



**UNIVERSIDADE FEDERAL DA BAHIA**

## **TRABALHO DE GRADUAÇÃO**

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

9p.: 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 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 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

<b>Capítulo 1—Introdução</b>	1
<b>Capítulo 2—Fundamentação Teórica</b>	3
2.1 Lógica Fuzzy . . . . .	3
2.2 Pré-Processamento . . . . .	3
2.3 Agrupamento Fuzzy . . . . .	3
2.4 Extração de descritores . . . . .	3
<b>Capítulo 3—Revisão Bibliográfica</b>	5
<b>Capítulo 4—Exemplos</b>	7



## LISTA DE FIGURAS

4.1	Figura UFBA . . . . .	7
-----	-----------------------	---



## LISTA DE TABELAS

4.1	Tabela Exemplo . . . . .	7
-----	--------------------------	---



## 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* ).



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

## **FUNDAMENTAÇÃO TEÓRICA**

### **2.1 LÓGICA FUZZY**

### **2.2 PRÉ-PROCESSAMENTO**

### **2.3 AGRUPAMENTO FUZZY**

### **2.4 EXTRAÇÃO DE DESCRITORES**



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

## **REVISÃO BIBLIOGRÁFICA**

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 um 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 estender 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* ).

Livro (DEMEYER; DUCASSE; NIERSTRASZ, 2008) e livro (RAYMOND, 1999).

## CAPÍTULO 4

## EXEMPLOS

Figura 4.1 e tabela 4.1.



**Figura 4.1** Figura UFBA

**Tabela 4.1** Tabela Exemplo

elemento 11	elemento 12
elemento 21	elemento 22
elemento 31	elemento 32



## REFERÊNCIAS BIBLIOGRÁFICAS

DEMEYER, S.; DUCASSE, S.; NIERSTRASZ, O. *Object-Oriented Reengineering Patterns*. Square Bracket Associates, 2008. (This book is available as a free download from <http://www.iam.unibe.ch/scg/OORP/>). ISBN 978-3-9523341-2-6. Disponível em: <http://scg.unibe.ch/download/oorp/>.

RAYMOND, E. S. *The Cathedral and the Bazaar*. 1st. ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 1999. ISBN 1565927249.