

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
- 5 Conclusão
- 6 Trabalhos futuros

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
- 5 Conclusão
- 6 Trabalhos futuros

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 e documentos HTML.

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 e documentos HTML.

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 e documentos HTML.

Introdução

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

Exemplos

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

Introdução

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

Exemplos

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

Introdução

As seguintes áreas vem explorando e propondo técnicas para otimizar esse processo:

- Mineração de Dados (MD)
- Aprendizado de Máquina
- Recuperação da Informação (RI)

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 na 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

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 Sistema de Recuperação de Informação (SRI)

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.**

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) - Graus de pertinência (Problemas com ruídos).
- *Possibilistic C-Means* (PCM) - Graus de tipicidade (Pode gerar grupos coincidentes).
- *Possibilistic Fuzzy C-Means* (PFCM) - Graus de pertinência e tipicidade (Híbrido).

Introdução

Foi então formulada a seguinte hipótese:

Hipótese

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

Para validar a hipótese definiu-se o como objetivo desta monografia:

Objetivo

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

Introdução

A partir das investigações conduzidas descobriu-se que os **graus de tipicidade afetam** a qualidade dos descritores dos grupos.

Essa descoberta motivou a proposição dos 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
- 5 Conclusão
- 6 Trabalhos futuros

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>termo₁</i>	<i>termo₂</i>	<i>termo₃</i>
<i>doc₁</i>	1	3	4
<i>doc₂</i>	9	2	0

Tabela: Exemplo matriz docs x termos



	<i>termo₁</i>	<i>termo₂</i>	<i>termo₃</i>
<i>doc₁</i>	0.1	0.6	1.0
<i>doc₂</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.

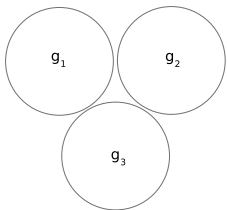


Imagem: Grupos crisp

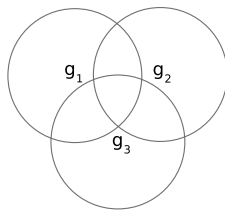


Imagem: Grupos fuzzy

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.

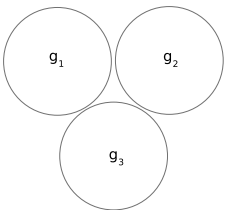


Imagem: Grupos crisp

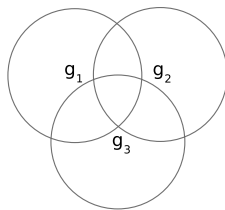


Imagem: Grupos fuzzy

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.

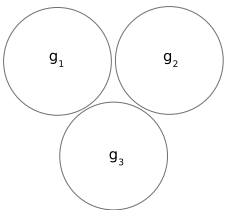


Imagem: Grupos crisp

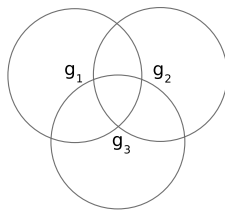


Imagem: Grupos fuzzy

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

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

	<i>grupo₁</i>	<i>grupo₂</i>	total
<i>doc₁</i>	0,5	0,5	1,0
<i>doc₂</i>	0,5	0,5	1,0

Tabela: Pertinências FCM

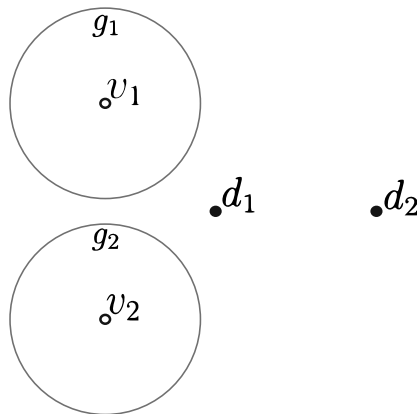


Imagem: Problema dos ruídos

Agrupamento (PCM) (Krishnapuram e Keller, 1993)

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

	<i>grupo₁</i>	<i>grupo₂</i>	total
<i>doc₁</i>	0,7	0,7	1,4
<i>doc₂</i>	0,2	0,2	0,4

Tabela: Tipicidades PCM

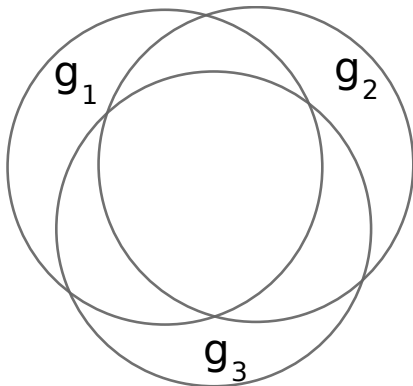


Imagem: Grupos coincidentes

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

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

	<i>grupo₁</i>	<i>grupo₂</i>	total
<i>doc₁</i>	0,5	0,5	1,0
<i>doc₂</i>	0,5	0,5	1,0

Tabela: Pertinências PFCM

	<i>grupo₁</i>	<i>grupo₂</i>	total
<i>doc₁</i>	0,7	0,7	1,4
<i>doc₂</i>	0,2	0,2	0,4

Tabela: Tipicidades PFCM

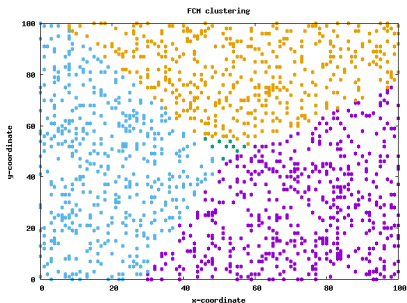
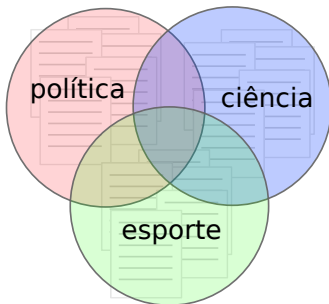


Imagem: Agrupamento de pontos.

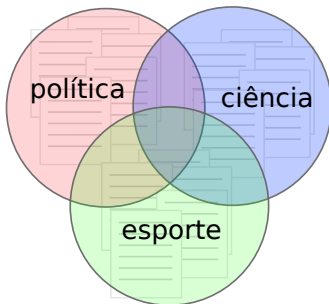
Extração de descritores

- 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).



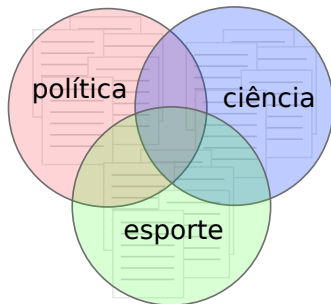
Extração de descritores

- 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).



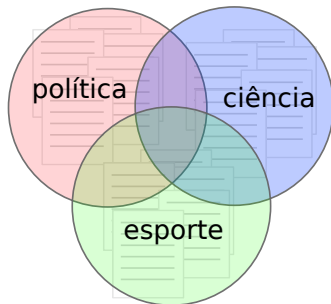
Extração de descritores

- 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).



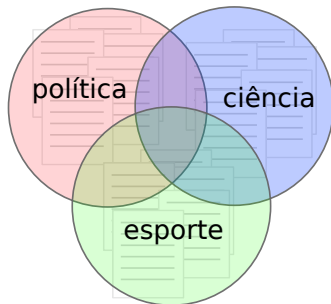
Extração de descritores

- 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

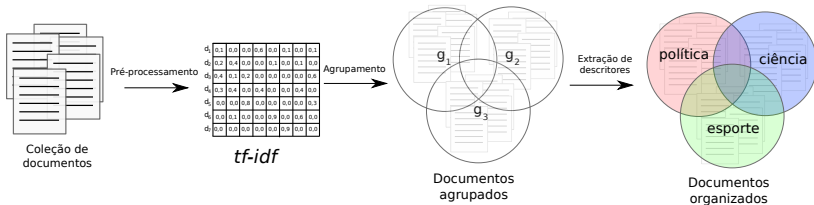


Imagem: Organização flexível de documentos.

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
- 5 Conclusão
- 6 Trabalhos futuros

What is haplotyping and why is it important?

You hopefully know this after the previous three talks...

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
- 5 Conclusão
- 6 Trabalhos futuros

Coleções textuais

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

Tabela: Características das coleções textuais utilizadas nesta pesquisa

What is haplotyping and why is it important?

You hopefully know this after the previous three talks...

What is haplotyping and why is it important?

You hopefully know this after the previous three talks...

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
- 5 Conclusão
- 6 Trabalhos futuros

What is haplotyping and why is it important?

You hopefully know this after the previous three talks...


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
- 5 Conclusão
- 6 Trabalhos futuros


What is haplotyping and why is it important?

You hopefully know this after the previous three talks...


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>.


 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>>.


 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.

Referências II


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


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

 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>>.

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

Referências III

 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.

 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.