Tratamento de imprecisão e incerteza na identificação de documentos textuais similares

Tatiane M. Nogueira[1], **Heloisa A. Camargo**[2], **Solange O. Rezende**[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brazil

{tatiane,solange}@icmc.usp.br

[2]Departamento de Ciência da Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brazil

heloisa@dc.ufscar.br

Abstract. *The management of uncertainty and imprecision in the identification of similar documents is a very important topic in Text Mining. Several times, the documents recovered by a search system do not present the expected relevance with respect to the user query. The clustering of similar documents considering documents belonging to more than one category allows the definition of more effective search mechanisms. This paper presents the analysis results of the behavior of two algorithms for documents clustering - Fuzzy C-Means and Expectation Maximization - that consider the possibility of documents to belong to more than one group/topic, with different degrees.*

Resumo. *O tratamento da imprecisão e incerteza na identificação de documentos similares é um tema de bastante importância na Mineração de Textos. Muitas vezes, os documentos recuperados por um sistema de busca não apresentam a relevância esperada com relação à consulta do usuário. O agrupamento de documentos similares que considera documentos pertencentes a mais de uma categoria pode propiciar a definição de mecanismos de busca mais efetivos. Este artigo apresenta os resultados da análise do comportamento de dois algoritmos para agrupamento de documentos - Fuzzy C-Means e Expectation Maximization - que consideram a possibilidade de documentos pertencerem a mais de um grupo/tópico, com diferentes graus.*

1. Introdução

O avanço e a popularização da tecnologia vivenciados ao longo dos anos tornaram comum o uso de sistemas de coleta e armazenamento digital de dados por parte das mais diversas organizações. Com isto, são geradas bases de dados com dimensões que crescem rapidamente, atingindo quantidades de dados que extrapolam a capacidade humana de, manualmente, analisá-las e compreendê-las por completo. A Mineração de Textos (MT), por sua vez, é o elemento intermediário entre os dados textuais e o conhecimento que pode ser extraído por meio da observação de padrões e regularidades presentes nessas bases, extraindo dos documentos informações novas e potencialmente úteis.

O agrupamento de documentos similares pode ser visto como parte da MT, ressaltando-o como operação básica para o gerenciamento de documentos, uma vez que

se um usuário está interessado em um determinado documento, também pode estar interessado em outros documentos similares. Por outro lado, é importante que a MT forneça um nível de detalhes, de informação sobre os grupos, suficientes para tratar a imprecisão e incerteza do conhecimento real, o que pode ser adquirido aproveitando as vantagens de abordagens fuzzy, as quais oferecem mecanismos mais poderosos para representação do conhecimento.

Entre as abordagens fuzzy que podem ser utilizadas na MT destaca-se a utilização de lógica fuzzy a fim de permitir a caracterização dos textos por meio do conceito de gradualidade [Zadrozny and Nowacka 2009]. Com isto, os principais níveis de aplicação da teoria de conjuntos fuzzy na MT estão na definição de extensões do modelo booleano em relação à representação de documentos, à consulta em modelos de Recuperação de Informação e à definição de mecanismos associativos, como fuzzy *thesaurus* e agrupamento fuzzy.

Este trabalho, especificamente, aborda a questão do tratamento de imprecisão e incerteza na identificação de documentos similares por meio de métodos de agrupamento de dados. Entende-se por documentos similares aqueles relacionados a um tópico ou conjunto de tópicos e assume-se que a categorização de documentos em grupos é inerentemente imprecisa, sendo adequado, portanto, considerar a possibilidade de um documento pertencer a mais de uma categoria, ou referir-se a mais de um tópico, com diferentes graus.

O estudo conduzido até o presente momento consiste da análise de dois algoritmos - Fuzzy C-Means e *Expectation Maximization* - relativos às abordagens fuzzy e probabilística para o tratamento da imprecisão no contexto de agrupamento de documentos. Visando fundamentar as avaliações dos resultados fornecidos pelos algoritmos, foram feitas análises empíricas das medidas de relevância dos documentos com relação aos diferentes grupos/tópicos ressaltando a importância de um documento pertencer à diferentes grupos com diferentes graus de relevância.

Para tanto, este artigo está organizado da seguinte maneira: Na Seção 2 são apresentados alguns trabalhos relacionados ao tratamento de imprecisão e incerteza de documentos de textos por meio de abordagens fuzzy. Na Seção 3 são apresentados brevemente os algoritmos utilizados. Na Seção 4 são apresentados os experimentos e análise dos resultados obtidos a partir da aplicação dos algoritmos sobre uma coleção de documentos textuais. Algumas conclusões e sugestão de trabalhos futuros são apresentados na Seção 5.

2. Trabalhos Relacionados

A possibilidade de um documento de texto pertencer a mais de uma categoria, ou referir-se a mais de um tópico, com diferentes graus pode ser tratada por meio de técnicas que acrescentam melhorias à abordagem booleana, já que a tradicional lógica booleana é intolerante em termos de incorporar imprecisão e incerteza. Entre essas abordagens, destacam-se aquelas que são baseadas no Processamento de Língua Natural [Smeaton 1992], na Teoria Probabilística [Crestani et al. 1998], em Redes Neurais[Crestani and Pasi 1999], na Lógica Fuzzy[Crestani and Pasi 1999] ou no Modelo Vetorial[Salton and McGill 1983].

Para este trabalho, considera-se o uso da lógica fuzzy para a identificação dos diferentes graus com que os documentos podem pertencer aos tópicos/grupos definidos

sobre uma coleção de textos. Portanto, a seguir, são apresentados alguns trabalhos que fazem uso da lógica fuzzy como principal meio para o tratamento de imprecisão e incerteza inerentes à documentos textuais.

Segundo [Rodrigues and Sacks 2005], tópicos que caracterizam um dado domínio de conhecimento são algumas vezes associados uns aos outros e podem, também, ser relacionados à tópicos de outros domínios. Logo, documentos podem conter informações relevantes para diferenciar domínios em algum grau. Utilizando métodos de agrupamento fuzzy, documentos são atribuídos a vários grupos simultaneamente e, assim, pode ser descoberto relacionamento útil entre os domínios, os quais de outro modo são desconsiderados pelos métodos de agrupamento *hard*, ou seja métodos que não permitem que um documento pertença a mais de um grupo simultaneamente.

Os algoritmos de agrupamento fuzzy podem ser utilizados a fim de encontrar grupos que melhor representem as informações contidas nos dados, uma vez que medem a pertinência dos padrões (documentos) pertencentes aos grupos. De maneira geral, os algoritmos de agrupamento fuzzy podem ser classificados em particionais, quando os dados são subdivididos em k grupos, ou em hierárquicos, pelos quais os dados são organizados em uma árvore de grupos. Os algoritmos de agrupamento fuzzy particionais mais utilizados são os algoritmos Fuzzy C-Means (FCM) [Bezdek 1981], Guztafson-Kessel (GK) [Guztafson and Kessel 1979] e Gath-Geva (GG) [Gath and Geva 1989]. Com base em modificações feitas nestes algoritmos, em especial o FCM, foram desenvolvidos algoritmos de agrupamento fuzzy hierárquicos de maneira a manipular bases textuais a fim de não só organizar hierarquicamente os documentos de texto mas, também, obter os graus com que um documento pode pertencer a diferentes grupos/tópicos.

Em [Rodrigues and Sacks 2004] foi proposta uma modificação do algoritmo FCM para agrupamento de documentos que utiliza o coeficiente de similaridade de cosseno ao invés da distância euclidiana. Esta modificação, por sua vez, foi aproveitada em [Rodrigues and Sacks 2005] para o desenvolvimento de um algoritmo de agrupamento fuzzy hierárquico chamado *Hierarchical Hyper-spherical c-Means Algorithm* (H^2-FCM) para utilização na mineração de textos por meio da construção de uma taxonomia de tópicos que explora a noção de similaridade assimétrica para organizar grupos fuzzy hierarquicamente formando uma hierarquia de tópicos significante baseada no centróide dos grupos.

Foi proposta em [Sraçoğlu et al. 2007] e posteriormente melhorada em [Sraçoğlu et al. 2008] uma abordagem com o uso da lógica fuzzy também para busca de similaridade entre documentos na tentativa de solucionar o problema de multi categorias. Segundo [Sraçoğlu et al. 2007], o maior problema dos atuais sistemas de busca é o resultado da busca, os quais disponibilizam documentos não relacionados ou diminuem ao máximo o número de documentos não relacionados como resultado da busca. Geralmente, nestes sistemas, os documentos de texto pertencem apenas a uma categoria.

[Torra 2005] apresenta ainda uma nova proposta de algoritmo de agrupamento fuzzy hierárquico, no qual alguns grupos são previamente definidos por meio do algoritmo de agrupamento FCM. Com isto, um processo iterativo é aplicado para a construção da hierarquia seguindo a estratégia *top-down*, na qual os grupos definidos anteriormente são

particionados utilizando um agrupamento hierárquico divisivo. O conceito de hierarquia fuzzy é também utilizado por [Lee 2001] para propor uma nova regra de associação fuzzy pela qual categorias de atributos são generalizadas. Além destes, outras abordagens para agrupamento fuzzy de documentos podem ser conferidas em [Krishnapuram et al. 2003, Horng et al. 2005, Bordogna et al. 2006].

O tratamento da imprecisão e incerteza de coleções de texto é também um ponto chave para o desenvolvimento de modelos de sistemas de Recuperação da Informação (RI) [Crestani and Pasi 1999]. Pesquisas de modelos de RI fuzzy e de generalizações fuzzy do modelo de RI booleano podem ser encontradas em [Bordogna and Pasi 2001]. No nível de indexação de documentos, algumas técnicas fuzzy tem sido definidas para prover representações mais específicas e personalizadas dos documentos do que aqueles gerados por procedimentos de indexação já existentes. A principal idéia é modelar explicitamente uma estratégia de indexação que adapta a representação formal de documentos para a personalidade do usuário de acordo com o conteúdo dos documentos [Bordogna and Pasi 2005].

Considerando-se a grande quantidade de abordagens fuzzy já exploradas para tratamento de imprecisão e incerteza na mineração de textos, neste trabalho é feita uma análise empírica do comportamento de um algoritmo de agrupamento fuzzy e um probabilístico, descritos a seguir, para a identificação de documentos similares.

3. Algoritmos de Agrupamento Utilizados

De maneira geral, agrupamentos de documentos são utilizados de modo a permitir identificar similaridade entre documentos, precisão na recuperação de informação, organização dos resultados recuperados por uma máquina de busca, exploração de uma coleção de documentos, hierarquia de documentos, construção de um taxonomia de tópicos e produção de um classificador de documentos. Neste trabalho foram realizados alguns experimentos a fim de obter conhecimento acerca de uma coleção de textos pela identificação de similaridades entres eles. Os experimentos consistem da aplicação do algoritmo de agrupamento Fuzzy C-Means e o probabilístico *Expectation Maximization*, brevemente apresentados nas subseções a seguir, para observar e comparar o comportamento dos mesmos na distribuição de graus/probabilidades dos documentos nos grupos.

Segundo [Dubois and Prade 1994], probabilidade e lógica fuzzy são usualmente contrastadas como dois veículos conceituais e computacionais distintos destinados à representação e processamento da incerteza. A incerteza não está diretamente associada a qualquer sistema do mundo real, mas está relacionada principalmente com o processo de descrever o sistema selecionado em si. Portanto, a maneira pela qual a incerteza se manifesta, bem como o método pelo qual ela pode ser captada corretamente dependem do observador. Logo, não há um método universal para o tratamento da incerteza e, ao contrário de muitas discussões, lógica fuzzy e probabilidades podem ser métodos complementares ao invés de opostos.

3.1. Fuzzy C-Means

O Fuzzy C-Means (FCM) é um algoritmo extremamente poderoso para o agrupamento de dados, cuja idéia é de ponderar a distância de um ponto ao centro do grupo por meio de um valor de pertinência. Com isto, um dado pertencerá a um certo grupo com um grau,

pois neste tipo de agrupamento os limites entre os grupos são imprecisos. Além disso, todo dado deverá pertencer ao menos a um grupo e nenhum grupo poderá conter todos os dados. O processo iterativo consiste na atualização dos centros de grupos, centróides, baseado na otimização de um critério de erro.

Considere um conjunto de dados $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, no qual $\mathbf{x}_k$ é um vetor de valores p-dimensional de atributos $\mathbf{x}_k = \left[x_{k_1}, x_{k_2}, ..., x_{k_p}\right] \in \mathbb{R}^p$, para $1 \leq k \leq n$.

Uma Pseudo-partição fuzzy, também conhecida como c-partição fuzzy, é uma família de conjuntos fuzzy de X denotados por $\mathrm{P} = \{A_1, A_2, ..., A_c\}$, que satisfaz as Equações (1) e (2), para todo $k = 1...n$ [Klir and Yuan 1995].

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

$$0 < \sum_{k=1}^{n} A_i(\mathbf{x}_k) < n, \; para\, i = 1, ..., c \tag{2}$$

O FCM é um algoritmo iterativo que atualiza os centros dos grupos definidos previamente com a definição de uma partição. Após cada atualização dos centróides, a partição é redefinida. Logo, a performance do FCM depende da forma de atualização destes centros e da redefinição das partições.

Os c vetores de centros dos grupos, $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c$, são calculados pela Equação (3), para todo $i \in \mathrm{N}_c$, sendo que $m > 1$ é um número real, chamado fator de fuzificação, que controla a influência dos graus de pertinência [Klir and Yuan 1995] e $A_i(\mathbf{x}_k)$ é o grau de pertinência do vetor $\mathbf{x}_k$ ao grupo i.

$$\mathbf{v}_i = \frac{\sum\limits_{k=1}^{n} \left[A_i(\mathbf{x}_k)\right]^m \mathbf{x}_k}{\sum\limits_{k=1}^{n} \left[A_i(\mathbf{x}_k)\right]^m} \tag{3}$$

Cada elemento da partição é redefinido pela Equação (4).

$$A_i(\mathbf{x}_k) = \frac{1}{\sum\limits_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|}^2 \right)^{\frac{1}{m-1}}} \tag{4}$$

A performance do FCM é baseada na otimização da função objetivo representada pela Equação (5) sobre a pseudo-partição P, na qual $\|\mathbf{x}_k - \mathbf{v}_i\|$ representa a distância entre $\mathbf{x}_k$ e $\mathbf{v}_i$, o qual deve ser atualizado conforme Equação (3) [Klir and Yuan 1995].

$$J_m(P) = \sum_{k=1}^{n} \sum_{i=1}^{c} \left[A_i(x_k)\right]^m \|x_k - v_i\|^2 \tag{5}$$

O objetivo do FCM é minimizar a função objetivo J_m descrita anteriormente, ou seja, minimizar a distância entre os padrões e os centros dos grupos. Ao iniciar sua

execução, o algoritmo FCM assume que uma determinada quantidade de grupos e um número pequeno ϵ como critério de parada são definidos previamente.

3.2. Expectation Maximization

O algoritmo de agrupamento probabilístico *Expectation Maximization* (EM) [Dempster et al. 1977] é um modelo estatístico que faz uso do modelo de misturas de gaussianas. Inúmeras melhorias vem sendo feitas nesse algoritmo e, para os experimentos aqui apresentados, utilizou-se a implementação do EM disponível na ferramenta WEKA [Garner 1995].

O algoritmo EM é similar ao procedimento do K-Means, no qual um conjunto de parâmetros é recomputado até que o valor de convergência desejado seja atingido. O modelo de misturas de gaussianas assume que todos os atributos são variáveis aleatórias independentes. Uma mistura é um conjunto de N distribuições probabilísticas na qual cada distribuição representa um grupo. Uma instância individual é assinalada com uma probabilidade dada por um certo conjunto de valores de atributos em um determinado grupo.

No caso mais simples $N = 2$, as distribuições de probabilidades são assumidas normais e os dados consistem de um único valor real de atributo. Neste caso, o algoritmo deve determinar o valor de cinco parâmetros: média e desvio padrão para o grupo 1, média e desvio padrão para o grupo 2, amostragem probabilística P para o grupo 1 (a probabilidade para o grupo 2 é $1 - P$). Com isto, o procedimento mais geral do algoritmo EM consiste de:

1. Estimar os valores iniciais para os cinco parâmetros.
2. Utilizar a função de densidade probabilística de uma distribuição normal para computar a probabilidade de cada instância nos grupos. No caso de uma única variável independente com média μ e desvio padrão σ, a fórmula de distribuição probabilística é dada pela Equação (6). Assumindo dois grupos, haverá duas fórmulas de distribuição probabilística cada qual com diferentes valores de média e desvio padrão.
3. Utilizar a pontuação probabilística para reestimar os cinco parâmetros.
4. Retornar ao passo 2.

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)e^{\frac{-(\sigma-\mu)^2}{2\sigma^2}}} \tag{6}$$

O algoritmo termina quando a fórmula que mede a qualidade dos grupos não apresenta mudanças significativas. Uma medida da qualidade do grupo pode ser a verossimilhança (*likelihood*) que os dados obtiveram a partir do conjunto de dados determinada pelo agrupamento. Assim, a verossimelhança é a multiplicação da soma das probabilidades de cada instância em cada grupo.

4. Experimentos e Análise dos Resultados

A fim de comparar os graus/probabilidades obtidos por meio dos algoritmos apresentados anteriormente, foi construída uma coleção de textos cujo domínio é controlado composta de 50 artigos científicos em português relacionados à subáreas da grande área de

Inteligência Artificial. Manualmente foram coletadas informações acerca do domínio o que possibilitou a construção de uma taxonomia *gold*, ou seja, aquela a ser comparada com os resultados obtidos automaticamente após a aplicação dos algoritmos de agrupamento fuzzy e probabilístico. A análise foi feita independente da hierarquia, ou seja, após a execução dos algoritmos observou-se os graus de pertinência/probabilidades que os algoritmos de agrupamento atribuíram a cada um dos textos nos grupos. Uma vez que o domínio é controlado, sabe-se de antemão se um texto tende a pertencer a mais de um grupo/tópico, ou faz parte de apenas um único tópico, bem como os documentos que tendem a pertencer à um mesmo grupo.

Com este domínio foi possível montar a taxonomia apresentada na Figura 1, na qual são apresentados os tópicos que mais caracterizam a coleção, bem como os documentos (identificados por um número) que pertencem à cada tópico. Por exemplo, o número *19* representa o documento de texto com Identificador 19 no tópico Mineração. Sobre este domínio controlado, os algoritmos FCM e EM foram executados para o número fixo de grupos igual a 6 (quantidade de nós do primeiro nível da taxonomia) sobre os 50 documentos da coleção após a extração da tabela atributo-valor na etapa de pré-processamento da mineração de textos.

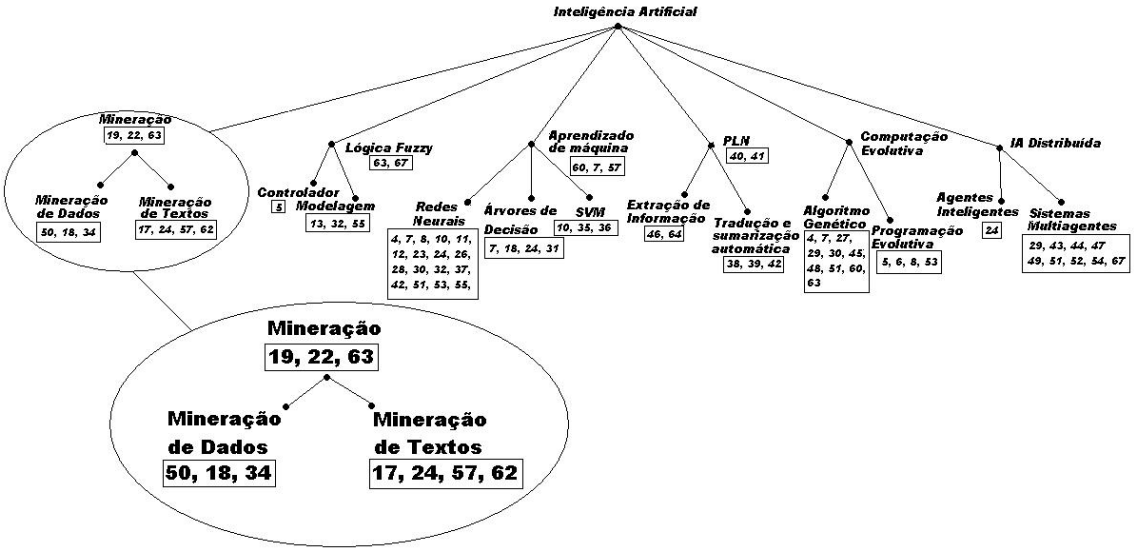

Figura 1. Taxonomia *gold*

Após a execução dos algoritmos, a análise dos resultados consiste em comparar a distribuição de graus de pertinência/probabilidades atribuídas às instâncias/textos em cada grupo e se estes graus refletem a distribuição manual feita na taxonomia. Logo, documentos que tendem a pertencer a mais de um grupo devem apresentar graus de pertinência/probabilidades distribuídas entre os grupos. Assim, foi possível observar três tipos de conjuntos de textos: (1) textos em que o FCM apresentou comportamento da distribuição de graus de pertinência de maneira equilibrada entre os grupos, (2) textos em que o EM apresentou comportamento da distribuição de probabilidades de maneira equilibrada entre os grupos e (3) o comportamento de ambos os algoritmos sobre os textos conhecidamente difíceis de definir o tópico.

Os resultados são apresentados em formato de tabela, cuja primeira coluna possui

o código identificador dos textos da coleção, e as demais representam os graus de pertinência dos textos nos 6 grupos obtidos pelo algoritmo FCM ou EM como destacado na primeira linha das tabelas.

Na Tabela 1 são apresentados os textos que o EM apresentou distribuição de probabilidades dos textos equilibrada entre os grupos comparados à distribuição de graus de pertinência feita pelo FCM. Observe que para todos os textos desta tabela o EM distribuiu as maiores probabilidades em dois grupos apenas, e nos demais grupos apresentou probabilidade bastante baixa. O mesmo ocorreu com o FCM, que apresentou graus de pertinência relativamente altos em dois grupos e graus baixos nos demais grupos. Destes, apenas no texto 19 o FCM apresentou comportamento diferente do EM, já que o FCM apresentou alto grau em apenas um grupo e baixo nos demais.

Tabela 1. EM x FCM

Textos	Expectation Maximization						Fuzzy C-Means					
19	3.19E-07	0.957	1.69E-05	3.57E-07	0.042	3.25E-06	5.95E-01	2.77E-03	2.03E-05	1.07E-06	6.69E-04	0.980
28	3.27E-07	6.22E-07	0.979	2.16E-07	0.020	4.26E-07	0.009	0.840	0.087	0.011	7.62E-04	0.046
29	4.52E-06	7.56E-06	0.640	2.91E-06	0.359	8.38E-06	7.55E-08	0.001	3.20E-11	0.165	0.054	0.005
35	1.05E-07	6.71E-07	2.55E-06	2.97E-07	0.012	0.987	0.335	0.485	0.037	0.036	0.022	0.082
36	1.72E-07	9.36E-07	4.36E-06	1.84E-07	0.009	0.990	0.264	0.657	0.020	0.011	0.028	0.017
38	0.994	7.79E-08	1.96E-06	2.75E-08	0.005	1.34E-07	0.002	0.503	0.006	7.08E-07	7.37E-06	0.321
45	1.11E-07	3.51E-07	5.66E-06	0.988	0.011	9.39E-07	0.281	0.496	0.190	0.007	0.006	0.012
51	1.08E-06	2.96E-06	0.909	9.20E-07	0.090	2.30E-06	0.001	0.991	1.70E-03	8.56E-05	7.23E-07	0.005
64	0.994	7.79E-08	1.96E-06	2.75E-08	0.005	1.34E-07	0.002	0.503	0.006	7.08E-07	7.37E-06	0.321

Já na Tabela 2 são apresentados os textos que o FCM apresentou maior distribuição de graus de pertinência dos textos equilibrada entre os grupos comparados à distribuição de probabilidades feita pelo EM. Observe que em todos os textos que o FCM distribuiu os graus de pertinência entre os grupos, o EM apresentou comportamento contínuo atribuindo probabilidade alta à todos estes textos em um único grupo.

Tabela 2. FCM x EM

Textos	Fuzzy C-Means						Expectation Maximization					
11	2.09E-05	5.39E-05	0.006	0.442	0.139	0.392	8.92E-06	4.67E-05	3.35E-04	6.85E-06	0.999	3.05E-05
13	0.188	0.156	0.098	0.214	0.264	0.075	1.05E-05	3.51E-05	2.41E-04	4.63E-06	0.999	1.14E-05
22	0.059	0.222	0.262	0.144	0.163	0.145	1.09E-05	2.71E-05	2.01E-04	4.15E-06	0.999	2.10E-05
34	0.030	0.205	0.086	0.163	0.192	0.318	1.33E-05	1.87E-05	1.99E-04	7.19E-06	0.999	2.08E-05
4	0.239	0.108	0.055	0.093	0.305	0.181	7.87E-06	3.14E-05	4.22E-04	1.40E-05	0.999	2.34E-05
40	0.006	0.122	0.036	0.163	0.175	0.462	2.09E-05	3.54E-05	2.65E-04	7.66E-06	0.999	3.04E-05
5	0.173	0.037	0.143	0.271	0.179	0.193	8.25E-06	7.61E-05	2.65E-04	6.31E-06	0.999	1.92E-05
60	0.095	0.119	0.207	0.180	0.117	0.277	1.17E-05	3.58E-05	2.04E-04	6.71E-06	0.999	2.81E-05

Uma vez que as avaliações experimentais são realizadas com um conjunto controlado de documentos, sabe-se de antemão quais documentos de texto são compatíveis com mais de um tópico da taxonomia. Logo, na Tabela 3 é apresentado o comportamento dos algoritmos FCM e EM para os textos conhecidamente ambíguos (caracterizam-se por mais de um tópico). Observe que na maioria dos textos, o FCM, embora apresente sempre um grau maior para um único grupo, este mesmo texto pertencerá há outros grupos com graus menores devido à característica inerente ao FCM em que todo dado pertencerá a um certo grupo com um determinado grau. Já o EM apresentou este comportamento apenas nos textos 51, 29, 28, porém com uma diferença de probabilidade muito baixa entre os grupos, o que não reflete que os textos podem ser compatíveis com mais de um tópico.

Diante do comportamento dos algoritmos de agrupamento EM e FCM apresentados nas tabelas anteriores, algumas conclusões são feitas na seção a seguir acerca das particularidades de cada algoritmo para o tratamento da imprecisão e incerteza de documentos textuais, bem como possíveis trabalhos a serem investigados futuramente.

Tabela 3. Textos compatíveis com mais de um tópico

Textos	Fuzzy C-Means						Expectation Maximization					
24	2.96E-01	1.55E-01	2.69E-01	0.076	7.65E-02	0.126	8.09E-06	1.34E-05	1.66E-04	2.55E-06	0.999	5.49E-06
27	6.46E-04	1.49E-06	0.641	2.27E-04	0.054	3.79E-04	9.17E-06	1.22E-05	9.99E-04	4.90E-06	0.998	1.20E-05
28	0.009	0.840	0.087	0.011	7.62E-04	0.046	3.27E-07	6.22E-07	0.979	2.16E-07	0.020	4.26E-07
29	7.55E-08	0.001	3.20E-11	0.165	0.054	0.005	4.52E-06	7.56E-06	0.640	2.91E-06	0.359	8.38E-06
40	0.006	0.122	0.036	0.163	0.175	0.462	2.09E-05	3.54E-05	2.65E-04	7.66E-06	0.999	3.04E-05
5	0.173	0.037	0.143	0.271	0.179	0.193	8.25E-06	7.61E-05	2.65E-04	6.31E-06	0.999	1.92E-05
51	0.001	0.991	1.70E-03	8.56E-05	7.23E-07	0.005	1.08E-06	2.96E-06	0.909	9.20E-07	0.090	2.30E-06
55	7.90E-06	5.02E-04	1.64E-02	2.73E-07	2.16E-04	0.586	1.26E-05	2.56E-05	8.52E-04	7.38E-06	0.999	2.12E-05
63	6.20E-05	0.005	0.014	0.024	0.086	0.836	9.97E-06	2.94E-05	2.58E-04	1.15E-05	0.999	3.52E-05
7	0.081	0.015	0.115	0.264	0.400	0.121	1.12E-05	3.45E-05	2.12E-04	4.85E-06	0.999	3.95E-05

5. Conclusão e Trabalhos Futuros

Neste trabalho foi apresentada uma análise comparativa entre os algoritmos de agrupamento Fuzzy C-Means e *Expectation Maximization*, ambos para o tratamento de imprecisão e incerteza no agrupamento de documentos textuais. Por meio da análise foi possível observar que as duas abordagens estipulam, de alguma maneira, a relevância de documentos dentro da coleção em determinados tópicos/grupos com base nos termos que ocorrem nos mesmos. Porém, as abordagens seguem critérios diferentes para esta estimativa. Enquanto o FCM busca distribuir, de maneira equilibrada, os documentos em todos os grupos, o EM tenta encontrar a maior probabilidade de um documento dentro de um único grupo. Diante disto, é importante ressaltar que é necessária uma avaliação objetiva em termos de cálculo do *Precision* e *Recall* sobre os documentos organizados por ambos os algoritmos, a fim de observar o comportamento da recuperação de informação mediante os graus de pertinência/probabilidades obtidos por ambos os algoritmos. Além disto, vale investigar abordagens que se beneficiam de ambos os algoritmos em conjunto já que cada algoritmo apresenta particularidades na definição de importância dos documentos nos tópicos.

Referências

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Bordogna, G., Pagani, M., and Pasi, G. (2006). *Soft Computing for Information Retrieval on the Web*. Springer Verlag.

Bordogna, G. and Pasi, G. (2001). *Lectures in Information Retrieval*. Springer Verlag.

Bordogna, G. and Pasi, G. (2005). Personalized indexing and retrieval of heterogeneous structured documents. *Information Retrieval*, 8(2):301–318.

Crestani, F., Lalmas, M., van Rijsbergen, C., and Campbell, I. (1998). Is this document relevant? probably. *ACM Computing Surveys*, 30(4):528–552.

Crestani, F. and Pasi, G. (1999). Soft information retrieval: Applications of fuzzy set theory and neural networks. In N.Kasabov and Kozma, R., editors, *Neuro-fuzzy Techniques for Intelligent Information Systems*, pages 287–313. Physica-Verlag, Springer-Verlag Group.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Joournal of the Royal Statistical Society, Series B*, 39(1):1–38.

Dubois, D. and Prade, H. (1994). Fuzzy sets - a convenient fiction for modeling vagueness and possibility. *IEEE Transactions on Fuzzy Systems*, 2:6–21.

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In *In Proc. of the New Zealand Computer Science Research Students Conference*, pages 57–64.

Gath, I. and Geva, B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:773–781.

Guztafson, E. E. and Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix. *IEEE CDC*, pages 503–516.

Horng, Y.-J., Chen, S.-M., , Chang, Y.-C., and Lee, C.-H. (2005). A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. *IEEE Transactions on Fuzzy Systems*, 13(2):216–228.

Klir, G. J. and Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: theory and applications*. Prentice-Hall, 1 edition.

Krishnapuram, R., Dhawale, A., and Kummamuru, K. (2003). Fuzzy co-clustering of documents and keywords. In *International Conference on Fuzzy Systems*, pages 772–777.

Lee, K.-M. (2001). Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies. *IEEE*, pages 2977–2982.

Rodrigues, E. M. and Sacks, L. (2005). Learning topic hierarchies from text documents using a scalable hierarchical fuzzy clustering method. In *International Conference on Recent Advances in Soft Computing*, pages 269–274.

Rodrigues, M. E. S. M. and Sacks, L. (2004). A scalable hierarchical fuzzy clustering algorithm for text mining. In *UK Workshop on Computational Intelligence*.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Smeaton, A. F. (1992). Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, 35(3):268–278.

Sraçoğlu, R., TüTüncü, K., and Allahverdi, N. (2007). A fuzzy clustering approach for finding similar documents using a novel similarity measure. *Expert Systems with Applications*, 33:600–605.

Sraçoğlu, R., TüTüncü, K., and Allahverdi, N. (2008). A new approach on search for similar documents with multiple categories using fuzzy clustering. *Expert Systems with Applications*, 34:2545–2554.

Torra, V. (2005). Fuzzy c-means for fuzzy hierarchical clustering. In *IEEE International Conference on Fuzzy Systems*, pages 646–651.

Zadrozny, S. and Nowacka, K. (2009). Fuzzy information retrieval model revisited. *Fuzzy Sets and Systems*, 160:2173–2191.