battery meter is hard to distinguish and provides little insight into the expected life of the battery.</i> |
| <i>Although the battery cannot be replaced as there are large number of ways to charge the device.</i>  |
| <i>The battery life of the Kindle is very long.</i>   |

## **Reuters**

A coleção Reuters-21578<sup>3</sup> é uma das coleções mais utilizadas para testes de pesquisas

---

<sup>2</sup><http://aws.amazon.com/mturk/>

<sup>3</sup><http://www.daviddlewis.com/resources/testcollections/>

sobre categorização de documentos. Essa coleção foi obtida pelo *Carnegie Group, Inc. and Reuters, Ltd* durante o desenvolvimento do sistema de categorização de documentos chamado CONSTRUE (Hayes e Weinstein, 1990). Essa coleção é composta por 21578 documentos em seu formato original distribuídos em 43 categorias. Nos experimentos realizados neste doutorado, foi utilizado um subconjunto composto por 1052 documentos obtidos da coleção original. Este subconjunto foi obtido pela seleção aleatória de documentos de cada classe da coleção.

### ***Wap***

A coleção Wap foi obtida por Moore et al. (1997) no projeto chamado WebACE (Han et al., 1998). Cada documento dessa coleção corresponde a uma página da web em um dos tópicos da hierarquia do Yahoo!<sup>4</sup>. A coleção é composta por 1560 documentos distribuídos em 20 categorias em seu formato original, dos quais todos foram utilizados nos experimentos.

### ***20Newsgroups***

A coleção 20Newsgroups tem se tornado uma das coleções mais populares para avaliação de aplicações de técnicas de aprendizado de máquina sobre documentos, tal como classificação de documentos e agrupamento de documentos. Essa coleção foi adquirida por Lang (1995) para a pesquisa chamada Newsweeder. A coleção original é composta por 20000 documentos, distribuídas em aproximadamente 20 categorias. Nos experimentos realizados neste doutorado, foram selecionados os documentos da categoria *science*, os quais são distribuídos nas classes *sci.crypt*, *sci.electronics*, *sci.med* and *sci.space*. Esse subconjunto de documentos da coleção 20Newsgroups é composto por 2000 documentos, o qual foi obtido pela seleção aleatória de documentos de cada classe da coleção.

### ***NSF***

A coleção NSF (*National Science Foundation*) foi adquirida por meio do repositório *UCI Machine Learning Repository* (Frank e Asuncion, 2010). A coleção original consiste de 129000 resumos relacionados aos prêmios oferecidos pela NSF para pesquisas básicas. Neste doutorado foram selecionados aleatoriamente 1600 documentos para a realização dos experimentos.

### ***Hitech***

A coleção Hitech foi obtida a partir da conferência *Text REtrieval Conference* (TREC)<sup>5</sup>. Seus documentos são compostos de notícias da revista *Jose Mercury News*, as quais são classificadas em diferentes tópicos. A coleção original consiste de 2301 documentos, dos quais foram selecionados aleatoriamente 600 documentos para os experimentos realizados neste doutorado.

---

<sup>4</sup><http://www.yahoo.com>

<sup>5</sup><http://trec.nist.gov>

#### 4.5.2 Pré-processamento dos documentos utilizados na avaliação dos métodos propostos

Para a extração de descritores de grupos, considera-se que a coleção de documentos é primeiramente pré-processada, de forma a obter a matriz documentos-termos, conforme apresentado na Seção 2.2 do Capítulo 2. Nos experimentos realizados para avaliação dos métodos propostos para extração de descritores de grupos fuzzy, a matriz documentos-termos foi gerada utilizando a ferramenta Pretext (Soares et al., 2008). Essa matriz contém em suas células a *tf-idf* (*Term Frequency-Inverse Document Frequency*) dos termos, a qual refere-se a frequência dos termos em um documento ponderada pela frequência com que os mesmos termos ocorrem na coleção inteira. Por meio desta medida, a importância dos termos em um documento é ponderada, de tal forma que termos presentes em muitos documentos tem um peso menor do que os termos que ocorrem raramente na coleção. Termos que ocorrem em muitos documentos não distinguem assuntos. Por outro lado, termos que ocorrem em poucos documentos podem distinguir assuntos.

Uma vez pré-processados, os documentos são agrupados por meio de algum dos algoritmos de agrupamento fuzzy apresentados no Capítulo 2. A medida de validação Silhueta Fuzzy, também apresentada no Capítulo 2, é utilizada em todos os experimentos para determinar a quantidade de grupos adequada para organizar a coleção de documentos.

A partir do agrupamento fuzzy de documentos tem-se a matriz documentos-grupos e os seus descritores podem ser extraídos por um dos métodos propostos.

#### 4.5.3 Algoritmos de classificação utilizados na avaliação dos métodos propostos

Para avaliar a acurácia dos descritores extraídos pelos métodos propostos, algoritmos de classificação conhecidos foram utilizados: SVM, Naive Bayes (NB), Multinomial Naive Bayes (M.Naive), KNN e C4.5.

A maior vantagem do SVM é sua habilidade de aprender independente da dimensionalidade do espaço de características. Entretanto, de acordo com Shanahan e Roma (2003), quando o SVM é aplicado para classificação de documentos ele provê excelente precisão, mas baixa revocação.

O classificador Naive Bayes (NB) é baseado nas regras de Bayes sobre probabilidade condicional. Ele usa todos os atributos contidos nos dados, e os analisa individualmente. De acordo com Schneider (2005), este método é frequentemente utilizado para aplicações e experimentos devido à sua simplicidade e eficiência.

O KNN (*K-Nearest Neighbor*) é um método de aprendizado baseado em instâncias (*Instance-Based Learning - IBL*) (Mitchell, 1997). Abordagens IBL podem construir aproximações diferentes da função objetivo para cada instância a ser classificada. O KNN constrói uma aproximação local para a função objetivo e a aplica nos vizinhos da instância a ser classificada. Isto possui significativas vantagens quando a função objetivo é muito complexa. A desvantagem do KNN é que ele tipicamente considera todos os

atributos das instâncias quando objetiva recuperar da memória exemplos de treinamento semelhantes (Joachims, 1998).

O algoritmo C4.5 (Quinlan, 1993) é um modelo de aprendizado de máquina preditivo que decide o valor objetivo de um novo exemplo baseado em vários valores de atributos dos dados disponíveis.

Resultados experimentais obtidos por Joachims (1998) mostraram que os SVMs consistentemente obtêm bom desempenho nas tarefas de categorização de documentos, superando os demais métodos. Contudo, Gabrilovich e Markovitch (2004) demonstraram que em algumas coleções o C4.5 supera o SVM e o KNN, embora o SVM seja considerado substancialmente superior para categorizações de documentos. De acordo com os autores, quando a seleção de características é executada, C4.5 constrói pequenas árvores de decisão que capturam o conceito melhor que o SVM e o KNN. Além disso, mesmo quando a seleção de características é otimizada para cada classificador, o C4.5 formula um poderoso modelo de classificação, o qual é significativamente superior ao KNN e marginalmente inferior ao SVM.

A implementação desses métodos está disponível na ferramenta Weka (Hall et al., 2009), cuja versão 3.6 foi utilizada para realização dos experimentos de avaliação dos métodos propostos para neste doutorado para extração de descritores de grupos. Os algoritmos Naive Bayes (NB), Multinomial Naive Bayes (M.Naive) e J48 (a implementação Weka do método C4.5) foram executados nos experimentos utilizando os parâmetros padrão da ferramenta. Apenas o desempenho do SVM foi refinado usando os *Normalized Polynomial Kernel* e o parâmetro de complexidade  $c=2.0$ . Esse refinamento foi necessário porque o SVM é conhecidamente sensível à inicialização de seus parâmetros. O IBk (implementação weka do método KNN) foi experimentado utilizando as quantidades de vizinhos variando de 1 a 7. Os melhores resultados foram obtidos utilizando 5 vizinhos.

Sendo assim, devido às características distintas de cada um, estes algoritmos de classificação são boas opções para avaliar os descritores de grupos obtidos pelos métodos propostos e verificar a acurácia dos descritores quando utilizados por diferentes tipos de algoritmos de classificação.

A seguir, são avaliados os métodos propostos neste doutorado, os quais foram apresentados nas Seções 4.2, 4.3 e 4.4, respectivamente.

#### 4.5.4 Avaliação do método SoftO-FDCL

Visando a organização flexível de documentos, o método SoftO-FDCL foi proposto para extrair descritores de grupos obtidos do agrupamento fuzzy *flat* de documentos. Sendo assim, é importante que os descritores sejam o mais representativos possíveis para a coleção de documentos. Para verificar se os descritores extraídos pelo método SoftO-FDCL satisfazem esta afirmação, foram realizadas quatro avaliações: avaliação qualitativa, avaliação quantitativa, teste possibilístico e comparação com métodos de seleção de atributos, as quais são apresentadas a seguir.

## Avaliação quantitativa

A avaliação quantitativa do método SoftO-FDCL foi realizada para verificar o poder preditivo dos descritores extraídos. Para tanto, os descritores de grupos são considerados atributos dos documentos e a classe do documento é considerado o grupo no qual ele possui maior grau de pertinência. Assim, a qualidade dos descritores é medida pelo desempenho dos algoritmos de classificação e de uma análise comparativa entre o desempenho dos mesmos algoritmos de classificação utilizando descritores obtidos pelo método centroide (Manning et al., 2008). Por meio desse método, os termos que ocorrem com maior frequência no vetor que corresponde ao centroide de um determinado grupo são selecionados como descritores deste grupo.

Utilizando os descritores de grupos extraídos pelos métodos Centróide e SoftO-FDCL foram obtidas, portanto, taxas de acerto pelos métodos de classificação: SVM, Naive Bayes (NB), Multinomial Naive Bayes (M.Naive), KNN e C4.5.

A avaliação quantitativa do método SoftO-FDCL foi realizada utilizando a coleção de documentos Opnosis, descrita anteriormente na Seção 4.5.1. Vale ressaltar que, para um vocabulário grande como o vocabulário da coleção Opnosis, uma quantidade razoável de descritores deve ser considerada para a categorização de documentos. O vocabulário da coleção Opnosis é considerado grande devido à diversidade dessa coleção. A mesma é composta por 51 documentos distribuídos em 3 categorias, ou seja, os documentos são compostos por palavras que distinguem cada uma das três categorias. Assim, considerando o tamanho dos documentos, a quantidade de 100 descritores foi escolhida arbitrariamente para representar cada grupo.

Um exemplo de descritores obtidos pelo método SoftO-FDCL, para a organização da coleção Opnosis, pode ser observado na Tabela 4.6. Observe que os termos foram stem-mizados durante o pré-processamento dos documentos, conforme apresentado na Seção 2.2 do Capítulo 2.

As taxas de acerto obtidas para cada classificador são apresentadas na Tabela 4.7. Os melhores resultados estão destacados em cinza. Com estas taxas, a representatividade dos descritores obtidos pelo método SoftO-FDCL e pelo método centroide foi analisada.

Estes resultados foram apresentados por Nogueira et al. (2011a) concluindo-se que o método SoftO-FDCL extraí bons descritores para os grupos obtidos da coleção Opnosis. É importante ressaltar que o alto desvio padrão foi obtido devido à pequena quantidade de documentos ter sido avaliada usando *10-fold cross validation*.

Além da extração dos descritores, foi obtida a organização flexível da coleção Opnosis pela distribuição dos documentos em mais de um grupo. Considerando que os graus de pertinência dos documentos nos grupos refletem a compatibilidade dos documentos com os tópicos representados pelos descritores de grupos, a organização flexível permite organizar documentos em vários tópicos simultaneamente, uma vez que documentos podem abordar vários tópicos ao mesmo tempo.

Tabela 4.6: Vinte descritores com maior valor de  $f_1$  obtidos pelo método SoftO-FDCL para cada grupo da coleção Opinosis

|                |   |
|----------------|---|
| <b>Grupo 1</b> | easi read, touch screen, post speed, gp unit, top notch, readi navig, coupl time, updat garmin, estim time arriv, miss turn, turn direct, year ago, speed limit, long trip, nice featur, time arriv, estim time, text speech, post speed limit, turn turn   |
| <b>Grupo 2</b> | place stai, staff friendli, servic room, great locat, hotel great, front desk, book room, conveni locat, great room, nice touch, room servic, comfort room, room larg, good servic, room nice, clean room, room small, locat great, bed comfort, room clean   |
| <b>Grupo 3</b> | increa batteri life, keyboard larg comfort type, keyboard larg, long batteri life, comfort type, usb port, decent size screen, speed perform, keyboard great, light weight, long batteri, full size, screen small, keyboard larg comfort, life great, make hard, batteri life great, remov batteri, keyboard layout, batteri life |
| <b>Grupo 4</b> | radio excel, sound qualiti, interior qualiti, screen bi, great sound, great screen, qualiti good, qualiti interior, love pedomet, batteri longer, sound system, ipod nano, gen nano, video camera, love fact, design flaw, ga mileag, love car, built speaker, batteri life   |

Tabela 4.7: Taxas de acerto obtidas pelos algoritmos de classificação utilizando os descritores extraídos pelos métodos Centroide e SoftO-FDCL

| Algoritmo de classificação | Centroide   | SoftO-FDCL  |
|----------------------------|-------------|-------------|
| SVM                        | 0,56 (0,21) | 0,70 (0,24) |
| NB                         | 0,44 (0,19) | 0,79 (0,17) |
| MM.Naive                   | 0,58 (0,23) | 0,77 (0,19) |
| KNN-5                      | 0,44 (0,16) | 0,60 (0,24) |
| J48                        | 0,55 (0,22) | 0,45 (0,18) |

O método SoftO-FDCL permite ainda a organização dos documentos em tópicos de maneira totalmente não-supervisionada, *i.e.*, sem participação do especialista de domínio. Logo, esta organização pode ser generalizada e explorada por diferentes usuários.

### Avaliação qualitativa

Em adição à avaliação quantitativa apresentada anteriormente, uma avaliação qualitativa do método SoftO-FDCL foi realizada também utilizando a coleção Opinosis. A orga-

nização flexível da coleção Opinosis apresentada no formato matriz documentos-grupos na Tabela 4.8 e os tópicos identificados por descritores de grupos já apresentados na Tabela 4.6 são os principais resultados dessa avaliação.

Na primeira coluna da Tabela 4.8 tem-se a identificação (ID) de cada documento da coleção Opinosis. Na segunda coluna, a característica do produto revisado em cada documento é apresentada, e nas demais colunas tem-se os graus de pertinência de cada documento nos quatro grupos obtidos a partir do agrupamento fuzzy (Grupo 1, Grupo 2, Grupo 3 e Grupo 4). Nesta tabela, os valores destacados em negrito representam o maior grau de pertinência do documento.

Para explorar a avaliação qualitativa, considere o documento D20 na Tabela 4.8. Esse documento é composto pela revisão de consumidores sobre o teclado do computador portátil 1005ha (“*keyboard of the netbook 1005ha*”). Esse documento tem grau de compatibilidade com os grupos 1, 2, 3 e 4, igual a 0.21, 0.19, 0.38, e 0.21, respectivamente. A organização do documento D20 pode apresentar a seguinte interpretação. Uma vez que ele possui maior compatibilidade com o grupo 3, ele é compatível com o tópico relacionado às características do produto teclado, uma vez que o grupo 3 é representado por descritores que identificam o tópico teclado (*keyboard*). Por outro lado, o documento D20 possui menor compatibilidade com o grupo 2. Como esperado, esse grupo é representado por descritores mais próximos em significado das palavras presentes nos documentos sobre hotéis, os quais não tem relação com o documento D20. Entretanto, o documento D20 também tem grau de compatibilidade com o grupo 4 e o grupo 1, os quais são representados por descritores que identificam os tópicos qualidade (*quality*) e tempo (*time*), respectivamente.

Os grupos 4 e 1 são representativos da organização flexível, uma vez que documentos compostos por revisões sobre diferentes características de um produto (hotel, carro ou produtos eletrônicos) podem abordar os tópicos identificados pelos descritores de ambos os grupos. Por exemplo, documentos sobre hotéis e/ou sobre carros podem ser compatíveis com o tópico relacionado a tempo, se os revisores avaliarem o tempo gasto para chegar, de carro, a um destino a partir de um hotel. Além disso, um produto eletrônico também pode ser revisado pelo tempo de sua bateria. Como outro exemplo, os produtos podem ser compatíveis com o tópico relacionado à qualidade, uma vez que todos eles podem ser revisados de acordo com a qualidade de suas características.

Uma vez obtida essa organização, foi avaliado se os descritores de grupos extraídos pelo método SoftO-FDCL capturaram a informação essencial da coleção Opinosis, no sentido de que as principais palavras utilizadas por humanos em seus sumários são similares aos descritores extraídos. Essa avaliação foi realizada observando se a frequência dos descritores, extraídos automaticamente, nos sumários dos documentos feitos por humanos aumenta quando a compatibilidade de um documento com um tópico também aumenta.

Na Figura 4.2 tem-se quatro gráficos, um para cada grupo, os quais representam os graus de compatibilidade (pertinência) (representados por linhas no gráfico) dos docu-

Tabela 4.8: Organização flexível da coleção Opinosis representada no formato documentos-grupos obtida pelo método SoftO-FDCL

| ID  | Característica do produto       | Grupo 1        | Grupo 2        | Grupo 3        | Grupo 4        |
|-----|---------------------------------|----------------|----------------|----------------|----------------|
| D1  | accuracy_garmin_nuvi_255W_gps   | <b>0,99998</b> | 0,00001        | 0,00001        | 0,00001        |
| D2  | bathroom_bestwestern_hotel_sfo  | 0,23280        | <b>0,30607</b> | 0,22693        | 0,23420        |
| D3  | battery-life_amazon_kindle      | 0,22154        | 0,22052        | <b>0,30602</b> | 0,25192        |
| D4  | battery-life_ipod_nano_8gb      | 0,00003        | 0,00002        | 0,00003        | <b>0,99992</b> |
| D5  | battery-life_netbook_1005ha     | 0,00000        | 0,00000        | <b>0,99999</b> | 0,00000        |
| D6  | buttons_amazon_kindle           | 0,24933        | 0,24888        | 0,25046        | <b>0,25133</b> |
| D7  | comfort_honda_accord_2008       | 0,25001        | 0,25066        | 0,24837        | <b>0,25096</b> |
| D8  | comfort_toyota_camry_2007       | <b>0,25170</b> | 0,24938        | 0,24978        | 0,24913        |
| D9  | directions_garmin_nuvi_255W_gps | <b>0,68597</b> | 0,10174        | 0,10417        | 0,10812        |
| D10 | display_garmin_nuvi_255W_gps    | <b>0,31475</b> | 0,22445        | 0,22675        | 0,23405        |
| D11 | eyesight-issues_amazon_kindle   | <b>0,25169</b> | 0,24765        | 0,25108        | 0,24959        |
| D12 | features_windows7               | 0,25086        | 0,24787        | <b>0,25153</b> | 0,24974        |
| D13 | fonts_amazon_kindle             | <b>0,25169</b> | 0,25008        | 0,24766        | 0,25057        |
| D14 | food_holiday_inn_london         | 0,20458        | <b>0,38809</b> | 0,20258        | 0,20475        |
| D15 | food_swissotel_chicago          | 0,24767        | <b>0,25456</b> | 0,24651        | 0,25126        |
| D16 | free_bestwestern_hotel_sfo      | 0,24750        | <b>0,25781</b> | 0,24646        | 0,24823        |
| D17 | gas_mileage_toyota_camry_2007   | 0,24928        | 0,24839        | <b>0,25297</b> | 0,24936        |
| D18 | interior_honda_accord_2008      | <b>0,25292</b> | 0,24903        | 0,24883        | 0,24922        |
| D19 | interior_toyota_camry_2007      | 0,24893        | 0,24999        | 0,24829        | <b>0,25279</b> |
| D20 | keyboard_netbook_1005ha         | 0,21047        | 0,18917        | <b>0,38453</b> | 0,21582        |
| D21 | location_bestwestern_hotel_sfo  | 0,24664        | <b>0,26283</b> | 0,24282        | 0,24771        |
| D22 | location_holiday_inn_london     | 0,20394        | <b>0,39104</b> | 0,20048        | 0,20455        |
| D23 | mileage_honda_accord_2008       | 0,25114        | 0,24802        | 0,24870        | <b>0,25214</b> |
| D24 | navigation_amazon_kindle        | <b>0,25013</b> | 0,25007        | 0,24982        | 0,24998        |
| D25 | parking_bestwestern_hotel_sfo   | 0,24879        | <b>0,26096</b> | 0,24422        | 0,24603        |
| D26 | performance_honda_accord_2008   | 0,24944        | <b>0,25175</b> | 0,24936        | 0,24945        |
| D27 | performance_netbook_1005ha      | 0,05800        | 0,05487        | <b>0,82842</b> | 0,05870        |
| D28 | price_amazon_kindle             | 0,24646        | <b>0,25125</b> | 0,25307        | 0,24922        |
| D29 | price_holiday_inn_london        | 0,20791        | <b>0,37717</b> | 0,20596        | 0,20896        |
| D30 | quality_toyota_camry_2007       | 0,24950        | <b>0,25054</b> | 0,24949        | 0,25048        |
| D31 | room_holiday_inn_london         | 0,00000        | <b>0,99999</b> | 0,00000        | 0,00000        |
| D32 | rooms_bestwestern_hotel_sfo     | 0,23290        | <b>0,31676</b> | 0,21243        | 0,23790        |
| D33 | rooms_swissotel_chicago         | 0,23404        | <b>0,31132</b> | 0,21400        | 0,24064        |
| D34 | satellite_garmin_nuvi_255W_gps  | <b>0,29992</b> | 0,23188        | 0,23124        | 0,23696        |
| D35 | screen_garmin_nuvi_255W_gps     | <b>0,26127</b> | 0,24102        | 0,24743        | 0,25027        |
| D36 | screen_ipod_nano_8gb            | 0,16818        | 0,15751        | 0,16332        | <b>0,51099</b> |
| D37 | screen_netbook_1005ha           | 0,13792        | 0,12259        | <b>0,59537</b> | 0,14413        |
| D38 | seats_honda_accord_2008         | 0,24917        | 0,25043        | 0,24985        | <b>0,25056</b> |
| D39 | service_bestwestern_hotel_sfo   | 0,24755        | <b>0,26357</b> | 0,24027        | 0,24860        |
| D40 | service_holiday_inn_london      | 0,05528        | <b>0,83669</b> | 0,05245        | 0,05558        |
| D41 | service_swissotel_hotel_chicago | 0,24612        | <b>0,27135</b> | 0,23661        | 0,24592        |
| D42 | size_asus_netbook_1005ha        | 0,18775        | 0,17004        | <b>0,44958</b> | 0,19263        |
| D43 | sound_ipod_nano_8gb             | 0,15830        | 0,14842        | 0,15039        | <b>0,54289</b> |
| D44 | speed_garmin_nuvi_255W_gps      | <b>0,45658</b> | 0,17701        | 0,17716        | 0,18924        |
| D45 | speed_windows7                  | 0,24717        | 0,24179        | <b>0,26343</b> | 0,24761        |
| D46 | staff_bestwestern_hotel_sfo     | 0,24674        | <b>0,26955</b> | 0,23401        | 0,24969        |
| D47 | staff_swissotel_chicago         | 0,24875        | <b>0,26227</b> | 0,23940        | 0,24957        |
| D48 | transmission_toyota_camry_2007  | 0,24891        | 0,25034        | <b>0,25064</b> | 0,25011        |
| D49 | updates_garmin_nuvi_255W_gps    | <b>0,29744</b> | 0,23134        | 0,23058        | 0,24064        |
| D50 | video_ipod_nano_8gb             | 0,16080        | 0,15033        | 0,15751        | <b>0,53136</b> |
| D51 | voice_garmin_nuvi_255W_gps      | <b>0,28940</b> | 0,23193        | 0,23555        | 0,24312        |

mentos da coleção Opinosis com o grupo no qual esse documento tem grau de pertinência maior ou igual ao limiar  $\delta = \frac{1}{c}$ , para  $c$  igual a quantidade de grupos, ou seja, documentos que possuem grau de pertinência maior ou igual a 0,25; e a frequência (representada por barras no gráfico) dos 50 melhores descritores do grupo nos sumários dos documentos. O eixo  $y$  dos gráficos representa a escala dos graus de compatibilidade e frequência. No eixo  $x$  dos gráficos tem-se os documentos que possuem maior grau de compatibilidade com o grupo representado pelo gráfico correspondente.

Nos gráficos apresentados na Figura 4.2 observa-se que o documento D3 possui grau de compatibilidade 0,31 com o grupo 3 e os descritores do grupo 3 ocorrem nos sumários do documento D3 com 80% de frequência, ou seja, 80% dos descritores extraídos pelo método SoftO-FDCL foram utilizados nos sumários do documento D3. Por outro lado, o documento D1 tem grau de compatibilidade 0,99 com o grupo 1 e os descritores do grupo 1 ocorrem nos sumários do documento D1 com 10% de frequência. Dessa observação, conclui-se que alguns descritores são mais frequentes nos sumários dos documentos que possuem baixa compatibilidade com um grupo do que nos sumários dos documentos que possuem alta compatibilidade com um grupo. Logo, os descritores que apresentam menor relevância para representar um determinado grupo são aqueles que ocorrem com alta frequência nos sumários dos documentos que possuem baixa compatibilidade com este grupo.

Por outro lado, uma característica específica é observada no documento D1. possuem grau de compatibilidade 0,99 com algum grupo. Ele possui grau de compatibilidade 0,99 com o grupo 1, mas os descritores do grupo 1 possuem baixa frequência nos sumários desse documento. Uma explicação para isto é que os descritores são obtidos a partir de toda a coleção de documentos, enquanto a frequência utilizada na avaliação apresentada na Figura 4.2 foi computada sobre os sumários escritos por humanos. Portanto, existem palavras que foram utilizadas nos sumários mas que não foram utilizadas nos documentos, especialmente palavras sinônimas. Além disso, os sumários são menores que os documentos.

Esses resultados foram apresentados por Nogueira et al. (2012a), concluindo-se que o método SoftO-FDCL extraí descritores de grupos que capturam a informação essencial dos documentos a serem organizados de maneira flexível, uma vez que grande parte dos descritores extraídos pelo método SoftO-FDCL foram também utilizados nos sumários escritos por humanos.

### Teste possibilístico

Uma vez que em métodos de extração de descritores do tipo DCL o agrupamento é separado da extração de descritores, qualquer algoritmo de agrupamento fuzzy pode ser utilizado para agrupar os documentos. No entanto, os graus de pertinência obtidos podem interferir na qualidade dos descritores, já que cada algoritmo de agrupamento fuzzy tem sua própria definição de grau de pertinência.

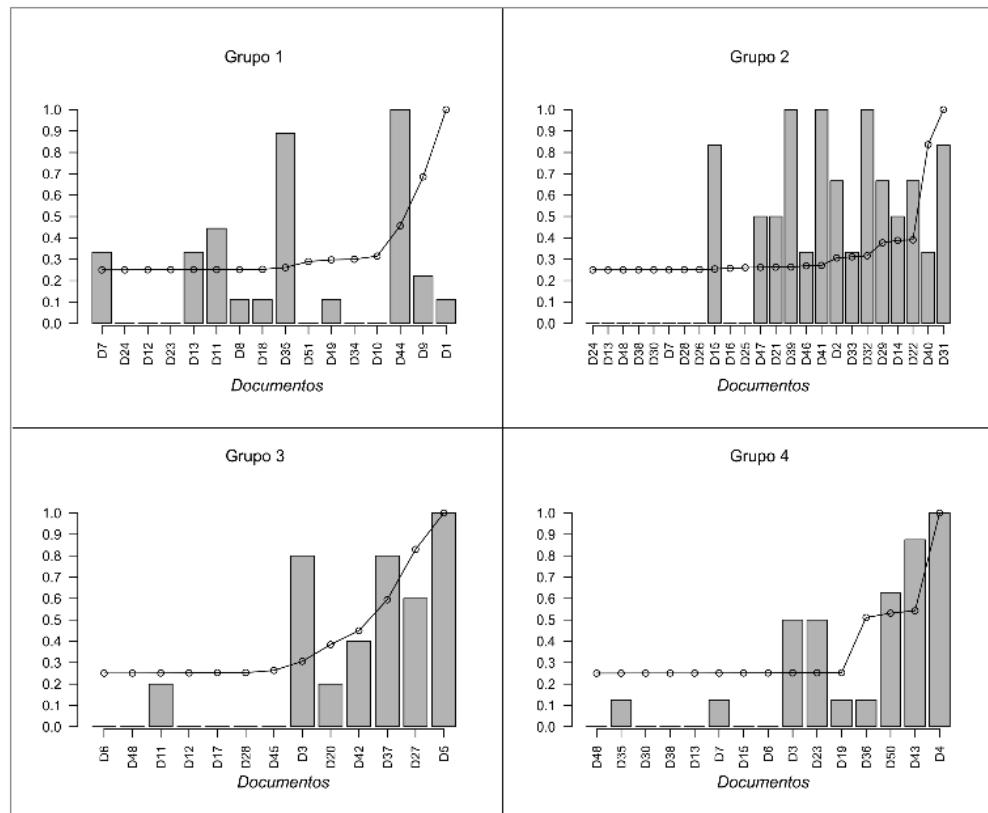


Figura 4.2: Frequênciade 50 descritores nos sumários dos documentos que possuem grau de pertinência no grupo representado por um gráfico maior ou igual ao limiar  $\delta=0,25$

Sendo assim, uma vez que o método SoftO-FDCL depende dos graus de pertinência obtidos do processo de agrupamento fuzzy para ponderar um termo candidato a descritor de grupo, é importante verificar o quanto os graus de pertinência influenciam na extração de descritores de grupos fuzzy.

Como exemplo de verificação de tal influência, considere a situação na qual existem 3 documentos compostos por 3 termos cada um, para a qual a matriz documentos-termos é ilustrada na Tabela 4.9. Esses documentos foram agrupados em 2 grupos utilizando os algoritmos de agrupamento Fuzzy C-Means (FCM) (Bezdek, 1981) e Possibilístico C-Means (PCM) (Pal et al., 2005), ambos apresentados no Capítulo 2. Considere a matriz documentos-grupos ilustrada na Tabela 4.10 obtida do algoritmo FCM. Para esse exemplo, o método SoftO-FDCL é inicialmente executado utilizando essas duas matrizes.

Tabela 4.9: Exemplo de matriz documentos-termos

| Documentos | Termos |       |       |
|------------|--------|-------|-------|
|            | $t_1$  | $t_2$ | $t_3$ |
| $d_1$      | 0      | 0     | 1     |
| $d_2$      | 1      | 1     | 1     |
| $d_3$      | 0      | 1     | 1     |

Tabela 4.10: Matriz documentos-grupos obtida do agrupamento FCM para o exemplo da Tabela 4.9

| Documentos | Grupos |       |
|------------|--------|-------|
|            | $g_1$  | $g_2$ |
| $d_1$      | 0,5    | 0,5   |
| $d_2$      | 0,3    | 0,7   |
| $d_3$      | 0,6    | 0,4   |

Considerando cada termo na Tabela 4.9 como candidato a descritor de grupo, conforme definição do método SoftO-FDCL apresentado anteriormente na Seção 4.2, a extração de descritores de um determinado grupo tem início com a ponderação dos termos candidatos a descritores pela medida  $f1$ . Para o exemplo apresentado, obteve-se os seguintes valores de  $f1$  para todos os termos em todos os grupos utilizando a informação de pertinência apresentada na Tabela 4.10 :  $f1(t_1, g_1) = 0$ ,  $f1(t_2, g_1) = 0.5$ ,  $f1(t_3, g_1) = 0.79$ ,  $f1(t_1, g_2) = 0.5$ ,  $f1(t_2, g_2) = 0.66$ ,  $f1(t_3, g_2) = 1.0$

Para verificar a influência do grau de pertinência no método SoftO-FDCL, considere a matriz documentos-grupos obtida do agrupamento PCM na Tabela 4.11.

Tabela 4.11: Matriz documentos-grupos obtida do agrupamento PCM para o exemplo da Tabela 4.9

| Documentos | Grupos |       |
|------------|--------|-------|
|            | $g_1$  | $g_2$ |
| $d_1$      | 0,3    | 0,3   |
| $d_2$      | 0,3    | 0,7   |
| $d_n$      | 0,6    | 0,4   |

Aplicando o método SoftO-FDCL utilizando a informação de pertinência apresentada na Tabela 4.11, obteve-se os seguintes valores de  $f1$  para todos os termos em todos os grupos:  $f1(t_1, g_1) = 0$ ,  $f1(t_2, g_1) = 1.0$ ,  $f1(t_3, g_1) = 0.49$ ,  $f1(t_1, g_2) = 1.0$ ,  $f1(t_2, g_2) = 0.66$ ,  $f1(t_3, g_2) = 0.49$ .

Sendo assim, a partir do exemplo apresentado, foram obtidos diferentes *rankings* de termos candidatos a descritores. Aplicando o método SoftO-FDCL após o agrupamento de documentos realizado por meio do algoritmo FCM, obteve-se um *ranking* de termos candidatos a descritores igual a  $t_1 < t_2 < t_3$  para o grupo 1 e  $t_1 < t_2 < t_3$  para o grupo 2. Aplicando o método SoftO-FDCL após o agrupamento de documentos realizado por meio do algoritmo PCM, obteve-se um *ranking* de termos candidatos a descritores igual a  $t_1 < t_3 < t_2$  para o grupo 1 e  $t_3 < t_2 < t_1$  para o grupo 2. Portanto, há indícios de que essa diferença pode afetar a organização de documentos por meio de descritores de grupos.

Diante disto, para verificar se a diferença de *rankings* de termos obtidos pelo método SoftO-FDCL após o agrupamento de documentos realizado por meio do algoritmo FCM e PCM influencia na qualidade dos descritores obtidos, alguns experimentos foram

realizados utilizando diferentes coleções de documentos. Para avaliar a qualidade dos descritores, o poder de predição dos descritores obtidos pelo método SoftO-FDCL foi medido considerando que os mesmos são bons atributos para categorização dos documentos. Para tanto, os descritores de grupos são considerados atributos dos documentos e a classe do documento é o grupo no qual ele possui maior grau de pertinência.

Para essa avaliação, algoritmos de classificação conhecidos e definidos na Seção 4.5.3 foram executados. Os resultados obtidos são apresentados nas Tabelas 4.12, 4.13, 4.14, 4.15, 4.16 e 4.17.

Tabela 4.12: Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção Opinosis)

| <b>Algoritmo de Classificação</b> | <b>PCM</b>   | <b>FCM</b>   |
|-----------------------------------|--------------|--------------|
| NB                                | 84,00(15,78) | 66,00(18,97) |
| M.Naive                           | 88,00(10,33) | 80,00(21,08) |
| KNN                               | 62,00(19,89) | 62,00(17,51) |
| SVM                               | 82,00(19,89) | 62,00(22,01) |
| J48                               | 68,00(21,50) | 54,00(28,36) |

Tabela 4.13: Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção 20Newsgroups)

| <b>Algoritmo de Classificação</b> | <b>PCM</b>  | <b>FCM</b>  |
|-----------------------------------|-------------|-------------|
| NB                                | 58,43(3,20) | 62,53(2,45) |
| M.Naive                           | 48,82(1,69) | 40,17(2,27) |
| KNN                               | 54,03(2,66) | 52,18(4,25) |
| SVM                               | 66,08(2,66) | 69,93(1,57) |
| J48                               | 64,63(2,74) | 62,83(2,82) |

Tabela 4.14: Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção Reuters-21578)

| <b>Algoritmo de Classificação</b> | <b>PCM</b>  | <b>FCM</b>  |
|-----------------------------------|-------------|-------------|
| NB                                | 20,37(3,54) | 61,47(3,40) |
| M.Naive                           | 50,71(1,21) | 97,34(1,25) |
| KNN                               | 59,85(3,75) | 98,10(0,44) |
| SVM                               | 55,38(6,00) | 98,67(0,92) |
| J48                               | 61,00(4,23) | 98,57(0,67) |

#### 4.5. Avaliação dos métodos propostos

---

Tabela 4.15: Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção WAP)

| <b>Algoritmo de Classificação</b> | <b>PCM</b>  | <b>FCM</b>  |
|-----------------------------------|-------------|-------------|
| NB                                | 53,37(3,75) | 51,19(3,37) |
| M.Naive                           | 56,77(1,91) | 92,62(0,88) |
| KNN                               | 53,63(3,64) | 93,14(0,87) |
| SVM                               | 71,84(3,50) | 91,79(0,58) |
| J48                               | 66,84(4,29) | 95,70(1,42) |

Tabela 4.16: Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção Hitech)

| <b>Algoritmo de Classificação</b> | <b>PCM</b>  | <b>FCM</b>  |
|-----------------------------------|-------------|-------------|
| NB                                | 40,72(5,45) | 38,21(5,80) |
| M.Naive                           | 41,82(4,95) | 41,27(3,43) |
| SVM                               | 46,98(5,63) | 46,14(5,14) |
| KNN                               | 43,46(6,30) | 35,05(5,36) |
| J48                               | 36,76(6,65) | 36,09(6,03) |

Tabela 4.17: Qualidade dos descritores extraídos pelo método SoftO-FDCL para os grupos obtidos pelos algoritmos PCM e FCM (Coleção NSF)

| <b>Algoritmo de Classificação</b> | <b>PCM</b>  | <b>FCM</b>  |
|-----------------------------------|-------------|-------------|
| NB                                | 90,62(1,55) | 93,56(1,82) |
| M.Naive                           | 85,90(0,57) | 93,94(0,16) |
| SVM                               | 98,25(0,97) | 99,94(0,13) |
| KNN                               | 93,37(0,78) | 97,59(0,87) |
| J48                               | 97,22(1,18) | 98,19(0,60) |

Devido a diferença de resultados obtidos utilizando o FCM e o PCM para essas bases, percebeu-se que o método SoftO-FDCL é dependente do conceito de pertinência embutido no agrupamento fuzzy utilizado. Logo, é importante considerar as especificações do algoritmo de agrupamento fuzzy escolhido para agrupar os documentos, de forma que o agrupamento obtido seja o mais apropriado para a coleção a ser organizada.

Considerando os resultados obtidos nessa avaliação, concluiu-se que o algoritmo FCM foi mais apropriado para a organização dos documentos das coleções NSF, WAP e Reuters, enquanto o algoritmo PCM foi mais apropriado para as coleções Opinosis, 20Newsgroups e Hitech. Portanto, observa-se que os descritores extraídos de grupos de coleções com um menor número de classes representam melhor os grupos obtidos pelo algoritmo PCM do que os grupos obtidos pelo algoritmo FCM. Embora a quantidade de classes das coleções não corresponda à quantidade de grupos obtida pelos algoritmos de agrupamento, nem

a informação de classe seja utilizada no processo de agrupamento, o algoritmo FCM tem melhor desempenho sobre uma quantidade maior de grupos porque o algoritmo PCM pode apresentar o problema de grupos coincidentes.

O problema de grupos coincidentes ocorre quando a inicialização do agrupamento, a qual está relacionada a matriz de tipicidades inicial (definições no Capítulo 2) não é suficientemente distinta, *i.e.*, a execução do algoritmo de agrupamento resulta em  $c$  grupos, embora a quantidade correta de grupos seja  $c' < c$ . Quando uma coleção apresenta muitas classes de documentos, os vetores de características que definem os documentos são usualmente esparsos, uma vez que documentos de diferentes classes são compostos por diferentes termos. Isso dificulta a inicialização do algoritmo PCM.

Na avaliação sobre os algoritmos FCM e PCM, a medida Silhueta Fuzzy (Campello e Hruschka, 2006), apresentada no Capítulo 2, foi utilizada para avaliar o agrupamento de documentos e escolher a quantidade apropriada de grupos para organizar as coleções. Além disso, o algoritmo PCM foi inicializado utilizando as saídas do algoritmo FCM, como sugerido por Krishnapuram e Keller (1993) para evitar o problema de inicialização. No entanto, os resultados obtidos nos experimentos indicam que não há garantia de que  $c' = c$  é a quantidade de grupos correta, inclusive quando o PCM e a medida FS sugerem isso.

### Comparação com métodos de seleção de atributos

No processo de classificação de documentos, a seleção de atributos é uma tarefa bastante importante. Por meio da seleção de atributos o subconjunto de termos que serão utilizados no processo de classificação é selecionado com dois propósitos: primeiro, ela reduz a quantidade de termos a serem analisados no processo de classificação; segundo, ela melhora a acurácia da classificação pela eliminação de ruídos, os quais são termos que levam a classificação errada dos documentos (Manning et al., 2008).

Sendo assim, uma vez que o método SoftO-FDCL extrai descritores de grupos, os quais são considerados como atributos importantes dos documentos distribuídos nos grupos, foi realizada uma comparação entre o desempenho do método SoftO-FDCL e dois conhecidos métodos de seleção de atributos: Informação Mútua (*Mutual Information* - MI) e Qui-quadrado (*chi-squared* -  $\chi^2$ ), os quais são também utilizados para extrair descritores de grupos (Popescul e Ungar, 2000; Treeratpituk e Callan, 2006; Manning et al., 2008; Chitsaz et al., 2009; Muhr et al., 2010).

O método MI para seleção de atributos mede o grau de dependência entre duas variáveis. No contexto da extração de descritores, uma variável refere-se ao termo  $t_j$  candidato a descritor do grupo  $g_l$ , o qual também é considerado uma variável. Assim, a informação mútua entre um termo  $t_j$  e um grupo  $g_l$  é medido conforme Equação (4.29).

$$MI(g_l, t_j) = \sum_{g_l \in \{0,1\}} \sum_{t_j \in \{0,1\}} prob(g_l, t_j) \log_2 \frac{prob(g_l, t_j)}{prob(g_l)prob(t_j)}, \quad (4.29)$$

no qual  $prob(g_l = 1)$  representa a probabilidade de um documento pertencer ao grupo  $g_l$ ,  $prob(g_l = 0)$  representa a probabilidade de um documento não pertencer ao grupo  $g_l$ ,  $prob(t_j = 1)$  representa a probabilidade do termo  $t_j$  ocorrer em um documento e  $prob(t_j = 0)$  representa a probabilidade do termo  $t_j$  não ocorrer em um documento. Sendo assim, considerando a matriz de contingência utilizada pelo método SoftO-FDCL apresentada na Tabela 4.1, as probabilidades  $prob(g_l, t_j)$ ,  $prob(g_l)$  e  $prob(t_j)$  são calculadas como segue, para  $n$  igual a quantidade de documentos da coleção.

- $prob(g_l = 1, t_j = 1) = \frac{ganhos(t_j, g_l)}{n}$
- $prob(g_l = 1, t_j = 0) = \frac{perdas(t_j, g_l)}{n}$
- $prob(g_l = 0, t_j = 1) = \frac{ruidos(t_j, g_l)}{n}$
- $prob(g_l = 0, t_j = 0) = \frac{rejeitos(t_j, g_l)}{n}$
- $prob(g_l = 1) = \frac{ganhos(t_j, g_l) + perdas(t_j, g_l)}{n}$
- $prob(g_l = 0) = \frac{ruidos(t_j, g_l) + rejeitos(t_j, g_l)}{n}$
- $prob(t_j = 1) = \frac{ganhos(t_j, g_l) + ruidos(t_j, g_l)}{n}$
- $prob(t_j = 0) = \frac{perdas(t_j, g_l) + rejeitos(t_j, g_l)}{n}$

Assim, o método MI mede quanta informação um determinado termo possui sobre um grupo. Portanto, quanto maior o valor de MI de um termo, mais o termo representa o grupo, pois isso significa que o termo contém bastante informação sobre o grupo.

Por outro lado, o método  $\chi^2$  de seleção de atributos mede a probabilidade que a ocorrência de um evento aproxima-se da expectativa inicial, ou seja, mede a independência entre dois eventos. No contexto da extração de descritores, um evento refere-se ao termo  $t_j$  candidato a descritor do grupo  $g_l$ , o qual também é considerado uma variável. Assim, a independência estatística entre um termo  $t_j$  e um grupo  $g_l$  é medido conforme Equação (4.30).

$$X^2(g_l, t_j) = \sum_{g_l \in \{0,1\}} \sum_{t_j \in \{0,1\}} \frac{(O_{g_l, t_j} - E_{g_l, t_j})^2}{E_{g_l, t_j}}, \quad (4.30)$$

no qual  $O_{g_l, t_j}$  refere-se à quantidade de documentos observados e  $E_{g_l, t_j}$  refere-se à quantidade de documentos esperados, os quais são calculados, considerando a matriz de contingência utilizada pelo método SoftO-FDCL apresentada na Tabela 4.1, como segue, para  $n$  igual a quantidade de documentos da coleção.

- $O_{g_l=0, t_j=0} = rejeitos(t_j, g_l)$
- $O_{g_l=0, t_j=1} = ruidos(t_j, g_l)$
- $O_{g_l=1, t_j=0} = perdas(t_j, g_l)$

- $O_{g_l=1,t_j=1} = ganhos(t_j, g_l)$
- $E_{g_l=0,t_j=0} = n \cdot prob(g_l = 0) \cdot prob(t_j = 0)$
- $E_{g_l=0,t_j=1} = n \cdot prob(g_l = 0) \cdot prob(t_j = 1)$
- $E_{g_l=1,t_j=0} = n \cdot prob(g_l = 1) \cdot prob(t_j = 0)$
- $E_{g_l=1,t_j=1} = n \cdot prob(g_l = 1) \cdot prob(t_j = 1)$
- $prob(g_l = 1) = \frac{ganhos(t_j, g_l) + perdas(t_j, g_l)}{n}$
- $prob(g_l = 0) = \frac{ruidos(t_j, g_l) + rejeitos(t_j, g_l)}{n}$
- $prob(t_j = 1) = \frac{ganhos(t_j, g_l) + ruidos(t_j, g_l)}{n}$
- $prob(t_j = 0) = \frac{perdas(t_j, g_l) + rejeitos(t_j, g_l)}{n}$

Assim, o método  $\chi^2$  mede o quanto um termo é independente de um grupo. Portanto, quanto menor o valor de  $\chi^2$  do termo, mais o termo representa o grupo, pois isso significa que o termo é dependente do grupo.

Ao obter os descritores de grupos extraídos pelos métodos MI,  $\chi^2$  e SoftO-FDCL, os mesmos foram comparados, considerando que os descritores extraídos são bons atributos para classificar os documentos nos grupos. Sendo assim, os descritores de grupos são considerados atributos dos documentos e a classe do documento é o grupo no qual ele possui maior grau de pertinência.

Os resultados obtidos do processo de classificação utilizando os algoritmos de classificação SVM, Naive Bayes (NB), Multinomial Naive Bayes (M.Naive), KNN e C4.5, são apresentados nas Tabelas 4.18, 4.19, 4.20 e 4.21.

Tabela 4.18: Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e  $\chi^2$  (Coleção Opinosis)

| Algoritmo de Classificação | SoftO-FDCL   | MI           | $\chi^2$     |
|----------------------------|--------------|--------------|--------------|
| NB                         | 66,00(18,97) | 72,00(13,98) | 74,00(16,47) |
| M.Naive                    | 80,00(21,08) | 80,00(18,86) | 80,00(18,86) |
| KNN                        | 62,00(17,51) | 52,00(23,48) | 56,00(24,59) |
| SVM                        | 62,00(22,01) | 58,00(23,94) | 62,00(22,01) |
| J48                        | 54,00(28,36) | 50,00(23,57) | 52,00(23,48) |

#### 4.5. Avaliação dos métodos propostos

---

Tabela 4.19: Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e  $\chi^2$  (Coleção 20NewsGroups)

| <b>Algoritmo de Classificação</b> | <b>SoftO-FDCL</b> | <b>MI</b>   | <b><math>\chi^2</math></b> |
|-----------------------------------|-------------------|-------------|----------------------------|
| NB                                | 62,53(2,45)       | 61,78(2,89) | 61,43(2,61)                |
| M.Naive                           | 40,17(2,27)       | 44,17(2,03) | 44,22(2,26)                |
| KNN                               | 52,18(4,25)       | 59,78(2,71) | 59,63(3,25)                |
| SVM                               | 69,93(1,57)       | 71,03(1,65) | 69,88(2,38)                |
| J48                               | 62,83(2,82)       | 61,68(2,69) | 61,53(2,91)                |

Tabela 4.20: Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e  $\chi^2$  (Coleção Reuters)

| <b>Algoritmo de Classificação</b> | <b>SoftO-FDCL</b> | <b>MI</b>   | <b><math>\chi^2</math></b> |
|-----------------------------------|-------------------|-------------|----------------------------|
| NB                                | 61,47(3,40)       | 50,81(4,89) | 39,68(5,75)                |
| M.Naive                           | 97,34(1,25)       | 98,00(0,94) | 98,38(0,64)                |
| KNN                               | 98,10(0,44)       | 98,10(0,44) | 98,10(0,44)                |
| SVM                               | 98,67(0,92)       | 98,67(0,92) | 98,29(1,33)                |
| J48                               | 98,57(0,67)       | 98,38(0,64) | 98,57(0,67)                |

Tabela 4.21: Comparação entre o método SoftO-FDCL e os métodos de seleção de atributos MI e  $\chi^2$  (Coleção WAP)

| <b>Algoritmo de Classificação</b> | <b>SoftO-FDCL</b> | <b>MI</b>   | <b><math>\chi^2</math></b> |
|-----------------------------------|-------------------|-------------|----------------------------|
| NB                                | 50,67(5,21)       | 57,02(4,49) | 37,40(5,67)                |
| M.Naive                           | 63,95(1,12)       | 69,02(1,33) | 69,53(1,78)                |
| KNN                               | 61,45(2,81)       | 71,97(3,11) | 72,42(2,97)                |
| SVM                               | 66,07(2,67)       | 78,64(2,50) | 81,46(3,08)                |
| J48                               | 62,48(3,97)       | 77,55(3,43) | 77,61(2,52)                |

Observando os resultados obtidos, concluiu-se que, de maneira geral, o método de extração de descritores SoftO-FDCL é semelhante aos métodos de seleção de atributos MI e  $\chi^2$  quando esses são também utilizados como métodos para extração de descritores de grupo. O que os diferencia é o significado associado à cada método: O método SoftO-FDCL seleciona termos que melhor recuperam os documentos de um determinado grupo, o método MI seleciona termos que apresentam melhor informação mútua com um determinado grupo e o método  $\chi^2$  seleciona termos que apresentam maior dependência estatística com um determinado grupo.

Neste doutorado, o diferencial em utilizar qualquer um desses métodos (SoftO-FDCL, MI e  $\chi^2$ ), conforme apresentado na avaliação comparativa, é a organização flexível de documentos, uma vez que a matriz de contingência apresentada na Tabela 4.1 permite

que um documento pertença a mais de um grupo e que, portanto, os descritores de grupos identifiquem tópicos da organização flexível proposta.

O método SoftO-FDCL é considerado o principal método proposto neste doutorado, uma vez que o mesmo possibilita a organização flexível de documentos de maneira simplificada. Por este motivo, um maior número de avaliações foi realizado sobre o método SoftO-FDCL do que sobre os demais métodos, os quais foram propostos a fim de viabilizar melhorias na organização flexível de documentos.

#### 4.5.5 Avaliação do método SoftO-wFDCL

Conforme apresentado na Seção 4.3, o método SoftO-wFDCL é uma extensão do método SoftO-FDCL desenvolvida para extração de descritores de grupos fuzzy *flat* com a inserção dos graus de pertinência no cálculo da medida  $f1$  dos termos. Com essa extensão pretende-se obter descritores que sejam mais representativos para os grupos de documentos dos que os descritores extraídos pelo método SoftO-FDCL e o método Centroide.

Visando verificar o poder preditivo dos descritores extraídos pelo método SoftO-wFDCL e compará-lo com os métodos SoftO-FDCL e Centroide, algoritmos de classificação foram executados. Para tanto, os descritores extraídos foram considerados como atributos dos documentos e a classe do documento como o grupo no qual ele possui maior grau de pertinência. Assim, o desempenho dos métodos Centroide, SoftO-FDCL e SoftO-wFDCL foram comparados por meio das taxas de acerto obtidas pelos métodos de classificação: SVM, Naive Bayes (NB), Multinomial Naive Bayes (M.Naive), KNN e C4.5. As coleções utilizadas nessa avaliação foram: Opinosis, 20NewsGroups, Reuters e WAP, cujos resultados são apresentados nas Tabelas 4.22, 4.23, 4.24 e 4.25, respectivamente. Os melhores resultados para cada caso estão destacados em cinza.

Tabela 4.22: Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção Opinosis

| <b>Algoritmo de Classificação</b> | <b>Centroide</b> | <b>SoftO-FDCL</b> | <b>SoftO-wFDCL</b> |
|-----------------------------------|------------------|-------------------|--------------------|
| NB                                | 48,00(19,32)     | 66,00(18,97)      | 60,00(23,09)       |
| M.Naive                           | 40,00(18,86)     | 80,00(21,08)      | 68,00(21,50)       |
| KNN                               | 30,00(25,39)     | 62,00(17,51)      | 48,00(21,50)       |
| SVM                               | 42,00(27,41)     | 62,00(22,01)      | 52,00(27,00)       |
| J48                               | 44,00(33,73)     | 54,00(28,36)      | 56,00(26,33)       |

#### 4.5. Avaliação dos métodos propostos

---

Tabela 4.23: Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção 20NewsGroups

| Algoritmo de Classificação | Centroide   | SoftO-FDCL  | SoftO-wFDCL |
|----------------------------|-------------|-------------|-------------|
| NB                         | 45,82(1,90) | 62,53(2,45) | 56,38(3,03) |
| M.Naive                    | 37,37(0,84) | 40,17(2,27) | 38,32(1,18) |
| KNN                        | 43,42(3,29) | 52,18(4,25) | 52,92(2,04) |
| SVM                        | 49,57(2,30) | 69,93(1,57) | 61,78(2,16) |
| J48                        | 43,92(3,25) | 62,83(2,82) | 56,93(2,61) |

Tabela 4.24: Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção Reuters

| Algoritmo de Classificação | Centroide   | SoftO-FDCL  | SoftO-wFDCL |
|----------------------------|-------------|-------------|-------------|
| NB                         | 56,99(3,61) | 61,47(3,40) | 61,47(3,40) |
| M.Naive                    | 98,00(0,70) | 97,34(1,25) | 97,34(1,25) |
| KNN                        | 97,34(0,87) | 98,10(0,44) | 98,10(0,44) |
| SVM                        | 98,10(0,63) | 98,67(0,92) | 98,67(0,92) |
| J48                        | 98,00(0,29) | 98,57(0,67) | 98,57(0,67) |

Tabela 4.25: Avaliação comparativa entre os métodos SoftO-wFDCL e SoftO-FDCL utilizando a coleção WAP

| Algoritmo de Classificação | Centroide   | SoftO-FDCL  | SoftO-wFDCL |
|----------------------------|-------------|-------------|-------------|
| NB                         | 28,54(7,44) | 50,67(5,21) | 24,18(2,33) |
| M.Naive                    | 64,14(1,07) | 63,95(1,12) | 63,18(0,76) |
| KNN                        | 58,24(3,92) | 61,45(2,81) | 59,91(3,22) |
| SVM                        | 63,25(1,18) | 66,07(2,67) | 63,57(1,71) |
| J48                        | 61,07(2,37) | 62,48(3,97) | 58,69(3,51) |

Assim como o método SoftO-FDCL, o método SoftO-wFDCL também foi comparado com métodos de seleção de atributos para extração de descritores. No entanto, para uma comparação mais justa, considerou-se que os métodos MI e  $\chi^2$  extraíram descritores utilizando a matriz de contingência utilizada pelo método SoftO-wFDCL apresentada na Tabela 4.2. Por meio dessa matriz, a informação de pertinência é embutida no processo de ponderação de um termo candidato a descritor de grupo. Considerando tal característica, na avaliação comparativa entre os métodos SoftO-wFDCL, MI e  $\chi^2$ , os métodos MI e  $\chi^2$  são denominados wMI e w $\chi^2$ , respectivamente. Os resultados dessa avaliação são apresentados nas Tabelas 4.26, 4.27, 4.28 e 4.29.

Tabela 4.26: Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e  $w\chi^2$  (Coleção Opinosis)

| Algoritmo de Classificação | SoftO-wFDCL  | wMI          | $w\chi^2$    |
|----------------------------|--------------|--------------|--------------|
| NB                         | 60,00(23,09) | 56,00(24,59) | 54,00(23,19) |
| M.Naive                    | 68,00(21,50) | 62,00(23,94) | 64,00(22,71) |
| KNN                        | 48,00(21,50) | 34,00(13,50) | 38,00(14,76) |
| SVM                        | 52,00(27,00) | 50,00(23,57) | 54,00(21,19) |
| J48                        | 56,00(26,33) | 54,00(21,19) | 52,00(21,50) |

Tabela 4.27: Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e  $w\chi^2$  (Coleção 20NewsGroups)

| Algoritmo de Classificação | SoftO-wFDCL | wMI         | $w\chi^2$   |
|----------------------------|-------------|-------------|-------------|
| NB                         | 56,38(3,03) | 57,83(2,66) | 57,83(2,66) |
| M.Naive                    | 38,32(1,18) | 38,62(1,87) | 38,62(1,87) |
| KNN                        | 52,92(2,04) | 51,68(3,09) | 51,68(3,09) |
| SVM                        | 61,78(2,16) | 58,83(2,93) | 58,83(2,93) |
| J48                        | 56,93(2,61) | 54,88(1,53) | 54,88(1,53) |

Tabela 4.28: Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e  $w\chi^2$  (Coleção Reuters)

| Algoritmo de Classificação | SoftO-wFDCL | wMI         | $w\chi^2$   |
|----------------------------|-------------|-------------|-------------|
| NB                         | 61,47(3,40) | 98,10(0,89) | 98,10(0,89) |
| M.Naive                    | 97,34(1,25) | 97,72(0,49) | 97,72(0,49) |
| KNN                        | 98,10(0,44) | 98,00(0,70) | 98,00(0,70) |
| SVM                        | 98,67(0,92) | 98,10(1,00) | 98,10(1,00) |
| J48                        | 98,57(0,67) | 97,72(0,49) | 97,72(0,49) |

Tabela 4.29: Avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e  $w\chi^2$  (Coleção WAP)

| Algoritmo de Classificação | SoftO-wFDCL | wMI          | $w\chi^2$    |
|----------------------------|-------------|--------------|--------------|
| NB                         | 24,18(2,33) | 62,73(11,12) | 62,03(10,86) |
| M.Naive                    | 63,18(0,76) | 69,85(2,29)  | 70,30(2,27)  |
| KNN                        | 59,91(3,22) | 81,66(2,28)  | 81,08(2,14)  |
| SVM                        | 63,57(1,71) | 83,45(2,09)  | 83,26(1,68)  |
| J48                        | 58,69(3,51) | 77,04(2,78)  | 77,29(3,09)  |

Assim como observado na avaliação comparativa entre o método SoftO-FDCL e os métodos de seleção de atributos MI e  $\chi^2$ , na avaliação comparativa entre o método SoftO-wFDCL e os métodos de seleção de atributos wMI e  $w\chi^2$  observou-se que os métodos

apresentam resultados semelhantes. Assim, a escolha de qual método utilizar para extrair descritores de grupos depende do problema abordado, uma vez que cada método apresenta um significado diferente: o método SoftO-wFDCL seleciona termos que melhor recuperam os documentos de um determinado grupo; o método MI seleciona termos que apresentam melhor informação mútua com um determinado grupo e o método  $\chi^2$  seleciona termos que apresentam maior dependência estatística com um determinado grupo. Neste doutorado, o diferencial em utilizar qualquer um desses métodos (SoftO-wFDCL, wMI e  $w\chi^2$ ), conforme apresentado na avaliação comparativa, é a organização flexível de documentos, uma vez que a matriz de contingência utilizada pelos mesmos permite que um documento pertença a mais de um grupo e que, portanto, os descritores de grupos identifiquem tópicos da organização flexível proposta. Além disso, nos métodos SoftO-wFDCL, wMI e  $w\chi^2$ , tem-se a informação de pertinência embutida no processo de extração de descritores. O uso dessa informação favorece a escolha de descritores de grupos que melhor identificam tópicos da organização flexível proposta, pela qual documentos podem abordar mais de um tópico com diferentes graus de intensidade.

#### 4.5.6 Avaliação do método HSoftO-FDCL

O método HSoftO-FDCL é uma extensão do método SoftO-FDCL desenvolvida para extração de descritores de grupos fuzzy hierárquicos. Geralmente, a estrutura hierárquica é utilizada para a organização de documentos em um SRI porque a mesma possibilita uma melhor visualização e exploração da coleção organizada. Assim, a organização flexível de documentos também pode beneficiar-se de tal estrutura.

Como exemplo de organização flexível hierárquica, observe a organização dos documentos da coleção Opnosis apresentada na Figura 4.3. Nesta figura, tem-se apenas os grupos dos níveis 1 e 2 da hierarquia obtida por meio do algoritmo de agrupamento Hierarchical Fuzzy C-Means (HFCM) (Pedrycz e Reformat, 2006), apresentado no Capítulo 2.

Nesse exemplo, é possível observar que os documentos podem pertencer a mais de um grupo no mesmo nível da hierarquia com diferentes graus de compatibilidade (em parenteses) e os descritores podem representar mais de um grupo com pesos de representatividade diferentes nos grupos. Por exemplo o termo *ga\_mileage* em destaque na Figura 4.3, foi escolhido como descritor de dois grupos, ambos no nível 2. No primeiro, esse descritor tem 0.581 de representatividade e no segundo ele tem 0.378 de representatividade. Essa ponderação indica que documentos sobre carros podem ser alocados em ambos os grupos. Porém, documentos nos quais o termo *ga\_mileage* é mais frequente são alocados com maior grau de compatibilidade no primeiro grupo.

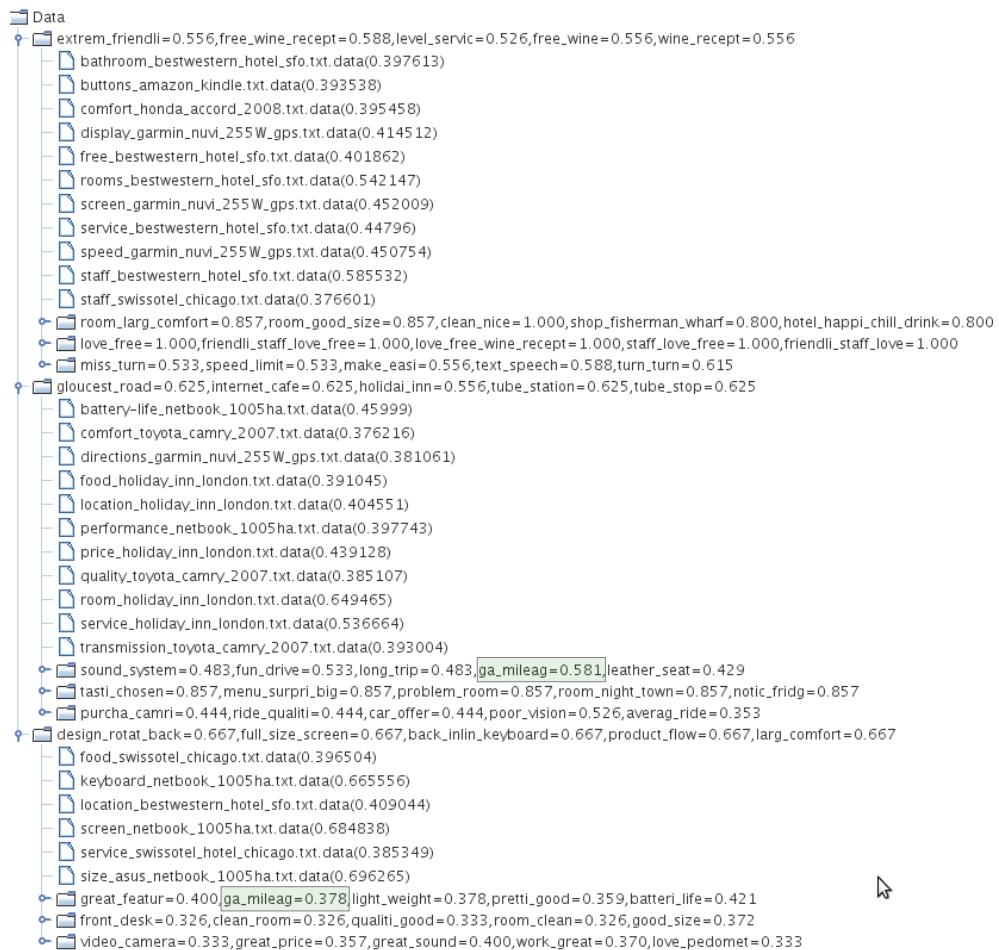


Figura 4.3: Organização flexível hierárquica de documentos da coleção Opinosis

Outra característica interessante da organização hierárquica é que, nessa estrutura, os grupos em um nível mais abaixo da hierarquia são especializações dos grupos em um nível acima. Por exemplo, observe a Figura 4.4. O primeiro grupo, representado pelos descritores *extrem\_friendli*, *free\_wine\_recept*, *level\_servic*, *free\_wine* e *wine\_recept* (indicado por Grupo 1), é composto por documentos sobre hotéis, carros e produtos eletrônicos. Por outro lado, um dos grupos que é extensão desse grupo, o qual é representado pelos descritores *room\_larg\_comfort*, *room\_good\_size*, *clean\_nice*, *shop\_fisherman\_wharf* e *hotel\_happi\_chill\_drink* (indicado por Grupo 1A), é composto somente por documentos sobre hotéis. Além disso, os descritores de grupos originados do Grupo 1 que não identificam um único tópico são estendidos para outros grupos. Por exemplo, o grupo originado do Grupo 1, representado pelos descritores *miss\_turn*, *speed\_limit*, *make\_easi*, *text\_speech* e *turn\_turn* (indicado por Grupo 1C), é estendido em outros grupos de documentos cujos descritores identificam tópicos mais específicos.

Sendo assim, para avaliar a qualidade dos descritores obtidos pelo método HSoftO-FDCL, comparou-se o poder preditivo dos descritores obtidos para os grupos hierárquicos e os descritores obtidos pelo método SoftO-FDCL para os grupos *flat*. Para que fosse possível a execução dos mesmos algoritmos de classificação utilizados para avaliação do

método SoftO-FDCL, foram feitos cortes na hierarquia nos níveis 2, 3, 4 e 5, e a análise comparativa foi realizada sobre esses níveis. Ou seja, os algoritmos de classificação foram executados sobre a matriz composta pelos descritores dos grupos de um determinado nível da hierarquia.

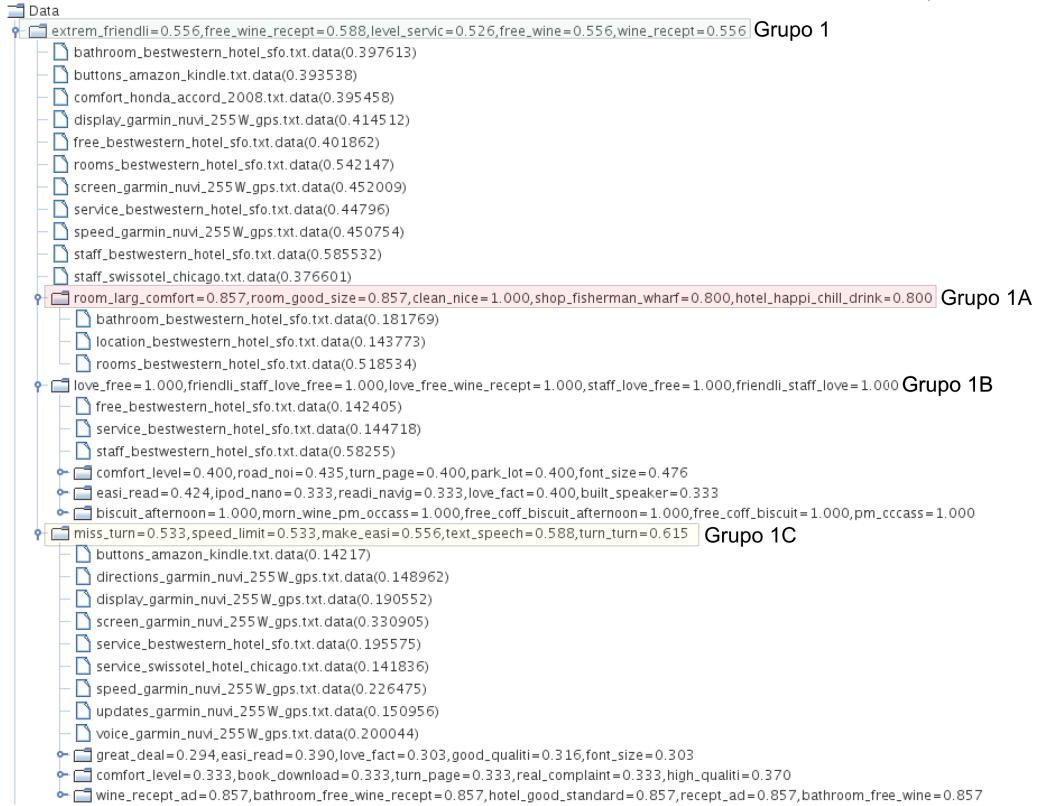


Figura 4.4: Visão Parcial da Organização flexível hierárquica de documentos da coleção Opinosis - observação da especialização/generalização dos tópicos identificados por descritores de grupos

Por exemplo, considere a Figura 4.5. Nesta figura tem-se a distribuição de 5 documentos na hierarquia. Cada documento é disposto no grupo no qual ele possui maior grau de pertinência.

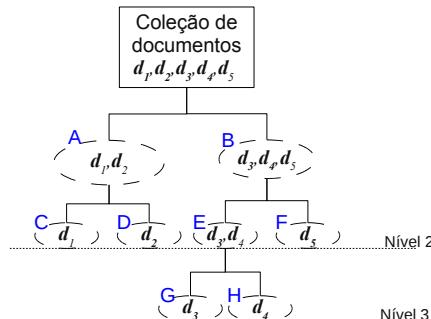


Figura 4.5: Exemplo de hierarquia fuzzy com cinco documentos

Considerando essa distribuição, a matriz obtida, por exemplo, do corte no nível 3

dessa hierarquia é apresentada na Tabela 4.30. Observe que a classe do documento é o grupo no qual ele possui maior grau de pertinência e o nome da classe remete ao caminho percorrido pelo mesmo ao longo da hierarquia. Por exemplo, o documento  $d_4$  no nível 3 pertence à classe *BEG*, pois ele possui maior grau de pertinência no grupo *B* do nível 1 da hierarquia, seguido do grupo *E* do nível 2 e do grupo *G* no nível 3.

Tabela 4.30: Matriz atributo-valor obtida do corte no nível 3 da hierarquia apresentada na Figura 4.5

| Documentos | Descriptor 1 | Descriptor 2 | ... | Classe     |
|------------|--------------|--------------|-----|------------|
| $d_3$      | ...          | ...          | ... | <i>BEG</i> |
| $d_4$      | ...          | ...          | ... | <i>BEH</i> |

As tabelas 4.32, 4.31 e 4.33, apresentam os resultados obtidos pelos métodos SoftO-FDCL e HSoftO-FDCL sobre as coleções Opnosis, Hitech e Reuters. Nesta avaliação foram utilizadas as coleções Opnosis, Hitech e Reuters de forma a observar os resultados obtidos sobre o agrupamento hierárquico de coleções de documentos de diferentes quantidades de classes e documentos: a coleção Opnosis é composta por 51 documentos distribuídos em 3 classes, a coleção Hitech é composta por 600 documentos distribuídos em 6 classes e a coleção Reuters é composta por 1052 documentos distribuídos em 43 classes.

Tabela 4.31: Comparação entre o método SoftO-FDCL e o método HSoftO-FDCL (Coleção Opnosis)

| Algoritmo de Classificação | SoftO-FDCL   | HSoftO-FDCL nível2 | HSoftO-FDCL nível3 | HSoftO-FDCL nível4 | HSoftO-FDCL nível5 |
|----------------------------|--------------|--------------------|--------------------|--------------------|--------------------|
| NB                         | 66,00(18,97) | 31,67(14,34)       | 40,00(25,50)       | 23,50(23,46)       | 3,33(10,54)        |
| M.Naive                    | 80,00(21,08) | 61,00(15,24)       | 60,50(21,92)       | 30,00(20,28)       | 13,33(23,31)       |
| SVM                        | 62,00(22,01) | 60,67(26,75)       | 49,50(24,55)       | 33,50(17,17)       | 19,17(31,93)       |
| KNN                        | 62,00(17,51) | 51,33(19,89)       | 37,50(20,72)       | 35,50(16,57)       | 22,50(15,74)       |
| J48                        | 54,00(28,36) | 39,33(13,50)       | 35,00(15,63)       | 31,00(16,30)       | 15,83(16,87)       |

Tabela 4.32: Comparação entre o método SoftO-FDCL e o método HSoftO-FDCL (Coleção Hitech)

| Algoritmo de Classificação | SoftO-FDCL  | HSoftO-FDCL nível2 | HSoftO-FDCL nível3 | HSoftO-FDCL nível4 | HSoftO-FDCL nível5 |
|----------------------------|-------------|--------------------|--------------------|--------------------|--------------------|
| NB                         | 38,21(5,80) | 23,17(6,16)        | 14,50(3,34)        | 9,67(3,22)         | 6,90(2,79)         |
| M.Naive                    | 41,27(3,43) | 38,50(5,90)        | 27,67(6,77)        | 17,17(5,03)        | 15,83(2,55)        |
| SVM                        | 35,05(5,36) | 52,17(3,77)        | 41,67(3,85)        | 30,50(3,69)        | 25,59(4,37)        |
| KNN                        | 46,14(5,14) | 47,67(5,04)        | 36,50(3,46)        | 24,50(3,93)        | 18,19(2,38)        |
| J48                        | 36,09(6,03) | 41,67(5,15)        | 31,50(4,04)        | 22,00(2,19)        | 14,81(4,93)        |

Tabela 4.33: Comparação entre o método SoftO-FDCL e o método HSoftO-FDCL (Coleção Reuters)

| Algoritmo de Classificação | SoftO-FDCL  | HSoftO-FDCL nível2 | HSoftO-FDCL nível3 | HSoftO-FDCL nível4 | HSoftO-FDCL nível5 |
|----------------------------|-------------|--------------------|--------------------|--------------------|--------------------|
| NB                         | 61,47(3,40) | 30,60(6,07)        | 19,77(3,14)        | 17,21(2,15)        | 18,35(1,73)        |
| M.Naive                    | 97,34(1,25) | 41,15(3,85)        | 23,00(3,05)        | 20,82(4,65)        | 17,40(4,05)        |
| SVM                        | 98,67(0,92) | 58,17(2,35)        | 47,05(3,09)        | 37,27(5,93)        | 32,51(3,63)        |
| KNN                        | 98,10(0,44) | 52,75(5,58)        | 43,91(4,39)        | 33,09(3,90)        | 28,81(3,79)        |
| J48                        | 98,57(0,67) | 62,45(3,63)        | 52,37(3,88)        | 45,16(5,27)        | 41,63(4,15)        |

Nesta avaliação, os descritores obtidos pelo método SoftO-FDCL apresentam, de maneira geral, resultados superiores aos descritores obtidos pelo método HSoftO-FDCL, os quais são destacados em cinza nas tabelas. A justificativa para tal resultado decorre do fato de que os algoritmos de classificação utilizados não consideram a estrutura hierárquica em seu processo e, portanto, a acurácia dos algoritmos é diminuída devido à perda de informação nos cortes da hierarquia, já que a quantidade de documentos em um nível mais abaixo da hierarquia é reduzida.

## 4.6 Considerações finais

Neste capítulo foram apresentados os métodos propostos neste doutorado. Esses métodos possibilitam a organização flexível de documentos pela extração de descritores de grupos após o agrupamento fuzzy de documentos, que permite a organização de documentos assumindo-se que eles podem abordar diferentes tópicos com diferentes graus de intensidade. Os descritores de grupos são importantes porque eles identificam os tópicos abordados pelos documentos.

Os métodos propostos possibilitam a organização flexível de documentos, uma vez que os mesmos extraem descritores de grupos considerando a imprecisão e a incerteza inerentes aos documentos. Embora os métodos propostos sejam independentes do algoritmo de agrupamento fuzzy utilizado, foram propostos três métodos porque considerou-se três perspectivas na organização: i) o método SoftO-FDCL extrai descritores de grupos fuzzy *flat*; ii) o método SoftO-wFDCL, é uma extensão do método SoftO-FDCL e também extrai descritores de grupos fuzzy *flat*, mas acrescenta o grau de pertinência obtido do agrupamento fuzzy, como uma informação adicional para a extração de descritores; e iii) o método HSoftO-wFDCL, o qual também é uma extensão do método SoftO-FDCL para extração de descritores de grupos hierárquicos, mas proporcionam uma visão dos documentos em diferentes níveis de abstração. Sendo assim, a utilização de um determinado método proposto depende da estrutura, *flat* ou hierárquica, que deseja-se obter com a organização flexível de documentos.

Também neste capítulo foram apresentados os resultados obtidos a partir dos experimentos realizados para avaliar cada um dos métodos propostos.

No próximo capítulo, uma aplicação do método SoftO-FDCL para um problema real de organização flexível de documentos será apresentada.



# Aplicação do Método SoftO-FDCL: organização flexível de comentários de médicos de família sobre um processo de avaliação da educação médica continuada canadense

## 5.1 Considerações iniciais

Neste capítulo, uma aplicação do método SoftO-FDCL sobre um problema real é apresentada. Esse problema refere-se à organização de comentários de médicos de família canadenses sobre recomendações de tratamento enviadas para os mesmos por meio de um recurso desenvolvido pela associação de farmacêuticos canadenses (*Canadian Pharmacists Association - CPhA*). Esse recurso é chamado de *e-Therapeutics*<sup>1</sup>, pelo qual é possível o gerenciamento de recomendações de tratamento baseado em evidências farmacológicas e não farmacológicas. Especificamente, *e-Therapeutics* ajudam os médicos de família canadenses (*Canadian Family Physicians - FPs*) a saber quais opções terapêuticas estão disponíveis em uma determinada situação clínica. Além disso, para sensibilizar os FPs de que a informação terapêutica pode ser útil no cuidado de seus pacientes, os editores da CPhA selecionam informações chave, chamadas de *e-Therapeutics Highlights*<sup>2</sup>. *Highlights* são enviados semanalmente por e-mail para os FPs. Um exemplo de email recebido pelos médicos de família canadenses pode ser observado na Figura 5.1.

Em parceria com a CPhA e a faculdade de médicos de família do Canadá (*College of Family Physicians of Canada - CFPC*), o grupo *Information Technology Primary Care*

<sup>1</sup><http://www.pharmacists.ca/index.cfm/function/store/PublicationDetail.cfm?pPub=9>

<sup>2</sup><http://www.pharmacists.ca/index.cfm/more-information/et-mcgill/>



Figura 5.1: Exemplo de email enviado para um médico de família canadense sobre um *e-Therapeutics+ Highlight*

*Research Group* (ITPCRG) da Universidade de McGill implementou o programa de educação médica continuada (*Continuing Medical Education - CME*) o qual faz uma conexão entre o método *Information Assessment Method* (IAM) e os *Highlights* (Pluye et al., 2009, 2010a,b). O método IAM avalia o valor da informação contida nos *Highlights* em quatro situações: relevância, impacto cognitivo, uso e benefícios esperados à saúde. Quando os FPs recebem um email informando sobre um novo *e-Therapeutics+ Highlights*, eles podem ler a informação contida no *e-Therapeutics+ Highlights* e avaliar um *Highlight*, o qual é um trecho importante dos *e-Therapeutics+ Highlights* destacados em verde no texto do email. Essa avaliação é realizada por meio do questionário IAM aberto ao pressionar o botão “Useful Info?”. Esse questionário objetiva avaliar a informação clínica contida nos *Highlights*. Um exemplo de *e-Therapeutics+ Highlights* pode ser observado na Figura 5.2. O questionário IAM pode ser observado na Figura 5.3.

Dados do IAM permitem aos pesquisadores do grupo ITPCRG coletar *feedbacks* dos FPs e então avaliar o conteúdo dos *Highlights* sob a perspectiva dos FPs. Uma grande quantidade de *feedbacks* dos FPs são comentários construtivos (*Constructive Feedback Comments - CFBs*), os quais possibilitam à CPhA melhorar o conteúdo dos *e-Therapeutics+*.

Atualmente, CFBs são identificados manualmente (Pluye et al., 2012). No entanto, a coleção de comentários cresce rapidamente, a qual pode atingir uma quantidade de comentários que aumenta o esforço de editores e pesquisadores para analisá-los. A extração automática de conhecimento a partir destes comentários é uma importante tarefa para

**Pharmacologic Choices**

**Acetaminophen** and **Ibuprofen** are the only therapeutic choices available for managing fever in children ([Table 1](#)). They have been well studied in large populations and are considered safe in therapeutic doses.<sup>4</sup>, [15](#) In some systematic reviews, ibuprofen appears to have a shorter time to reduction of fever and longer duration of defervescence.<sup>4</sup>, [16](#) The clinical significance of these marginal benefits is unclear. Both medications have analgesic properties.

**ASA** is not recommended in children under 15 years of age because of the possible enhanced risk of Reye's syndrome.<sup>17</sup>

**Naproxen sodium** has not been studied in children for the treatment of fever and therefore is not recommended in children under 12 years of age.

A response to antipyretic therapy does not exclude the possibility that serious underlying illness is present. Clinical decision-making should not be based on response to fever treatment.<sup>18</sup>

**It is a common practice to alternate acetaminophen and ibuprofen to "normalize" temperature.<sup>19,20</sup> However, there is insufficient evidence to support the routine practice of alternating or combining acetaminophen and ibuprofen.<sup>21,22</sup> While alternating or combining acetaminophen and ibuprofen may result in a greater period of time without fever, the clinical significance of this difference is uncertain. In addition, it has not been shown to be either safe or more effective in improving discomfort than a single antipyretic. There may be an increased risk of parental confusion and subsequent dosing errors when this strategy is employed.** [\[Useful Info?\]](#)

**Therapeutic Tips** [Back to Top](#)

- Use doses of acetaminophen and ibuprofen based on the child's weight, not on age. Maximum dose per day should be specified.
- Concentrations of liquid acetaminophen and ibuprofen preparations vary according to product. Remind caregivers to check the concentration of a product each time medication is given.
- Acetaminophen is the most common cause of analgesic overdose in children under 6 years old. Store antipyretics in locked cabinets to prevent inappropriate access. Instruct parents to use a calibrated measuring device, and educate them about the many formulations available and the potential for error with substitution.

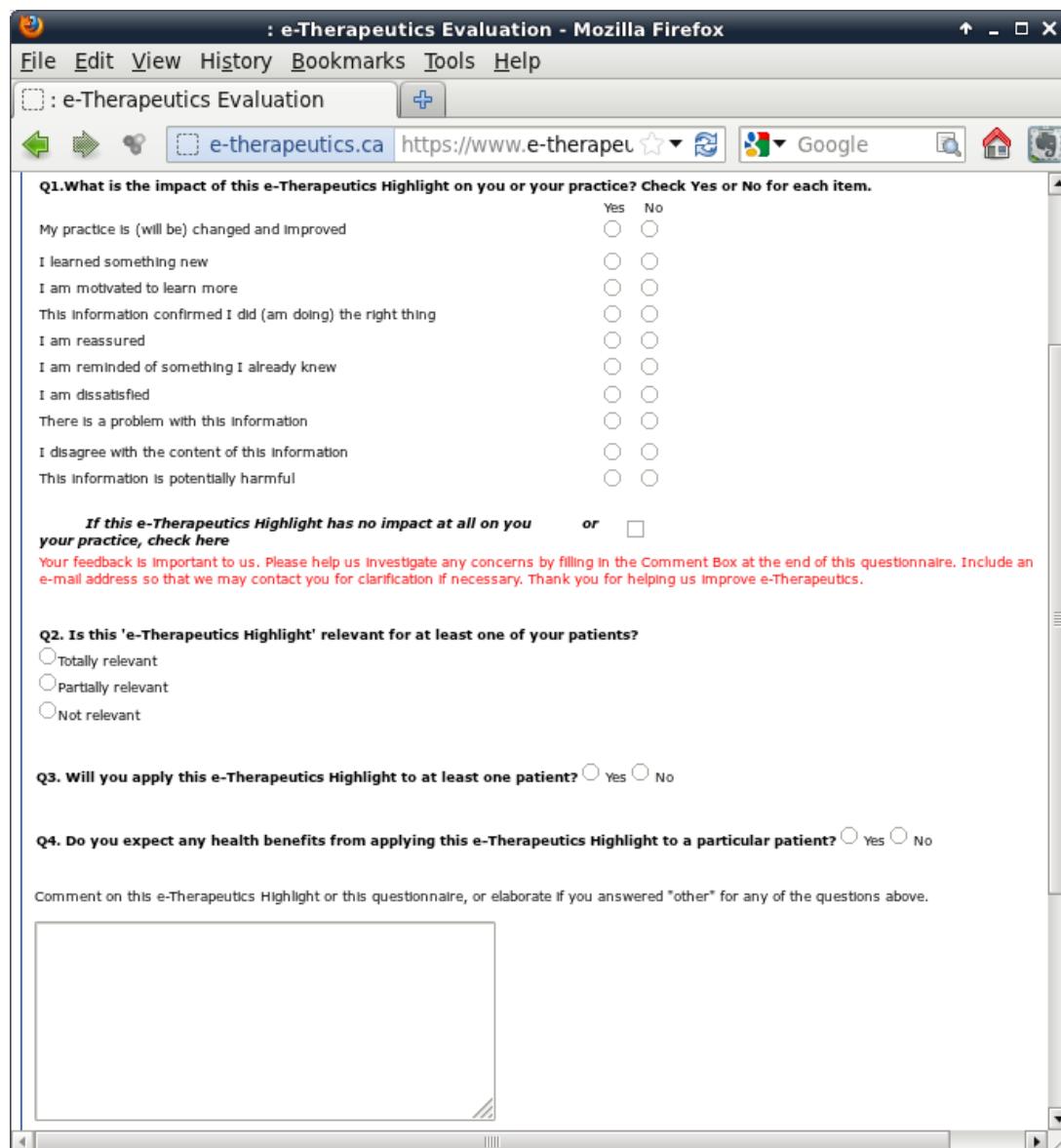
Figura 5.2: Exemplo de um *e-Therapeutics + Highlight*. O trecho destacado em verde corresponde a um *Highlight*

encontrar conhecimento útil que ajude no processo de tomada de decisão dos editores da CPhA.

Portanto, o principal objetivo da aplicação do método SoftO-FDCL proposto neste doutorado é auxiliar a seleção de CFBs e, consequentemente, otimizar o gerenciamento dos *Highlights*.

Utilizando o método SoftO-FDCL para a extração de descritores de grupos de comentários enviados pelos médicos de família, a organização flexível da coleção de comentários é obtida. Essa organização é considerada apropriada para o problema apresentado porque comentários construtivos (CFBs) e não-construtivos (non-CFBs) apresentam características similares, já que um mesmo termo pode ser utilizado em ambos os tipos de comentários. Ou seja, ao organizar os comentários em dois grupos, um mesmo comentário pode estar alocado no grupo cujos descritores identificam CFBs e no grupo cujos descritores identificam non-CFBs. A partir dos graus de pertinência dos comentários nos grupos, os pesquisadores podem decidir se um determinado comentário é considerado construtivo ou não.

Sendo assim, a organização flexível de comentários não elimina a participação dos pesquisadores na seleção de comentários, mas reduz o esforço realizado pelos mesmos, já que a organização flexível dos comentários dá um indício do tipo de cada comentário.



The screenshot shows a Mozilla Firefox browser window with the title bar "e-Therapeutics Evaluation - Mozilla Firefox". The address bar displays "e-Therapeutics Evaluation" and the URL "https://www.e-therapeutics.ca". The main content area contains the following questionnaire:

**Q1. What is the impact of this e-Therapeutics Highlight on you or your practice? Check Yes or No for each item.**

|   | Yes                   | No                    |
|---|-----------------------|-----------------------|
| My practice is (will be) changed and improved               | <input type="radio"/> | <input type="radio"/> |
| I learned something new                                     | <input type="radio"/> | <input type="radio"/> |
| I am motivated to learn more                                | <input type="radio"/> | <input type="radio"/> |
| This information confirmed I did (am doing) the right thing | <input type="radio"/> | <input type="radio"/> |
| I am reassured  | <input type="radio"/> | <input type="radio"/> |
| I am reminded of something I already knew                   | <input type="radio"/> | <input type="radio"/> |
| I am dissatisfied   | <input type="radio"/> | <input type="radio"/> |
| There is a problem with this information                    | <input type="radio"/> | <input type="radio"/> |
| I disagree with the content of this information             | <input type="radio"/> | <input type="radio"/> |
| This information is potentially harmful                     | <input type="radio"/> | <input type="radio"/> |

**If this e-Therapeutics Highlight has no impact at all on you or your practice, check here**

Your feedback is important to us. Please help us investigate any concerns by filling in the Comment Box at the end of this questionnaire. Include an e-mail address so that we may contact you for clarification if necessary. Thank you for helping us improve e-Therapeutics.

**Q2. Is this 'e-Therapeutics Highlight' relevant for at least one of your patients?**

- Totally relevant
- Partially relevant
- Not relevant

**Q3. Will you apply this e-Therapeutics Highlight to at least one patient?**  Yes  No

**Q4. Do you expect any health benefits from applying this e-Therapeutics Highlight to a particular patient?**  Yes  No

Comment on this e-Therapeutics Highlight or this questionnaire, or elaborate if you answered "other" for any of the questions above.

(A large empty text area is provided for comments.)

Figura 5.3: Questionário IAM

## 5.2 Seleção manual de comentários construtivos

Por meio do questionário IAM, médicos de família canadenses (*Canadian Family Physicians* - FPs) submetem milhares de comentários de *feedback* relacionados aos *e-Therapeutics+ Highlights* recebidos por e-mail. Por exemplo, de 20 de Janeiro de 2010 a 19 de Janeiro de 2011, 51 *Highlights* foram enviados para cerca de 17000 FPs, dos quais 5346 submeteram 31429 questionários IAM (avaliações dos *Highlights*). Desses questionários, 4166 (13.3%) contém comentários, gerando 682 (2.2%) CFBs.

FPs podem ler, avaliar o *Highlight* e escrever um comentário com sua opinião sobre o assunto do *Highlight* utilizando uma caixa de texto no questionário IAM. Os comentários são, então, selecionados como CFB ou non-CFB. Os pesquisadores do grupo ITPCRG selecionam como CFBs os comentários que referem-se a um “comentário textual que requer atenção para disparar uma investigação futura” (Pluye et al., 2012). Dois comentários sobre o *Highlight* - Exemplo apresentado a seguir, um construtivo (CFB) e um não construtivo (non-CFB), são apresentados na Tabela 5.1.

### ***Highlight - Exemplo:***

*The dual-action serotonin and noradrenaline reuptake inhibitor (SNRI) duloxetine is beneficial in improving pain, stiffness, fatigue and overall quality of life in fibromyalgia patients; these effects appear to be independent of the drug's effect on depression. Duloxetine is generally well tolerated. There is little evidence to support the efficacy of venlafaxine in patients with fibromyalgia.*

Tabela 5.1: CFB e non-CFB para o *Highlight*- Exemplo

| Comentário  | Tipo de comentário |
|---|--------------------|
| <i>In my own practice I've seen little/no benefit from pregabalin or duloxetine. I wonder about the quality of the studies supporting their use.</i>  | CFB                |
| <i>Good review of disease process, current nonpharm and pharm treatment options. SNRIs are very well tolerated, evidence that Duloxetine is useful apart from antidepressant effects for treatment of fibromyalgia excellent knowledge - will modify my practice.</i> | non-CFB            |

O comentário CFB apresentado na Tabela 5.1 demonstra a insatisfação de um FP, o qual questiona os editores do *e-Therapeutics+ Highlights* para verificar os estudos referenciados no *Highlight* - Exemplo. Por outro lado, o comentário non-CFB é uma repetição e elogio à informação contida no *Highlight* - Exemplo.

De maneira geral, a rotulação dos comentários em CFB e non-CFB é uma tarefa difícil e custosa por duas razões: (a) comentários são compostos de sentenças com alta subjetividade, imprecisão e incerteza; e (b) diferentes tipos de comentários sobre um mesmo *Highlight* podem ser compostos por sentenças similares (Hripcsak et al., 2007).

A fim de rotular manualmente os comentários e selecionar aqueles que podem prover melhorias para um *Highlight*, pesquisadores do grupo ITPCRG desenvolveram 7 regras para selecionar manualmente CFBs, pelas quais um comentário é considerado CFB quando ele é:

1. Um comentário que corresponde em significado a uma avaliação IAM de “Desacordo”, “Dano potencial”, “Insatisfação” ou “Problema com a informação”;
2. Um comentário sobre uma informação ausente ou que indique a necessidade de mais informação;
3. Um comentário que inclui nuances ou reservas (por exemplo, “Eu concordo, mas...”);
4. Um comentário que revela que o leitor não ficou ‘convencido’ pela informação;
5. Um comentário dizendo que a informação relevante não foi encontrada;
6. Um comentário afirmando que o *Highlight* é ‘conhecimento antigo’ ou equivalente;
7. Um comentário negativo no questionário IAM ou no processo de avaliação.

Embora as regras sejam simples e úteis, algumas inconsistências podem ocorrer. Por exemplo, o comentário “*concise summary on the state of evidence for different dementia treatments*” foi selecionado manualmente como um CFB. No entanto, esse comentário não condiz com nenhuma regra apresentada anteriormente, o que indica uma limitação do processo manual de seleção de CFBs. Visando solucionar este tipo problema, a organização automática de comentários é proposta. Tal organização foi realizada por meio do método SoftO-FDCL, no qual as palavras mais frequentes nos comentários ajudam a organizar os comentários em construtivos e não-construtivos com algum grau de compatibilidade.

## 5.3 Identificação automática de comentários

Considerando que comentários construtivos e não-construtivos são documentos que podem apresentar características similares, já que um mesmo termo pode ser utilizado em ambos os tipos de comentários, a organização flexível é considerada apropriada para este caso. Sendo assim, o método SoftO-FDCL proposto neste doutorado foi utilizado para a extração de descritores de grupos de comentários de maneira a identificar comentários construtivos e não-construtivos.

Para que seja possível a utilização do método SoftO-FDCL, os comentários, considerados documentos, são pré-processados, conforme apresentado no Capítulo 2 Seção 2.2,

para poderem ser agrupados por meio de um algoritmo de agrupamento fuzzy. Nesta aplicação, foi utilizado o algoritmo de agrupamento Fuzzy C-Means, também apresentado no Capítulo 2.

O método SoftO-FDCL foi aplicado sobre duas coleções de documentos reais: (1) comentários enviados por FPs no período entre 01 de Janeiro de 2011 e 31 de Dezembro de 2011; e (2) comentários enviados por FPs no período entre 03 de Janeiro de 2012 e 14 de Fevereiro de 2012. Todos os comentários que compõem essas duas coleções foram manualmente rotuladas em CFBs ou non-CFBs por pesquisadores do grupo ITPCRG utilizando as regras apresentadas anteriormente. Essa rotulação manual foi utilizada para avaliar o resultado obtido pela aplicação do método SoftO-FDCL. Considerando este conhecimento prévio, cada uma das coleções deu origem a três outras coleções: uma coleção composta por todos os comentários, uma coleção composta por somente CFBs e uma coleção composta por somente non-CFBs. As características dessas coleções são apresentadas na Tabela 5.2.

Tabela 5.2: Coleções utilizadas na aplicação do método SoftO-FDCL para organização flexível dos comentários de médicos de família canadenses. As coleções são identificadas pela coluna “ID” e a quantidade de comentários que compõe cada coleção é identificada pela coluna “# comentários”. A porcentagem de CFBs e non-CFBs obtida a partir das coleções de 2011 e 2012 são também apresentadas na coluna “# comentários”

| Comentários recebidos  | Coleção              | ID               | # comentários |
|--|----------------------|------------------|---------------|
| de 01 de Janeiro de 2011 à 31 de Dezembro de 2011 (12 meses) | Todos os comentários | <i>all-2011</i>  | 4998          |
|  | CFBs                 | <i>cfb-2011</i>  | 1183 (23,67%) |
|  | non-CFBs             | <i>ncfb-2011</i> | 3815 (76,33%) |
| de 03 de Janeiro de 2012 à 14 de Fevereiro de 2012 (1 mês)   | Todos os comentários | <i>all-2012</i>  | 656           |
|  | CFBs                 | <i>cfb-2012</i>  | 117 (17,84%)  |
|  | non-CFBs             | <i>ncfb-2012</i> | 539 (82,16%)  |

O objetivo principal da aplicação do método SoftO-FDCL sobre essas coleções é organizar de maneira flexível os documentos da coleção *all-2011* por meio de grupos cujos descritores identifiquem dois tipos de comentários: CFBs e non-CFBs.

Os resultados obtidos com a aplicação do método SoftO-FDCL sobre as coleções apresentadas na Tabela 5.2, bem como a discussão sobre os mesmos, são apresentados a seguir.

## 5.4 Resultados obtidos

Para comparar a organização flexível obtida pelo método SoftO-FDCL com a categorização manual de comentários, além da coleção *all-2011*, as coleções *cfb-2011* e *ncfb-2011* também tiveram seus documentos agrupados. Embora essas duas últimas coleções sejam

compostas por comentários de apenas um tipo (CFB ou non-CFB), elas também foram agrupadas em uma quantidade mínima de grupos porque o método SoftO-FDCL requer que documentos sejam organizados em grupos para que, posteriormente, seja possível a extração de descritores.

Uma organização flexível foi obtida para cada uma das coleções *all-2011*, *cfb-2011* e *ncfb-2011* para observar se os descritores de grupos obtidos a partir da coleção composta por todos os comentários, coleção *all-2011*, compartilham alguma similaridade com os descritores de grupos obtidos a partir das coleções que foram manualmente rotuladas como CFB (coleção *cfb-2011*) ou non-CFB (coleção *ncfb-2011*). Consequentemente, foi observado também se os descritores identificam adequadamente os tipos de comentários de cada grupo da coleção *all-2011*.

Nas Tabelas 5.3, 5.4 e 5.5, os descritores de dois grupos encontrados para cada coleção são apresentados. Para uma melhor observação manual dos descritores extraídos pelo método SoftO-FDCL, os 10 melhores descritores foram selecionados. Essa quantidade foi escolhida arbitrariamente.

Tabela 5.3: Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção *all-2011*

| Descritores do grupo 1  | Descritores do grupo 2  |
|---|---|
| <i>patient, inform, practic, help, excel, review, interest, relev, treatment, topic</i> | <i>good, info, remind, sum- mari, manag, common, problem, articl, nice, great</i> |

Tabela 5.4: Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção *cfb-2011*

| Descritores do grupo 1   | Descritores do grupo 2  |
|--|---|
| <i>inform, effect, good, dose, treatment, interest, benefit, risk, medic, parent</i> | <i>patient, recommend, practic, help, drug, treat, review, evid, don, studi</i> |

Tabela 5.5: Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção *ncfb-2011*

| Descritores do grupo 1  | Descritores do grupo 2   |
|---|--|
| <i>patient, inform, practic, help, excel, review, relev, interest, treatment, topic</i> | <i>good, common, remind, articl, overview, nice, sum- mari, manag, problem, info</i> |

Os resultados indicam que é mais fácil identificar non-CFBs do que CFBs. Para tanto, observou-se se os descritores dos grupos 1 e 2 da coleção *all-2011*, apresentados na Tabela 5.3, são iguais aos descritores de ambos os grupos das coleções *cfb-2011* e *ncfb-2011*, apresentados nas Tabelas 5.4 e 5.5, respectivamente.

A maioria dos descritores (*patient, inform, practic, help, review, interest e treatment*) do grupo 1 da coleção *all-2011*, apresentado na Tabela 5.3, aparecem entre os descritores de um dos grupos de ambas as coleções: CFBs (*cfb-2011*) e non-CFBs (*ncfb-2011*). Assim, não se encontram evidências de que o grupo 1 da coleção *all-2011* representa completamente uma das duas coleções.

Por outro lado, exceto o descritor *great*, todos os descritores do grupo 2 da coleção *all-2011* foram também extraídos como descritores do grupo 2 da coleção *ncfb-2011*. Além disso, é possível observar que, exceto os descritores *manag* e *great*, os descritores do grupo 2 da coleção *all-2011* são muito mais frequentes nos comentários que foram manualmente rotulados como non-CFBs (coleção *ncfb-2011*) do que nos comentários da coleção *cfb-2011*. Essa informação é apresentada na Figura 5.4. Esse resultado sugere que o grupo 2 da coleção *all-2011* contém, em sua maioria, comentários não construtivos.

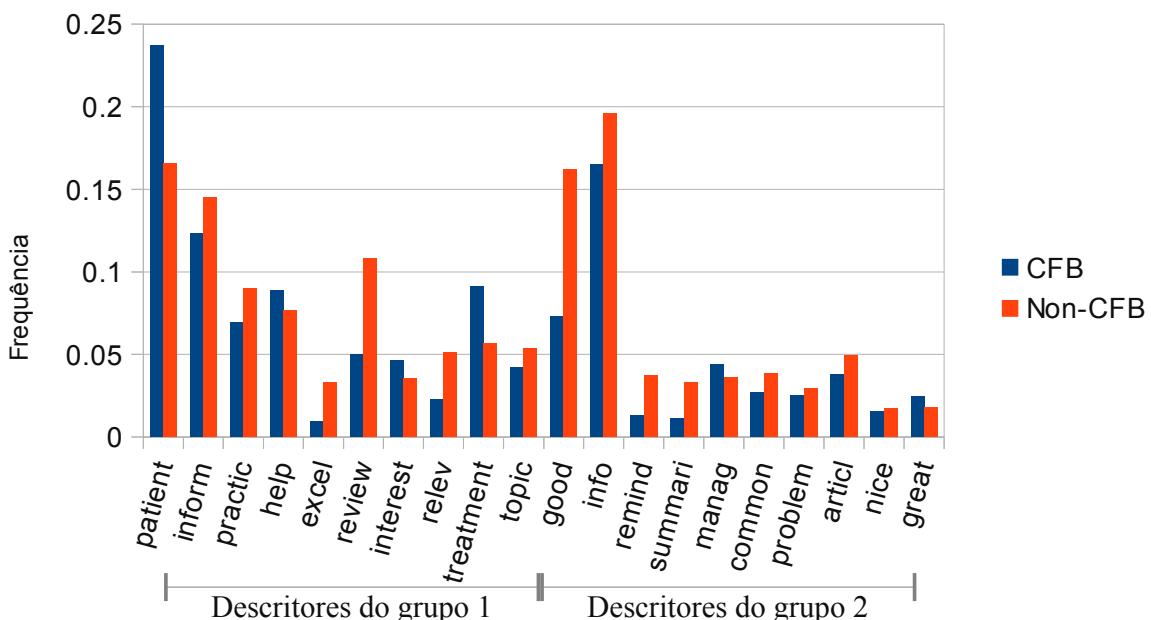


Figura 5.4: Frequência dos descritores obtidos da aplicação do método SoftO-FDCL sobre a coleção *all-2011* nas coleções *cfb-2011* e *ncfb-2011*

Para observar como um determinado descritor pode ocorrer tanto em comentários construtivos, quanto não construtivos, observe a ocorrência do descritor “*good*” em um CFB e em um non-CFB, ambos apresentados na Tabela 5.6.

Tabela 5.6: Exemplo de comentários em que o descritor “*good*” ocorre

| non-CFB   | CFB  |
|---|--|
| <b><i>Good</i></b> to have a reputable source confirm what pharmaceutical companies are claiming about their product. | <i>I enjoy these highlights, concise, easy to read and useful, keep doing this <b>good</b> work.</i> |

Os mesmos experimentos realizados sobre os comentários enviados por FPs em 2011 foram realizados sobre os comentários recebidos em 2012. O principal objetivo dessa observação é encontrar um padrão de escrita de comentários que permanece ao longo do tempo e que pode auxiliar no processo de organização dos comentários independente do período no qual os comentários são escritos pelos médicos de família. Nas Tabelas 5.7, 5.8 e 5.9, os descritores de cada grupo são apresentados.

Tabela 5.7: Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção *all-2012*

| Descritores do grupo 1   | Descritores do grupo 2  |
|--|---|
| <i>good, patient, inform, practic, medic, drug, interest, articl, info, help</i> | <i>review, treatment, learn, excel, treat, manag, highlight, good_review, topic, adhd</i> |

Tabela 5.8: Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção *cfb-2012*

| Descritores do grupo 1  | Descritores do grupo 2   |
|---|--|
| <i>treatment, read, drug, patient, find, med, inform, potentii, articl, practic</i> | <i>medic, start, cost, effect, review, children, chang, specialist, pt</i> |

Tabela 5.9: Descritores de grupos obtidos da aplicação do método SoftO-FDCL sobre a coleção *ncfb-2012*

| Descritores do grupo 1  | Descritores do grupo 2  |
|---|---|
| <i>good, review, inform, patient, practic, interest, excel, info, articl, treat</i> | <i>learn, treatment, help, manag, highlight, medic, drug, topic, adhd, ib</i> |

O grupo 2 da coleção *all-2012*, apresentada na Tabela 5.7, contém, em sua maioria, descritores que ocorrem somente nos grupos da coleção *ncfb-2012*: *review, learn, excel, treat, manag, highlight* e *adhd*. Esses descritores, exceto o descritor *treat*, possuem mais alta frequência na coleção *ncfb-2012* do que na coleção *cfb-2012*, como pode ser observado na Figura 5.5.

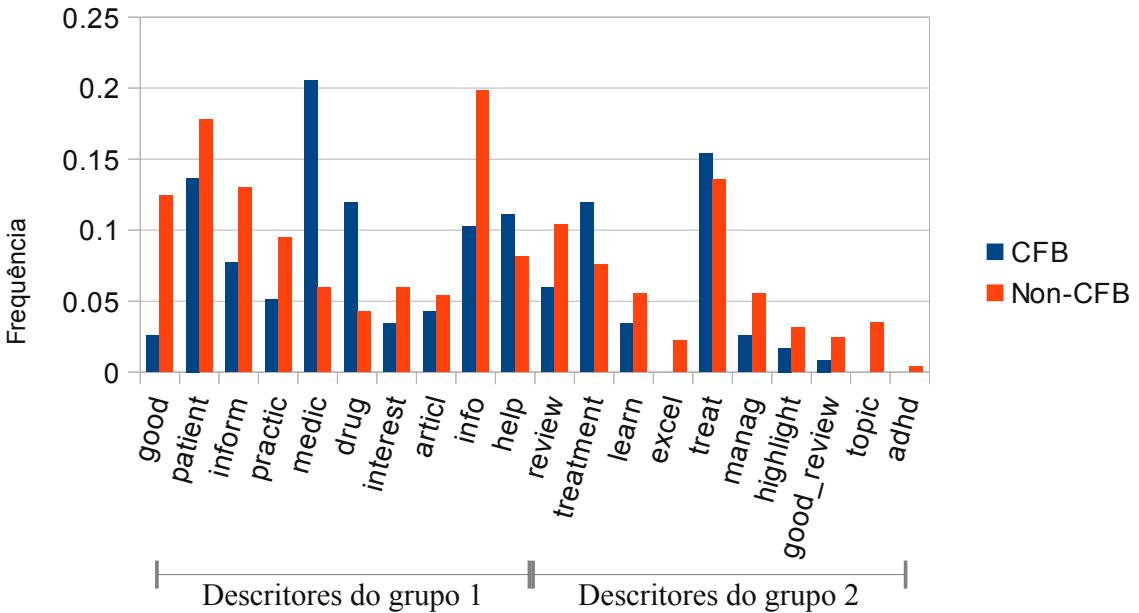


Figura 5.5: Frequência dos descritores obtidos da aplicação do método SoftO-FDCL sobre a coleção *all-2012* nas coleções *cfb-2012* e *ncfb-2012*

Observando os resultados obtidos da aplicação do método SoftO-FDCL sobre comentários enviados em 2011 e 2012, embora alguns descritores extraídos das coleções *all-2011* e *all-2012* sejam diferentes, eles apresentam características similares: é mais fácil identificar non-CFBs do que CFBs.

Sendo assim, os descritores de grupos encontrados nos experimentos apresentam um indicativo acerca do grupo no qual os comentários são predominantemente non-CFBs. Essa informação pode auxiliar os pesquisadores do grupo ITPCGR na tarefa de separar comentários construtivos dos não construtivos. A fim de tornar mais clara esta observação, considere dois grupos,  $g_1$  e  $g_2$ . Suponha que os descritores obtidos a partir do grupo  $g_1$  indicam que este grupo é composto em sua maioria por non-CFBs. Por outro lado, os descritores do grupo  $g_2$  não garantem que seus comentários são construtivos ou não. Considere também três comentários,  $\{d_1, d_2, d_3\}$ , e seus graus de pertinência nos grupos obtidos do agrupamento fuzzy  $\{\mu(d_1, g_1), \mu(d_1, g_2), \mu(d_2, g_1), \mu(d_2, g_2), \mu(d_3, g_1), \mu(d_3, g_2)\}$ , conforme apresentado na Tabela ??.

Tabela 5.10: Graus de pertinência de três comentários em dois grupos

| Comentários | Grupos |       |
|-------------|--------|-------|
|             | $g_1$  | $g_2$ |
| $d_1$       | 0.9    | 0.1   |
| $d_2$       | 0.1    | 0.9   |
| $d_3$       | 0.5    | 0.5   |

Analizando os graus de pertinência, é possível descartar o comentário  $d_1$  porque ele

é considerado um non-CFB, uma vez que ele possui maior grau de pertinência no grupo  $g_1$ ,  $\mu(\mathbf{d}_1, g_1) = 0.9$ . No entanto, o comentário  $\mathbf{d}_2$  deve ser manualmente analisado para verificar se ele é um CFB ou non-CFB porque seu grau de pertinência é maior no grupo  $g_2$ ,  $\mu(\mathbf{d}_2, g_2) = 0.9$ . Já o comentário  $\mathbf{d}_3$  também dever ser manualmente analisado porque seu grau de pertinência é igualmente distribuído em ambos os grupos,  $\mu(\mathbf{d}_3, g_1) = \mu(\mathbf{d}_3, g_2) = 0.5$ . Assim, percebeu-se que a organização flexível não elimina a intervenção humana, mas a apóia, uma vez que a quantidade de comentários a ser analisada é reduzida pela eliminação de comentários que apresentam alto grau de pertinência nos grupos cujos descritores identificam non-CFBs.

## 5.5 Considerações finais

Neste capítulo foi apresentada a aplicação do método SoftO-FDCL sobre um problema real. Tal aplicação resultou na organização flexível de comentários de médicos de família canadenses sobre recomendações de tratamento recebidos por e-mail pelos mesmos. Essa organização auxilia os pesquisadores do grupo ITPCRG na seleção de CFBs pela redução da quantidade de comentários a serem manualmente analisados.

Essa aplicação foi desenvolvida durante estágio de doutorado realizado no exterior e teve participação direta dos pesquisadores envolvidos com o projeto de *e-Therapeutics+ Highlights*. Como trabalho futuro, é importante desenvolver um software que automatize a organização de comentários proposta, por meio de uma interface amigável ao usuário, de forma que os comentários sejam avaliados em CFB ou non-CFB a medida que os mesmos forem sendo enviados pelos FPs.

Além disso, o algoritmo de agrupamento utilizado nos experimentos é um algoritmo não-supervisionado, o qual busca por uma estrutura de grupos em dados não rotulados. Ou seja, o algoritmo de agrupamento utilizado não considera nenhuma informação prévia em seu processo de agrupamento. Portanto, a informação acerca da rotulação manual dos comentários foi utilizada somente na avaliação dos descritores obtidos após o processo de agrupamento. Como pesquisa futura, estuda-se a possibilidade de inserção deste conhecimento prévio no processo de agrupamento pela utilização de um algoritmo de agrupamento semi-supervisionado. Em geral, neste tipo de agrupamento, uma pequena quantidade de documentos da coleção a ser organizada é rotulada de forma a calibrar o algoritmo de agrupamento a ser executado sobre os demais documentos.

Conforme apresentado, o método SoftO-FDCL foi aplicado para a organização flexível de comentários de médicos de família canadenses, obtendo de forma simplificada uma organização flexível adequada para o problema abordado. No entanto, é possível, em um trabalho futuro, aplicar também o método wSoftO-FDCL, esperando-se obter descritores que melhor identifiquem os comentários construtivos e não construtivos. Por outro lado, a aplicação do método HSoftO-FDCL para o problema em questão precisa ser melhor investigada, pois a organização flexível de comentários de médicos de família canadenses é

suficientemente boa utilizando apenas dois grupos *flat*: um para os comentários construtivos e outro para os não construtivos. Assim, é preciso investigar em um trabalho futuro a necessidade de generalização/especialização dos grupos de comentários que obtém-se da organização flexível hierárquica com a aplicação do método HSoftO-FDCL.



## CAPÍTULO

### 6

# Conclusões

Métodos de agrupamento de documentos têm sido bastante utilizados para obter conhecimento útil sobre documentos (Manning et al., 2008; Baeza-Yates e Ribeiro-Neto, 2011). Esse conhecimento é obtido à medida que documentos que abordam assuntos semelhantes são alocados em um mesmo grupo. No entanto, existem situações em que a escolha de um único grupo para um dado documento não é a mais apropriada, uma vez que esse documento pode abordar diversos assuntos, possuindo relacionamentos com diversos grupos simultaneamente. Sendo assim, conforme apresentado ao longo desta tese, por meio do agrupamento fuzzy de documentos é possível obter uma organização flexível de documentos, cujo diferencial é a possibilidade de um documento abordar diferentes assuntos com diferentes graus de intensidade, caracterizando a imprecisão e incerteza típicas de situações reais. Visto que documentos são melhor interpretados quando organizados em grupos cujos descritores identificam tópicos da coleção de documentos, a organização flexível de documentos proposta neste doutorado é alcançada pela extração de descritores de grupos fuzzy de documentos.

Este cenário motivou a verificação da hipótese desta tese de doutorado:

*A extração de descritores de grupos fuzzy de documentos possibilita a organização flexível de documentos, a qual permite que usuários de sistemas de recuperação de informação accessem o conteúdo dos documentos organizados considerando a imprecisão e incerteza típicas de situações reais.*

A partir dos estudos e experimentos realizados e da hipótese estabelecida, definiu-se como objetivo desta tese:

*Investigar e desenvolver métodos para a extração de descritores de grupos fuzzy que permitam a organização flexível de documentos.*

Para atender o objetivo estabelecido, diversas atividades de pesquisa foram realizadas, dentre as quais se destacam a proposta e desenvolvimento de três métodos de extração de descritores de grupos fuzzy: SoftO-FDCL, Soft-wFDCL e HSoftO-FDCL. Esses métodos contribuem para o estado da arte, extraindo descritores de grupos fuzzy separadamente do processo de agrupamento.

Neste capítulo, um resumo das contribuições deste doutorado, as parcerias com grupos de pesquisa de outras instituições, as limitações encontradas e os trabalhos futuros são apresentados.

## 6.1 Resumo das contribuições

As contribuições ao estado da arte obtidas neste doutorado estão diretamente relacionados com a proposta de uma organização flexível de documentos. De maneira geral, essas contribuições consistem de estudo, proposta, desenvolvimento e avaliação de métodos para extração de descritores de grupos fuzzy. Além disso, uma das contribuições consiste da aplicação de um dos métodos propostos em uma aplicação real no contexto da educação médica continuada canadense.

A primeira contribuição consiste da proposta e desenvolvimento do método SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*). Por meio desse método, descritores de grupos fuzzy *flat* são extraídos após o processo de agrupamento, visando identificar tópicos da organização flexível de documentos. Os experimentos realizados mostraram que a flexibilidade da organização de documentos é alcançada a partir da utilização dos graus de pertinência obtidos do agrupamento fuzzy, indicando a compatibilidade entre documentos e grupos. Além disso, por meio desse método, a avaliação dos termos candidatos a descritores mede quão representativo um descritor é para um grupo e os descritores identificam tópicos da organização flexível (Nogueira et al., 2011a), (Nogueira et al., 2012a), (Nogueira et al., 2012b), (Nogueira et al., 2013).

A segunda contribuição consiste da proposta e desenvolvimento do método SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*), pelo qual descritores de grupos fuzzy *flat* também são extraídos após o processo de agrupamento, porém o grau de pertinência dos documentos em cada grupo, obtidos do agrupamento fuzzy, é diretamente utilizado na avaliação dos termos candidatos a descritores. Essa nova forma de avaliação considera que os graus de pertinência carregam uma informação adicional acerca da representatividade dos termos, a qual pode contribuir para uma avaliação mais precisa da importância de um termo candidato a descritor de grupo. Nos experimentos realizados, o método SoftO-wFDCL apresentou resultados iguais ou superiores ao método SoftO-FDCL, ressaltando a importância dos mesmos para a organização flexível de documentos.

A terceira contribuição consiste da proposta e desenvolvimento do método HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*). Por meio desse

método, descritores de grupos fuzzy hierárquicos são extraídos após o processo de agrupamento, identificando tópicos da organização hierárquica flexível de documentos. Os experimentos realizados mostraram que a organização hierárquica flexível de documentos permite que a coleção de documentos seja visualizada e explorada iterativamente, já que por meio desta organização dois tópicos podem ser a especialização ou generalização um do outro. Além disso o agrupamento fuzzy hierárquico e a extração de descritores pelo método HSoftO-FDCL garantem que os documentos pertençam a mais de um grupo no mesmo nível da hierarquia com diferentes graus de compatibilidade, uma vez que os descritores podem representar mais de um grupo com pesos de representatividade diferentes nos grupos.

Uma quarta contribuição foi obtida neste doutorado pela aplicação do método SoftO-FDCL no contexto da educação médica continuada canadense. Desta aplicação obteve-se a organização flexível de comentários de médicos de família canadenses (*Canadian Family Physicians* - FPs) sobre recomendações de tratamento recebidos por e-mail. A associação de farmacêuticos canadenses (*Canadian Pharmacists Association* - CPhA) tem utilizado um recurso chamado *e-Therapeutics+*<sup>1</sup>, pelo qual é possível o gerenciamento de recomendações de tratamento baseado em evidências farmacológicas e não farmacológicas. Especificamente, *e-Therapeutics+* ajuda os FPs a saber quais opções terapêuticas estão disponíveis em uma determinada situação clínica. Para sensibilizar os FPs de que a informação terapêutica pode ser útil em cuidar de seus pacientes, os editores da CPhA selecionam informações chave, chamadas de *e-Therapeutics+ Highlights*<sup>2</sup>. *Highlights* são enviados semanalmente por e-mail para os FPs. Assim, a aplicação do método SoftO-FDCL foi realizada neste contexto possibilitando a organização dos comentários dos FPs acerca de uma nova informação contida no *Highlight* recebido, de forma que os editores da CPhA consigam acrescentar melhorias nos *Highlights* pela seleção de comentários construtivos.

Além dessas quatro principais contribuições, foram obtidas duas outras contribuições que não fazem parte da proposta principal de extração de descritores de grupos para a organização flexível proposta neste doutorado, mas que foram importantes para o entendimento do problema abordado no mesmo.

A primeira consiste na representação de documentos por meio de agrupamento fuzzy e a geração de regras fuzzy a partir desse agrupamento. Os resultados obtidos com esta abordagem de geração de regras fuzzy mostraram-se promissores para a classificação de documentos, permitindo a redução da dimensionalidade da representação usual de documentos (Nogueira et al., 2010), (Yaguinuma et al., 2010a), (Yaguinuma et al., 2010b), (Nogueira et al., 2011b), (Yaguinuma et al., 2012). Além disso, essa contribuição propicia o tratamento de imprecisão e incerteza de documentos, aspecto chave da proposta de organização flexível de documentos abordada neste doutorado.

A segunda está relacionada à Recuperação de Informação (RI) flexível. Para tanto, foi

<sup>1</sup><http://www.pharmacists.ca/index.cfm/function/store/PublicationDetail.cfm?pPub=9>

<sup>2</sup><http://www.pharmacists.ca/index.cfm/more-information/et-mcgill/>

realizado um estudo sobre a aplicação de regras fuzzy no nível da consulta de um Sistema de Recuperação de Informação (SRI). A partir deste estudo e da revisão da literatura, concluiu-se que para que um documento seja satisfatoriamente recuperado por um SRI, considerando que a imprecisão e incerteza são típicas de documentos textuais, a coleção da qual ele faz parte deve ser organizada de maneira flexível.

Além dessas contribuições, resume-se a seguir contribuições adicionais desta tese de doutorado.

- Levantamento bibliográfico de trabalhos relacionados à extração de descritores de grupos (Capítulo 2);
- Exploração e implementação de algoritmos de agrupamento fuzzy (Capítulo 2);
- Exploração e implementação de algoritmos de agrupamento fuzzy hierárquicos ((Eico et al., 2012));
- Levantamento bibliográfico de trabalhos relacionados à organização flexível de documentos (Capítulo 3);
- Exploração da influência do pré-processamento de documentos no agrupamento de documentos ((Nogueira et al., 2010), (Yaguminha et al., 2010a));
- Proposta de um projeto científico intitulado “Tratamento de imprecisão e incerteza na representação, organização e recuperação de informação textual utilizando abordagem fuzzy”, o qual foi submetido e aprovado no programa de mobilidade internacional Santander. Este projeto foi desenvolvido na Universidade de Granada - Espanha no período de Maio a Setembro de 2010.
- Proposta de um projeto científico intitulado “Gerenciamento de Imprecisão e Incerteza para Organização Flexível de Documentos Textuais”, o qual foi submetido e aprovado no Programa Institucional de Doutorado Sanduíche no Exterior (PDSE-CAPES). Este projeto foi desenvolvido na Universidade McGill - Canadá no período de Fevereiro a Agosto de 2012.
- Participação em um projeto regular de pesquisa intitulado “Métodos de Agrupamento Hierárquico para Organização Automática de Resultados de Motores de Busca”, o qual foi aprovado pela FAPESP (processo 2011/19850-9). O método HSoftO-FDCL proposto neste doutorado tem sido investigado e avaliado junto a outros pesquisadores que participam deste projeto de pesquisa.

## 6.2 Publicações provenientes deste doutorado

Durante o período deste doutorado, vários resultados foram obtidos e alguns artigos foram publicados. A lista de artigos provenientes da pesquisa realizada neste doutorado é apresentada a seguir.

1. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O. . Descriptor extraction of overlapped document clusters: a comparison between a fuzzy and a possibilistic approach proposed for flexible document organization. *Applied Soft Computing*, 2013 (Artigo em fase final de revisão para submissão).
2. NOGUEIRA, T.; CAMARGO, H.; ROSSI, R.; PLUYE, P; GRAD, R.; TANG, D.; JOHNSON-LAFLEUR, J; LEWIS, D. ; REZENDE, S. . Automatic organization of family physicians textual comments about treatment recommendations can help to identify non-constructive comments. *Computers in Biology and Medicine*. 2013 (Artigo submetido em outubro de 2012).
3. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O. . Fuzzy-DDE: a fuzzy method for the extraction of document cluster descriptors. *International Journal of Computer Information Systems and Industrial Management Applications*, 2013, v. 5, pp. 472-479.
4. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O. . Fuzzy cluster descriptors improve flexible organization of documents. In: International Conference on Intelligent Systems Design and Applications (ISDA), 2012, Kochi, Índia, pp. 616-621.
5. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O. . Fuzzy Rules for Document Classification to Improve Information Retrieval. *International Journal of Computer Information Systems and Industrial Management Applications*, v. 3, p. 210-217, 2011.
6. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O. . Fuzzy cluster descriptor extraction for flexible organization of documents. In: 11th International Conference on Hybrid Intelligent Systems (HIS), 2011, Melacca-Malásia. 11th International Conference on Hybrid Intelligent Systems (HIS), 2011. p. 528-533.
7. EICO, CYNTIA H. N. ; NOGUEIRA, T. M. ; REZENDE, S. O. . Apoio ao gerenciamento de imprecisão e incerteza em documentos textuais utilizando agrupamento fuzzy. In: SIICUSP, 2011, São Carlos - SP. SIICUSP, 2011.
8. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O. . On The Use of Fuzzy Rules to Text Document Classification. In: 10th International Conference on Hybrid Intelligent Systems, 2010, Atlanta-Georgia-USA. 10th International Conference on Hybrid Intelligent Systems, 2010.
9. NOGUEIRA, T. M. ; CAMARGO, H. A. ; REZENDE, S. O.. Tratamento de imprecisão e incerteza na identificação de documentos textuais similares. In: Congresso da Academia Trinacional de Ciências, 2009, Foz do Iguaçu-PR. Anais do C3N, 2009. v. 1.

Além desses trabalhos, houve também a colaboração em outros trabalhos relacionados:

1. YAGUINUMA, C. A., CAMARGO, H. A., SANTOS, M. T. P., NICOLETTI, M. C., e NOGUEIRA, T. M.. Fuzz-onto: A meta-ontology for representing fuzzy elements and supporting fuzzy classification rules. International Conference on Intelligent Systems Design and Applications (ISDA), p 166-171, 2012.
2. YAGUINUMA, C. A., NOGUEIRA, T. M., FERRAZ, V. R. T., SANTOS, M. T. P., e CAMARGO, H. A.. A model for representing vague linguistic terms and fuzzy rules for classification in ontologies. International Conference on Enterprise Information Systems (ICEIS), p. 438-442, 2010.
3. YAGUINUMA, C. A., SANTOS, M. T. P., CAMARGO, H. A. e NOGUEIRA, T. M.. A meta-ontology approach for representing vague linguistic terms and fuzzy rules for classification in ontologies. International Enterprise Distributed Object Computing Conference Workshops, EDOCW '10, p. 263-271, 2010.

### 6.3 Parcerias

Durante o desenvolvimento desta tese, várias parcerias foram realizadas com grupos cujas pesquisas possuem alguma relação com o foco deste trabalho. As parcerias possibilitaram a interação com pesquisadores do grupo de pesquisa no qual este doutorado está inserido e de outras instituições de ensino. Essas parcerias são citadas a seguir.

Este doutorado foi realizado no grupo de pesquisadores da área de Mineração de Textos do LABIC (Laboratório de Inteligência Computacional) do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP). Este grupo é coordenado pela Profa. Dra. Solange Oliveira Rezende e composto por alunos de iniciação científica, mestrandos, doutorandos, pós-doutorandos e pesquisadores externos, todos realizando pesquisas em alguma etapa do processo de mineração de textos. Para este doutorado, a interação com o grupo foi fundamental para a pesquisa relacionada ao gerenciamento de documentos textuais nas cinco etapas da mineração de textos: identificação do problema, pré-processamento, extração de padrões, pós-processamento e uso do conhecimento. Essa interação obteve como resultado a participação na orientação de uma iniciação científica. Os resultados desse trabalho foi publicado e apresentado no Simpósio Internacional de Iniciação Científica (SIICUSP) 2011 (Eico et al., 2012), pelo qual foi realizado o estudo de algoritmos de agrupamento fuzzy hierárquicos para organização de documentos. Também como resultado da interação com o grupo, foi proposto e aprovado o projeto regular de pesquisa à FAPESP intitulado “Métodos de Agrupamento Hierárquico para Organização Automática de Resultados de Motores de Busca”, processo 2011/19850-9, coordenado pela orientadora deste doutorado Profa. Dra. Solange Oliveira Rezende e com participação dos demais doutorandos do grupo, cujo objetivo geral é investigar novos métodos de agrupamento hierárquico *hard* e *soft* para resultados provenientes

de motores de busca, e assim desenvolver um ambiente que permita explorar, de maneira mais efetiva, os resultados obtidos por sistemas de recuperação de informação.

Este doutorado também foi realizado em parceria com o grupo CIG (Grupo de Inteligência Computacional) do Departamento de Ciência da Computação da Universidade Federal de São Carlos (DC-UFSCar). Este grupo é coordenado pela Profa. Dra. Heloisa de Arruda Camargo, coorientadora deste doutorado, e também composto por alunos de iniciação científica, mestrandos, doutorandos e pesquisadores, todos realizando pesquisas da área de teoria de conjuntos fuzzy. A interação com este grupo foi importante para discussões dos conceitos relacionados à teoria de conjuntos fuzzy, da qual obteve-se como resultado publicações de um estudo desenvolvido para a representação de regras fuzzy de documentos via ontologias fuzzy (Yaguinuma et al., 2010a) (Yaguinuma et al., 2010b) (Yaguinuma et al., 2012).

Por meio do programa de mobilidade internacional Santander, foi possível também o desenvolvimento de um trabalho colaborativo com o grupo de pesquisa da Universidade de Granada - Espanha. Este grupo de pesquisa vem trabalhando com computação flexível e sistemas de informação inteligentes (*Soft Computing and Intelligent Information Systems*<sup>3</sup>) destacando-se como um grupo reconhecido mundialmente na área de sistemas fuzzy. Fazem parte do grupo mais de 40 pesquisadores, entre pesquisadores seniores e estudantes de doutorado, sob a coordenação do professor Francisco Herrera. No geral, o grupo tem como foco de suas pesquisas o desenvolvimento de técnicas de computação flexível: sistemas fuzzy, algoritmos genéticos, sistemas fuzzy genéticos, aprendizado evolutivo e computação bioinspirada, bem como aplicações nos campos da mineração de dados, tomada de decisão, recuperação de informação e outros. Em parceria com esse grupo, as pesquisas relacionadas a este doutorado foram realizadas no laboratório SECABA<sup>4</sup>, coordenado pelo Prof. Dr. Enrique Herrera Viedma juntamente com o coordenador geral do grupo Prof. Dr. Francisco Herrera, ambos trabalham juntos na recuperação de informação e computação com palavras. As pesquisas realizadas junto à este grupo, deram origem aos experimentos iniciais sobre o tratamento de imprecisão e incerteza da recuperação de informação apresentados no Apêndice B.

Por fim, por meio do Programa Institucional de Doutorado Sanduíche no Exterior (PDSE-CAPES), foi estabelecida uma parceria com o grupo *Information Technology Primary Care Research Group* (ITPCRG) da Universidade McGill - Canadá, sob a supervisão do Prof. Dr. Pierre Pluye. Esse grupo de pesquisa desenvolveu um método de avaliação da informação (*Information Assessment Method* (IAM)) (Pluye et al., 2009)) que é regularmente enviada aos médicos de família canadenses por e-mail como parte da educação médica continuada dada aos mesmos. Por meio deste método, os pesquisadores do grupo avaliam o conteúdo destes e-mails na perspectiva dos profissionais de saúde canadenses que recebem a educação médica continuada na forma de *e-Therapeutics+ Highlights*. *e-Therapeutics+* é um recurso exclusivo do Canadá para a prescrição e gestão terapêutica

<sup>3</sup><http://sci2s.ugr.es/>

<sup>4</sup><http://sci2s.ugr.es/secaba/>

farmacológica no momento do atendimento, o qual provê aos farmacêuticos ou outros profissionais de saúde o acesso on-line a uma base de evidências sobre medicamentos canadenses e informações terapêuticas confiáveis. Assim, *e-Therapeutics+* auxilia os profissionais a saberem qual medicamento funciona em qual situação. Com isto, de forma a fornecer acesso fácil e atual à informação prática e relevante que podem ser úteis no cuidado de seus pacientes, editores da associação de farmacêuticos canadenses (Canadian Pharmacists Association) selecionam destaques interessantes de cada um dos tópicos em *e-Therapeutics+*, os quais são denominados *e-Therapeutics+ Highlights*. Estes destaques são enviados duas vezes por semana por e-mail aos usuários do *e-Therapeutics+*. Assim, o grupo ITPCRG desenvolveu um projeto que visa avaliar a forma como informações clínicas no *e-Therapeutics+* são aplicadas na prática pelos médicos de família canadenses e é patrocinado pelos Institutos Canadenses de Pesquisa em Saúde (*Canadian Institutes of Health Research (CIHR)*) e o Fundo de Investigação de Saúde do Quebec (*Fonds de Recherche en Santé du Québec (FRSQ)*). Nesse contexto, um dos métodos propostos neste doutorado foi aplicado e os resultados obtidos mostraram-se satisfatórios ao problema real abordado pelo grupo ITPCRG. Como resultado da interação com o grupo, uma documentação dos experimentos realizados e dos resultados obtidos foi submetido para o periódico *Computers in Biology and Medicine* e encontra-se em fase de avaliação pelo seu corpo editorial.

## 6.4 Limitações

Segundo Zhang et al. (2009), a qualidade dos descritores de grupos deve ser avaliada considerando 4 propriedades: (i) concisão, o qual significa que eles devem ser os mais curtos possíveis, mas suficientes para abordar o tópico do grupo; (ii) comprehensibilidade, também conhecida como transparência, o qual significa que eles devem mapear o conteúdo dos grupos; (iii) acurácia, que significa que eles devem refletir o tópico que corresponde ao grupo; e, (iv) distinção, o qual significa que eles devem ser mais frequentes em um grupo do que em outros.

Neste doutorado os descritores foram avaliados pela sua acurácia, por meio do uso de algoritmos de classificação. Por outro lado, embora a comprehensibilidade tenha sido avaliada sobre coleção Opinosis, já que esta coleção é composta por poucos documentos e um conjunto de sumários escritos por humanos, permitindo realizar uma avaliação qualitativa, a comprehensibilidade é considerada uma limitação, pois os métodos propostos extraem descritores na forma stemizada, que é difícil de entender, e para que sejam úteis de alguma forma, os descritores tem que ser usados em número elevado. Já a avaliação de concisão e distinção dos descritores extraídos pelos métodos propostos é considerada uma limitação deste doutorado. No geral, a avaliação da concisão dos descritores extraídos é bastante subjetiva e dependente do conhecimento de especialistas, o que seria oneroso para avaliação de um grande volume de documentos. Por outro lado, a distinção de des-

critores não condiz com o foco principal deste doutorado que é a organização flexível de documentos. Neste tipo de organização entende-se que um descritor pode ser igualmente representativo para mais de um grupo, já que não apenas um descritor identifica um tópico, mas um conjunto de descritores.

Outra limitação está relacionada a análises comparativas com o estado da arte. Os métodos mais citados na literatura, dos quais alguns foram citados no Capítulo 2, realizam a avaliação dos descritores por meio do desempenho do algoritmo de agrupamento. Isto ocorre porque a maioria dos métodos de extração de descritores disponíveis na literatura são do tipo DCF (*Description Comes First*) em que a extração de descritores ocorre antes, ou ao mesmo tempo, do processo de agrupamento. Sendo assim, não foram realizados experimentos comparativos com os métodos disponíveis na literatura porque os métodos propostos apresentam um mecanismo diferente dos métodos citados no estado da arte, concluindo-se que a comparação não seria adequada.

## 6.5 Trabalhos futuros

Considerando as limitações apresentadas anteriormente, é necessário investir esforços em realizar experimentos que permitam uma avaliação dos descritores extraídos com relação à concisão dos mesmos. Embora oneroso, estes experimentos podem reforçar ainda mais a utilidade dos métodos propostos.

No desenvolvimento deste doutorado, a maioria dos experimentos e avaliações concentraram-se sobre a validação do método SoftO-FDCL. No entanto, faz-se necessário executar uma quantidade maior de experimentos e avaliações sobre os métodos HSoftO-FDCL e SoftO-wFDCL, dadas suas potenciais utilidades. Em especial, sobre o método HSoftO-FDCL, o qual destaca-se como um método promissor para organização flexível de documentos, visto que a estrutura hierárquica tem se destacado no estado da arte como organização que permite auxiliar o usuário em uma busca exploratória dos resultados obtidos em Sistema de Recuperação de Informação, em diversos níveis de granularidade. A organização hierárquica flexível de documentos facilita a busca pela informação de interesse do usuário, obtendo-se uma visão complementar ao modelo baseado em uma simples lista ordenada de documentos de acordo com a relevância definida pelo usuário.

Relacionada com a aplicação do método SoftO-FDCL no contexto do grupo *Information Technology Primary Care Research Group* (ITPCRG) da Universidade McGill - Canadá, espera-se ainda desenvolver um software que automatize a organização de comentários proposta, por meio de uma interface amigável ao usuário, de forma que os comentários sejam avaliados em construtivos ou não construtivos a medida que os mesmos forem sendo enviados pelos médicos de família.

Adicionalmente, a comprovação da hipótese deste doutorado motiva a continuação das pesquisas com a realização de trabalhos futuros, como os citados anteriormente, por meio de um estágio de pós-doutorado.

Por fim, este doutorado possibilita a abertura de uma linha de pesquisa pela qual é possível a realização de explorações relativas à organização flexível de dados. Por meio desta linha de pesquisa, novas investigações podem ser consideradas sobre problemas relacionados ao tratamento de imprecisão e incerteza de dados com diferentes estruturas e padrões, desenvolvendo ações que levem à produção científica inovadora na área de Recuperação de Informação.

# Referências Bibliográficas

---

---

- Aggarwal, C. C. e Zhai, C. (2012). *Mining Text Data*. Springer. Citado na página 1.
- Akinribido, C. T., Afolabi, B. S., Akhigbe, B. I., e Udo, I. J. (2011). A fuzzy-ontology based information retrieval system for relevant feedback. *International Journal of Computer Science Issues*, 1:382–389. Citado nas páginas 2 e 37.
- Anaya-Sánchez, H., Pons-Porrata, A., e Berlanga-Llavori, R. (2008). A new document clustering algorithm for topic discovering and labeling. Em *Proceedings of the 13th Iberoamerican congress on Pattern Recognition: Progress in Pattern Recognition, Image Analysis and Applications*, páginas 161–168. Citado nas páginas 4 e 40.
- Baeza-Yates, R. A. e Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Addison-Wesley Professional, 2<sup>a</sup> edição. Citado nas páginas 7, 9, 35, e 95.
- Berry, M. e Kogan, J. (2010). *Text Mining: Applications and Theory*. Wiley InterScience. Wiley. Citado na página 1.
- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA. Citado nas páginas 10, 17, 34, 38, 40, 63, e 118.
- Bezdek, J. C. e Pal, N. R. (1992). *Fuzzy Models for Pattern Recognition*. IEEE Press, 1<sup>a</sup> edição. Citado na página 17.
- Bordogna, G., Pagani, M., e Pasi, G. (2006). A dynamic hierarchical fuzzy clustering algorithm for information filtering. Em Herrera-Viedma, E., Pasi, G., e Crestani, F., editores, *Soft Computing in Web Information Retrieval*, volume 197, páginas 3–23. Citado nas páginas 4, 10, 38, e 39.
- Bordogna, G. e Pasi, G. (2001). Modeling vagueness in information retrieval. Em Agosti, M., Crestani, F., e Pasi, G., editores, *Lectures on information retrieval*, páginas 207–241. Citado na página 2.

- Bordogna, G. e Pasi, G. (2004). Soft fusion of information access. *Fuzzy Sets and Systems*, 148:205–218. Citado na página 2.
- Bordogna, G. e Pasi, G. (2011). Soft clustering for information retrieval applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(2):138–146. Citado nas páginas 3, 9, e 39.
- Bordogna, G. e Pasi, G. (2012). A quality driven hierarchical data divisive soft clustering for information retrieval. *Knowledge-Based Systems*, 26:9–19. Citado na página 39.
- Campello, R. e Hruschka, E. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157(21):2858 – 2875. Citado nas páginas 17, 19, e 67.
- Campello, R. J., Hruschka, E. R., e Alves, V. S. (2009). On the efficiency of evolutionary fuzzy clustering. *Journal of Heuristics*, 15:43–75. Citado na página 17.
- Carmel, D., Roitman, H., e Zwerdling, N. (2009). Enhancing cluster labeling using Wikipedia. Em *Proceedings of the 32nd International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval*, páginas 139–146. Citado na página 20.
- Chen, C.-L., Tseng, F. S., e Liang, T. (2010a). An integration of WordNet and fuzzy association rule mining for multi-label document clustering. *Data & Knowledge Engineering*, 69(11):1208 – 1226. Citado nas páginas 22 e 23.
- Chen, C.-L., Tseng, F. S. C., e Liang, T. (2010b). Mining fuzzy frequent itemsets for hierarchical document clustering. *Information Processing and Management*, 46:193–211. Citado nas páginas 22 e 23.
- Chi, Z., Yan, H., e Pham, T. (1996). *Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition*. World Scientific. Citado nas páginas 33 e 119.
- Chin, O. S., Kulathuramaiyer, N., e Yeo, A. W. (2006). Automatic discovery of concepts from text. Em *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, páginas 1046–1049. Citado na página 20.
- Chitsaz, E., Taheri, M., Katebi, S. D., e Jahromi, M. Z. (2009). An improved fuzzy feature clustering and selection based on chi-squared-test. Em *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, páginas 1–6. Citado na página 67.
- Chli, M. e Wilde, P. D. (2006). Internet search: subdivision-based interactive query expansion and the soft semantic web. *Applied Soft Computing*, 6(4):372 – 383. Citado na página 35.

- Chowdhury, C. e Bhuyan, P. (2010). Information retrieval using fuzzy c-means clustering and modified vector space model. Em *Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology*, volume 1, páginas 696–700. Citado nas páginas 4, 37, e 39.
- Conrado, M. S. (2009). O efeito do uso de diferentes formas de extração de termos na comprehensibilidade e representatividade dos termos em coleções textuais na língua portuguesa. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Carlos - SP. Disponível em <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-19012010-112047/pt-br.php>. Citado na página 8.
- Crestani, F., Lalmas, M., van Rijsbergen, C., e Campbell, I. (1998). Is this document relevant? Probably. *ACM Computing Surveys*, 30(4):528–552. Citado na página 28.
- Crestani, F. e Pasi, G. (1999). Soft information retrieval: Applications of fuzzy set theory and neural networks. Em N.Kasabov e Kozma, R., editores, *Neuro-fuzzy Techniques for Intelligent Information Systems*, páginas 287–313. Physica-Verlag, Springer-Verlag Group. Citado na página 28.
- Crestani, F. e Pasi, G. (2000). *Soft Computing in Information Retrieval: Techniques and Applications*. Physica Verlag. Citado na página 2.
- Cutting, D. R., Karger, D. R., Pedersen, J. O., e Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. Em *Proceedings of the 15th Annual International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval*, páginas 318–329. Citado na página 20.
- Dae-Young e Choi (2003). Enhancing the power of web search engines by means of fuzzy query. *Decision Support Systems*, 35(1):31 – 44. Citado na página 35.
- Deng, J., Hu, J., Chi, H., e Wu, J. (2010). An improved fuzzy clustering method for text mining. Em *Proceedings of the 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing*, volume 1, páginas 65–69. Citado na página 12.
- Eico, C. H. N., Nogueira, T., Rezende, S., e Camargo, H. (2012). Apoio ao gerenciamento de imprecisão e incerteza em documentos textuais utilizando agrupamento fuzzy. Em *Anais do Simpósio Internacional de Iniciação Científica (SIICUSP)*. Citado nas páginas 98 e 100.
- Fayyad, U. M., Shapiro, G. P., e Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34. Citado na página 1.

- Feldman, R. e Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press. Citado nas páginas 4, 7, e 40.
- Frank, A. e Asuncion, A. (2010). UCI machine learning repository. <http://archive.ics.uci.edu/ml>. Citado nas páginas 53 e 55.
- Fung, B., Wang, K., e Ester, M. (2003). Hierarchical document clustering using frequent itemsets. Em *Proceedings of the International Conference on Data Mining*, páginas 59–70. Citado na página 22.
- Gabrilovich, E. e Markovitch, S. (2004). Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. Em *Proceedings of the 21st International Conference on Machine Learning*, páginas 41–49. Citado na página 57.
- Gabrilovich, E. e Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Em *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, páginas 1606–1611. Citado na página 20.
- Ganesan, K., Zhai, C., e Han, J. (2010). Opinosis: a graph based approach to abstractive summarization of highly redundant opinions. Em *Proceedings of the 23rd International Conference on Computational Linguistics*, páginas 340–348. Citado na página 54.
- Gath, I. e Geva, B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:773–781. Citado na página 10.
- Geraci, F., Pellegrini, M., Maggini, M., e Sebastiani, F. (2006). Cluster generation and cluster labelling for web snippets: A fast and accurate hierarchical solution. Em Crespani, F., Ferragina, P., e Sanderson, M., editores, *String Processing and Information Retrieval*, volume 4209, páginas 25–36. Citado na página 20.
- Gomez-Skarmeta, A. F., Delgado, M., e Vila, M. A. (1999). About the use of fuzzy clustering techniques for fuzzy model identification. *Fuzzy Sets and Systems*, 106(2):179–188. Citado na página 17.
- Guztafson, E. E. e Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix. Em *Proceedings of the IEEE Conference on Decision and Control*, páginas 761–766. Citado na página 10.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., e Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD (Special Interest Group on Knowledge Discovery & Data Mining) Explorations Newsletter*, 11(1). Citado nas páginas 57 e 120.

- Han, E.-H., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., e Moore, J. (1998). WebACE: a web agent for document categorization and exploration. Em *Proceedings of the 2nd International Conference on Autonomous Agents*, páginas 408–415. Citado na página 55.
- Hayes, P. J. e Weinstein, S. P. (1990). Construe/TIS: a system for content-based indexing of a database of news stories. Em *Proceedings of the 2nd Annual Conference on Innovative Applications of Artificial Intelligence*, páginas 1–5. Citado na página 55.
- Herrera-Viedma, E., Pasi, G., e Crestani, F. (2006). *Soft Computing in Web Information Retrieval: Models and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc. Citado na página 2.
- Horng, Y.-J., Chen, S.-M., , Chang, Y.-C., e Lee, C.-H. (2005). A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. *IEEE Transactions on Fuzzy Systems*, 13(2):216–228. Citado nas páginas 4 e 39.
- Hotho, A., Staab, S., e Stumme, G. (2003). Wordnet improves text document clustering. Em *In Proceedings of the SIGIR (Special Interest Group on Information Retrieval) 2003 Semantic Web Workshop*, páginas 541–544. Citado na página 20.
- Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., e Melton, G. B. (2007). Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia. *Computers in Biology and Medicine*, 37(3):296–304. Citado na página 86.
- Hruschka, E. R., Castro, L. N. d., e Campello, R. J. G. B. (2004). Evolutionary algorithms for clustering gene-expression data. Em *Proceedings of the 4th IEEE International Conference on Data Mining*, páginas 403–406. Citado na página 19.
- Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., e Chen, Z. (2008). Enhancing text clustering by leveraging Wikipedia semantics. Em *Proceedings of the 31st Annual International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval*, páginas 179–186. Citado na página 20.
- Hüllermeier, E. (2011). Fuzzy sets in machine learning and data mining. *Applied Soft Computing*, 11(2):1493 – 1505. Citado nas páginas 4 e 29.
- Ishibuchi, H., Nakashima, T., e Murata, T. (1999). Voting fuzzy rule-based systems for pattern classification problems. *Fuzzy Sets and Systems*, 103:223–238. Citado na página 33.

- Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. Em *Proceedings of the 10th European Conference on Machine Learning*, páginas 137–142. Citado na página 57.
- Kaufman, L. e Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Citado nas páginas 10 e 17.
- Klir, G. J. e Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: theory and applications*. Prentice-Hall, 1<sup>a</sup> edição. Citado nas páginas 3, 11, 32, e 33.
- Kozielski, M. (2007). Multilevel conditional fuzzy c-means clustering of XML documents. Em Kok, J., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenić, D., e Skowron, A., editores, *Knowledge Discovery in Databases*, volume 4702, páginas 532–539. Springer Berlin Heidelberg. Citado nas páginas 4 e 39.
- Kraft, D. H., Pasi, G., e Bordogna, G. (2006). Vagueness and uncertainty in information retrieval: how can fuzzy sets help? Em *Proceedings of the International Workshop on Research Issues in Digital Libraries*, páginas 1–10. Citado nas páginas 2, 3, e 28.
- Krishnapuram, R. e Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110. Citado nas páginas 12, 13, e 67.
- Lang, K. (1995). Newsweeder: learning to filter netnews. Em *Proceedings of the 20th International Conference on Machine Learning*, páginas 331–339. Citado na página 55.
- Lee, K.-M. (2001). Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies. *Proceedings of the Joint 9th IFSA (International Fuzzy Systems Association) World Congress and 20th NAFIPS (North American Fuzzy Information Processing Society) International Conference*, 5:2977–2982. Citado nas páginas 4 e 39.
- Lopez-Herrera, A., Herrera-Viedma, E., e Herrera, F. (2009). Applying multi-objective evolutionary algorithms to the automatic learning of extended boolean queries in fuzzy ordinal linguistic information retrieval systems. *Fuzzy Sets and Systems*, 160:2192–2205. Citado nas páginas 2 e 37.
- Luger, G. (2004). *Artificial Intelligence: Structures and strategies for complex problem solving*. Addison Wesley Longman, 5<sup>a</sup> edição. Citado na página 35.
- Lynn, S. e Ng, Y.-K. (2008). Using vagueness measures to re-rank documents retrieved by a fuzzy set information retrieval model. Em *Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, páginas 39–43. Citado na página 37.
- Manning, C. D., Raghavan, P., e Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press. Citado nas páginas 7, 9, 35, 58, 67, e 95.

- Matsumoto, T. e Hung, E. (2012). A transduction-based approach to fuzzy clustering, relevance ranking and cluster label generation on web search results. *Journal of Intelligent Information Systems*, 38(2):419–448. Citado nas páginas 22 e 23.
- Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38:39–41. Citado na página 20.
- Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Education. Citado na página 56.
- Moore, J., Han, E.-H., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., e Mobasher, B. (1997). Web page categorization and feature selection using association rule and principal component clustering. Em *Proceedings of the 7th Workshop on Information Technologies and Systems*. Citado na página 55.
- Muhr, M., Kern, R., e Granitzer, M. (2010). Analysis of structural relationships for hierarchical cluster labeling. Em *Proceedings of the 33rd International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval*, páginas 178–185. Citado na página 67.
- Nogueira, T., Rezende, S., e Camargo, H. (2012a). Fuzzy cluster descriptors improve flexible organization of documents. Em *Proceedings of the 12th International Conference on Intelligent Systems Design and Applications*, páginas 616–621. Citado nas páginas 62 e 96.
- Nogueira, T. M., Camargo, H. A., e Rezende, S. O. (2011a). Fuzzy cluster descriptor extraction for flexible organization of documents. *Proceedings of the 11th International Conference on Hybrid Intelligent Systems*, páginas 528–533. Citado nas páginas 58 e 96.
- Nogueira, T. M., Camargo, H. A., e Rezende, S. O. (2011b). Fuzzy rules for document classification to improve information retrieval. *International Journal of Computer Information Systems and Industrial Management Applications*, 3:1–8. Citado nas páginas 34 e 97.
- Nogueira, T. M., Camargo, H. A., e Rezende, S. O. (2013). Fuzzy-DDE: a fuzzy method for the extraction of document cluster descriptors. *International Journal of Computer Information Systems and Industrial Management Applications*, 5:472–479. Citado na página 96.
- Nogueira, T. M., de A. Camargo, H., Rossi, R. G., Pluye, P., Grad, R., Tang, D. L., Johnson-Lafleur, J., Lewis, D., e Rezende, S. O. (2012b). Automatic organization of family physicians textual comments about treatment recommendations can help to identify non-constructive comments. *Computers in Biology and Medicine*. (Submetido em outubro de 2012). Citado na página 96.

- Nogueira, T. M., Rezende, S. O., e Camargo, H. A. (2010). On the use of fuzzy rules to text document classification. Em *Proceedings of the 10th International Conference on Hybrid Intelligent Systems*, páginas 19–24. Citado nas páginas 34, 97, e 98.
- Oliveira, J. V. d. e Pedrycz, W. (2007). *Advances in Fuzzy Clustering and its Applications*. John Wiley & Sons, Inc. Citado nas páginas ix e 13.
- Osinski, S. e Weiss, D. (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54. Citado nas páginas 20 e 22.
- Pal, N. R., Pal, K., Keller, J. M., e Bezdek, J. C. (2005). A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530. Citado nas páginas 10, 12, 13, 14, 40, e 63.
- Pasi, G. (2002). Flexible information retrieval: some research trends. *Mathware and Soft Computing*, 9:107–121. Citado na página 37.
- Pedrycz, A. e Reformat, M. (2006). Hierarchical FCM in a stepwise discovery of structure in data. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 10:244–256. Citado nas páginas 10, 14, 40, e 74.
- Pedrycz, W. (1996). Conditional fuzzy c-means. *Pattern Recognition Letter*, 17(6):625–631. Citado nas páginas ix e 14.
- Pedrycz, W. (1998). *Computational Intelligence: An Introduction*. Boca Raton, FL. Citado na página 2.
- Pedrycz, W. e Gomide, F. (2007). *Fuzzy Systems Engineering: Toward Human-Centric Computing*. Wiley-IEEE Press, 1<sup>a</sup> edição. Citado na página 11.
- Pluye, P., Grad, R., Granikov, V., Jagosh, J., e Leung, K. H. (2010a). Evaluation of email alerts in practice: part 1 - review of the literature on clinical emailing channels. *Journal of Evaluation Clinical Practice*, 16(6):1227–1235. Citado na página 82.
- Pluye, P., Grad, R., Johnson-Lafleur, J., Bambrick, T., Burnand, B., e Mercer, J. (2010b). Evaluation of email alerts in practice: part 2 - validation of the information assessment method (IAM). *Journal of Evaluation Clinical Practice*, 16(6):1236–1243. Citado na página 82.
- Pluye, P., Grad, R., Repchinsky, C., Farrell, B., Johnson-Lafleur, J., e Bambrick, T. (2009). IAM: A comprehensive and systematic information assessment method for electronic knowledge resources. Em *Handbook of Research on IT Management and Clinical Data Administration in Healthcare*, páginas 521–548. Citado nas páginas 82 e 101.

- Pluye, P., Grad, R., Repchinsky, C., Jovaisas, B., Lewis, D., Tang, D., Granikov, V., Bonar, J., e Marlow, B. (2012). Better than best evidence? The information assessment method can help information providers to use family physicians' feedback for 2-way knowledge translation. *Canadian Family Physician*. (Aceito para publicação). Citado nas páginas 82 e 85.
- Popescul, A. e Ungar, L. (2000). Automatic labeling of document clusters. Disponível em <http://citeseer.nj.nec.com/popescul00automatic.html>. Citado na página 67.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. Citado na página 57.
- Radecki, T. (1979). Fuzzy set theoretical approach to document retrieval. *Information Processing and Management*, 15(5):247–260. Citado na página 35.
- Rodrigues, E. M. e Sacks, L. (2005). Learning topic hierarchies from text documents using a scalable hierarchical fuzzy clustering method. Em *Proceedings of the International Conference on Recent Advances in Soft Computing*, páginas 269–274. Citado nas páginas 4, 10, e 38.
- Rodrigues, M. E. S. M. e Sacks, L. (2004). A scalable hierarchical fuzzy clustering algorithm for text mining. Em *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, páginas 1–6. Citado na página 38.
- Salton, G. e McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill. Citado na página 44.
- Santos, R. T., Nievola, J. C., Freitas, A. A., e Lopes, H. S. (1999). Extração de regras de redes neurais via algoritmos genéticos. Em *Anais do IV Congresso Brasileiro de Redes Neurais*, páginas 158–163. Citado na página 3.
- Saraçoglu, R., Tütüncü, K., e Allahverdi, N. (2007). A fuzzy clustering approach for finding similar documents using a novel similarity measure. *Expert Systems with Applications*, 33:600–605. Citado nas páginas 4, 38, e 39.
- Saraçoglu, R., Tütüncü, K., e Allahverdi, N. (2008). A new approach on search for similar documents with multiple categories using fuzzy clustering. *Expert Systems with Applications*, 34:2545–2554. Citado nas páginas 4, 29, e 38.
- Schneider, K. (2005). Techniques for improving the performance of naïve bayes for text classification. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, 3406:682–693. Citado na página 56.
- Shanahan, J. e Roma, N. (2003). Improving SVM text classification performance through threshold adjustment. *Machine Learning, Lecture Notes in Computer Science*, 2837:361–372. Citado na página 56.

- Soares, M. V. B., Prati, R. C., e Monard, M. C. (2008). PRETEXT II: descrição da reestruturação da ferramenta de pré-processamento de textos. Relatório Técnico 333, ICMC-USP. Citado nas páginas 56 e 117.
- Song, S., Guo, Z., e Chen, P. (2011). Fuzzy document clustering using weighted conceptual model. *Information Technology*, 10:1178–1185. Citado na página 39.
- Tan, P.-N., Steinbach, M., e Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. Citado na página 1.
- Tjhi, W.-C., Chen, L., e Member, S. (2009). Dual fuzzy-possibilistic co-clustering for categorization of documents. *IEEE Transactions on Fuzzy Systems*, 17(3):532–543. Citado na página 39.
- Toda, H. e Kataoka, R. (2005). A clustering method for news articles retrieval system. Em *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, páginas 988–989. Citado na página 20.
- Torra, V. (2005). Fuzzy c-means for fuzzy hierarchical clustering. Em *Proceedings of the IEEE International Conference on Fuzzy Systems*, páginas 646–651. Citado na página 38.
- Treeratpituk, P. e Callan, J. (2006). Automatically labeling hierarchical clusters. Em *Proceedings of the International Conference on Digital Government Research*, páginas 167–176. Citado nas páginas 20 e 67.
- Wang, L. e Mendel, J. (1992). Generating fuzzy rules by learning from examples. *IEEE Transaction on Fuzzy Systems, Man and Cybernetics*, 22:414–1427. Citado nas páginas 34 e 118.
- Yaginuma, C. A., Camargo, H. A., Santos, M. T. P., Nicoletti, M. C., e Nogueira, T. M. (2012). Fuzz-onto: a meta-ontology for representing fuzzy elements and supporting fuzzy classification rules. Em *Proceedings of the 12th International Conference on Intelligent Systems Design and Applications*, páginas 166–171. Citado nas páginas 97 e 101.
- Yaginuma, C. A., Nogueira, T. M., Ferraz, V. R. T., Santos, M. T. P., e Camargo, H. A. (2010a). A model for representing vague linguistic terms and fuzzy rules for classification in ontologies. Em *Proceedings of the International Conference on Enterprise Information Systems*, páginas 438–442. Citado nas páginas 97, 98, e 101.
- Yaginuma, C. A., Santos, M. T. P., Camargo, H. A., e Nogueira, T. M. (2010b). A meta-ontology approach for representing vague linguistic terms and fuzzy rules for classification in ontologies. Em *Proceedings of the 14th IEEE International Enterprise*

- Distributed Object Computing Conference Workshops*, páginas 263–271. Citado nas páginas 97 e 101.
- Yan, Y., Chen, L., e Tjhi, W.-C. (2012). Fuzzy semi-supervised co-clustering for text documents. *Fuzzy Sets and Systems*, páginas 1–16. Citado nas páginas 2 e 39.
- Zadeh, L. (1997). *What is Soft Computing*. Springer-Verlag. Citado na página 2.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353. Citado na página 2.
- Zadrożny, S. e Nowacka, K. (2008). Interpretation of the keywords weights in information retrieval: fuzzy logic based approaches. Em *Proceedings of the 19th International Workshop on Database and Expert Systems Application*, páginas 657–661. Citado nas páginas 2 e 37.
- Zadrożny, S. e Nowacka, K. (2009). Fuzzy information retrieval model revisited. *Fuzzy Sets and Systems*, 160:2173–2191. Citado nas páginas 2 e 37.
- Zamir, O. e Etzioni, O. (1998). Web document clustering: a feasibility demonstration. Em *Proceedings of the 21st Annual International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval*, páginas 46–54. Citado na página 22.
- Zhang, C. (2009). Document clustering description based on combination strategy. Em *Proceedings of the International Conference on Innovative Computing, Information and Control*, páginas 1084–1088. Citado nas páginas 21, 24, e 43.
- Zhang, C., Wang, H., Liu, Y., e Xu, H. (2009). Document clustering description extraction and its application. Em *Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, páginas 370–377. Citado nas páginas 24, 52, e 102.
- Zimmermann, H. J. (1991). *Fuzzy Set Theory and Its Applications*. Kluwer Academic Publishers, Boston, USA, 2<sup>a</sup> edição. Citado na página 32.



# Classificação de Documentos Utilizando Regras Fuzzy

---



---

Neste apêndice é apresentado um mecanismo para classificação de documentos utilizando regras fuzzy. Este mecanismo, ilustrado na Figura A.1, é composto de duas etapas.

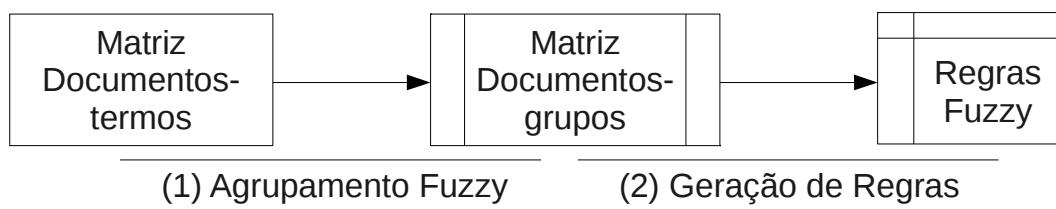


Figura A.1: Método fuzzy para classificação de documentos

A primeira etapa consiste no agrupamento fuzzy. Nessa etapa ocorre a transformação da matriz documentos-termos em uma matriz de menor dimensionalidade, chamada de matriz documentos-grupos.

Para melhor compreender esta etapa, considere a matriz documentos-termos representada em  $M_{n \times k}$  (Matriz A.1), para  $n$  igual a quantidade de documentos e  $k$  igual a quantidade de termos. Esta matriz é obtida do pré-processamento dos documentos de uma determinada coleção utilizando a ferramenta Pretext (Soares et al., 2008). Cada linha desta matriz representa um documento  $d_i$ , com  $1 \leq i \leq n$ , e cada coluna representa um termo  $t_j$ , com  $1 \leq j \leq k$ . Cada célula da matriz é composta pela frequência  $\sigma(t_j, d_i)$  do termo  $t_j$  no documento  $d_i$ . A última coluna da matriz identifica a classe dos documentos, considerando que os mesmos foram previamente rotulados.

$$M_{n \times k} = \begin{pmatrix} 0.4 & 0.1 & 0.3 & 0.5 & 0.1 & 0.2 & 0.7 & 0.5 & 0.6 & a \\ 0.3 & 0.5 & 0.2 & 0.7 & 0.5 & 0.2 & 0.7 & 0.3 & 0.2 & b \\ 0.1 & 0.2 & 0.7 & 0.3 & 0.7 & 0.4 & 0.5 & 0.3 & 0.1 & c \\ 0.4 & 0.5 & 0.3 & 0.7 & 0.3 & 0.2 & 0.7 & 0.5 & 0.1 & c \\ 0.2 & 0.1 & 0.2 & 0.5 & 0.1 & 0.2 & 0.7 & 0.3 & 0.8 & a \\ 0.5 & 0.3 & 0.4 & 0.3 & 0.1 & 0.4 & 0.5 & 0.3 & 0.1 & a \\ 0.8 & 0.9 & 0.3 & 0.7 & 0.1 & 0.2 & 0.7 & 0.5 & 0.0 & b \\ 0.4 & 0.5 & 0.3 & 0.5 & 0.3 & 0.4 & 0.3 & 0.1 & 0.9 & c \\ 0.4 & 0.5 & 0.3 & 0.3 & 0.3 & 0.2 & 0.7 & 0.5 & 0.1 & c \\ 0.4 & 0.5 & 0.2 & 0.5 & 0.7 & 0.4 & 0.3 & 0.1 & 0.9 & a \end{pmatrix} \quad (A.1)$$

Na primeira etapa do mecanismo proposto, os documentos são agrupados por meio do algoritmo de agrupamento Fuzzy C-Means (FCM) (Bezdek, 1981) (apresentado no Capítulo 2) e obtém-se uma matriz documentos-grupos representada em  $W_{n \times c}$  (Matriz A.2), para  $n$  igual a quantidade de documentos e  $c$  igual a quantidade de grupos. Cada linha dessa matriz representa um documento  $d_i$ , com  $1 \leq i \leq n$ , e cada coluna representa um grupo  $g_l$ , com  $1 \leq l \leq c$ . Cada célula da matriz é composta pelo grau de pertinência  $\mu(d_i, g_l)$  do documento  $d_i$  no grupo  $g_l$ . A última coluna desta matriz também identifica a classe dos documentos. É importante ressaltar que as classes são obtidas em um processo de rotulação dos documentos anterior ao mecanismo proposto e que as mesmas não são consideradas no processo de agrupamento, mas no processo de geração das regras.

$$W_{n \times c} = \begin{pmatrix} 0.3 & 0.6 & 0.1 & a \\ 0.3 & 0.5 & 0.2 & b \\ 0.1 & 0.2 & 0.7 & c \\ 0.3 & 0.4 & 0.3 & c \\ 0.2 & 0.2 & 0.6 & a \\ 0.4 & 0.5 & 0.1 & a \\ 0.1 & 0.8 & 0.1 & b \\ 0.3 & 0.3 & 0.4 & c \\ 0.1 & 0.1 & 0.8 & c \\ 0.2 & 0.1 & 0.7 & a \end{pmatrix} \quad (A.2)$$

Na segunda etapa do mecanismo proposto, as regras fuzzy são geradas a partir da matriz documentos-grupos. Neste trabalho foi utilizado o método de Wang&Mendell (Wang e Mendel, 1992) para geração de regras fuzzy pela sua facilidade de implementação. As regras geradas para classificação de documentos assumem o seguinte formato:

**SE**  $G_1$  é  $a_1$  **E**  $G_2$  é  $a_2$  **E**  $\dots$  **E**  $G_c$  é  $a_c$  **ENTÃO** *Class*

Nesse formato de regra,  $G_1, G_2 \dots G_c$  são variáveis linguísticas que representam os  $c$  grupos formados pelo agrupamento de documentos, as quais foram granularizadas nos

termos linguísticos  $A = \{a_1, a_2, a_3\}$ . Por exemplo, o grupo  $g_1$ , é representado na Figura A.2 como uma variável linguística  $G_1$  granularizada nos termos linguísticos  $a_1 = Baixo$ ,  $a_2 = Médio$ ,  $a_3 = Alto$ , os quais são caracterizados por uma função de pertinência triangular. Esta função considera os valores mínimo  $\min\{\mu(\mathbf{d}_i, g_1) | \forall \mathbf{d}_i, 1 \leq i \leq n\}$  e máximo  $\max\{\mu(\mathbf{d}_i, g_1) | \forall \mathbf{d}_i, 1 \leq i \leq n\}$  dos graus de pertinência dos documentos no grupo  $g_1$ .

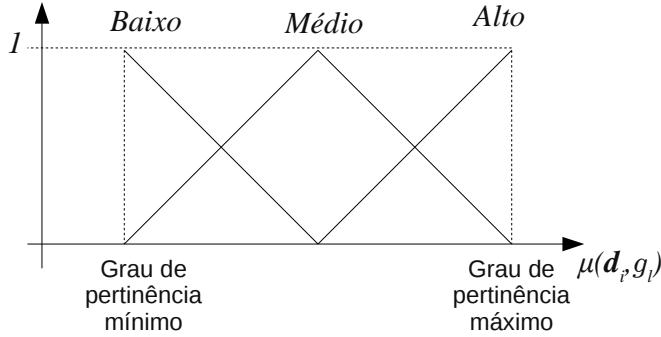


Figura A.2: Variável linguística  $G_1$

No consequente da regra,  $Class$  é uma das classes nas quais os documentos são previamente rotulados.

Uma vez geradas as regras, as mesmas podem ser utilizadas para classificar os documentos por meio de um Sistema de Classificação Fuzzy (SCF), o qual consiste de um Sistema Fuzzy Baseado em Regras (SFBR) desenvolvido com o objetivo específico de executar a tarefa de classificação. Um SCF utiliza métodos de inferência que são específicos para esta tarefa.

Na classificação de documentos utilizando o mecanismo proposto, o método de inferência utilizado funciona como segue. Considere  $\mathbf{d}_i = [\mu(\mathbf{d}_i, g_1), \mu(\mathbf{d}_i, g_2), \dots, \mu(\mathbf{d}_i, g_c)]$  a representação de um documento a ser classificado, no qual  $\mu(\mathbf{d}_i, g_l)$ ,  $1 \leq i \leq n$  e  $1 \leq l \leq c$ , é o grau de pertinência do documento  $\mathbf{d}_i$  no grupo  $g_l$ . Considere ainda  $\{R_1, R_2, \dots, R_z\}$  o conjunto de  $z$  regras do sistema de classificação. Cada regra possui  $c$  antecedentes, uma vez que todos os grupos são considerados como variáveis linguísticas das regras. Com base no método de Raciocínio Fuzzy Clássico (RFC), também conhecido como método da regra vencedora (Chi et al., 1996), o processo de inferência utilizado para classificar o documento  $\mathbf{d}_i$  é:

1. Calcular o grau de compatibilidade entre o documento de entrada  $\mathbf{d}_i$  e cada regra  $R_s$ , para  $s = 1, \dots, z$ :

$$C_{compat}(R_s, \mathbf{d}_i) = t(A_q(\mu(\mathbf{d}_i, g_1)), A_q(\mu(\mathbf{d}_i, g_2)), \dots, A_q(\mu(\mathbf{d}_i, g_c))),$$

no qual  $t$  denota uma  $t$ -norma e  $A_q(\mu(\mathbf{d}_i, g_l))$  denota o grau de pertinência do grau de pertinência do documento  $\mathbf{d}_i$  no grupo  $g_l$  no conjunto fuzzy  $a_q$ , o qual é um

---

termo linguístico com  $1 \leq q \leq |A|$ , para  $A = \{a_1, a_2, a_3\}$  é o conjunto de termos linguísticos da regra  $R_s$ .

2. Encontrar a regra  $R_{smax}$  que possui o maior grau de compatibilidade com o documento  $d_i$ , isto é,

$$C_{ompat}(R_s, d_i) = \max \{Compat(R_s, d_i)\}, s = 1, \dots, z$$

3. Atribuir a classe  $Class_{smax}$  ao documento  $d_i$ , na qual  $Class_{smax}$  é a classe do consequente da regra  $R_{smax}$  que possui o maior grau de compatibilidade com o documento, encontrada no passo anterior.

Uma vez geradas as regras a partir de grupos fuzzy de documentos e definido o método de inferência a ser utilizado pelo SCF, foram realizados cinco testes para avaliar este mecanismo. Em todos os testes, foram comparadas as taxas de classificação correta obtidas a partir do método proposto e de outros quatro métodos bastante conhecidos na literatura para classificação, disponíveis na ferramenta WEKA (Hall et al., 2009): KNN, J48, Naive Bayes e OneR.

Além dos testes relacionados ao desempenho da classificação, também foi analisada a influência do pré-processamento na geração de regras fuzzy obtidas do agrupamento de documentos. Essa análise foi realizada porque, ao final da etapa de pré-processamento, a matriz documentos-termos é de alta dimensionalidade e altamente esparsa. Tais características, em alguns casos, podem fazer com que o processo de agrupamento seja de alto custo computacional ou mesmo impossível, afetando negativamente o resultado da extração de conhecimento.

Para os experimentos de avaliação foram utilizadas cinco coleções de documentos extraídos de anais da biblioteca digital da ACM<sup>1</sup>. Cada coleção é composta de 90 documentos (instâncias) escolhidos aleatoriamente e distribuídos em 5 classes.

Utilizando a ferramenta Pretext para pré-processamento dos documentos, foi possível reduzir a quantidade de termos representativos da coleção pela frequência. Esta tarefa de redução, chamada de seleção de termos, foi realizada de acordo com cinco diferentes condições (testes) apresentadas na Tabela A.1. Estes testes foram definidos variando o valor mínimo e máximo de frequência dos termos nas coleções de documentos. Estes valores particulares de frequências foram escolhidos com o objetivo de projetar as condições de teste variando as frequências mínimas e o número de termos, considerando que o objetivo da avaliação era analisar o quanto é possível fazer uma seleção de termos sem perda de informação. A quantidade de documentos e termos de cada coleção em cada teste é apresentada na Tabela A.2.

A dimensionalidade da matriz documentos-termos também foi reduzida pelo agrupamento de documentos, na qual a quantidade de grupos escolhida foi a quantidade de

---

<sup>1</sup><http://portal.acm.org/dl.cfm?coll=portal&dl=ACM&CFID=25286060&CFTOKEN=97119727>

| Testes | Frequências |        |
|--------|-------------|--------|
|        | Mínimo      | Máximo |
| 1      | 50          | 500    |
| 2      | 100         | 300    |
| 3      | 50          | 100    |
| 4      | 500         | 1000   |
| 5      | 50          | 1000   |

Tabela A.1: Variação de frequência para seleção de termos

| Coleção | Documentos | Quantidade de termos |         |         |         |         |
|---------|------------|----------------------|---------|---------|---------|---------|
|         |            | Teste 1              | Teste 2 | Teste 3 | Teste 4 | Teste 5 |
| Exp1    | 399        | 3132                 | 1357    | 1436    | 1713    | 3398    |
| Exp2    | 410        | 2722                 | 1166    | 1299    | 1442    | 2945    |
| Exp3    | 424        | 3073                 | 1371    | 1356    | 1741    | 3326    |
| Exp4    | 394        | 3072                 | 1313    | 1430    | 1653    | 3352    |
| Exp5    | 471        | 3471                 | 1522    | 1577    | 1916    | 3807    |

Tabela A.2: Coleções utilizadas nos experimentos e respectivas quantidades de documentos e termos

classes em cada coleção. Após o agrupamento dos documentos, as regras foram geradas e a quantidade de variáveis das regras é igual a quantidade de grupos, uma vez que as regras são geradas a partir dos graus de pertinência de cada documento em cada grupo.

Os resultados obtidos a partir dos métodos de classificação em cada teste e a média de cada método testado são apresentados nas Tabelas A.3, A.4, A.5, A.6 e A.7. O *ranking* de cada método em cada coleção é apresentado entre parênteses.

Para testar se há ou não diferença significativa entre os métodos, o teste de Friedman com o pós-teste de Bonferroni-Dunn foi realizado com a hipótese nula de que o desempenho dos cinco métodos, medidos em termos de taxas de classificação correta, são comparáveis. A hipótese nula foi rejeitada com 95% de confiança e os resultados mostraram que o método proposto, o método J48 e o método Naive Bayes apresentam resultados similares e a diferença entre os métodos não é estatisticamente significante. Nos testes, o método Naive Bayes foi sempre o primeiro no *ranking*, e o método proposto e o método J48 alternaram entre o segundo e o terceiro lugar no *ranking*. Apesar de apresentar resultados semelhantes aos métodos Naive Bayes e J48, o método proposto apresenta como vantagem a redução da dimensionalidade da matriz documentos-termos pela geração de regras fuzzy a partir do agrupamento fuzzy de documentos. Assim, o método proposto é uma abordagem interessante para a classificação de documentos porque a alta dimensionalidade da matriz documentos-termos dificulta a comprehensibilidade dos documentos a serem classificados.

| Coleção | Método Proposto | KNN       | J48       | Naive Bayes | OneR      |
|---------|-----------------|-----------|-----------|-------------|-----------|
| Exp1    | 88.0 (1)        | 50.60 (4) | 77.40 (3) | 87.90 (2)   | 39.80 (5) |
| Exp2    | 84.0 (3)        | 50.70 (4) | 91.50 (1) | 90.40 (2)   | 50.40 (5) |
| Exp3    | 48.0 (3)        | 27.10 (5) | 78.70 (2) | 88.90 (1)   | 39.60 (4) |
| Exp4    | 88.0 (2)        | 51.00 (4) | 87.30 (3) | 97.90 (1)   | 46.70 (5) |
| Exp5    | 96.0 (1)        | 43.90 (4) | 83.80 (3) | 92.30 (2)   | 40.30 (5) |
| Average | 80.8            | 44.66     | 83.74     | 91.48       | 43.36     |
| Rank    | 2.00            | 4.20      | 2.40      | 1.60        | 4.80      |

Tabela A.3: Teste 1 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR

Além dos testes estatísticos, os resultados obtidos em cada teste foram analisados sob

| Coleção | Método Proposto | KNN       | J48       | Naive     | OneR       |
|---------|-----------------|-----------|-----------|-----------|------------|
| Exp1    | 60.0 (3)        | 41.10 (4) | 77.40 (2) | 87.90 (1) | 36.090 (5) |
| Exp2    | 40.0 (5)        | 56.09 (3) | 86.09 (2) | 90.00 (1) | 49.260 (4) |
| Exp3    | 60.0 (3)        | 33.01 (5) | 77.83 (2) | 86.08 (1) | 36.500 (4) |
| Exp4    | 72.0 (3)        | 61.16 (4) | 87.05 (2) | 94.41 (1) | 41.870 (5) |
| Exp5    | 68.0 (3)        | 54.98 (4) | 77.70 (2) | 88.95 (1) | 36.090 (5) |
| Rank    | 3.40            | 4.00      | 2.00      | 1.00      | 4.600      |

Tabela A.4: Teste 2 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR

| Coleção | Método Proposto | KNN       | J48       | Naive     | OneR      |
|---------|-----------------|-----------|-----------|-----------|-----------|
| Exp1    | 84.0 (1)        | 37.09 (4) | 71.92 (3) | 80.20 (2) | 31.82 (5) |
| Exp2    | 96.0 (1)        | 50.48 (4) | 84.39 (2) | 82.68 (3) | 40.97 (5) |
| Exp3    | 88.0 (1)        | 23.82 (5) | 69.57 (3) | 79.24 (2) | 30.66 (4) |
| Exp4    | 48.0 (4)        | 48.47 (3) | 83.75 (2) | 91.37 (1) | 39.59 (5) |
| Exp5    | 60.0 (3)        | 32.27 (4) | 73.03 (2) | 83.65 (1) | 29.08 (5) |
| Rank    | 2.00            | 4.00      | 2.40      | 1.80      | 4.80      |

Tabela A.5: Teste 3 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR

| Coleção | Método Proposto | KNN       | J48       | Naive     | OneR      |
|---------|-----------------|-----------|-----------|-----------|-----------|
| Exp1    | 88.0 (2)        | 55.63 (4) | 76.69 (3) | 88.47 (1) | 40.35 (5) |
| Exp2    | 84.0 (3)        | 55.12 (4) | 86.58 (2) | 90.73 (1) | 50.48 (5) |
| Exp3    | 68.0 (3)        | 42.68 (4) | 81.60 (2) | 89.38 (1) | 39.62 (5) |
| Exp4    | 52.0 (4)        | 60.65 (3) | 89.84 (2) | 94.92 (1) | 46.70 (5) |
| Exp5    | 64.0 (3)        | 51.38 (4) | 82.16 (2) | 90.87 (1) | 40.33 (5) |
| Rank    | 3.00            | 3.80      | 2.20      | 1.00      | 5.00      |

Tabela A.6: Teste 4 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR

| Coleção | Método Proposto | KNN       | J48       | Naive     | OneR      |
|---------|-----------------|-----------|-----------|-----------|-----------|
| Exp1    | 84.0 (2)        | 53.63 (4) | 81.95 (3) | 87.96 (1) | 33.08 (5) |
| Exp2    | 80.0 (3)        | 51.21 (4) | 93.17 (1) | 91.07 (2) | 50.48 (5) |
| Exp3    | 68.0 (3)        | 32.54 (5) | 77.59 (2) | 89.38 (1) | 40.09 (4) |
| Exp4    | 72.0 (3)        | 53.55 (4) | 87.56 (2) | 95.93 (1) | 45.93 (5) |
| Exp5    | 60.0 (3)        | 42.46 (4) | 82.37 (2) | 91.93 (1) | 40.12 (5) |
| Rank    | 2.80            | 4.20      | 2.00      | 1.20      | 4.80      |

Tabela A.7: Teste 5 - Taxas de classificação corretas obtidas pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR

três diferentes pontos de vista, representados por gráficos. Nos gráficos, cada barra representa uma coleção ou método, dependendo da análise realizada e conforme apresentados nas legendas, e o eixo vertical representa os resultados obtidos em cada teste. A fim de dispor múltiplas linhas em um mesmo gráfico, mas separadas umas das outras, um gráfico de linhas empilhadas (*Stacked Line Chart*) foi utilizado.

Na primeira análise, o objetivo foi verificar se as quantidades de termos derivadas da seleção de termos apresentadas na Tabela 1, interfere no método proposto. Assim, na Figura A.3 é apresentada a comparação entre os resultados obtidos pelo método proposto em cada teste para cada coleção. Nesta figura é possível observar que o pré-processamento que mais interfere nos resultados, diminuindo a qualidade da classificação, é a seleção de termos do Teste 2, em que a quantidade de termos foi reduzida em relação ao Teste 1. Este resultado sugere que uma grande redução na quantidade de termos conduz a perda de informação.

O objetivo da segunda análise foi observar os resultados obtidos a partir de todos os algoritmos de classificação em função das condições de pré-processamento. Assim, na

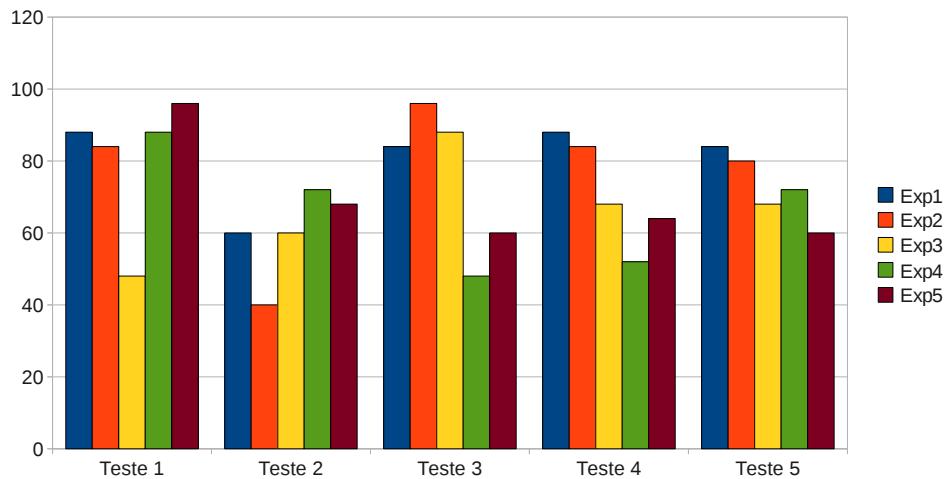


Figura A.3: Influência da quantidade de termos no desempenho da classificação obtida pelo método proposto

Figura A.4 são apresentadas apenas a média de resultados obtida por cada algoritmo em cada teste. Essa análise reforça a conclusão obtida na primeira análise, de que o método proposto é sensível à redução de termos. Observa-se que os testes 2, 3 e 4, nos quais houve uma grande redução na quantidade de termos, o método proposto também apresentou desempenho inferior quando comparado com os outros algoritmos.

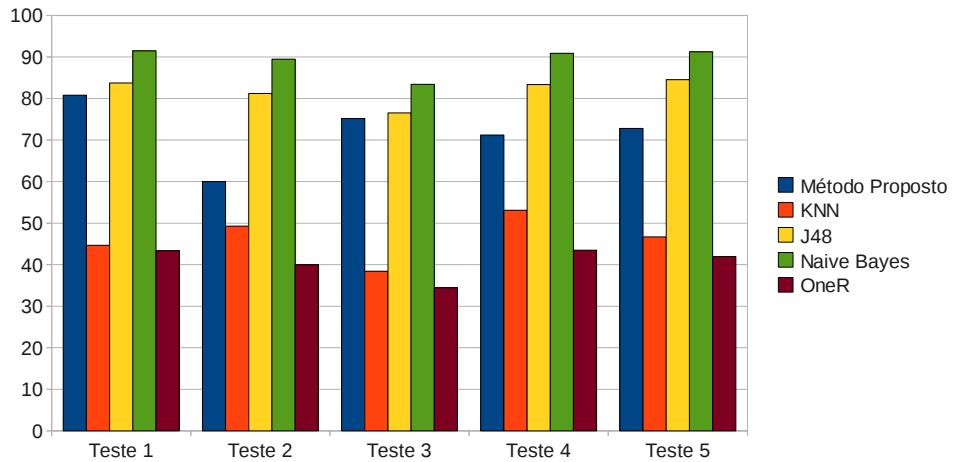


Figura A.4: Influência da quantidade de termos no desempenho da classificação pelo método proposto e pelos métodos KNN, J48, Naive Bayes e OneR

Por fim, na terceira análise, os testes 1 a 5 foram organizados em ordem crescente da frequência mínima e renomeados como testes de A a E, como mostrado na Tabela A.8. Nesta análise, apresentada na Figura A.5, o objetivo é mostrar o que mais influencia nos resultados do método proposto: a mudança na frequência mínima (Testes A a E) ou a quantidade de termos em um dado intervalo de frequências (Testes 1 a 5). Como é possível observar no gráfico, a mudança da frequência mínima, a qual foi reduzida do Teste C para o Teste D e do Teste D para o Teste E, implicam mudanças no resultado. Do Teste C para o Teste D, a frequência mínima foi aumentada e a frequência máxima

foi reduzida, resultando em uma redução da quantidade de termos. Do Teste D para o Teste E, as frequências mínima e máxima foram aumentadas, resultando no aumento da quantidade de termos. As taxas de classificação, neste caso, voltaram ao mesmo nível do Teste C. Sendo assim, pode-se observar que as taxas de classificação são diretamente influenciadas pela quantidade de termos, já que a redução da quantidade de termos leva à perda de informação.

| Testes      | Frequências |        |
|-------------|-------------|--------|
|             | Mínimo      | Máximo |
| A (Teste 3) | 50          | 100    |
| B (Teste 1) | 50          | 500    |
| C (Teste 5) | 50          | 1000   |
| D (Teste 2) | 100         | 300    |
| E (Teste 4) | 500         | 1000   |

Tabela A.8: Configuração das frequências dos Testes 1 a 5 organizados em ordem crescente da frequência mínima e renomeados como testes de A a E

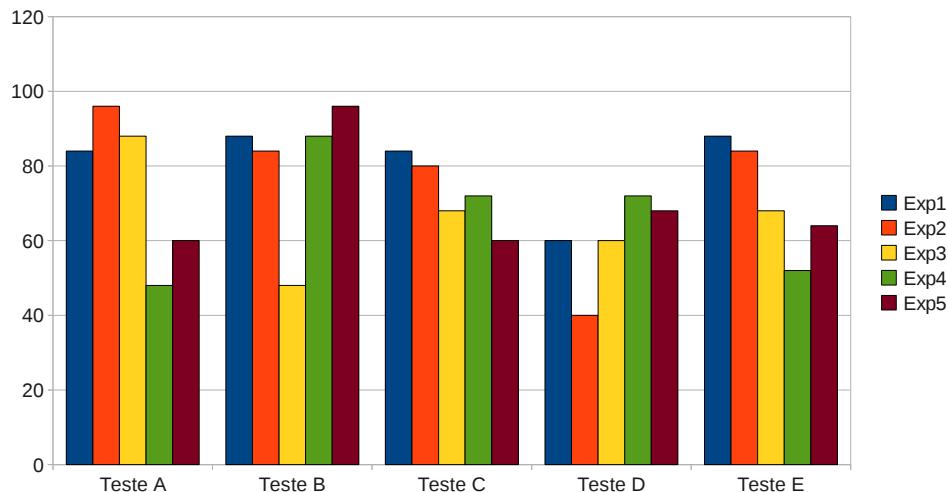


Figura A.5: Resultados obtidos pela mudança na frequência mínima

A partir destas análises, conclui-se que embora o método proposto tenha um bom desempenho, é necessário considerar o pré-processamento dos documentos porque este interfere diretamente nos resultados.

## Estratégia de *matching* para Recuperação Flexível de Documentos

---



---

Na estratégia de *matching* desenvolvida para considerar a flexibilidade de um Sistema de Recuperação de Informação (SRI) no nível da consulta, a relevância dos documentos com relação aos termos utilizados nas consultas é representada por meio de termos linguísticos. Por exemplo, um termo pode ser “muito importante” ou “pouco importante” para uma consulta. Esta representação se assemelha a indicação de importância dada pelos seres humanos.

Neste contexto, considere  $k$ ,  $n$  e  $c$  a quantidade de termos, documentos e classes, respectivamente. Seja o documento  $\mathbf{d}_i$ ,  $i = 1 \dots n$ , representado pelas frequências  $\sigma(t_j, \mathbf{d}_i)$  do termo  $t_j$  no documento  $\mathbf{d}_i$ ,  $1 \leq j \leq k$ ; um conjunto de regras fuzzy  $\{R_1, R_2, \dots, R_z\}$  geradas conforme a abordagem apresentada no Apêndice A; um subconjunto do conjunto de termos representativos da coleção de documentos, os quais são selecionados pelo sistema como palavras-chave a serem utilizadas na consulta do usuário  $\{\iota_1, \iota_2, \dots, \iota_b\}$ , com  $b \leq k$ ; um conjunto de classes dos documentos escolhidos como relevantes para o usuário  $\{\kappa_1, \kappa_2, \dots, \kappa_x\}$ , com  $x \leq c$ , e os graus de relevância linguísticos atribuídos pelo usuário a cada palavra-chave e classe dos documentos  $L = \{(u_1, u_2, \dots, u_\gamma)\}$ , com  $\gamma = b + x$ . Uma vez executadas as etapas de classificação de documento e encontrada a classe de saída  $\kappa_l$ ,  $1 \leq l \leq c$ , para o documento  $\mathbf{d}_i$  utilizando regras fuzzy conforme apresentado no Apêndice A, a estratégia de *matching* proposta para recuperar o documento  $\mathbf{d}_i$  é:

1. Verifique o grau de relevância da classe de documentos requisitada pelo usuário  $u_l$ ,  $1 \leq l \leq \gamma$  e compare-a com o grau de relevância obtido no processo de classificação  $C_{compat}(R_s, \mathbf{d}_i)$ ,  $1 \leq s \leq z$ , sendo  $R_s$  a regra pela qual obteve-se a classe  $\kappa_l$ . Se o grau de relevância corresponde ao mesmo grau determinado pelo usuário, isto é,  $C_{compat}(R_s, \mathbf{d}_i) = u_l(\kappa_l)$ , i.e., o grau de pertinência da classe  $\kappa_l$  no conjunto  $u_l$ ,

---

execute a próxima etapa. Caso contrário, o documento não é recuperado, pois nenhuma regra classifica o documento com o grau de relevância requisitado pelo usuário.

2. Encontre o grau de pertinência de cada palavra-chave escolhida pelo usuário como relevantes para a consulta. Considere  $u_l(\iota_l)$  o grau de pertinência da palavra-chave  $\iota_l$  escolhida pelo usuário com algum grau de relevância  $u_l$ .
3. Por fim, se  $\min(u_1(\iota_1), u_2(\iota_2), \dots, u_\gamma(\iota_\gamma)) \geq \Xi$ , no qual  $\Xi$  é um limiar, recupere o documento  $d_i$ .

Esta estratégia de *matching* e a utilização de regras fuzzy para classificação dos documentos são as atividades principais para a recuperação de documentos que correspondam às preferências do usuário. Entende-se por preferências os graus de relevância dados pelo usuário aos termos e classes dos documentos.

Para melhor compreender o funcionamento da estratégia proposta, observe o exemplo apresentado a seguir.

Para a execução da estratégia de *matching* proposta, considere um conjunto de regras geradas a partir da matriz documentos-grupos obtida do processo de agrupamento fuzzy, conforme apresentado na Tabela B.1. Observe que o antecedente das regras correspondem aos grupos dos documentos.

Tabela B.1: Base de regras geradas a partir da matriz documentos-grupos

| Regras geradas |   |
|----------------|---|
| $R_1$          | <b>SE</b> $G_1$ is <i>baixo</i> <b>E</b> $G_2$ is <i>medio</i> <b>E</b> $G_3$ is <i>medio</i> <b>ENTÃO</b> $\kappa_1$ |
| $R_2$          | <b>SE</b> $G_1$ is <i>alto</i> <b>E</b> $G_2$ is <i>baixo</i> <b>E</b> $G_3$ is <i>baixo</i> <b>ENTÃO</b> $\kappa_2$  |
| $R_3$          | <b>SE</b> $G_1$ is <i>medio</i> <b>E</b> $G_2$ is <i>baixo</i> <b>E</b> $G_3$ is <i>alto</i> <b>ENTÃO</b> $\kappa_3$  |
| $R_4$          | <b>SE</b> $G_1$ is <i>alto</i> <b>E</b> $G_2$ is <i>baixo</i> <b>E</b> $G_3$ is <i>medio</i> <b>ENTÃO</b> $\kappa_2$  |

Uma vez geradas as regras, as palavras-chave previamente selecionadas para serem utilizadas na consulta são consideradas como variáveis linguísticas fuzzy e granularizadas em termos linguísticos fuzzy: *relevante*, *pouco relevante* e *muito relevante*. Desta maneira, o usuário pode escolher a relevância linguística da palavra-chave utilizada na sua consulta. Portanto, o SRI flexível no nível da consulta deve apresentar ao usuário a opção de escolher não somente as palavras-chave mais importantes para sua consulta, mas também a opção de indicar qual a relevância de cada uma delas por meio de termos linguísticos.

Por fim, após a definição de relevância apresentada pelo usuário, pode-se aplicar a estratégia de *matching* proposta para recuperar os documentos que melhor satisfazem a requisição do usuário.

Para ilustrar este processo, considere os critérios de relevância do usuário como apresentados na Tabela B.2. As classes dos documentos são  $\kappa_1$ ,  $\kappa_2$  e  $\kappa_3$ . As palavras-chave escolhidas pelo usuário são  $\iota_1$  e  $\iota_3$ .

Uma vez definidos os critérios de relevância do usuário, o SRI acessa sua atividade de filtragem e faz a correspondência (*matching*) entre a requisição do usuário e um documento. Para isto, são considerados os graus de compatibilidade entre a classe relevante e a

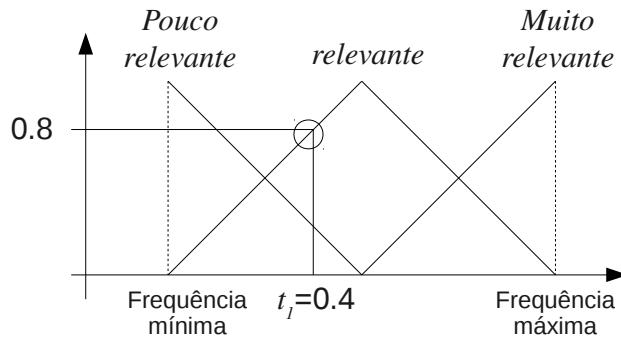
Tabela B.2: Critérios de relevância definidos pelo usuário

| Classes e palavras-chave | Graus de relevância definidos pelo usuário |
|--------------------------|--|
| $\kappa_1$               | <i>muito relevante</i>                     |
| $\kappa_2$               | <i>pouco relevante</i>                     |
| $\kappa_3$               | <i>irrelevante</i>                         |
| $\iota_1$                | <i>relevante</i>                           |
| $\iota_3$                | <i>pouco relevante</i>                     |

regra disparada no processo de classificação  $C_{compat}(R_s, \mathbf{d}_i)$  no qual  $1 \leq s \leq z$  e  $1 \leq i \leq n$ , com  $z$  igual à quantidade de regras e  $n$  igual à quantidade de documentos. Após checar a relevância da classe dos documentos, deve-se checar a relevância das palavras-chave.

Continuando com o exemplo, o documento  $\mathbf{d}_1$ , representado pelo vetor  $\mathbf{d}_1 = [0.3, 0.6, 0.1, a]$  na matriz documentos-grupos, é corretamente classificado pela regra  $R_1$ , a primeira regra da base de regras apresentada na Tabela B.1. Supondo que o grau de compatibilidade  $C_{compat}(R_1, \mathbf{d}_1)$  corresponde a relevância da classe requisitada pelo usuário, *muito relevante* para a classe  $\kappa_1$ , então a classe do documento corresponde a requisição do usuário. O documento  $\mathbf{d}_1$ , representado pelo vetor  $\mathbf{d}_1 = [0.4, 0.1, 0.3, 0.5, 0.1, 0.2, 0.7, 0.5, 0.6, a]$  na matriz na matriz documentos-termos e as palavras-chave escolhidas pelo usuário possuem frequências  $\iota_1 = 0.4$  e  $\iota_3 = 0.3$ .

Com a informação de frequência das palavras-chave escolhidas como relevantes para a consulta é possível encontrar o grau de pertinência de cada uma delas na função de pertinência que representa a relevância da palavra-chave requisitada pelo usuário. Esta função de pertinência considera as frequências mínima e máxima da palavra-chave na coleção para definição dos conjuntos fuzzy. Logo, como ilustrado na Figura B.1, a palavra-chave  $\iota_1$ , cuja frequência é 0.4, possui grau de pertinência 0.8 no conjunto fuzzy *relevante*, o qual foi o critério (relevância linguística) definido pelo usuário para a recuperação dos documentos. Do mesmo modo, a palavra-chave  $\iota_3$ , cuja frequência é 0.3, possui grau de pertinência 0.4 no conjunto fuzzy *pouco relevante* também definido pelo usuário.

Figura B.1: Função de pertinência da palavra-chave  $\iota_1$ 

Por fim, considerando o limiar  $\Xi = 0.5$ , o Sistema de Recuperação de Informação (SRI) não vai recuperar o documento  $\mathbf{d}_1$  para o usuário, pois de acordo com a etapa 4

---

da estratégia de *matching*  $\min(0.8, 0.4) = 0.4$  e portanto menor que o limiar. Observe que a palavra-chave  $\iota_2$  é desconsiderada porque a mesma é vista como irrelevante para o usuário.

A estratégia de *matching* apresentada<sup>1</sup> foi idealizada considerando que o uso de modelos linguísticos fuzzy em problemas de recuperação de informação é útil quando os valores de relevância não podem ser expressos por meio de valores numéricos. A abordagem linguística de um SRI apresenta uma maneira mais natural para o usuário fazer as suas requisições. Assim, como trabalho futuro, pretende-se realizar investigações a fim de reduzir a participação do especialista na seleção das palavras-chave a serem utilizadas na consulta, pois esta seleção limita o espaço de busca do SRI.

---

<sup>1</sup>A exploração sobre esta estratégia de *matching* foi realizada durante um estágio no exterior com o grupo de pesquisa da Universidade de Granada - Espanha. Este grupo de pesquisa vem trabalhando com computação flexível e sistemas de informação inteligentes (*Soft Computing and Intelligent Information Systems*)<sup>2</sup> destacando-se como um grupo reconhecido mundialmente na área de sistemas fuzzy. Durante o estágio foi realizada a atualização do levantamento bibliográfico e investigação de uma abordagem fuzzy para recuperação de informação.