UNIVERSIDADE FEDERAL DA BAHIA

TRABALHO DE GRADUAÇÃO

Uma abordagem híbrida para organização flexível de documentos

Nilton Vasques Carvalho Junior

Programa de Graduação em Ciência da Computação

Salvador
2 de junho de 2016

NILTON VASQUES CARVALHO JUNIOR

UMA ABORDAGEM HÍBRIDA PARA ORGANIZAÇÃO FLEXÍVEL DE DOCUMENTOS

Este Trabalho de Graduação foi apresentado ao Programa de Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Tatiane Nogueira Rios

Salvador
2 de junho de 2016

Ficha catalográfica.

Carvalho, Nilton Vasques Jr.

Uma abordagem híbrida para organização flexível de documentos/ Nilton Vasques Carvalho Junior– Salvador, 2 de junho de 2016.

67p.: il.

Orientadora: Profa. Dra. Tatiane Nogueira Rios.
Monografia (Graduação)– UNIVERSIDADE FEDERAL DA BAHIA, INSTITUTO DE MATEMÁTICA, 2 de junho de 2016.

"1. Fuzzy C Means. 2. Organização flexível de documents. 3. Lógica Fuzzy. 4. Mineração de dados.".
I. Rios, Tatiane Nogueira. II. UNIVERSIDADE FEDERAL DA BAHIA. INSTITUTO DE MATEMÁTICA. III Título.

NUMERO CDD

TERMO DE APROVAÇÃO

NILTON VASQUES CARVALHO JUNIOR

UMA ABORDAGEM HÍBRIDA PARA ORGANIZAÇÃO FLEXÍVEL DE DOCUMENTOS

Este Trabalho de Graduação foi julgado adequado à obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo Programa de Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 02 de junho de 2016

Profa. Dra. Tatiane Nogueira Rios
Universidade Federal da Bahia

Prof. Dr. Maurício Pamplona Segundo
Universidade Federal da Bahia

Prof. Dr. Ricardo Rios
Universidade Federal da Bahia

Coloque sua DEDICATÓRIA AQUI.

AGRADECIMENTOS

Coloque seus AGRADECIMENTOS AQUI.

O que sabemos é uma gota, o que ignoramos é um oceano.

—ISAAC NEWTON (1687)

RESUMO

O presente trabalho de conclusão de curso intenta realizar, à luz da literatura da área de mineração de textos e campos do saber correlatos, uma investigação dos impactos dos algoritmos de agrupamento na composição da organização flexível de documentos textuais. No seu princípio desta monografia, discute-se a importância e as escolhas utilizadas na etapa pré-processamento, anterior à aplicação do agrupamento, além dos critérios de validação do agrupamento que são utilizados no etapa do pós-processamento, a saber, implementado pela medida da silheuta fuzzy. Para contextualizar a prática da atividade de organização textual flexível proposta, destacam-se ao longo dos capítulos os desafios inerentes à organização de textos, a exemplo o problema dos altos custos computacionais de busca de graus de similaridade semântica em matrizes esparsas do tipo atributo valor, assim como os possíveis mecanismos para mitigar ou reduzir os efeitos negativos dessas dificuldades no processo de atribuição de significado aos grupos produzidos pelo agrupamento, através da extração de termos descritores relevantes. A estratégia defendida aqui nesta monografia trata-se de uma abordagem flexível e híbrida de organização de documentos, mesclando os benefícios concedidos pela adequada interpretação de partições fuzzy e possibilísticas existentes no método de agrupamento Possibilistic C-Means e Possibilistic Fuzzy C-Means (PFCM). Como fruto das análises experimentais aqui desenvolvidas, foram propostos os métodos de extração de descritores Possibilistic Description Comes Last (PDCL) e Mixed - Possibilistic Fuzzy Description Comes Last (PFDCL). Ambos mostraram-se relevantes através de evidências experimentais e análises subjetivas à adequação dos métodos, para a organização flexível de documentos, contribuindo com descobertas originais para o estado da arte da área. Os resultados da pesquisa ainda estimulam novas implementações cuja execução pode se transcorrer em trabalhos futuros.

Palavras-chave: agrupamento fuzzy, agrupamento possibilístico, organização flexível de documentos, mineração de textos

ABSTRACT

The present coursework aims to conduct, under the knowledge of Text Mining and related areas, an investigation about the impacts of clustering algorithms in the development of the flexible organization of textual documents. In the beginning of this coursework, it is discussed the relevance and the choices on preprocessing stage, that is prior to clustering proper, besides, the validation criteria of clustering during the post processing stage, namely, fuzzy silhouette. To contextualize the piratical of the proposed flexible textual organization, it is remarked along chapters the intrinsic challenges of text organization, for example, the higher computational costs problem in seeking semantic similarities degrees on sparse matrices value-attribute matrices, as well as possible ways to mitigate or reduce the negative effects of these difficult on the process meaning attribution groups that are outcome from clustering though relevant descriptors extraction The strategy of the present course work consists in an approach of flexible and hybrid document organization, mixing the benefits of the known fuzzy partition interpretation and possibilistic one on clustering method of Possibilistic C-Means e Possibilistic Fuzzy C-Means (PFCM). As an output of here developed experiments, it were proposed the Possibilistic Description Comes Last (PDCL) and Mixed - Possibilistic Fuzzy Description Comes Last (PFDCL) descriptors' extraction methods. Both were confirmed by experimental evidences and analysis that are subjective to methods adjusts for the flexible organization of documents collaborating for original discoveries on the state of arts. The results of the present research, after all, stimulate new implementations whose execution may be executed on future works.

Keywords: fuzzy clustering, possibilistic clustering, flexible organization, documents, text mining

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

Uma breve introdução sobre do que se trata esta monografia e a maneira como o texto está organizado.

INTRODUÇÃO

O avanço da computação pessoal, em particular a computação móvel, tem proporcionado um gigantesco aumento da quantidade de dados armazenados pela humanidade ao longo dos anos. A critério de exemplo, a popular plataforma de rede social Facebook[1], produz diariamente mais de 25 *terabytes* de informação (Havens et al., 2012). De acordo com Huang et al. (2015), a tendência com o avanço das tecnologias, é que tudo seja integrado a internet, de tal modo que os pesquisadores já chegam a dizer que os dados são o novo recurso natural do planeta. Ainda segundo os autores, entre as maiores fontes de geração de dados estão os sistemas governamentais, plataformas de mídias sociais, assim como arquivos armazenados pelas corporações, como por exemplo, formulários médicos, opiniões de consumidores, relatórios e etc.. Entretanto, Muggleton (2006) ressalta que todos esses dados excedem os limites humanos para o uso e compreensão destes.

Diante desse cenário, instituições públicas e privadas estão sobrecarregadas com a tarefa de processar essa imensa quantidade de informação em bases de dados com documentos não estruturados e em diversos formatos (Kobayashi e Aono, 2008). Estes documentos usualmente são de diversos tipos, como por exemplo, textos, áudios, imagens, vídeos, documentos HTML, podendo estar, inclusive, em diferentes idiomas.

Nesse contexto, diversas pesquisas tem objetivado a proposição ou refinamento de técnicas para automatização do processo de análise e aquisição de conhecimento útil desse montante de informações armazenadas. Porém, devido a multi disciplinaridade inerente desse campo de estudo, o mesmo tem sido estudado pelas comunidades de mineração de dados, aprendizado de máquina e recuperação de informação.

A Mineração de Dados (MD) é um campo de estudo que vem obtendo rápidos avanços nos últimos anos, e segundo Aggarwal e Zhai (2012), isto se deve aos avanços das tecnologias de *hardware* e *software*, o qual possibilitou o massivo armazenamento de diferentes tipos de dados, inclusive os dados textuais. Portanto, como resultado desse aumento na quantidade de documentos disponíveis na forma textual, existe uma demanda crescente

[1] ⟨https://facebook.com/⟩

no desenvolvimento e aprimoramento de métodos e algoritmos que possam efetivamente processar e extrair padrões dos dados de maneira dinâmica e escalável.

Por outro lado, enquanto os dados estruturados já possuem mecanismos bem eficientes de armazenamento e recuperação, os dados textuais são geralmente gerenciados através de mecanismos de buscas para suprir essa falta de estruturação. Esses mecanismos de busca possibilitam aos usuários uma conveniente maneira para recuperar informações em coleções textuais através de consultas com palavras chaves. Desse modo, compete ao campo de estudo da Recuperação de Informação (RI) a tarefa de explorar, investigar e propor métodos para otimização da eficiência e efetividade de sistemas de buscas (Baeza-Yates e Ribeiro-Neto, 2011).

Mas segundo Baeza-Yates e Ribeiro-Neto (2011), as pesquisas de recuperação de informação tem focado tradicionalmente, em formas de facilitar o acesso à informação, do que realizar a descoberta de novos padrões em documentos textuais, o qual se destaca como sendo o objetivo principal da mineração de textos. A mineração de textos, por sua vez, é uma especialização da mineração de dados, que busca incorporar atividades de estruturação dos documentos em formatos apropriados, facilitando a aplicação dos tradicionais métodos de extração de padrões da MD, minimizando as perdas durante a conversão do formato original não estruturado (Nogueira, 2013).

Contudo, uma série de características diferenciam os documentos textuais de outras formas de dados. O que por sua vez, afeta o desempenho das clássicas técnicas da MD. Dentre essas características peculiares, destacam-se como mais importantes os fatos de que os dados são esparsos e possuem alta dimensionalidade. Por exemplo, uma coleção de documentos (corpus) pode conter 100.000 palavras (termos), enquanto um único documento desse corpus pode conter somente algumas centenas de palavras (Aggarwal e Zhai, 2012). Essa discrepância, tem implicações diretas em várias técnicas de identificação de padrões, e especialmente no agrupamento textual, que deriva de clássicas técnicas de agrupamento da mineração de dados, aplicados à conjuntos de baixa dimensionalidade.

Para cumprir a tarefa de extrair informações relevantes de documentos textuais e identificar as estruturas inerentes aos mesmos. A mineração de textos emprega uma variedade de técnicas, as quais se destacam aquelas usualmente desenvolvidas para efetuar tarefas de coleta, pré-processamento, agrupamento textual e seleção de termos descritores para o agrupamento.

O agrupamento pode, de maneira geral, ser definido como a tarefa de agrupar uma coleção de objetos, de acordo algum critério de similaridade. É possível distinguir os tipos de agrupamento em função da lógica empregada por eles. Com isso, tem-se os algoritmos que derivam da lógica clássica ou da lógica fuzzy. Na lógica clássica, após a conclusão do agrupamento, cada elemento só pertence à apenas um grupo, enquanto que na lógica fuzzy, a pertinência do elemento será distribuída entre os grupos.

Ao se analisar a diversidade de conteúdo em dados textuais, é trivial notar que frequentemente um texto aborda um ou mais temas. O que implica que o agrupamento clássico, ao atribuir um objeto a apenas um grupo, não irá representar bem a imprecisão e incerteza natural dos documentos.

Deste modo, os métodos de agrupamento derivados da lógica fuzzy se mostram mais capacitados para lidar com essa imprecisão e incerteza da realidade multi temática dos

documentos textuais. Assim sendo, uma organização flexível de documentos pode ser definida como o processo que compreende a estruturação dos dados, a adição de flexibilidade proporcionada pelo agrupamento fuzzy, a extração de descritores dos grupos de maneira flexível e a recuperação de informação através de um Sistema de Recuperação de Informação (SRI).

Ao se observar o processo de organização flexível de documentos, percebe-se que o mesmo abrange várias etapas, cada uma delas com suas particularidades. No entanto, apesar da importância desempenhada por cada etapa do processo, o agrupamento em si pode ser visto como uma das peças chaves, pois ele é diretamente responsável por organizar os documentos de acordo com as suas similaridades. Adicionalmente, é preciso desconsiderar ou reduzir a influência de documentos ruidosos, que destoam do restante da coleção nos grupos finais.

Os algoritmos Fuzzy C-Means (FCM) (Bezdek et al., 1984), que deriva do clássico K-Means (MacQueen et al., 1967), e o Possibilistic C-Means (PCM) (Krishnapuram e Keller, 1993), são exemplos de métodos de agrupamento capazes de organizar de maneira automatizada uma coleção de documentos em um conjunto de grupos (Mei et al., 2014; Tjhi e Chen, 2009; Boughanem et al., 2008; Saracoglu et al., 2008). Ambos distribuem os documentos de uma coleção textual em um conjunto de grupos, de modo que cada documento possa pertencer a diferentes grupos com diferentes graus de pertinência, considerando assim a flexibilidade necessária para tratar a imprecisão e incerteza do processo.

No entanto, o FCM apresenta alguns resultados indesejados, diante da presença de dados ruidosos na coleção. Em se tratando de coleções textuais, um dado ruidoso pode ser considerado como um documento que possua uma temática bastante diferente dos demais documentos da coleção. Com o objetivo de atribuir valores de pertinências mais realísticos aos elementos a serem agrupados e penalizar com baixas pertinências os elementos ruidosos, o algoritmo PCM foi proposto. Porém, o PCM é muito sensível à inicialização, o que pode resultar em grupos coincidentes, onde não há uma separação muito bem definida dos elementos.

Visando contemplar os benefícios de ambos os métodos, foi proposto o método de agrupamento Possibilistic Fuzzy C-Means (PFCM) (Pal et al., 2005), como uma versão híbrida dos algoritmos FCM e o PCM, objetivando adicionar robustez à tarefa de agrupamento.

Após o agrupamento dos documentos, é necessário realizar a extração dos termos que melhor descrevem os grupos. Para realização dessa tarefa, tem-se alguns métodos na literatura do tipo DCF (*Description Comes First*), que realizam a extração de maneira embutida no processo de agrupamento. Porém, essa abordagem torna o processo de extração de descritores dependente do algoritmo de agrupamento. Com o propósito de contornar esse cenário, foi proposto em Nogueira (2013) o método *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) (Nogueira, 2013), o qual extrai os termos descritores após a etapa de agrupamento de maneira independente do algoritmo de agrupamento utilizado. Permitindo avaliar diretamente os impactos dos métodos de agrupamento, na extração de descritores e por consequência na qualidade da organização flexível de documentos.

Entretanto, o método SoftO-FDCL foi pensado inicialmente para interpretar as per-

tinências produzidas na partição do FCM, que difere da partição resultante produzida pelo PCM. A principal contribuição do PCM foi uma alteração no modo de atribuição da pertinência de uma elemento a um grupo, o que impacta diretamente na partição dos grupos resultantes.

Diante deste contexto, e tendo em vista o crescente aumento de informações produzidas além da capacidade humana de analisar. Com a demanda crescente no desenvolvimento e aprimoramento das técnicas de extração e identificação de conhecimento útil em dados textuais, bem como a necessidade de se organizar esses dados de maneira flexível, tratando a imprecisão e incerteza natural desses dados e considerando as particularidades existentes nos métodos de agrupamento, foi formulada a seguinte hipótese para o desenvolvimento desse trabalho:

> *A utilização de uma estratégia híbrida de agrupamento e extração de descritores, entre os graus de pertinência e tipicidade providos pelo método de agrupamento PFCM, permitem o aumento da robustez e resiliência contra ruídos na organização flexível de documentos, aumentando assim a relevância dos grupos obtidos.*

Para demonstrar a validade da hipótese formulada, com base na exploração de estratégias existentes na literatura para o aprimoramento do processo de organização flexível de documentos, definiu-se o seguinte objetivo:

> *Conduzir uma investigação em torno dos métodos de agrupamento FCM, PCM e PFCM, para compreender e interpretar corretamente as peculiaridades de se extrair descritores a partir de um agrupamento híbrido.*

A fim de atender a esse objetivo, foram realizadas as seguinte tarefas ao longo do desenvolvimento desta monografia: estudo dos fundamentos teóricos necessários para a organização flexível de documentos, revisão das estratégias recentes utilizadas por pesquisadores para aprimorar a organização flexível de documentos, condução de diversos experimentos para analisar os impactos da aplicação do algoritmo PFCM no processo de agrupamento e na extração de descritores.

Considerando-se os resultados dos experimentos realizados, foi descoberto que as alterações existentes no PCM, impactam diretamente na qualidade dos descritores extraídos pelo método SoftO-FDCL. Essa descoberta motivou a proposição de dois novos métodos de extração de descritores: *Possibilistic Descriptor Comes Last* (PDCL) e *Mixed - Possibilistic Fuzzy Descriptor Comes Last* (Mixed-PFDCL). Os quais apresentaram resultados que contribuem de maneira significativa para o estado da arte da extração de descritores dos grupos fuzzy e para o aprimoramento da organização flexível de documentos.

Detalhes de cada tarefa realizada, bem como a comprovação da hipótese levantada, são apresentados ao longo deste trabalho como segue:

Capítulo 2: Neste capítulo, os fundamentos teóricos necessários para compreender o processo de organização flexível de documentos são apresentados, os quais contemplam a descrição da etapa de pré-processamento e estruturação dos dados; a definição dos

principais algoritmos de agrupamento, capazes de proporcionar flexibilidade ao processo de organização de documentos; e a descrição da extração de descritores.

Capítulo 3: Este capítulo aborda uma breve revisão do estado da arte encontrado na literatura, referente às diversas estratégias propostas pelos pesquisadores. com o objetivo de aprimorar todas as etapas da organização flexível de documentos.

Capítulo 4: Neste capítulo, um estudo dos impactos da utilização do método de agrupamento híbrido na organização flexível de documentos é apresentado por meio da realização de análises experimentais. Neste capítulo, as influências das tipicidades presentes na partição de pertinências do PCM também são apresentadas. Seguidas pela proposta dos métodos PDCL e Mixed-PFDCL.

Capítulo 5: Por fim, este capítulo contempla as conclusões obtidas de todo o estudo realizado nesta monografia, assim como discussões a respeito dos resultados obtidos nos experimentos. Aqui também está apontada uma série de possibilidades de extensões que derivam desta pesquisa.

Capítulo

2

Este capítulo tem como objetivo fundamentar as bases necessárias dos campos de estudos utilizados nesta monografia.

FUNDAMENTAÇÃO TEÓRICA

2.1 CONSIDERAÇÕES INICIAIS

A lógica fuzzy foi introduzida por Lofti Zadeh em 1965, onde o autor inicia a discussão definindo os conjuntos fuzzy, as quais constituem classes de objetos com valores contínuos de pertinência. Cada conjunto é caracterizado por uma função de pertinência, a qual atribui a cada objeto do conjunto um grau de pertinência que varia entre zero e um. As operações matemáticas da teoria dos conjuntos, como inclusão, união, intersecção, complemento, relação, etc., também são estendidas aos conjuntos fuzzy, assim como várias propriedades dessas notações são definidas.

Uma das motivações para o uso da lógica fuzzy, vem da maneira como nosso cérebro classifica e rotula o mundo real. Por exemplo, ao rotularmos uma pessoa como alta, estamos atribuindo ela ao grupo de pessoas altas. Assim como quando nos expressamos sobre o quanto um determinado dia está fazendo calor ou frio. O conjunto de pessoas altas ou dias frios, não se enquadra na sua totalidade na lógica clássica, pois essa forma imprecisa de descrever o mundo a nossa volta desempenha um papel fundamental na forma de pensar humana, assim como também nas áreas de reconhecimento de padrões, comunicação e abstração (Zadeh, 1965). Neste sentido, esta seção tem como propósito contextualizar os principais aspectos da lógica fuzzy que a torna tão importante no contexto da organização flexível de documentos. Ressalta-se, portanto, que definições mais aprofundadas sobre a lógica fuzzy fogem do escopo desse texto. Adicionalmente, além dos principais fundamentos necessários para compreensão da abordagem apresentada neste documento, estão a atividade de pré-processamento dos dados, cuja finalidade é filtrar e estruturar os dados para serem processados nas etapas seguintes; os principais algoritmos de agrupamento fuzzy, presentes na literatura, com as suas definições matemáticas e pseudo códigos; e, por fim, a tarefa de rotular os grupos encontrados na etapa de agrupamento com os termos que melhor os representem, permitindo assim a realização de consultas em Sistemas de Recuperação da Informação (SRI).

2.2 CONJUNTOS E LÓGICA FUZZY

2.2.1 Definição de conjuntos fuzzy

Seja X um espaço de objetos, com um elemento genérico x. Sendo $X = \{x\}$. Um conjunto fuzzy A em X é caracterizado por uma função de pertinência $f_A(x)$, a qual associa a cada elemento de X um número real presente no intervalo de $[0, 1]$, sendo o valor de $f_A(x)$ a representação do grau de pertinência de x em A.

2.2.2 Lógica fuzzy

A lógica fuzzy é uma lógica multi valorada, onde os valores das variáveis pertencem ao intervalo de $[0, 1]$, enquanto na lógica clássica os valores verdades só possuem os estados 0 ou 1 (também conhecido como valores *crisp*). Uma da mais importantes aplicações está no tratamento de precisão e incerteza. O que nos permite modelar soluções mais adequadas para ambientes imprecisos e incertos. Antes da lógica fuzzy ser introduzida por Zadeh (1965), em 1930 Lukasiewics (Chen, 2000) desenvolveu a lógica n-valorada para $3 < n < \infty$, utilizando apenas os operadores lógicos de negação $-$ e implicação $\Rightarrow$. Dado então um inteiro positivo, $n > 3$, a lógica n-valorada assume valores verdade pertencente ao intervalo $[0, 1]$, definidos pela seguinte partição igualmente espaçada:

$$0 = \frac{0}{n-1}, \frac{1}{n-1}, \frac{2}{n-1}, ..., \frac{n-2}{n-1}, \frac{n-1}{n-1} = 1$$

Para estender a lógica n-valorada para uma lógica com infinitos valores $2 \leq n \leq \infty$, Zadeh (1965) modificou a lógica de Lukasiewics definindo os seguintes operadores lógicos:

$$\bar{a} = 1 - a$$
$$a \wedge b = min\{a, b\}$$
$$a \vee b = max\{a, b\}$$
$$a \Rightarrow b = min\{1, 1 + b - a\}$$
$$a \Leftrightarrow b = 1 - |a - b|$$

O objetivo da lógica fuzzy é prover mecanismos para tratar imprecisão e incerteza, se baseando na teoria de conjuntos fuzzy e usando proposições imprecisas, de modo similar a lógica clássica usando proposições precisas baseadas na teoria dos conjuntos. Para entender essa noção, observe a seguir um mesmo exemplo pela ótica do raciocínio clássico e pela ótica das ferramentas usadas para descrever imprecisão da lógica fuzzy.

a) Todo texto com 100 palavras ou mais da área jurídica, tem como assunto o direito.

b) O texto A com título “as manifestações de junho”, tem 100 palavras da área jurídica.

c) O texto B com título “política nas universidades”, tem 99 palavras da área jurídica.

d) O texto A tem como assunto o direito e o texto B não tem como assunto o direito.

Essa série de proposições ilustra o raciocínio empregado na lógica clássica e, seguindo as regras de inferência, conseguimos verificar que as sentenças estão corretas. No entanto, é fácil notar que a sentença d) não expressa muito bem o nosso entendimento sobre a temática dos textos. Seria comum alguém substituir a sentença d), por: O texto B fala um pouco sobre direito. Sendo assim é possível adicionar a imprecisão comum no mundo real às sentenças anteriores, conforme reescrito a seguir.

a) Todo texto que tem entre 50 e 100 palavras da área jurídica fala um pouco sobre direito. Enquanto todo texto que contenha 100 ou mais palavras da área jurídica fala bastante sobre direito.

b) O texto A com título "as manifestações de junho", tem 100 palavras da área jurídica.

c) O texto B com título "política nas universidades", tem 99 palavras da área jurídica.

d) O texto A fala bastante sobre direito, enquanto o texto B fala um pouco sobre direito.

Esse tipo de dedução comumente utilizada no nosso dia a dia, não tem como ser tratada pela lógica clássica. No entanto, podemos lidar com esse tipo de inferência imprecisa, empregando a lógica fuzzy, a qual permite o uso de alguns termos linguísticos imprecisos como:

- Predicados fuzzy: antigo, raro, caro, alto, rápido
- Quantificadores fuzzy: muito, pouco, quase, alguns
- Graus de verdade fuzzy: totalmente verdadeiro, verdadeiro, parcialmente falso, falso, definitivamente falso

2.3 PRÉ-PROCESSAMENTO

Pré-processamento dos dados é o processo de limpeza e preparação dos dados para extração de padrões. Para este trabalho, especificamente, considera-se dado como sendo um documento textual e a tarefa de extração de padrões a ser considerada é o agrupamento.

Essa etapa é importante porque algumas palavras em um documento podem causar pouco ou nenhum impacto no significado geral do documento (Haddi et al., 2013). Soma-se a isso o enorme custo computacional do processo de mineração de textos, devido à grande quantidade de atributos presente em dados textuais, visto que quanto maior for a coleção de textos, maior será a quantidade de palavras distintas. Tal dimensionalidade eleva bastante o custo computacional de qualquer tarefa de extração de padrões. Por isso, vários pesquisadores propuseram métodos para tentar simplificar, sintetizar e eliminar redundâncias desnecessárias nas coleções de textos.

A fase de pré-processamento voltada para a mineração de textos, as quais visam preparar dados estruturados para as clássicas operações de mineração de dados, requer técnicas muito específicas para o preparo dos dados não estruturados (Feldman e Sanger, 2007).

Segundo Feldman e Sanger (2007), é possível categorizar de maneira clara as técnicas de pré-processamento de textos em duas categorias, de acordo com as tarefas realizadas pela técnica e através dos algoritmos e frameworks que a mesma utiliza. Por sua vez, as técnicas categorizadas pelas suas tarefas, geralmente visam realizar a estruturação do documento através de tarefas e sub tarefas. Como por exemplo, realizar a extração de título e sub título de documentos no formato PDF. No entanto, as demais técnicas de pré-processamento são derivadas de métodos formais, e incluem esquemas de classificação, modelos probabilísticos e sistemas baseado em regras.

O processo de pré-processamento de dados textuais, inicia com um documento parcialmente estruturado e avança incrementando a estrutura através do refinamento das características do documento e adicionando novas (Feldman e Sanger, 2007). No contexto da mineração de textos, as características dos documentos são as suas palavras(Haddi et al., 2013). Ao final do processo, as palavras mais relevantes são utilizadas, e as demais são descartadas.

O processo como um todo envolve várias etapas, as quais pode-se elencar a remoção de espaços, expansão de abreviações, remoção de *stopwords*, que são palavras que não possuem relevância no significado geral do texto e geralmente são compostas por proposições, pronomes, artigos, interjeições dentre outras(Nogueira, 2013). Assim como também o processo de *stemming* ou lematização, onde se busca encontrar o radical da palavra, visando assim remover palavras que possuam significados similares. Ainda é possível usar as técnicas de NLP ($NaturalLanguageProcessing$) para eliminar sinônimos. Por fim, é realizada a seleção de termos mais característicos para toda a coleção (Haddi et al., 2013).

Diversos métodos foram propostos visando capturar a importância dos termos em coleções textuais. Sendo o método *Term Frequency Inverse Document Frequency* (TF-IDF) um dos mais importantes e frequentemente utilizado na literatura (Haddi et al., 2013). A definição da TF-IDF está na Equação (2.1), onde n é o número de documentos da coleção, DF o total de documentos que possuem este termo e FF (*frequency feature*) a frequência do termo no documento.

$$\varphi(t,d) = FF * log(\frac{N}{DF}) \tag{2.1}$$

Como resultado final de todo o processo de pré-processamento, obtém-se a matrix D. Onde D representa os n documentos da coleção, sendo cada documento d_i, com $1 \leq i \leq n$, uma linha da matriz D, definido como sendo $d_i = [\varphi(t_1, d_i), \varphi(t_2, d_i), \varphi(t_3, d_i), ..., \varphi(t_k, d_i)]$, onde t_j é um termo presente na coleção, com $1 \leq j \leq k$.

2.4 AGRUPAMENTO FUZZY

O agrupamento é um processo não supervisionado cujo o objetivo é organizar os objetos similares no mesmo grupo e os objetos com grau de dissimilaridade elevado em grupos distintos (Nogueira, 2013). Este processo é de grande utilidade para diversos campos de estudo da inteligência computacional, como a mineração de dados, recuperação de informação, segmentação de imagens e classificação de padrões. Neste trabalho, os objetos

a serem agrupados são os documentos textuais.

O problema de organizar os documentos de maneira a maximizar a similaridade entre os membros de um mesmo grupo, e minimizar a similaridade entre documentos de grupos distintos, é essencialmente um problema de otimização. Sendo assim, pretende-se otimizar a escolha dos grupos, entre todas as possibilidades de agrupamento, dada uma função objetivo que captura a qualidade dos grupos. Esta função é responsável por atribuir ao conjunto de possíveis grupos um número real, de maneira que quanto melhor for os grupos, maior será o seu valor(Feldman e Sanger, 2007).

A medida de similaridade desempenha um papel fundamental no agrupamento, uma vez que ela precisa expressar o quão distante está um elemento do outro na coleção. Assim sendo, para obtermos bons resultados durante o processo de agrupamento é de grande importância a escolha adequada da medida de similaridade, e esta escolha precisa ser feita de acordo com o tipo dos dados. Na literatura, a medida de similaridade mais popular é a distância euclidiana (Equação 2.2), que tem se mostrado bastante adequada em dados com baixa dimensionalidade.

$$D(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2} \tag{2.2}$$

No entanto, em coleções textuais a matriz documentos x termos é naturalmente esparsa, devido a grande variedade de termos em uma coleção, o que faz com que um determinado documento d_i não contenha diversos termos presentes em um outro documento d_j. Assim, o vetor de características de cada documento acaba sendo preenchido com vários zeros, reduzindo a eficácia da distância euclidiana (Equação 2.2) (Nogueira, 2013).

A medida de similaridade mais comum para coleções textuais é o coeficiente de similaridade de cosseno (Feldman e Sanger, 2007). Por sua vez o coeficiente de similaridade de cosseno, desconsidera os diversos zeros presentes nos vetores de termos dos documentos, levando em conta apenas o ângulo formado entre eles(Nogueira, 2013). Na Equação (2.3) temos a definição do coeficiente de similaridade de cosseno, onde d_1 e d_2, são dois documentos quaisquer da coleção de documentos, e $1 \leq t \leq k$, onde k é a quantidade total de termos da coleção, e d_{ik} a frequência do termo t no documento d_i.

$$scos(d_1, d_2) = cos\theta = \frac{d_1 \cdot d_2}{|d_1||d_2|} = \sum_{t=1}^{k} \varphi(d_{1t}, d_1) \cdot \varphi(d_{2t}, d_2) \in [0, 1] \tag{2.3}$$

Os grupos resultantes de um processo de agrupamento, podem possuir algumas características que estão diretamente relacionadas com o método de agrupamento empregado. Estes podem ser *hard* ou *crisp*, caso o método de agrupamento seja baseado na lógica clássica, assim como podem ser $soft$, caso o método seja baseado na lógica fuzzy. No agrupamento *hard*, cada documento d_i só poderá pertencer a um único grupo g_j (Bezdek et al., 1984). Enquanto em grupos $soft$, cada documento d_i pode pertencer a um ou mais grupos g, com grau de pertinência variados. Além destes, os grupos ainda podem ser $flat$ ou hierárquicos, onde no agrupamento $flat$ todos os grupos estão no mesmo nível,

enquanto no modelo hierárquico os grupos podem estar dispostos em uma hierarquia, de modo que uma relação de parentesco é definida entre eles.

Para o agrupamento *hard*, considere $G = \{g_1, g_2, g_3, ..., g_c\}$ os grupos resultantes do agrupamento, sendo c o total de grupos, e a pertinência de cada documento d_i à um grupo g_l, $1 \leq l \leq c$, pode ser representada pela função característica $\kappa(d_i, g_j) \in \{0, 1\}$. Um dos mais populares algoritmos a implementar essa abordagem é o K-Means. Nos trabalhos de Bezdek et al. (1984), Nogueira (2013), Feldman e Sanger (2007), é apontada uma falha inerente dessa abordagem, pois quando um documento só pode pertencer a um único grupo, fica evidenciado que o mesmo não compartilha nenhuma similaridade com os documentos dos demais grupos, o que não expressa a imprecisão intrínseca da sobreposição dos assuntos em documentos de texto.

Com o objetivo de tratar essa falha da abordagem *hard* e adicionar o tratamento de imprecisão e incerteza no agrupamento, Bezdek et al. (1984) utilizou o modelo de partições fuzzy definido em Zadeh (1965), para permitir pertinências parciais de um elemento a um grupo, propondo assim o algoritmo Fuzzy C Means (FCM). Sendo assim, a função característica passa a ser considerada como uma função de pertinência de um documento d_i em um grupo g_l, a qual é definida por $\mu(d_i, g_l) \in [0, 1]$, tal que $\sum_{l=1}^{c} \mu(d_i, g_j) = 1$.

Um desafio sempre presente em métodos de agrupamento é a descoberta do número ideal de grupos em uma coleção. O método de organização flexível proposto em Nogueira (2013), utiliza a *Fuzzy Silhouette* (FS) para realizar a validação do agrupamento fuzzy, e por conseguinte encontrar o número de grupos ideal. A função FS é uma adaptação do método de critério de largura média (*Average Silhuette Width Criterion* - ASWC), desenvolvido para o agrupamento *crisp*. A definição da silhueta fuzzy está nas Equações (2.4) e (2.5), onde $\alpha(d_i, g_l)$ é a distância média entre o documento d_i e todos os documentos presentes no grupo g_l, enquanto $\beta(d_i, g_l) = min\{\alpha(d_i, g_h) | 1 \leq h \leq c; h \neq l\}$, é a medida de dissimilaridade de d_i ao grupo vizinho mais próximo de g_l, tal que c é a quantidade de grupos.

$$S(d_i) = \frac{\beta(d_i, g_l) - \alpha(d_i, g_l)}{max\{\alpha(d_i, g_l), \beta(d_i, g_l)\}} \tag{2.4}$$

$$FS = \frac{\sum_{i=1}^{n} (\mu_1(d_i) - \mu_2(d_i)) S(d_i)}{\sum_{i=1}^{n} (\mu_1(d_i) - \mu_2(d_i))} \tag{2.5}$$

Na Equação (2.5), $\mu_1(d_i)$ é a maior pertinência do documento d_i em um grupo, enquanto $\mu_2(d_i)$ é a segunda maior. Quanto maior então for o valor da função FS, melhor será o agrupamento. Deste modo, para encontrar o número de grupos ideal, basta executar a função FS variando o número de grupos, e selecionar o agrupamento que tiver o valor máximo de FS.

Toda investigação realizada neste trabalho tomou como base os métodos de agrupamento que derivam do algoritmo FCM(Bezdek et al., 1984), descrito a seguir, para se beneficiar da capacidade de tratar imprecisão e incerteza da lógica fuzzy, e por conseguinte permitir que um mesmo documento seja categorizado em mais de um tópico (grupo), refletindo a realidade dos documentos textuais.

Em todos os experimentos a medida de similaridade de cosseno (Equação 2.3) foi utilizada, e a quantidade de grupos ideal foi escolhida utilizando o método da silhueta fuzzy (Equação 2.5).

2.4.1 Algoritmo Fuzzy C Means (FCM)

Bezdek et al. (1984) descreve um método de agrupamento fuzzy que produz como saída partições fuzzy e protótipos dos grupos. Esse algoritmo desempenha um papel importante no contexto do agrupamento fuzzy, devido ao seu pioneirismo no campo de estudo, possuindo diversas extensões, sendo considerado um dos mais amplamente utilizados métodos de agrupamento fuzzy da literatura(Pal et al., 2005). A maioria dos métodos de agrupamento fuzzy são derivações do FCM(Krishnapuram e Keller, 1993).

Para compreensão do algoritmo, considere $V = \{v_1, v_2, v_3, ..., v_c\}$ os protótipos dos grupos $G = \{g_1, g_2, g_3, ...g_c\}$ definidos por

$$V = \left\{ v_j | v_j = \frac{\sum_{i=1}^{n} [\mu(d_i, g_j)]^m d_i}{\sum_{i=1}^{n} [\mu(d_i, g_j)]^m}, 1 < j \leq c \right\}, \tag{2.6}$$

tal que v_j seja o protótipo de g_j, c o número de grupos gerados no agrupamento e n o número de documentos presentes na coleção.

Seja $U_{c \times n}$ uma partição fuzzy conforme a Equação:

$$U_{c \times n} = \{\mu(d_i, g_k) | \mu(d_i, g_k) \in [0, 1], 1 < i \leq n, 1 < k \leq c\} \tag{2.7}$$

Esta partição apresenta todas as pertinências dos documentos aos grupos, cuja função de pertinência $\mu(d_i, g_k)$ de cada documento na coleção, é definida por

$$\mu(d_i, g_k) = \frac{1}{\sum_{j=1}^{n} (\frac{dist(d_i, v_k)}{dist(d_i, v_j)})^{\frac{1}{m-1}}} \tag{2.8}$$

e está sujeita as restrições definidas nas equações (2.9) e (2.10).

Considera-se m o fator de fuzificação, que regula o quão fuzzy será as partições finais. De modo que para $m = 1$ a partição resultante é totalmente *crisp* (Figura 2.1) e para $m \to \infty$ a interseção entre os grupos tende a aumentar (Figura 2.2)(Pal et al., 2005).

Segundo Bezdek et al. (1984) nenhuma teoria ou evidência computacional aponta para um valor ótimo de m, contudo o autor aponta que a faixa de valores ideais aparenta ser $[1, 30]$. Sendo assim, se existir um conjunto de dados para teste, uma boa estratégia para a escolha de m é a realização de testes experimentais. Caso contrário, o intervalo de $[1.5, 3.0]$ aparenta trazer bons resultados para a maior parte dos dados.

Onde usualmente no contexto do agrupamento de coleções textuais $dist(d_i, g_k) = scos(d_i, g_k)$. Adicionalmente, o algoritmo FCM estabelece algumas restrições: deve-se evitar que o FCM produza grupos vazios (Equação (2.9)) e deve-se impor que a soma das pertinências seja sempre igual a um (Equação (2.10)).

$$\sum_{k=1}^{c} \mu(d_i, g_k) = 1 \tag{2.9}$$

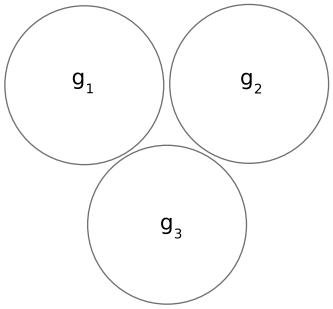

Figura 2.1: Ilustração denotando os grupos g_1, g_2, g_3 organizados sem sobreposição, para $m = 1$.

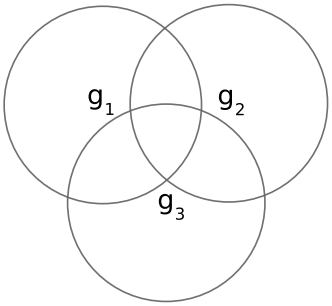

Figura 2.2: Ilustração denotando os grupos g_1, g_2, g_3 organizados de maneira fuzzy, com sobreposição, quando $m \to \infty$.

$$0 < \sum_{i=1}^{n} \mu(d_i, g_k) < n \tag{2.10}$$

.

A restrição imposta pela Equação (2.10), no entanto, produz um problema em elementos equidistantes aos grupos. Ou seja, quando temos o caso $\mu(d_i, g_1) = \mu(d_i, g_2) = ... = \mu(d_i, g_c)$. Nessa situação, o grau de pertinência de um elemento a cada grupo será a pertinência média, ou seja $\mu(d_i, g_1) = \mu(d_i, g_2) = ... = \mu(d_i, g_c) = \frac{1}{c}$. Supondo agora um segundo documento d_2, que seja mais distante do que d_1, porém assim como d_2, também equidistante aos grupos, temos que $\mu(d_2, g_j) = \mu(d_1, g_j) = \frac{1}{c}$, para $1 < j \leq c$. Nesse contexto, a pertinência de d_2 e d_1, não expressa a distância relativa desses documentos aos grupos. Esse problema está ilustrado na Figura 2.3.

Segundo Bezdek et al. (1984) várias funções de otimização da partição fuzzy produzida no agrupamento foram propostas, sendo a minimização da função objetivo $J(U_{c\times n}, G, V, D)$, definida na Equação (2.11) a mais popular. Onde considere $U_{c\times n}$ como os graus de pertinência fuzzy, V os protótipos dos grupos G e D o conjunto de documentos.

$$min\{(J(U_{c\times n}, G, V, D) = \sum_{i=1}^{n}\sum_{j=1}^{c}[\mu(d_i, g_j)]^m dist(d_i, v_j)\} \tag{2.11}$$

O algoritmo mais comumente utilizado para prover soluções aproximadas dessa minimização (2.11), é a através da iteração de Picard[1](Pal et al., 2005) entre as equações (2.6) e (2.7). O ciclo de aproximações se dá por $V_{t-1} \Rightarrow U_t \Rightarrow V_t$, onde, ao final de cada

[1]Método iterativo para construção de soluções aproximadas, atribuído ao matemático francês Charles Emile Picard (1856-1941). ⟨http://mathfaculty.fullerton.edu/mathews/n2003/PicardIterationMod.html⟩

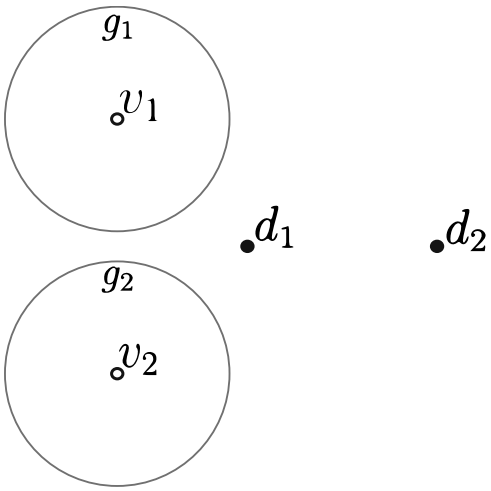

Figura 2.3: Problema dos elementos equidistantes do algoritmo FCM. Na imagem g_1 e g_2 são grupos, com os seus respectivos protótipos v_1 e v_2. Enquanto d_1 e d_2 são documentos equidistantes aos protótipos v_1 e v_2. Portanto $\mu(d_1, g_1) = \mu(d_1, g_2) = \mu(d_2, g_1) = \mu(d_2, g_2) = 0.5$.

iteração, é verificado se $\| V_{t-1} - V_t \| < \varepsilon$. A literatura também expressa que os ciclos podem começar pela partição fuzzy, fazendo então $U_{t-1} \Rightarrow V_t \Rightarrow U_t$, e, ao final do ciclo, checando o erro mínimo com $\| U_{t-1} - U_t \| < \varepsilon$, sendo t o contador de iterações. Contudo, existem benefícios em termos de processamento e memória ao utilizar a iteração iniciando e finalizando com V(Pal et al., 2005). Por outro lado, Bezdek et al. (1984) e Pal et al. (2005) afirmam que a convergência desse modelo iterativo ocorre em ambos os tipos de ciclo. Geralmente, a partição fuzzy inicial U_0 é comumente inicializada com valores aleatórios ou com o resultado de um agrupamento previamente executado (Pal et al., 2005; Krishnapuram e Keller, 1993). Nas demais iterações a atualização dos protótipos é realizada a partir da Equação (2.6).

O pseudo código utilizando a abordagem iterativa descrita a cima, está listado no Algoritmo 1, no qual a função **inicializa-particao-fuzzy(D,G)**, pode ser uma das duas formas de inicialização descritas anteriormente.

fcm(D, c, m, ε)
início
  G $\leftarrow [g_1, g_2, ..., g_c]$;
  $U_0 \leftarrow$ **inicializa-particao-fuzzy(D,G)**;
  $t \leftarrow 0$;
  faça
    $V_t \leftarrow$ calcula usando (2.6);
    $t \leftarrow t + 1$;
    $U_t \leftarrow$ calcula usando (2.7);
  enquanto $\| U_{t-1} - U_t \| > \varepsilon$;
  retorne(U_t, V_t);
fim

Algoritmo 1: Pseudo código da implementação iterativa do método FCM

Por fim, está ilustrado na Figura 2.4, os resultados produzidos pelo algoritmo FCM, em dois conjuntos de coordenadas no R^2. Na figura, os pontos foram pintados com a cor correspondente ao grupo em que o mesmo obteve o maior valor de pertinência.

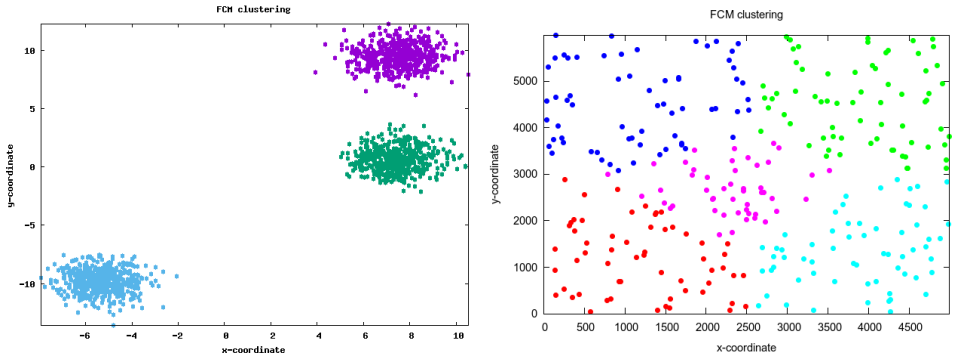

Figura 2.4: Resultado do agrupamento de dois conjuntos de coordenadas no R^2 usando o algoritmo FCM[2].

2.4.2 Algoritmo Possibilistic C Means (PCM)

A restrição probabilística do FCM, apresentada na Equação (2.10), que obriga a soma das pertinências de um elemento ser igual a um, nem sempre resulta em pertinências que representam bem a realidade dos dados, conforme exemplificado na Figura 2.3. Esse problema se agrava ainda mais, em bases com muitos dados ruidosos (*outliers*). Portanto, com o objetivo de contornar esses problemas do FCM, foi proposto em Krishnapuram e Keller (1993) o algoritmo *Possiblistic C Means* (PCM).

Ao contrário do FCM, o PCM não atribui pertinências dos documentos aos grupos,

mas sim tipicidades, as quais podem ser interpretadas como graus de possibilidade de um elemento pertencer a um determinado grupo. Como consequência, a partição resultante é possibilística. Para se adequar a essa abordagem possibilística, a função objetivo do PCM deriva da Equação (2.11) do FCM. Tendo também as funções de atualização de protótipos e atribuição de pertinências modificadas.

Na teoria de conjuntos fuzzy, a pertinência de um elemento a um grupo fuzzy não depende da pertinência desse mesmo elemento em outro grupo. No entanto, no modelo FCM, a restrição apresentada pela Equação (2.10) torna dependente a pertinência dos elementos aos grupos. De maneira que se um elemento obtiver um grau elevado de pertinência em um dado grupo, ele não poderá ter uma pertinência também elevada em outro grupo. Ou seja $\mu(d_1, g_1) = 1 - \mu(d_1, g_2)$, supondo um agrupamento com dois grupos. Nesse contexto Krishnapuram e Keller (1993) sugere relaxamento da restrição, permitindo assim, que a pertinência dependa unicamente da distância do elemento ao grupo. Logo, as restrições apresentadas nas Equações (2.9) e (2.10), são redefinidas nas Equações (2.12) e (2.13), nas quais $\lambda(d_i, g_j)$ representa a tipicidade do documento d_i em relação ao grupo g_j.

$$\lambda(d_i, g_j) \in [0, 1], \forall i, j \tag{2.12}$$

$$0 < \sum_{j=1}^{n} \lambda(d_i, g_j) \leq n, \forall j \tag{2.13}$$

Com isso, se mantiver-se a função objetivo do FCM (Equação 2.11), seria possível obter uma solução trivial, bastando atribuir 0 à todas as pertinências para minimizar se a função objetivo $J(U_{c\times n}, G, V, D)$ (Krishnapuram e Keller, 1993). Portanto, para evitar essa solução, e manter a característica de se atribuir pertinências elevadas aos elementos representativos e penalizar os elementos não representativos, os autores reformularam a função objetivo do FCM conforme está apresentado na Equação 2.14. No qual o parâmetro γ_j regula o limiar da distância dos documentos ao grupo g_j, de modo que se um documento tiver distância maior do que γ_j a sua tipicidade em relação ao $grupo_j$ será menor do que 0,5, e se a distância for menor do que γ_j, o seu grau de compatibilidade no $grupo_j$ será maior do que 0,5.

$$K_m(P_{c\times n}, G, V, D) = \sum_{j=1}^{c}\sum_{i=1}^{n}[\lambda(d_i, g_j)]^m dist(d_i, v_j) + \sum_{j=1}^{c}\gamma_j \sum_{i=1}^{n}[1 - \lambda(d_i, g_j)]^m \tag{2.14}$$

De acordo com Krishnapuram e Keller (1993), o valor de γ_j deve ser escolhido a depender da faixa de possibilidades (tipicidades) desejada para um grupo. Por exemplo, γ_j pode ser igual para todos os grupos, quando se deseja que o formato dos grupos seja

similar. Contudo, na maioria dos casos, se espera que γ_j reflita o formato e tamanho particular de cada grupo. Assim sendo, os autores indicam que a definição apresentada na Equação (2.15) tem se mostrado adequada para maior parte dos dados, onde L é usualmente 1.

$$\gamma_j = L\frac{\sum_{i=1}^{n}\lambda(d_i,g_j)^m dist(d_i,v_j)}{\sum_{i=1}^{n}\lambda(d_i,g_j)^m} \tag{2.15}$$

A partição de graus de compatibilidade possibilísticos do PCM está definida na Equação (2.16), a qual $\lambda(d_i,g_k)$ é sujeita as restrições apresentadas nas Equações (2.12) e (2.13).

$$P_{c\times n} = \{\lambda(d_i,g_k)|\lambda(d_i,g_k)\in[0,1], 1<i\leq n, 1<k\leq c\} \tag{2.16}$$

$$\lambda(d_i,g_j) = \frac{1}{1+\left(\frac{dist(d_i,g_j)}{\gamma_j}\right)^{\frac{1}{m-1}}} \tag{2.17}$$

Por sua vez, a atualização de protótipos do PCM, ocorre de maneira similar a Equação (2.6) do FCM, apenas alterando a pertinência $\mu(d_i,g_j)$ por $\lambda(d_i,g_j)$.

A síntese do algoritmo PCM está apresentada em forma de pseudo código no Algoritmo 2.

pcm(D,c,m,ε)
início
  G $\leftarrow [g_1, g_2, ..., g_c]$;
  $P_0 \leftarrow$ **inicializa-particao-fuzzy(D,G)**;
  $\gamma_j \leftarrow$ calcula utilizando (2.15);
  $t \leftarrow 0$;
  faça
    $V_t \leftarrow$ calcula usando (2.6);
    $t \leftarrow t+1$;
    $P_t \leftarrow$ calcula usando (2.16);
  enquanto $\| P_{t-1} - P_t \| > \varepsilon$;
  retorne(P_t, V_t);
fim

Algoritmo 2: Pseudo código da implementação iterativa do método PCM

Por fim, está ilustrado na Figura 2.5 (a), o resultado do agrupamento obtido pelo método FCM; e, em (b), o agrupamento gerado pelo algoritmo PCM em um conjunto de coordenadas no R^2. Na figura, os pontos foram pintados com a cor correspondente ao grupo em que o mesmo obteve o maior valor de pertinência. Observa-se nessa comparação simplificada que o algoritmo PCM tentou maximizar a pertinência dos pontos aos grupos maiores, ocasionando uma maior quantidade de pontos com pertinência elevada em dois grupos, ao contrário do FCM que distribuiu de maneira uniforme os pontos em 4 grupos.

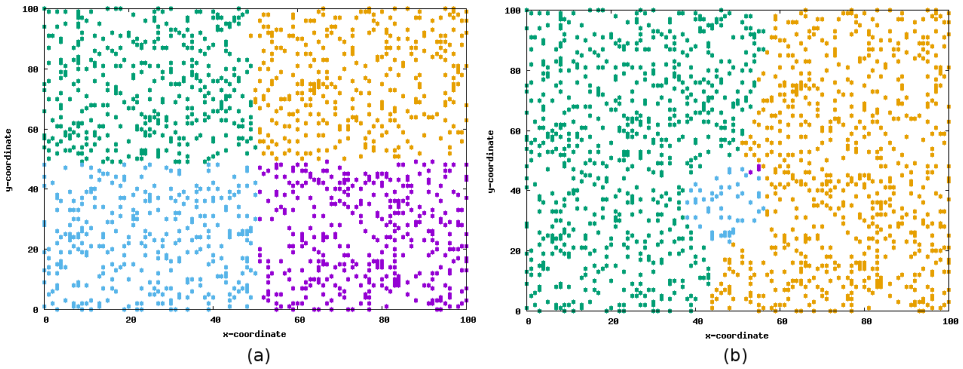

Figura 2.5: Demonstração de agrupamentos obtidos com os algoritmos FCM[3](a) e PCM[3](b) .

2.4.3 Algoritmo Possibilistic Fuzzy C Means (PFCM)

De acordo com Pal et al. (2005) o algoritmo PCM pode levar os resultados do agrupamento a conter grupos coincidentes. Ou seja, quando o protótipo v_i está muito próximo de outro protótipo v_j. Segundo os autores, isto ocorre quando a inicialização da partição inicial não possui protótipos suficientemente separados. Esse problema não é causado por uma escolha ruim da penalidade presente na função objetivo do PCM, o que ocorre é uma falta de restrições para evitar que isso aconteça.

Carvalho et al. (2016) cita que as pertinências do FCM e as tipicidades do PCM são ambas importantes para a correta interpretação das sub estruturas dos dados. Quando se tem dados que precisam ser agrupados de maneira *hard*, as pertinências se mostram como uma escolha adequada, de modo que é intuitivo atribuir o elemento ao grupo em que o mesmo possua a maior pertinência. Por outro lado, durante a atualização dos protótipos, as tipicidades desempenham um papel fundamental para aliviar os efeitos indesejados dos dados ruidosos.

Com o propósito de aproveitar os benefícios de ambas as abordagens, Pal et al. (2005) propôs o algoritmo PFCM, que utiliza as pertinências $\mu(d_i, g_j)$ do FCM e as tipicidades $\lambda(d_i, g_j)$ do PCM. Cabe ao usuário definir a proporção de cada uma das contribuições com parâmetros que ponderam o peso de ambos. Para tanto, é realizada uma mistura entre as funções objetivo apresentadas nas Equações (2.11) e (2.14) resultando na minimização da

[3]Resultados obtidos baseados na implementação dos algoritmo FCM e PCM, produzida como parte deste trabalho disponível em: ⟨https://github.com/niltonvasques/fcm⟩

função objetivo apresentada na Equação (2.18), a qual está sujeita as condições impostas pela Equação (2.19), onde $a, b > 0$ e $m, n > 1$. Por sua vez, os parâmetros a e b, representam a importância relativa dos valores de pertinência e tipicidades, os quais ficam a critério do usuário a sua definição de acordo com o contexto dos dados. De maneira geral, os autores sugerem que b seja maior que a, porém não muito maior, para não eliminar completamente os benefícios do FCM.

$$L_m(U_{c\times n}, P_{c\times n}, G, V, D) = \sum_{j=1}^{c}\sum_{i=1}^{n}[a\mu(d_i, g_j)^n + b\lambda(d_i, g_j)^m]dist(d_i, v_j) + \sum_{j=1}^{c}\gamma_j \sum_{i=1}^{n}[1-\lambda(d_i, g_j)]^m \tag{2.18}$$

$$\sum_{j=i}^{c}\mu(d_i, g_j) = 1, \forall i, 0 < \mu(d_i, g_j), \lambda(d_i, g_j) \leq 1 \tag{2.19}$$

A mistura e as ponderações adicionados no algoritmo PFCM também são agregadas a função de atualização dos protótipos apresentada na Equação (2.20), a qual passa a se beneficiar das características de ambos os algoritmos.

$$V = \left\{ v_j | v_j = \frac{\sum_{i=1}^{n}[a\mu(d_i, g_j)^n + b\lambda(d_i, g_j)^m]d_i}{\sum_{i=1}^{n}[a\mu(d_i, g_j)^n + b\lambda(d_i, g_j)^m]}, 1 < j \leq c \right\} \tag{2.20}$$

Desta maneira, é minimizado os efeitos dos dados ruidosos do FCM, assim como também o problema dos protótipos coincidentes do PCM e a singularidade do FCM é evitada.

O pseudo código do PFCM é apresentado no Algoritmo 3, onde a função **inicializa-prototipos**(D, G) é responsável por gerar os protótipos iniciais da partição V_0.

Como resultado demonstrativo desse algoritmo, é possível observar na Figura 2.6, que os grupos produzidos são em certa perspectiva um intermédio entre o agrupamento produzido pelo FCM e PCM no mesmo conjunto de dados, apresentados na Figura 2.5.

[4]Resultados obtidos baseados na implementação dos algoritmo FCM e PCM, produzida como parte este trabalho disponível em: ⟨https://github.com/niltonvasques/fcm⟩

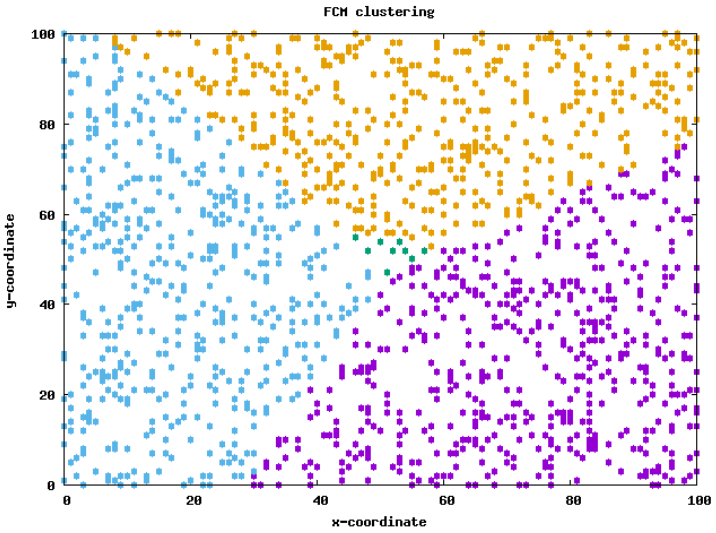

Figura 2.6: Demonstração do agrupamento obtido com os algoritmo PFCM[4]em um conjunto de coordenadas de pontos no R^2.

pfcm(D, c, m, ε)
início
  G $\leftarrow [g_1, g_2, ..., g_c]$;
  $V_0 \leftarrow$ **inicializa-prototipos(D,G)**;
  $\gamma_j \leftarrow$ calcula utilizando a Equação (2.15);
  $t \leftarrow 0$;
  faça
    $U_t \leftarrow$ calcula com Equação (2.7) usando V_{t-1};
    $P_t \leftarrow$ calcula com Equação (2.16) usando V_{t-1};
    $V_t \leftarrow$ calcula com Equação (2.20);
    $t \leftarrow t + 1$;
  enquanto $\| V_{t-1} - V_t \| > \varepsilon$;
  retorne(U_t, P_t, V_t);
fim

Algoritmo 3: Pseudo código da implementação iterativa do método PFCM

2.4.4 Algoritmo Hierarchic Fuzzy C Means (HFCM)

Documentos de texto tratam de vários temas, como política, esporte, tecnologia e etc. E os métodos de agrupamento *soft*, como FCM e PCM, quando aplicados a coleções textuais, buscam encontrar semelhanças entre os documentos e agrupar por temas relacionados. Adicionalmente um tema pode se dividir em sub temas, como por exemplo esporte, que pode se dividir em futebol, vôlei, tênis e etc. Deste modo, os temas presentes em uma coleção textual podem ser organizados em uma hierarquia de tópicos conforme a Figura

2.7. Neste contexto, construir hierarquias utilizando métodos de agrupamento fuzzy é o propósito principal do algoritmo HFCM proposto em Pedrycz e Reformat (2006).

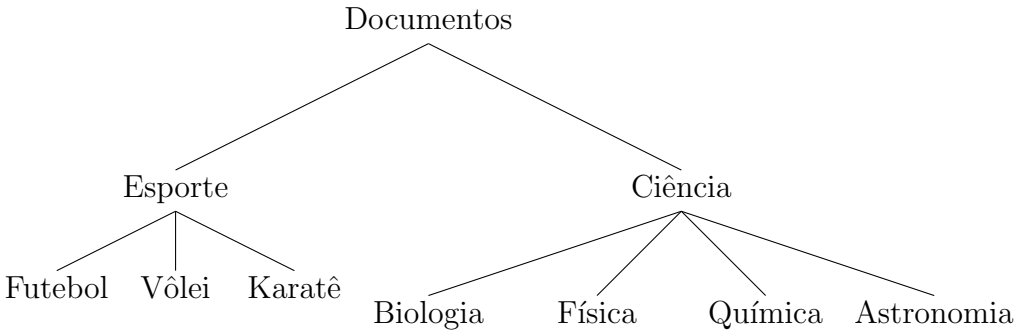

Figura 2.7: Exemplo de hierarquia de tópicos presentes em uma coleção de textos.

O HFCM consegue realizar essa tarefa expandindo sucessivamente as folhas presentes na hierarquia em subgrupos mais detalhados (Pedrycz e Reformat, 2006). Essa expansão é realizada através de novos agrupamentos com o algoritmo *Conditional Fuzzy C-Means* (CFCM), que é uma versão condicional do FCM. O primeiro nível da hierarquia é o resultado direto do algoritmo FCM, enquanto os demais níveis são agrupamentos obtidos com o CFCM, sobre a coleção de documentos filtrada no grupo a ser expandido.

O HFCM é executado sobre um conjunto de n documentos, $D = \{d_1, d_2, ..., d_n\}$, no qual inicialmente é aplicado o algoritmo FCM para produzir o primeiro nível da hierarquia. Como resultado é gerada a partição $U_{c\times n}[1]$ e os protótipos $V[1] = \{v_1[1], v_2[1], ..., v_c[1]\}$ são gerados, onde [1] representa o primeiro nível da hierarquia.

A expansão ocorre sempre nas folhas da hierarquia, e a decisão de expandir um determinado grupo, se dá pela avaliação do agrupamento, que é realizado através do índice de desempenho Q apresentados nas Equações (2.21) e (2.22), onde Q representa a qualidade dos protótipos gerados pelo agrupamento. Quanto melhor for um grupo, mais próximo de zero será o resultado da medida de desempenho Q. E quanto maior for o desempenho, pior será o grupo. Portanto, o grupo que obtiver o maior valor de Q, será escolhido para a expansão condicional através do CFCM.

$$Q_j[1] = \sum_{d_i \in g_j} dist(d'_i, d_i)^2 \tag{2.21}$$

$$Q'_j[2] = \sum_{d_i \in g_j} dist(d'_i, d_i)^2 \tag{2.22}$$

Considere então j o grupo com maior valor de Q do nível l da hierarquia e $D_j[l]$ (Equação 2.23) a coleção de documentos selecionados do grupo j com pertinência maior que a pertinência média.

$$D_j[l] = \left\{ d_k | \mu(d_k, g_j[l]) \leq \frac{1/l}{c} \right\} \tag{2.23}$$

O agrupamento com o algoritmo CFCM é então executado sobre a coleção $D_j[l]$ produzindo c novos grupos para o nível $l+1$ da hierarquia. Todo o processo se repete então para o nível $l+1$ da hierarquia, e assim sucessivamente. Segundo os autores, o ponto de parada da expansão hierárquica pode ser uma profundidade predefinida pelo usuário , com a estabilização das medidas de desempenho dos grupos ou supervisionada, de modo que o usuário observa a hierarquia que está sendo produzida e interrompe o processo quando desejar (Pedrycz e Reformat, 2006).

$$d'_i = \sum_{h=1}^{c} \mu(d_i, g_h)[1] v_h[1] \tag{2.24}$$

$$d'_i = \sum_{h=1}^{c} \mu'(d_i, g_h)[j, 2] v_h[2] \tag{2.25}$$

Segundo Nogueira (2013), os protótipos dos grupos representam uma versão condensada dos documentos agrupados. Portanto, d_i também pode ser representado pela combinação linear das pertinências de d_i com os protótipos, resultando em d'_i. Logo, é esperado que d'_i seja o mais próximo possível do documento original d_i. Consequentemente, é utilizado, essa noção para estimar a qualidade de um grupo através das Equações (2.21) para o nível inicial da hierarquia e (2.22) nos demais níveis, onde Q calcula a soma total das distâncias dos documentos d_i de um grupo com d'_j.

A atualização dos protótipos no algoritmo CFCM ocorre da mesma forma definida no algoritmo FCM. Contudo, a função de pertinência é redefinida para a Equação (2.26), a qual é imposta a restrição apresentada na Equação (2.27). Considere, que o valor de l corresponde ao nível da hierarquia e g_j seja o grupo expandido no nível $l-1$.

$$\mu(d_i, g_h[l]) = \frac{\mu(d_i, g_j[l-1])}{\sum_{k=1}^{c} \left(\frac{dist(d_i, v_h[l])}{dist(d_i, v_k[l])} \right)^{\frac{1}{m-1}}}, 1 < h \leq c, d_i \in D_j[l-1] \tag{2.26}$$

Ao observar-se o numerador da Equação (2.26), percebe-se que a pertinência de um documento d_i em um grupo $g_h[l]$, será no máximo a pertinência de d_i ao grupo imediatamente superior na hierarquia. Essa, noção está explicitamente definida na Equação (2.27), a qual é uma adaptação da restrição probabilística do FCM (Equação 2.10). A Equação (2.27), estabelece, que a soma das pertinências de um documento d_i no nível l

da hierarquia terá que ser igual a pertinência desse documento no grupo $g_j[l-1]$ que foi expandido no nível anterior$(l-1)$.

$$\sum_{h=1}^{c} \mu(d_i, g_h[l]) = \mu(d_i, g_j[l-1]), d_i \in D_j[l] \tag{2.27}$$

O pseudo código do método CFCM é apresentado no Algoritmo 4, de modo a deixar uma representação mais objetiva de como estruturar esses elementos.

cfcm($D_j[l]$, **U[l-1], l, c, m,** ε)
início
  G[l] $\leftarrow [g_1, g_2, ..., g_c]$;
  $V_0[l] \leftarrow$ **inicializa-prototipos(D,G)**;
  $t \leftarrow 0$;
  faça
    $U_t[l] \leftarrow$ (Equação 2.26) usando V_{t-1};
    $V_t[l] \leftarrow$ (Equação 2.6);
    $t \leftarrow t+1$;
  enquanto $\| V_{t-1}[l] - V_t[l] \| > \varepsilon$;
  retorne$(U_t[l], V_t[l])$;
fim

Algoritmo 4: Pseudo código do método CFCM

No Algoritmo 5 está apresentado o pseudo código do método HFCM, exemplificando como o mesmo reúne o FCM e o CFCM para produzir uma hierarquia de tópicos. Onde o critério de parada adotado foi a profundidade máxima da hierarquia, representado com o parâmetro $lmax$.

Na seção a seguir, tem-se uma breve descrição da extração dos descritores dos grupos, possibilitando assim a interpretação dos seus significados.

2.5 EXTRAÇÃO DE DESCRITORES

A tarefa de atribuir significados à grupos é um dos problemas chave do agrupamento de textos, pois ao final do processo de agrupamento, os grupos precisam apresentar alguma relevância para o usuário(Zhang e Xu, 2008). Portanto, é imprescindível que sejam extraídos descritores significativos para representar os documentos que compõem os grupos.

A etapa de extração de descritores etapa pode ser realizada manualmente, com o usuário guiando o processo, ou de forma automatizada, que por sua vez é mais interessante para a proposta de organização flexível de documentos, visto que, para grandes bases de dados textuais, a tarefa de extrair descritores para todos os grupos encontrados durante o agrupamento, pode ser bastante exaustiva para o usuário.

Dentre os métodos automatizados, é encontrado na literatura dois tipos de abordagens, uma baseada em conhecimento interno e a outra baseada em conhecimento externo.

hfcm(D, c, m, ε, lmax)
início
$l \leftarrow 0$;
Hierarquia $\leftarrow$;
G[l],V[l],U[l] $\leftarrow$ **fcm**(D,c,m,ε);
Hierarquia $\leftarrow Hierarquia + \{U[l], V[l]\}$;
Q[l] $\leftarrow$ calcula desempenho dos grupos (Equação 2.21);
$g_{max} \leftarrow$ escolhe o grupo g_j com maior valor de Q;
$D_{max} \leftarrow$ seleciona documentos de g_{max} (Equação 2.23);
$l \leftarrow l + 1$;
faça
Q[l] $\leftarrow$ calcula desempenho dos grupos (Equação 2.22);
$g_{max} \leftarrow$ escolhe o grupo g_j com maior valor de Q;
$D_{max} \leftarrow$ seleciona documentos de g_{max} (Equação 2.23);
G[l],V[l],U[l] $\leftarrow$ **cfcm**$(D_{max}, U[l-1], l, c, m, \varepsilon)$;
Hierarquia $\leftarrow Hierarquia + \{U[l], V[l]\}$;
$l \leftarrow l + 1$;
enquanto $l \leq lmax$;
retorne$(Hierarquia)$;
fim

Algoritmo 5: Pseudo código do método HFCM

A primeira se utiliza somente de informações que podem ser obtidas na coleção de documentos, como por exemplo a frequência do termo, localização do termo na estrutura do documento. Enquanto a abordagem de conhecimento externo, leva em consideração fontes de informação externas, como por exemplo a consulta a extensa base de termos na língua inglesa WordNet[5], para auxiliar a escolha dos termos mais representativos.

Em ambas abordagens, a literatura fornece uma ampla gama de métodos, com o objetivo de obter bons descritores dos grupos. Os descritores podem ser extraídos com os termos mais frequentes dos documentos no grupo. No entanto, o resultado pode ser genérico demais (Treeratpituk e Callan, 2006). Dessa maneira uma outra estratégia mais adequada pode ser a extração dos descritores dos documentos que estão mais próximos do centróide do grupo.

Nogueira (2013) destaca que grande parte dos métodos de extração de descritores encontrados na literatura são embutidos na fase de agrupamento. O que justifica a avaliação dos mesmos em função do desempenho do agrupamento. No entanto, essa junção da extração de descritores na fase de agrupamento, dificulta a combinação de diferentes técnicas de agrupamento e consequentemente a escolha de bons descritores. Logo, os métodos onde a extração é realizada após a fase de agrupamento, de maneira independente, permitem uma melhor adaptação da proposta de organização flexível de documentos para diferentes contextos.

Nesse contexto, percebe-se à existência de algumas estratégias para extração de descri-

[5]http://wordnet.princeton.edu/

tores, utilizando ou não conhecimento externo, e embutida ou independente do processo de agrupamento. A partir da avaliação dessas abordagens, e de acordo com o objetivo definido nesta monografia, considerou-se que a abordagem independente do algoritmo de agrupamento é mais pertinente ao presente estudo, pois ela viabiliza a condução de experimentos com vários métodos de agrupamento. Nesse sentido, foi escolhido o método *Soft Organization - Fuzzy Description Comes last* (SoftO-FDCL) proposto por Nogueira (2013), devido o mesmo possuir essas características necessárias para a investigação dos impactos do agrupamento na qualidade dos descritores e da organização flexível. Optou-se então para descrever este método no Capítulo 4, pois, as motivações para os métodos propostos nesta monografia se baseiam nas descobertas de propriedades do método SoftO-FDCL.

Na seção a seguir, está apresentada as considerações extraídas desse capítulo e a conexão dos temas aqui apresentados com a pesquisa realizada nesta monografia.

2.6 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado e fundamentado a teoria necessária para se compreender os temas dissecados nesta monografia. De maneira geral, foi visto que a proposição da lógica fuzzy, proporcionou a capacidade no tratamento de imprecisão e incerteza inerentes do mundo real. E em particular, esses benefícios permitiram o surgimento de mecanismos para organizar de maneira flexível os documentos textuais.

Foi visto que a organização flexível possui uma série de etapas no seu processo, tendo cada uma suas peculiaridades e teoria relacionada. Sendo que a etapa de pré-processamento, desempenha o fundamental papel de coletar e estruturar os dados textuais, preparando-os para a etapa de agrupamento. O agrupamento por sua vez, é realizado com métodos clássicos já existentes na Mineração de Dados (MD), porém ligeiramente adaptados para compreender a alta dimensionalidade dos das coleções textuais. Ao final desse processo, a extração dos descritores desempenha a tarefa de atribuir significado relevante à grupos, finalizando a organização flexível de documentos. A partir então dessa organização produzida, é possível um Sistema de Recuperação de Informação (SRI), utilizar essa organização para indexar e recuperar informações.

Considerando o objetivo definido nessa monografia, será investigado no Capítulo 4, os impactos dos algoritmos de agrupamento FCM, PCM e PFCM na qualidade da organização flexível de documentos.

No próximo capítulo, será apresentado uma breve revisão das estratégias recentes adotadas por pesquisadores para otimizar as etapas da organização flexível de documentos.

Capítulo 3

Trabalhos relacionados a organização flexível de documentos e sistemas de recuperação de informação.

TRABALHOS RELACIONADOS

3.1 CONSIDERAÇÕES INICIAIS

A proposta de organização flexível de documentos está relacionada a vários campos de estudo . Por isso a literatura existente para essa proposta é bastante rica e densa. Com o propósito de otimizar a atividade de pesquisa e seleção do conhecimento científico produzido a respeito do tema, foram utilizadas algumas técnicas de revisão sistemática de literatura ($SLR - Sistematic\ Literature\ Review$) utilizadas em Rios e Mello (2010). Com o objetivo de estabelecer critérios mais precisos na fase inicial da descoberta de conteúdo científico relacionado ao tema. Especificamente foi adotada uma técnica comum ao método SLR, que consiste na elaboração de uma string de busca, usando operadores lógicos. Estabelecendo assim uma maneira mais objetiva para a obtenção de resultados relevantes à proposta nesta monografia. Levando em consideração os tópicos chave e a proposta desse trabalho, foi construída a seguinte string de busca:

$$(clustering\ OR\ "cluster\ label*"\ OR\ "cluster\ descriptors")\ AND\ fuzzy\ AND\ (document\ OR\ "text\ mining"\ OR\ "document\ organization"\ OR\ "soft\ document"\ OR\ "text\ data") \quad (3.1)$$

Devido o acervo de publicações científicas possuir grande diversidade, assim como também a possibilidade de se utilizar operadores lógicos e buscas parametrizadas, foi realizado então uma busca no repositório IEEExplore[1], com resultados dos anos de 2010 e 2016, permitindo assim a obtenção de artigos mais recentes.

Com base nos resultados obtidos, foi realizada a leitura dos títulos e resumos dos artigos, com o propósito de descartar resultados com baixa relevância para a pesquisa apresentada nesta monografia. Durante a fase de leitura parcial dos resultados da busca,

[1] http://ieeexplore.ieee.org/

foram agrupados os artigos em três categorias: agrupamento fuzzy, extração de descritores e organização flexível de documentos. As publicações selecionadas e direcionadas para a categoria de agrupamento fuzzy, foram as que possuíam propostas de alteração de métodos de agrupamento existentes ou novos métodos. Enquanto artigos que tinham como conteúdo a análise dos termos de uma coleção, critério de seleção de termos ou atribuição de termos a grupos de documentos, foram agrupados na categoria de extração de descritores. Por fim, artigos mais gerais, propondo métodos ou realizando revisões de métodos, pertinentes ao processo de organização de documentos textuais, foram categorizados no grupo de organização flexível de documentos.

Para complementar os resultados obtidos foram adicionados artigos de alta relevância para o tema, e que apesar de serem antigos, ainda são amplamente citados em pesquisas recentes. Muitos desses artigos como é o caso do método FCM proposto em Bezdek et al. (1984), são pilares fundamentais para o tema.

As próximas seções contém a revisão das pesquisas selecionadas, onde é elucidado os pontos chave de cada pesquisa, a definição das propostas contidas nos artigos e, por fim, a conexão com o objetivo dessa monografia.

3.2 ORGANIZAÇÃO FLEXÍVEL DE DOCUMENTOS

A lógica fuzzy foi proposta por Lofti Zadeh para lidar com a incerteza e imprecisão em diversos problemas do mundo real. A partir desse trabalho desenvolvido por Zadeh, várias pesquisas se propuseram a explorar os benefícios concedidos pela flexibilidade proporcionada pela lógica fuzzy. Em particular, a mineração de textos (MT), desenvolveu um método para organizar de maneira flexível uma coleção de documentos, baseado na lógica fuzzy. Baseados nessa especialização da MT, uma série de pesquisas tem sido desenvolvidos, com o propósito de aprimorar a organização flexível de documentos. A seguir está apresentado as abordagens recentes encontradas na literatura a respeito deste tema.

Segundo Matsumoto e Hung (2010), os mecanismos adotados em sistemas de recuperação de informação (SRI), tais como buscadores web, estão dispostos em duas abordagens. A primeira abordagem, consiste do usuário realizando a busca, a qual é comumente chamada de busca web personalizada. Nessa abordagem, os resultados obtidos são ordenados de acordo com a relevância do resultado para o usuário. Para calcular essa relevância, os buscadores realizam tarefas de coleta de dados dos usuários e comparação das preferências com demais usuários do sistema. Já na segunda abordagem os resultados da busca são categorizados, permitindo assim que o usuário decida em qual categoria ele pretende visualizar as informações. Por exemplo, quando um usuário pesquisar pelo termo java, os resultados poderiam ser agrupados nas seções: máquina virtual, linguagem java, programas em java, oracle e etc. Seguindo essa linha de categorização de resultados em SRIs, Marcacini e Rezende (2010) propõem uma abordagem de agrupamento incremental e hierárquico para construção dos tópicos dos documentos, a qual permite a atualização das categorias a medida que novos documentos são adicionados sem realizar a etapa de agrupamento novamente. A visualização dessa abordagem de categorização hierárquica,

é possível através da ferramenta online Torch[2].

O surgimento de várias tecnologias, como mídias sociais, computação ubíqua, internet das coisas e principalmente os dispositivos móveis, que ultrapassou os 7 bilhões de dispositivos[3] no ano de 2015, produzindo uma imensa quantidade de dados não estruturados diariamente. Tem dificultando a tarefa de métodos de mineração de dados, e por consequência os métodos de organização de documentos já existentes via agrupamento. A esse cenário é usualmente atribuído o nome de *Big Data*. Nesse contexto, as pesquisas apresentadas em Havens et al. (2012) e Kumar et al. (2015), são focadas em bases com imensas quantidades de dados . Segundo Havens et al. (2012) existem duas abordagens principais para otimizar o agrupamento de dados que se encaixam na categoria *Very Large*. Conforme apresentado na 3.1), a primeira consiste na técnica de agrupamento distribuído incremental e a segunda no agrupamento por amostragem progressiva ou aleatória. Nos métodos que usam a técnica de amostragem, primeiramente é selecionado uma amostra com os dados representativos da coleção, depois é realizado o agrupamento, e em seguida é generalizado o agrupamento para o restante dos dados. Um dos métodos mais populares baseado em amostragem é o algoritmo *generalized extensible fast FCM* (geFFCM)(Havens et al., 2012). O geFFCM utiliza amostragem progressiva para se obter uma versão reduzida dos dados, de maneira que a mesma preserve as características da base original. Porém segundo Havens et al. (2012), a técnica de amostragem do geFFCM é ineficiente para dados na categoria *Very Large*, o que levou os autores a propor uma extensão do geFFCM com uma melhoria na forma de realizar a amostragem dos dados, utilizando uma metodologia de seleção aleatória.

Bytes	10^6	10^8	10^{10}	10^{12}	$10^{>12}$
"tamanho"	medium	large	huge	monster	very large

Tabela 3.1: Classificação das bases de dados de acordo com o seu tamanho (Havens et al., 2012)

De acordo com Deng et al. (2010), a organização flexível de dados através do algoritmo FCM possui uma falta de estabilidade, pois como a inicialização do FCM depende da aleatoriedade, o resultado final do agrupamento pode variar a cada inicialização. Assim como os dados presentes em bases de dados textuais são de alta dimensionalidade. Os autores propuseram então um modelo de inicialização da partição que extrai da coleção medidas de peso, raio e objetos mais representativos para orientar a inicialização da partição inicial. A respeito do problema da dimensionalidade, Deng et al. (2010) sugere a redução da matriz documentos x termos, usando uma medida estatística para avaliar a qualidade dos termos presentes na coleção, descartando assim os termos considerados de baixa qualidade e consequentemente reduzindo a largura da matriz.

Karami et al. (2015) propõe um modelo para análise textual de documentos médicos. Um dos pontos interessantes propostos pelo autor é a utilização do agrupamento fuzzy na

[2]http://sites.labic.icmc.usp.br/torch/webcluster/

[3]Segundo o relatório do The Mobile Economy disponível em ⟨http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf⟩, a quantidade de dispositivos móveis (smartphones e tablets) atingiu o total de 7,517 bilhões no ano de 2015.

etapa de pré-processamento e ponderação dos termos, antes de realizar o agrupamento e classificação. O agrupamento fuzzy é aplicado a coleção de termos presentes na coleção, e ao contrário do agrupamento na etapa pós processamento, a pertinência ocorre da palavra a um tópico ou grupo, de maneira que termos com alta pertinência possuem significados semânticos mais próximos. Essa aproximação semântica é realizada com base em um vocabulário predefinido.

Conforme foi definido no capítulo 2, os algoritmos de agrupamento fuzzy apresentados são baseados na otimização de funções objetivos. Contudo, Cunha et al. (2012) descrevem que a otimização de funções que possuam vários mínimos locais sem nenhuma tendência global, utilizando as clássicas técnicas de otimização, podem simplesmente não só demorar a convergir, como a convergência pode nunca ser encontrada, ou ainda convergir para um mínimo local. Portanto diante da necessidade de se otimizar funções com tais características, desenvolveu-se estratégias de otimização baseadas em heurísticas, os quais permitem uma análise das características globais da função, de maneira que se possa realizar sucessivas buscas do mínimo global. Por outro lado, métodos heurísticos não garantem a convergência para o mínimo global, porém de modo geral apresentam boas aproximações.

Dentre estas, tem-se heurísticas evolutivas inspiradas no processo de adaptação dos seres vivos. Que é o caso dos algoritmos genéticos, que são métodos de busca de soluções probabilísticos inspirados nos mecanismos genéticos de adaptação dos seres vivos através da seleção natural.

Assim sendo Jiang et al. (2013) traz um proposta para enriquecer a tarefa de agrupamento textual, baseada na combinação dos algoritmos genéticos, com o método de agrupamento FCM, para evitar o problema de convergência para mínimos locais. O problema de convergência do FCM , ocorre quando algumas condições iniciais são satisfeitas, levando o FCM para convergir para um mínimo local (Bezdek et al., 1984).

O autor destaca ainda que os algoritmos genéticos possuem grande potencial na realização de buscas paralelas globais, principalmente quando trata-se de grande volumes de dados, com elevados requisitos de classificação e computação paralela, restrições as quais o algoritmo iterativo para otimizar a função objetivo do FCM não é capaz de atender. No artigo, explora as qualidades e deficiências do algoritmo *Immnune Genetic Algorithm*(IGA), que é uma versão mais atual do algoritmo genético (GA), que garante a diversidade dos indivíduos, aprimorando assim os potenciais de busca pela solução ótima global (Jiang et al., 2013). Porém, o IGA demanda muito custo computacional para garantir a diversidade da população, assim como também existe uma dificuldade em estimar os parâmetros iniciais, levando o IGA a uma finalização precoce, ou uma execução indefinida. Por conta dessas deficiências, o algoritmo *Partheno Genetic Algorithm*(PGA) é investigado , o qual não necessita da inicialização da população inicial, e também não sofre do problema de finalização prematura do IGA. Para solucionar esse problema, o autor propõe o método PGA-FCM, que utiliza o PGA para encontrar solução aproximada global da função de minimização do FCM, e em seguida utiliza o próprio FCMi, tendo como entrada a partição de saída do PGA, para encontrar a solução ótima, de maneira a reduzir o tempo de convergência.

Segundo Saranya e Arunpriya (2014), está presente na literatura uma outra visão do

agrupamento textual, que realiza o agrupamento das sentenças contidas nos documentos. Tal abordagem tem-se mostrado fundamental para, por exemplo, evitar a sobreposição de informações em técnicas de sumarização de documentos. Assim como também pode ser útil para a obtenção de novas informações a partir de um conjunto de resultados em uma busca na web por exemplo. A sumarização de sentenças tem como objetivo, produzir um sumário que contenha informações relevantes dos documentos. Onde por sua vez, estratégias de ordenação das sentenças(*ranking*) de acordo com sua relevância, medidas de similaridade de sentenças utilizando informações da semântica dos termos são utilizadas para cumprir o objetivo. O autor também descreve que o agrupamento fuzzy de sentenças, pode ser utilizado para realizar tarefas de análise de contradição do conteúdo com o assunto da coleção de documentos o qual este está inserido. De modo que se possa identificar dados ruidosos que não contribuem com a ideia geral da coleção.

Ainda seguindo essa linha de otimização dos resultados apresentados após uma busca em um SRI, Nogueira et al. (2012) cita que uma das principais limitações de um SRI, está na maneira rígida de interpretar as strings de busca do usuário. Pode ocasionar que alguns documentos relevantes para o usuário, não sejam retornados na busca. O autor informa que a literatura usualmente aborda esse problema de duas formas, em que ambas se baseiam na reformulação das strings de busca. Sendo que a primeira procura reformular os termos de uma busca, levando em conta características semânticas dos termos, de modo a encontrar versões mais apropriadas, permitindo então que uma maior quantidade de documentos relevantes seja retornado pela busca. A segunda forma de abordar esse problema consiste em ir ajustando os termos da busca, a partir dos primeiros documentos encontrados.

Seguindo a abordagem de reformulação semântica, em Murali e Damodaram (2015) é apresentado um sistema de recuperação de informação e organização flexível de documentos, de maneira a considerar a informação semântica contida nos termos. A abordagem proposta pelos autores torna mais precisa a similaridade entre os documentos, assim como também atua na reformulação da busca. A proposta usa identificação de sinônimos baseado em ontologias da WordNet, hiperônimos[4] e hipônimos[5], para realizar tarefas de desambiguação semântica entre os termos, durante o pré-processamento. Em seguida os autores utilizam o método de agrupamento PFCM, devido ao potencial de agregar as qualidades e minimizar as deficiências dos algoritmos FCM e PCM. Adicionalmente a extração de termos descritores dos grupos é realizada, construindo-se um índice de termos para cada grupo. Um dos diferenciais apresentados na proposta de Murali e Damodaram (2015) ocorre na fase final, quando o usuário realiza uma busca. As palavras chave inseridas pelo usuário, são processadas na ontologia do WordNet visando encontrar palavras com significados similares. A partir dessas palavras resultantes a similaridade com os termos descritores dos grupos, é calculado e por fim os documentos que estiverem no grupo que obteve a maior similaridade com a busca digitada pelo usuário são retornados.

Com o objetivo de apresentar uma alternativa a essas duas abordagens para prover

[4]Palavras de sentido genérico, ou seja, possuem significado bem amplo. Por exemplo ferramenta é um hiperônimo de chave de fenda.

[5]Palavras de sentido específico, sendo a palavra que está abaixo de um hiperônimo na hierarquia. Por exemplo, chave de fenda é hipônimo de ferramenta.

flexibilidade em um SRI, Nogueira et al. (2012) propõem um método para gerenciar a precisão e incerteza inerentes a documentos textuais, ainda na fase de organização dos documentos. Evitando assim a dependência da busca e a participação do usuário no processo. Um dos diferenciais apresentados pelos autores é a possibilidade de obtenção dos termos de busca ainda na fase de organização de documentos, de maneira não supervisionada. O que traz uma grande independência ao processo, possibilitando a automatização da organização flexível de documentos com pouca ou nenhuma intervenção do usuário. Os termos de busca encontrados na etapa de organização dos documentos são os descritores dos grupos, que são utilizados para distribuir os documentos em tópicos. Uma vez que durante a etapa de organização é utilizado o agrupamento fuzzy, os documentos podem ser atribuídos mais de um tópico. Essa abordagem fortalece ainda mais a flexibilidade na organização de documentos.

A extração de termos que representem bem os grupos obtidos na fase de agrupamento é de fundamental importância, pois é a partir dos termos extraídos que é possível a recuperação dos documentos através de uma consulta realizada por um usuário. Dada essa importância, Nogueira (2013) conduziu uma investigação acerca dos principais métodos de extração de descritores, que resultou na proposta de um modelo de SRI para organização flexível de documentos, com a adição de três métodos de extração de descritores. Os métodos de extração de descritores propostos, se baseiam em medidas na clássicas na literatura, que medem a efetividade da recuperação da informação de um SRI, as quais são a precisão, revocação e a medida-F. A medida de precisão procura identificar dentre os documentos recuperados a proporção de documentos relevantes. Enquanto a revocação calcula a taxa de documentos relevantes recuperados a partir dos documentos que são previamente definidos como relevantes durante uma consulta em um SRI. A metodologia de extração adotada pelo autor nos métodos propostos, é do tipo *Description Comes Last* (DCL), que realiza a extração de descritores separado do processo de agrupamento, em contraste com outros métodos de extração do tipo *Description Comes First* (DCF), no qual os descritores são extraídos na fase de pré-processamento ou simultaneamente com o agrupamento. Nogueira (2013) destaca que utilizar métodos do tipo DCL, traz o benefício de tornar independente o algoritmo de extração de descritores do algoritmo de agrupamento utilizado.

O primeiro método proposto em Nogueira (2013) é o SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*), que tem como propósito extrair descritores de partições fuzzy de documentos, de maneira a flexibilizar um SRI. A avaliação dos termos candidatos a descritores é feita então com base na pertinência do documento ao qual o termo pertence, onde os termos só são considerados se o documento tiver pertinência maior que um dado limiar, definido como sendo $\delta = \frac{1}{c}$. Ao fazer isso o método penaliza os termos de documentos com baixa pertinência ao grupo e ao mesmo tempo considera os termos de documentos que pertencem também a outros grupos, de modo a conservar a flexibilidade que o agrupamento fuzzy proporciona.

Como extensão ao SoftO-FDCL, foi também proposto o método Soft-wFDCL(*Soft Organization - weighted Fuzzy Description Comes Last*), onde é levado em consideração também o grau de pertinência do documento no cálculo de relevância do termo. A

adição da pertinência na equação, é justificada como sendo um fator de ponderação da importância de um termo ao grupo. Deste modo os termos que forem considerados relevantes para compor os descritores, ou seja os que estiverem a cima do limiar definido no método SoftO-FDCL, serão ponderados em função da pertinência.

Os dois métodos anteriores são aplicados a uma organização flexível de documentos organizados de modo $flat$. Com a finalidade de adicionar flexibilidade também a uma organização de documentos hierárquica, como por exemplo a obtida no método de agrupamento HFCM, Nogueira (2013) traz o método HSoftO-FDCL (*Hierarchical Soft Organization - Fuzzy Description Comes Last*), como outra extensão do SoftO-FDCL. A modificação realizada nesse método ocorre no limiar de corte dos termos, pois no algoritmo HFCM a pertinência de um documento em um dado grupo, é relativa a pertinência desse mesmo documento no grupo imediatamente superior na hierarquia, conforme foi definido na equação 2.27 na página 22. Assim sendo o limiar do SoftO-FDCL é redefinido como sendo a pertinência do documento d_i no grupo superior dividido pela quantidade de grupos, ou seja $\zeta = \dfrac{\mu(d_i, g_j[l-1])}{c}$, tal que c seja a quantidade de grupos e l o nível do grupo que está sendo extraído os descritores. Para os grupos do primeiro nível da hierarquia os descritores são extraídos usando o δ.

Em Yan et al. (2013) é informado que em estudos recentes os métodos de co-agrupamento fuzzy, tem apresentado resultados superiores a extensões do FCM, e métodos de co-agrupamento *crisp*, para algumas bases de dados com alta dimensionalidade. O diferencial dos métodos de co-agrupamento fuzzy em relação ao agrupamento fuzzy clássico (FCM), está no agrupamento simultâneo de documentos e termos, assim como a literatura sugere a relevância desse método para categorização de dados com alta dimensionalidade (Yan et al., 2013). Esse potencial dos algoritmos de co-agrupamento é devido à sua inerente característica de redução da dimensão dos dados, através do agrupamento também dos termos, que permite melhor capturar a estrutura dos dados. Entretanto base de dados com elevada sobreposição de termos entre os documentos, ou seja termos que aparecem em vários documentos, podem afetar significativamente os métodos de co-agrupamento, assim como a organização dos termos produzidos podem não condizer com a realidade dos dados (Tjhi e Chen, 2008). Assim sendo o método HFCR definido em Tjhi e Chen (2008), tenta contornar esses problemas, substituindo o *ranking* dos termos, pelo agrupamento fuzzy, gerando assim duas partições fuzzy, uma de termos e uma de documentos. Segundo os autores, essa estratégia permite o HFCR contornar os problemas comuns dos métodos de co-agrupamento, mantendo a capacidade de melhor agrupar dados de alta dimensionalidade e reduzindo a complexidade computacional em relação a outros algoritmos de co-agrupamento presentes na literatura. No entanto, a função objetivo do HFCR possui complexidade de $O(c * n * k)$, que é maior que a complexidade da função objetivo (equação 2.11) do FCM clássico $O(c * n)$, onde c é o número de grupos, n é o número de documentos e k o número de termos distintos em toda coleção.

3.3 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentadas pesquisas relacionadas ao tema desta monografia, em que foi possível observar a variedade de estratégias adotadas para aumentar a eficiência de sistemas de recuperação de informação quando utilizados no contexto da organização flexível de documentos. As pesquisas aqui discutidas apresentam propostas de melhorias em todas as etapas presentes na organização flexível de documentos. Na fase de pré-processamento é possível utilizar abordagens semânticas para a compactação da quantidade de termos. Na estratégia de avaliar a semântica das palavras também pode ser utilizada na etapa final durante a recuperação da informação. Outras estratégias concentram-se na etapa de agrupamento, com inúmeras possibilidades de otimização. As pesquisas aqui apresentadas, exploram a adaptação dos algoritmos fuzzy existentes para o problema do *BigData*, ou propõe a utilizam de outras heurísticas no processo, como a adição dos algoritmos genéticos no agrupamento. Outra alternativa, é a extrapolação do agrupamento, para abordar também os termos, realizando assim um melhor reconhecimento da estrutura dos grupos, no entanto, essa extrapolação vem com uma carga de processamento adicional. Outro conjunto de autores focam seus esforços em melhorar a extração de descritores dos grupos obtidos, afinal mesmo com um bom agrupamento realizado, se bons termos não forem obitidos para descrever os grupos, a etapa de recuperação dos documentos pode ficar prejudicada.

As pesquisas relacionadas ao tema demonstram que a organização de documentos não é um problema exaurido, e que não possui uma solução canônica. Consequentemente, percebe-se que existe bastante trabalho a ser feito para otimizar a organização flexível de documentos. No próximo capítulo então, as abordagens propostas como resultado da investigação conduzida nesse trabalho, assim como os resultados dos experimentos são apresentadas.

Capítulo

4

Investigação e refinamento do método de organização flexível de documentos

ABORDAGEM PROPOSTA

4.1 CONSIDERAÇÕES INICIAS

A organização de uma coleção de documentos em vários tópicos, de modo que exista sobreposição entre os grupos, é um importante problema em sistemas de recuperação de informação(SRIs). Na literatura, diversas estratégias são utilizadas visando otimizar a organização flexível de documentos, conforme foi abordado no capítulo anterior. Soma-se a isso o fato de que a maioria dos métodos que adicionam flexibilidade ao processo, como por exemplo os métodos de agrupamento, nem sempre são desenvolvidos com o foco em documentos textuais, os quais possuem características que acrescentam algumas dificuldades no processo, tais como a alta dimensionalidade dos dados, e o armazenamento de maneira não estruturada (Steinbach et al., 2003). Adicionalmente com o crescente aumento do uso de tecnologias de produção de conteúdo, a quantidade de dados textuais tem alcançado grandes volumes de dados, o que os enquadra no contexto do *Big Data*, fortalecendo a importância de se conduzir pesquisas e investigações em torno da organização flexível de documentos.

Nesse contexto, este capítulo tem como objetivo detalhar as contribuições desta monografia para organização flexível de documentos, através da investigação dos impactos de se utilizar uma estratégia híbrida de agrupamento fuzzy e possibilístico para a organização de documentos. Tal estratégia dar-se-á pelo uso do algoritmo *Possibilistic Fuzzy C-Means*, o qual pretende resolver os problemas dos elementos equidistantes e dos grupos coincidentes, apresentados nas partições fuzzy e possibilística respectivamente.

Conforme observa-se no capítulo 2, o algoritmo PFCM produz duas partições, sendo uma fuzzy e outra possibilística, o que induziu o presente trabalho a propor duas extensões do método de extração de descritores *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) proposto por Nogueira (2013). O primeiro método proposto, denominado PDCL(*Possibilistic Descriptor Comes Last*), aborda uma nova estratégia de interpretação dos graus de compatibilidade da partição possibilística. Enquanto a

segunda abordagem proposta, Mixed-PFDCL(*Mixed - Possibilistic Fuzzy Descriptor Comes Last*), é uma adaptação do PDCL, utilizando uma abordagem híbrida para mesclar as duas partições presentes no algoritmo PFCM.

Nas seções a seguir são apresentadas informações das bases de dados utilizadas, com as suas características, origem e composição dos documentos; um estudo sobre a interpretação dos graus de compatibilidade possibilísticos durante a extração de descritores; as propostas sugeridas por essa monografia, que derivam deste estudo; e, por fim, os resultados obtidos com os experimentos realizados.

4.2 COLEÇÕES TEXTUAIS

Na mineração de textos e consequentemente nos trabalhos relacionados à organização flexível de documentos, é comum se realizar a avaliação dos métodos propostos, conduzindo-se experimentos sobre coleções textuais existentes na literatura com essa finalidade (Rossi et al., 2013). Para isso, as coleções precisam se estarem estruturadas. Assim sendo, nesta pesquisa foi adotada a estrutura de representação de documentos textuais *tf-idf* apresentada no Capítulo 2, Equação (2.1), como forma de estruturar os dados presentes nas coleções, de modo a capturar a importância relativa dos termos nos documentos e nas coleções, montando assim ao final do pré-processamento uma matriz documentos x termos.

A base Opinosis[1]é composta de opiniões de consumidores a respeito das características de alguns produtos, obtidas dos portais amazon.com, tripadvisor e edmunds.com. As opiniões presentes na base, abordam tópicos como serviços de hospedagem, dispositivos eletrônicos e carros. Sendo que no total as sentenças presentes na coleção estão distribuídas em 51 categorias, onde cada categoria possui 100 aproximadamente. Os dados dessa base foram obtidos no repositório *UCI Machine Learning Repository* (Frank e Asuncion, 2010), que mantém várias coleções de dados que são utilizados pela comunidade de aprendizado de máquina.

A coleção de documentos 20Newsgroup[2]contém aproximadamente 20000 documentos de notícias, particionados em mais ou menos 20 temas. Para os experimentos realizados nesta pesquisa, foi utilizado uma amostragem da coleção, contendo 2000 documentos pertencentes ao tema ciência, a qual contém os tópicos sci.space, sci.electronics e sci.med. Esta base tem-se mostrado bastante popular em aplicações textuais de aprendizado de máquina, tais como agrupamento e classificação de textos. Essa base foi coletada originalmente por Ken Lang para a pesquisa Newsweeder apresentada em (Lang, 1995).

Os documentos presentes na base de dados Reuters-21578[3]apareceram inicialmente na Reuters newswire em 1987. Sendo que os documentos foram coletados e indexados diretamente por membros da Reuters e da *Carnegie Group, Inc.* também em 1987 para o desenvolvimento do CONSTRUE (Hayes e Weinstein, 1990), que foi um sistema de categorização de documentos. No ano de 1990 essa base de dados foi tornada pública pela Reuters, para ser utilizada em pesquisas de recuperação de informação. No entanto, as

[1] ⟨http://archive.ics.uci.edu/ml/datasets/Opinosis+Opinion+26frasl3B+Review⟩
[2] ⟨http://qwone.com/~jason/20Newsgroups/⟩
[3] ⟨https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection⟩

versões inicias dessa base continham documentos repetidos e ambíguos, o que motivou um grupo de pesquisadores de categorização textual durante a conferência *ACM SIGIR '96*, a realizar uma limpeza na base, possibilitando uma melhor comparação dos resultados entre diferentes estudos. Essa versão final ficou com o total de 21578 documentos, distribuídos entre 43 diferentes categorias. Nesta pesquisa foi utilizada uma amostragem da coleção contendo 1052 documentos selecionados aleatoriamente de cada classe da coleção.

A base de dados WAP(*WebACE Project*) é composta de um conjunto de páginas web coletadas por Moore et al. (1997), para um projeto de pesquisa de agrupamento, seleção e recuperação de páginas web. Os dados presentes nesta coleção foram obtidos pelos autores do artigo em 98 páginas web, onde posteriormente foram distribuídos em 20 diferentes categorias, que abrangem tópicos como negócios e finanças, tecnologias, trabalho e indústria. O conteúdo obtido está disposto em 1560 documentos na sua versão original e todos os documentos foram utilizados nesta pesquisa.

A coleção de documentos NSF[4](*National Science Foundation*) foi obtida do repositório de dados para pesquisas de aprendizado de máquina *UCI Machine Learning Repository* (Frank e Asuncion, 2010). O conteúdo dos dados presentes na base é composto de 129000 resumos, sendo um resumo por documento, descrevendo prêmios da NSF para pesquisas básicas. Para os experimentos descritos nesta monografia, foram selecionados 1600 documentos de maneira aleatória entre as categorias apresentadas na coleção.

A base de dados Hitech adquirida em Karypis (2015), é parte de uma coleção de bases da conferência TREC (*Text REtrieval Conference*)[5]. Esta base é composta de um conjunto de notícias da revista *Jose Mercury News*[6], as quais são distribuídas em categorias distintas. As notícias presentes na coleção abordam temas como computadores, eletrônicos, saúde, medicina, pesquisa e tecnologia. A base originalmente possui 2301 documentos e para esta pesquisa foram selecionados 600 documentos aleatoriamente.

Outro aspecto não menos importante, são as características particulares das coleções textuais. Pois ressalta-se que para uma mais apurada análise dos resultados, é pertinente considerar as particularidades de cada coleção, com a finalidade de encontrar possíveis justificativas para os resultados apresentados, realizando-se indagações comparativas às peculiaridades sabidamente conhecidas dos métodos analisados. O conjunto de características particulares de cada coleção obtidos em Rossi et al. (2013) e adaptados a esta pesquisa, dar-se à como apresentado na Tabela 4.1.

Uma análise objetiva das características presentes nas seis coleções utilizadas nos experimentos pode ser observada na Tabela 4.2, onde é possível notar de maneira bem objetiva ao se observar a coluna com o percentual de zeros da tabela, que todas as coleções apresentam uma quantidade de zeros em mais de 90% das frequências dos termos presentes na $tf - idf$, o que deixa explícito o peculiar problema dos dados esparsos já caracterizado ao longo do texto, como algo inerente aos dados textuais e que afeta negativamente grande parte dos resultados do processo de mineração de textos.

A seguir, um refinamento organização flexível é abordado utilizando o algoritmo PFCM.

[4] ⟨https://archive.ics.uci.edu/ml/datasets/NSF+Research+Award+Abstracts+1990-2003⟩
[5] ⟨http://trec.nist.gov/data.html⟩
[6] ⟨http://www.mercurynews.com/⟩

documentos	número de documentos presentes na coleção
termos	número de termos existentes na coleção após o pré-processamento
% zeros	número relativo de zeros na *tf-idf*, quantificando o quanto a matriz é esparsa
classes	número de classes presentes na coleção
n-gramas	quantidade de termos considerados sequencialmente na coleção

Tabela 4.1: Descrição das características objetivas presentes em coleções textuais elencadas para este trabalho

coleção	docs	termos	classes	% zeros	n-gramas
Opinosis	51	842	3	95,73%	1-grama
20newsgroups	2000	11028	4	99,11%	1-grama
Hitech	600	6925	6	97,93%	1-grama
NSF	1600	2806	16	99,76%	1-grama
WAP	1560	8070	20	98,51%	1-grama
Reuters-21578	1052	3925	43	98,55%	1-grama

Tabela 4.2: Características das coleções textuais utilizadas nesta pesquisa

4.3 REFINAMENTO COM O ALGORITMO PFCM

Conforme ficou evidenciado, a tarefa de organizar de maneira flexível um conjunto de documentos textuais, possui diversos desafios. Em particular, ao se agrupar um conjunto de documentos é esperado que os grupos resultantes possuam significado relevante, ou seja o algoritmo de agrupamento precisa detectar a estrutura natural dos documentos (Steinbach et al., 2003). Alguns desses desafios estão na dificuldade em escalar os métodos usuais para coleções textuais na categoria *Very Large* conforme a escala apresentada na Tabela 3.1; na obtenção de mecanismos efetivos para se avaliar a qualidade dos grupos produzidos; nas técnicas para se medir a interpretabilidade dos resultados; na capacidade para estimar os parâmetros dos algoritmos; na possibilidade de funcionar de maneira incremental, reduzindo o custo computacional durante a atualização dos grupos com novos dados; e, também na capacidade de continuar a produzir bons resultados em cenários compostos de documentos ruidosos (Carvalho et al., 2016).

Portanto, para Steinbach et al. (2003):

> *[...] there is no reason to expect that one type of clustering approach will be suitable for all types of data, even all high dimensional data. Statisticians and other data analysts are very cognizant of the need to apply different tools for different types of data, and clustering is no different.*

Diante dos desafios propostos, e com a evidência de que é possível aprimorar os resultados ao se utilizar novas estratégias de agrupamento, a investigação apresentada

nesta seção tem como objetivo analisar de qual forma a organização de documentos pode ser otimizada, ao aplicar na etapa de agrupamento uma estratégia que misture as partições possibilística e fuzzy, por meio do algoritmo PFCM.

Vale ressaltar, que a escolha desse algoritmo foi feita devido o seu potencial para absorver as qualidades presentes no *Fuzzy C-Means* (FCM) contrabalanceando as suas deficiências ao agregar também o *Possibilistic C-Means* (PCM) e sua partição possibilística. Além disso, existem diversas pesquisas na literatura abordando o desempenho do PFCM, como por exemplo em Pal et al. (2005), Yan e Chen (2009), Subhashini e Kumar (2010), Grover (2014) e Popescu et al. (2015).

Sendo assim, foram conduzidos experimentos adaptando a estratégia de organização flexível de documentos definida em Nogueira (2013), utilizando na etapa de agrupamento o método PFCM. Uma vez que esse método produz duas partições, uma possibilística e uma fuzzy, foi aplicado o método de extração de descritores SoftO-FDCL na partição fuzzy e também na partição possibilística, produzindo assim dois grupos de descritores.

Com essa adaptação espera-se uma melhor organização dos documentos, de forma que melhores descritores sejam escolhidos para caracterizar grupos. Tal processo de organização é ilustrado na Figura 4.1.

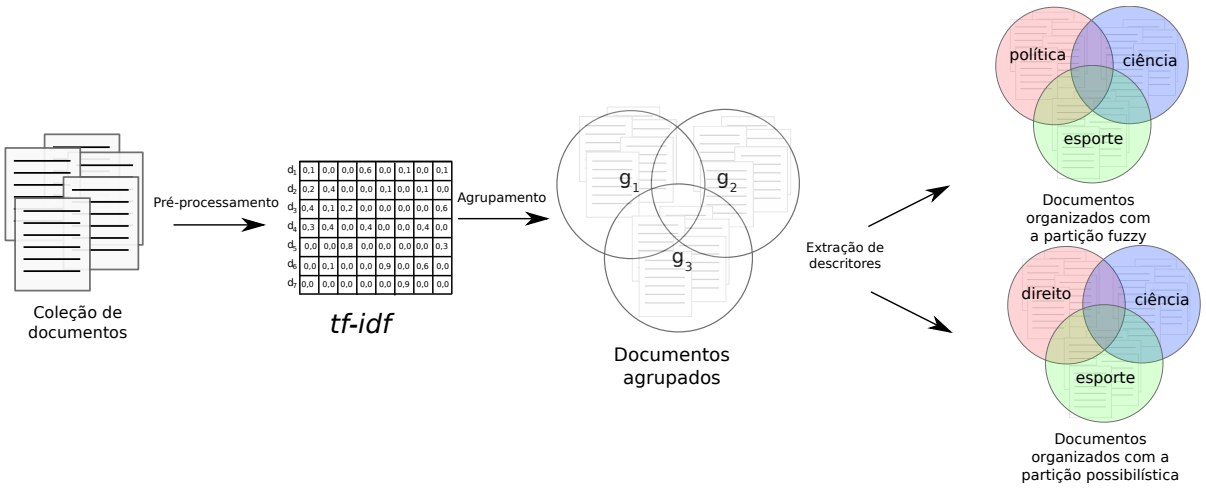

Figura 4.1: Estratégia de organização flexível de documentos adotada ao se misturar abordagens fuzzy e possibilísticas no agrupamento

Para se calcular a quantidade ótima de grupos, para cada coleção foi utilizado o método da silhueta fuzzy (Equação 2.5), método bastante utilizado com o propósito de avaliar o agrupamento de documentos. Assim sendo, o número ideal de grupos é determinado após a execução da silhueta fuzzy variando o número de grupos entre 2 e o número de classes de cada coleção. Ressalta-se que em coleções que os documentos não possuem rótulos, ou seja o número de classes é desconhecido, ainda é possível usar o método da silhueta fuzzy para definir o número ótimo de grupos. No entanto, a quantidade máxima de grupos deve ser definida de modo empírico ou com base em alguma informação prévia a respeito dos dados.

Para permitir uma análise comparativa dos resultados, o experimento foi realizado

também com os algoritmos FCM e PCM. Como resultado do agrupamento das coleções, está disposto na Tabela 4.11 a comparação do número de grupos obtidos por cada algoritmo de agrupamento. Nessa comparação nota-se que os algoritmos FCM e PFCM foram os que alcançaram uma quantidade de partições mais próxima da quantidade de classes existentes em cada coleção. Enquanto o PCM manteve uma tendência a produzir uma quantidade menor de grupos em relação aos demais.

coleção	**# classes**	**FCM**	**PCM**	**PFCM**
Opinosis	3	3	3	3
20Newsgroup	4	2	2	2
Hitech	6	6	5	5
NSF	16	11	2	16
WAP	20	14	5	16
Reuters-21578	43	22	11	36

Tabela 4.3: Quantidade ótima de grupos determinada através do método da silhueta fuzzy para cada algoritmo de agrupamento

Após agrupar os dados utilizando os métodos FCM, PCM e PFCM, foi aplicado o método de extração de descritores SoftO-FDCL. Para avaliar os descritores produzidos, foi verificado o potencial preditivo dos mesmos, possibilitando assim quantificar a qualidade dos termos selecionados para nomear os grupos.

A avaliação do potencial preditivo dos descritores foi realizada, realizando se uma defuzificação dos grupos produzidos pelo agrupamento, ou seja, se durante o agrupamento foi gerado o conjunto de grupos $G = \{g_1, g_2, ..., g_c\}$, temos então o conjunto de grupos *crisp* $C' = \{crisp_1, crisp_2, ..., crisp_c\}$, onde cada $crisp_i$ corresponde ao grupo g_i. A função de defuzificação adotada para se converter as as partições fuzzy e possibilística, as quais permitem que um documento pertença a um ou mais grupos, considera o grupo *crisp* de um documento d_i, como sendo o grupo $crisp_j$ do respectivo $grupo_j$, no qual d_i possui a maior pertinência/tipicidade. Esta definição está formalizada na Equação (4.1).

$$crisp(d_i) = \begin{cases} crisp_j, & \mu(d_i, g_j) = \max\limits_{\forall g \in G} \mu(d_i, g), \text{se a partição for fuzzy} \\ crisp_j, & \lambda(d_i, g_j) = \max\limits_{\forall g \in G} \lambda(d_i, g), \text{se a partição for possibilística} \end{cases} \tag{4.1}$$

Após se atribuir os documentos aos grupos *crisp*, é produzida uma outra matriz, considerando apenas os termos descritores dos grupos. Logo, essa matriz documentos x descritores $D'_{n \times m}$, é uma versão condensada da matriz documentos x termos $D_{n \times k}$, onde n corresponde a quantidade de documentos, k à quantidade de termos e m à quantidade de descritores. O conteúdo dessa matriz condensada, assim como na matriz original, é a frequência dos descritores nos documentos.

Visando avaliar a qualidade dos descritores e permitir uma comparação direta dos impactos dessa abordagem com os resultados publicados em Nogueira (2013) e Nogueira et al. (2015). Submeteu-se a matriz D' aos algoritmos de classificação SVM, Naive Bayes,

Multinomial Naive Bayes, KNN e C4.5, que são bem comuns na avaliação de métodos de aprendizado de máquina, e foram os mesmos utilizados em Nogueira et al. (2015).

Nesse contexto, foi utilizada a implementação dos algoritmos de classificação anteriormente citados presentes na ferramenta WEKA (Hall et al., 2009). Os algoritmos Naive Bayes (NB), Multinomial Naive Bayes (NB-Multinomial) e o J48 (que é a implementação do C4.5 existente no WEKA), foram executados com os parâmetros padrão da ferramenta. Por outro lado, o SVM foi ajustado para usar o *Normalized Polynomial Kernel* com o parâmetro de complexidade sendo $c = 2.0$. O algoritmo IBk (implementação do KNN presente no WEKA) foi executado 7 vezes, variando o parâmetro de vizinhos de 1 até 7, sendo escolhido o melhor resultado. Ressalta-se que foi adotada a técnica *10-fold cross validation* no experimento para melhor capturar a capacidade de generalização do modelo. Os resultados dessa avaliação estão apresentados nas Figuras 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7.

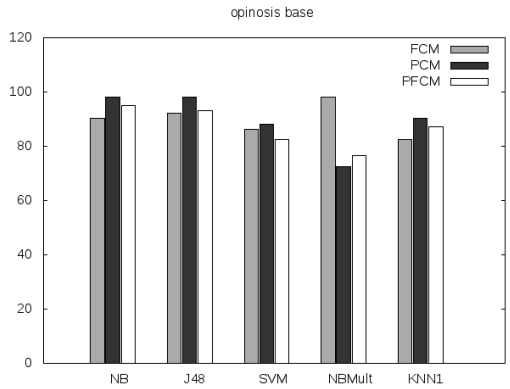

Figura 4.2: Desempenho obtido com os descritores extraídos com o algoritmo SoftO-FDCL a partir dos métodos de agrupamento FCM, PCM e PFCM executados na coleção Opinosis

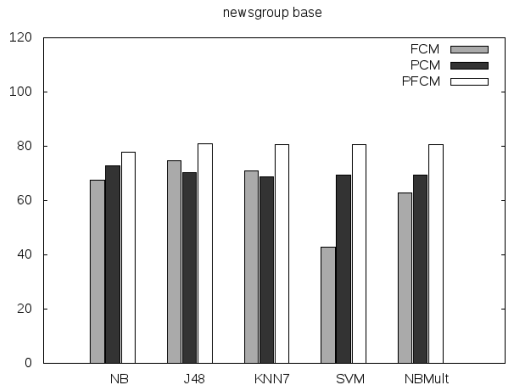

Figura 4.3: Desempenho obtido com os descritores extraídos com o algoritmo SoftO-FDCL a partir dos métodos de agrupamento FCM, PCM e PFCM executados na coleção 20Newsgroup

O resumo dos resultados do desempenho dos descritores extraídos após o agrupamento com os algoritmos FCM, PCM e PFCM é apresentado na Tabela 4.4. Na tabela, a marcação (✓) denota qual método de agrupamento obteve a maior taxa de classificação dentre os demais.

Esses resultados obtidos reforçam a flexibilidade e adaptação do método SoftO-FDCL (Nogueira, 2013), a novos algoritmos de agrupamento, demonstrando-se promissor na tarefa de extrair termos relevantes dos grupos produzidos na etapa de agrupamento. Tal expectativa é demonstrada por meio do potencial preditivo evidenciado na Tabela 4.4, com as taxas máximas de classificação de mais de 80% para quase todas as coleções, com exceção da base Hitech, a qual obteve a taxa máxima de 58.67%.

Adicionalmente, ressalta-se a importância também de avaliar de maneira subjetiva os descritores selecionados dos grupos, permitindo compreender se os termos obtidos fazem sentido para a organização de documentos em grupos.

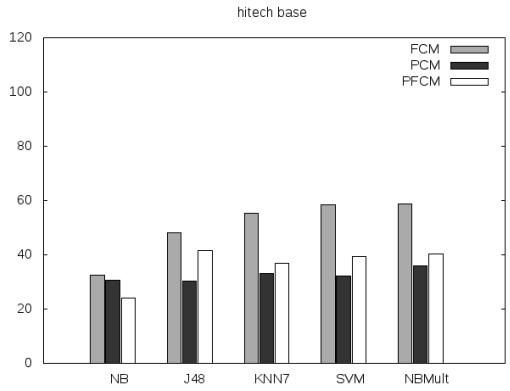

Figura 4.4: Desempenho obtido com os descritores extraídos com o algoritmo SoftO-FDCL a partir dos métodos de agrupamento FCM, PCM e PFCM executados na coleção Hitech

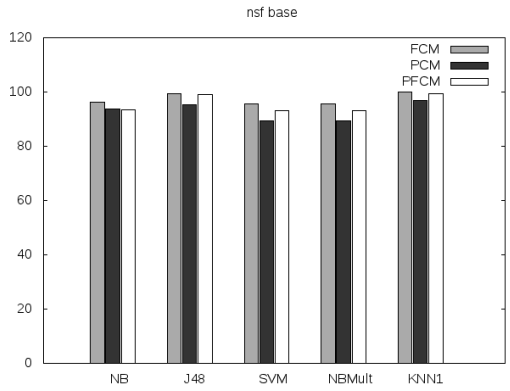

Figura 4.5: Desempenho obtido com os descritores extraídos com o algoritmo SoftO-FDCL a partir dos métodos de agrupamento FCM, PCM e PFCM executados na coleção NSF

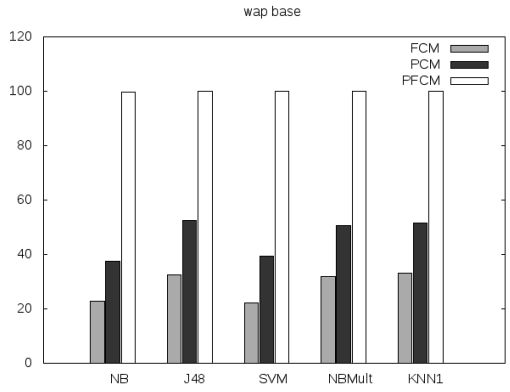

Figura 4.6: Desempenho obtido com os descritores extraídos com o algoritmo SoftO-FDCL a partir dos métodos de agrupamento FCM, PCM e PFCM executados na coleção WAP

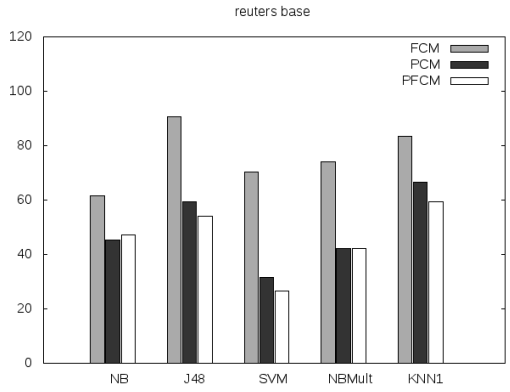

Figura 4.7: Desempenho obtido com os descritores extraídos com o algoritmo SoftO-FDCL a partir dos métodos de agrupamento FCM, PCM e PFCM executados na coleção Reuters-21578

nome	docs	termos	classes	% zeros	FCM	PCM	PFCM
Opinosis	51	842	3	95,73%		✓	
20newsgroups	2000	11028	4	99,11%			✓
Hitech	600	6925	6	97,93%	✓		
NSF	1600	2806	16	99,76%	✓		
WAP	1560	8070	20	98,51%			✓
Reuters-21578	1052	3925	43	98,55%	✓		

Tabela 4.4: Sumário dos resultados da classificação dos descritores

método	$crisp_1$	$crisp_2$	$crisp_3$
FCM	easy, clear, drive, display, control, car, version, nice, work, perfect	fact, import, isn't, model, problem, unit, design, don't, doesn't, found	breakfast, nearby, concierge, eat, bottle, coffee, floor, food, inn, friendly
PCM	easy, read, problem, version, don't, small, nice, car, work, found	fact, back, turn, expect, size, close, quality, review, min, feature	feel, amazing, isn't, extreme, drive, include, point, reason, give, run
PFCM μ	easy, drive, control, don't, version, nice, car, work, perfect, lot	fact, isn't, read, complete, device, display, size, doesn't, found	breakfast, nearby, pleasant, concierge, eat, coffee, floor, clean, friendly, food
PFCM λ	club, immaculate, send, towel, basic, exception, spotl, pillow, typical, fridge	pub, housekeep, holiday, tourist, tea, smoke, pm, renovate, facilite, london	usual, central, forum, bottle, modern, adult, supply, food, reserve, dinner

Tabela 4.5: Descritores extraídos com os métodos de agrupamento FCM, PCM e PFCM da coleção Opinosis, onde μ e λ se referem as partições fuzzy e possibilística respectivamente, da qual os descritores foram extraídos.

Sendo assim, para uma análise subjetiva dos resultados, os descritores da coleção Opinosis, foram obtidos por possuir poucos grupos e assim facilitar a análise e a visualização. A coleção Opinosis contém opiniões dos usuários a respeito de serviços de hospedagem, dispositivos eletrônicos e carros e espera-se que os descritores de grupos se aproximem semanticamente de tais categorias. Na Tabela 4.5 tem-se a seleção de descritores escolhidos para cada grupo, extraídos pelos algoritmos FCM, PCM e PFCM. Ao analisar os descritores selecionados é possível notar uma tendência geral, do grupo $crisp_1$ conter descritores relacionadas a carros, o $crisp_2$ conter descritores sobre dispositivos eletrônicos e o grupo $crisp_3$ descritores sobre hospedagem e alimentação. Contudo, nota-se que os descritores do PCM e do PFCMλ (descritores da partição possibilística do PFCM) estão um pouco mais misturados, não apresentando uma tendência geral bem definida. Uma explicação possível a esse resultado pode se encontrar na própria partição possibilística a qual permite que um mesmo documento possua um grau de tipicidade elevado em todos os grupos. Neste contexto, uma solução possível pode ser uma adaptação do método de extração de descritores SoftO-FDCL voltado para a partição possibilística, assim como também para algoritmos híbridos com duas partições, que é o caso do PFCM.

De maneira complementar, é importante salientar que o método de extração de descritores SoftO-FDCL é totalmente influenciado pelos valores contidos nas partições fuzzy e possibilística. Portanto, é também importante realizar uma análise dos métodos de agrupamento utilizados nos experimentos, com a finalidade de entender qual método é mais apropriado em determinados contextos. Os resultados apontam que a dimensionalidade

das bases foi um fator determinante no desempenho dos métodos de agrupamento para a organização flexível de documentos, e, consequentemente, a extração de descritores. Por exemplo, do sumário de resultados apresentados na Tabela 4.4, é possível observar que o método PCM obteve o melhor resultado na coleção Opinosis, que possui a menor dimensionalidade (842 termos), enquanto que o algoritmo FCM superou os demais métodos na coleção NSF (2806 termos), Reuters-21578 (3925 termos), Hitech (6925 termos), e por fim o algoritmo PFCM atingiu melhores resultados para as coleções WAP (8070 termos) e 20Newsgroup (11028 termos), que são as coleções de maior dimensionalidade.

Na próxima seção, motivado pelos resultados desses experimentos será explorado, uma adaptação do método SoftO-FDCL para evitar o processo de extração dupla de descritores em algoritmos que possuam partições fuzzy e possibilística.

4.4 UMA ABORDAGEM HÍBRIDA PARA EXTRAÇÃO DE DESCRITORES

Nos experimentos anteriores foi identificado um possível problema ao realizar a extração dos descritores de maneira separada em cada partição do PFCM, assim como também foi apontado que o método pode não capturar toda essência da partição possibilística, que difere da partição fuzzy do FCM por não possuir a restrição que obriga a soma das pertinências de um grupo ser igual a um (Equação 2.9). Logo, é intuitivo indagar que para uma melhor interpretação dos grupos produzidos em um método de agrupamento híbrido, seja pertinente utilizar também uma abordagem mista de extração de descritores. Aproveitando-se assim dos benefícios existentes na partição possibilística, a qual penaliza os elementos ruidosos, com baixos valores de tipicidade, sem abrir mão das vantagens presentes na partição fuzzy. Para isso é necessário compreender os mecanismos de funcionamento do método SoftO-FDCL, para que seja possível propor uma adaptação para este contexto.

4.4.1 Investigações na extração de descritores em partições possibilísticas

O método SoftO-FDCL (*Soft Organization - Fuzzy Description Comes Last*) proposto em Nogueira (2013) é baseado em uma adaptação das medidas clássicas de Recuperação de Informação (RI), para quantificar a relevância dos termos candidatos à descritores dos grupos obtidos na etapa de agrupamento, utilizando a informação de pertinência ou tipicidade advinda do algoritmo de agrupamento. Esse grau de compatibilidade entre um documento e um grupo, desempenha um papel fundamental na seleção dos termos candidatos a descritores de um determinado grupo, pois através deste, é possível penalizar os termos de documentos que possuam baixo grau de compatibilidade com o grupo no qual o descritor está sendo extraído.

Para realizar a extração dos descritores, inicialmente todos os termos que permanecem na coleção após a etapa de pré-processamento são considerados como candidatos a descritores. Posteriormente o método realiza uma avaliação quantitativa da relevância de cada termo t_k para um grupo g_j, utilizando a medida $f1$ apresentada na Equação 4.4, que é a média harmônica da precisão (Equação 4.2) e da revocação (Equação 4.3). A medida de precisão checa a quantidade de documentos significantes entre os documentos

recuperados. Enquanto a medida de revocação calcula a proporção de documentos relevantes recuperados entre todos os documentos relevantes da coleção. Tanto a medida de precisão quanto a de revocação, tomam como base as informações obtidas a partir da matriz de contingência apresentada na Tabela 4.6. Adicionalmente tem-se que um documento d_i é considerado como parte do grupo g_j, caso $\mu(d_j, g_j) \geq \delta$ para valores de pertinências ou $\lambda(d_i, g_j) \geq \delta$ para partição possibilística, onde $\delta = \frac{1}{c}$ e c a quantidade de grupos. O limiar δ é uma parte relevante do método SoftO-FDCL, pois ele possibilita considerar como candidatos a descritores, os termos presentes em documentos que pertençam a mais de um grupo, ao mesmo tempo que também penaliza os termos presentes em documentos com baixa compatibilidade em um grupo.

	Documentos do grupo g_j com grau de compatibilidade maior ou igual a δ	Documentos do grupo g_j com grau de compatibilidade menor do que δ
Documentos que possuem o descritor candidato t_k	*ganhos*	*ruídos*
Documentos que não possuem o descritor candidato t_k	*perdas*	*rejeitos*

Tabela 4.6: Matriz de contingência do termo t_k para o grupo g_j para as medidas de recuperação de informação

$$precisão(t_k, g_j) = \frac{ganhos}{ganhos + ruídos} \tag{4.2}$$

$$recuperação(t_k, g_j) = \frac{ganhos}{ganhos + perdas} \tag{4.3}$$

$$f1(t_k, g_j) = \frac{2 * precisão(t_k, g_j) * recuperação(t_k, g_j)}{precisão(t_k, g_j) + recuperação(t_k, g_j)} \tag{4.4}$$

Para efetuar a seleção dos termos descritores, é construído um *ranking* dos termos candidatos de cada grupo, ordenados pela pontuação obtida com a medida $f1$ (Equação 4.4). A partir dessa pontuação, os termos com as maiores pontuações em cada grupo são selecionados. Nogueira (2013) destaca que a quantidade de descritores a ser selecionada fica a critério do usuário.

Em síntese, o método SoftO-FDCL cria uma tabela de pontuação dos termos candidatos aos grupos, e seleciona os que obtiverem melhores valores de $f1$.

Formalizando essa percepção, é possível definir as duas propriedades que derivam dessa discussão, as quais são as Equações (4.5) e (4.6). A primeira propriedade apresentada na Equação (4.5) expressa que se um documento possuir pertinência maior do que o limiar δ em um grupo g_1 qualquer, obrigatoriamente existirá ao menos um outro grupo g_2 no qual esse mesmo documento terá pertinência inferior ao limiar δ. Portanto, essa propriedade nos indica que na maioria das vezes um ou mais documentos serão descartados na análise

dos termos candidatos do grupo, o que reforça a adequação desse limiar para a partição de pertinências. Por outro lado, a segunda propriedade apresentada na Equação (4.6) denota o único caso particular, no qual um documento d_i não será descartado em nenhum grupo. Isto ocorre apenas se d_i possuir pertinência igual ao limiar em todos os grupos, o que só acontece em dados ruidosos, que apresentam o problema do elemento equidistante detalhado na Figura 2.3 do Capítulo 2.

$$\mu(d_1, g_1) > \delta \rightarrow \exists \mu(d_i, g_2) < \delta \tag{4.5}$$

$$(\mu(d_1, g_1) = \delta) \wedge (\mu(d_i, g_2) = \delta) \wedge ... \wedge (\mu(d_i, g_{c-1}) = \delta) \rightarrow (\mu(d_i, g_c) = \delta) \tag{4.6}$$

Contudo, para a partição possibilística, essas duas propriedades não são satisfeitas. Isso ocorre devido a remoção da restrição da Equação (2.9), o que permite o grau de compatibilidade da partição possibilística variar de maneira independente entre o intervalo de $[0, 1]$, sem ser influenciado pelo grau de compatibilidade do documento nos demais grupos.

Para tornar claro essas análises, considere uma situação onde tem-se uma coleção de textos com 3 documentos (Tabela 4.7), onde cada documento possui 3 termos. Essa coleção de documentos foi agrupada, usando o método PFCM, o qual produziu 2 grupos, com as suas respectivas partições de pertinência (Tabela 4.8) e possibilística (Tabela 4.9).

	t_1	t_2	t_3
d_1	0	0	1
d_2	1	1	0
d_2	1	1	0

Tabela 4.7: Exemplo de matriz documentos x termos

	g_1	g_2
d_1	0.5	0.5
d_2	0.3	0.7
d_3	0.3	0.7

Tabela 4.8: Exemplo de matriz documents x grupos com graus de pertinência

	g_1	g_2
d_1	0.7	0.7
d_2	0.5	0.8
d_3	0.5	0.9

Tabela 4.9: Exemplo de matriz documents x grupos com graus de possibilidade

A partir desse agrupamento é possível aplicar a extração de descritores, utilizando as partições apresentadas nas Tabelas 4.8 e 4.9. Seguindo o procedimento do método SoftO-FDCL, inicialmente devemos considerar todos os descritores presentes na Tabela 4.7, como candidatos a descritores. Então, para promover os descritores candidatos do grupo g_1 utilizando os graus de pertinências, deve-se observar os documentos que possuem pertinência maior ou igual ao limiar, onde no exemplo é igual a $\delta = \frac{1}{2} = 0.5$. Portanto, o único documento que se enquadra no filtro do limiar é o documento d_1 e o termo t_3, que é o

único presente no documento d_1, é promovido a descritor candidato e consequentemente a descritor do grupo g_1. Já o grupo g_2 possui 3 documentos com pertinência maior ou igual ao limiar δ. Sendo assim, os termos de d_1, d_2 e d_3 são promovidas a descritores candidatos do grupo g_2. Ao calcular-se a medida $f1$ para cada um deles temos: $f1(t_1, g_2) = 0,8$, $f1(t_2, g_2) = 0,8$, $f1(t_3, g_2) = 0,5$. Logo, os descritores selecionados para cada grupo foram $g_1 = \{t_3\}$ e $g_2 = \{t_1, t_2\}$.

Ao aplicar a extração de descritores sobre a partição possibilística, tem-se agora os documentos d_1, d_2 e d_3 considerados parte do grupo g_1, de acordo com a regra do limiar. Logo, deve-se considerar os termos dos 3 documentos como descritores candidatos. Os valores de $f1$ para os 3 termos candidatos do grupo g_1 são: $f1(t_1, g_1) = 0,8$, $f1(t_2, g_1) = 0,8$, $f1(t_3, g_1) = 0,5$. Para o grupo g_2 obtém-se os mesmos termos candidatos, pois d_1, d_2 e d_3 também possuem valores de tipicidade maiores ou igual ao limiar δ, portanto os valores de $f1$ para o grupo g_2 são: g_1: $f1(t_1, g_1) = 0,8$, $f1(t_2, g_1) = 0,8$, $f1(t_3, g_2) = 0,5$. Foi obtido a mesma pontuação para os mesmos termos candidatos nos dois grupos. Com isso, fica claro as causas que levaram a extração dos descritores confusos apresentados na Tabela 4.5. Ao se utilizar a mesma intepretação dos graus de pertinência do FCM, nas tipicidades, o critério chave de descarte dos documentos de baixa relevância em um grupo deixou de fazer efeito.

É importante ressaltar que esta situação se agrava ainda mais com o aumento do número de grupos produzidos pelo agrupamento, pois quanto maior for a quantidade de grupos, menor será o valor do limiar δ. Esta noção fica explícita na terceira propriedade do limiar apresentada na Equação (4.7). Ou seja, mais facilmente os graus de compatibilidade possibilísticos irão passar no filtro do limiar, e, consequentemente todos os documentos serão considerados relevantes para todos os grupos.

$$\lim_{c \to \infty} \delta = 0 \tag{4.7}$$

Na próxima seção será descrito uma possível proposta para interpretação dos graus de compatibilidade das partições possibilística.

4.4.2 Interpretando os graus de compatibilidade das partições possibilísticas

A interpretação direta dos graus de compatibilidade possibilísticos gera uma série de problemas na extração de descritores, conforme ficou demonstrado na seção anterior. Com isso, podemos formular a seguinte pergunta: Como interpretar corretamente os graus de compatibilidade possibilísticos para corretamente identificar os documentos relevantes de um dado grupo? Sabe-se que o valor de tipicidade pode variar livremente entre o intervalo $[0, 1]$, sem a restrição probabilística (Equação 2.9) do FCM. Essa é uma característica positiva introduzida em Krishnapuram e Keller (1993), a qual atribui valores de pertinência mais justos aos grupos fuzzy, em consonância com a teoria de conjuntos fuzzy proposta em Zadeh (1965) e brevemente contextualizada no Capítulo 2. Para melhor compreendermos esse conceito de valores mais justos, podemos analisar o que os autores defenderam na publicação original:

> Since our membership functions correspond more closely to the notion of typicality, the resulting algorithms are naturally more immune to

> noise. Noise points will have low degrees of compatibility in all clusters, which makes their effect on the clustering negligible (Krishnapuram e Keller, 1993).

Portanto, a vantagem em deixar os graus de compatibilidade independentes é a de se expressar a relevância/importância real de um elemento em relação a um grupo, o que consequentemente torna o método mais robusto e menos suscetível a ruídos. Para identificar a importância de um documento em um grupo, pode-se por exemplo escolher em qual grupo um documento d_i deveria ser considerado relevante, considerando o grupo em que esse documento d_i possuísse o maior grau de compatibilidade. No entanto, essa estratégia resultaria em uma extração de descritores *crisp*, perdendo assim toda a flexibilidade proporcionada nas partições $soft$ (Nogueira, 2013).

Sendo assim, é necessária uma estratégia que consiga interpretar bem os graus de compatibilidade, de modo a se conservar a flexibilidade inerente à partição fuzzy, sem sacrificar a robustez contra ruídos da tipicidade.

Nesse contexto, propõe-se realizar tal interpretação em duas etapas. A primeira será constituída de uma conversão da tipicidade oriunda do PCM para a pertinência do FCM, de maneira a se satisfazer a restrição probabilística do FCM (Equação 2.9). No entanto, ao apenas realizar a conversão perde-se a robustez contra ruídos do PCM. Por isso, é possível contornar essa situação adicionando uma penalidade ao cálculo da pontuação dos termos.

A conversão proposta dos valores de tipicidade para pertinência, dar-se a como apresentado na Equação (4.8), a qual satisfaz a condição necessária (Equação 4.9) para que o limiar δ seja aplicado, sem sofrer o problema de considerar um documento como relevante em todos os grupos.

$$\lambda'(d_i, g_j) = \frac{\lambda(d_i, g_j)}{\sum_{k=1}^{c} \lambda(d_i, g_k)} \tag{4.8}$$

$$\sum_{k=1}^{c} \lambda'(d_i, g_k) = 1 \tag{4.9}$$

Entretanto, apenas realizar a conversão não atende aos dois requisitos exigidos. Para isso vamos adaptar a matriz de contingência apresentada na Tabela 4.6, adicionando uma ponderação as medidas, de modo a adicionar uma gratificação nos termos de documentos com elevada tipicidade e penalizar os termos de documentos com baixo valor de tipicidade, uma estratégia de ponderação similar é encontrada no método SoftO-wFDCL (*Soft Organization - weighted Fuzzy Description Comes Last*) em Nogueira (2013), porém esse método, assim como o SoftO-FDCL não se propõem a interpretar as tipicidades da partição possibilística de maneira diferenciada.

A adaptação sugerida da matriz de contingência está apresentada na Tabela 4.10, na qual os valores de ganhos, perdas, ruídos e rejeitos foram mapeados para as Equações (4.10), (4.11), (4.12) e (4.13) respectivamente. Portanto ao invés de realizar uma contagem discreta dos ganhos, perdas, ruídos e rejeitos, é realizado uma soma contínua das contribuições de cada documento ao grupo, considerando o valor de tipicidade. Desse

modo, conseguimos reduzir a contribuição de documentos com baixa tipicidade no grupo e aumentar a contribuição de documentos com alta tipicidade no grupo. Resultando assim, em uma pontuação mais justa e mais coerente dos termos extraídos dos grupos. Os ajustes necessários nas medidas de precisão e recuperação estão detalhados nas equações (4.14) e (4.15), enquanto o cálculo da medida $f1(t_k, g_j)$ permanece como definido na Equação (4.4). Destaca-se que nestas equações o limiar δ passa a ser verificado com os valores de tipicidades convertidos, ilustrados pela função $\lambda'(d_i, g_j)$ (Equação 4.8) e $\varphi(t_i, d)$ refere-se a frequência do termo t_i no documento d, de acordo com a Equação (2.1).

	Documentos do grupo g_j com grau de compatibilidade maior ou igual a δ	Documentos do grupo g_j com grau de compatibilidade menor do que δ
Documentos que possuem o descritor candidato t_k	$ganhos(t_k, g_j)$	$ruídos(t_k, g_j)$
Documentos que não possuem o descritor candidato t_k	$perdas(t_k, g_j)$	$rejeitos(t_k, g_j)$

Tabela 4.10: Matriz de contingência do termo t_k para o grupo g_j adaptada para a partição possibilística

$$ganhos(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \lambda'(d, g_j) \geq \delta, \varphi(t_i, d) > 0\} \tag{4.10}$$

$$perdas(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \lambda'(d, g_j) \geq \delta, \varphi(t_i, d) = 0\} \tag{4.11}$$

$$ruídos(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \lambda'(d, g_j) < \delta, \varphi(t_i, d) > 0\} \tag{4.12}$$

$$rejeitos(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \lambda'(d, g_j) < \delta, \varphi(t_i, d) = 0\} \tag{4.13}$$

$$precisão(t_k, g_j) = \frac{ganhos(t_k, g_j)}{ganhos(t_k, g_j) + ruídos(t_k, g_j)} \tag{4.14}$$

$$recuperação(t_k, g_j) = \frac{ganhos(t_k, g_j)}{ganhos(t_k, g_j) + perdas(t_k, g_j)} \tag{4.15}$$

Nas próximas duas sessões é caracterizado as duas abordagens que derivam dessa interpretação aqui detalhada. A primeira, é uma proposta de extensão do método SoftO-FDCL voltado para a partição possibilística do PCM, com a interpretação apresentada nessa sessão. Enquanto a segunda abordagem, apresenta uma estratégia de extração de descritores para o algoritmo PFCM, interpretando em conjunto as duas partições presentes no algoritmo.

4.4.3 O método PDCL

A agora que a proposta de interpretação da partição possibilística está concluída, vamos reunir todos esses passos em um método para extração de descritores para a partição possibilística, a qual será denominado aqui de PDCL (*Possibilistic Descriptor Comes Last*). Para realizar a extração de descritores o método PDCL, considera inicialmente todos os termos como candidatos, em seguida para cada grupo g_j é calculado a precisão (Equação 4.14) e recuperação (Equação 4.15) de todos os termos t_k. A partir então destes valores é calculada a pontuação de cada termo t_k no grupo g_j, com a medida $f1(t_k, g_j)$ (Equação 4.4). Obtido então essa pontuação por grupo dos termos candidatos, deve-se selecionar os m descritores de maior pontuação em cada grupo. No qual a quantidade m de descritores é definida pelo usuário. A síntese do método PDCL está apresentado no Algoritmo 6.

4.4.4 O método Mixed-PFDCL

Na discussão do experimento da sessão 4.3, onde é apresentado os resultados do refinamento com o método PFCM (*Possibilistic Fuzzy C Means*) na proposta de organização flexível de documentos, foi salientado a importância de se montar uma estratégia que conseguisse capturar a filosofia híbrida do algoritmo PFCM, levando em consideração suas duas partições.

Uma das características presentes no método PFCM, é a adição dos parâmetros a e b que atuam como reguladores da influência do FCM e do PCM no agrupamento obtido. Portanto há de destacar, a importância de se considerar tais parâmetros no processo de extração de descritores, objetivando assim mais coerência com o algoritmo.

Nesse contexto, é apresentado a Equação (4.16), a qual combina os valores de pertinência e de tipicidade convertidos, em um único grau de compatibilidade. Essa combinação nada mais é do que a média ponderada pelos parâmetros a e b dos valores de pertinência e tipicidade convertido respectivamente.

$$\mu'(d_i, g_j) = \frac{a\mu(d_i, g_j) + b\lambda'(d_i, g_j)}{a + b} \tag{4.16}$$

Como resultado dessa combinação precisamos apenas adaptar as medidas de ganhos, perdas, ruídos e rejeitos, para considerarem a relevância de um dado documento d em relação ao limiar δ, a partir da pertinência híbrida $\mu'(d_i, g_j)$. Essa adaptação está então disposta nas equações (4.17), (4.18), (4.19) e (4.20). Ressalta-se que foi mantido no somatório de ganhos, perdas, ruídos e rejeitos, o valor de tipicidade, devido a capacidade do mesmo em expressar melhor a realidade do agrupamento. As demais medidas de precisão, recuperação e $f1$, permanecem com a mesma definição apresentada anteriormente. Para uma melhor compreensão de como se encaixa todas essas partes, um pseudo código do método Mixed-PFDCL está apresentado no Algoritmo 6.

$$ganhos(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \mu'(d, g_j) \geq \delta, \varphi(t_i, d) > 0\} \tag{4.17}$$

$$perdas(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \mu'(d, g_j) \geq \delta, \varphi(t_i, d) = 0\} \quad (4.18)$$

$$ruídos(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \mu'(d, g_j) < \delta, \varphi(t_i, d) > 0\} \quad (4.19)$$

$$rejeitos(t_i, g_j) = \sum_{d \in D'} \lambda(d, g_j), D' = \{d | d \in D, \mu'(d, g_j) < \delta, \varphi(t_i, d) = 0\} \quad (4.20)$$

extrair-descritores(U, P, D, G, m)
início
descritores ← ∅;
para cada $g \in G$ **faça**
candidatos ← $[t_1, t_2, ..., t_k]$;
ranking ← ∅;
para cada $t \in$ *candidatos* **faça**
precisao ← Equação (4.2) para o PDCL ou (4.14) para o Mixed-PFDCL;
recuperacao ← Equação (4.3) para o PDCL ou (4.15) para o Mixed-PFDCL;
pontuacao ← Equação (4.4);
ranking[t] ← pontuacao;
fim
descritores[g] ← m termos do ranking com maior pontuacao;
fim
retorne (descritores);
fim

Algoritmo 6: Pseudo código da extração de descritores com os métodos PDCL e Mixed-PFDCL. Onde considere U a partição de pertinências fuzzy (Equação 2.7), P a partição possibilística (Equação 2.16), D a coleção de documentos da coleção, G os grupos produzidos pelo método de agrupamento e m a quantidade descritores desejada por grupo.

4.4.5 Resultados

Para mensurar os impactos das duas estratégias híbridas para extração de descritores apresentadas aqui nesta seção, foi realizado outro experimento, com os algoritmos PCM e PFCM, com as bases apresentadas na seção 4.2. Durante o experimento fora utilizado os métodos de extração de descritores SoftO-FDCL, PDCL, Mixed-PFDCL.

Desde que as propostas aqui apresentadas, pretendem otimizar os resultados apresentados anteriormente, foi adotada uma metodologia similar ao experimento anterior. Ou seja, o agrupamento final obtido para cada base, é resultado dos grupos que obtiveram o maior valor na medida de silhueta fuzzy. Onde a quantidade de grupos para cada base, variou entre 2 e o número de classes de cada base (Tabela 4.2). Ressalta-se ainda que para minimizar os efeitos da aleatoriedade da partição inicial nesse experimento, o agrupamento foi executado 5 vezes para cada quantidade de grupos na silhueta fuzzy.

Os parâmetros m e n que regulam a variação entre fuzzy e *crisp* das partições resultantes conforme ilustrados nas Figuras 2.1 e 2.2, foram definidos para 1.2 de maneira empírica. De maneira geral, observou-se que as bases de maior dimensionalidade, estavam produzindo grupos coincidentes mais facilmente, quando os valores de m e n eram maior do que 1.5. Por sua vez os parâmetros a e b do algoritmo PFCM, o qual define a influência das pertinências fuzzy do FCM e os graus de compatibilidade possibilísticos do PCM respectivamente, foram definidos como sendo $a = 1,0$ e $b = 1,2$. Essa escolha deriva das conclusões apresentadas em Pal et al. (2005), o qual nos informa que de acordo com a função objetivo (Equação 2.18) do PFCM, se utilizarmos valores de a maiores que b, os protótipos dos grupos são mais influenciados pelas pertinências. Contudo, como em coleções textuais naturalmente esparsas, é mais comum haver documentos ruidosos, o qual não se encaixe totalmente em nenhum grupo. Sendo assim, foi adotado o valor de b maior do que a, para reduzir os efeitos indesejados dos ruídos.

As coleções foram então agrupadas com os métodos PCM e PFCM, utilizando a metodologia descrita. A quantidade ótima de grupos, obtida com o método da silhueta fuzzy, está apresentado na Tabela 4.11. Os resultados apresentado nesta tabela, reforçam as conclusões apresentadas experimento anterior, de que a quantidade de grupos ótima do método PFCM tende a se aproximar mais da quantidade original de classes de cada coleção, enquanto o método PCM possui uma tendência em obter um número de grupos bem inferior a quantidade original de classes.

Coleção	**# classes**	**PCM**	**PFCM**
Opinosis	3	2	3
20Newsgroup	4	4	4
Hitech	6	2	6
NSF	16	2	8
WAP	20	2	17
Reuters-21578	43	4	40

Tabela 4.11: Quantidade ótima de grupos determinada através do método da silhueta fuzzy para cada algoritmo de agrupamento no segundo experimento conduzido com os métodos PCM e PFCM

Em seguida, foi realizado a extração dos descritores sobre as partições ótimas encontradas por cada algoritmo de agrupamento, sobre as coleções textuais. Como a motivação desse experimento, foi avaliar a qualidade dos descritores produzidos pelos métodos PDCL e Mixed-PFDCL propostos, a extração de descritores foi também realizada com o método SoftO-FDCL, possibilitando assim compararmos os resultados. E assim como no experimento anterior, essa análise quantitativa dos descritores produzidos, foi feita, utilizando a mesma estratégia de avaliação preditiva, com os 5 algoritmos de classificação do experimento anterior. No entanto, como a proposta de interpretação das duas partições produzidas pelo algoritmo PFCM, contempla um grau de compatibilidade misto entre as duas partições, foi pertinente generalizar essa interpretação também para função de defuzificação dos grupos, apresentada na Equação (4.1). Essa adaptação está apresentada

na Equação (4.21), onde foi adicionado, o grau de compatibilidade híbrido proposto na Equação (4.16).

$$crisp(d_i) = \begin{cases} crisp_j, & \mu(d_i, g_j) = \max\limits_{\forall g \in G} \mu(d_i, g), \text{se a partição for fuzzy} \\ crisp_j, & \lambda(d_i, g_j) = \max\limits_{\forall g \in G} \lambda(d_i, g), \text{se a partição for possibilística} \\ crisp_j, & \mu'(d_i, g_j) = \max\limits_{\forall g \in G} \mu'(d_i, g), \text{se houver duas partições} \end{cases} \quad (4.21)$$

Sendo assim, é realizada uma redução na matrix documentos x termos original para a matriz documentos x descritores, conforme descrito na seção 4.3, onde a pertinência de cada documento é submetida a uma defuzificação com a Equação (4.21). Posteriormente, essa matriz de dimensionalidade reduzida é submetida aos mesmos classificadores do experimento anterior, com os mesmos parâmetros. Os resultados dessa classificação estão apresentados nas Figuras 4.8, 4.9, 4.10, 4.11, 4.12 e 4.13.

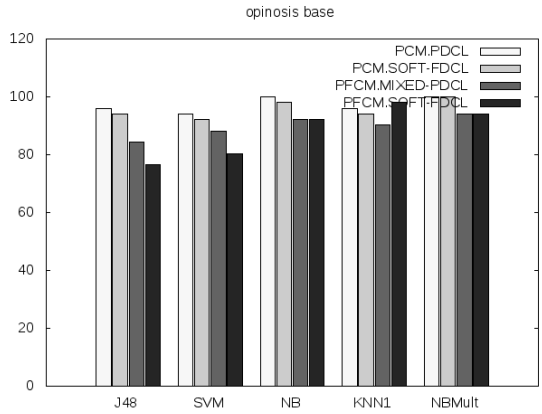

Figura 4.8: Desempenho obtido dos descritores extraídos com os algoritmos SoftO-FDCL, Mixed-PFDCL e PDCL sobre o agrupamento produzido pelos métodos PCM e PFCM na coleção Opinosis

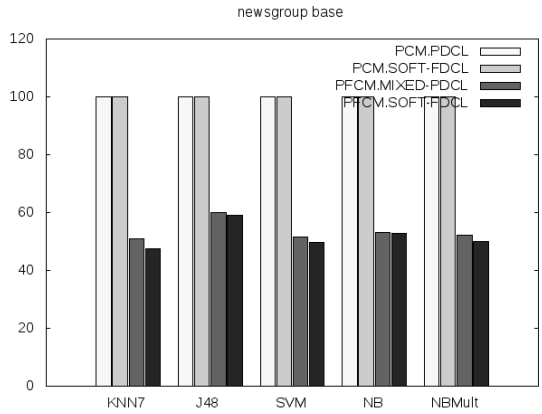

Figura 4.9: Desempenho obtido dos descritores extraídos com os algoritmos SoftO-FDCL, Mixed-PFDCL e PDCL sobre o agrupamento produzido pelos métodos PCM e PFCM na coleção 20Newsgroup

O sumário desses resultados consta na Tabela 4.12, onde o marcador (✓) denota que o método obteve maiores taxas de acerto entre os 5 algoritmos de classificação utilizados. Como nessa investigação o propósito foi comparar os métodos de extração de descritores, dividiu-se o sumário de resultados de acordo com o método de agrupamento, PCM e PFCM respectivamente. Ressalta-se ainda, que garantir que a extração fosse realizada sobre os mesmos grupos, ambos os métodos de extração de descritores foram aplicados simultaneamente ao agrupamento. Portanto, os métodos SoftO-FDCL e PDCL foram aplicados ao mesmo agrupamento produzido pelo PCM, enquanto o SoftO-FDCL e o Mixed-PFDCL foram executados no mesmo agrupamento gerado pelo PFCM.

Os resultados dispostos nesse sumário, corroboram a hipótese formulada a respeito da interpretação das partições possibilísticas e híbridas no contexto da extração de descri-

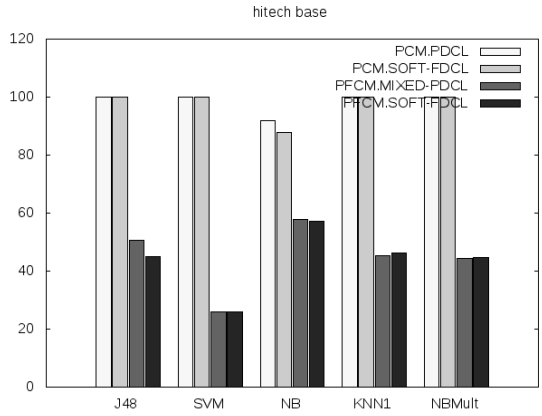

Figura 4.10: Desempenho obtido dos descritores extraídos com os algoritmos SoftO-FDCL, Mixed-PFDCL e PDCL sobre o agrupamento produzido pelos métodos PCM e PFCM na coleção Hitech

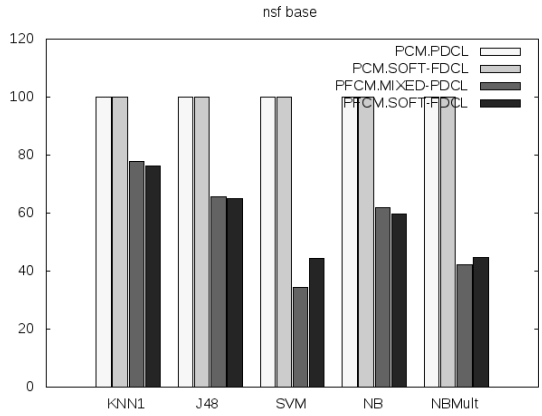

Figura 4.11: Desempenho obtido dos descritores extraídos com os algoritmos SoftO-FDCL, Mixed-PFDCL e PDCL sobre o agrupamento produzido pelos métodos PCM e PFCM na coleção NSF

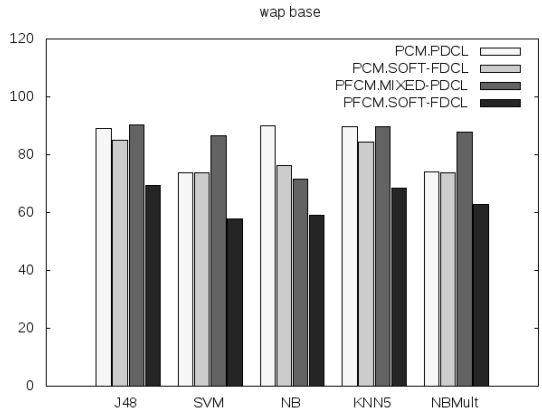

Figura 4.12: Desempenho obtido dos descritores extraídos com os algoritmos SoftO-FDCL, Mixed-PFDCL e PDCL sobre o agrupamento produzido pelos métodos PCM e PFCM na coleção WAP

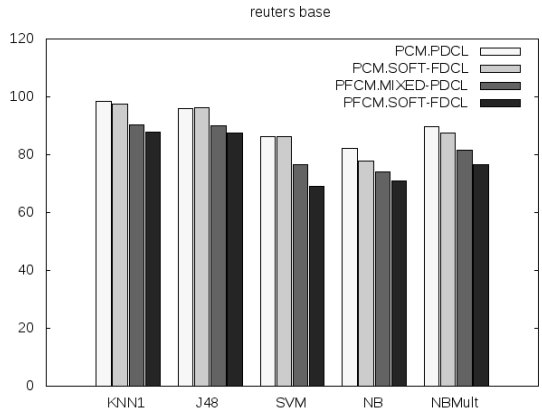

Figura 4.13: Desempenho obtido dos descritores extraídos com os algoritmos SoftO-FDCL, Mixed-PFDCL e PDCL sobre o agrupamento produzido pelos métodos PCM e PFCM na coleção Reuters-21578

tores para a organização flexível de documentos. Pois, como se observa na Tabela 4.12, o método PDCL e o Mixed-PFDCL, superam os resultados do método SoftO-FDCL, em ambos os algoritmos de agrupamento. Embora, tenha existido 2 empates na comparação entre os métodos SoftO-FDCL e PDCL, para as coleções 20newsgroups e NSF.

Ao se analisar os resultados, comparando as taxas de acerto obtidas entre os métodos de agrupamento, é possível se observar uma relevante tendência do PCM superar o PFCM. Entretanto, as taxas de acerto próximas de 100% no PCM, desperta atenção. Nesse sentido, vamos analisar na Tabela 4.13, os grupos majoritários da matriz documentos

	PCM		PFCM	
Coleção	SoftO-FDCL	PDCL	SoftO-FDCL	Mixed-PFDCL
Opinosis		✓		✓
20newsgroups	✓	✓		✓
Hitech		✓		✓
NSF	✓	✓		✓
WAP		✓		✓
Reuters-21578		✓		✓

Tabela 4.12: Sumário dos resultados da classificação dos descritores extraídos com os métodos SoftO-FDCL, PDCL e Mixed-PFDCL

x descritores, obtidos após a defuzificação do agrupamento com a Equação (4.21), para entender o por que de taxas tão elevadas.

	PCM	PFCM	
Coleção	SoftO-FDCL e PDCL	SoftO-FDCL	Mixed-PFDCL
Opinosis	84,31%	43,13%	43,13%
20newsgroups	99,9%	49,55%	51,65%
Hitech	95,16%	25,83%	25,83%
NSF	100%	44,43%	34,21%
WAP	93,3%	57,69%	86,53%
Reuters-21578	97,81%	68,91%	76,42%

Tabela 4.13: Informações das classes majoritárias obtidas através da defuzificação dos grupos fuzzy, com a Equação (4.21)

É demonstrado então na Tabela 4.13, que a defuzificação dos agrupamentos produzidos pelo método PCM em cada coleção, resultou em grupos majoritários próximo de 100%, o que por sua vez facilita a tarefa do classificador. Desse modo, fica claro as causas de taxas de acertos tão elevadas no método PCM, e assim como também nos mostra a importância de não analisarmos os resultados da classificação isoladamente. Por outro lado, percebe-se na Tabela 4.13, a discrepância dos grupos majoritários no algoritmo PFCM em comparação ao PCM. Isso por sua vez, indica que o agrupamento produzido pelo PFCM, gerou grupos mais coerentes, o que reforça a adequação desse algoritmo para coleções textuais.

Nos estudos do limiar δ, detalhados na seção 4.4.1, foi pontuado que ao se utilizar esse limiar em partições possibilísticas, poderia resultar em listas similares de descritores entre os grupos. Portanto, para comprovar essas conclusões obtidas dessa investigação, é pertinente analisar as listas dos descritores com maior pontuação em cada grupo, produzidas pelos métodos Soft-FDCL, PDCL e Mixed-PFDCL nesse experimento. Contudo, como são muitas bases, e em alguns casos a quantidade de grupos é demasiadamente elevada, está apresentado nas Tabelas 4.14 e 4.15, somente as 5 maiores pontuações dos termos candidatos aos grupos obtidos pelos algoritmos PCM e PFCM na base Opinosis.

Assim como no experimento anterior, foi escolhido apresentar essa base, para facilitar a interpretação dos resultados, pois a mesma contém uma quantidade baixa de documentos e de grupos resultantes do agrupamento.

	$grupo_1$		$grupo_2$	
método	termo	pontuação	termo	pontuação
SoftO-FDCL	caf	0,923077	caf	0,923077
	floor	0.888889	floor	0.888889
	food	0.880000	food	0.880000
	coffe	0.857143	coffe	0.857143
	concierge	0.846154	concierge	0.846154
PDCL	bathro	0.894716	make	0.800980
	food	0.888785	time	0.789846
	concierge	0.860127	nice	0.779338
	supermarket	0.856632	feature	0.778138
	chain	0.856632	easy	0.768564

Tabela 4.14: Lista ordenada com 5 termos candidatos de maior pontuação, obtidos após a extração de descritores com os métodos Soft-FDCL e PDCL aplicados ao agrupamento da coleção Opinosis com o algoritmo PCM

Portanto, ao se analisar na Tabela 4.14, as pontuações e os termos do $grupo_1$ e do $grupo_2$, obtidos com o método SoftO-FDCL, nota-se uma repetição dos termos e pontuações para ambos os grupos. Essa evidência, por usa vez, fortalece a problemática na interpretação dos graus de compatibilidade possibilísticos apresentada na seção 4.4.1. Por outro lado, os termos obtidos com o método proposto, não se repetem em ambos os grupos, assim como também estão mais coerentes entre si. Conforme foi detalhado, a base Opinosis é composta de uma série de opiniões sobre eletrônicos, hotéis e carros, sendo assim é possível observar nos termos encontrados pelo método PDCL, uma proximidade semântica com esses assuntos.

Já Tabela 4.15, a qual contém os 5 termos candidatos de maior pontuação no $grupo_1$, $grupo_2$ e $grupo_3$, não apresentou o problema da repetição dos mesmos termos nos grupos para o método SoftO-FDCL. Isto se deve por conta da extração de descritores com o método SoftO-FDCL nesse experimento, utilizar somente os graus de pertinências, os são adequados a este método, conforme foi destacado durante a investigação desse método. Contudo, no experimento anterior, foi levantado a importância de se utilizar as pertinências e tipicidades presentes no algoritmo PFCM, para melhor interpretar o agrupamento produzido por essa abordagem híbrida. Nesse contexto, observa-se que os termos de maior pontuação do método Mixed-PFDCL, diferem parcialmente dos termos obtidos com o método SoftO-FDCL, porém há de se observar que os termos dos grupos SoftO-FDCL podem ser semanticamente relacionados aos termos do método Mixed-PFDCL. Ainda é possível notar, que os 3 grupos de termos são similares aos 3 tópicos presentes na coleção Opinosis. O que por sua vez, reforça a adequadação do método proposto para interpretar de maneira híbrida ambas as partições geradas no pelo método PFCM.

	$grupo_1$		$grupo_2$		$grupo_3$	
método	termo	pontuação	termo	pontuação	termo	pontuação
SoftO-FDCL	text	0.850000	coffe	0.933333	ga	0.833333
	featur	0.758621	eat	0.903226	camr	0.818182
	device	0.750000	floor	0.896552	transmission	0.818182
	user	0.750000	inn	0.896552	seat	0.769231
	unit	0.727273	lobby	0.896552	rir	0.769231
Mixed-PFDCL	featur	0.876540	friendl	0.951225	fun	0.891245
	text	0.869230	bed	0.950665	ga	0.868032
	find	0.830332	coffe	0.948341	rir	0.847864
	isnt	0.817311	stay	0.945294	seat	0.843576
	easy	0.809918	night	0.933519	engine	0.840700

Tabela 4.15: Lista ordenada com 5 termos candidatos de maior pontuação, obtidos após a extração de descritores com os métodos Soft-FDCL e PDCL aplicados ao agrupamento da coleção Opinosis com o algoritmo PFCM

Na próxima seção será apresentado uma breve sintése dos experimentos conduzidos nesse capítulo, pontuando as principais contribuições e descobertas encontradas ao longo desse processo de investigação.

4.5 CONSIDERAÇÕES FINAIS

Neste capítulo foi detalhado todo o processo de investigação realizado nessa monografia, assim como a metodologia utilizada para a realização dos experimentos, informações das coleções textuais utilizadas e as devidas análises dos resultados. A abordagem híbrida proposta por essa monografia constitui a adição da robustez proporcionada pelo algoritmo PFCM o qual é uma versão híbrida dos algoritmos FCM e PCM. Durante esse processo, foi identificado no primeiro experimento, que era preciso propor uma maneira de melhor interpretar as tipicidades do algoritmo PCM. Outra conclusão obtida do primeiro experimento, foi também a necessidade de realizar a extração de descritores dos grupos produzidos pelo método PFCM utilizando as duas partições. Nesse sentido, foi aqui apresentada uma abordagem para corretamente se interpretar as tipicidades, assim como também uma estratégia híbrida para se realizar a extração de descritores do método PFCM.

No próximo capítulo, será abordada as conclusões a respeito de toda a pesquisa apresentada nessa monografia.

Capítulo

5

Síntese da investigação e dos experimentos realizados nesta monografia

CONCLUSÃO

A pesquisa desenvolvida nesta monografia investiga e elabora uma solução voltada para organização de uma coleção de documentos textuais de maneira flexível. O processo de organização em si, conforme discutido ao longo dos capítulos, envolve uma gama de desafios para ser executado com eficiência, como, por exemplo, os impactos negativos da elevada dimensionalidade da matriz documentos x termos, que é comumente muito esparsa em coleções textuais. Por conta disso, a tarefa de se calcular a similaridade entre dois documentos quaisquer, a partir daquela mesma matriz, apresenta alto grau de dificuldade. Outra problemática fundamental está no processo de agrupamento, através do qual se espera que os grupos resultantes consigam capturar a estrutura natural das coleções e que esses grupos possuam relevância para os usuários finais. Atendendo a esses princípios, o agrupamento consegue cumprir o papel de aquisição e descoberta de conhecimento. Coleções textuais podem também conter documentos ruidosos, que destoam do restante da coleção, sendo que é esperado o emprego de técnicas no processo de organização flexível para que ele não seja prejudicado pela presença desses documentos. Com o aumento massivo da quantidade de dados produzidos pela humanidade, se faz também necessário que todo o processo seja capaz de se adequar a coleções com grande volumes de dados. Assim, a organização flexível de documentos proposta fundamenta-se em um conhecimento bastante intensivo, por meio do qual se aplica um conjunto de diversas técnicas os desafios apontados acima bem como outros associados a esse processamento de textos.

Todo esse contexto apresentado se mostra inviável de ser abordado de maneira aprofundada em uma única pesquisa. Por isso, as investigações conduzidas nessa monografia foram focadas no aumento da robustez do processo, reduzindo-se os impactos dos dados ruidosos ao se utilizar uma estratégia híbrida com o algoritmo PFCM. Portanto, a hipótese formulada e verificada nesta monografia foi:

> *A utilização de uma estratégia híbrida de agrupamento e extração de descritores, entre os graus de pertinência e tipicidade providos pelo método de agrupamento PFCM, permitem o aumento da robustez e resiliência contra ruídos*

na organização flexível de documentos, aumentando assim a relevância dos grupos obtidos.

Portanto, com base na exploração das estratégias existentes na literatura para aprimoramento do processo de organização flexível de documentos e da avaliação da hipótese formulada, o objetivo desta monografia é definido como segue:

Conduzir uma investigação em torno dos métodos de agrupamento FCM, PCM e PFCM, para se compreender e interpretar corretamente as peculiaridades de se extrair descritores em um agrupamento híbrido.

A fim de atender a esse objetivo, foi realizado nesta monografia um estudo dos fundamentos necessários para a organização flexível de documentos, uma revisão das estratégias recentes utilizadas por pesquisadores para aprimorar a organização flexível de documentos. E, por fim, como resultado, o estudo dos impactos de se utilizar o algoritmo PFCM no processo e na extração de descritores, o qual derivaram a proposição de dois métodos de extração de descritores: PDCL e Mixed-PFDCL. Tais métodos são extensões do método SoftO-FDCL apresentado em (Nogueira, 2013), e diante dos resultados, contribuem de maneira significativa para o estado da arte da extração de descritores dos grupos fuzzy.

Na seção seguinte, apresentam-se as principais contribuições fornecidas por esta monografia e os possíveis trabalhos futuros que derivam desta pesquisa.

5.1 RESUMO DAS CONTRIBUIÇÕES

A investigação realizada nesta monografia, produziu algumas importantes contribuições para o processo de se organizar documentos de maneira flexível. Essas contribuições são distribuídas no estudo e fundamentação da teoria relacionada a esse campo de estudo, investigação e apresentação das possibilidades de aprimoramento e otimização do processo explorados recentemente na literatura e a investigação das peculiaridades de adotar uma estratégia híbrida o que posteriormente motivou a proposição de dois métodos de extração de descritores.

A partir do estudo da teoria e dos fundamentos necessários para o tema, é apresentado nesta monografia um rico conteúdo abordando os detalhes dos métodos de agrupamento FCM, PCM, PFCM e HFCM, juntamente com os seus pseudo-códigos, para auxiliar novos pesquisadores na rápida implementação de tais métodos.

Outra importante contribuição, consiste na apresentação das pesquisas que se situam no estado da arte das etapas presentes no processo de organização flexível de documentos. Ficou evidenciado, a partir das pesquisas encontradas na literatura, a diversidade as estratégias existentes na literatura para mitigar os desafios existentes. Foi visto que ainda é proposto abordagens para aprimorar todas as etapas existentes no processo, as quais contemplam o pré-processamento, agrupamento, extração de descritores e recuperação da informação, com isso conclui-se que a organização flexível não é um problema resolvido na literatura, o que o torna um problema de pesquisa bastante promissor.

A primeira contribuição experimental dessa monografia, consistiu no estudo dos impactos de se adicionar o método PFCM no processo de organização flexível de documentos.

A partir desse estudo foi observado que o algoritmo PFCM possui uma tendência para aumentar a eficiência do agrupamento produzido em coleções textuais de maior dimensionalidade, o que foi comprovado a partir das evidências contidas na Tabela 4.4. No entanto, se destaca aqui a importância da realização de novos estudos com um maior número de coleções textuais de baixa e alta dimensionalidade para se confirmar esta tendência.

Por outro lado, constatou-se nesse experimento a capacidade de adaptação do método SoftO-FDCL a novos algoritmos de agrupamento. No entanto, este último método, segundo as suas pesquisas iniciais, considera somente uma única partição no processo de pontuação dos termos candidatos, o que por sua vez não consegue capturar toda a essência do agrupamento produzido pelo PFCM. Ainda nesse experimento foi observado um problema na interpretação das tipicidades contidas em partições possibilísticas, no processo de extração de descritores. Esse problema deriva diretamente da natureza probabilística dos graus de pertinência da partição fuzzy do FCM, que não se aplica às tipicidades, a qual influencia direta adequação ou não do limiar δ do método SoftO-FDCL.

Portanto, a partir da identificação desse problema de interpretação dos graus de compatibilidade possibilísticos, uma outra importante contribuição dessa monografia, consiste na formulação das propriedades do limiar apresentadas nas equações 4.5, 4.6 e 4.7. Essas propriedades expressam características importantes, que os graus de compatibilidade devem possuir, para o limiar δ do método do método SoftO-FDCL conseguir extrair bons descritores dos grupos.

Feitas essas considerações, apresentam-se as duas principais contribuições dessa monografia. A primeira consiste no método PDCL, que propõe uma abordagem para interpretar os graus de compatibilidade possibilísticos, respeitando as propriedades das equações 4.5, 4.6 e 4.7, sem deixar de lado a resiliência contra ruídos inerente das tipicidades. Os experimentos conduzidos com esse método demonstraram a qualidade dos descritores extraídos com o PDCL em comparação com o método SoftO-FDCL, cujos comparativos estão apresentados na Tabela 4.14. No entanto, observou-se neste experimento que o método PCM produziu uma quantidade baixa de grupos em comparação ao número total de classes em cada coleção, assim como também a defuzificação resultou em grupos majoritários com elevado percentual. Com isso, é importante salientar, a necessidade de se conduzir estudos experimentais a respeito dos parâmetros ideais dos métodos de agrupamento para coleções textuais.

A segunda grande contribuição desta monografia consiste na proposta de um método de extração de descritores híbrido, capaz de interpretar as duas partições presentes no algoritmo PFCM de maneira bastante satisfatória. Um segundo experimento foi então conduzido para atestar os benefícios desse abordagem híbrida de extração de descritores. Os resultados apresentados no sumário da Tabela 4.12, atestam através de uma avaliação preditiva com clássicos algoritmos de classificação, que os descritores extraídos através do método Mixed-PFDCL no agrupamento resultante do método PFCM, foram melhores que os descritores do método SoftO-FDCL. Ainda na Tabela 4.15, se demonstrou que os descritores obtidos com o método Mixed-PFDCL preservam uma proximidade semântica, com os demais termos do mesmo grupo, assim como os descritores do método SoftO-FDCL.

Durante as pesquisas realizadas neste projeto de conclusão de curso, houve a publicação de um artigo com os resultados obtidos no primeiro experimento, em uma das mais importantes conferências de sistemas fuzzy. Isso reforça a importância e relevância do tema estudado nesta monografia para a comunidade científica. Segue o artigo publicado:

CARVALHO, N. V. J. et al. Flexible Document Organization by Mixing Fuzzy and Possibilistic Clustering algorithms. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, p. 1–8, 2016.

5.2 TRABALHOS FUTUROS

Ao considerar todas as conclusões apresentadas anteriormente, existem diversas possibilidades que podem derivar deste trabalho. Em particular, foi notado que é necessário se realizar uma avaliação mais apurada dos impactos na variação dos parâmetros dos algoritmos de agrupamento, de modo que se possa obter, senão uma forma automatizada dos parâmetros com base nas características das bases, pelo menos indicações de quais parâmetros utilizar.

Nesse sentido, está apresentado na Figura 5.1, um exemplo dos impactos da variação da quantidade de grupos e do parâmetro m, sobre a medida de silhueta fuzzy. Ao analisar o gráfico, observa-se uma clara tendência da medida de silhueta fuzzy (FS) aumentar, com o incremento no número de grupos e para a faixa de valores de m entre 1,0 e 1,4. Contudo, esse tipo de análise requer sucessivas execuções de todo o agrupamento, o que por sua vez é um processo bastante oneroso computacionalmente. Sendo assim, pretende-se estender esse estudo para outras bases e para os demais algoritmos de agrupamento, assim como mensurar a qualidade do agrupamento produzido através de outras medidas de validação existentes na literatura, como, por exemplo, as medidas de entropia e pureza do agrupamento apresentadas em Deepa e Revathy (2012) ou a medida de validação proposta em Zhang et al. (2008), que explora as pertinências e tipiciadades existentes no PFCM.

Nesta monografia, também foi observado diversas vezes que um dos grandes desafios no agrupamento de documentos está na dimensionalidade dos dados. Soma-se a isso o fato de que os principais algoritmos de agrupamento possuem como ponto crítico de decisão no processo de separação dos elementos em grupos, a medida de similaridade. Sendo assim, é importante se investigar também as possibilidades de aprimoramento da organização flexível de documentos com as recentes medidas de similaridade propostas na literatura (Lin et al., 2014; Nagwani, 2015) para dados de alta dimensionalidade e em particular para coleções textuais.

Outro aspecto de grande relevância é a escalabilidade da proposta de organização flexível de documentos, para coleções textuais que se enquadram na categoria *Very Large* de acordo com a Tabela 3.1. Portanto, se pretende realizar estudos objetivando escalar o processo utilizado nesta pesquisa para o cenário de *Big Data*, utilizando como base as pesquisas com grandes quantidades de dados apresentadas em Honda et al. (2014), Wang et al. (2014) e Havens et al. (2012). De antemão, postula-se que, no que diz respeito a implementação dos algoritmos aqui utilizados, todo o processo foi paralelizado utilizando

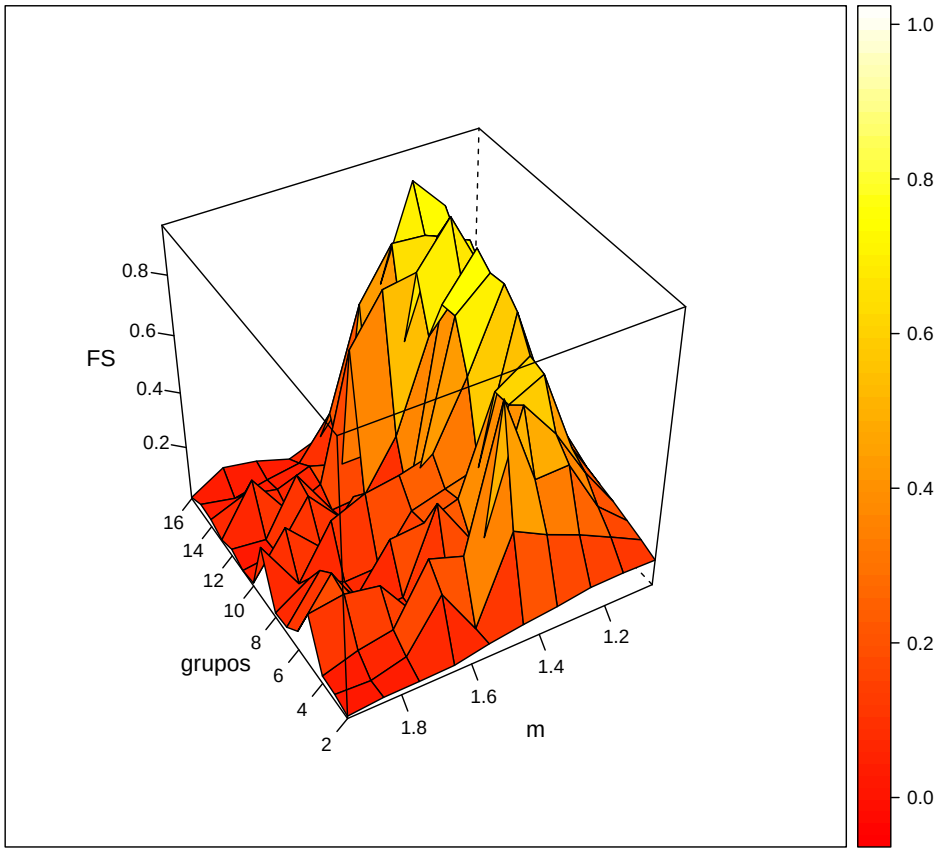

Figura 5.1: Gráfico das influências da variação da quantidade de grupos e do parâmetro m, na pontuação obtida pela medida de silhueta fuzzy para o algoritmo PFCM na base NSF

o framework de computação paralela OpenMP (*Open Multi-Processing*)[1], o que habilita a realização de experimentos em computadores com elevado potencial de processamento.

[1] ⟨http://openmp.org/⟩

REFERÊNCIAS BIBLIOGRÁFICAS

AGGARWAL, C. C.; ZHAI, C. An introduction to text mining. In: *Mining Text Data*. Springer Science + Business Media, 2012. p. 1–10. Disponível em: ⟨http://dx.doi.org/10.1007/978-1-4614-3223-4_1⟩.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern information retrieval: the concepts and technology behind search. *Choice Reviews Online*, American Library Association, v. 48, n. 12, p. 48–6950–48–6950, aug 2011. Disponível em: ⟨http://dx.doi.org/10.5860/choice.48-6950⟩.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, v. 10, n. 2, p. 191 – 203, 1984. ISSN 0098-3004. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/0098300484900207⟩.

BOUGHANEM, M.; PRADE, H.; BOUIDGHAGHEN, O. Extracting topics in texts: Towards a fuzzy logic approach. In: *Proceedings of the Information Processing and Management of Uncertainty (IPMU)*. [S.l.: s.n.], 2008. p. 1733–1740.

CARVALHO, N. V. J. et al. Flexible Document Organization by Mixing Fuzzy and Possibilistic Clustering algorithms. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, p. 1–8, 2016.

CHEN, G. *Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems*. Hoboken, NJ: CRC Press, 2000. Disponível em: ⟨https://cds.cern.ch/record/1250131⟩.

CUNHA, A. G.; TAKAHASHI, R.; ANTUNES, C. H. *Manual de computação evolutiva e metaheurística*. Coimbra University Press, 2012. Disponível em: ⟨http://dx.doi.org/10.14195/978-989-26-0583-8⟩.

DEEPA, M.; REVATHY, P. Validation of document clustering based on purity and entropy measures. *IJARCCE International Journal of Advanced Research in Computer and Communication Engineering*, v. 1, may 2012. Disponível em: ⟨http://ijarcce.com/upload/may/Validation%20of%20Document%20Clustering%20based%20on%20Purity%20and%20Entropy%20measures.pdf⟩.

DENG, J. et al. An improved fuzzy clustering method for text mining. In: *The 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing (NSWCTC), 2010*. [S.l.: s.n.], 2010. v. 1, p. 65–69.

FELDMAN, R.; SANGER, J. *The text mining handbook: Advanced approaches in analyzing unstructured data*. [S.l.]: Cambridge University Press, 2007.

FRANK, A.; ASUNCION, A. *UCI Machine Learning Repository*. 2010. Disponível em: ⟨[http://archive.ics.uci.edu/ml]⟩.

GROVER, N. A study of various fuzzy clustering algorithms. *International Journal of Engineering Research*, v. 3, p. 177–181, 2014.

HADDI, E.; LIU, X.; SHI, Y. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, v. 17, p. 26 – 32, 2013. ISSN 1877-0509. First International Conference on Information Technology and Quantitative Management. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S1877050913001385⟩.

HALL, M. et al. The WEKA data mining software: an update. *SIGKDD Explorations*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009.

HAVENS, T. et al. Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems*, v. 20, n. 6, p. 1130–1146, 2012.

HAYES, P. J.; WEINSTEIN, S. P. Construe/TIS: A system for content-based indexing of a database of news stories. In: *2nd Annual Conference on Innovative Applications of Artificial Intelligence*. [S.l.: s.n.], 1990. p. 1–5.

HONDA, K.; TANAKA, D.; NOTSU, A. Incremental algorithms for fuzzy co-clustering of very large cooccurrence matrix. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.: s.n.], 2014. p. 2494–2499.

HUANG, G. et al. A data as a product model for future consumption of big stream data in clouds. In: *2015 IEEE International Conference on Services Computing*. Institute of Electrical & Electronics Engineers (IEEE), 2015. Disponível em: ⟨http://dx.doi.org/10.1109/SCC.2015.43⟩.

JIANG, H. et al. An improved method of fuzzy clustering algorithm and its application in text clustering. *JOURNAL OF INFORMATION & COMPUTATIONAL SCIENCE*, JOURNAL OF INFORMATION & COMPUTATIONAL SCIENCE, v. 10, n. 2, p. 519, 2013. Disponível em: ⟨http://manu35.magtech.com.cn/Jwk_ics/EN/abstract/article_1507.shtml⟩.

KARAMI, A. et al. FLATM: A fuzzy logic approach topic model for medical documents. In: *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*. Institute of Electrical & Electronics Engineers (IEEE), 2015. Disponível em: ⟨http://dx.doi.org/10.1109/NAFIPS-WConSC.2015.7284190⟩.

KARYPIS, G. *Cluto - software for clustering high-dimensional datasets*. 2015. ⟨http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download.⟩

KOBAYASHI, M.; AONO, M. Vector space models for search and cluster mining. In: *Survey of Text Mining II*. Springer Science + Business Media, 2008. p. 109–127. Disponível em: ⟨http://dx.doi.org/10.1007/978-1-84800-046-9_6⟩.

KRISHNAPURAM, R.; KELLER, J. M. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, v. 1, n. 2, p. 98–110, 1993. ISSN 1063-6706.

KUMAR, D. et al. A hybrid approach to clustering in big data. *IEEE Transactions on Cybernetics*, Institute of Electrical & Electronics Engineers (IEEE), p. 1–1, 2015. Disponível em: ⟨http://dx.doi.org/10.1109/TCYB.2015.2477416⟩.

LANG, K. NewsWeeder: Learning to filter netnews. In: *Machine Learning Proceedings 1995*. Elsevier BV, 1995. p. 331–339. Disponível em: ⟨http://dx.doi.org/10.1016/B978-1-55860-377-6.50048-7⟩.

LIN, Y.-S.; JIANG, J.-Y.; LEE, S.-J. A similarity measure for text classification and clustering. *IEEE Trans. Knowl. Data Eng.*, Institute of Electrical & Electronics Engineers (IEEE), v. 26, n. 7, p. 1575–1590, jul 2014. Disponível em: ⟨http://dx.doi.org/10.1109/TKDE.2013.19⟩.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MARCACINI, R. M.; REZENDE, S. O. Incremental construction of topic hierarchies using hierarchical term clustering. In: *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE'2010), Redwood City, San Francisco Bay, CA, USA, July 1 - July 3, 2010*. [S.l.]: Knowledge Systems Institute Graduate School, 2010. p. 553. ISBN 1-891706-26-8.

MATSUMOTO, T.; HUNG, E. Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation. In: *FUZZ-IEEE*. IEEE, 2010. p. 1–8. ISBN 978-1-4244-6919-2. Disponível em: ⟨http://dblp.uni-trier.de/db/conf/fuzzIEEE/fuzzIEEE2010.html#MatsumotoH10⟩.

MEI, J.-P. et al. Incremental fuzzy clustering for document categorization. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.: s.n.], 2014. p. 1518–1525.

MOORE, J. et al. Web page categorization and feature selection using association rule and principal component clustering. *IBM shared research report/University of Minnesota (Minneapolis, Mn.)*, Citeseer, v. 98, p. 3, 1997.

MUGGLETON, S. H. 2020 computing: Exceeding human limits. *Nature*, Nature Publishing Group, v. 440, n. 7083, p. 409–410, mar 2006. Disponível em: ⟨http://dx.doi.org/10.1038/440409a⟩.

MURALI, D.; DAMODARAM, A. Semantic document retrieval system using fuzzy clustering and reformulated query. In: *2015 International Conference on Advances in Computer Engineering and Applications*. Institute of Electrical & Electronics Engineers (IEEE), 2015. Disponível em: ⟨http://dx.doi.org/10.1109/ICACEA.2015.7164788⟩.

NAGWANI, N. K. A comment on "a similarity measure for text classification and clustering". *IEEE Trans. Knowl. Data Eng.*, Institute of Electrical & Electronics Engineers (IEEE), v. 27, n. 9, p. 2589–2590, sep 2015. Disponível em: ⟨http://dx.doi.org/10.1109/TKDE.2015.2451616⟩.

NOGUEIRA, T. M. *Organização Flexível de Documentos*. Tese (Doutorado) — ICMC-USP, 2013.

NOGUEIRA, T. M.; CAMARGO, H. A.; REZENDE, S. O. Flexible document organization: Comparing fuzzy and possibilistic approaches. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.: s.n.], 2015. p. 1–8.

NOGUEIRA, T. M.; REZENDE, S. O.; CAMARGO, H. A. Fuzzy cluster descriptors improve flexible organization of documents. In: *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*. Institute of Electrical & Electronics Engineers (IEEE), 2012. Disponível em: ⟨http://dx.doi.org/10.1109/ISDA.2012.6416608⟩.

PAL, N. R. et al. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, IEEE Press, v. 13, n. 4, p. 517–530, 2005. ISSN 1063-6706.

PEDRYCZ, A.; REFORMAT, M. Hierarchical FCM in a stepwise discovery of structure in data. *Soft Comput.*, v. 10, n. 3, p. 244–256, 2006. Disponível em: ⟨http://dx.doi.org/10.1007/s00500-005-0478-8⟩.

POPESCU, M. et al. Random projections fuzzy c-means (rpfcm) for big data clustering. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.: s.n.], 2015. p. 1–6.

RIOS, A. R.; MELLO, F. R. A systematic literature review on decomposition approaches to estimate time series components. *Journal of Computer Science*, 2010.

ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. *ICMC TECHNICAL REPORT - Benchmarking Text Collections for Classification and Clustering Tasks*. São Carlos, SP, Brasil, 2013. Disponível em: ⟨http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_395.pdf⟩.

SARACOGLU, R.; TUTUNCU, K.; ALLAHVERDI, N. A new approach on search for similar documents with multiple categories using fuzzy clustering. *Expert Systems with Applications*, v. 34, p. 2545–2554, 2008.

SARANYA, J.; ARUNPRIYA, C. Survey on clustering algorithms for sentence level text. *International Journal of Computer Trends and Technology*, Seventh Sense Research Group Journals, v. 10, n. 2, p. 61–66, apr 2014. Disponível em: ⟨http://dx.doi.org/10.14445/22312803/IJCTT-V10P111⟩.

STEINBACH, M.; ERTöZ, L.; KUMAR, V. The challenges of clustering high-dimensional data. In: *In New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics, and Pattern Recognition*. [S.l.]: Springer-Verlag, 2003. ISBN 978-3-642-07739-5.

SUBHASHINI, R.; KUMAR, V. Evaluating the performance of similarity measures used in document clustering and information retrieval. In: *First International Conference on Integrated Intelligent Computing (ICIIC)*. [S.l.: s.n.], 2010. p. 27–31.

TJHI, W.-C.; CHEN, L. A heuristic-based fuzzy co-clustering algorithm for categorization of high-dimensional data. *Fuzzy Sets and Systems*, Elsevier BV, v. 159, n. 4, p. 371–389, feb 2008. Disponível em: ⟨http://dx.doi.org/10.1016/j.fss.2007.10.003⟩.

TJHI, W.-C.; CHEN, L. Dual fuzzy-possibilistic coclustering for categorization of documents. *IEEE Transactions on Fuzzy Systems*, v. 17, n. 3, p. 532–543, 2009. ISSN 1063-6706.

TREERATPITUK, P.; CALLAN, J. Automatically labeling hierarchical clusters. In: FORTES, J. A. B.; MACINTOSH, A. (Ed.). *DG.O*. Digital Government Research Center, 2006. (ACM International Conference Proceeding Series, v. 151), p. 167–176. Disponível em: ⟨http://dblp.uni-trier.de/db/conf/dgo/dgo2006.html#TreeratpitukC06⟩.

WANG, Y.; CHEN, L.; MEI, J.-P. Stochastic gradient descent based fuzzy clustering for large data. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.: s.n.], 2014. p. 2511–2518.

YAN, Y.; CHEN, L. Hyperspherical possibilistic fuzzy c-means for high-dimensional data clustering. In: *Proceedings of the 7th International Conference on Information, Communications and Signal Processing*. [S.l.: s.n.], 2009. (ICICS'09), p. 637–641. ISBN 978-1-4244-4656-8.

YAN, Y.; CHEN, L.; TJHI, W.-C. Fuzzy semi-supervised co-clustering for text documents. *Fuzzy Sets and Systems*, Elsevier BV, v. 215, p. 74–89, mar 2013. Disponível em: ⟨http://dx.doi.org/10.1016/j.fss.2012.10.016⟩.

ZADEH, L. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338 – 353, 1965. ISSN 0019-9958. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S001999586590241X⟩.

ZHANG, C.; XU, H. Clustering description extraction based on statistical machine learning. *Intelligent Information Technology Applications, 2007 Workshop on*, IEEE Computer Society, Los Alamitos, CA, USA, v. 2, p. 22–26, 2008.

ZHANG, C.; ZHOU, Y.; MARTIN, T. A validity index for fuzzy and possibilistic c-means algorithm. In: *Proc. IPMU*. [S.l.: s.n.], 2008. p. 877–882.