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.

18p.: 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

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

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 $G = \{g_1, g_2, g_3, ..., g_m\}$ os grupos resultantes do agrupamento, sendo m 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 $\sum_{j=1}^{m} \kappa(d_i, g_j) = 1$. 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) fez uso da *FuzzySilhouette* (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 (*AverageSilhuetteWidthCriterion* - ASWC), desenvolvido para o agrupamento *crisp* (NOGUEIRA, 2013). A definição da silhueta fuzzy (adaptado de (NOGUEIRA, 2013)) 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 $\beta(d_i, g_l) = min\{\alpha(d_i, g_h) | 1 \leq h \leq m; h \neq l\}$, é a medida de dissimilaridade de d_i ao grupo vizinho mais próximo de g_l, tal que m 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).

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

$$v_j = \frac{\sum_{i=1}^{n}[\mu(d_i, g_j)]^m d_i}{\sum_{i=1}^{n}[\mu(d_i, g_j)]^m} \tag{2.6}$$

, tal que v_i seja o protótipo de g_i e c o número de grupos gerados no agrupamento. Enquanto m é um parâmetro chamado de fator de fuzificação, que regula o quão fuzzy será as partições finais. De modo que para $m = 1$ a partição resultante é totalmente *hard* e para $m \to \infty$ a interseção entre os grupos tende a aumentar(PAL et al., 2005)(NOGUEIRA, 2013). O algoritmo FCM realiza então uma série de atualizações nos protótipos dos grupos, a partir de uma uma pseudo partição fuzzy. O termo pseudo partição é empregado pois os elementos em uma partição fuzzy, não estão associados a um único grupo. Assim, os valores de $g_1, g_2, g_3, ...g_n$ são comumente inicializados com valores aleatórios(NOGUEIRA, 2013) ou com o resultado de outro agrupamento previamente executado(PAL et al., 2005)(KRISHNAPURAM; KELLER, 1993). Nas demais iterações a atualização dos protótipos é realizada a partir do grau de pertinência $\mu(d_i, g_k)$ de cada documento na coleção, conforme definido na equação (2.7), sendo m um parâmetro chamado de fator de fuzificação. No contexto do agrupamento de documentos $dist(d_i, g_k) = scos(d_i, g_k)$.

$$\mu(d_i, g_k) = \frac{1}{\sum_{j=1}^{n}(\frac{dist(d_i, v_k)}{dist(d_i, v_j)})^{\frac{1}{m-1}}} \tag{2.7}$$

O método FCM sujeita a equação (2.7) as restrições (2.8) (2.9), o que impõe algumas limitações ao método, que por sua vez são exploradas por alguns métodos que estendem do FCM, como o PCM e o PFCM.

$$\sum_{k=1}^{c} \mu(d_i, g_k) = 1 \tag{2.8}$$

$$0 < \sum_{i=1}^{n} \mu(d_i, g_k) < n \tag{2.9}$$

2.3.2 Algoritmo Possibilistic C Means (PCM)

2.3.3 Algoritmo Possibilistic Fuzzy C Means (PFCM)

2.3.4 Algoritmo Hierarchic Fuzzy C Means (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 pretend-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 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

Revisão de todo material utilizado desde a fase de pesquisa e implementação até a execução dos experimentos.

REVISÃO BIBLIOGRÁFICA

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.

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

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. 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. Baseada nessa abordagem 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

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

autores do artigo.

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⟩.

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⟩.

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⟩.

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.

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.

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.