# *Relatório Final*

***PIBIC & PIBIC-AF, PIBITI E VOLUNTÁRIOS***

<Observação: Favor não alterar o layout desta página de rosto. Apenas preencha os dados nos campos solicitados. A partir da segunda página estão os demais itens do modelo a serem preenchidos.>

| EDITAL / PROGRAMA |
|---|
| **EDITAL PROPCI/UFBA_01_PIBIC 2015-2016** |

| ESTUDANTE IC<br>(Digitar nome completo, sem abreviações). |
|---|
| **Nilton Vasques Carvalho Junior** |

| Título do Plano de Trabalho do Estudante<br>(Digitar o título completo, sem abreviações, exatamente igual ao título do plano de trabalho aprovado). |
|---|
| **Extração de padrões para gerenciamento de imprecisão e incerteza na organização flexível de documentos de textos** |

| ORIENTADOR<br>(Digitar nome completo, sem abreviações). |
|---|
| **Tatiane Nogueira Rios** |

| Título do Projeto do Orientador<br>(Digitar o título completo, sem abreviações, exatamente igual ao título do projeto do orientador). |
|---|
| **Gerenciamento de imprecisão e incerteza para organização flexível de documentos de texto utilizando a teoria de conjuntos fuzzy** |

Salvador
Agosto de 2016

**RESUMO**

Este relatório visa apresentar uma investigação dos impactos dos algoritmos de agrupamento na composição da organização flexível de documentos textuais. Discutindo aspectos relevantes da etapa de pré-processamento dos documentos a serem organizados; da realização de agrupamento para extração de padrões nestes documentos e dos critérios de validação da organização obtida. A pesquisa realizada para desenvolvimento deste trabalho tem como principal contribuição uma abordagem híbrida para a organização flexível de documentos, mesclando os benefícios concedidos pela adequada interpretação de partições fuzzy e possibilísticas existentes no método de agrupamento Possibilistic C-Means e Possibilistic Fuzzy C-Means (PFCM). Como fruto das análises aqui realizadas, foram propostos os métodos de extração de descritores Possibilistic Description Comes Last (PDCL) e Mixed - Possibilistic Fuzzy Description Comes Last (PFDCL). Ambos mostraram-se relevantes através de evidências experimentais e análises subjetivas à adequação dos métodos, para a organização flexível de documentos, contribuindo com descobertas originais para o estado da arte. Os resultados da pesquisa ainda estimulam novas implementações cuja execução pode transcorrer em trabalhos futuros.

## 1. INTRODUÇÃO

O avanço da computação pessoal, em particular a computação móvel, tem proporcionado um gigantesco aumento da quantidade de dados armazenados pela humanidade ao longo dos anos. A critério de exemplo, a popular plataforma de rede social Facebook 1 , produz diariamente mais de 25 terabytes de informação [1]. A tendência com o avanço das tecnologias, é que tudo seja integrado a internet, de tal modo que os pesquisadores já chegam a dizer que os dados são o novo recurso natural do planeta [2]. Ainda segundo os autores, entre as maiores fontes de geração de dados estão os sistemas governamentais, plataformas de mídias sociais, assim como arquivos armazenados pelas corporações, como por exemplo, formulários médicos, opiniões de consumidores, relatórios e etc.. Entretanto, é importante ressaltar que todos esses dados excedem os limites humanos para o uso e compreensão destes [3].

Diante desse cenário, instituições públicas e privadas estão sobrecarregadas com a tarefa de processar essa imensa quantidade de informação em bases de dados com documentos não estruturados e em diversos formatos [4]. Estes documentos usualmente são de diversos tipos, como por exemplo, textos, áudios, imagens, vídeos, documentos HTML, podendo estar, inclusive, em diferentes idiomas.

Nesse contexto, diversas pesquisas tem objetivado a proposição ou refinamento de técnicas para automatização do processo de análise e aquisição de conhecimento útil desse montante de informações armazenadas. Porém, devido a multi disciplinaridade inerente desse campo de estudo, o mesmo tem sido estudado pelas comunidades de mineração de dados, aprendizado de máquina e recuperação de informação.

A Mineração de Dados (MD) é um campo de estudo que vem obtendo rápidos avanços nos últimos anos, e, isto se deve aos avanços das tecnologias de hardware e sof tware, o qual possibilitou o massivo armazenamento de diferentes tipos de dados, inclusive os dados textuais [5]. Portanto, como resultado desse aumento na quantidade de documentos disponíveis na forma textual, existe uma demanda crescente no desenvolvimento e aprimoramento de métodos e algoritmos que possam efetivamente processar e extrair padrões dos dados de maneira dinâmica e escalável.

Por outro lado, enquanto os dados estruturados já possuem mecanismos bem eficientes de armazenamento e recuperação, os dados textuais são geralmente gerenciados através de mecanismos de buscas para suprir essa falta de estruturação. Esses mecanismos de busca possibilitam aos usuários uma conveniente maneira para recuperar informações em coleções textuais através de consultas com palavras chaves. Desse modo, compete

ao campo de estudo da Recuperação de Informação (RI) a tarefa de explorar, investigar e propor métodos para otimização da eficiência e efetividade de sistemas de buscas [6].

Mas segundo Baeza-Yates e Ribeiro-Neto (2011), as pesquisas de recuperação de informação tem focado tradicionalmente, em formas de facilitar o acesso à informação, do que realizar a descoberta de novos padrões em documentos textuais, o qual se destaca como sendo o objetivo principal da mineração de textos. A mineração de textos, por sua vez, é uma especialização da mineração de dados, que busca incorporar atividades de estruturação dos documentos em formatos apropriados, facilitando a aplicação dos tradicionais métodos de extração de padrões da MD, minimizando as perdas durante a conversão do formato original não estruturado [7].

Contudo, uma série de características diferenciam os documentos textuais de outras formas de dados. O que por sua vez, afeta o desempenho das clássicas técnicas da MD. Dentre essas características peculiares, destacam-se como mais importantes os fatos de que os dados são esparsos e possuem alta dimensionalidade. Por exemplo, uma coleção de documentos (corpus) pode conter 100.000 palavras (termos), enquanto um único documento desse corpus pode conter somente algumas centenas de palavras [5]. Essa discrepância, tem implicações diretas em várias técnicas de identificação de padrões, e especialmente no agrupamento textual, que deriva de clássicas técnicas de agrupamento da mineração de dados, aplicados à conjuntos de baixa dimensionalidade.

Para cumprir a tarefa de extrair informações relevantes de documentos textuais e identificar as estruturas inerentes aos mesmos. A mineração de textos emprega uma variedade de técnicas, as quais se destacam aquelas usualmente desenvolvidas para efetuar tarefas de coleta, pré-processamento, agrupamento textual e seleção de termos descritores para o agrupamento.

O agrupamento pode, de maneira geral, ser definido como a tarefa de agrupar uma coleção de objetos, de acordo algum critério de similaridade. É possível distinguir os tipos de agrupamento em função da lógica empregada por eles. Com isso, tem-se os algoritmos que derivam da lógica clássica ou da lógica fuzzy. Na lógica clássica, após a conclusão do agrupamento, cada elemento só pertence à apenas um grupo, enquanto que na lógica fuzzy, a pertinência do elemento será distribuída entre os grupos.

Ao se analisar a diversidade de conteúdo em dados textuais, é trivial notar que frequentemente um texto aborda um ou mais temas. O que implica que o agrupamento clássico, ao atribuir um objeto a apenas um grupo, não irá representar bem a imprecisão e incerteza natural dos documentos.

Deste modo, os métodos de agrupamento derivados da lógica fuzzy se mostram mais capacitados para lidar com essa imprecisão e incerteza da realidade multi temática dos documentos textuais. Assim sendo, uma 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).

Ao se observar o processo de organização flexível de documentos, percebe-se que o mesmo abrange várias etapas, cada uma delas com suas particularidades. No entanto, apesar da importância desempenhada por cada etapa do processo, o agrupamento em si pode ser visto como uma das peças chaves, pois ele é diretamente responsável por organizar os documentos de acordo com as suas similaridades. Adicionalmente, é preciso desconsiderar ou reduzir a influência de documentos ruidosos, que destoam do restante da coleção nos grupos finais.

O algoritmo Fuzzy C-Means (FCM) [8], que deriva do clássico K-Means [9], e o Possibilistic C-Means (PCM) [10], são exemplos de métodos de agrupamento capazes de organizar de maneira automatizada uma coleção de documentos em um conjunto de grupos. Ambos distribuem os documentos de uma coleção textual em um conjunto de grupos, de modo que cada documento possa pertencer a diferentes grupos com diferentes graus de pertinência, considerando assim a flexibilidade necessária para tratar a imprecisão e incerteza do processo.

No entanto, o FCM apresenta alguns resultados indesejados, diante da presença de

dados ruidosos na coleção. Em se tratando de coleções textuais, um dado ruidoso pode ser considerado como um documento que possua uma temática bastante diferente dos demais documentos da coleção. Com o objetivo de atribuir valores de pertinências mais realísticos aos elementos a serem agrupados e penalizar com baixas pertinências os elementos ruidosos, o algoritmo PCM foi proposto. Porém, o PCM é muito sensível à inicialização, o que pode resultar em grupos coincidentes, onde não há uma separação muito bem definida dos elementos.

Visando contemplar os benefícios de ambos os métodos, foi proposto o método de agrupamento Possibilistic Fuzzy C-Means (PFCM) [11], como uma versão híbrida dos algoritmos FCM e o PCM, objetivando adicionar robustez à tarefa de agrupamento.

Após o agrupamento dos documentos, é necessário realizar a extração dos termos que melhor descrevem os grupos. Para realização dessa tarefa, tem-se alguns métodos na literatura do tipo DCF (Description Comes First), que realizam a extração de maneira embutida no processo de agrupamento. Porém, essa abordagem torna o processo de extração de descritores dependente do algoritmo de agrupamento. Com o propósito de contornar esse cenário, foi proposto em Nogueira (2013) o método Soft Organization Fuzzy Description Comes Last (SoftO-FDCL) (Nogueira, 2013), o qual extrai os termos descritores após a etapa de agrupamento de maneira independente do algoritmo de agrupamento utilizado. Permitindo avaliar diretamente os impactos dos métodos de agrupamento, na extração de descritores e por consequência na qualidade da organização flexível de documentos.

Entretanto, o método SoftO-FDCL foi pensado inicialmente para interpretar as pertinências produzidas na partição do FCM, que difere da partição resultante produzida pelo PCM. A principal contribuição do PCM foi uma alteração no modo de atribuição da pertinência de uma elemento a um grupo, o que impacta diretamente na partição dos grupos resultantes.

Diante deste contexto, e tendo em vista o crescente aumento de informações produzidas além da capacidade humana de analisar. Com a demanda crescente no desenvolvimento e aprimoramento das técnicas de extração e identificação de conhecimento útil em dados textuais, bem como a necessidade de se organizar esses dados de maneira flexível, tratando a imprecisão e incerteza natural desses dados e considerando as particularidades existentes nos métodos de agrupamento, foi formulada a seguinte hipótese para o desenvolvimento desse trabalho:

> *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 demonstrar a validade da hipótese formulada, com base na exploração de estratégias existentes na literatura para o aprimoramento do processo de organização flexível de documentos, definiu-se o seguinte 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.*

Considerando-se os resultados dos experimentos realizados, foi descoberto que as alterações existentes no PCM, impactam diretamente na qualidade dos descritores extraídos pelo método SoftO-FDCL. Essa descoberta motivou a proposição de dois novos métodos de extração de descritores: Possibilistic Descriptor Comes Last (PDCL) e Mixed - Possibilistic Fuzzy Descriptor Comes Last (Mixed-PFDCL). Os quais apresentaram resultados que contribuem de maneira significativa para o estado da arte da extração de descritores dos grupos fuzzy e para o aprimoramento da organização flexível de documentos.

## 2. MATERIAIS E MÉTODOS

Entre os principais fundamentos necessários para compreensão da abordagem apresentada neste documento, estão a atividade de pré-processamento dos dados, cuja finalidade é filtrar e estruturar os dados para serem processados nas etapas seguintes; os principais algoritmos de agrupamento fuzzy, presentes na literatura, com as suas definições matemáticas e pseudo códigos; e, por fim, a tarefa de rotular os grupos encontrados na etapa de agrupamento com os termos que melhor os representem, permitindo assim a realização de consultas em Sistemas de Recuperação da Informação (SRI).

### 2.1 PRÉ PROCESSAMENTO

Pré-processamento dos dados é o processo de limpeza e preparação dos dados para extração de padrões. Para este trabalho, especificamente, considera-se dado como sendo um documento textual e a tarefa de extração de padrões a ser considerada é o agrupamento.

Essa etapa é importante porque algumas palavras em um documento podem causar pouco ou nenhum impacto no significado geral do documento [12]. Soma se a isso o enorme custo computacional do processo de mineração de textos, devido à grande quantidade de atributos presente em dados textuais, visto que quanto maior for a coleção de textos, maior será a quantidade de palavras distintas. Tal dimensionalidade eleva bastante o custo computacional de qualquer tarefa de extração de padrões. Por isso, vários pesquisadores propuseram métodos para tentar simplificar, sintetizar e eliminar redundâncias desnecessárias nas coleções de textos.

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 [13]. No contexto da mineração de textos, as características dos documentos são as suas palavras [12]. Ao final do processo, as palavras mais relevantes são utilizadas, e as demais são descartadas.

### 2.2 AGRUPAMENTO FUZZY

O agrupamento é um processo não supervisionado cujo o objetivo é organizar os objetos similares no mesmo grupo e os objetos com grau de dissimilaridade elevado em grupos distintos [7]. 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. Neste trabalho, os objetos a serem agrupados são os documentos textuais.

#### 2.2.1 Fuzzy C Means (FCM)

Bezdek et al. (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 ao seu pioneirismo no campo de estudo, possuindo diversas extensões, sendo considerado um dos mais amplamente utilizados métodos de agrupamento fuzzy da literatura [11]. A maioria dos métodos de agrupamento fuzzy são derivações do FCM [10].

O algoritmo fuzzy c means é baseado no algoritmo crisp K-means, com uma adaptação para a lógica fuzzy. Ou seja, ao contrário do algoritmo K-means, o FCM passa a permitir que um documento pertença a mais de um grupo. Nos experimentos, foi utilizada a versão iterativa do FCM, onde a cada iteração, a pseudo partição (matriz com as informações de pertinência dos documentos nos grupos) fuzzy é atualizada.

O FCM busca atualizar a pseudo partição fuzzy e minimizar a função objetivo apresentado na Figura 1, a cada iteração. O critério de parada é o parâmetro epsilon, que limita o grau de refinamento na partição.

$$J(P) = \sum_{i=1}\sum_{j=1}[u(di, gl)]^n||di - vl||$$

*Figura 1*

Por conta da maneira como a função objetivo é definida, o FCM tem uma restrição que obriga que a soma das pertinências de um documento em todos os grupos seja sempre 1 (um). Isso leva ao problema dos elementos equidistantes, onde um documento que tenha a mesma similaridade entre os grupos, receberá a pertinência média, mesmo que a similaridade deste documento entre os grupos não seja relevante.A partição inicial do FCM é gerada distribuindo aleatoriamente a pertinência dos documentos nos grupos. Vale ressaltar que todos os demais algoritmos que foram utilizados nessa pesquisa, possuem essa característica, por serem extensões do FCM.

### 2.2.2 Possibilistic C Means (PCM)

A restrição probabilística do FCM, que obriga a soma das pertinências de um elemento ser igual a um, nem sempre resulta em pertinências que representam bem a realidade dos dados. Esse problema se agrava ainda mais, em bases com muitos dados ruidosos (outliers). Portanto, com o objetivo de contornar esses problemas do FCM, foi proposto em Krishnapuram e Keller (1993) o algoritmo Possiblistic C Means (PCM).

Ao contrário do FCM, o PCM não atribui pertinências dos documentos aos grupos, mas sim tipicidades, as quais podem ser interpretadas como graus de possibilidade de um elemento pertencer a um determinado grupo. Como consequência, a partição resultante é possibilística.

A tipicidade do documento em relação ao grupo, tem como objetivo ponderar os elementos que tenham mais relevância em relação ao grupo, e penalizar os documentos menos relevantes. De modo a remover completamente a restrição imposta pelo FCM. A função objetivo do FCM é então modificada pela equação (Figura 2).

$$K(P) = \sum_{i=1}\sum_{j=1}[\varphi(di, gl)]^n||di - vl|| + \sum_{l=1}\gamma i\sum_{i=1}(1 - \varphi(di, gl))^m$$

*Figura 2*

Apesar das melhorias propostas no PCM, existe também uma peculiaridade neste algoritmo, conhecida na literatura como problema dos clusters coincidentes [6]. Este problema ocorre, quando os grupos fuzzy são dispostos muito juntos, levando assim a uma alta interseção entre os conjuntos.

### 2.2.3 Possibilistic Fuzzy C Means (PFCM)

Com o propósito de aproveitar os benefícios de ambas as abordagens, Pal et al. (2005) propôs o algoritmo PFCM, que utiliza as pertinências do FCM e as tipicidades do PCM. Cabe ao usuário definir a proporção de cada uma das contribuições com parâmetros que ponderam o peso de ambos. Para tanto, é realizada uma mistura entre as funções objetivo apresentadas nas Figuras 1 e 2, resultando na minimização da função objetivo apresentada na Figura 3.

$$L(P) = \sum_{i=1}\sum_{j=1}[a[\phi(di, gl)]^n + b[\varphi(di, gl)]^m]||di - vl|| + \sum_{l=1}\gamma i\sum_{i=1}(1 - \varphi(di, gl))^m$$

*Figura 3*

## 2.3 EXTRAÇÃO DE DESCRITORES

A tarefa de atribuir significados à grupos é um dos problemas chave do agrupamento de

textos, pois ao final do processo de agrupamento, os grupos precisam apresentar alguma relevância para o usuário [14]. Portanto, é imprescindível que sejam extraídos descritores significativos para representar os documentos que compõem os grupos.

A etapa de extração de descritores 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, visto que, para grandes bases de dados textuais, a tarefa de extrair descritores para todos os grupos encontrados durante o agrupamento, pode ser bastante exaustiva para o usuário.

Dentre os métodos automatizados, dois tipos de abordagens são encontrados na literatura, uma baseada em conhecimento interno e a outra baseada em conhecimento externo. A primeira se utiliza somente de informações que podem ser obtidas na coleção de do- cumentos, como por exemplo a frequência do termo, localização do termo na estrutura do documento. Enquanto a abordagem de conhecimento externo, leva em consideração fontes de informação externas, como por exemplo a consulta a extensa base de termos na língua inglesa WordNet , para auxiliar a escolha dos termos mais representativos.

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 descritores 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.

Nesse contexto, percebe-se à existência de algumas estratégias para extração de descritores, utilizando ou não conhecimento externo, e embutida ou independente do processo de agrupamento. A partir da avaliação dessas abordagens, e de acordo com o objetivo definido, considerou-se que a abordagem independente do algoritmo de agrupamento é mais pertinente ao presente estudo, pois ela viabiliza a condução de experimentos com vários métodos de agrupamento. Nesse sentido, foi escolhido o método Soft Organization - Fuzzy Description Comes last (SoftO-FDCL) proposto por Nogueira (2013), devido o mesmo possuir essas características necessárias para a investigação dos impactos do agrupamento na qualidade dos descritores e da organização flexível.

## 2.4 COLEÇÕES TEXTUAIS

Na mineração de textos e consequentemente nos trabalhos relacionados à organização flexível de documentos, é comum se realizar a avaliação dos métodos propostos, conduzindo se experimentos sobre coleções textuais existentes na literatura com essa finalidade. Para isso, as coleções precisam se estarem estruturadas. Assim sendo, nesta pesquisa foi adotada a estrutura de representação de documentos textuais tf-idf, como forma de estruturar os dados presentes nas coleções, de modo a capturar a importância relativa dos termos nos documentos e nas coleções, montando assim ao final do pré-processamento uma matriz documentos x termos.

Outro aspecto não menos importante, são as características particulares das coleções textuais. Pois ressalta-se que para uma mais apurada análise dos resultados, é pertinente considerar as particularidades de cada coleção, com a finalidade de encontrar possíveis justificativas para os resultados apresentados, realizando-se indagações comparativas às peculiaridades sabidamente conhecidas dos métodos analisados. O conjunto de características particulares de cada coleção, dar-se à como apresentado na Tabela 1. Uma análise objetiva das características presentes nas seis coleções utilizadas nos experimentos pode ser observada na Tabela 2, onde é possível notar de maneira bem objetiva ao se observar a coluna com o percentual de zeros da tabela, que todas as coleções apresentam uma quantidade de zeros em mais de 90% das frequências dos termos presentes na $tf - idf$, o que deixa explícito o peculiar problema dos dados esparsos já caracterizado ao longo do texto, como algo inerente aos dados textuais e que

afeta negativamente grande parte dos resultados do processo de mineração de textos.

| | |
|---|---|
| **documentos** | número de documentos presentes na coleção |
| **termos** | número de termos existentes na coleção após o pré-processamento |
| **% zeros** | número relativo de zeros na *tf-idf*, quantificando o quanto a matriz é esparsa |
| **classes** | número de classes presentes na coleção |
| **n-gramas** | quantidade de termos considerados sequencialmente na coleção |

*Tabela 1: Descrição das características objetivas presentes em coleções textuais elenca-das para este trabalho*

| 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 2: Características das coleções textuais utilizadas nesta pesquisa*

## 3. RESULTADOS

### 3.1 REFINAMENTO COM O ALGORITMO PFCM

Conforme ficou evidenciado, a tarefa de organizar de maneira flexível um conjunto de documentos textuais, possui diversos desafios. Em particular, ao se agrupar um conjunto de documentos é esperado que os grupos resultantes possuam significado relevante, ou seja o algoritmo de agrupamento precisa detectar a estrutura natural dos documentos [15]. Alguns desses desafios estão na dificuldade em escalar os métodos usuais para coleções textuais de grande dimensionalidade; na obtenção de mecanismos efetivos para se avaliar a qualidade dos grupos produzidos; nas técnicas para se medir a interpretabilidade dos resultados; na capacidade para estimar os parâmetros dos algoritmos; na possibilidade de funcionar de maneira incremental, reduzindo o custo computacional durante a atualização dos grupos com novos dados; e, também na capacidade de continuar a produzir bons resultados em cenários compostos de documentos ruidosos.

Diante dos desafios propostos, e com a evidência de que é possível aprimorar os resultados ao se utilizar novas estratégias de agrupamento, a investigação apresentada nesta seção tem como objetivo analisar de qual forma a organização de documentos pode ser otimizada, ao aplicar na etapa de agrupamento uma estratégia que misture as

partições possibilística e fuzzy, por meio do algoritmo PFCM.

Vale ressaltar, que a escolha desse algoritmo foi feita devido o seu potencial para absorver as qualidades presentes no Fuzzy C-Means (FCM) contrabalanceando as suas deficiências ao agregar também o Possibilistic C-Means (PCM) e sua partição possibilística. Além disso, existem diversas pesquisas na literatura abordando o desempenho do PFCM, como por exemplo em [11].

Sendo assim, foram conduzidos experimentos adaptando a estratégia de organização flexível de documentos definida em Nogueira (2013), utilizando na etapa de agrupamento o método PFCM. Uma vez que esse método produz duas partições, uma possibilística e uma fuzzy, foi aplicado o método de extração de descritores SoftO-FDCL na partição fuzzy e também na partição possibilística, produzindo assim dois grupos de descritores.

Com essa adaptação espera-se uma melhor organização dos documentos, de forma que melhores descritores sejam escolhidos para caracterizar grupos. Tal processo de organização é ilustrado na Figura 4.

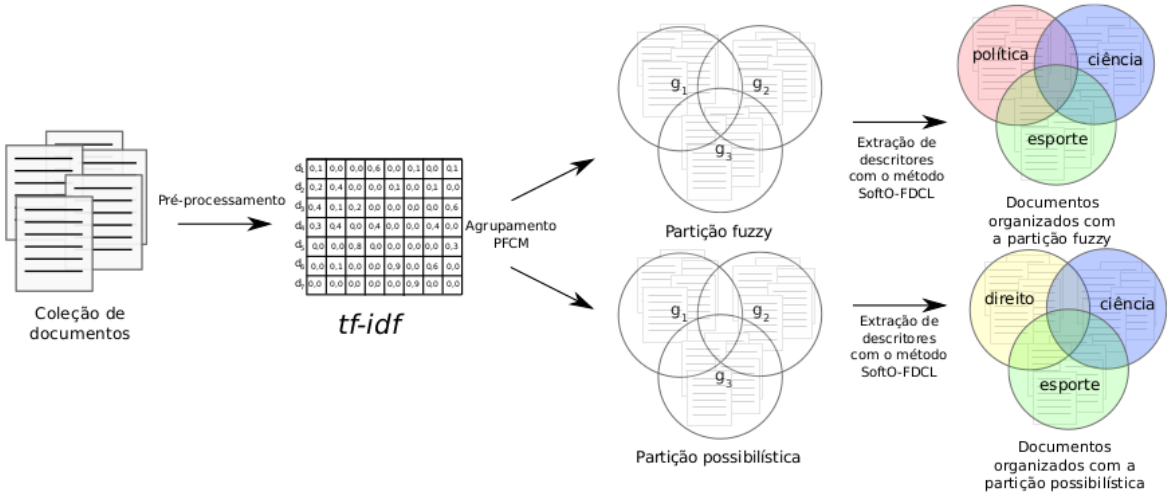


*Figura 4: Estratégia de organização flexível de documentos adotada ao se misturar abordagens fuzzy e possibilísticas no agrupamento.*

Para se calcular a quantidade ótima de grupos, para cada coleção foi utilizado o método da silhueta fuzzy, método bastante utilizado com o propósito de avaliar o agrupamento de documentos. Assim sendo, o número ideal de grupos é determinado após a execução da silhueta fuzzy variando o número de grupos entre 2 e o número de classes de cada coleção. Ressalta-se que em coleções que os documentos não possuem rótulos, ou seja o número de classes é desconhecido, ainda é possível usar o método da silhueta fuzzy para definir o número ótimo de grupos. No entanto, a quantidade máxima de grupos deve ser definida de modo empírico ou com base em alguma informação prévia a respeito dos dados.

Para permitir uma análise comparativa dos resultados, o experimento foi realizado também com os algoritmos FCM e PCM. Como resultado do agrupamento das coleções, está disposto na Tabela 3 a comparação do número de grupos obtidos por cada algoritmo de agrupamento. Nessa comparação nota-se que os algoritmos FCM e PFCM foram os que alcançaram uma quantidade de partições mais próxima da quantidade de classes existentes em cada coleção. Enquanto o PCM manteve uma tendência a produzir uma quantidade menor de grupos em relação aos demais.

| 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 3: Quantidade ótima de grupos determinada através do método da silhueta fuzzy para cada algoritmo de agrupamento*

Após agrupar os dados utilizando os métodos FCM, PCM e PFCM, foi aplicado o método de extração de descritores SoftO-FDCL. Para avaliar os descritores produzidos, foi verificado o potencial preditivo dos mesmos, possibilitando assim quantificar a qualidade dos termos selecionados para nomear os grupos.

Visando avaliar a qualidade dos descritores e permitir uma comparação direta dos impactos dessa abordagem com os resultados publicados em Nogueira (2013). O resultado dos agrupamentos foi submetido aos algoritmos de classificação SVM, Naive Bayes, Multinomial Naive Bayes, KNN e C4.5, que são bem comuns na avaliação de métodos de aprendizado de máquina.

Nesse contexto, foi utilizada a implementação dos algoritmos de classificação anteriormente citados presentes na ferramenta WEKA. Os algoritmos Naive Bayes (NB), Multinomial Naive Bayes (NB-Multinomial) e o J48 (que é a implementação do C4.5 existente no WEKA), foram executados com os parâmetros padrão da ferramenta. Por outro lado, o SVM foi ajustado para usar o Normalized Polynomial Kernel com o parâmetro de complexidade sendo c = 2.0. O algoritmo IBk (implementação do KNN presente no WEKA) foi executado 7 vezes, variando o parâmetro de vizinhos de 1 até 7, sendo escolhido o melhor resultado. Ressalta-se que foi adotada a técnica 10-fold cross validation no experimento para melhor capturar a capacidade de generalização do modelo.

O resumo dos resultados do desempenho dos descritores extraídos após o agrupamento com os algoritmos FCM, PCM e PFCM é apresentado na Tabela 4. Na tabela, está marcado os métodos de agrupamento que obtiveram a maior taxa de classificação dentre os demais. Esses resultados obtidos reforçam a flexibilidade e adaptação do método SoftO-FDCL [7] a novos algoritmos de agrupamento, demonstrando-se promissor na tarefa de extrair termos relevantes dos grupos produzidos na etapa de agrupamento.

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

*Tabela 4: Sumário dos resultados da classificação dos descritores*

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

*Tabela 5: Sumário dos resultados da classificação dos descritores*

Adicionalmente, ressalta-se a importância também de avaliar de maneira subjetiva os descritores selecionados dos grupos, permitindo compreender se os termos obtidos fazem sentido para a organização de documentos em grupos.

| Método | $crisp_1$ | $crisp_2$ | $crisp_3$ |
|---|---|---|---|
| **FCM** | easy, clear, drive, display, control, car, version, nice, work, perfect | fact, import, isn't, model, problem, unit, design, don't, doesn't, found | breakfast, nearby, concierge, eat, bottle, coffee, floor, food, inn, friendly |
| **PCM** | easy, read, problem, version, don't, small, nice, car, work, found | fact, back, turn, expect, size, close, quality, review, min, feature | feel, amazing, isn't, extreme, drive, include, point, reason, give, run |
| **PFCM** $\mu$ | easy, drive, control, don't, version, nice, car, work, perfect, lot | fact, isn't, read, complete, device, display, size, doesn't, found | breakfast, nearby, pleasant, concierge, eat, coffee, floor, clean, friendly, food |
| **PFCM** $\lambda$ | club, immaculate, send, towel, basic, exception, spotl, pillow, typical, fridge | pub, housekeep, holiday, tourist, tea, smoke, pm, renovate, facilite, london | usual, central, forum, bottle, modern, adult, supply, food, reserve, dinner |

*Table 6: Descritores extraídos com os métodos de agrupamento FCM, PCM e PFCM da coleção Opinosis, onde μ e λ se referem as partições fuzzy e possibilística respectivamente, da qual os descritores foram extraídos.*

Sendo assim, para uma análise subjetiva dos resultados, os descritores da coleção Opinosis, foram selecionados por possuirem poucos grupos facilitando a análise e a visualização. A coleção Opinosis contém opiniões dos usuários a respeito de serviços de hospedagem, dispositivos eletrônicos e carros e espera-se que os descritores de grupos se aproximem semanticamente de tais categorias. Na Tabela 4 tem-se a seleção de descritores escolhidos para cada grupo, extraídos pelos algoritmos FCM, PCM e PFCM. Ao analisar os descritores selecionados é possível notar uma tendência geral, do grupo crisp1 conter descritores relacionadas a carros, o crisp2 conter descritores sobre dispositivos eletrônicos e o grupo crisp3 descritores sobre hospedagem e alimentação. Contudo, nota-se que os descritores do PCM e do PFCMλ (descritores da partição possibilística do PFCM) estão um pouco mais misturados, não apresentando uma tendência geral bem definida. Uma explicação possível a esse resultado pode se encontrar na própria partição possibilística a qual permite que um mesmo documento possua um grau de tipicidade elevado em todos os grupos. Neste contexto, uma solução possível pode ser uma adaptação do método de extração de descritores SoftO-FDCL voltado para a partição possibilística, assim como também para algoritmos híbridos com duas partições, que é o caso do PFCM.

### 3.2 UMA ABORDAGEM HÍBRIDA PARA EXTRAÇÃO DE DESCRITORES

Nos experimentos anteriores foi identificado um possível problema ao realizar a extração dos descritores de maneira separada em cada partição do PFCM, assim como também foi apontado que o método pode não capturar toda essência da partição possibilística, que difere da partição fuzzy do FCM por não possuir a restrição que obriga a soma das pertinências de um grupo ser igual a um. Logo, é intuitivo indagar que para uma melhor interpretação dos grupos produzidos em um método de agrupamento híbrido, seja pertinente utilizar também uma abordagem mista de extração de descritores. Aproveitando-se assim dos benefícios existentes na partição possibilística, a qual penaliza os elementos ruidosos, com baixos valores de tipicidade, sem abrir mão das vantagens presentes na partição fuzzy. Para isso é necessário compreender os mecanismos de fun-

cionamento do método SoftO-FDCL, para que seja possível propor uma adaptação para este contexto.

### 3.2.1 INTERPRETANDO CORRETAMENTE AS TIPICIDADES

A interpretação direta dos graus de compatibilidade possibilísticos gera uma série de problemas na extração de descritores. Com isso, podemos formular a seguinte pergunta: Como interpretar corretamente os graus de compatibilidade possibilísticos para corretamente identificar os documentos relevantes de um dado grupo? Sabe-se que o valor de tipicidade pode variar livremente entre o intervalo [0, 1], sem a restrição probabilística do FCM. Essa é uma característica positiva introduzida em Krishnapuram e Keller (1993), a qual atribui valores de pertinência mais justos aos grupos fuzzy, em consonância com a teoria de conjuntos fuzzy, para melhor compreendermos esse conceito de valores mais justos.

Nesse contexto, propõe-se realizar tal interpretação em duas etapas. A primeira será constituída de uma conversão da tipicidade oriunda do PCM para a pertinência do FCM, de maneira a se satisfazer a restrição probabilística do FCM. No entanto, ao apenas realizar a conversão perde-se a robustez contra ruídos do PCM. Por isso, é possível contornar essa situação adicionando uma penalidade ao cálculo da pontuação dos termos.

A conversão proposta dos valores de tipicidade para pertinência, dar-se a como apresentado na Figura 5, a qual resolve o problema de considerar um documento como relevante em todos os grupos.

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

*Figura 5*

### 3.2.2 MÉTODO DE EXTRAÇÃO DE DESCRITORES PDCL

Agora que a proposta de interpretação da partição possibilística está concluída, o método para extração de descritores para a partição possibilística, o qual será denominado aqui de PDCL (Possibilistic Descriptor Comes Last) é proposto. Para realizar a extração de descritores, o método PDCL considera inicialmente todos os termos como candidatos. Em seguida, para cada grupo os valores de precisão e recuperação de todos os termos são calculados. A partir destes valores, a pontuação de cada termo no grupo é calculada, utilizando-se a medida f. A partir dessa pontuação por grupo dos termos candidatos, deve-se selecionar os descritores de maior pontuação em cada grupo. A quantidade de descritores é definida pelo usuário.

### 3.2.3 MÉTODO DE EXTRAÇÃO DE DESCRITORES Mixed-PFDCL

Uma das características presentes no método PFCM, é a adição dos parâmetros *a* e *b* que atuam como reguladores da influência do FCM e do PCM no agrupamento ob- tido. Portanto, é importante destacar a relevância de tais parâmetros no processo de extração de descritores, objetivando assim mais coerência com o algoritmo e exatidão nos resultados.

Nesse contexto, é também proposta a combinação dos valores de pertinência e de tipicidade convertendo-os em um único grau de compatibilidade conforme Figura 6. Essa combinação refere-se à média ponderada dos valores de pertinência e tipicidade pelos parâmetros a e b do método PFCM.

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

*Figura 6*

Como resultado dessa combinação, a estratégia definida anteriormente na Figura 4 é alterada, eliminando a dupla extração de descritores do método SoftO-FDCL do agrupamento produzido por meio do PFCM, na abordagem híbrida de extração de descritores com o método Mixed-PFDCL proposto. Essa nova estratégia proposta está contextualizada na Figura 7.

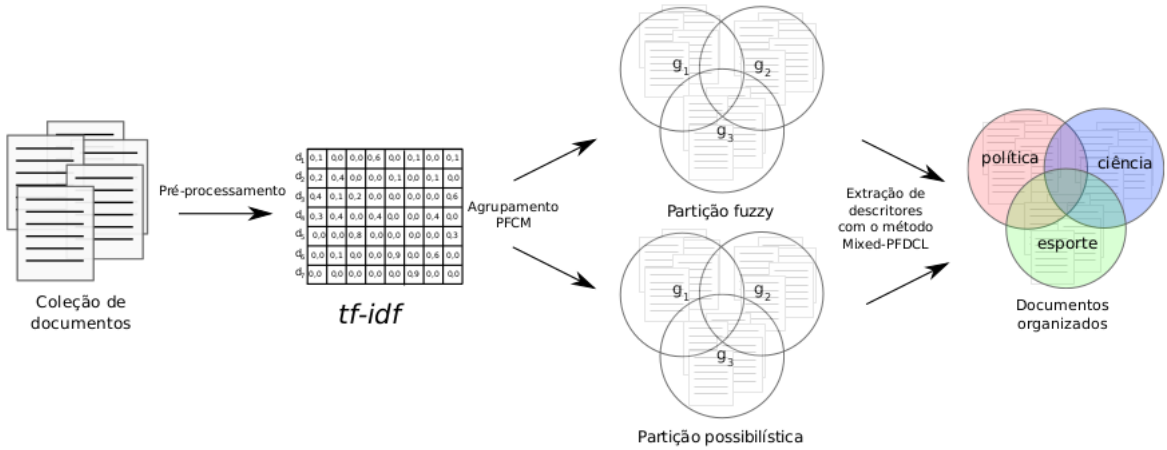


*Figura 7: Estratégia híbrida proposta para uma organização flexível de documentos com o agrupamento com o método PFCM e a extração de descritores com o método Mixed-PFDCL.*

Para mensurar os impactos das duas propostas (PDCL e Mixed-PFDCL) para extração de descritores apresentadas aqui nesta seção, foi realizado outro experimento, com os algoritmos PCM e PFCM. Durante o experimento fora utilizado os métodos de extração de descritores SoftO-FDCL, PDCL, Mixed-PFDCL.

Uma vez que as propostas aqui apresentadas têm por objetivo otimizar os resultados apresentados anteriormente, foi adotada uma metodologia similar aos experimentos anteriores. Ou seja, o agrupamento final obtido para cada base, é resultado dos grupos que obtiveram o maior valor na medida de silhueta fuzzy. Onde a quantidade de grupos para cada base, variou entre 2 e o número de classes de cada base (Tabela 7). Ressalta-se ainda que para minimizar os efeitos da aleatoriedade da partição inicial nesse experimento, o agrupamento foi executado 5 vezes para cada quantidade de grupos na silhueta fuzzy.

As coleções foram então agrupadas com os métodos PCM e PFCM, utilizando a metodologia descrita. A quantidade ótima de grupos, obtida com o método da silhueta fuzzy, está apresentado na Tabela 7. Os resultados apresentado nesta tabela, reforçam as conclusões apresentadas experimento anterior, de que a quantidade de grupos ótima do método PFCM tende a se aproximar mais da quantidade original de classes de cada coleção, enquanto o método PCM possui uma tendência em obter um número de grupos bem inferior a quantidade original de classes.

| Coleção | # classes | PCM | PFCM |
|---|---|---|---|
| Opinosis | 3 | 2 | 3 |
| 20Newsgroup | 4 | 4 | 4 |
| Hitech | 6 | 2 | 6 |
| NSF | 16 | 2 | 8 |
| WAP | 20 | 2 | 17 |
| Reuters-21578 | 43 | 4 | 40 |

*Tabela 7: Quantidade ótima de grupos determinada através do método da silhueta fuzzy para cada algoritmo de agrupamento no segundo experimento conduzido com os métodos PCM e PFCM*

Em seguida, foi realizado a extração dos descritores sobre as partições ótimas encontradas por cada algoritmo de agrupamento, sobre as coleções textuais. Como a motivação desse experimento, foi avaliar a qualidade dos descritores produzidos pelos métodos PDCL e Mixed-PFDCL propostos, a extração de descritores foi também realizada com o método SoftO-FDCL, possibilitando assim compararmos os resultados. E assim como no experi- mento anterior, essa análise quantitativa dos descritores produzidos, foi feita, utilizando a mesma estratégia de avaliação preditiva, com os 5 algoritmos de classificação do ex- perimento anterior.

O sumário desses resultados consta na Tabela 8, onde está marcado o método que obteve maiores taxas de acerto entre os 5 algoritmos de classificação utilizados. Como nessa investigação o propósito foi comparar os métodos de extração de descritores, dividiu-se o sumário de resultados de acordo com o método de agrupamento, PCM e PFCM respectivamente. Ressalta-se ainda, que garantir que a extração fosse realizada sobre os mesmos grupos, ambos os métodos de extração de descritores foram aplicados simultaneamente ao agrupamento. Portanto, os métodos SoftO-FDCL e PDCL foram aplicados ao mesmo agrupamento produzido pelo PCM, enquanto o SoftO-FDCL e o Mixed-PFDCL foram executados no mesmo agrupamento gerado pelo PFCM.

Os resultados dispostos nesse sumário, corroboram a hipótese formulada a respeito da interpretação das partições possibilísticas e híbridas no contexto da extração de descritores para a organização flexível de documentos. Pois, como se observa na Tabela 8, o método PDCL e o Mixed-PFDCL, superam os resultados do método SoftO-FDCL, em ambos os algoritmos de agrupamento. Embora, tenha existido 2 empates na comparação entre os métodos SoftO-FDCL e PDCL, para as coleções 20newsgroups e NSF.

| | PCM | | PFCM | |
|---|---|---|---|---|
| **Coleção** | **SoftO-FDCL** | **PDCL** | **SoftO-FDCL** | **Mixed-PFDCL** |
| Opinosis | | ✓ | | ✓ |
| 20newsgroups | ✓ | ✓ | | ✓ |
| Hitech | | ✓ | | ✓ |
| NSF | ✓ | ✓ | | ✓ |
| WAP | | ✓ | | ✓ |
| Reuters-21578 | | ✓ | | ✓ |

*Tabela 8: Sumário dos resultados da classificação dos descritores extraídos com osmétodos SoftO-FDCL, PDCL e Mixed-PFDCL*

## 4. DISCUSSÃO

Ficou evidenciado, a partir das pesquisas encontradas na literatura, a diversidade de estratégias existentes na literatura para mitigar os desafios existentes. Foi visto ainda que tem sido frequentemente propostas abordagens para aprimorar todas as etapas existentes no processo de organização de documentos, as quais contemplam o pré-processamento, agrupamento, extração de descritores e recuperação da informação.

A partir do estudo da teoria e dos fundamentos necessários para o tema, foi apresentado o conteúdo abordando os detalhes dos métodos de agrupamento FCM, PCM, PFCM.

A primeira contribuição experimental dessa pesquisa, consistiu no estudo dos impactos de se adicionar o método PFCM no processo de organização flexível de documentos. A partir desse estudo, foi observado que o algoritmo PFCM possui uma tendência para aumentar a eficiência do agrupamento produzido em coleções textuais de maior dimensionalidade, o que foi comprovado a partir das evidências contidas na Tabela 4. No entanto, se destaca aqui a importância da realização de novos estudos com um maior número de coleções textuais de baixa e alta dimensionalidade para se confirmar esta tendência.

Por outro lado, constatou-se nesse experimento a capacidade de adaptação do método SoftO-FDCL a novos algoritmos de agrupamento. No entanto, este último método, segundo as suas pesquisas iniciais, considera somente uma única partição no processo de pontuação dos termos candidatos, o que por sua vez não consegue capturar toda a essência do agrupamento produzido pelo PFCM. Ainda nesse experimento foi observado um problema na interpretação das tipicidades contidas em partições possibilísticas, no processo de extração de descritores. Esse problema deriva diretamente da natureza probabilística dos graus de pertinência da partição fuzzy do FCM, que não se aplica às tipicidades, a qual influencia diretamente na adequação do limiar do método SoftO-FDCL.

Feitas essas considerações, apresentam-se as duas principais contribuições dessa pesquisa. A primeira consiste no método PDCL, que propõe uma abordagem para interpretar os graus de compatibilidade possibilísticos, sem deixar de lado a resiliência contra ruídos inerente das tipicidades. Os experimentos conduzidos com esse método demonstraram a qualidade dos descritores extraídos com o PDCL em comparação com o método SoftO-FDCL, cujos comparativos estão apresentados na Tabela 8. Também observou-se neste experimento que o método PCM produziu uma quantidade baixa de grupos em comparação ao número total de classes em cada coleção.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS (máximo 15)

[1] 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.

[2] HUANG, G. et al. A data as a product model for future consumption of big stream data in clouds. In: 2015 IEEE International Conference on Services Computing. Institute of Electrical & Electronics Engineers (IEEE), 2015. Disponível em: hhttp://dx.doi.org/10.1109/SCC.2015.43i.

[3] MUGGLETON, S. H. 2020 computing: Exceeding human limits. Nature, Nature Publishing Group, v. 440, n. 7083, p. 409–410, mar 2006. Disponível em: hhttp://dx.doi.org/10.1038/440409ai.

[4] 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: hhttp://dx.doi.org/10.1007/978-1-84800-046-9 6i.

[5] AGGARWAL, C. C.; ZHAI, C. An introduction to text mining. In: Mining Text Data. Springer Science + Business Media, 2012. p. 1–10. Disponível em: hhttp://dx.doi.org/10.1007/978-1-4614-3223-4 1i.

[6] BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern information retrieval: the concepts and technology behind search. Choice Reviews Online, American Library Association, v. 48, n. 12, p. 48–6950–48–6950, aug 2011. Disponível em: hhttp://dx.doi.org/10.5860/choice.48-6950i.

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

[8] 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: hhttp://www.sciencedirect.com/science/article/pii/0098300484900207i.

[9] MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. [S.l.], 1967. v. 1, n. 14, p. 281–297.

[10] 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.

[11] 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.

[12] 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: hhttp://www.sciencedirect.com/science/article/pii/S1877050913001385i.

[13] FELDMAN, R.; SANGER, J. The text mining handbook: Advanced approaches in analyzing unstructured data. [S.l.]: Cambridge University Press, 2007.
[14] 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.
[15] 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.

## 6. ATIVIDADES REALIZADAS NO PERÍODO

- Leitura de artigos relacionados ao tema proposto.
- Desenvolvimento de métodos de agrupamento derivados do Fuzzy C Means.
- Desenvolvimento de métodos de métodos de extração de descritores.
- Implementação dos métodos de agrupamento clássicos da literatura.
- Realização de diversos experimentos comparativos.
- Escrita de artigo científico.
- Participação de reuniões do grupo de pesquisa.

## 7. PARTICIPAÇÃO EM REUNIÕES CIENTÍFICAS E PUBLICAÇÕES

Os resultados desta pesquisa foram detalhados no artigo científico em anexo, cuja submissão teve ótimas revisões e foi aprovado para publicação em uma das mais importantes conferências da área de sistemas fuzzy:

*Carvalho, N. V. J., Rezende S. O., Camargo H. A., Nogueira T. M. Flexible Document Organization by Mixing Fuzzy and Possibilistic Clustering algorithms. IEEE International Conference on Fuzzy Systems (FUZZ- IEEE), p. 1–8, 2016.*

## 8. DIFICULDADES ENCONTRADAS / CAUSAS E PROCEDIMENTOS PARA SUPERÁ-LAS

Por conta da excessiva carga de processamento requisitada nos experimentos, cada sequência de testes demandava demasiado tempo para conclusão, causando assim um atraso na obtenção dos resultados. Como solução de contorno foi alugado um serviço na nuvem para realizar o processamento. No entanto, ressalta-se que idealmente seria necessário o acesso a computadores com elevado potencial de processamento para a redução do tempo dos experimentos, possibilitando assim a realização de mais experimentos.

Salvador, _____ de_____ de 2016.

| Estudante | Orientador (a) |
|---|---|

Secretaria do Programa
Rua Basílio da Gama, 06. Canela.
Salvador – BA. 40.110-040.
Tel.: 71 3283-7968 Fax: 71 3283-7964
E-mail: pibic@ufba.br

# Flexible Document Organization by Mixing Fuzzy and Possibilistic Clustering algorithms

Nilton V. Carvalho Jr.
Department of Computer Science
Federal University of Bahia– Brazil
{niltonvasques@dcc.ufba.br}

Solange O. Rezende
Institute of Mathematics and Computer Science
University of São Paulo – Brazil
{solange@icmc.usp.br}

Heloisa A. Camargo
Department of Computer Science
Federal University of São Carlos – Brazil
{heloisa@dc.ufscar.br}

Tatiane M. Nogueira
Department of Computer Science
Federal University of Bahia– Brazil
{tatianenogueira@dcc.ufba.br}

***Abstract*—A powerful and flexible organization of documents can be obtained by mixing fuzzy and possibilistic clustering. In such organization, documents can belong to more than one cluster simultaneously with different compatibility degrees. Clusters represent topics, which are identified by one or more descriptors extracted by a proposed method. In this manuscript, we investigated whether or not the descriptors extracted after applying possibilistic fuzzy clustering improve the flexible organization of documents. Experiments were carried out on real-world document collections and we evaluated the ability of descriptors to capture the essential information in every dataset. Results have shown the effectiveness of extracting possibilistic fuzzy cluster descriptors, improving the flexible organization of documents.**

***Keywords*—fuzzy clustering, possibilistic clustering, flexible organization, documents, text mining**

## I. Introduction

A lot of technologies as social media, mobile computing, and Internet of Things (IoT) comprises the big data problem [1], which has become a challenging problem in the proposal of computational methods capable of dealing with an exorbitant amount of data generated every day [2], [3], [4]. This problem is even more challenging when carried out the flexible organization of the data stored in textual format. Such data covers a wide set of topics that are constantly updated[5] and the organization of them in categories can not be predefined.

According to [6], system flexibility means the ability of a system to manage imprecise and/or uncertain information inherent to the data. To illustrate the usefulness of such a flexibility in the management of data represented by textual documents, consider a context in which news are organized in categories according to their main topic. Consider a news (textual document) with the title "*Experts affirm the adventure sport strengthens heart health*", which addresses complementary topics: *Sports* and *Health*. This news can be assigned to distinct categories: the categories related to *Sports* topic or the categories related to the *Health* topic. Nevertheless, the cited news deals with both topics simultaneously, which suggests that the assignment of this news to categories that represent both topics would be more appropriate than choosing predefined categories that represents just one of them.

Therefore, supposing that a user is requiring documents of the *Sports* topic, if the cited document is assigned only to the *Health* topic, this document would not be recovered for the user, despite being useful for his/her requirements.

One way to develop this kind of management of documents is by means of clustering. Clustering algorithms group documents that share many terms, what is an indication that the content of these documents is similar. Document clustering is used in a variety of applications because if there is a document in a cluster that is relevant to a user, then it is likely that other documents from the same cluster are also relevant [7], [8], [9].

Furthermore, to overcome the drawback concerning multi-topic documents, there are clustering algorithms designed to produce overlapping clustering solutions [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Overlapping clustering algorithms scatter a document collection so that each document may belong to different clusters with different membership degrees. The interpretation of these membership degrees can be used to quantify the compatibility of a document with a topic, which is identified by cluster representatives.

Usually, the cluster representatives are probabilistic models or cluster prototypes. However, in the document clustering, representatives such as the cluster prototype are not very useful to identify the topic addressed by the documents in each cluster. We have proposed a method to automatically discover overlapping cluster descriptors [20][21], which are terms present in the documents and significant to the topic addressed in the documents. Since documents are represented by a high dimensional feature space, the extraction of good descriptors is a big problem to be solved. The extraction of cluster descriptors is a problem even bigger in the flexible organization of documents using overlapping clustering, since the same descriptor can be representative for more than one cluster with different weights of representativeness.

The method proposed in [20] extracts the best descriptors of a cluster from a rank of descriptor candidates. It can extract descriptors after the fuzzy/possibilistic document clustering in order to achieve the flexible organization of documents.

In addition, we have proposed in [22] an investigation comparing the performance of the extraction of meaningful

overlapping cluster descriptors when using the well known Fuzzy C-Means (FCM) [23] and Possibilistic C-Means (PCM) [24] clustering algorithms, since the method proposed to organize documents in a flexible way is dependent on degrees of compatibility of each document with each cluster there are obtained after the clustering process. The proposed investigation was conducted using document cluster descriptors as features for text categorization.

In [22], we have concluded that each algorithm should be used depending on the collection characteristics. Collections composed by a large number of terms have better results when the cluster descriptors are extracted after the FCM in stead of after PCM. This is due to the fact that the PCM presents the problem of coincident clusters which is increased by the big number of terms. At the same time, the PCM is a good alternative when the collection presents noise data.

Considering that a balance between noise data identification and classification results is the key for the choice of a good clustering algorithm for flexible documents organization, we have compared the already obtained results with results obtained from clustering algorithms that mixes both FCM and PCM.

The fuzzy possibilistic c-means (FPCM) [25] is one of that algorithms. It generates both membership and typicality values when clustering unlabeled data. FPCM constrains the typicality values so that the sum over all data points of typicalities to a cluster is one. However, according to the authors, the row sum constraint produces unrealistic typicality values for large data sets. Therefore, later, they proposed a new model called possibilistic fuzzy c-means (PFCM) [26] clustering algorithm. PFCM produces memberships and possibilities simultaneously, along with the usual point prototypes or cluster centers for each cluster. PFCM is a hybridization of possibilistic c-means (PCM) and fuzzy c-means (FCM) that often avoids various problems of PCM, FCM and FPCM, such as the noise sensitivity defect of FCM, overcomes the coincident clusters problem of PCM and eliminates the row sum constraints of FPCM.

For this reason, we have investigated whether or not the PFCM can be used for flexible document organization.

To present such investigation, this paper is organized as follows. In Section II, basic concepts concerning flexible organization of documents and overlapping cluster descriptor extraction are reviewed. In Section III, the experimental results concerning the performance of the descriptors extracted after FCM, PCM and PFCM are presented, followed by discussions about the achieved results. Finally, in Section IV, the conclusion and the future directions of this research is also presented.

## II. Flexible organization of documents

In this section, we review the basic concepts and the methods used in our approach proposed in [20] and improved in [21] to organize documents in a flexible way.

### A. Document preprocessing

The preprocessing of documents is necessary to structure the documents in order to make them processable by the algorithms of pattern extraction. The most common output of a document preprocessing is the representation of a document collection in a vector space in the form of a document-term matrix. Each matrix row corresponds to one document in the collection and each matrix column (attribute) corresponds to one term in the entire collection of documents.

The terms in the document-term matrix are first examined in an initial effort to disregard terms that do not represent useful knowledge. In this step of examination, three tasks are very common: (1) Elimination of stopwords, which are words that are not relevant in the analysis of documents and usually consist of prepositions, pronouns, articles, interjections, among others; (2) Stemming, a technique that reduce the words to their root form in order to reduce the number of terms needed to represent the document collection; (3) $n$-gram extraction, which is the extraction of terms represented by $n$ consecutive words, since words that occur in sequence in the document may contain more information than isolated words.

After selecting the terms that represent the document collection, for the proposed approach, the document-term matrix contains in its cells the ratio between the frequency of a particular term in a document and the inverse of the frequency of this term in the document collection ($tf$-$idf$ Term Frequency-Inverse Document Frequency). By this measure, the importance of the terms in a document is weighted, so that terms which are present in a lot of documents have a smaller weight than the terms that occur more rarely in the collection.

The definition of $tf$-$idf$ and preprocessed document-term matrix is presented next.

*Definition 2.1:* Let $D$ be a document collection and $d$ a general document in $D$. The frequency of a term $t$ in document $d$, denoted by $tf(t,d)$, is the number of times that $t$ occurs in $d$. The inverse of the frequency of the term $t$ in the collection $D$ is given by $idf(t) = log\frac{n}{d(t)}$, where $n$ is the number of documents in $D$ and $d(t)$ is the number of documents in $D$ where $t$ occurs. The measure $tf - idf$ (term frequency-inverse document frequency) of a term $t$ in a document $d$ from a colection $D$ is defined as $tf - idf(t,d) = tf(t,d) \times idf(t,D)$.

*Definition 2.2:* Consider a document collection $D = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_n\}$ and let T be the number of terms in the collection. A document-term matrix $W = [d_{kj}]$ is composed by a document $\mathbf{d}_k$ in each row such that each column corresponds to a term $t_j$, $j = 1, ..., T$. A document $\mathbf{d}_k$ is represented by a vector $[d_{k1}, d_{k2}, ..., d_{kT}]$, $1 \leq k \leq n$. This vector comprises the frequency of each term $t$ in the document $\mathbf{d}_k$, weighted by how often this term occurs in the collection, i.e., $d_{kj} = tf - idf(t_j, d_k)$ (see Definition 2.1).

The document-term matrix is inherently high dimensional and sparse, which sometimes can make the document organization computationally very expensive or even impossible. This negatively affects the outcome of some knowledge extraction algorithms.

Therefore, to make the flexible organization of documents possible, the preprocessed documents are clustered by means of overlapping clustering algorithms described next.

### B. Overlapping document clustering

The most used clustering algorithms to organize data into overlapping clusters is the FCM, since by it a document can

belongs to more than one cluster with different membership degrees.

In addition, many other clustering algorithms that are extensions from it was suggested by researchers, trying solve some issues in FCM, as for example PCM[24] and PFCM[26].

One issue that needs to be considered when performing cluster analysis is the "curse of dimensionality" [27], which refering to any problem in data analysis that results from a large number of attributes. This problem is very frequent in text mining processing and fuzzy clustering, since the document-term matrix is too sparse with a high dimensionality [28].

The FCM [23] was proposed to be used to cluster low dimensional data and, because of that, uses the Euclidean distance in its process. However, this metric distance is not appropriate for high-dimensional and sparse data. According to [29], the similarity measures plays an important role to documents clustering and the use of cosine coefficient similarity is more appropriate.

Therefore, the FCM and its extensions (PCM and PFCM) clustering algorithms was slightly modified to handle the flexible organization of documents facing the high dimensionality problem. The modification was done in the similarity measure norm function, where was used the cosine coefficient similarity, and this measure is defined as follows.

*Definition 2.3:* Let $n$ be the number of documents in the collection and $c$ be the number of clusters. Consider a document $\mathbf{d}_k$, $k = 1, ..., n$, and a cluster prototype $\mathbf{v}_i$, $i = 1, ..., c$. The dissimilarity between a document and a prototype $\|\mathbf{d}_k - \mathbf{v}_i\|$ is measured using the cosine coefficient similarity according to Equation(1) and Equation (2).

$$sim(\mathbf{d}_k, \mathbf{v}_i) = cos\theta = \frac{\mathbf{d}_k \cdot \mathbf{v}_i}{|\mathbf{d}_k|\,|\mathbf{v}_i|} \in [0, 1] \tag{1}$$

$$\|\mathbf{d}_k - \mathbf{v}_i\| = 1 - sim(\mathbf{d}_k, \mathbf{v}_i) \in [0, 1] \tag{2}$$

Although by means of FCM, PCM and PFCM algorithms it is possible to obtain the compatibility of documents in more than one cluster, they represent different concepts of overlapping, which influence the compatibility degrees found in their clustering process. Therefore, each algorithm has its particularities and is performed as follows.

*1) Fuzzy C-Means:* The FCM [23] algorithm used for document clustering is an iterative process that updates the prototypes of the clusters defined initially from a fuzzy pseudo-partition and the partition matrix giving the membership degree of each document to each cluster. This update tries to minimize the dissimilarity between a document and a cluster prototype. The pseudo-partition is defined as follows [30].

*Definition 2.4:* Let $c$ be the number of clusters and $A_i(\mathbf{d}_k)$ the membership degree of the document $\mathbf{d}_k$ in the cluster $i$, $k = 1, ..., n$, $i = 1, ..., c$. A fuzzy pseudo-partition $U = [A_i(\mathbf{d}_k)]$ is a family of fuzzy sets of $D$ (see Definition 2.2) denoted by $\mathrm{P} = \{A_1, A_2, ..., A_c\}$, which satisfies the Equations (3) and (4).

$$\sum_{i=1}^{c} A_i(\mathbf{d}_k) = 1 \tag{3}$$

$$0 < \sum_{k=1}^{n} A_i(\mathbf{d}_k) < n \tag{4}$$

During the clustering procedure, the prototypes and the partition matrix are updated until a stopping criterion is satisfied. Let $n$ be the number of documents in the collection and $c$ be the number of clusters. The document cluster prototypes $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_c\}$ are calculated according to Equation (5), where $m > 1$ is a real number, called fuzzification factor, that controls the influence of the membership degrees in the fuzzy clustering. In the experiments presented in this paper, $m = 2.5$ was used for clustering all datasets.

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} [A_i(\boldsymbol{d}_k)]^m \boldsymbol{d}_k}{\sum_{k=1}^{n} [A_i(\boldsymbol{d}_k)]^m},\ i = 1, ..., c, k = 1, ..., n. \tag{5}$$

Further, based on Equation (5) and the definition of dissimilarity presented in Definition 2.3, the FCM algorithm updates the fuzzy pseudo-partition according to Equation (6).

$$\mu_{ik} = A_i(\boldsymbol{d}_k) = \frac{1}{\sum_{j=1}^{c} \left( \frac{\|\boldsymbol{d}_k - \boldsymbol{v}_i\|}{\|\boldsymbol{d}_k - \boldsymbol{v}_j\|} \right)^{\frac{1}{m-1}}} \tag{6}$$

The goal of FCM is to minimize the optimization function $J_m$, defined in Equation (7). The performance of FCM is based on the $J_m$ optimization under the fuzzy pseudo-partition $U$ defined in Definition 2.4.

$$J_m(U) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m \|\boldsymbol{d}_k - \boldsymbol{v}_i\| \tag{7}$$

Furthermore, before FCM starts running, the number of groups $c$, a small number $\epsilon$ as stopping criteria and a fuzzification factor $m$ must be defined.

*2) Possibilistic C-Means:* The membership degree $A_i(\boldsymbol{d}_k)$ that FCM assigns to a document $\boldsymbol{d}_k$ is related to the relative distance of $\boldsymbol{d}_k$ to the cluster prototype $\boldsymbol{v}_i$, $i = 1, ..., c$. If $\mathbf{d}_k$ is equidistant to two prototypes, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, the membership degree of $\boldsymbol{d}_k$ in each cluster will be the same: $A_1(\boldsymbol{d}_k) = 0.5$ and $A_2(\mathbf{d}_k) = 0.5$.

Let us consider a noise data as a document that is far but equidistant from the prototypes of two clusters. By means of FCM, noise data can be assigned to both clusters with the same membership degrees as the documents closer and also equidistant to the cluster prototypes. In Figure 1 we illustrate such situation, in which $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ have both the same membership degree, $0.5$, in the clusters $A_1$ and $A_2$, although document $\boldsymbol{d}_1$ is closer to the clusters prototypes than $\boldsymbol{d}_2$.

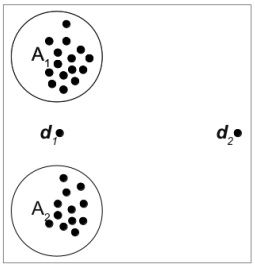


Fig. 1. Position of documents $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ related to clusters $A_1$ and $A_2$ (Adapted from [31])

According to Pal et al. in [26], such a situation illustrate the basic notion of probabilistic partitioning of data sets of FCM, which has the constraint $\sum_{i=1}^{c} A_i(\boldsymbol{d}_k) = 1$, i.e, the sum of a document membership degree in all clusters must be equal to 1. Therefore, the PCM algorithm was developed to relax this constraint of FCM, considering the absolute value of the distance of $\boldsymbol{d}_k$ from the cluster prototypes. Considering such looseness, the $A_i(\boldsymbol{d}_k)$ obtained by means of PCM should be interpreted as the typicality of a document $\boldsymbol{d}_k$ relative to cluster $i$.

In a similar way as FCM, the PCM [24] algorithm used for document clustering is an iterative process that updates the prototypes of the clusters defined initially from a pseudo-partition (degrees of typicality of every object in all clusters). This update tries to minimize the dissimilarity between a document and a cluster prototype. Its cluster prototype update is identical to that in Equation (5). Further, based on the definition of dissimilarity presented in Definition 2.3, the PCM updates the pseudo-partition according to Equation (8) [26].

$$\sigma_{ik} = A_i(\boldsymbol{d}_k) = \frac{1}{1 + \left( \frac{\|\boldsymbol{d}_k - \boldsymbol{v}_i\|}{\gamma_i} \right)^{\frac{1}{m-1}}} \tag{8}$$

The user-defined constant $\gamma_i > 0$ is considered to minimize the singularity problem of FCM. Therefore, the distance $\|\boldsymbol{d}_k - \boldsymbol{v}_i\|$ can be zero, relaxing the constraint in Equation (3). As recommended by Krishnapuram and Keller in [24], we have chosen $\gamma_i$ according to Equation (9).

$$\gamma_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m \|\boldsymbol{d}_k - \boldsymbol{v}_i\|}{\sum_{k=1}^{n} \mu_{ik}^m} \tag{9}$$

where $\mu_{ik}$ is a terminal FCM partition of $D$ according to Equation (6).

The goal of PCM is to minimize the optimization function $P_m$, according to Equation (10) with the tipicalities matrix $H = [\sigma_{ik}]$.

$$P_m(H) = \sum_{k=1}^{n} \sum_{i=1}^{c} {\sigma_{ik}}^m \|\boldsymbol{d}_k - \boldsymbol{v}_i\| + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} (1 - \sigma_{ik})^m \tag{10}$$

The first term of $P_m$ is just the function $J_m$, and in the absence of the second term, unconstrained optimization will lead to the trivial solution $\sigma_{ik} = 0$, $i = 1, ..., c$ , $k = 1, ..., n$. The second term of $P_m$ acts as a penalty which tries to bring $\sigma_{ik}$ toward 1 [26].

*3) Possibilistic Fuzzy C-Means:* According to [26], memberships (or relative typicalities) and possibilities (or absolute typicalities) are both important for correct interpretation of data substructure. When a data point needs to be crisply labeled, membership is a plausible choice as it is natural to assign a point to the cluster whose prototype is closest to that point. On the other hand, while estimating the centroids, typicality is an important means for alleviating the undesirable effects of outliers.

For that reason, [26] suggested the Possibilistic Fuzzy C-Means (PFCM) to improve PCM, by mixing the benefits from FCM and PCM. To solve the coincident cluster problem that happens in some databases when using the PCM, the PFCM uses the tipicality function from PCM and membership function from FCM, and assign a weight to each, in way that the user can adjust how much from each algorithm will be used. In Equation (11) follows the definition of the goal function that PFCM try to minimize.

The goal function presented in Equation (11) is constrained by $\sum_{i=1}^{c} \mu_{ik} = 1 \forall k$, $0 < \mu_{ik}, t_{ik} \leq 1$, $0 < a, b$ and $m, n > 1$. The membership function $\mu_{ik}$ is the same that is defined in Equation (6). The tipicality function $\sigma_{ik}$ was slightly modified, adding the constant $b$ as presentend in Equation (12) and, as recommended by [24] we also choose $\gamma_i$ according to Equation (9). Finally, the constants $a$ and $b$ define the relative importance of the membership values and typicality values respectively, and must be choose by user according with the problem.

$$\sigma_{ik} = \frac{1}{1 + (\frac{b}{\gamma_i} \|\boldsymbol{d}_k - \boldsymbol{v}_i\|)^{1/m-1}} \tag{12}$$

The equation that updates the PFCM cluster prototypes is adapted to take advantages of its strategy of mixing fuzzy and possibilistic clustering, presented in Equation (13).

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} [a(\mu_{ik})^n + b(\sigma_{ik})^m] \boldsymbol{d}_k}{\sum_{k=1}^{n} [a(\mu_{ik})^n + b(\sigma_{ik})^m]}, \; i = 1, ..., c, k = 1, ..., n. \tag{13}$$

According to Equation (13), if we use a high value of $b$ compared to $a$, then the centroids will be more influenced by the typicality values than the membership values. On the other hand, if we use a higher value of $a$ then the centroids will be more influenced by the membership values. Thus, to reduce the effect of outliers, we should use a bigger value for $b$ than $a$ [26].

Finally, once the documents are clustered, the overlapping cluster descriptors can be extracted by the method described next.

$$L(P) = \sum_{i=1}^{n}\sum_{j=1}^{c}[a(\mu_{ik})^n + b(\sigma_{ik})^m]\,\|\boldsymbol{d}_k - \boldsymbol{v}_i\| + \sum_{l=1}^{c}\gamma i \sum_{i=1}^{n}(1-\sigma_{ik})^m \tag{11}$$

### C. Overlapping cluster descriptor extraction

The method proposed in [20] carries out a procedure that uses an adaptation of the classical measures of information retrieval [32] namely precision, recall, and $f1$-measure, which is the weighted harmonic mean of precision and recall.

In the fuzzy and possibilistic cluster descriptor extraction, all the terms found in the document preprocessing step are initially considered as descriptor candidates. Additionally, a document $\mathbf{d}_k$ is considered to belong to cluster $i$ if it has a membership degree $A_i(\mathbf{d}_k) \geq s$, where $s = \frac{1}{c}$. The threshold $s$ is considered for two reasons. Firstly, its use allows the selection of descriptor candidates from documents that belong to more than one cluster with different compatibility degrees, instead of considering only the cluster with the highest compatibility degree. Secondly, using this threshold it is possible to penalize the descriptor candidates that occur in documents with low compatibility degree in a cluster.

A rank of terms weighted by their $f1$-measure is obtained for each cluster as follows, considering the contingency matrix presented in Table I.

TABLE I. CONTINGENCY MATRIX FOR INFORMATION RETRIEVAL MEASUREMENT

| | Documents of cluster $c$ (documents with compatibility degree in cluster $c$ higher than or equal to $s$) | Documents that are not in cluster $c$ (documents with compatibility degree in cluster $c$ lower than $s$) |
|---|---|---|
| Documents which have the descriptor candidate $t$ | *hits* | *noises* |
| Documents which do not have the descriptor candidate $t$ | *losses* | *rejects* |

i) Calculate the precision of a descriptor candidate $t$ in a cluster $c$:

$$p(t,c) = \frac{hits}{hits + noises} \tag{14}$$

ii) Calculate the recall of a descriptor candidate $t$ in a cluster $c$:

$$r(t,c) = \frac{hits}{hits + losses} \tag{15}$$

iii) Calculate the $f1$-measure of a descriptor candidate $t$ in a cluster $c$:

$$f1(t,c) = \frac{2 \cdot p(t,c) \cdot r(t,c)}{p(t,c) + r(t,c)} \tag{16}$$

Since the ranking of descriptor candidates is obtained, the descriptors are selected. We have presented in [22] how the compatibility degree influences the extraction of overlapped cluster descriptors. The number of descriptors to be selected depends on the application.

Next, some experimental results are presented concerning the improvement that can be obtained in the flexible document organization, and consequently the meaningful cluster descriptors extraction, by mixing fuzzy and possibilistic clustering algorithms.

## III. EXPERIMENTAL RESULTS

A variety of issues in cluster analysis is well-known, such as scalability to large data sets, effective means of evaluating the validity of clusters that are produced, easy interpretability of the results, ability to estimate any parameters required by the clustering technique, ability to function in an incremental manner, and robustness in the presence of different underlying data and cluster characteristics. Therefore, there is not expected that one type of clustering approach will be suitable for all types of data, even all high dimensional data [28].

The intention of the investigation presented here is to analyse how the flexible document organization can be improved by mixing fuzzy and possibilistic document clustering. Investigations concerning to the PFCM clustering algorithm performance, and its extensions, are presented in many other papers in the state of the art, as for example in [26], [33], [29], [34], [19].

We carried out the experiments using six different real datasets (document collections), whose general characteristics are summarized in Table II. To ensure diversity in the collections, they were obtained from different sources and they are about different subjects. Details about each dataset can be obtained in the references and/or by a summary presented in [22]. All collections were preprocessed using the Pretext[1] tool [35]. Any term that occurs in fewer than two documents was eliminated and 1-gram terms were selected, i.e, terms composed of one word.

TABLE II. DOCUMENT COLLECTIONS

| Dataset | # docs | # terms | Subject |
|---|---|---|---|
| Opinosis | 51 | 842 | Costumer reviews[36] |
| NSF | 1600 | 2806 | Abstracts[37] |
| Reuters-21578 | 1052 | 3925 | Newswire stories[38] |
| Hitech | 600 | 6925 | Newspaper stories[2] |
| WAP | 1560 | 8070 | Web pages[39] |
| 20Newsgroups | 2000 | 11028 | Usenet news[40] |

When the organization of documents is achieved using document clusters, it is important to evaluate its interpretability, also known as transparency, which means that the organization should map the content of the collection. Furthermore, as highlighted by Havens $et.al$ in [41], clustering algorithms are meant to find the natural groupings in unlabeled data (or to discover unknown trends in labeled data).

In this context, the number of clusters for FCM, PCM and PFCM algorithms was defined using the number of classes of each dataset as input for the Fuzzy Silhouette (FS) [42] method. Such a method is commonly used to evaluate document clustering and choose the best number of clusters for the document organization. Therefore, the number of clusters

[1] http://sites.labic.icmc.usp.br/pretext2/

is determined considering the best value of silhuette obtained from a number of clusters between 2 and the number of classes of each dataset. In situations where the number of classes are unknown, the number of clusters can also be determined by using the FS. However, the maximum number of clusters need to be defined empirically. The number of clusters compared with the number of classes is showed in table III. By this comparison, we check that FCM and PFCM were better successful in producing similar partitions to the labeled data than PCM.

TABLE III. OPTIMAL NUMBER OF CLUSTERS

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

After clustering the documents using FCM, PCM and PFCM, we have extracted fuzzy cluster descriptors from the obtained clusters. Next, we have checked the power prediction of the descriptors in the sense that they can be used as good attributes for text categorization.

We evaluated the predictive power of the descriptors considering each cluster as a class and the descriptors as document attributes. Since in fuzzy/possibilistic clustering the documents can belong to more than one cluster, the document class is the cluster in which it has the highest membership/typicality degree. After labeling each document in the collection with the corresponding cluster, an attribute-value matrix was created with each descriptor being an attribute. The matrix entries are the frequency of the descriptors in each document.

Using such attribute-value matrix, we have performed well-known classification algorithms of machine learning (SVM, Naive Bayes, Multinomial Naive Bayes, KNN and C4.5) in the same way as investigated in [22], since we intent to extend such investigation by the results presented here.

The classification was carried out using the WEKA tool [43]. The Naive Bayes (NB), Multinomial Naive Bayes (NB-Multinominal) and J48 algorithms (the weka implementation of the C4.5 classification method) were executed using the default parameters of the tool. However, the performance of the SVM was tuned up using the Normalized Polynomial Kernel and the complexity parameter c=2.0. The IBk (the weka implementation of the KNN classification method) was experimented ranging the number of neighbors from 1 to 7. The best result was obtained using 5 neighbors. The 10-fold cross validation method was used in all experiments.

The performance rates (correct classification rate) obtained from each classifier over each document collection are presented in Figures 2, 3, 4, 5, 6, and 7.

As the PFCM mixes membership values and typicality values, the performance rates were obtained by PFCM results from two extraction of descriptors: one using the membership values and the other using typicality values. After that, the classifiers is performed about PFCM typicality descriptors and PFCM membership descriptors. Thereafter, the average classification rate obtained between both descriptors was calculated and showed in Figures below.

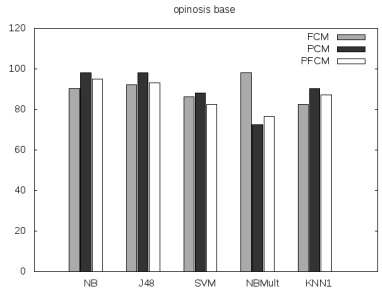


Fig. 2. Performance of cluster descriptors obtained from PCM, FCM, PFCM clustering algorithms (Opinosis collection)

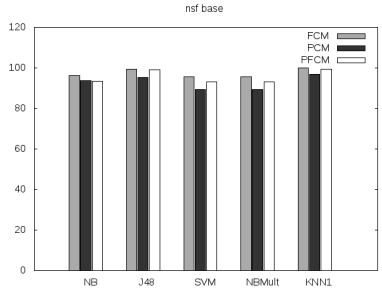


Fig. 3. Performance of cluster descriptors obtained from PCM, FCM, PFCM clustering algorithms (NSF collection)

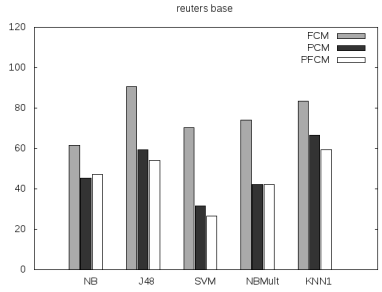


Fig. 4. Performance of cluster descriptors obtained from PCM, FCM, PFCM clustering algorithms (Reuters-21578 collection)

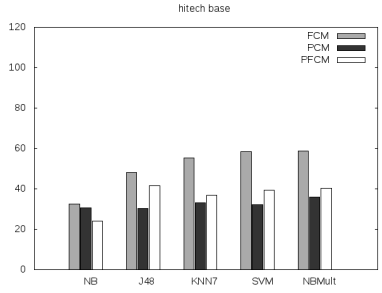


Fig. 5. Performance of cluster descriptors obtained from PCM, FCM, PFCM clustering algorithms (Hitech collection)

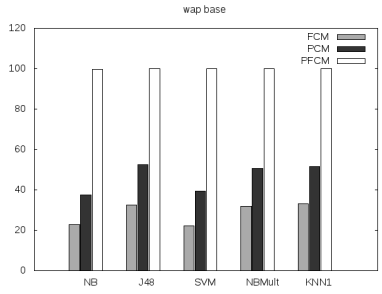


Fig. 6. Performance of cluster descriptors obtained from PCM, FCM, PFCM clustering algorithms (WAP collection)

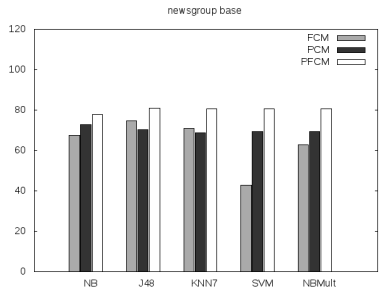


Fig. 7. Performance of cluster descriptors obtained from PCM, FCM, PFCM clustering algorithms (20Newsgroups collection)

In summary, the performance of the descriptors obtained after clustering by means of FCM, PCM and PFCM is presented in Table IV. The check mark (✓) shows from which clustering method the descriptors of each collection have obtained the best results.

TABLE IV. SUMMARY RESULTS

| Dataset | # terms | FCM | PCM | PFCM |
|---|---|---|---|---|
| Opinosis | 842 | | ✓ | |
| NSF | 2806 | ✓ | | |
| Reuters-21578 | 3925 | ✓ | | |
| Hitech | 6925 | ✓ | | |
| WAP | 8070 | | | ✓ |
| 20Newsgroups | 11028 | | | ✓ |

Next, we discuss the results obtained by the experiments and draw some conclusions.

## IV. DISCUSSION AND CONCLUSIONS

In our investigation we are facing the challenge of organizing an amount of data stored in a textual format. For that, we presented comparative results concerning the performance of cluster descriptors obtained after FCM, PCM and PFCM clustering algorithms for flexible document organization.

An important goal of flexible document organization is to improve a system with new information concerning to its amount of data. Converting these data into performance information that can be used by any corporation to improve the lives of clients and the effectiveness and efciency of their system is not a straightforward task.

Therefore, the key contribution of this paper is to show that the method for extracting cluster descriptors, proposed in [20], is a promising method in the sense that the descriptors can be used as good attributes for text categorization in a flexible document organization. Additionally, considering that such a method is dependent on the compatibility degrees of each document in each cluster, in this paper we also have analysed which overlapping clustering algorithm can be considered more appropriated to map the content of a document collection.

A good mechanism to extract cluster descriptors is critical to document organization, specially involving the end-user, since this step is responsible for the presentation of data to him/her. Traditional approaches identify cluster descriptors selecting the prominent and dominating terms in the documents of that cluster. These approaches provide a statistically important set of terms as descriptors. However, they often fail to provide descriptors aid in identifying overlapping information concerning to the contents of the cluster.

We can conclude from the results that the high dimensionality of the document-term matrix, obtained when we have a big amount of data, have been directed influenced in the performance of clustering algorithms and, consequently, the cluster descriptors extraction.

In Table IV we can observe this conclusion by the number of terms, in which, when the dimensionality of a document collection increases, the result obtained by a clustering algorithm has been changed. This situation is very common in data mining and specifically in text mining. As commented in Section II-B, the "curse of dimensionality" [27] needs to be considered when performing cluster analysis, since a large number of terms are hard to think in, impossible to visualize, and an enumeration of all subspaces becomes intractable with increasing dimensionality.

In this context, the conclusions obtained in [22] could be reinforced and improved by the experiments carried out in this paper.

The reason for the obtained result is that, when the number of terms of a document collection increases, the space in which the documents are represented becomes higher and sparser, becoming difficult for clustering find interesting relations between the terms and documents. Consequently, the extraction of meaningful cluster descriptors used for text categorization is affected.

Therefore, we recommend the use of PFCM for clustering high dimensional data. The PFCM avoids the noise sensitivity defect of FCM and overcomes the coincident clusters problem of PCM, two situations very common in big data problems.

In the future, we will continue investigating the performance of high dimensional data clustering, since finding meaningful and useful information in a flexible document organization depends on the selection of the appropriate clustering technique. One technique that we wish to examine is the co-clustering[16], [44], by which the duality between documents and terms can be exploited.

## ACKNOWLEDGMENT

Authors acknowledge the Brazilian National Research Agency (CNPq) for their support with the UFBA_PIBIC grant 2015-2016/8611.

## REFERENCES


[1] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*.

[2] D. Kumar, J. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. Havens, "A hybrid approach to clustering in big data," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–1, 2015.

[3] J. Yang and J. Ma, "A big-data processing framework for uncertainties in transportation data," in *International Conference on Fuzzy systems (FUZZ-IEEE)*, 2015, pp. 1–6.

[4] S. Chotipant, F. Hussain, and O. Hussain, "An automated and fuzzy approach for semantically annotating services," in *International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015, pp. 1–7.

[5] J. A. Iglesias, A. Tiemblo, A. Ledezma, and A. Sanchis, "Web news mining in an evolving framework," *Information Fusion*, vol. 28, pp. 90–98, 2016.

[6] D. H. Kraft, G. Pasi, and G. Bordogna, "Vagueness and uncertainty in information retrieval: How can fuzzy sets help?" *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*, pp. 1–10, 2006.

[7] B. Simhachalam and G. Ganesan, "Possibilistic fuzzy c-means clustering on medical diagnostic systems," in *International Conference on Contemporary Computing and Informatics (IC3I)*, 2014, pp. 1125–1129.

[8] I.-J. Chiang, C.-H. Liu, Y.-H. Tsai, and A. Kumar, "Discovering latent semantics in web documents using fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 2122–2134, 2015.

[9] M. Naik, H. Prajapati, and V. Dabhi, "A survey on semantic document clustering," in *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2015, pp. 1–10.

[10] W.-C. Tjhi and L. Chen, "Dual fuzzy-possibilistic coclustering for categorization of documents," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 3, pp. 532–543, 2009.

[11] G. Bordogna and G. Pasi, "Soft clustering for information retrieval applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 138–146, 2011.

[12] A. Garcia-Plaza, V. Fresno, and R. Martinez, "Fitting document representation to specific datasets by adjusting membership functions," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, June 2012, pp. 1–8.

[13] W. L. Chang, K. M. Tay, and C. P. Lim, "Enhancing an evolving tree-based text document visualization model with fuzzy c-means clustering," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2013, pp. 1–6.

[14] H. Jiang, F. Ye, J. Gu, Y. Liu, M. Zhu, and D. Chen, "An improved method of fuzzy clustering algorithm and its application in text clustering," *Journal of Information & Computational Science*, vol. 10, no. 2, pp. 519–526, 2013.

[15] Y. Wang, L. Chen, and J.-P. Mei, "Stochastic gradient descent based fuzzy clustering for large data," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2014, pp. 2511–2518.

[16] K. Honda, D. Tanaka, and A. Notsu, "Incremental algorithms for fuzzy co-clustering of very large cooccurrence matrix," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2014, pp. 2494–2499.

[17] J.-P. Mei, Y. Wang, L. Chen, and C. Miao, "Incremental fuzzy clustering for document categorization," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2014, pp. 1518–1525.

[18] P. Fazendeiro and J. V. Oliveira, "Observer-biased fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 1, pp. 85–95, 2015.

[19] M. Popescu, J. Keller, J. Bezdek, and A. Zare, "Random projections fuzzy c-means (rpfcm) for big data clustering," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015, pp. 1–6.

[20] T. Nogueira, S. Rezende, and H. Camargo, "Fuzzy cluster descriptor extraction for flexible organization of documents," in *Proceedings of the 11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011, pp. 528–533.

[21] T. Nogueira, H. Camargo, and S. Rezende, "Fuzzy cluster descriptors improve flexible organization of documents," in *Proceedings of the 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2012, pp. 616–621.

[22] T. M. Nogueira, H. A. Camargo, and S. O. Rezende, "Flexible document organization: Comparing fuzzy and possibilistic approaches," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015, pp. 1–8.

[23] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[24] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[25] N. Pal, K. Pal, and J. Bezdek, "A mixed c-means clustering model," in *International Conference on Fuzzy Systems (FUZZ-IEEE)*, vol. 1, 1997, pp. 11–21 vol.1.

[26] N. R. Pal, K. Pal, J. M. Keller, and B. J.C., "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 2005.

[27] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, New Jersey, USA: Princeton University Press, 1961.

[28] M. Steinbach, L. Ertz, and V. Kumar, "The challenges of clustering high dimensional data," in *New Directions in Statistical Physics*. Springer Berlin Heidelberg, 2004, pp. 273–309.

[29] R. Subhashini and V. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," in *First International Conference on Integrated Intelligent Computing (ICIIC)*, 2010, pp. 27–31.

[30] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: theory and applications*, 1st ed. Prentice-Hall, 1995.

[31] J. V. d. Oliveira and W. Pedrycz, *Advances in Fuzzy Clustering and its Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2007.

[32] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[33] Y. Yan and L. Chen, "Hyperspherical possibilistic fuzzy c-means for high-dimensional data clustering," in *Proceedings of the 7th International Conference on Information, Communications and Signal Processing*, ser. ICICS'09, 2009, pp. 637–641.

[34] N. Grover, "A study of various fuzzy clustering algorithms," *International Journal of Engineering Research*, vol. 3, pp. 177–181, 2014.

[35] M. V. B. Soares, R. C. Prati, and M. C. Monard, "PRETEXT II: Description of restructuring tool preprocessing of texts," ICMC-USP, Tech. Rep. 333, 2008, (in portuguese).

[36] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING), 2010*, 2010, pp. 340–348.

[37] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: [http://archive.ics.uci.edu/ml]

[38] P. J. Hayes and S. P. Weinstein, "Construe/TIS: A system for content-based indexing of a database of news stories." in *2nd Annual Conference on Innovative Applications of Artificial Intelligence.*, 1990, pp. 1–5.

[39] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Webace: a web agent for document categorization and exploration," in *Proceedings of the second international conference on Autonomous agents*, 1998, pp. 408–415.

[40] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 331–339.

[41] T. Havens, J. Bezdek, C. Leckie, L. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 6, pp. 1130–1146, 2012.

[42] R. Campello and E. Hruschka, "A fuzzy extension of the silhouette width criterion for cluster analysis," *Fuzzy Sets and Systems*, vol. 157, no. 21, pp. 2858 – 2875, 2006.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[44] W.-C. Tjhi and L. Chen, "Possibilistic fuzzy co-clustering of large document collections," *Pattern Recognition*, vol. 40, pp. 3452–3466, 2007.