

# Uma abordagem híbrida para organização flexível de documentos

## Apresentação de Monografia

Nilton Vasques Carvalho Junior

Universidade Federal da Bahia  
Departamento de Ciência da Computação  
**Orientadora:** Profa. Dra. Tatiane Nogueira Rios  
Contato: niltonvasques {arroba} dcc.ufba.br

2 de Junho de 2016

# Conteúdo

- 1 Introdução
- 2 Fundamentação Teórica
  - Pré-processamento
  - Agrupamento (FCM,PCM,PFCM)
  - Extração de descritores
- 3 Trabalhos relacionados
- 4 Abordagem proposta
  - Refinamento com PFCM
  - Método PDCL
  - Método Mixed-PFDCL
  - Resultados
- 5 Conclusão

# Conteúdo

- 1 Introdução
- 2 Fundamentação Teórica
  - Pré-processamento
  - Agrupamento (FCM,PCM,PFCM)
  - Extração de descritores
- 3 Trabalhos relacionados
- 4 Abordagem proposta
  - Refinamento com PFCM
  - Método PDCL
  - Método Mixed-PFDCL
  - Resultados
- 5 Conclusão

# Introdução

- O avanço da tecnologia tem proporcionado um **aumento gigantesco** na quantidade de **dados armazenados**.
- A rede social Facebook produz mais de *25 terabytes/dia* (Havens et al., 2012).
- Governos e corporações também produzem milhares de **documentos** todos os dias, tais como relatórios, formulários, pesquisas de opiniões e etc.
- Muggleton (2006) ressalta que este cenário está além dos limites humanos para o uso e compreensão.

# Introdução

- O avanço da tecnologia tem proporcionado um **aumento gigantesco** na quantidade de **dados armazenados**.
- A rede social Facebook produz mais de **25 terabytes/dia** (Havens et al., 2012).
- Governos e corporações também produzem milhares de **documentos** todos os dias, tais como relatórios, formulários, pesquisas de opiniões e etc.
- Muggleton (2006) ressalta que este cenário está além dos limites humanos para o uso e compreensão.

# Introdução

- O avanço da tecnologia tem proporcionado um **aumento gigantesco** na quantidade de **dados armazenados**.
- A rede social Facebook produz mais de **25 terabytes/dia** (Havens et al., 2012).
- Governos e corporações também produzem milhares de **documentos** todos os dias, tais como relatórios, formulários, pesquisas de opiniões e etc.
- Muggleton (2006) ressalta que este cenário está além dos limites humanos para o uso e compreensão.

# Introdução

- O avanço da tecnologia tem proporcionado um **aumento gigantesco** na quantidade de **dados armazenados**.
- A rede social Facebook produz mais de **25 terabytes/dia** (Havens et al., 2012).
- Governos e corporações também produzem milhares de **documentos** todos os dias, tais como relatórios, formulários, pesquisas de opiniões e etc.
- Muggleton (2006) ressalta que este cenário está além dos limites humanos para o uso e compreensão.

# Introdução

- Kobayashi e Aono (2008) enfatizam que instituições estão sobrecarregadas com o processamento desse montante de dados.
- Os dados possuem diversos tipos e formatos, sendo armazenados de forma estruturada ou **não estruturada**.

## Exemplos

documentos de textos, planilhas, áudios, imagens, vídeos, documentos HTML e etc.



# Introdução

- Kobayashi e Aono (2008) enfatizam que instituições estão sobrecarregadas com o processamento desse montante de dados.
- Os dados possuem diversos tipos e formatos, sendo armazenados de forma estruturada ou **não estruturada**.

## Exemplos

documentos de textos, planilhas, áudios, imagens, vídeos, documentos HTML e etc.

# Introdução

- Kobayashi e Aono (2008) enfatizam que instituições estão sobrecarregadas com o processamento desse montante de dados.
- Os dados possuem diversos tipos e formatos, sendo armazenados de forma estruturada ou **não estruturada**.

## Exemplos

documentos de textos, planilhas, áudios, imagens, vídeos, documentos HTML e etc.

# Introdução

- Dados estruturados já possuem mecanismos eficientes de armazenamento e recuperação.
- Documentos textuais por serem não estruturados são recuperados através de Sistemas de Recuperação da Informação (SRI).

## Exemplos

Duckduckgo, Jus Brasil, IEEExplore, ACM, Google e etc

# Introdução

- Dados estruturados já possuem mecanismos eficientes de armazenamento e recuperação.
- **Documentos textuais** por serem **não estruturados** são recuperados através de Sistemas de Recuperação da Informação (SRI).

## Exemplos

Duckduckgo, Jus Brasil, IEEExplore, ACM, Google e etc

# Introdução

- Demanda crescente para desenvolvimento e aprimoramento de métodos que possam processar e **extrair padrões de dados textuais**.
- A extração de padrões de documentos textuais é o principal objetivo da Mineração de Textos (MT).

# Introdução

Vários desafios estão presentes no processo de extração de padrões de documentos textuais, entre eles destaca-se:

- Não estruturados.
- Naturalmente **imprecisos** e **incertos**.
- Abordam um ou mais temas.
- **Alta dimensionalidade**.
- Dados **esparsos**.

## Exemplos

Uma coleção de documentos pode conter 100.000 palavras, enquanto um documento pode conter apenas algumas centenas (Aggarwal e Zhai, 2012).

# Introdução

Vários desafios estão presentes no processo de extração de padrões de documentos textuais, entre eles destaca-se:

- Não estruturados.
- Naturalmente **imprecisos** e **incertos**.
- Abordam um ou mais temas.
- **Alta dimensionalidade**.
- Dados **esparsos**.

## Exemplos

Uma coleção de documentos pode conter 100.000 palavras, enquanto um documento pode conter apenas algumas centenas (Aggarwal e Zhai, 2012).

# Introdução

Vários desafios estão presentes no processo de extração de padrões de documentos textuais, entre eles destaca-se:

- Não estruturados.
- Naturalmente **imprecisos** e **incertos**.
- Abordam um ou mais temas.
- **Alta dimensionalidade**.
- Dados **esparsos**.

## Exemplos

Uma coleção de documentos pode conter 100.000 palavras, enquanto um documento pode conter apenas algumas centenas (Aggarwal e Zhai, 2012).



# Introdução

Vários desafios estão presentes no processo de extração de padrões de documentos textuais, entre eles destaca-se:

- Não estruturados.
- Naturalmente **imprecisos** e **incertos**.
- Abordam um ou mais temas.
- **Alta dimensionalidade**.
- Dados **esparsos**.

## Exemplos

Uma coleção de documentos pode conter 100.000 palavras, enquanto um documento pode conter apenas algumas centenas (Aggarwal e Zhai, 2012).

# Introdução

Vários desafios estão presentes no processo de extração de padrões de documentos textuais, entre eles destaca-se:

- Não estruturados.
- Naturalmente **imprecisos** e **incertos**.
- Abordam um ou mais temas.
- **Alta dimensionalidade**.
- Dados **esparsos**.

## Exemplos

Uma coleção de documentos pode conter 100.000 palavras, enquanto um documento pode conter apenas algumas centenas (Aggarwal e Zhai, 2012).









# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- Grupos com significado relevante.
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- Considerar a imprecisão e a incerteza.
- Reduzir a influência de documentos ruidosos.

# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- **Grupos com significado relevante.**
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- Considerar a imprecisão e a incerteza.
- Reduzir a influência de **documentos ruidosos**.



# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- **Grupos com significado relevante.**
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- Considerar a imprecisão e a incerteza.
- Reduzir a influência de **documentos ruidosos**.

# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- **Grupos com significado relevante.**
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- Considerar a imprecisão e a incerteza.
- Reduzir a influência de **documentos ruidosos**.

# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- **Grupos com significado relevante.**
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- Considerar a imprecisão e a incerteza.
- Reduzir a influência de **documentos ruidosos**.

# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- **Grupos com significado relevante.**
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- **Considerar a imprecisão e a incerteza.**
- Reduzir a influência de **documentos ruidosos**.

# Introdução

O agrupamento é muito importante neste processo e possui uma série de desafios:

- Agrupar de acordo com a similaridade.
- **Grupos com significado relevante.**
- Escalável para grandes coleções (*Big Data*).
- Baixo custo computacional.
- Estimar os parâmetros dos algoritmos.
- **Considerar a imprecisão e a incerteza.**
- **Reduzir a influência de documentos ruidosos.**

# Introdução

## Citação

*[...] não é esperado que um único método de agrupamento atenda todas as exigências para todos os conjuntos de dados [...]  
(Steinbach et al., 2003).*

# Introdução

Existem diversos métodos de agrupamento na literatura, os quais destacam-se:

- *Fuzzy C-Means* (FCM)
- *Possibilistic C-Means* (PCM)
- *Possibilistic Fuzzy C-Means* (PFCM)

# Introdução

A partir das investigações conduzidas foi proposto dois métodos de extração de descritores:

- **Possibilistic Description Comes Last (PDCL)**
- **Mixed - Possibilistic Fuzzy Description Comes Last (Mixed-PFDCL) (Híbrido)**



# Conteúdo

- 1 Introdução
- 2 Fundamentação Teórica
  - Pré-processamento
  - Agrupamento (FCM,PCM,PFCM)
  - Extração de descritores
- 3 Trabalhos relacionados
- 4 Abordagem proposta
  - Refinamento com PFCM
  - Método PDCL
  - Método Mixed-PFDCL
  - Resultados
- 5 Conclusão

# Pré-processamento

- Remoção de espaços.
- Expansão de abreviações.
- Remoção de *stopwords* (pronomes, artigos e etc.).
- Lematização (Casa  $\rightarrow$  Cas).
- Estruturação dos documentos (TF-IDF).

	<i>termo<sub>1</sub></i>	<i>termo<sub>2</sub></i>	<i>termo<sub>3</sub></i>
<i>doc<sub>1</sub></i>	1	3	4
<i>doc<sub>2</sub></i>	9	2	0

**Tabela:** Exemplo matriz docs x termos



	<i>termo<sub>1</sub></i>	<i>termo<sub>2</sub></i>	<i>termo<sub>3</sub></i>
<i>doc<sub>1</sub></i>	0.1	0.6	1.0
<i>doc<sub>2</sub></i>	0.9	0.4	0.0

**Tabela:** Exemplo matriz tf-idf

# Agrupamento

- Organizar objetos similares em um mesmo grupo.
- Grupos crisp x fuzzy
- Coeficiente de similaridade de cosseno.
- Validação do agrupamento com o método silhueta fuzzy.

# Agrupamento

- Organizar objetos similares em um mesmo grupo.
- Grupos crisp x fuzzy
- Coeficiente de similaridade de cosseno.
- Validação do agrupamento com o método silhueta fuzzy.

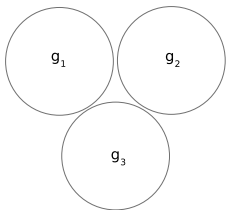


Imagem: Grupos crisp

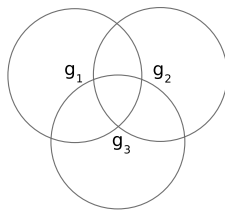


Imagem: Grupos fuzzy

# Agrupamento

- Organizar objetos similares em um mesmo grupo.
- Grupos crisp x fuzzy
- Coeficiente de similaridade de cosseno.
- Validação do agrupamento com o método silhueta fuzzy.

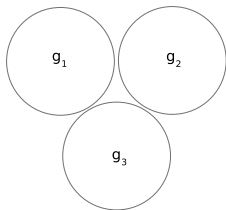


Imagem: Grupos crisp

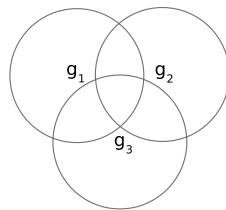


Imagem: Grupos fuzzy

# Agrupamento

- Organizar objetos similares em um mesmo grupo.
- Grupos crisp x fuzzy
- Coeficiente de similaridade de cosseno.
- Validação do agrupamento com o método silhueta fuzzy.

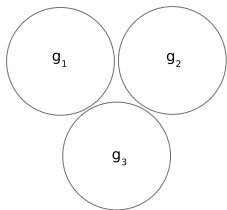


Imagem: Grupos crisp

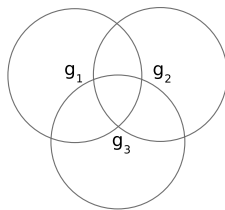


Imagem: Grupos fuzzy

# Agupamento (FCM) (Bezdek et al., 1984)

- Graus de pertinência.
- **Restrição probabilística.**
- Problema com ruídos.

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	<b>total</b>
<i>doc<sub>1</sub></i>	0,5	0,5	1,0
<i>doc<sub>2</sub></i>	0,5	0,5	1,0

Tabela: Pertinências FCM

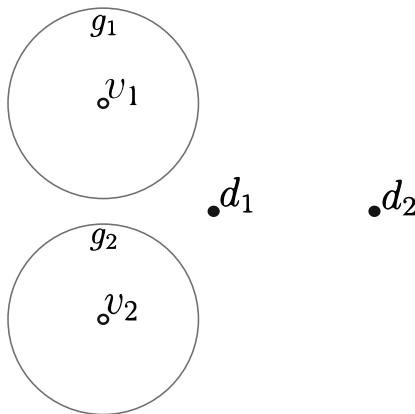


Imagem: Problema dos ruídos

# Agrupamento (PCM) (Krishnapuram e Keller, 1993)

- Graus de tipicidade.
- Remoção da restrição probabilística.
- Problema dos grupos coincidentes.

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	<b>total</b>
<i>doc<sub>1</sub></i>	0,7	0,7	1,4
<i>doc<sub>2</sub></i>	0,2	0,2	0,4

Tabela: Tipicidades PCM

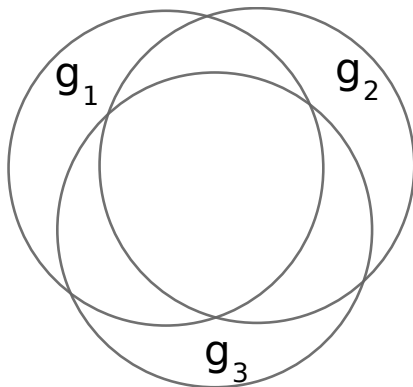


Imagem: Grupos coincidentes



# Agrupamento (PFCM) (Pal et al., 2005)

- Pertinências e tipicidades.
- Robustez.
- Parâmetros de ponderação  $a$  e  $b$ .

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	<b>total</b>
<i>doc<sub>1</sub></i>	0,5	0,5	1,0
<i>doc<sub>2</sub></i>	0,5	0,5	1,0

Tabela: Pertinências PFCM

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	<b>total</b>
<i>doc<sub>1</sub></i>	0,7	0,7	1,4
<i>doc<sub>2</sub></i>	0,2	0,2	0,4

Tabela: Tipicidades PFCM

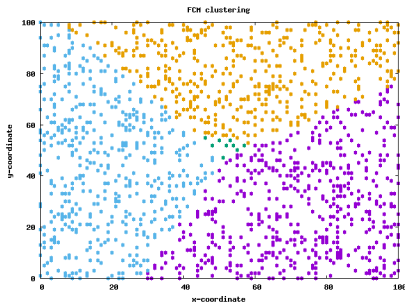
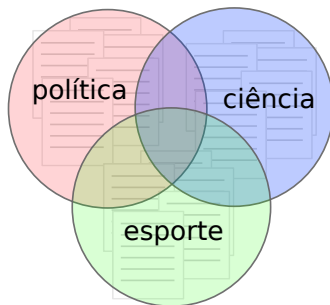


Imagem: Agrupamento de pontos.

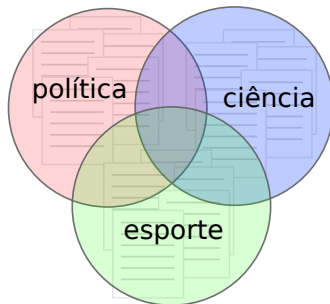
# Extração de descritores

- Atribuir significados aos grupos.
- Manual ou **Automatizada**.
- Abordagens de conhecimento interno e externo.
- Durante o agrupamento (*Description Comes First* - DCF)
- **Após o agrupamento** (*Description Comes Last* - **DCL**).
- Método *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) (Nogueira, 2013).



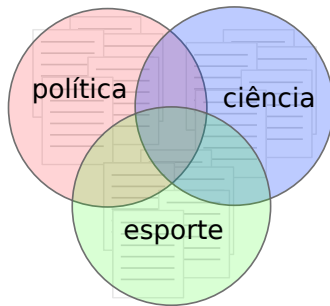
# Extração de descritores

- Atribuir significados aos grupos.
- Manual ou **Automatizada**.
- Abordagens de conhecimento interno e externo.
- Durante o agrupamento (*Description Comes First* - DCF)
- **Após o agrupamento** (*Description Comes Last* - **DCL**).
- Método *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) (Nogueira, 2013).



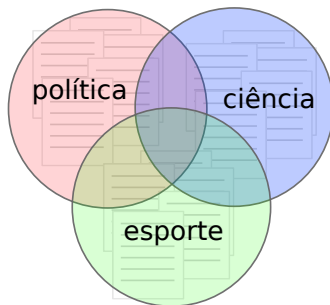
# Extração de descritores

- Atribuir significados aos grupos.
- Manual ou **Automatizada**.
- Abordagens de conhecimento interno e externo.
- Durante o agrupamento (*Description Comes First* - DCF)
- **Após o agrupamento** (*Description Comes Last* - **DCL**).
- Método *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) (Nogueira, 2013).



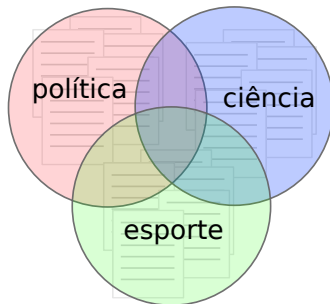
# Extração de descritores

- Atribuir significados aos grupos.
- Manual ou **Automatizada**.
- Abordagens de conhecimento interno e externo.
- Durante o agrupamento (*Description Comes First* - DCF)
- **Após o agrupamento** (*Description Comes Last* - **DCL**).
- Método *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) (Nogueira, 2013).



# Extração de descritores

- Atribuir significados aos grupos.
- Manual ou **Automatizada**.
- Abordagens de conhecimento interno e externo.
- Durante o agrupamento (*Description Comes First* - DCF)
- **Após o agrupamento** (*Description Comes Last* - **DCL**).
- Método *Soft Organization - Fuzzy Description Comes Last* (SoftO-FDCL) (Nogueira, 2013).



# Organização Flexível de Documentos

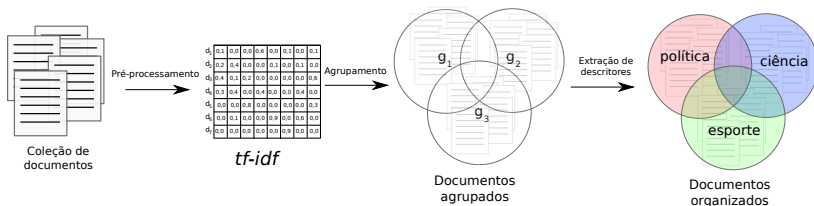


Imagem: Organização flexível de documentos.

## Definição

A **organização flexível de documentos** pode ser definida como o processo que compreende a **estruturação dos dados**, a adição de flexibilidade proporcionada pelo **agrupamento fuzzy**, a **extração de descritores** dos grupos de maneira flexível e a recuperação de informação através de um SRI.

# Organização Flexível de Documentos

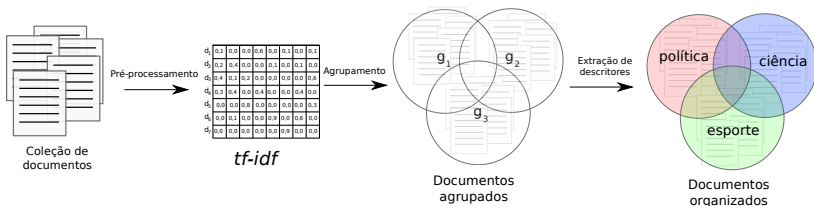


Imagem: Organização flexível de documentos.

## Definição

A **organização flexível de documentos** pode ser definida como o processo que compreende a **estruturação dos dados**, a adição de flexibilidade proporcionada pelo **agrupamento fuzzy**, a **extração de descritores** dos grupos de maneira flexível e a recuperação de informação através de um SRI.



# Conteúdo

- 1 Introdução
- 2 Fundamentação Teórica
  - Pré-processamento
  - Agrupamento (FCM,PCM,PFCM)
  - Extração de descritores
- 3 **Trabalhos relacionados**
- 4 Abordagem proposta
  - Refinamento com PFCM
  - Método PDCL
  - Método Mixed-PFDCL
  - Resultados
- 5 Conclusão

# Trabalhos relacionados

- Marcacini e Rezende (2010) propões uma abordagem incremental hierárquica para construção de tópicos.
- Havens et al. (2012) e Kumar et al. (2015) explora otimizações para *Big Data*.
- Deng et al. (2010) propõe uma maneira de estabilizar a inicialização do agrupamento.
- Karami et al. (2015) utiliza o agrupamento ainda na fase de pré-processamento.

# Trabalhos relacionados

- Marcacini e Rezende (2010) propões uma abordagem incremental hierárquica para construção de tópicos.
- Havens et al. (2012) e Kumar et al. (2015) explora otimizações para *Big Data*.
- Deng et al. (2010) propõe uma maneira de estabilizar a inicialização do agrupamento.
- Karami et al. (2015) utiliza o agrupamento ainda na fase de pré-processamento.

# Trabalhos relacionados

- Marcacini e Rezende (2010) propões uma abordagem incremental hierárquica para construção de tópicos.
- Havens et al. (2012) e Kumar et al. (2015) explora otimizações para *Big Data*.
- Deng et al. (2010) propõe uma maneira de estabilizar a inicialização do agrupamento.
- Karami et al. (2015) utiliza o agrupamento ainda na fase de pré-processamento.

# Trabalhos relacionados

- Marcacini e Rezende (2010) propões uma abordagem incremental hierárquica para construção de tópicos.
- Havens et al. (2012) e Kumar et al. (2015) explora otimizações para *Big Data*.
- Deng et al. (2010) propõe uma maneira de estabilizar a inicialização do agrupamento.
- Karami et al. (2015) utiliza o agrupamento ainda na fase de pré-processamento.

# Trabalhos relacionados

- Jiang et al. (2013) combina os algoritmos genéticos no agrupamento para evitar os mínimos locais.
- Murali e Damodaram (2015) propõe uma medida de similaridade com informações semânticas.
- Nogueira (2013) traz uma abordagem de extração de descritores independente do agrupamento.

# Trabalhos relacionados

- Jiang et al. (2013) combina os algoritmos genéticos no agrupamento para evitar os mínimos locais.
- Murali e Damodaram (2015) propõe uma medida de similaridade com informações semânticas.
- Nogueira (2013) traz uma abordagem de extração de descritores independente do agrupamento.

# Trabalhos relacionados

- Jiang et al. (2013) combina os algoritmos genéticos no agrupamento para evitar os mínimos locais.
- Murali e Damodaram (2015) propõe uma medida de similaridade com informações semânticas.
- Nogueira (2013) traz uma abordagem de extração de descritores independente do agrupamento.



# Conteúdo

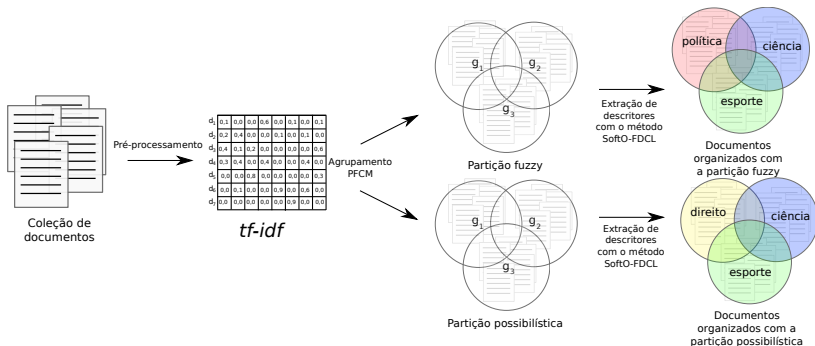
- 1 Introdução
- 2 Fundamentação Teórica
  - Pré-processamento
  - Agrupamento (FCM,PCM,PFCM)
  - Extração de descritores
- 3 Trabalhos relacionados
- 4 Abordagem proposta
  - Refinamento com PFCM
  - Método PDCL
  - Método Mixed-PFDCL
  - Resultados
- 5 Conclusão

# Coleções textuais

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

**Tabela:** Características das coleções textuais utilizadas nesta pesquisa

# Refinamento com PFCM



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

# Refinamento com PFCM

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

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

# Refinamento com PFCM

<b>Coleção</b>	<b>docs</b>	<b>termos</b>	<b>FCM</b>	<b>PCM</b>	<b>PFCM</b>
Opinosis	51	842		✓	
20newsgroups	2000	11028			✓
Hitech	600	6925	✓		
NSF	1600	2806	✓		
WAP	1560	8070			✓
Reuters-21578	1052	3925	✓		

**Tabela:** Sumário dos resultados da classificação dos descritores

# Refinamento com PFCM

Método	<i>crisp<sub>1</sub></i>	<i>crisp<sub>2</sub></i>	<i>crisp<sub>3</sub></i>
FCM	drive, display, control, <b>car</b> , work	import, <b>model</b> , problem, unit, design	breakfast, <b>con-</b> <b>cierge</b> , coffee, food, inn
PCM	read, problem, <b>car</b> , work, found	turn, size, qua- lity, review, fea- ture	extreme, drive, point, reason, run
PFCM $\mu$	drive, control, version, <b>car</b> , work	read, complete, <b>device</b> , display, size	breakfast, plea- sant, <b>concierge</b> , coffee, clean
PFCM $\lambda$	club, immacu- late, towel, pil- low, fridge	housekeep, tourist, tea, smoke, london	bottle, adult, food, reserve, dinner

**Tabela:** Descritores extraídos com os métodos de agrupamento FCM, PCM e PFCM da coleção Opinions .

# Refinamento com PFCM - Discussão

- FCM e PFCM capturaram melhor a estrutura das coleções.
- Capacidade de adaptação do método SoftO-FDCL.
- Descritores fuzzy mais significativos.
- **Descritores possibilísticos pouco significativos.**
- Dimensionalidade aparenta influenciar os resultados.

# Método SoftO-FDCL (Nogueira, 2013)

	$\mu(d_i, g_j) \geq \delta, \forall d_i$	$\mu(d_i, g_j) < \delta, \forall d_i$
$t_k \in d_i, \forall d_i$	<i>ganhos</i>	<i>ruídos</i>
$t_k \notin d_i, \forall d_i$	<i>perdas</i>	<i>rejeitos</i>

**Tabela:** Matriz de contingência do método SoftO-FDCL

$$\text{precisão}(t_k, g_j) = \frac{\text{ganhos}}{\text{ganhos} + \text{ruídos}} \quad (1)$$

$$\text{recuperação}(t_k, g_j) = \frac{\text{ganhos}}{\text{ganhos} + \text{perdas}} \quad (2)$$

$$f1(t_k, g_j) = \frac{2 * \text{precisão}(t_k, g_j) * \text{recuperação}(t_k, g_j)}{\text{precisão}(t_k, g_j) + \text{recuperação}(t_k, g_j)} \quad (3)$$



# Método SoftO-FDCL (Nogueira, 2013)

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>
<i>termo<sub>1</sub></i>	0.85	0.75
<i>termo<sub>2</sub></i>	0.35	0.95
<i>termo<sub>3</sub></i>	0.25	0.55
<i>termo<sub>4</sub></i>	0.80	0.65
<i>termo<sub>5</sub></i>	0.50	0.50
<i>termo<sub>6</sub></i>	0.30	0.24
<i>termo<sub>7</sub></i>	0.10	0.83

**Tabela:** Pontuação dos termos obtidas com a medida f1

# Método SoftO-FDCL (Nogueira, 2013)

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>
<i>termo<sub>1</sub></i>	0.85	0.75
<i>termo<sub>2</sub></i>	0.35	0.95
<i>termo<sub>3</sub></i>	0.25	0.55
<i>termo<sub>4</sub></i>	0.80	0.65
<i>termo<sub>5</sub></i>	0.50	0.50
<i>termo<sub>6</sub></i>	0.30	0.24
<i>termo<sub>7</sub></i>	0.10	0.83

Tabela: Descritores de maior pontuação em cada grupo

# O limiar é adequado?

$$\delta = \frac{1}{\text{total de grupos}} = \frac{1}{2} = 0,5 \quad (4)$$

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	<b>total</b>
<i>doc<sub>1</sub></i>	0,4	0,6	1,0
<i>doc<sub>2</sub></i>	0,8	0,2	1,0

Tabela: Pertinências PFCM

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	<b>total</b>
<i>doc<sub>1</sub></i>	0,6	0,9	1,5
<i>doc<sub>2</sub></i>	0,4	0,1	0,5

Tabela: Tipicidades PFCM

# Convertendo tipicidades em pertinências

## Tipicidades → Pertinências

$$\lambda'(d_i, g_j) = \frac{\lambda(d_i, g_j)}{\sum_{k=1}^c \lambda(d_i, g_k)} \quad (5)$$

	<i>grupo<sub>1</sub></i>	<i>grupo<sub>2</sub></i>	total
<i>doc<sub>1</sub></i>	$\frac{0,6}{1,5} = 0,4$	$\frac{0,9}{1,5} = 0,6$	1,0
<i>doc<sub>2</sub></i>	$\frac{0,4}{0,5} = 0,8$	$\frac{0,1}{0,5} = 0,2$	1,0

Tabela: Tipicidades → Pertinências

# Método PDCL

	$\lambda'(d_i, g_j) \geq \delta, \forall d_i$	$\lambda'(d_i, g_j) < \delta, \forall d_i$
$t_k \in d_i, \forall d_i$	<i>ganhos</i>	<i>ruídos</i>
$t_k \notin d_i, \forall d_i$	<i>perdas</i>	<i>rejeitos</i>

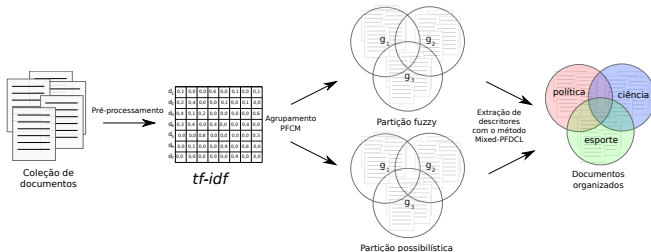
**Tabela:** Matriz de contingência do método PDCL

Ponderando os ganhos, ruídos, perdas e rejeitos

$$ganhos(t_k, g_j) = \sum_{d_i \in D'} \lambda(d_i, g_j) \quad (6)$$

$$D' = \{d_i \mid SE \lambda'(d_i, g_j) \geq \delta \ E \ t_k \in d_i \ PARA \ \forall d_i\} \quad (7)$$

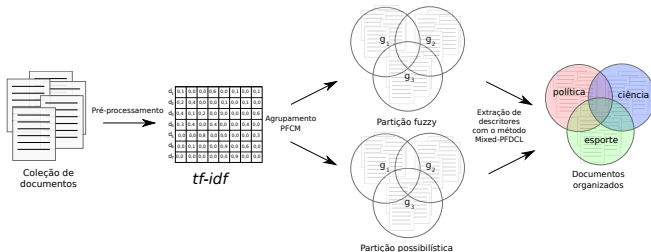
# Método Mixed-PFDCL



## Pertinência híbrida

$$\mu'(d_i, g_j) = \frac{a\mu(d_i, g_j) + b\lambda'(d_i, g_j)}{a + b} \quad (8)$$

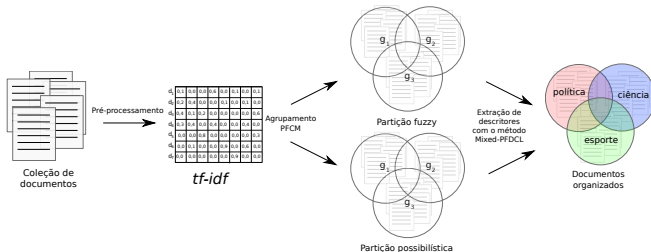
# Método Mixed-PFDCL



## Pertinência híbrida

$$\mu'(d_i, g_j) = \frac{a\mu(d_i, g_j) + b\lambda'(d_i, g_j)}{a + b} \quad (8)$$

# Método Mixed-PFDCL



## Pertinência híbrida

$$\mu'(d_i, g_j) = \frac{a\mu(d_i, g_j) + b\lambda'(d_i, g_j)}{a + b} \quad (8)$$

	$\mu'(d_i, g_j) \geq \delta, \forall d_i$	$\mu'(d_i, g_j) < \delta, \forall d_i$
$t_k \in d_i, \forall d_i$	ganhos	ruídos
$t_k \notin d_i, \forall d_i$	perdas	rejeitos

**Tabela:** Matriz de contingência do método Mixed-PFDCL



# Resultados

	PCM		PFCM	
<b>Coleção</b>	<b>SoftO-FDCL</b>	<b>PDCL</b>	<b>SoftO-FDCL</b>	<b>Mixed</b>
Opinosis		✓		✓
20newsgroups	✓	✓		✓
Hitech		✓		✓
NSF	✓	✓		✓
WAP		✓		✓
Reuters-21578		✓		✓

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

# Resultados

	<i>grupo<sub>1</sub></i>		<i>grupo<sub>2</sub></i>	
Método	termo	pontuação	termo	pontuação
SoftO-FDCL	caf	0,923077	caf	0,923077
	floor	0.888889	floor	0.888889
	food	0.880000	food	0.880000
	coffe	0.857143	coffe	0.857143
	concierge	0.846154	concierge	0.846154
PDCL	bathro	0.894716	make	0.800980
	food	0.888785	time	0.789846
	concierge	0.860127	nice	0.779338
	supermarket	0.856632	feature	0.778138
	chain	0.856632	easy	0.768564

**Tabela:** 5 termos de maior pontuação obtidos extraídos com os métodos Soft-FDCL e PDCL da coleção Opinosis com o algoritmo PCM

# Resultados - Discussão

- Demonstram a adequação da interpretação proposta.
- O método SoftO-FDCL pode gerar pontuações similares a partir das tipicidades.
- Os métodos propostos resolvem o problema de pontuações similares.
- Os métodos PDCL e Mixed-PFDCL superaram o método SoftO-FDCL.
- O método Mixed-PFDCL se mostrou adequado para a organização híbrida de documentos.

# Conteúdo

- 1 Introdução
- 2 Fundamentação Teórica
  - Pré-processamento
  - Agrupamento (FCM,PCM,PFCM)
  - Extração de descritores
- 3 Trabalhos relacionados
- 4 Abordagem proposta
  - Refinamento com PFCM
  - Método PDCL
  - Método Mixed-PFDCL
  - Resultados
- 5 Conclusão

# Conclusão

- A organização flexível de documentos envolve muitos campos de estudo.
- Detalhamento dos métodos de agrupamento FCM, PCM, PFCM e HFCM.
- É possível aprimorar todas as etapas do processo.
- Impactos do PFCM na organização flexível de documentos.

# Conclusão

- A organização flexível de documentos envolve muitos campos de estudo.
- Detalhamento dos métodos de agrupamento FCM, PCM, PFCM e HFCM.
- É possível aprimorar todas as etapas do processo.
- Impactos do PFCM na organização flexível de documentos.

# Conclusão

- A organização flexível de documentos envolve muitos campos de estudo.
- Detalhamento dos métodos de agrupamento FCM, PCM, PFCM e HFCM.
- É possível aprimorar todas as etapas do processo.
- Impactos do PFCM na organização flexível de documentos.

# Conclusão

- A organização flexível de documentos envolve muitos campos de estudo.
- Detalhamento dos métodos de agrupamento FCM, PCM, PFCM e HFCM.
- É possível aprimorar todas as etapas do processo.
- Impactos do PFCM na organização flexível de documentos.



# Conclusão

- **Propriedades do limiar  $\delta$ .**
- A estratégia híbrida se mostrou adequada, produzindo bons descritores.
- Os métodos propostos obtiveram bons resultados.
- Publicação de artigo científico na conferência FUZZ-IEEE.

# Conclusão

- **Propriedades do limiar  $\delta$ .**
- **A estratégia híbrida se mostrou adequada, produzindo bons descritores.**
- Os métodos propostos obtiveram bons resultados.
- Publicação de artigo científico na conferência FUZZ-IEEE.

# Conclusão

- **Propriedades do limiar  $\delta$ .**
- **A estratégia híbrida se mostrou adequada, produzindo bons descritores.**
- **Os métodos propostos obtiveram bons resultados.**
- Publicação de artigo científico na conferência FUZZ-IEEE.

# Conclusão

- **Propriedades do limiar  $\delta$ .**
- **A estratégia híbrida se mostrou adequada, produzindo bons descritores.**
- **Os métodos propostos obtiveram bons resultados.**
- **Publicação de artigo científico na conferência FUZZ-IEEE.**

# Referências I



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





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>>.




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.

# Referências II


 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.


 JIANG, H. et al. An improved method of fuzzy clustering algorithm and its application in text clustering. *JOURNAL OF INFORMATION & COMPUTATIONAL SCIENCE*, JOURNAL OF INFORMATION & COMPUTATIONAL SCIENCE, v. 10, n. 2, p. 519, 2013. Disponível em: <[http://manu35.magtech.com.cn/Jwk\\_ics/EN/abstract/article\\_1507.shtml](http://manu35.magtech.com.cn/Jwk_ics/EN/abstract/article_1507.shtml)>.


 KARAMI, A. et al. FLATM: A fuzzy logic approach topic model for medical documents. In: *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*. Institute of Electrical &

# Referências III

Electronics Engineers (IEEE), 2015. Disponível em: <<http://dx.doi.org/10.1109/NAFIPS-WConSC.2015.7284190>>.

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

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

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

# Referências IV




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.





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



# Referências V

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

 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.

# Referências VI



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