UNIVERSIDADE FEDERAL DA BAHIA

TRABALHO DE GRADUAÇÃO

Uma abordagem fuzzy híbrida para organização flexível de documentos, utilizando os algoritmos de agrupamento possibilístico e fuzzy c means

Nilton Vasques Carvalho Junior

Programa de Graduação em Ciência da Computação

Salvador
1 de junho de 2016

NILTON VASQUES CARVALHO JUNIOR

UMA ABORDAGEM FUZZY HÍBRIDA PARA ORGANIZAÇÃO FLEXÍVEL DE DOCUMENTOS, UTILIZANDO OS ALGORITMOS DE AGRUPAMENTO POSSIBILÍSTICO E FUZZY C MEANS

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
1 de junho de 2016

Ficha catalográfica.

Carvalho, Nilton Vasques Jr.

Uma abordagem fuzzy híbrida para organização flexível de documentos, utilizando os algoritmos de agrupamento possibilístico e fuzzy c means / Nilton Vasques Carvalho Junior– Salvador, 1 de junho de 2016.

27p.: il.

Orientadora: Profa. Dra. Tatiane Nogueira Rios.
Monografia (Graduação)– UNIVERSIDADE FEDERAL DA BAHIA, INSTITUTO DE MATEMÁTICA, 1 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 FUZZY HÍBRIDA PARA ORGANIZAÇÃO FLEXÍVEL DE DOCUMENTOS, UTILIZANDO OS ALGORITMOS DE AGRUPAMENTO POSSIBILÍSTICO E FUZZY C MEANS

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, DIA de MES de ANO

Profa. Dra. Tatiane Nogueira Rios
Universidade Federal da Bahia

Prof. Dr. Professor 2
Universidade 123

Profa. Dra. Professora 3
Universidade ABC

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

Diante da grande quantidade de informações geradas e armazenadas pela humanidade na atualidade, vários métodos foram propostos visando processar esses dados. Dentre esses dados, temos uma imensa quantidade de dados textuais, que por sua vez são não estruturados. Com isso é notória a importância, de organizar de maneira automatizada, esses documentos pelos assuntos ao qual se tratam. Em particular temos um conjunto de técnicas pertencentes ao campo de estudo da mineração de textos, que visam realizar a tarefa de extrair informações relevantes de documentos textuais. Esta tarefa de análise e extração de informações é comumente segmentada nas tarefas de coleta, pré-processamento dos documentos, agrupamento dos dados e por fim a extração de descritores dos grupos obtidos na etapa de agrupamento. Os métodos de agrupamento podem ser separados então pela lógica matemática utilizada, que pode ser a lógica clássica ou a lógica fuzzy. Na lógica clássica, após o agrupamento, cada documento só poderá pertencer a um grupo, enquanto na lógica fuzzy, a pertinência do documento será distribuída entre os grupos. Se analisarmos a diversidade de conteúdo em documentos textuais, é trivial notar que frequentemente um texto aborda um ou mais temas. Com isso é evidente a necessidade de desenvolver-se técnicas para organizar de maneira flexível os documentos. Percebe-se então, que os métodos de agrupamento fuzzy, se mostram coerentes com a realidade multi temática dos documentos textuais. Por sua vez, o método FCM(fuzzy c means), que é uma adaptação do clássico k means, se propõe a identificar e separar uma coleção de documentos em grupos, respeitando a lógica multi valorada, permitindo assim que um documento pertença a um ou mais grupos. No entanto o FCM possui algumas falhas conhecidas, o que motivou a pesquisa e desenvolvimento de métodos alternativos e baseados no FCM, com o propósito de sanar estes problemas. Este é o caso dos métodos PCM(Possibilístico C Means) e PFCM(Possibilístico C Means). Para então avaliarmos corretamente o resultado do agrupamento e a qualidade da organização flexível de documentos, é preciso extrair corretamente os descritores dos grupos obtidos, levando em consideração a relevância de determinado termo para cada grupo. Com isso temos um cenário no qual é preciso combinar métodos de agrupamento fuzzy com métodos de extração de descritores, para obtermos uma bom resultado no processo de organização dos documentos. A investigação e refinamento dessa combinação de métodos, foi a motivação do presente trabalho. Como resultado desse trabalho foi, proposto extender os experimentos referentes a organização flexível de documentos, utilizando novos métodos de agrupamento fuzzy existentes na literatura, como o PCM e o PFCM. Assim como também foi proposto os métodos de extração de descritores: i) Mixed-PFDCL (*Mixed - Possibilistic Fuzzy Descriptor Comes Last*), que se utiliza da abordagem híbrida do algoritmo PFCM, misturando assim descritores fuzzy e possibilísticos. ii) MixedW-PFDCL (*Mixed Weight - Possibilistic Fuzzy Descriptor Comes Last*), onde além de misturar

descritores fuzzy e possibilístico, leva em consideração os parâmetros de ponderação do método PFCM. Além dos métodos de extração de descritores, foi conduzido um estudo dos impactos de se utilizar o algoritmo PCM, no método de agrupamento hierárquico HFCM, o que resultou no método de agrupamento hierárquico HPCM (*Hierarchical Possibilistic C Means*).

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

ABSTRACT

A new powerful and flexible organization of documents can be obtained by mixing fuzzy and possibilistic clustering, in which documents can belong to more than one cluster simultaneously with different compatibility degrees with a particular topic. The topics are represented by clusters and the clusters are identified by one or more descriptors extracted by a proposed method. We aim to investigate whether the descriptors extracted after fuzzy and possibilistic clustering improves the flexible organization of documents. Experiments were carried using a collection of documents and we evaluated the descriptors ability to capture the essential information of the used collection. The results prove that the fuzzy possibilistic clusters descriptors extraction is effective and can improve the flexible organization of documents.

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

SUMÁRIO

Capítulo 1—Introdução 1

Capítulo 2—Fundamentação Teórica 3

2.1 Conjuntos e Lógica Fuzzy . . . 3
2.1.1 Considerações iniciais . . . 3
2.1.2 Definição de conjuntos fuzzy . . . 3
2.1.3 Lógica fuzzy . . . 4
2.2 Pré-Processamento . . . 5
2.3 Agrupamento Fuzzy . . . 6
2.3.1 Algoritmo Fuzzy C Means (FCM) . . . 8
2.3.2 Algoritmo Possibilistic C Means (PCM) . . . 11
2.3.3 Algoritmo Possibilistic Fuzzy C Means (PFCM) . . . 13
2.3.4 Algoritmo Hierarchic Fuzzy C Means (HFCM) . . . 15
2.4 Extração de descritores . . . 18

Capítulo 3—Trabalhos Relacionados 20

3.1 Considerações Iniciais . . . 20
3.2 Organização Flexível de Documentos . . . 21

Capítulo 4—Experimentos 24

4.1 Refinamento com os algoritmos PCM e PFCM . . . 24
4.2 Refinamento da extração de descritores usando uma abordagem mista . . . 24
4.3 Abordagem possibilística hierárquica . . . 24

Capítulo 5—Conclusão 25

LISTA DE FIGURAS

2.1 Ilustração denotando os grupos g_1, g_2, g_3 organizados sem sobreposição, para $m = 1$. 9
2.2 Ilustração denotando os grupos g_1, g_2, g_3 organizados de maneira fuzzy, com sobreposição, quando $m \to \infty$. 9
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 $\mu(d_1, g_1) = \mu(d_1, g_2) = \mu(d_2, g_1) = \mu(d_2, g_2) = 0.5$. 10
2.4 Resultado do agrupamento de dois conjuntos de coordenadas no R^2 usando o algoritmo FCM[1]. 11
2.5 Demonstração de agrupamentos obtidos com os algoritmos FCM[2](a) e PCM[2](b) 13
2.6 Demonstração do agrupamento obtido com os algoritmo PFCM[3]em um conjunto de coordenadas de pontos no R^2. 15
2.7 Exemplo de hierarquia de tópicos presentes em uma coleção de textos. . . 16

LISTA DE TABELAS

3.1 Classificação das bases de dados de acordo com o seu tamanho(HAVENS et al., 2012) 22

Capítulo 1

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

INTRODUÇÃO

Diante da grande quantidade de informações geradas e armazenadas pela humanidade na atualidade, vários métodos foram propostos visando processar esses dados. Dentre esses dados, temos uma imensa quantidade de dados textuais, que por sua vez são não estruturados. Com isso é notória a importância, de organizar de maneira automatizada, esses documentos pelos assuntos ao qual se tratam. Em particular temos um conjunto de técnicas pertencentes ao campo de estudo da mineração de textos, que visam realizar a tarefa de extrair informações relevantes de documentos textuais. Esta tarefa de análise e extração de informações é comumente segmentada nas tarefas de coleta, pré-processamento dos documentos, agrupamento dos dados e por fim a extração de descritores dos grupos obtidos na etapa de agrupamento. Os métodos de agrupamento podem ser separados então pela lógica matemática utilizada, que pode ser a lógica clássica ou a lógica fuzzy. Na lógica clássica, após o agrupamento, cada documento só poderá pertencer a um grupo, enquanto na lógica fuzzy, a pertinência do documento será distribuída entre os grupos. Se analisarmos a diversidade de conteúdo em documentos textuais, é trivial notar que frequentemente um texto aborda um ou mais temas. Com isso é evidente a necessidade de desenvolver-se técnicas para organizar de maneira flexível os documentos. Percebe-se então, que os métodos de agrupamento fuzzy, se mostram coerentes com a realidade multi temática dos documentos textuais. Por sua vez, o método FCM(fuzzy c means), que é uma adaptação do clássico k means, se propõe a identificar e separar uma coleção de documentos em grupos, respeitando a lógica multi valorada, permitindo assim que um documento pertença a um ou mais grupos. No entanto o FCM possui algumas falhas conhecidas, o que motivou a pesquisa e desenvolvimento de métodos alternativos e baseados no FCM, com o propósito de sanar estes problemas. Este é o caso dos métodos PCM(Possibilístico C Means) e PFCM(Possibilístico C Means). Para então avaliarmos corretamente o resultado do agrupamento e a qualidade da organização flexível de documentos, é preciso extrair corretamente os descritores dos grupos obtidos, levando em consideração a relevância de determinado termo para cada grupo. Com isso temos um cenário no qual é preciso combinar métodos de agrupamento fuzzy com métodos de

extração de descritores, para obtermos uma bom resultado no processo de organização dos documentos. A investigação e refinamento dessa combinação de métodos, foi a motivação do presente trabalho. Como resultado desse trabalho foi, proposto extender os experimentos referentes a organização flexível de documentos, utilizando novos métodos de agrupamento fuzzy existentes na literatura, como o PCM e o PFCM. Assim como também foi proposto os métodos de extração de descritores: i) Mixed-PFDCL (*Mixed - Possibilistic Fuzzy Descriptor Comes Last*), que se utiliza da abordagem híbrida do algoritmo PFCM, misturando assim descritores fuzzy e possibilísticos. ii) MixedW-PFDCL (*Mixed Weight - Possibilistic Fuzzy Descriptor Comes Last*), onde além de misturar descritores fuzzy e possibilístico, leva em consideração os parâmetros de ponderação do método PFCM. Além dos métodos de extração de descritores, foi conduzido um estudo dos impactos de se utilizar o algoritmo PCM, no método de agrupamento hierárquico HFCM, o que resultou no método de agrupamento hierárquico HPCM (*Hierarchical Possibilistic C Means*).

Capítulo

2

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

FUNDAMENTAÇÃO TEÓRICA

2.1 CONJUNTOS E LÓGICA FUZZY

2.1.1 Considerações iniciais

Primeiramente introduzida em (ZADEH, 1965), onde o autor inicia a discussão definindo os conjuntos fuzzy, sendo uma classe de objetos com valores contínuos de pertinência. Cada conjunto é então 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 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). Portanto esta seção tem como propósito contextualizar os principais aspectos da lógica fuzzy que a torna tão importante no contexto da organização flexível de documentos. Portanto definições mais aprofundadas sobre fuzzy fogem do escopo desse texto .

2.1.2 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.

2.1.3 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 a modelar soluções mais adequadas para ambientes imprecisos e incertos. Antes da lógica fuzzy ser introduzida em (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 entendermos essa noção, vejamos então um mesmo exemplo pela ótica do raciocínio clássico e em seguida usando as ferramentas 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 e) O texto B fala um pouco sobre direito. Vamos então adicionar a imprecisão comum no mundo real as sentenças anteriores.

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

2.2 PRÉ-PROCESSAMENTO

Pré-processamento dos dados é o processo de limpeza e preparação do texto para classificação. Assim como muitas palavras em um texto não causam nenhum impacto no significado geral do documento(HADDI; LIU; SHI, 2013). Soma se a isso o enorme custo computacional do processo de mineração de textos, devido a grande quantidade de verbetes presente em dados textuais. Portanto quanto maior for a coleção de textos, maior será a quantidade de palavras distintas. Elevando bastante o custo computacional das tarefas de agrupamento e classificação, que por sua vez são baseadas na análise do vocabulário dos documentos. Com isso, vários pesquisadores propuseram métodos para tentar simplificar, sintetizar e eliminar redundâncias desnecessárias nas coleções de textos. Pois, quanto mais compacto for a quantidade de verbetes da coleção de documentos, menor o custo computacional e a quantidade de memória utilizada nas fases de agrupamento, extração de descritores e classificação. A esse conjunto de técnicas realizadas inicialmente sobre os documentos, denominamos de pré-processamento.

A fase de pré-processamento voltada para a mineração de textos, requer técnicas muito diferentes no preparo dos dados não estruturados para as fases posteriores, do que as técnicas comumente encontradas nos métodos de descoberta de informação. As quais visam preparar dados estruturados para as clássicas operações de mineração de dados (FELDMAN; SANGER, 2007).

Segundo (FELDMAN; 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; SANGER, 2007). No contexto da mineração de textos as características dos documentos são as suas palavras(HADDI; LIU; SHI, 2013). Ao final do processo, as palavras mais relevantes são utilizadas, e as demais são descartadas. Uma vez que manter estas palavras torna a dimensionalidade do problema maior, pois cada palavra no texto é tratada como uma dimensão(HADDI; LIU; SHI, 2013).

O processo como um todo envolve várias etapas, as quais podemos 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 (HADDI; LIU; SHI, 2013).

Diversos métodos foram então propostos para se capturar a importância dos termos em coleções textuais. Sendo o método *Term Frequency Inverse Document Frequency* (TF-IDF) um dos mais importantes(HADDI; LIU; SHI, 2013) e frequentemente utilizado na literatura. 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 $1 \leq n \leq N$, 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 $1 \leq j \leq k$.

2.3 AGRUPAMENTO FUZZY

O agrupamento é um processo não supervisionado (FELDMAN; SANGER, 2007), onde o objetivo é organizar os documentos similares no mesmo grupo e os documentos com grau de dissimilaridade elevado em grupos distintos(NOGUEIRA, 2013)(FELDMAN; SANGER, 2007). 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 (FELDMAN; SANGER, 2007).

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 (FELDMAN; SANGER, 2007). Então pretende-se otimizar a escolha dos grupos, entre todos 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; SANGER, 2007).

A medida de similaridade desempenha um papel fundamental no agrupamento, uma vez que ela precisa expressar o quanto distante está um elemento do outro na coleção. Assim sendo, para obtermos bons resultados durante a organização dos elementos 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(FELDMAN; SANGER, 2007) 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 verbetes em uma coleção, o que faz com que um determinado documento d_i, não contenha diversos termos presentes em um outro documento d_j. Resultando assim que o vetor de características de cada documento, seja preenchido com vários zeros. Reduzindo então a eficácia da distância euclidiana (Equação 2.2) (NOGUEIRA, 2013). Consequentemente a medida de similaridade mais comum para coleções textuais é o coeficiente de similaridade de cosseno (NOGUEIRA, 2013)(FELDMAN; 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 desse processo, 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; EHRLICH; FULL, 1984). Enquanto em grupos $soft$, cada documento d_i pode pertencer a um ou mais grupos g_j, 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.

Portanto, seja $G = \{g_1, g_2, g_3, ..., g_c\}$ os grupos resultantes do agrupamento, sendo c o total de grupos. No agrupamento *hard*, a pertinência de cada documento d_i pode ser

representada pela função de pertinência $\kappa(d_i, g_j) \in \{0, 1\}$, tal que $\sum_{j=1}^{c} \kappa(d_i, g_j) = 1$. Um dos mais populares algoritmos a implementar essa abordagem *hard* é o K Means. Em (BEZDEK; EHRLICH; FULL, 1984)(NOGUEIRA, 2013)(FELDMAN; SANGER, 2007), é apontado 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; EHRLICH; FULL, 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 de pertinência de um documento d_i em um grupo g_j, pode ser definida como sendo $\mu(d_i, g_j) \in [0, 1]$, tal que $\sum_{j=1}^{m} \mu(d_i, g_j) = 1$.

Outro 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* (NOGUEIRA, 2013). 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 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; EHRLICH; FULL, 1984), 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, refletindo a realidade dos documentos textuais. Utilizando como medida de similaridade o coeficiente de similaridade de cosseno (Equação 2.3). E por fim a quantidade de grupos ideal foi escolhida utilizando o método da silhueta fuzzy (Equação 2.5).

2.3.1 Algoritmo Fuzzy C Means (FCM)

(BEZDEK; EHRLICH; FULL, 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 seu pioneirismo no campo de estudo, possuindo diversas extensões. Assim como também é considerado um dos mais amplamente utilizados métodos de agrupamento fuzzy da literatura(PAL et al., 2005). Sendo a maioria dos métodos de agrupamento fuzzy, derivações do FCM(KRISHNAPURAM; KELLER, 1993).

Seja então $V = \{v_1, v_2, v_3, ..., v_c\}$ 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_i seja o protótipo de g_i, c o número de grupos gerados no agrupamento e n o número de documentos presentes na coleção. Bem como seja

$$U_{c \times n} = \{\mu(d_i, g_k) | \mu(d_i, g_k) \in [0, 1], 1 < i \leq n, 1 < k \leq c, eqs(2.9)(2.10)\} \tag{2.7}$$

uma partição fuzzy, com todas as pertinências dos documentos aos grupos. Sendo 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)(NOGUEIRA, 2013). Segundo (BEZDEK; EHRLICH; FULL, 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.

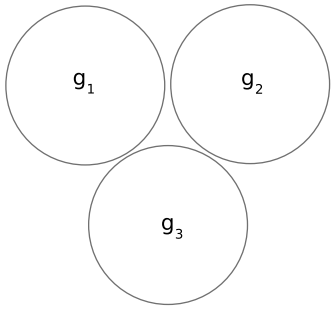

Figura 2.1 Ilustração denotando os grupos g_1, g_2, g_3 organizados sem sobreposição, para $m = 1$.

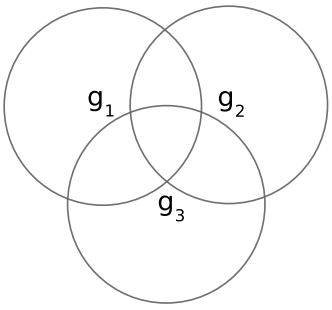

Figura 2.2 Ilustração denotando os grupos g_1, g_2, g_3 organizados de maneira fuzzy, com sobreposição, quando $m \to \infty$.

E ademais a função de pertinência $\mu(d_i, g_k)$ de cada documento na coleção, é definida como sendo

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

, tal que $\mu(d_i, g_k)$ está sujeita as equações

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

e

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

. Onde usualmente no contexto do agrupamento de coleções textuais $dist(d_i, g_k) = scos(d_i, g_k)$. Sendo assim a restrição (2.10) da equação (2.8) tem como função evitar que algoritmo FCM produza grupos vazios(NOGUEIRA, 2013). Enquanto a equação (2.9) impõe que a soma das pertinências seja sempre igual a um. Entretanto a restrição (2.10) 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 do 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).

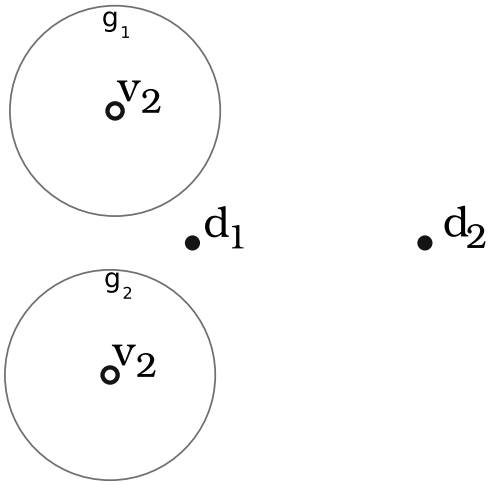

Figura 2.3 Problema dos elementos equidistantes do algoritmo FCM. Na imagem g_1 e g_2 são grupos, com os seus respectivos protótipos v_1 e v_2. Enquanto d_1 e d_2 são documentos equidistantes aos protótipos v_1 e v_2. Portanto $\mu(d_1, g_1) = \mu(d_1, g_2) = \mu(d_2, g_1) = \mu(d_2, g_2) = 0.5$.

Segundo (BEZDEK; EHRLICH; FULL, 1984) várias funções de otimização das partições fuzzy produzidas no agrupamento foram propostas. Sendo a minimização da função objetivo $J(U_{c\times n}, G, V, D)$, que visa minimizar a distância entre os documentos e os protótipos dos grupos(NOGUEIRA, 2013), definida na equação (2.11) a mais popular.

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

O mais utilizado algoritmo para soluções aproximadas da minimização (2.11), é a iteração de Picard[1](PAL et al., 2005) através das equações (2.6)(2.7). Sendo assim o ciclo de aproximações se dá por $V_{t-1} \Rightarrow U_t \Rightarrow V_t$, onde ao final de cada 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). (BEZDEK; EHRLICH; FULL, 1984) e (PAL et al., 2005) afirmam que a convergência desse modelo iterativo ocorre em ambos os tipos de ciclo. A partição fuzzy inicial U_0 é comumente inicializada com valores aleatórios(NOGUEIRA, 2013) ou com o resultado de um agrupamento previamente executado(PAL et al., 2005)(KRISHNAPURAM; KELLER, 1993). Nas demais iterações a atualização dos protótipos é realizada a partir da equação (2.6). O pseudo código utilizando essa abordagem iterativa está listado no algoritmo (1), no qual **inicializa-particao-fuzzy(D,G)**, pode ser uma das duas formas de inicialização descritas anteriormente.

Data: D, c, m, ε
Result: V, U
G $\leftarrow [g_1, g_2, ..., g_c]$;
$U_0 \leftarrow$ **inicializa-particao-fuzzy(D,G)**;
$t \leftarrow 0$;
do
  $V_t \leftarrow$ calcula usando (2.6);
  $t \leftarrow t + 1$;
  $U_t \leftarrow$ calcula usando (2.7);
while $\| U_{t-1} - U_t \| > \varepsilon$;
retorne(U_t, V_t);

Algoritmo 1: Pseudo código da implementação iterativa do método FCM

Por fim está ilustrado na figura (2.4), os resultados produzido 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.

2.3.2 Algoritmo Possibilistic C Means (PCM)

A restrição probabilística (2.10) do FCM 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*). Visando então contornar esses problemas do FCM, foi proposto em (KRISHNAPURAM; KELLER, 1993) o algoritmo *Possiblistic C Means* (PCM).

Ao contrário do FCM, o PCM não atribui pertinências dos documentos aos grupos, e sim tipicidades. Que podem ser interpretadas como graus de possibilidade de um elemento

[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⟩

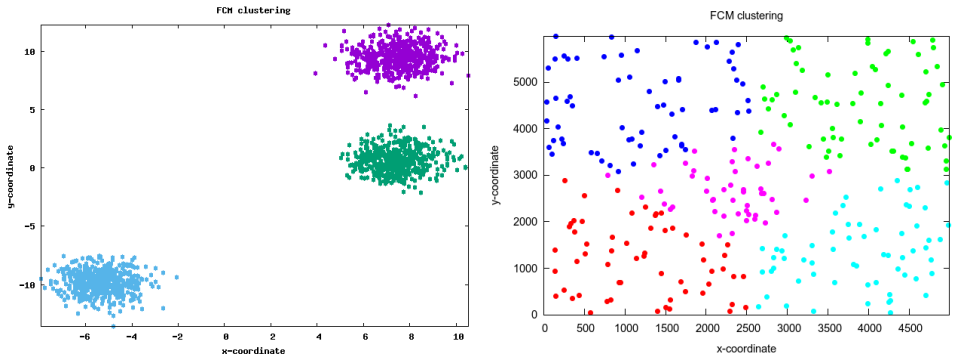

Figura 2.4 Resultado do agrupamento de dois conjuntos de coordenadas no R^2 usando o algoritmo FCM[3].

pertencer ao 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 prótotipos 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 (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. Portanto (KRISHNAPURAM; KELLER, 1993) relaxa a restrição (2.10), fazendo com que a pertinência dependa unicamente da distância do elemento ao grupo. Logo as restrições (2.9) e (2.10), são redefinidas nas equações (2.12)(2.13), onde $\lambda(d_i, g_j)$ representa a tipicidade do documento d_i em relação ao grupo g_j.

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

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

Após relaxar a restrição (2.10) se mantivermos a função objetivo do FCM (equação 2.11), teríamos uma solução trivial, bastando atribuir 0 a todas as pertinências para minimizar $J(U_{c\times n}, G, V, D)$(KRISHNAPURAM; KELLER, 1993)(NOGUEIRA, 2013). Portanto buscando evitar essa solução trivial, e manter a característica de atribuir aos elementos representativos pertinências elevadas aos grupos e penalizar os elementos não representativos, a função objetivo é reformulada como sendo

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

, onde $\gamma_j > 0$ são parâmetros que determinam a distância a qual o valor de pertinência (tipicidade) se torna 0.5. (KRISHNAPURAM; KELLER, 1993) explica que γ_j seja esco-

lhido a depender da faixa de possibilidades (pertinência) desejada para um grupo. Por exemplo, γ_j pode ser igual para todos os grupos, quando se deseja que a forma 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, o autor indica que a definição (2.15) se mostra 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 pertinências possibilísticas produzidas pelo PCM é então definida como sendo

$$P_{c\times n} = \{\lambda(d_i,g_k)|\lambda(d_i,g_k)\in[0,1], 1<i\leq n, 1<k\leq c, eqs(2.12)(2.13)\} \tag{2.16}$$

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

, enquanto a atualização de protótipos 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).

Data: D, c, m, ε
Result: V, P
G $\leftarrow [g_1, g_2, ..., g_c]$;
$P_0 \leftarrow$ **inicializa-particao-fuzzy(D,G)**;
$\gamma_j \leftarrow$ calcula utilizando (2.15);
$t \leftarrow 0$;
do
  $V_t \leftarrow$ calcula usando (2.6);
  $t \leftarrow t+1$;
  $P_t \leftarrow$ calcula usando (2.16);
while $\| P_{t-1} - P_t \| > \varepsilon$;
retorne(P_t, V_t);

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.

2.3.3 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

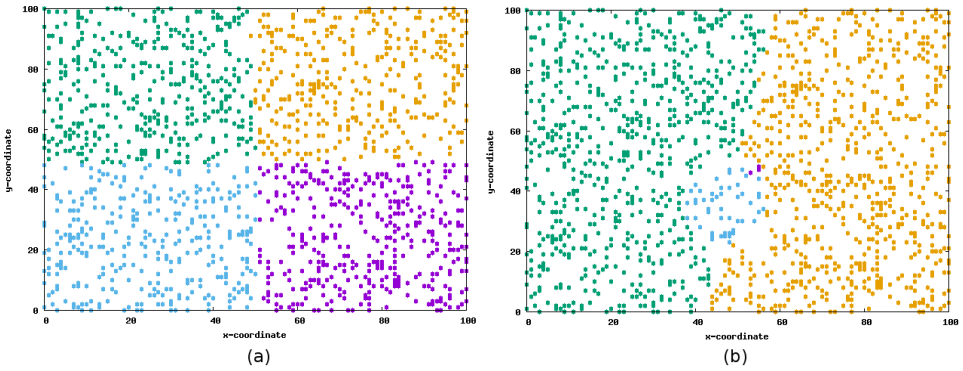

Figura 2.5 Demonstração de agrupamentos obtidos com os algoritmos FCM[5](a) e PCM[5](b) .

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. Pois quando se tem dados que precisam ser rotulados 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 menor distância. 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 então 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. Deixando o usuário definir a proporção de cada uma das contribuições com parâmetros que ponderam o peso de ambos. Portanto é realizado uma mistura entre as funções objetivo (2.11) e (2.14) resultando na minimização da função objetivo (2.18), que está sujeita as condições de (2.19), onde $a, b > 0$ e $m, n > 1$. Os parâmetros a, b representam a importância relativa dos valores de pertinência e tipicidades e devem ser definidos pelo usuário de acordo com o problema. 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 (2.20), a qual passa a se beneficiar das características de ambos os algoritmos. De maneira a reduzir os efeitos dos dados ruidosos, minimizar o

problema dos protótipos coincidentes e evitar a singularidade do FCM. O pseudo código do PFCM é apresentado no Algoritmo 3, onde a função **inicializa-prototipos**(D, G) gera os protótipos iniciais da partição V_0. Como resultado demonstrativo desse algoritmo está ilustrado na figura (2.6), onde é possível observar 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).

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

Data: D, c, m, ε
Result: V, P
G $\leftarrow [g_1, g_2, ..., g_c]$;
$V_0 \leftarrow$ **inicializa-prototipos(D,G)**;
$\gamma_j \leftarrow$ calcula utilizando (2.15);
$t \leftarrow 0$;
do
  $U_t \leftarrow$ calcula com (2.7) usando V_{t-1};
  $P_t \leftarrow$ calcula com (2.16) usando V_{t-1};
  $V_t \leftarrow$ calcula com (2.20);
  $t \leftarrow t + 1$;
while $\| V_{t-1} - V_t \| > \varepsilon$;
retorne(U_t, P_t, V_t);

Algoritmo 3: Pseudo código da implementação iterativa do método PFCM

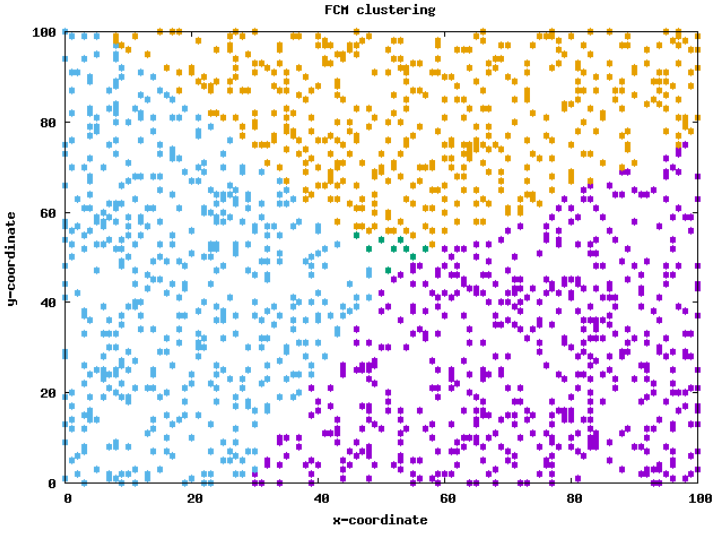

Figura 2.6 Demonstração do agrupamento obtido com os algoritmo PFCM[7]em um conjunto de coordenadas de pontos no R^2.

2.3.4 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 tópicos. Porém acontece, que um tópico 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. Portanto construir hierarquias utilizando métodos de agrupamento fuzzy é o propósito principal do algoritmo HFCM proposto em (PEDRYCZ; REFORMAT, 2006).

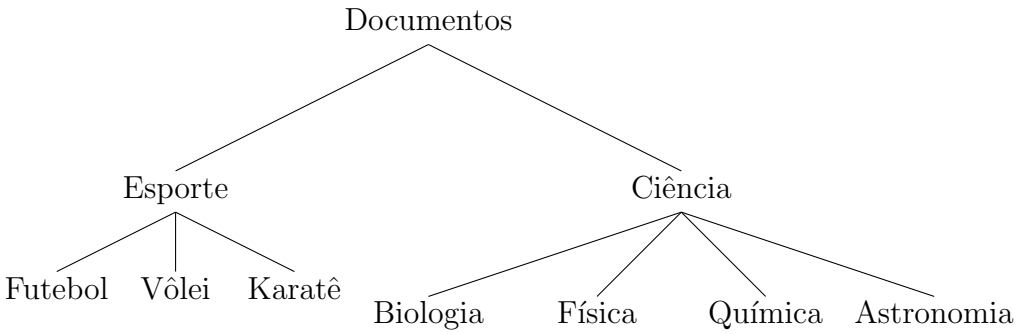

Figura 2.7 Exemplo de hierarquia de tópicos presentes em uma coleção de textos.

O HFCM consegue realizar essa tarefa, expandindo sucessivamente as folhas presentes na hierarquia em subgrupos mais detalhados(PEDRYCZ; REFORMAT, 2006). Onde a expansão é realizada através de novos agrupamentos com o algoritmo CFCM, que é uma versão condicional do FCM. No entanto 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 grupo a ser expandido.

Seja então $D = \{d_1, d_2, ..., d_n\}$ um conjunto com n documentos, então o HFCM executa sobre D o algoritmo FCM produzindo o primeiro nível da hierarquia o método. Como resultado é gerada uma partição fuzzy $U_{c\times n}[1]$ e os protótipos $V[1] = \{v_1[1], v_2[1], ..., v_c[1]\}$, onde [1] representa o 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 (2.22)(2.23) , onde Q representa a qualidade dos protótipos gerados pelo agrupamento. De maneira que 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. Considere então j o grupo com maior valor de Q do nível l da hierarquia e

$$D_j[l] = \left\{ d_k | \mu(d_k, g_j[l]) \leq \frac{1/l}{c} \right\} \tag{2.21}$$

a coleção de documentos selecionados do grupo j com pertinência maior que a pertinência média. 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; REFORMAT, 2006).

$$Q_j[1] = \sum_{d_i \in g_j} dist(d'_i, d_i)^2 \tag{2.22}$$

$$Q'_j[2] = \sum_{d_i \in g_j} dist(d'_i, d_i)^2 \tag{2.23}$$

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

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

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 2.22 para o nível inicial da hierarquia e 2.23 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 maneira que o a equação (2.6) do FCM, contudo a função de pertinência é redefinida como sendo

$$\mu(d_i, g_h[l]) = \frac{\mu(d_i, g_j[l-1])}{\sum_{k=1}^{c} \left(\frac{dist(d_i, v_h[l])}{dist(d_i, v_k[l])} \right)^{\frac{1}{m-1}}}, 1 < h \leq c, d_i \in D_j[l-1], eq(2.27) \tag{2.26}$$

, cujo o valor de l corresponde ao nível da hierarquia e g_j seja o grupo expandido no nível $l-1$. Portanto percebe-se que a pertinência de um documento d_i em um grupo $g_h[l]$, será no máximo a pertinência do documento ao grupo que foi expandido. Com isso temos a restrição 2.10 do FCM adaptada no CFCM para

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

. Logo 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$(l-1)$.

O pseudo código do método CFCM é apresentado no Algoritmo 4, de modo a deixar uma representação mais objetiva de como estruturar esses elementos. Enquanto no Algoritmo 5 consta 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 $lmax$.

Data: $D_j[l]$, U[l-1], l, c, m, ε
Result: V[l], U[l]
G[l] $\leftarrow [g_1, g_2, ..., g_c]$;
$V_0[l] \leftarrow$ **inicializa-prototipos(D,G)**;
$t \leftarrow 0$;
do
  $U_t[l] \leftarrow$ calcula com (2.26) usando V_{t-1};
  $V_t[l] \leftarrow$ calcula com (2.6);
  $t \leftarrow t + 1$;
while $\| V_{t-1}[l] - V_t[l] \| > \varepsilon$;
retorne$(U_t[l], V_t[l])$;

Algoritmo 4: Pseudo código do método CFCM

Data: D, c, m, ε, lmax
Result: Hierarquia
$l \leftarrow 0$;
Hierarquia $\leftarrow$ G[l],V[l],U[l] $\leftarrow$ **fcm**(D,c,m,ε); Algoritmo 1;
Hierarquia $\leftarrow Hierarquia + \{U[l], V[l]\}$;
Q[l] $\leftarrow$ calcula desempenho dos grupos com a equação (2.22);
$g_{max} \leftarrow$ escolhe o grupo g_j com maior valor de Q;
$D_{max} \leftarrow$ seleciona documentos de g_{max} com equação (2.21);
$l \leftarrow l + 1$;
do
  Q[l] $\leftarrow$ calcula desempenho dos grupos com a equação (2.23);
  $g_{max} \leftarrow$ escolhe o grupo g_j com maior valor de Q;
  $D_{max} \leftarrow$ seleciona documentos de g_{max} com equação (2.21);
  G[l],V[l],U[l] $\leftarrow$ **cfcm**$(D_{max}, U[l-1], l, c, m, \varepsilon)$; Algoritmo 4;
  Hierarquia $\leftarrow Hierarquia + \{U[l], V[l]\}$;
  $l \leftarrow l + 1$;
while $l \leq lmax$;
retorne$(Hierarquia)$;

Algoritmo 5: Pseudo código do método HFCM

2.4 EXTRAÇÃO DE DESCRITORES

A tarefa de rotular grupos é um dos problemas chaves do agrupamento de textos, pois ao final do processo de agrupamento, os grupos precisam apresentar alguma relevância para o usuário(ZHANG; XU, 2008). Assim como pretende-se que os descritores escolhidos também sejam significativos para os documentos presentes no grupo a ser rotulado.

Essa etapa pode ser realizada manualmente, com o usuário guiando o processo, ou de

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

forma automatizada, que por sua vez é mais interessante para a proposta de organização flexível de documentos. Uma vez que para grandes bases de dados textuais, a tarefa de rotular 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(NOGUEIRA, 2013). 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, levam em considerações também fontes de informação externas, 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 no grupo, , no entanto o resultado pode ser genérico demais(TREERATPITUK; CALLAN, 2006), ou os descritores podem ser extraídos dos grupos que estão mais próximos do centroide do grupo.

Contudo (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 rótulos 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. Essa flexibilidade possibilitou que a investigação misturasse diferentes técnicas, permitindo obter melhores resultados.

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, como ficou evidenciado na fundamentação teórica. Por isso a literatura existente para essa proposta é bastante rica e densa. Portanto com o propósito de otimizar a atividade de pesquisa e seleção do conhecimento científico produzido a respeito do tema, foram utilizados algumas técnicas de revisão sistemática de literatura ($SLR - Sistematic\ Literature\ Review$) utilizadas em (RIOS; MELLO, 2010). Com o objetivo de estabelecer critérios mais precisos na fase inicial da descoberta de conteúdo científico relacionado ao tema. Foi então adotada uma técnica comum ao método SLR, que consiste na elaboração de uma uma string de busca, usando operadores lógicos. Estabelecendo assim uma maneira mais objetiva para a obtenção de resultados relevantes a proposta dessa monografia. Portanto, levando em consideração os tópicos chaves 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 amplo acervo de publicações científicas presentes no repositório IEEExplore[1], assim como também a possibilidade de se utilizar operadores lógicos e buscas parametrizadas. Foi realizado então uma busca no repositório IEEExplore, restringindo o período de resultados entre os anos de 2010 e 2016, permitindo então que os resultados obtidos fossem 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 essa pesquisa.

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

Durante a fase de leitura parcial dos resultados da busca, 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; EHRLICH; FULL, 1984), são pilares fundamentais para o tema.

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

3.2 ORGANIZAÇÃO FLEXÍVEL DE DOCUMENTOS

Após a proposição da lógica fuzzy que se propunha a lidar com a incerteza e imprecisão em (ZADEH, 1965), foi possível a elaboração de diversos métodos que se utilizassem dos benefícios da lógica fuzzy e aplicassem a diversos problemas do mundo real. Este é o caso da organização de documentos, que por não ser uma tarefa precisa, necessita de uma certa flexibilidade no processo.

(MATSUMOTO; HUNG, 2010) informa que os mecanismos adotados em sistemas de recuperação de informação (SRI), tais como buscadores web, estão dispostos em dois grupos. Sendo que o primeiro tem como foco o 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. Enquanto na segunda abordagem os resultados da busca é categorizado em grupos, permitindo assim que o usuário decida em qual grupo 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; REZENDE, 2010) propõe 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. É possível a visualização dessa abordagem de categorização hierárquica, através da ferramenta online Torch[2], dos autores do artigo.

Com o surgimento de várias tecnologias, como mídias sociais, computação ubíqua, internet das coisas e principalmente os dispositivos móveis, que ultrapassou os 7 bilhões

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

de dispositivos[3] no ano de 2015. Onde por sua vez, todas essas tecnologias produzem uma abundante quantidade de dados não estruturados, dificultando a tarefa de métodos de mineração de dados, e por consequência também os métodos de organização de documentos já existentes. A esse cenário é usualmente atribuído o nome de *Big Data*. Sendo assim novas pesquisas como (HAVENS et al., 2012) e (KUMAR et al., 2015) tem sido conduzidas, 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 $Very\ Large$ (Tabela 3.2, a primeira consiste na técnica de agrupamento distribuído incremental e o 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 matrix 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 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

[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.

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.

Em (NOGUEIRA, 2013) é apresentado um modelo de SRI para organização flexível de documentos, com a adição de três métodos de extração descritores na etapa final do processo. 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 permitem a recuperação dos documentos, através de uma consulta realizada por um usuário. Os métodos de extração de descritores propostos, se baseiam nas medidas de precisão, revocação e a medida-F, que são índices bem estabelecidos na área de mineração de textos.

Capítulo

4

Investigação e refinamento do método de organização flexível de documentos

EXPERIMENTOS

4.1 REFINAMENTO COM OS ALGORITMOS PCM E PFCM

4.2 REFINAMENTO DA EXTRAÇÃO DE DESCRITORES USANDO UMA ABORDAGEM MISTA

4.3 ABORDAGEM POSSIBILÍSTICA HIERÁRQUICA

Capítulo

5

Síntese da investigação e dos experimentos realizados nesta monografia

CONCLUSÃO

REFERÊNCIAS BIBLIOGRÁFICAS

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⟩.

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⟩.

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.

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⟩.

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.

KARAMI, A. et al. FLATM: A fuzzy logic approach topic model for medical documents. In: *Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015 Annual Conference of the North American*. [s.n.], 2015. p. 1–6. Disponível em: ⟨http://dx.doi.org/10.1109/NAFIPS-WConSC.2015.7284190⟩.

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*, PP, n. 99, p. 1–1, 2015.

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⟩.

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

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⟩.

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

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⟩.

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.