FERNANDO SIMEONE

AGRUPAMENTO HIERÁRQUICO INCREMENTAL DE TEXTOS COM INCORPORAÇÃO DE ENTIDADES NOMEADAS

LAVRAS – MG

2016

FERNANDO SIMEONE

AGRUPAMENTO HIERÁRQUICO INCREMENTAL DE TEXTOS COM INCORPORAÇÃO DE ENTIDADES NOMEADAS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Inteligência Computacional, para a obtenção do título de Mestre.

Prof. Dr. Ahmed Ali Abdalla Esmin

Orientador

LAVRAS – MG

2016

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Simeone, Fernando.
Agrupamento hierárquico incremental de textos com incorporação de entidades nomeadas / Fernando Simeone. – Lavras : UFLA, 2016.
127 p. : il.

Dissertação(mestrado acadêmico)–Universidade Federal de Lavras, 2016.
Orientador: Ahmed Ali Abdalla Esmin.
Bibliografia.

1. Mineração de dados. 2. Agrupamento hierárquico. 3. Agrupamento incremental. I. Universidade Federal de Lavras. II. Título.

FERNANDO SIMEONE

AGRUPAMENTO HIERÁRQUICO INCREMENTAL DE TEXTOS COM INCORPORAÇÃO DE ENTIDADES NOMEADAS

INCREMENTAL HIERARCHICAL GROUPING OF TEXTS WITH THE INCORPORATION OF NAMED ENTITIES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Inteligência Computacional, para a obtenção do título de Mestre.

APROVADA em 18 de abril de 2016.

Prof. Dr. Leonardo Andrade Ribeiro UFG

Prof. Dr. Wilian Soares Lacerda UFLA

Prof. Dr. Ahmed Ali Abdalla Esmin
Orientador

LAVRAS – MG

2016

AGRADECIMENTOS

Agradeço a Deus pela oportunidade de adquirir novos conhecimentos e enfrentar novos desafios, aos professores pelo o conhecimento transmitido, aos meus amigos pelo apoio e principalmente a minha família pelo apoio incondicional e pela compreensão em todos os momentos.

RESUMO

O crescente uso da internet tem gerado uma grande quantidade de informações, que são disponibilizadas de forma cada vez mais frequente para os usuários. Este cenário pode ser facilmente verificado em sites de notícias e mídias sociais. Com este grande volume de dados, surge a necessidade de organizar os dados de forma automatizada, com o intuito de proporcionar aos usuários a fácil navegação e localização destas informações. As técnicas de agrupamento hierárquico de documentos textuais visam a organização de textos em vários níveis de granularidade, resultando em uma estrutura que atende a tais necessidades. As técnicas de agrupamento incremental visam trabalhar com a dinamicidade dos dados, adaptando as estruturas de agrupamento conforme dados são adicionados ou removidos. Vários dos documentos disponíveis para análise podem conter nomes de entidades, como pessoas, empresas e locais. Tais informações podem ser incorporadas no processo de agrupamento com o intuito de auxiliar estas técnicas a serem mais eficazes na organização destes documentos. Este trabalho tem como objetivo o estudo de técnicas de agrupamento hierárquico incremental e a proposta e investigação de usos de entidades nomeadas no processo de agrupamento.

Palavras-chave: Mineração de dados. Agrupamento hierárquico. Agrupamento incremental. Mineração de textos. Entidades nomeadas.

ABSTRACT

The increasing use of the internet has generated a large amount of information made available to the users ever more frequently. This scenario can be easily verified in websites of news and social media. With this large amount of data, the need to organize data in an automatic manner emerges, with the intent of providing easy navigation and localization of this information to the users. The hierarchical grouping techniques of text documents aim at organizing texts in many levels of granularity, resulting in a structure that meets such needs. The incremental grouping techniques aim at working with the dynamicity of the data, adapting the grouping structures according to data added or removed. Many of the documents available for analysis can contain names of entities such as people, companies and locations. Such information can be incorporated to the grouping process with the objective of aiding these techniques in their efficiency and document organization. This work had the objective of studying incremental hierarchical grouping techniques, as well as proposing and investigating the use of entities named in the grouping process.

Keywords: Data mining. Hierarchical grouping. Incremental grouping. Text mining. Named entities.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O uso da Internet teve grande crescimento na última década, acarretando também o crescimento da quantidade de dados disponibilizados e a frequência de atualização destes dados. Este cenário pode ser facilmente verificado em sites de notícias e mídias sociais. Com o crescente volume de informações disponíveis, surge a necessidade de organizá-las de forma que possam ser facilmente encontradas, exploradas e analisadas. Neste sentido, uma das principais formas que o ser humano utiliza para lidar com grandes quantidades de dados é classificá-los ou agrupá-los em categorias de acordo com suas características (JAIN; MURTY; FLYNN, 1999). Observa-se que a organização manual dos dados pode ser uma tarefa de alto custo, visto que, além de exigir grande esforço, pode ser necessário o auxílio de um especialista no contexto das informações.

Para resolver tais questões, esforços têm sido dedicados no estudo de técnicas automatizadas para classificar ou agrupar dados que possuem características similares. Tais técnicas são conhecidas como classificação supervisionada e agrupamento. Como explicam Jain, Murty e Flynn (1999), na classificação supervisionada há um conjunto de dados já classificado, que é utilizado para treinamento de um algoritmo. Efetuado o treinamento, o algoritmo passa a conhecer os padrões das classes dos dados e é capaz de classificar novos dados. Nas técnicas de agrupamento, também conhecidas como classificação não supervisionada, não há um conjunto de dados de treinamento, sendo a categorização dos dados obtida utilizando apenas os próprios dados. Neste caso, os dados são agrupados de acordo com o grau similaridade entre si (CHERKASSKY; MULIER, 2007). Esta segunda técnica se mostra adequada para cenários onde não se possui conhecimento prévio dos dados ou há obstáculos para a construção de um conjunto de dados significativo para treinamento.

Visto o grande volume de dados presente na Web, um simples agrupamento pode não ser suficiente para tornar viável a exploração dos dados. Para suprir esta demanda, existem os algoritmos de agrupamento hierárquico, que buscam organizar os dados em diversos níveis de abstração. Neste tipo de algoritmo, os agrupamentos resultantes são organizados como uma árvore, facilitando a exploração e visualização dos dados (ZHAO; KARYPIS; FAYYAD, 2005).

Outro obstáculo presente em dados disponibilizados na internet é a sua dinamicidade. Em contextos como um site de notícias, constantemente há inserção ou remoção de informações. Neste cenário, algoritmos tradicionais de agrupamento não são eficazes, pois necessitam de toda a base de dados disponível em sua execução. Porém, existem os algoritmos de agrupamento incremental, que conseguem lidar com cenários dinâmicos. Estes estão aptos a processar a adição ou a remoção de dados, adaptando a estrutura de agrupamentos existente conforme necessário (GIL-GARCÍA; PONS-PORRATA, 2010a).

O contexto investigado neste trabalho é a organização hierárquica de documentos textuais em ambientes dinâmicos. Este é justamente o cenário no qual se enquadram os diversos sites de notícias da internet. Assim, este trabalho visa estudar as técnicas de agrupamento hierárquico incremental, compreender e comparar os trabalhos recentes propostos neste campo e investigar o uso de entidades nomeadas no processo de agrupamento, a fim de propor novas abordagens que possam ser aplicadas a este contexto.

1.1 Justificativa

Com relação ao cenário ao qual as técnicas estudadas podem ser aplicadas, a primeira motivação para o desenvolvimento deste projeto é o fato de este ser um cenário existente na atualidade. Dados textuais estão presentes em

sites de notícias, redes sociais e outros contextos da internet, sendo a sua atualização cada vez mais frequente. A organização automatizada destes dados possui aplicação prática, como facilitar a visualização e navegação dos dados, a construção de sistemas de recomendação e sistemas de busca.

A respeito das técnicas de agrupamento hierárquico incremental, ainda há diversos desafios que não foram totalmente superados, como a qualidade dos agrupamentos gerados, o grande volume e dimensionalidade dos dados, a obtenção de hierarquias de fácil navegação, a definição sobre o número de grupos a serem gerados, a geração de descritores de qualidade para os grupos, a inviabilidade de releitura dos dados já processados e a sensibilidade do algoritmo à ordem de entrada dos dados.

No que diz respeito ao volume de trabalhos, considerando os problemas de agrupamento hierárquico e agrupamento incremental separadamente, sabe-se da existência de vasta quantidade de trabalhos que buscam resolvê-los. Porém, considerando o cenário que envolve os dois problemas, observa-se um conjunto reduzido de trabalhos, sendo a maioria destes desenvolvidos nos últimos anos. Também não há uma compilação ou comparativo entre as técnicas apresentadas em trabalhos recentes. Tal fato apresenta uma oportunidade para estudo e comparação dos trabalhos relacionados ao tema e também para a investigação de novas abordagens.

A principal motivação deste trabalho é a hipótese de melhoria das técnicas de agrupamento hierárquico incremental por meio do uso de entidades nomeadas extraídas dos documentos. Esta foi uma técnica utilizada em um número reduzido de trabalhos, não havendo uma conclusão sobre sua eficácia neste contexto.

1.2 Objetivos

Esta seção apresenta os objetivos deste trabalho, onde na Seção 1.2.1 é apresentado o objetivo geral e na Seção 1.2.2 são apresentados os objetivos específicos.

1.2.1 Objetivo geral

O objetivo geral deste trabalho é estudar e comparar as soluções já existentes para agrupamentos hierárquicos de documentos textuais em ambientes dinâmicos e propor novas abordagens para resolver tal problema. Tais propostas serão testadas e avaliadas utilizando as bases de dados mais utilizadas nos trabalhos recentes, a fim de validar sua eficiência e eficácia.

1.2.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

a) Estudar e compreender os principais conceitos relacionados ao agrupamento hierárquico de documentos em ambientes dinâmicos;
b) Efetuar um levantamento dos trabalhos recentes sobre o tema, estudá-los e compará-los com o intuito de compreender as principais soluções existentes para o problema. Escrever artigo contendo os comparativos resultantes e uma análise sobre os trabalhos levantados;
c) Selecionar e implementar ao menos um dos algoritmos recentes levantados e realizar experimentos aplicando-os às principais bases de dados;
d) Construir um ambiente para a execução dos experimentos, que deve facilitar o teste e a comparação entre os algoritmos implementados, provendo métricas para comparação dos algoritmos assim como mecanismos para a compilação e visualização de resultados. Este

ambiente deve prover também facilidades para a inclusão de novos algoritmos, disponibilizando estruturas e sub-rotinas de uso comum;

e) Propor e implementar abordagens de uso de entidades nomeadas no processo de agrupamento. Efetuar experimentos e comparativos a fim de validar a eficácia do uso de tais informações para o agrupamento hierárquico incremental de documentos;
f) Publicar os resultados obtidos, assim como disponibilizar informações necessárias para a reprodução dos experimentos.

1.3 Organização do trabalho

Este trabalho está estruturado no seguinte formato: no Capítulo 2 é apresentada uma revisão de literatura contendo o embasamento teórico necessário para o entendimento do restante do trabalho; em seguida, o Capítulo 3 apresenta os conceitos relacionados à entidades nomeadas e as abordagens adotadas para seu uso em algoritmos de agrupamento; no Capítulo 4, é apresentado um levantamento dos trabalhos recentes relacionados ao tema estudado neste projeto; o Capítulo 5 apresenta a metodologia do trabalho, contendo as etapas do projeto e as ferramentas a serem utilizadas; no capítulo 6 são exibidos os experimentos realizados e os resultados obtidos; por fim, as conclusões e trabalhos futuros são apresentados no Capítulo 7.

2 AGRUPAMENTO HIERÁRQUICO DE DOCUMENTOS TEXTUAIS

2.1 Técnicas de agrupamento

O problema de agrupamento consiste basicamente na separação de dados em grupos, também conhecidos como *clusters*, baseada em uma medida de similaridade (CHERKASSKY; MULIER, 2007). Pode-se também definir tal tarefa como a classificação não supervisionada de padrões, como observações ou amostras de dados, em grupos. Estes dados são normalmente representados como vetores ou pontos em um espaço multidimensional, sendo agrupados de acordo com suas similaridades (JAIN; MURTY; FLYNN, 1999). Desta forma, os agrupamentos resultantes (*clusters*) são compostos por itens que são similares entre si e dissimilares dos itens inseridos em outros agrupamentos (BERKHIN, 2006). Um exemplo de agrupamento exposto na Figura 1, onde na Figura 1(a) é exibido o conjunto de dados analisado em um plano, e na Figura 1(b) são exibidos os mesmos dados, porém, agrupados conforme sua proximidade.

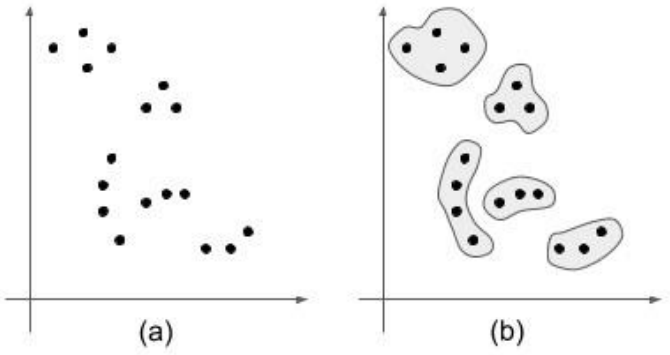


Figura 1 - Exemplo de agrupamento

Como explicado por Feldman e Sanger (2006), há diferenças entre as tarefas de classificação supervisionada e de agrupamento. Na tarefa de classificação supervisionada já é conhecido um conjunto de dados classificados

e rotulados, e tal conjunto é utilizado no processo de treinamento do algoritmo, para que ele aprenda os padrões contidos nestes dados e, assim, consiga classificar novos dados. Já a tarefa de agrupamento consiste em, a partir de um conjunto de dados não rotulados e sem nenhuma informação prévia, gerar grupos que tenham algum significado.

Segundo Jain (2010), as técnicas de agrupamento têm sido utilizadas com três principais objetivos: (1) identificação de informações implícitas, como anomalias, características salientes nos dados; (2) classificação natural, para identificação do grau de similaridades entre os dados e (3) facilitação da compreensão, como forma de organização dos dados e sumarização dos mesmos através de seus protótipos.

Protótipos são representações dos grupos, que descrevem seus elementos de forma compacta, podendo tanto facilitar o entendimento humano sobre o agrupamento quanto otimizar o processamento posterior de tais dados. Tal informação é de grande importância, visto que as técnicas de agrupamento são aplicadas em conjuntos de dados que não são previamente rotulados (JAIN; MURTY; FLYNN, 1999). Quando os dados são representados como pontos em um espaço, o protótipo pode ser representado como o ponto central do grupo. No caso de agrupamento de documentos de texto, o protótipo pode possuir uma lista de palavras mais frequentes do grupo.

Para a formação dos agrupamentos é necessário que haja um critério para estabelecer quais elementos devem estar em um mesmo grupo. Neste sentido, existem diversas medidas que podem ser utilizadas para determinar o grau de similaridade entre os dados. Como explicado por Cherkassky e Mulier (2007), tais medidas são conhecidas como medidas de similaridade, e são selecionadas subjetivamente com o intuito de alcançar agrupamentos que sejam interessantes para o contexto dos dados. Definir como será medida a similaridade entre dois itens do conjunto de dados analisado pode não ser uma

tarefa trivial. Como expõe Huang (2008), nem sempre é clara a definição de similaridade ou dissimilaridade entre dois documentos, visto que ela normalmente varia de acordo com o contexto do problema real. Como exemplos de medidas de similaridade pode-se citar a Distância Euclidiana, Similaridade Cosseno e Distância Minkowsky (XU et al., 2005).

Conforme explicam Jain, Murty e Flynn (1999) e Xu e Wunsch (2009), as técnicas de agrupamento são geralmente divididas em duas categorias: agrupamento particional e agrupamento hierárquico. Algoritmos de agrupamentos particionais dividem os dados em um número pré-estabelecido de grupos sem estrutura hierárquica, identificando as partições que otimizam o critério de agrupamento. Conforme definem Hansen e Jaumard (1997), dada uma amostra $O = \{O_1, O_2, ..., O_N\}$ de N elementos a serem agrupados, o agrupamento particional busca estabelecer um conjunto de M partições $P_m = \{C_1, C_2, ..., C_M\}$, tal que:

a) $C_j \neq \emptyset \qquad j = 1, 2, \dots, M$

b) $C_i \cap C_j = \emptyset \qquad i, j = 1, 2, \dots, M \qquad i \neq j$

c) $\bigcup_{j=1}^{M} C_j = O$

Desta forma, os agrupamentos gerados não podem ser vazios, nem possuir intersecções entre si. Já os algoritmos de agrupamentos hierárquicos, geram agrupamentos aninhados, como conjuntos e subconjuntos, sendo tal resultado atingido a partir de critérios para unir ou dividir cada agrupamento. Como também definido por Hansen e Jaumard (1997), técnicas de agrupamento hierárquico buscam organizar uma amostra O em uma estrutura hierárquica de partições $H = \{P_1, P_2, ..., P_q\}$, $q \leq N$, tal que $C_i \in P_k$, $C_j \in P_l$ e $k > l$ implica que $C_i \subset C_j$ ou $C_i \cap C_j = \emptyset$ para todo $i, j \neq i,k,l = 1...q$.

Como explicado por Xu et al. (2005) e como mostra a Figura2, o processo de agrupamento se divide em quatro etapas: seleção e extração de atributos, projeto ou seleção do algoritmo de agrupamento, validação dos agrupamentos e interpretação dos resultados.

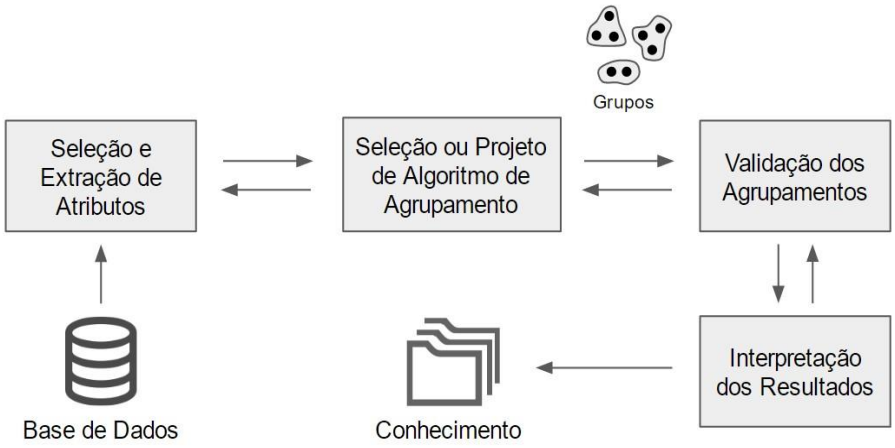


Figura 2 - O processo de agrupamento

Fonte: Inspirada na ilustração apresentada por Xu e Wunsch (2009).

A primeira etapa consiste na extração e seleção de atributos. A seleção de atributos consiste em, a partir do conjunto de atributos dos dados analisados, identificar um subconjunto que será utilizado futuramente pelos algoritmos. Já a extração de atributos consiste no processamento dos atributos originais existentes com o intuito de gerar um novo conjunto de atributos (JAIN; MURTY; FLYNN, 1999).

A segunda etapa consiste no projeto ou seleção de um algoritmo de agrupamento. Esta escolha está quase sempre relacionada à escolha da medida de similaridade a ser utilizada. Em geral, esta etapa apresenta obstáculos pois não há um algoritmo universal que possa resolver satisfatoriamente todos os problemas de agrupamento, sendo necessário anteriormente efetuar uma

investigação do contexto do problema para assim selecionar (ou desenvolver) o algoritmo apropriado (XU et al., 2005).

A terceira etapa consiste na validação dos agrupamentos gerados. Para garantir a confiabilidade dos resultados obtidos em técnicas de agrupamentos, é essencial a utilização de critérios e padrões de avaliação efetivos, visto que os agrupamentos gerados podem variar conforme diversos fatores, como o algoritmo utilizado, parâmetros e até mesmo a ordem em que os dados são analisados (XU et al., 2005).

A última etapa consiste na interpretação dos resultados. Os agrupamentos por si só podem não expressar informações totalmente evidentes, sendo necessária a análise de um especialista no contexto do problema, para que este possa extrair o significado dos agrupamentos gerados.

2.2 Representação dos dados

Após obtidos o conjunto de dados a ser agrupado e selecionados quais são seus atributos a serem considerados pelos algoritmos, é necessário, então, organizar tais dados de forma que possam ser trabalhados. Para tal, é normalmente utilizada a representação vetorial. Nesta abordagem, considerando N o número de elementos e p atributos a serem analisados nos algoritmos, cada objeto de dados consiste em um vetor p-dimensional, sendo suas coordenadas medidas a partir de cada um dos atributos. Desta forma, o conjunto de dados pode ser representado por uma matriz $N \times p$, onde cada linha representa um objeto e cada coluna representa um atributo (XU; WUNSCH, 2009). Esta matriz é conhecida como matriz padrão (JAIN; DUBES, 1988).

Como explicam Jain e Dubes (1988), os dados que compõem a matriz padrão, ou seja, os valores dos atributos dos objetos de dados da amostra, podem ser categorizados segundo seu tipo e sua escala. Reconhecer tais características

dos dados é um fator importante para a seleção do algoritmo de agrupamento a ser utilizado.

O tipo dos dados diz respeito ao seu grau de quantização, podendo ser de três tipos: contínuos, discretos ou binários. Os atributos contínuos são aqueles que podem assumir incontáveis e infinitos valores e normalmente são resultados de medições, como temperatura, peso e altura. Os atributos discretos possuem um conjunto contável ou limitado de possíveis valores, como cores, categoria de filme e modelos de carros. Dados binários são aqueles que podem assumir apenas dois valores, como sexo, sendo estes considerados uma subcategoria especial de dados discretos (JAIN; DUBES, 1988; XU; WUNSCH, 2009).

A escala dos dados diz respeito à significância relativa dos números, podendo os dados ser categorizados em qualitativos (nominais e ordinais) e quantitativos (intervalos e de proporção). Dentre os valores qualitativos, os nominais são na verdade rótulos ou nomes, que não possuem nenhum significado no sentido matemático, não podendo ser comparados em relação a ordem ou tamanho. Por exemplo, os valores (sim, não) podem até ser convertidos para os números (1, 0), porém não possuem um significado quantitativo. Os valores ordinais, também são representados por rótulos ou nomes, porém, possuem uma ordem. Esta ordem pode ser utilizada apenas na comparação de dois valores, não sendo possível quantificar sua diferença. Por exemplo: pequeno, médio, grande (JAIN; DUBES, 1988; XU; WUNSCH, 2009).

Já no contexto dos valores quantitativos, os valores de intervalo são aqueles para os quais já é possível calcular a diferença entre dois valores, porém a sua interpretação é dependente da unidade de medida, e normalmente não existe o conceito de zero. Por exemplo, a interpretação de uma nota 10 em uma avaliação depende se o intervalo utilizado é de 0 a 10 ou de 0 a 100. Outro exemplo é a temperatura, que em graus Celsius têm uma interpretação diferente

de graus Fahrenheit, e o valor 0 não representa de fato a ausência de calor. Nesta categoria os valores também não são significativamente proporcionais entre si, por exemplo, 30ºC não possuem de fato duas vezes a quantidade de calor que 15ºC (JAIN; DUBES, 1988; XU; WUNSCH, 2009).

Os valores de proporção são os valores onde os números possuem significado absoluto. Existe o conceito de zero e também pode ser considerada a proporção entre dois valores. Por exemplo, a distância entre dois pontos em metros. Neste caso 0m significa que não há distância, e 20m é exatamente o dobro da distância de 10m. Tais propriedades permanecem válidas mesmo no caso da mudança de unidade de medida (JAIN; DUBES, 1988; XU; WUNSCH, 2009).

Para o problema de agrupamento de textos, uma abordagem comumente utilizada é aquela onde os atributos dos documentos são os diferentes termos que eles possuem, sendo que seus valores consistem na frequência de cada termo dentro do documento, também conhecida como *tf* (*term frequency*). Por exemplo, dados os termos t_1, t_2, t_3, t_4, t_5, t_6 e os seguintes documentos:

a) d_1 = “t_1 t_2 t_1 t_1 t_5 t_6”;
b) d_2 = “t_2 t_5 t_3 t_4 t_3 t_4 t_5 t_5”;
c) d_3 = “t_6 t_6 t_1 t_2 t_3”.

Os vetores destes documentos, utilizando *tf*, podem ser representados conforme a Figura 3.

	t_1	t_2	t_3	t_4	t_5	t_6
d_1	3	1	0	0	1	1
d_2	0	1	2	2	3	0
d_3	1	1	1	0	0	2

Figura 3 - Exemplo de representação vetorial para documentos textuais utilizando frequência dos termos

2.3 Medidas de similaridade e distância

As técnicas de agrupamento baseiam-se em medidas de proximidade para optar por inserir ou não elementos em um mesmo grupo (CHERKASSKY; MULIER, 2007). Isto torna as medidas de proximidade um ponto chave para o funcionamento destes procedimentos, sendo a medida a ser utilizada uma escolha que requer cuidado, visto que existem diferentes tipos e escalas de dados (JAIN; MURTY; FLYNN, 1999).

Para medir a proximidade entre dois elementos, podem ser utilizadas medidas de similaridade ou medidas de dissimilaridade (distância), sendo que estas duas medidas são frequentemente conversíveis entre si (XU; WUNSCH, 2009). No caso das medidas de similaridade, quanto maior seu valor, mais semelhantes são os dois documentos e maior a probabilidade de serem inseridos em um mesmo grupo. Para as medidas de distância ocorre o inverso, quanto maior seu valor menor a semelhança entre os documentos, sendo menor a probabilidade de serem inseridos em um mesmo grupo.

No contexto de agrupamento de textos, a eficácia das diversas medidas de similaridade e dissimilaridade não está totalmente clara (HUANG, 2008). Desta forma, serão descritas a seguir algumas das medidas populares e comumente utilizadas no contexto de agrupamentos de textos. Para tal, deve-se considerar d_i e d_j dois documentos, x_i e x_j seus respectivos vetores de termos, p o

número de dimensões destes vetores e α o ângulo formado por estes dois vetores.

Uma medida de distância muito popular, utilizada para valores contínuos, é a distância euclidiana, apresentada na Equação 2.1. Esta medida é bastante intuitiva, visto que é utilizada para o cálculo da distância de objetos no espaço (distância geométrica) (JAIN; MURTY; FLYNN, 1999). A distância euclidiana é a distância padrão utilizada no algoritmo *K-means*.

$$D_{euc}(x_i, x_j) = \left(\sum_{k=1}^{d}\left(x_{i,k} - x_{j,k}\right)^2\right)^{1/2} \quad (2.1)$$

Outra medida de proximidade é a similaridade cosseno (EQUAÇÃO 2.2), que é bastante utilizada no agrupamento de documentos de textos. Nesta medida não é considerado o tamanho dos vetores x_i e x_j, mas apenas suas direções e o ângulo formado por eles. No contexto de agrupamentos de textos, isso significa que documentos compostos pelos mesmos termos (e a mesma proporção entre eles), porém com diferentes totais de termos, serão considerados semelhantes (HUANG, 2008).

$$D_{cos}\left(x_i, x_j\right) = \frac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|} \quad (2.2)$$

Estas são apenas alguns exemplos de medidas de similaridade e distância utilizadas em técnicas de agrupamento. Outros exemplos são sumarizados por (XU et al., 2005).

2.4 Agrupamento hierárquico

As técnicas de agrupamento hierárquico caracterizam-se por gerar grupos de dados com partições aninhadas (subgrupos) (XU; WUNSCH, 2009).

Tal hierarquia de grupos pode ser representada como uma árvore, que provê a visualização dos dados em diferentes níveis de abstração. Esta estrutura contendo agrupamentos de dados em diferentes níveis de granularidade é ideal para a exploração interativa e para a sua visualização (ZHAO; KARYPIS; FAYYAD, 2005).

Os algoritmos utilizados para tal agrupamento são divididos em duas categorias: divisivos e aglomerativos. Algoritmos aglomerativos iniciam com a criação de um grupo por documento e, ao longo das iterações do algoritmo, estes grupos vão sendo unidos de maneira que se forme um único grupo no final, que contém todos os subgrupos e documentos. Os algoritmos divisivos trabalham de forma oposta, iniciando com um único grupo, contendo todos os documentos e, ao longo das iterações, os grupos são divididos em subgrupos até que haja apenas um único elemento por grupo (FELDMAN; SANGER, 2006). A Figura 4 mostra um exemplo de hierarquia gerada por técnicas de agrupamento. Neste exemplo, os algoritmos aglomerativos gerariam a hierarquia partindo de baixo para cima no dendrograma, já os divisivos trabalhariam no sentido oposto.

Os algoritmos aglomerativos são mais amplamente utilizados para agrupamento hierárquico. O comportamento geral destes algoritmos pode ser descrito pelos seguintes passos (XU et al., 2005):

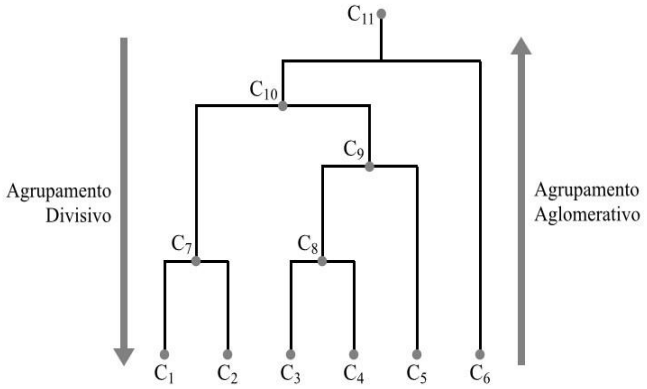


Figura 4 - Exemplo de dendrograma gerado a partir de agrupamento hierárquico

Fonte: adaptado de Xu e Wunsch (2009).

1. Iniciar com *N* grupos, cada um contendo um documento. Calcular a matriz de proximidade para os *N* grupos, contendo a distância de cada um deles para todos os demais.
2. Dada uma função de distância *Dist*(∗,∗), efetuar uma busca, na matriz de proximidade, pela distância mínima:

$$Dist\left(C_i, C_j\right) = \min_{\substack{1 \le m, \; l \le N \\ m \ne 1}} Dist\left(C_m, C_j\right)$$

   Então, combine C_i e C_j, para formar um novo grupo.
3. Atualizar a matriz de proximidade para que o novo grupo gerado seja incluído e suas distâncias para todos os outros grupos já existentes sejam calculadas.
4. Repetir os passos 2 e 3 até que todos os documentos estejam em um mesmo grupo.

Para entendimento, segundo Hansen e Jaumard (1997), a matriz de similaridade, também conhecida como matriz de dissimilaridade, é definida como $D = (d_{kl})$, uma matriz $N \times N$, sendo N o número de grupos e $d_{k \times l}$ a distância entre os grupos k e l. As distâncias contidas na matriz normalmente respeitam as seguintes propriedades: $d_{kl} \geq 0$, $d_{kk} = 0$ e $d_{kl} = d_{lk}$, para $k,l = 1,2,...,N$.

2.5 Agrupamento incremental

A maioria dos algoritmos de agrupamento presentes na literatura não trabalham com dados dinâmicos, ou seja, necessitam de toda a base de dados disponível para efetuar o agrupamento. Com o advento da web, onde diariamente surgem novas informações, tais algoritmos podem não ser eficazes (SAHOO et al., 2006). Para sanar esta necessidade, foram desenvolvidos algoritmos de agrupamento incremental, capazes de trabalhar com *data streams*. *Data streams* são sequências ordenadas de dados que são gerados continuamente (GUHA et al., 2003; SILVA et al., 2013). Como formalizado por Silva et al. (2013), um *data stream S* é uma sequência de objetos de dados d_1, d_2, ..., d_N, sendo $S = \{d_i\}_{i=1}^{N}$, podendo esta sequência pode ser ilimitada ($N \rightarrow \infty$).

Neste cenário, pode-se observar que o algoritmo não tem a base de dados completa a sua disposição, sendo necessário efetuar uma leitura linear destes dados (GUHA et al., 2003), onde eles são lidos em sequências, conforme disponibilizados, evitando a releitura de dados já processados. Tais restrições fazem com que trabalhar com dados dinâmicos seja bem mais complexo que com dados estáticos (PHRIDVIRAJ; SRINIVAS; GURURAO, 2014). Assim, a principal característica de agrupamentos incrementais é que, ao ter um objeto de dados adicionado ou removido, o algoritmo consegue adaptar suas estruturas de agrupamento, sem a necessidade de reagrupar todos os dados (GIL-GARCÍA; PONS-PORRATA, 2010a). Estes algoritmos também são utilizados em contextos onde há um grande volume de dados e não há como armazená-los na

memória, e também em situações onde os dados em si não são armazenados no disco, e sim apenas uma sumarização dos mesmos (GUHA et al., 2003).

Segundo Silva et al. (2013), para o desenvolvimento de algoritmos que mineram *data streams*, deve-se considerar as seguintes restrições:

a) Novos dados serão obtidos continuamente;
b) Não há uma ordem definida para que os dados sejam processados;
c) O *data stream* pode ser ilimitado;
d) Os dados são descartados após serem processados, evitando sua releitura;
e) O processo de geração dos dados é desconhecido e não-estacionário.

2.6 Algoritmos

Tendo sido já apresentados os principais conceitos relativos ao problema de agrupamento hierárquico de documentos, esta seção visa apresentar os dois algoritmos utilizados neste trabalho: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) e Dynamic Hierarchical Compact (DHC).

2.6.1 Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

O algoritmo UPGMA, apresentado por Jain e Dubes (1988), é um método de agrupamento hierárquico não incremental, ou seja, necessita de toda a base de dados disponível para a sua execução. A ideia principal deste algoritmo é buscar os dois grupos mais similares, uni-los como subgrupos de um novo grupo e removê-los da lista de grupos a serem verificados. Tal processo é repetido até que sobre apenas um único grupo, que corresponde ao grupo raiz da hierarquia.

O Algoritmo 1 descreve os passos executados pelo UPGMA. As linhas 1 e 2 efetuam a leitura de todos os documentos e o cálculo da matriz de similaridade. Neste procedimento, são calculadas as similaridades entre todos os

documentos, preenchendo completamente a matriz. Na linha 3 é inicializada uma lista para armazenar os grupos a serem verificados pelo algoritmo. Nas linhas de 4 a 7 cada documento é inserido em um novo grupo, e cada um destes grupos são adicionados à lista de grupos a serem verificados. Estes grupos iniciais são as folhas da hierarquia a ser formada. Assim, enquanto houver mais de um grupo restante a ser verificado, os grupos mais similares são selecionados (linha 9), são unidos gerando um novo grupo (linha 10) e este novo grupo é adicionado no conjunto de grupos a serem verificados (linha 11), de onde seus subgrupos são removidos (linha 12). Após tais operações, a matriz de similaridade é atualizada (linha 13), removendo valores de similaridade referentes aos grupos removidos e adicionando os valores para o novo grupo criado. É possível observar que a cada iteração (linhas de 8 a 14) a lista de grupos a serem verificados terá um elemento a menos, até que reste apenas um (raiz da hierarquia), finalizando o algoritmo.

Algoritmo 1 Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

1: Ler todos os documentos da base de dados
2: Calcular matriz de similaridade completa
3: Criar lista vazia de grupos a serem verificados
4: **para cada** documento **faça**
5:Criar um novo grupo contendo o documento
6:Adicionar o grupo na lista de grupos a serem verificados
7: **fim para**
8: **enquanto** existe mais de um grupo a ser verificado **faça**
9:Encontrar os dois grupos de maior similaridade, c_i e c_j
10:Criar novo grupo c_N contendo c_i e c_j como subgrupos
11:Adicionar c_N à lista de grupos a serem verificados
12:Remover c_i e c_j da lista de grupos a serem verificados
13:Atualizar matriz de similaridade
14: **fim enquanto**
15: O grupo restante na lista de grupos a serem verificados corresponde à raiz da hierarquia

Este algoritmo possui a característica de gerar muitos grupos e hierarquias muito verticais. Isto se deve ao fato de, na estrutura gerada ao final do algoritmo, cada grupo possuirá apenas dois subgrupos.

2.6.2 Dynamic Hierarchical Compact (DHC)

O DHC, proposto por Gil-García e Pons-Porrata (2010a), é um algoritmo de agrupamento hierárquico incremental, ou seja, a cada documento que lê, adapta suas estruturas a fim de inserir este documento na hierarquia, não havendo a necessidade de toda a base de dados disponível inicialmente.

Para a compreensão do algoritmo, é necessário primeiramente conhecer algumas estruturas utilizadas por ele. Ao longo da execução do algoritmo, dois grafos são criados para cada nível da hierarquia. O primeiro é o grafo de β-similaridade, onde os vértices são os grupos do nível correspondente da hierarquia e existe uma aresta entre dois vértices se a similaridade entre os grupos for maior ou igual a um parâmetro β. Neste caso, os dois grupos são chamados de β-similares. O Segundo grafo é o *max-S*, que é um subgrafo do β-similaridade, contendo todos os seus vértices, mas nem todas as arestas. No grafo *max-S* cada grupo possui uma aresta para o outro grupo mais similar a ele. A Figura 5, inspirada no exemplo apresentado por Gil-García e Pons-Porrata (2010a), apresenta um grafo de β-similaridade e seu correspondente *max-S*. O valor de β é um parâmetro passado para o algoritmo e, neste exemplo, $\beta = 0{,}4$.

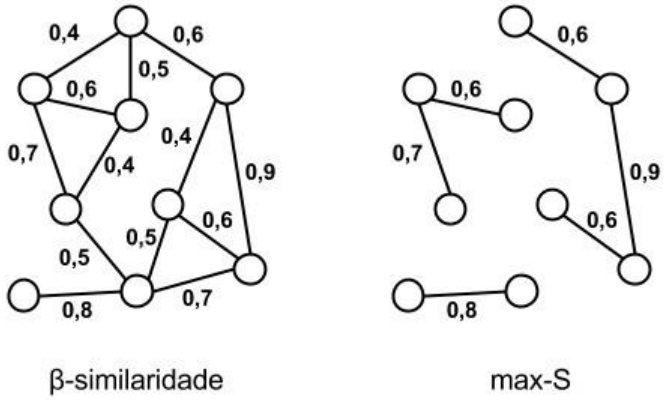


Figura 5 - Exemplo de grafo de β-similaridade e seu correspondente *max-S* (β= 0,4)

Utilizando estas estruturas, Gil-García e Pons-Porrata (2010a) propõem o *Dynamic Hierarchical Agglomerative Framework*, um conjunto de passos capaz de construir uma hierarquia de documentos. Dado um novo documento a ser adicionado, este *framework* consiste em, para cada nível da hierarquia, (1) atualizar os grafos *β-sim* e *max-S* e (2) encontrar uma cobertura de vértices para o grafo *max-S*, cujos elementos vão compor o grafo *β-sim* do próximo nível da hierarquia.

O Algoritmo 2 apresenta os passos executados pelo *Dynamic Hierarchical Agglomerative Framework*. Dados um novo documento a ser adicionado (linha 1), a similaridade deste para os demais documentos é calculada (linha 2), ele é inserido em um novo grupo (linha 3) e adicionado no grafo de β-similaridade do nível mais baixo da hierarquia de grupos (linha 4). A adição do grupo neste grafo consiste na criação de um vértice e arestas (no caso de similaridades maiores ou iguais que β). Em seguida, o algoritmo irá propagar esta alteração ao longo dos níveis da hierarquia (iteração das linhas 6 a 11). Esta iteração é executada enquanto o grafo de β-similaridade do nível analisado não for completamente desconectado (ou seja, possuir arestas). Quando é alcançado um nível completamente desconectado significa que não há mais agrupamentos

a serem realizados. Para cada nível, primeiramente é atualizado o grafo *max-S* (linha 7). É importante observar que, adicionado um vértice no grafo *max-S*, outros vértices e arestas podem ser impactados, visto que podem surgir graus de similaridade entre os vértices maiores do que os previamente existentes. O algoritmo 3 descreve com mais detalhes tal procedimento. No segundo passo da iteração é identificada uma cobertura de vértices para o grafo *max-S* (linha 8). É importante destacar que "cobertura de vértices" aqui consiste em um conjunto de subconjuntos de vértices cuja união resulta na totalidade dos vértices do grafo; cada um destes subconjuntos se tornará um grupo no próximo nível da hierarquia. Este passo não é detalhado no *framework*, sendo que diferentes algoritmos podem ser propostos apenas implementando o procedimento de identificar a cobertura de vértices. No trabalho de Gil-García e Pons-Porrata (2010a) são propostos os algoritmos DHS e DHC (utilizado neste trabalho). Por último, dados os subconjuntos de vértices da cobertura, o grafo de *β*-similaridade do próximo nível é atualizado (linha 9), para que o próximo nível da hierarquia possa ser analisado. Terminadas as iterações, o algoritmo para no último nível que deve existir na hierarquia, sendo os superiores a ele descartados (linha 12).

Algoritmo 2 Dynamic Hierarchical Agglomerative Framework

1: Ler novo documento a ser adicionado
2: Calcular similaridade do novo documento para os demais
3: Criar novo grupo contendo o novo documento
4: Atualizar grafo de *β*-similaridade do nível mais baixo da hierarquia $\beta\text{-}sim_0$
5: *nível* ← 0
6: **enquanto** o grafo $\beta\text{-}sim_{nível}$ não for completamente desconectado **faça**
7: Atualizar grafo $max\text{-}S_{nível}$
8: Atualizar a cobertura de vértices para grafo $max\text{-}S_{nível}$
9: Atualizar o grafo de β-similaridade do próximo nível, $\beta\text{-}sim_{nível+1}$
10: *nível* ← *nível* +1
11: **fim enquanto**
12: Remover os níveis da hierarquia maiores que *nível*, caso existam

Uma etapa complexa deste *framework* é a atualização do grafo *max-S* pois, uma vez adicionado um novo vértice, outros vértices e arestas podem ser impactados. Neste sentido, é razoável que apenas as estruturas afetadas sejam atualizadas, ao invés de reconstruir todo o grafo. Para tal, Gil-García e Pons-Porrata (2010a) apresentam um procedimento, apresentado no Algoritmo 3.

Apresentado o *Dynamic Hierarchical Agglomerative Framework*, o algoritmo DHC consiste em uma implementação deste *framework* onde a cobertura de vértices do grafo *max-S* corresponde ao conjunto de componentes conectadas no grafo. A Figura 6 apresenta um exemplo de cobertura de vértices considerada pelo algoritmo DHC.

Algoritmo 3 Atualização do grafo *max-S*

1: $N \leftarrow$ conjunto de vértices a serem adicionados no grafo
2: $R \leftarrow$ conjunto de vértices a serem removidos do grafo
3: $M \leftarrow$ conjunto de vértices cujo vértice mais β-similar está em R
4: Remover do grafo todos os vértices contidos em R
5: Adicionar ao grafo todos os vértices contidos em N
6: Encontrar o vértices mais β-similar a cada um dos vértices contidos em $N \cup M$ e adicionar as arestas correspondentes
7: Encontrar os vértices para os quais o vértice mais β-similar está contido em N e atualizar as arestas correspondentes

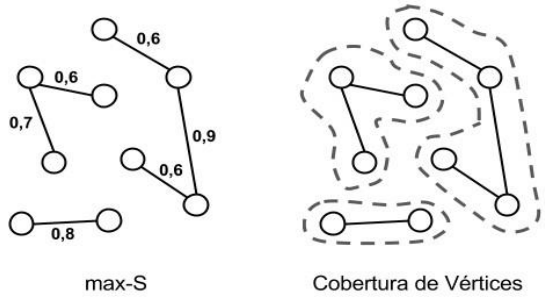


Figura 6 - Exemplo de grafo de *max-S* e a cobertura de vértices considerada pelo algoritmo DHC

Cada componente conectada do grafo se tornará um grupo no nível acima da hierarquia. É importante que, durante a atualização da cobertura de vértices, sejam atualizadas apenas as estruturas necessárias. Uma reconstrução total da cobertura acarretaria a reconstrução total de todos os níveis acima na hierarquia. Assim, no trabalho de Gil-García e Pons-Porrata (2010a) é apresentado um procedimento para atualização da cobertura de vértices, detalhado aqui no Algoritmo 4. Este procedimento consiste em, dadas as atualizações efetuadas no grafo *max-S* (linhas 1 a 4), identificar quais os vértices afetados e a quais grupos eles pertencem e, então, adicionar em uma fila *Q* todos os vértices destes grupos (linhas 6 a 8). Estes grupos afetados são removidos da lista de grupos existentes, em outras palavras, serão excluídos da hierarquia. Aqui, quando mencionado o grupo ao qual um vértice pertence, refere-se à componente conectada, ou seja, o grupo do próximo nível da hierarquia. Em seguida, os novos vértices também são adicionados à fila *Q* para análise (linha 9). Por último, são identificadas as componentes conectadas dos vértices contidos em *Q* e estas serão novos grupos a serem adicionados na hierarquia.

Algoritmo 4 Atualização do grafo *max-S*

1: *N* ← conjunto de vértices adicionados ao grafo *max-S*
2: *R* ← conjunto de vértices removidos do grafo *max-S*
3: *NE* ← conjunto de arestas adicionadas ao grafo *max-S*
4: *RE* ← conjunto de arestas removidas do grafo *max-S*
5: *Q* ← fila vazia de vértices a serem processados
6: Remover todos os vértices de *R* de seus grupos e inserir os demais vértices destes grupos em *Q*.
7: Inserir em *Q* todos vértices de grupos que possuem pelo menos um vértice com incidência de uma aresta em *RE*
8: Remover da lista de grupos existentes todos grupos cujos vértices foram afetados nos dois passos anteriores
9: Adicionar em *Q* todos os vértices de *N*
10: Construir grupos a partir das componentes conectadas dos vértices contidos em *Q* e adicioná-lo na lista de grupos existentes. Para cada aresta em *NE*, unir os dois grupos aos quais os vértices envolvidos pertencem

Pode-se observar que apenas os vértices afetados pelas atualizações em *max-S* são reprocessados. Vértices que não foram adicionados em *Q* são mantidos em seus grupos sem nenhuma alteração.

É importante enfatizar que, os grupos (componentes) que aqui forem adicionados e removidos são os grupos que serão adicionados e removidos do grafo de β-similaridade do próximo nível da hierarquia (linha 9 do Algoritmo 2).

2.7 Métricas de avaliação dos agrupamentos

Segundo Rijsbergen (1979), a avaliação dos sistemas de recuperação de informação (incluindo os algoritmos de agrupamento) é de grande importância para os usuários, tanto em questões sociais, quanto econômicas. Primeiramente, estas avaliações auxiliam os usuários a verificar a necessidade de utilizar um destes sistemas, substituindo os meios já utilizados e, em seguida, selecionar qual deles a ser utilizado. Por outro lado, a avaliação também auxilia os usuários a verificar o custo de um destes sistemas e se vale a pena utilizá-lo. Dentre as características de um sistema de recuperação de informação, *Precision* e *Recall*, e as métricas derivadas destas se destacam quando o objetivo é demostrar a efetividade do sistema.

Esta seção descreve detalhadamente algumas das métricas utilizadas para avaliação de agrupamentos. As métricas aqui descritas foram identificadas a partir dos trabalhos recentes descritos no Capítulo 4. Serão detalhadas nesta seção apenas as métricas utilizadas nos experimentos deste trabalho, sendo algumas outras descritas no Apêndice B.

2.7.1 *Precision e Recall*

Para a avaliação de algoritmos de agrupamento, algumas das métricas mais amplamente utilizadas são aquelas baseadas nos valores de *Precision* e *Recall*. Em sistemas de recuperação de informação, estas são medidas de relevância utilizadas para avaliar os dados recuperados. *Precision* representa

fração de instâncias recuperadas que são relevantes e *Recall* representa a fração das instâncias relevantes que foram recuperadas. Assim, obter um alto valor de *Precision* significa que a maior parte das instâncias retornadas são, de fato, relevantes enquanto, obter um alto valor de *Recall* significa que a maior parte dos dados relevantes foram retornados (RIJSBERGEN, 1979).

No contexto de técnicas de agrupamento, *Precision* e *Recall* são utilizados em testes onde a base de dados utilizada é previamente classificada, a fim de comparar esta classificação com os agrupamentos gerados. Para entendimento, deve-se considerar $C = \{C_1,...,C_n\}$ o conjunto de n grupos de documentos gerados por um algoritmo de agrupamento e $L = \{L_1,...,L_m\}$ o conjunto de m classes reais dos documentos (previamente atribuídas).

O valor de *Precision* (EQUAÇÃO 2.3) do grupo C_i em relação a classe L_j representa a fração de documentos inseridos no grupo C_i que pertencem a classe L_j.

Seu valor pode variar de 0 a 1 e, se $P(C_i, L_j) = 1$, significa que todos os documentos inseridos no grupo C_i pertencem a classe L_j. O Valor de *Recall* (EQUAÇÃO 2.4) de C_i em relação a L_j e representa a fração de documentos que pertencem a classe L_j que estão inseridos no grupo C_i. Seu valor também varia de 0 a 1 e, se $R(C_i, L_j) = 1$, significa que o grupo C_i contém todos os documentos da classe L_j (MARCACINI; HRUSCHKA; REZENDE, 2012).

$$P(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|} \tag{2.3}$$

$$R(C_i, L_j) = \frac{|C_i \cap L_j|}{|L_j|} \tag{2.4}$$

Se ambos os valores de *Precision*(C_i, L_j) e *Recall*(C_i, L_j) forem iguais a 1, significa que o grupo C_i corresponde exatamente a classe L_j. É possível

observar que estes valores, de forma independente, não são capazes de representar a correspondência entre o grupo C_i e a classe L_j e, por isso, diversas métricas de avaliação de agrupamentos utilizam uma combinação entre estes valores. Algumas dessas métricas serão descritas a seguir.

2.7.2 *F-Measure*

F-measure (LARSEN; AONE, 1999; RIJSBERGEN, 1979)(também conhecida como F_{SCORE} e F_1) é uma métrica amplamente utilizada na avaliação de algoritmos de agrupamento. O valor de *F-measure* do grupo C_i em relação a classe L_j é definido pela equação 2.5, que representa a média harmônica entre *Precision* e *Recall*. O valor de *F-measure* pode variar de 0 a 1 e, quanto maior o seu valor, maior a é a correspondência entre o grupo C_i e a classe L_j.

$$F(C_i, L_j) = \frac{2 \times P(C_i,L_j) \times R(C_i,L_j)}{P(C_i,L_j) + R(C_i,L_j)} \tag{2.5}$$

Para obter o valor geral de F de uma classe L_j, é utilizado o maior valor de $F(C_i, L_j)$ (EQUAÇÃO 2.6), considerando todos os grupos $C_i \in C$, ou seja, o valor de $F(L_j)$ considerado é calculado com o grupo que possui a maior correspondência com a classe L_j.

Assim, o valor geral de *F-measure* para o problema de agrupamento, considerando todos os grupos $C_1,...,C_n$, gerados a partir de um conjunto de documentos, é dados pela equação 2.7 (MARCACINI; HRUSCHKA; REZENDE, 2012; ZHAO; KARYPIS; FAYYAD, 2005). Esta é a média dos valores de $F(L_j)$ ponderada pelo número de documentos contidos nas classes. Alguns autores chamam este cálculo geral de *Overall* F_1 (GIL-GARCÍA; PONS-PORRATA, 2010a, 2010b), outros de *Overall F-measure* (CORREA-MORRIS et al., 2010; LARSEN; AONE, 1999), alguns de F_{SCORE} (MARCACINI; HRUSCHKA; REZENDE, 2012; MARCACINI; REZENDE, 2013; SINOARA

et al., 2014) e outros apenas de *F-measure* (HUANG et al., 2011; PENG; LIU, 2015).

$$F(L_j) = \max_{C_i \in C} F(C_i, L_j) \tag{2.6}$$

$$Overall\ F\text{-}measure = \frac{\sum_{j=1}^{m}\left(|L_j|\ F(L_j)\right)}{\sum_{j=1}^{m}|L_j|} \tag{2.7}$$

Um ponto a ser observado é que, para o cálculo de *F −measure* na avaliação de algoritmos de agrupamento hierárquicos, cujo resultado forma grupos e subgrupos, ao calcular os valores de *Recall* e *Precision* de um grupo, são considerados também os documentos de seus subgrupos (LARSEN; AONE, 1999).

2.7.3 *Overall F1-Travel*

A métrica *Overall F1-Travel*, apresentada por Gil-García e Pons-Porrata (2010b), também é adequada para a avaliação de agrupamentos hierárquicos. Ela é bastante similar a *Overall F-Measure*, porém, busca avaliar a navegabilidade da hierarquia gerada. Navegabilidade consiste na facilidade de encontrar um determinado grupo na hierarquia. Tanto hierarquias muito verticais, quanto as muito horizontais podem dificultar a busca por um documento e esta métrica busca avaliar o custo de navegação para encontrar os tópicos. A Equação 2.8 apresenta o cálculo de *F1-Travel*(L_j), que representa o custo de navegação para encontrar o tópico L_j. Nesta equação, *n* representa o número total de documentos e *v* o número de grupos visitados partindo da raiz da hierarquia até o grupo $\sigma(L_j)$. O grupo $\sigma(L_j)$ é aquele que mais se parece com a classe L_j; em outras palavras, é o grupo C_i que maximiza o valor de $F(C_i, L_j)$ (EQUAÇÃO 2.5). Para a contabilização de *v* é utilizada uma busca "*best-first*", detalhada por Gil-García e Pons-Porrata (2010b).

$$F1\text{-}Travel(L_j) = F\left(L_j, \sigma(L_j)\right)\left(1 - \frac{v}{2n}\right) \quad (2.8)$$

Assim, pode definir o valor de *Overall F1-travel* pela equação 2.9.

$$Overall\ F1\text{-}Travel = \frac{\sum_{j=1}^{m}\left(|L_j|\ F1\text{-}Travel(L_j)\right)}{\sum_{j=1}^{m}|L_j|} \quad (2.9)$$

3 ENTIDADES NOMEADAS

Entidades nomeadas são informações como nomes de pessoas, locais, organizações e expressões monetárias e datas (AL-ONAIZAN; KNIGHT, 2002) e podem ser comumente encontradas em páginas na Web, notícias, postagens em blogs e mídias sociais (HOFFART et al., 2011). Tais informações têm sido utilizadas com o intuito de melhorar a eficácia de técnicas de mineração de texto, como algoritmos de agrupamento (SINOARA et al., 2014) e detecção de novos eventos (ZHANG; ZI; WU, 2007). Como explica Sinoara et al. (2014), entidades nomeadas são mais informativas que outros atributos em documentos textuais, o que reforça a hipótese de que utilizá-las em algoritmos de agrupamento pode trazer melhora na qualidade dos resultados.

Visto que as entidades nomeadas são informações de grande valor que podem ser extraídas do conteúdo dos documentos, torna-se necessário investigar a melhor forma de utilizar tais informações no processo de agrupamento destes documentos. Neste trabalho são exploradas diferentes abordagens de utilização das entidades para auxiliar no cálculo de similaridade entre os documentos. É razoável que dois documentos que possuam entidades em comum estejam relacionados a um mesmo tópico, porém, não se pode desconsiderar os termos do documento, pois nem todos os documentos possuirão entidades em seu conteúdo. Assim é necessário conciliar estas duas informações: termos e entidades dos documentos.

Neste trabalho a extração de entidades nomeadas dos documentos foi efetuada com o uso da ferramenta Apache Stanbol (APACHE..., 2016), que consiste em um *software* para gerenciamento de conteúdo semântico. O Apache Stanbol fornece serviços HTTP para acesso a suas funcionalidades, dentre elas, a extração de entidades. Em outras palavras, é possível efetuar uma requisição HTTP para um dos serviços do Apache Stanbol e receber como resposta as entidades extraídas. Diversas informações sobre uma entidade são fornecidas,

como o nome da entidade, tipo, descrição, entre outras. Neste trabalho foram utilizados apenas os nomes e tipos das entidades.

As Seções 3.1, 3.2, 3.3 e 3.4 descrevem diferentes abordagens para uso destas informações. É importante destacar que a Abordagem 1 (Múltiplos vetores) e a Abordagem 2 (Entidades como termos de maior peso) são abordagens já descritas na literatura, propostas por outros trabalhos. Já a Abordagem 3 (Entidades como termos com peso em função do vetor de termos) e a Abordagem 4 (Entidades como termos com peso em função do vetor de termos e número de entidades) são proposta deste trabalho, elaboradas com o intuito de atingir melhores resultados no processo de agrupamento de documentos.

3.1 Abordagem 1: Múltiplos vetores

Esta abordagem foi proposta por Cao, Tang e Chau (2012) e consiste em, para cada documento, construir os seguintes vetores:

a) T: vetor de frequência de termos (tradicionalmente utilizado nos algoritmos);
b) EN: vetor de nomes das entidades;
c) ET: vetor de tipos das entidades;
d) ENT: vetor resultante das combinações “nome+tipo” das entidades.

Os vetores EN, ET e ENT contém os pesos atribuídos às entidades e tipos, de forma similar à frequência no vetor de termos.

Desta forma, dados dois documentos d_i e d_j, seus respectivos vetores T_i, EN_i, ENT_i e T_j, EN_j, ENT_j e uma função de similaridade $sim(v_1, v_2)$, o cálculo de similaridade entre os documentos é dado pela equação 3.1.

$$sim(d_i, d_j) = w_1 \cdot sim(T_i, T_j) + w_2 \cdot sim(EN_i, EN_j) + w_3 \cdot sim(ET_i, ET_j) + w_4 \cdot sim(ENT_i, ENT_j) \quad (3.1)$$

Na equação 3.1, os valores w_1, w_2, w_3 e w_4 são pesos dados a cada um dos vetores, onde $w_1 + w_2 + w_3 + w_4 = 1$. Estes pesos devem ser parâmetros para o algoritmo e a definição destes valores pode ser um desafio.

3.2 Abordagem 2: Entidades como termos de maior peso

Esta abordagem foi utilizada no trabalho de Huang et al. (2011) e consiste em adicionar os nomes das entidades no vetor de termos, sendo sua frequência t_e dada por uma constante W (EQUAÇÃO 3.2). A constante W é um parâmetro para o algoritmo e representa o peso das entidades em relação aos demais termos.

$$t_e = W \quad (3.2)$$

O propósito de estabelecer a frequência das entidades através de um parâmetro é atribuir a ela peso maior que os demais termos do documento. Ou seja, uma entidade seria considerada um termo mais frequente, tendo maior influência nos cálculos de similaridade. Porém, ao estabelecer o valor de um parâmetro W, não se sabe o conteúdo dos documentos a serem agrupados, a magnitude de seus vetores nem a frequência de seus termos, o que torna difícil encontrar um valor adequado para W. Isto motivou a proposta de outras duas abordagens, apresentadas nas Seções 3.3 e 3.4.

3.3 Abordagem 3: Entidades como termos com peso em função do vetor de termos

Visto o objetivo de tornar as entidades nomeadas termos de maior relevância que os demais, esta abordagem é similar à abordagem 2, porém a frequência atribuída à uma entidade no vetor de termos é proporcional à magnitude deste vetor. Assim, dada uma entidade *e*, o parâmetro *W* e o vetor *v* de um documento, o cálculo da frequência t_e da entidade é dada pela equação 3.3.

$$t_e = W \cdot \|v\| \tag{3.3}$$

Utilizando esta abordagem, em documentos maiores (contendo mais termos) maior peso será dados as entidades. Desta forma, uma entidade terá o mesmo peso proporcional em dois documentos de tamanhos distintos.

3.4 Abordagem 4: Entidades como termos com peso em função do vetor de termos e número de entidades

Esta última abordagem surge da hipótese de que se um documento obtiver poucas entidades, elas terão grande relevância para aquele documento, mas se o documento tiver um número maior de entidades, cada uma delas terá menor relevância. Ou seja, quanto maior o número de entidades contidas em um documento, menor é o peso atribuído a elas. Assim, considerando uma entidade *e*, o parâmetro *W*, o vetor *v* de um documento e *ne* o número de entidades contidas neste documento, o cálculo da frequência t_e da entidade é dada pela equação 3.4.

$$t_e = \frac{W \cdot \|v\|}{ne} \tag{3.4}$$

4 TRABALHOS RELACIONADOS

Compreendendo a importância de se observar as pesquisas em andamento relacionadas ao tema estudado neste projeto, este capítulo apresenta alguns trabalhos relacionados recentes, descrevendo de forma breve cada um deles. Os trabalhos aqui apresentados foram obtidos através de uma pesquisa sistemática, com critérios bem definidos, aplicada à algumas bibliotecas digitais selecionadas. A metodologia utilizada para efetuar este levantamento é detalhada no Apêndice A.

Os trabalhos aqui apresentados estão divididos em dois tipos: desenvolvimento de técnica (desenvolve novas técnicas ou abordagens de agrupamento) e aplicações (aplica técnicas de agrupamento em um contexto específico). É importante observar que os trabalhos do tipo “aplicações" não utilizam, necessariamente, as técnicas propostas nos trabalhos do tipo “desenvolvimento da técnica". A Tabela 1 sumariza os trabalhos encontrados.

A Seção 4.1 descreve os trabalhos que apresentam aplicações de técnicas de agrupamento; a Seção 4.2 descreve os trabalhos que propõem novas técnicas de agrupamento e a Seção 4.3 apresenta algumas características relativas a estes trabalhos; a Seção 4.4 apresenta uma discussão relacionada aos trabalhos estudados neste capítulo.

4.1 Trabalhos que apresentam aplicações das técnicas de agrupamento

Nesta seção são apresentadas algumas aplicações da utilização de agrupamentos hierárquico de documentos em ambientes dinâmicos. Estes são trabalhos que foram encontrados durante a pesquisa, porém, apesar de relacionados com o tema, não têm como foco principal a geração da hierarquia e sim a sua utilização em um contexto específico.

Tabela 1- Sumarização dos trabalhos encontrados

Título	**Ano**	**Tipo**	**Biblioteca**
Dynamic hierarchical algorithms for document clustering (GIL-GARCÍA; PONS-PORRATA, 2010a)	2010	Desenvolvimento da técnica	SD, ACM
Improving the dynamic hierarchical compact clustering algorithm by using feature selection (GIL-GARCÍA; PONS-PORRATA, 2010b)	2010	Desenvolvimento da técnica	ACM
Document update summarization using incremental hierarchical clustering (WANG; LI, 2010)	2010	Aplicação	ACM
An incremental nested partition method for data clustering (CORREA-MORRIS et al., 2010)	2010	Desenvolvimento da técnica	SD, ACM
Incremental document clustering using Multi-representation Indexing Tree (WANG; SONG; LIU, 2010)	2010	Desenvolvimento da técnica	IEEE
News topic detection based on hierarchical clustering and named entity (HUANG et al., 2011)	2011	Desenvolvimento da técnica	IEEE
Tracking and Connecting Topics via Incremental Hierarchical Dirichlet Processes (GAO et al., 2011)	2011	Aplicação	ACM, IEEE
Dynamic categorization of clinical research eligibility criteria by hierarchical clustering (LUO; YETISGEN-YILDIZ; WENG, 2011)	2011	Aplicação	SD
On the use of consensus clustering for incremental learning of topic hierarchies (MARCACINI; HRUSCHKA; REZENDE, 2012)	2012	Desenvolvimento da técnica	ACM
Incrementally Clustering Legislative Interpellation Documents (LIN; HUANG; LIAO, 2012)	2012	Aplicação	ACM, IEEE
HSPKNN: An effective and practical framework for hot topic detection of Internet news (LU et al., 2012)	2012	Aplicação	IEEE
Incremental hierarchical text clustering with privileged information (MARCACINI; REZENDE, 2013)	2013	Desenvolvimento da técnica	ACM
A general framework of hierarchical clustering and its applications (CAI et al., 2014)	2014	Desenvolvimento da técnica	SD
Named entities as privileged information for hierarchical text clustering (SINOARA et al., 2014)	2014	Desenvolvimento da técnica	ACM
ALGOR A novel incremental conceptual hierarchical text clustering method using CFu-tree (PENG; LIU, 2015)	2015	Desenvolvimento da técnica	SD

O trabalho de Wang e Li (2010) utiliza um método de agrupamentos hierárquicos com o foco na sumarização dos documentos, efetuando a atualização dos sumários no mesmo momento em que um novo documento é inserido. É utilizado o algoritmo COBWEB, adaptado para trabalhar com uma hierarquia de sentenças. O algoritmo proposto conseguiu produzir sumários de qualidade superior à outras técnicas utilizadas nos experimentos.

O trabalho de Gao et al. (2011) possui foco diferente do anterior: o relacionamento entre os tópicos. Uma vez agrupados os documentos em tópicos, este trabalho busca monitorar eventos nestes tópicos em um ambiente dinâmico, como a junção de dois tópicos para formar apenas um ou a divisão de um tópico resultando em dois deles. Também foi desenvolvida uma forma de visualização da evolução dos tópicos ao longo do tempo.

No trabalho de Luo, Yetisgen-Yildiz e Weng (2011) é proposto um método para categorização de critérios de elegibilidade de pacientes para pesquisas clínicas. Estes critérios são sentenças que representam características, como faixas etárias. Foi utilizada uma técnica semi-supervisionada, combinando agrupamento supervisionado e agrupamento hierárquico, com o objetivo final de associar os critérios de elegibilidade com dados reais de pacientes, sendo que estes já são estruturados.

No estudo de Lin, Huang e Liao (2012) foi proposta uma aplicação de agrupamento hierárquico incremental para documentos do site da biblioteca do Legislativo Yuan. No método apresentado foram extraídos apenas nomes de entidades (pessoas, locais e organizações) para serem utilizados no agrupamento. O agrupamento utilizado foi dividido em duas etapas, onde primeiramente um algoritmo de agrupamento hierárquico é utilizado para identificar o número ideal de grupos e em seguida o algoritmo *k-means* é utilizado para agrupar de fato os documentos. Foram realizadas comparações

com categorizações efetuadas manualmente por especialistas, obtendo uma taxa de acerto satisfatória.

No trabalho de Lu et al. (2012) foi desenvolvido um framework para detecção de tópicos em destaque (hot topics) em notícias da internet. Neste framework, um algoritmo de agrupamento incremental é utilizado com o intuito de identificar os tópicos candidatos a partir de notícias de uma janela de tempo. Em seguida um algoritmo baseado nos algoritmos KNN e Single-Pass é utilizado para selecionar os tópicos dentre os candidatos. Em um último passo, é apresentado também um algoritmo para geração de descritores para os tópicos identificados. O algoritmo obteve bom desempenho em comparação a outros.

4.2 Trabalhos que apresentam novas técnicas de agrupamento

No trabalho de Gil-García e Pons-Porrata (2010a) é apresentado um framework para agrupamento hierárquico incremental de documentos, onde é utilizado um grafo para armazenar cada nível da hierarquia. A partir deste framework são apresentados dois algoritmos: Dynamic Hierarchical Compact (DHC) e Dynamic Hierarchical Star (DHS). O DHC constrói agrupamentos disjuntos enquanto o DHS permite sobreposições entre os agrupamentos. Os resultados apresentados mostram superioridade de performance dos métodos propostos em relação aos algoritmos UPGMA e BKM e as novas técnicas têm como principal característica a geração de hierarquias de fácil navegação, pois procuram equilibrar sua largura com sua profundidade.

Evoluindo este último trabalho, Gil-García e Pons-Porrata (2010b) apresenta uma alteração no algoritmo DHC com o intuito de melhorar seu custo computacional. Visto que as atualizações nas estruturas utilizadas no algoritmo requerem diversos cálculos de similaridade entre os agrupamentos, é adicionado no processo um passo para a seleção local de termos mais relevantes para os

grupos, o que resulta na redução da dimensionalidade destes dados. Desta forma, obteve-se considerável redução do custo computacional do algoritmo DHC com pequena perda de qualidade da hierarquia gerada. É apresentada também uma nova métrica para avaliação da qualidade da hierarquia gerada, que busca medir o custo de navegação para encontrar o tópico desejado, auxiliando os experimentos para geração de hierarquias com boa navegabilidade.

No trabalho de Correa-Morris et al. (2010) é proposto um algoritmo para o cálculo de partições aninhadas chamado INPM. Este método utiliza diferentes critérios de agrupamento em cada nível da hierarquia, estruturando seus documentos em grafos de forma a poder utilizar as propriedades matemáticas desta estrutura. O algoritmo proposto não é sensível à ordem de inserção dos documentos e é possível obter os níveis da hierarquia de forma independente. Outra propriedade interessante é a possibilidade de calcular o número de partições em um nível sem a necessidade de execução de todo o procedimento.

Wang, Song e Liu (2010) apresenta um algoritmo baseado em árvores de indexação dinâmicas, chamado Multi-Representation Indexing Tree (MRIT), que é uma árvore utilizada para representar a hierarquia de documentos. A técnica apresentada tem como ponto positivo o fato de não ser sensível à ordem de inserção dos documentos e evita a descentralização do centro do agrupamento.

Huang et al. (2011) propõe uma estratégia para a extração de tópicos em notícias que se divide em duas etapas: o agrupamento retrospectivo, que utiliza um algoritmo de agrupamento hierárquico aglomerativo para agrupar as notícias já contidas na base de dados (notícias passadas) e o agrupamento online, que utiliza o algoritmo *Single-Pass* para inserir as novas notícias na hierarquia de tópicos. Neste segundo passo, as notícias são separadas em janelas de tempo de 24 horas. Então, elas são agrupadas em microgrupos para depois estes sejam inseridos na hierarquia, em grupos já existentes ou novos grupos. Este trabalho

utilizou a extração de entidades nomeadas para auxiliar nas tarefas de agrupamento, tendo esses termos um peso maior. Em seus experimentos, a estratégia utilizando entidades nomeadas mostrou-se mais efetiva quando comparada com a estratégia de considerar palavras do título com maior peso.

No trabalho de Marcacini, Hruschka e Rezende (2012) é apresentada uma abordagem do algoritmo *Buckshot* utilizando *consensus clustering*. A geração do modelo inicial apresenta-se como uma fase crítica de algoritmos de agrupamento incremental, visto que erros ocorridos nesta fase podem ser propagados ao longo processo de agrupamento dos demais documentos. Na busca por solucionar este problema, neste trabalho é apresentado o algoritmo Buckshot Consensus Clustering. Este utiliza um primeiro subconjunto dos dados para gerar diferentes soluções de agrupamento, para então integrá-las em uma única solução, gerando assim um modelo inicial mais robusto, menos sensível à degradação. Este método apresentou bons resultados quando comparado com algoritmo sem *consensus clustering* e também com algoritmos não incrementais.

No trabalho de Marcacini e Rezende (2013) é apresentado um método para utilizar informações privilegiadas para favorecer o agrupamento dos documentos. Informações privilegiadas são informações adicionais específicas do domínio que estão contidas em apenas um subconjunto inicial dos documentos. Normalmente tais informações são fornecidas por especialistas. A técnica apresentada, chamada LIHC (*LUPI-based Incremental Hierarchical Clustering*) utiliza o subconjunto que possui estas informações privilegiadas e aplica vários algoritmos de agrupamento a ele, utilizando *consensus clustering* para gerar um modelo inicial de particionamento dos dados. A partir desse modelo, os demais documentos são inseridos incrementalmente nos grupos, atualizando a estrutura. O método apresentou melhora na qualidade das hierarquias geradas em comparação ao algoritmo *Buckshot*.

Cai et al. (2014) apresenta um framework de agrupamento hierárquico onde seu foco principal é obter níveis da hierarquia que são significativos. Este framework é capaz de trabalhar com qualquer medida de proximidade que respeite algumas restrições, detalhadas no trabalho, como a inequalidade triangular com um fator de relaxamento. Este também foi adaptado para trabalhar em contextos dinâmicos. Nos experimentos conduzidos, o framework foi combinado com os algoritmos k-means e k-medians apresentando bons resultados quando comparado a outros algoritmos.

Utilizando o algoritmo LIHC (MARCACINI; REZENDE, 2013), no trabalho de Sinoara et al. (2014) é proposto o uso de entidades nomeadas contidas no texto como informações privilegiadas. Assim, para os conjuntos de dados utilizados, foram utilizadas ferramentas para a extração de entidades de uma parte dos dados e estas foram utilizadas para geração do modelo inicial de agrupamentos. Os resultados obtidos não demonstraram melhora significativa com o uso de entidades nomeadas em todos os testes realizados. Porém, o uso destas informações contribuiu para a geração de descritores de grupos mais inteligíveis. Constatou-se também que a qualidade das entidades extraídas podem afetar diretamente a qualidade dos agrupamentos obtidos.

Peng e Liu (2015) apresenta um método aglomerativo para agrupamento hierárquico de documentos chamado ICHTC-CF. Nele é utilizada a árvore CFu-Tree, que representa a hierarquia de agrupamentos e se mostra eficaz no contexto de agrupamento incremental. Uma característica importante deste método é a ausência de um parâmetro que estabeleça o número máximo de clusters (como o parâmetro K no algoritmo *K-means*), visto que a escolha do valor para este parâmetro pode ser um obstáculo. Assim, é utilizada a medida *Comparison Variation* (CV) para determinar se dois grupos devem ser unidos ou não durante o processo, o que garante a eficácia do método. Esta técnica demonstrou resultados superiores a outros algoritmos (*K-means*, CLIQUE,

Single Linkage, *Complete Linkage*) e um aumento de eficácia com relação ao aumento de termos (*features*).

4.3 Características dos trabalhos

Durante o estudo dos trabalhos apresentados neste capítulo, os trabalhos que apresentam novas técnicas de agrupamento foram observados de forma mais detalhada e as seguintes informações foram coletadas:

a) Medidas de similaridade e distância utilizadas;
b) Bases de dados utilizadas nos experimentos;
c) Métricas utilizadas para a avaliação dos algoritmos;
d) Algoritmos utilizados em comparações.

Tais características foram observadas com o intuito de comparar estes trabalhos e são apresentadas nesta seção. A primeira delas são as medidas de similaridade utilizadas pelos algoritmos propostos em cada um dos trabalhos. A Tabela 2 lista as medidas utilizadas por cada trabalho, onde é possível observar uma predominância do uso da similaridade de cosseno.

Tabela 2 - Medidas de similaridade utilizadas nos trabalhos.

Measure	**Type**	**Trabalhos que Utilizazam**
Similaridade de Cosseno (HUANG, 2008; XU; WUNSCH, 2005)	Similarity	Gil-García e Pons-Porrata (2010a, 2010b), Huang et al. (2011), Marcacini, Hruschka e Rezende (2012), Sinoara et al. (2014) e Wang e Li (2010)
Distância Euclidiana (HUANG, 2008; XU; WUNSCH, 2005)	Distance	Cai et al. (2014) e Peng e Liu (2015)
Soma ponderada dos *δ-kernels* (SCHOLKOPF; SMOLA, 2001; SHAWE-TAYLOR; CRISTIANINI, 2004)	Similarity	Correa-Morris et al. (2010)
***Polynomial kernel of degree* 1** (SCHOLKOPF; SMOLA, 2001; SHAWE-TAYLOR; CRISTIANINI, 2004)	Similarity	Correa-Morris et al. (2010)

As bases de dados utilizadas são apresentadas na Tabela 3. Esta Tabela apresenta apenas as bases de dados utilizadas em pelo menos dois trabalhos e os valores apresentados foram obtidos a partir dos trabalhos estudados. É importante destacar que alguns trabalhos utilizam apenas um subconjunto das bases de dados.

As principais métricas utilizadas nos trabalhos encontrados são listadas na Tabela 4. Algumas métricas utilizam os valores de *Precision* e *Recall*, sendo então estas utilizadas, indiretamente, por diversos trabalhos. Nesta Tabela, *Precision* e *Recall* foram inseridas de forma separada para destacar os trabalhos que utilizaram os valores destas métricas separadamente em suas comparações.

Tabela 3 - Bases de dados utilizadas em ao menos dois trabalhos

Bases de Dados	**Fonte**	**Num. Doc.**	**Num. Termos**	**Num. Classes**	**Trabalhos que Utilizaram**
Afp	TREC-5	695	12575	25	Correa-Morris et al. (2010) e Gil-García e Pons-Porrata (2010a)
Eln	TREC-4	5829	84344	50	Gil-García e Pons-Porrata (2010a, 2010b)
Tdt	TDT2	9824	55112	193	Correa-Morris et al. (2010) e Gil-García e Pons-Porrata (2010a, 2010b)
Reu	Reuters-21578	10369	35297	120	Gil-García e Pons-Porrata (2010a,2010b)
Ohscal	Ohsumed	9200	13512	12	Gil-García e Pons-Porrata (2010a, 2010b)
Reviews	San Jose Mercury	4069	22927	5	Gil-García e Pons-Porrata (2010a) e Marcacini, Hruschka e Rezende (2012)
Hitech	San Jose Mercury	2301	12942	6	Gil-García e Pons-Porrata (2010a, 2010b) e Marcacini, Hruschka e Rezende (2012)
20ng	20ng	18808	45434	20	Marcacini, Hruschka e Rezende (2012), Marcacini e Rezende (2013), Peng e Liu (2015) e Sinoara et al. (2014)
Re8	Reuters-21578	7674	8901	8	Marcacini, Hruschka e Rezende (2012) e Marcacini e Rezende (2013)
Reuters-21578	Reuters-21578	21578	Não Obtido	135	Cai et al. (2014), Peng e Liu (2015) e Sinoara et al. (2014)

Tabela 4 - Métricas utilizadas nos trabalhos.

Métricas	Trabalhos que Utilizaram
Precision (MARCACINI; HRUSCHKA; REZENDE, 2012; RIJSBERGEN, 1979)	Huang et al. (2011)
Recall (MARCACINI; HRUSCHKA; REZENDE, 2012; RIJSBERGEN, 1979)	Huang et al. (2011)
Overall F-measure (LARSEN; AONE, 1999; MARCACINI; HRUSCHKA; REZENDE, 2012; RIJSBERGEN, 1979; ZHAO; KARYPIS; FAYYAD, 2005)	Correa-Morris et al. (2010), Gil-García e Pons-Porrata (2010a, 2010b), Huang et al. (2011), Marcacini, Hruschka e Rezende (2012), Marcacini e Rezende (2013), Peng e Liu (2015) e Sinoara et al. (2014)
FCubed (AMIGÓ et al., 2009; GIL-GARCÍA; PONS-PORRATA, 2010a)	Gil-García e Pons-Porrata (2010a, 2010b)
HF1 (GIL-GARCÍA; PONS-PORRATA, 2010a)	Gil-García e Pons-Porrata (2010a)
Overall F1-Travel (GIL-GARCÍA; PONS-PORRATA, 2010b)	Gil-García e Pons-Porrata (2010b)
CF-Feature (HAN; ZHAO, 2009)	Wang e Li (2010)
Accuracy (WANG; LI, 2010)	Wang e Li (2010)
Relative Cost Ratio (CAI et al., 2014)	Cai et al. (2014)

Outro ponto divergente entre os experimentos realizados nos trabalhos são os outros algoritmos utilizados para comparação e avaliação do método proposto. A Tabela 5 exibe os algoritmos utilizados para comparações nos trabalhos encontrados.

É possível observar que os trabalhos apresentados são bastante divergentes neste aspecto, dificultando uma comparação direta entre eles. Um ponto a ser observado é que nem todos os algoritmos presentes na Tabela 5 são incrementais ou dinâmicos, por exemplo o algoritmo K-means. Apesar de haver métricas como Overall F-measure, que podem ser aplicadas a algoritmos particionais ou hierárquicos, há outras que não podem ser utilizadas para algoritmos particionais (como HF1 e F1-travel), limitando assim as possibilidades de comparações.

Tabela 5 - Algoritmos utilizados em comparações nos experimentos

Algoritmo / Estratégia	Trabalhos que Utilizaram
UPGMA (JAIN; DUBES, 1988)	Gil-García e Pons-Porrata (2010a, 2010b)
Bisecting K-means (KARYPIS; KUMAR; STEINBACH, 2000; FAYYAD, 2005) ZHAO; KARYPIS;	Gil-García e Pons-Porrata (2010a, 2010b), Marcacini, Hruschka e Rezende (2012) e Sinoara et al. (2014)
Single-link (JAIN; DUBES, 1988; SIBSON, 1973)	Correa-Morris et al. (2010) e Peng e Liu (2015)
Complete-link (DEFAYS, 1977; JAIN; DUBES, 1988)	Correa-Morris et al. (2010) e Peng e Liu (2015)
Star (ASLAM; PELEKHOV; RUS, 2004)	Correa-Morris et al. (2010)
Extended Star (GIL-GARCÍA; BADÍA-CONTELLES; PONS-PORRATA, 2003)	Correa-Morris et al. (2010)
ACONS (ALONSO; SUÁREZ; PAGOLA, 2007)	Correa-Morris et al. (2010)
Incremental clustering using Indexing Trees (ZHANG; ZI; WU, 2007)	Wang, Song e Liu (2010)
Title words based clustering (DAI et al., 2010)	Huang et al. (2011)
Buckshot (CUTTING et al., 1992)	Marcacini, Hruschka e Rezende (2012) e Marcacini e Rezende (2013)
Average Random Clustering (MARCACINI; HRUSCHKA; REZENDE, 2012)	Marcacini, Hruschka e Rezende (2012)
GFIO (K-Medians e K-means) (LIN et al., 2006)	Cai et al. (2014)
Small-Space (GUHA et al., 2003)	Cai et al. (2014)
CLIQUE (AGRAWAL et al., 1998)	Peng e Liu (2015)
K-means (MACQUEEN et al., 1967)	Peng e Liu (2015)

4.4 Discussão

Analisando as características dos trabalhos aqui estudados, é possível observar divergências em diversos aspectos. Nos trabalhos que consistem em aplicação da técnica, há uma variedades de objetivos finais para a utilização de agrupamentos. Há trabalhos que buscam a sumarização de documentos, organização destes, identificação de relacionamento entre tópicos ou identificação de hot topics.

Os trabalhos que buscam o desenvolvimento de algoritmos utilizam diferentes estratégias com diferentes focos. Em geral, os trabalhos buscam gerar hierarquias de qualidade, porém, alguns ainda propõem gerar hierarquias de fácil navegação, ou gerar apenas os níveis significativos da hierarquia, ou ainda um agrupamento não sensível a ordem de leitura dos dados. Com relação às estratégias adotadas, alguns focam na seleção de features, ou no uso de diferentes critérios para os níveis da hierarquia, ou ainda o uso de informações privilegiadas e de entidades nomeadas.

Com relação às medidas de similaridade e distância utilizadas, como descrito anteriormente, a escolha de qual medida utilizar pode não ser trivial e ser relativa ao contexto. Porém, há uma predominância do uso da Similaridade de Cosseno, visto que suas propriedades são favoráveis ao agrupamentos de textos.

As bases de dados utilizadas nos trabalhos também são bastante divergentes. A mais utilizada é a Reuters-21578 e seus subconjuntos Reu e Re8. Esta possui um grande volume de dados, o que pode auxiliar no desenvolvimento e testes de algoritmos mais robustos.

Diversas métricas de avaliação de agrupamentos são utilizadas pelos trabalhos, havendo um destaque para a Overall F-measure, que foi utilizada pela maioria deles. Esta métrica pode ser utilizada para algoritmos de agrupamento hierárquico ou particional, o que facilita a comparação entre algoritmos. Destacam-se também as métricas HF1 e Overall F1-Travel, que foram apresentadas por dois dos trabalhos estudados aqui, e focam na avaliação de hierarquias.

Observando os algoritmos utilizados em comparações pelos trabalhos, é possível verificar que nenhum destes trabalhos apresenta comparações com os demais, apenas com trabalhos anteriores ao período dos últimos cinco anos.

Um ponto importante a ser observado é que apenas dois dos trabalhos que buscam desenvolver novas técnicas de agrupamento utilizaram entidades nomeadas. No trabalho de Huang et al. (2011), entidades nomeadas foram utilizadas como termos de maior peso, porém foi efetuado apenas o comparativo com a abordagem utilizando termos do títulos com maior peso. Já no trabalho de Sinoara et al. (2014), entidades foram utilizadas como informação privilegiada em uma etapa de geração de um modelo inicial e não foi obtida melhora nos resultados. Isto mostra que o uso de entidades ainda pode ser estudado com maior profundidade a fim de compreender a contribuição que tais informações podem fornecer às técnicas de agrupamento.

5 METODOLOGIA

Para o desenvolvimento deste projeto foi necessária a execução de algumas etapas, desde a compreensão do problema até a proposta de novas abordagens e obtenção de resultados. A Figura 7 apresenta um fluxo das etapas executadas.

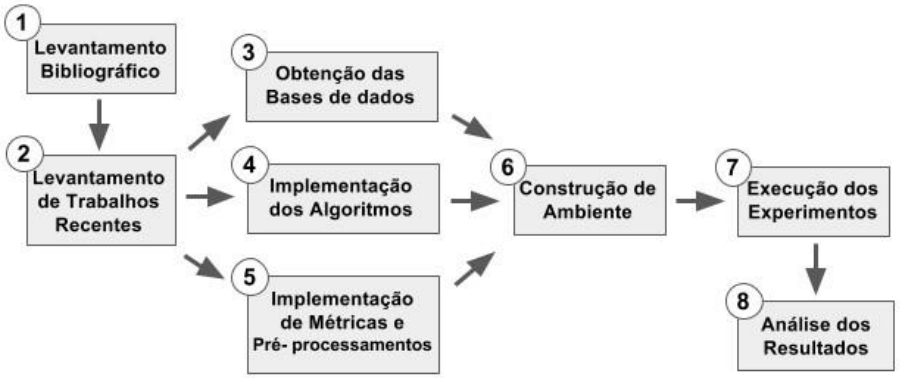


Figura 7 - Fluxo de etapas de execução do projeto

A Etapa 1 consistiu no levantamento bibliográfico com o intuito de compreender o problema de agrupamento hierárquico de documentos em ambientes dinâmicos. Tal estudo visou o entendimento dos principais conceitos de agrupamento, seus desafios e peculiaridades referentes a ambientes dinâmicos. O conhecimento obtido foi compilado e compõe o Capítulo 2 deste trabalho.

Uma vez estudados os conceitos relacionados ao tema, a Etapa 2 consistiu em uma busca pelos trabalhos recentes cujo foco é o problema estudado neste projeto. Para tal, foi efetuada uma pesquisa sistemática, utilizando critérios bem definidos de busca, aplicados à diversas bibliotecas digitais. Desta forma, foi possível analisar os trabalhos encontrados e selecionar

aqueles que são de interesse para este projeto. O processo de obtenção destes trabalhos é detalhado no Apêndice A.

Utilizando o resultado da etapa anterior, na Etapa 3 verificou-se as bases de dados mais utilizadas nos experimentos de outros autores, sendo estas priorizadas na escolha de quais bases utilizar neste trabalho. Mais detalhes sobre as bases de dados selecionadas são apresentados na seção 5.1.

Tendo em mãos os trabalhos recentes que propõem novos métodos para agrupamento hierárquico incremental de documentos, na Etapa 4 foi selecionado um dos algoritmos a ser implementado com o intuito de ser utilizado para validar o uso de entidades nomeadas para o agrupamento de textos. O algoritmo selecionado foi o Dynamic Hierarchical Compact (DHC), proposto por Gil-García e Pons-Porrata (2010a). Também se optou pela implementação de um algoritmo não incremental, neste caso o Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (JAIN; DUBES, 1988), que também foi utilizado por Gil-García e Pons-Porrata (2010a) em seus experimentos.

Para agilizar e evitar erros durante a execução dos experimentos, na Etapa 5, foram implementadas rotinas de pré-processamento dos dados, selecionadas de acordo com aquelas mais comumente utilizadas nos trabalhos estudados. Também foram implementadas as métricas de avaliação dos agrupamentos hierárquicos resultantes dos algoritmos. As métricas selecionadas também foram aquelas mais utilizadas nos trabalhos estudados e que se adequavam a este trabalho.

Uma vez obtidas as bases de dados e implementados os algoritmos e as métricas de avaliação, a Etapa 6 consistiu na construção de um ambiente para execução dos experimentos. Este ambiente tem como objetivo: (1) padronizar o código dos algoritmos já implementados, de forma a facilitar a implementação de novos algoritmos, disponibilizando subrotinas comuns que puderem ser utilizadas em diferentes situações; (2) formatar e padronizar os resultados dos

algoritmos e (3) agilizar a análise de resultados, aplicando as métricas selecionadas e facilitando a visualização de comparativos. Mais detalhes sobre a construção do ambiente para execução dos experimentos são apresentados na Seção 5.2.

A Etapa 7 consistiu na investigação, implementação e avaliação experimental de novas abordagens para o problema utilizando entidades nomeadas. Na execução desta atividade foram necessários repetidos ajustes e testes até que fossem alcançados resultados satisfatórios e que se pudesse validar a hipótese a respeito do uso de entidades para agrupamento hierárquico em ambientes dinâmicos.

Utilizando os resultados obtidos nos experimentos, na Etapa 8 foi efetuada uma compilação e interpretação dos resultados obtidos na Etapa 7. Tais resultados são apresentados em detalhes na Seção 6.

As seções seguintes deste capítulo apresentam as bases de dados utilizadas nos experimentos (Seção 5.1), o ambiente desenvolvido para a execução dos experimentos (Seção 5.2) e, por fim, as tecnologias utilizadas (Seção 5.3).

5.1 Bases de dados

No capítulo 4 são identificadas as bases de dados mais utilizadas nos trabalhos recentes relacionados ao tema. Apesar de haver diversas bases de dados, muitas delas já são disponibilizadas pré-processadas, o que dificulta o processo de extração de entidades, impossibilitando os experimentos aqui propostos. Assim, optou-se por utilizar bases de dados que estivessem inseridas no conjunto das mais utilizadas e que puderam ser obtidas em seu texto original. As bases selecionadas foram Reuters-21578 (LEWIS, 2004), Ohsumed (HERSH et al., 1994) e 20 newsgroups (RENNIE, 2008).

A base de dados Reuters-21578 (LEWIS, 2004) é composta por 21578 notícias do ano de 1987 da agência Reuters. Neste trabalho foram utilizadas apenas notícias que foram manualmente classificadas, para possibilitar a avaliação dos resultados, e foi utilizado no processo de agrupamento apenas o texto do corpo da notícia. Assim, notícias que possuíam o seu corpo vazio também foram excluídas. O conteúdo desta base de dados é disponibilizado em documentos no formato XML, sendo então consideradas apenas notícias com o atributo "TOPICS=YES" e com uma parte <BODY>.

Outra base de dados bastante selecionada é a 20 newsgroups ou 20ng (RENNIE, 2008), que é composta por 18808 e-mails enviados em listas de discussão na internet, divididos em 20 categorias. Para esta base de dados foi utilizado no processo de agrupamento apenas o corpo dos emails. Devido ao grande volume da base 20ng, afim de facilitar a execução dos experimentos, foram selecionadas 10 classes e foram utilizados apenas documentos pertencentes à elas.

A terceira base de dados selecionada é a Ohsumed (HERSH et al., 1994), que contém artigos de medicina dos anos de 1987 a 1991. Também com o intuito de facilitar a execução dos experimentos, foram selecionados apenas documentos de 10 classes.

Antes da utilização destas bases de dados, elas foram submetidas a um pré-processamento, passando pelas seguintes rotinas:

a) Extração de entidades;
b) Remoção de caracteres especiais, numerais e caracteres de pontuação;
c) *Stemming*;
d) Remoção de *Stop Words*;
e) Remoção de termos que ocorrem em apenas um documento.

A Tabela 6 apresenta detalhes sobre as bases de dados após efetuados os processamentos. Ao efetuar os pré-processamentos, foram extraídas entidades nomeadas, cujas informações não foram incluídas juntamente aos termos dos documentos, sendo armazenadas separadamente. Assim, as colunas referentes ao número total de termos e de termos distintos, dizem respeito apenas aos termos dos próprios documentos, sem considerar as entidades extraídas.

Optou-se por armazenar as entidades separadamente para que os algoritmos possam utilizá-las de diferentes formas. Informações referentes às entidades extraídas destas bases de dados são apresentadas na Tabela 7.

Tabela 6 - Bases de dados a serem utilizadas.

Base de dados	Documentos	Total de Termos	Termos Distintos	Classes
Reuters-21578	10377	719331	21079	119
Ohsumed	9830	910154	29453	10
20 Newgroups	9909	1205051	85949	10

Tabela 7 - Bases de dados a serem utilizadas.

Base de dados	Total de Entidades	Entidades Distintas	Total de Tipos	Tipos Distintos	Entidades / Documento
Reuters-21578	28214	1995	148299	118	2,71
Ohsumed	4419	998	23820	106	0,44
20 Newgroups	36122	2929	195023	133	3,64

5.2 Ambiente para execução de experimentos

Com o intuito de validar o uso de entidade nomeadas em algoritmos de agrupamentos, foram implementados os algoritmos, diversas estruturas de dados e rotinas que poderão ser utilizadas e estendidas em trabalhos futuros. Tais implementações foram reunidas em uma aplicação, que consiste no ambiente de execução de experimentos. Como apresentado na Figura 8, o ambiente desenvolvido é composto por dois módulos: algoritmos e visualização.

O primeiro módulo, o de algoritmos, consiste um uma aplicação contendo a implementação dos algoritmos de agrupamento (DHC, UPGMA), rotinas de pré-processamento dos dados (*Stemming*, remoção de *stop words*, extração de entidade, entre outros), métricas de avaliação dos resultados (*Precision*, *Recall*, *F-measure*, *F1-Travel*), mecanismos de leitura de bases de dados em diferentes formatos, medidas de similaridade e estruturas de dados comumente utilizadas em algoritmos de agrupamento. Todas as implementações foram efetuadas utilizando o paradigma Orientado a Objetos e de forma que novas rotinas e algoritmos possam ser adicionados.

Uma vez acionado com os devidos parâmetros, o módulo de algoritmos imprime os resultados em diversos arquivos, contendo as configurações dos experimentos, as hierarquias geradas e resultados das métricas de avaliação. A análise e comparação manual de tais resultados exigiria grande esforço e seria passível de erros.

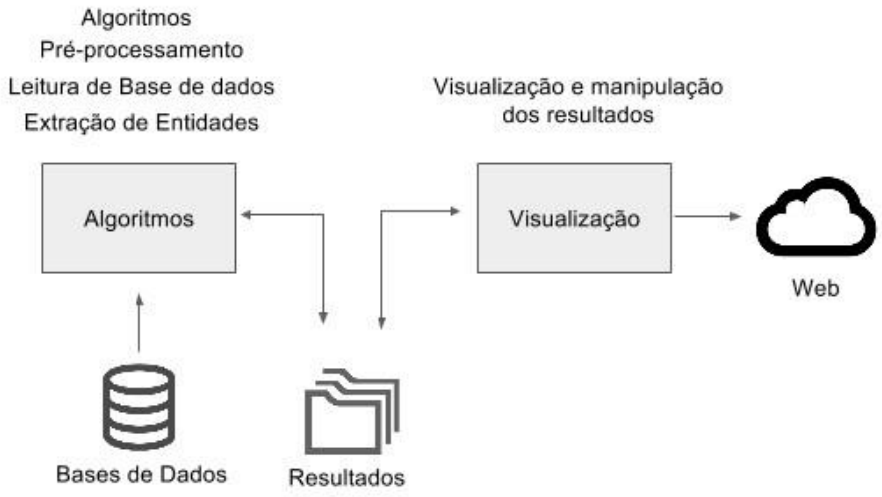


Figura 8 - Ambiente desenvolvido para execução dos experimentos

Para facilitar este trabalho, foi desenvolvido o módulo de visualização. Este módulo é capaz de carregar os resultados dos arquivos e apresentar em uma interface web tabelas comparativas, fornecendo também a visualização da hierarquia de documentos e também manipulações simples sobre os resultados.

5.3 Tecnologias utilizadas

A linguagem de programação utilizada para o desenvolvimento dos algoritmos foi a linguagem JavaSE 1.7. Esta foi a opção escolhida devido a familiaridade da equipe do projeto com esta tecnologia e também a grande disponibilidade de documentação e bibliotecas. Os experimentos foram executados em um computador com processador Intel Core i7, 2.4 GHz com memória RAM de 16 GB, utilizando o sistema operacional Ubuntu 14.04.

6 AVALIAÇÃO EXPERIMENTAL E RESULTADOS

Com o intuito de validar o uso de entidades nomeadas para o agrupamento hierárquico de documentos, foram realizados diversos testes utilizando as quatro abordagens apresentadas no Capítulo 3 com diferentes configurações. Para a abordagem 1 (múltiplos vetores), considerando os vetores de termos (T), nomes de entidades (EN), tipos de entidades (TE) e combinações entre nomes e tipos de entidades (ENT), foram utilizadas 21 combinações de pesos para estes vetores, apresentadas na Tabela 8.

Tabela 8 - Configurações de pesos dos vetores utilizadas na abordagem 1 do uso de entidades.

Num.	T	EN	ET	ENT
1	0.8	0.0	0.1	0.1
2	0.8	0.1	0.1	0.0
3	0.8	0.1	0.0	0.1
4	0.7	0.1	0.1	0.1
5	0.6	0.2	0.2	0.0
6	0.6	0.2	0.0	0.2
7	0.6	0.0	0.2	0.2
8	0.5	0.0	0.5	0.0
9	0.5	0.0	0.0	0.5
10	0.5	0.5	0.0	0.0
11	0.4	0.4	0.2	0.0
12	0.4	0.4	0.0	0.2
13	0.4	0.2	0.4	0.0
14	0.4	0.2	0.0	0.4
15	0.4	0.0	0.4	0.2
16	0.4	0.2	0.2	0.2
17	0.4	0.0	0.2	0.4
18	0.33	0.33	0.0	0.33
19	0.33	0.33	0.33	0.0
20	0.33	0.0	0.33	0.33
21	0.25	0.25	0.25	0.25

Para a abordagem 2 (entidades como termos de maior peso) foram utilizados os pesos $W = 1,2,...,10$. Para as abordagens 3 (entidades como termos com peso em função do vetor de termos) e 4 (entidades como termos com peso

em função do vetor de termos e número de entidades) foram utilizados os pesos $W = 0.1,0.2,...,2.0$.

Considerando tais configurações a Tabela 9 apresenta o número de testes feitos para cada um dos algoritmos (DHC e UPGMA) utilizando cada uma das abordagens de uso de entidades, para as bases de dados Reuters-21578, Ohsumed e 20 newsgroups.

Tabela 9 - Contabilização dos testes efetuados para cada um dos algoritmos

	Base de Dados	**Configurações**	**Execuções**
Sem uso de entidades	3	1	3
Abordagem 1	3	21	63
Abordagem 2	3	10	30
Abordagem 3	3	20	60
Abordagem 4	3	20	60
Total			216

6.1 Avaliação da Abordagem 1

Esta seção apresenta os testes realizados utilizando a Abordagem 1 (múltiplos vetores) para uso de entidades. São exibidos aqui os resultados obtidos para as diversas configurações para as bases de dados Reuters-21578 (Seção 6.1.1), Ohsumed (Seção 6.1.2) e 20 newsgroups (Seção 6.1.3). As tabelas apresentadas estão ordenadas pelo valor da métrica *F-measure* (F1), do maior para o menor. Desta forma, os melhores resultados estão localizados no topo da tabela. Em cada tabela foi adicionado um registro com as configurações vazias, que representa a execução do algoritmo sem o uso de entidades.

6.1.1 Testes executados para a base de dados Reuters-21578

A Tabela 10 apresenta os resultados dos testes efetuados para o algoritmo DHC para a base de dados Reuters-21578. Pode-se observar que apenas duas configurações para esta abordagem geraram hierarquias com maior *F-measure* que a gerada sem o uso de entidades, e apenas uma configuração

alcançou melhor valor para a métrica *F1-travel*. O tempo de execução da abordagem sem o uso de entidades chega a ser quase três vezes menor do que os tempos utilizando a abordagem 1.

Tabela 10 - Testes do algoritmo DHC executados para a base de dados Reuters-21578 utilizando a Abordagem 1 de uso de entidades nomeadas

Configurações						**Resultados**				**Tempo**
Num.	**T**	**EN**	**ET**	**ETN**	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
8	0.5	0.0	0.5	0.0	5	2003	6.1776	0.6732	0.5101	94:31
4	0.7	0.1	0.1	0.1	5	2166	5.7882	0.6665	0.5257	109:48
					5	2162	5.7970	0.6629	0.5214	38:42
1	0.8	0.0	0.1	0.1	5	2154	5.8148	0.6600	0.5188	99:10
2	0.8	0.1	0.1	0.0	5	2155	5.8126	0.6595	0.5191	97:07
3	0.8	0.1	0.0	0.1	5	2204	5.7057	0.6591	0.5150	109:16
11	0.4	0.4	0.2	0.0	5	2153	5.8171	0.6453	0.4912	94:50
15	0.4	0.0	0.4	0.2	5	2056	6.0442	0.6424	0.5006	93:54
5	0.6	0.2	0.2	0.0	5	2108	5.9199	0.6257	0.4856	94:48
16	0.4	0.2	0.2	0.2	5	2154	5.8148	0.6236	0.4727	94:19
17	0.4	0.0	0.2	0.4	5	2154	5.8148	0.6193	0.4761	96:08
21	0.25	0.25	0.25	0.25	5	2143	5.8396	0.6189	0.4847	100:23
9	0.5	0.0	0.0	0.5	5	2214	5.6844	0.6059	0.4714	97:24
6	0.6	0.2	0.0	0.2	5	2225	5.6613	0.6050	0.4627	98:03
7	0.6	0.0	0.2	0.2	5	2109	5.9175	0.5948	0.4610	95:49
12	0.4	0.4	0.0	0.2	5	2190	5.7357	0.5854	0.4526	96:15
13	0.4	0.2	0.4	0.0	4	2052	6.0541	0.5678	0.4493	99:37
14	0.4	0.2	0.0	0.4	5	2189	5.7379	0.5671	0.4364	97:02
18	0.33	0.33	0.0	0.33	5	2194	5.7271	0.5601	0.4381	94:13
10	0.5	0.5	0.0	0.0	5	2218	5.6760	0.5505	0.3866	101:19
19	0.33	0.33	0.33	0.0	4	2100	5.9386	0.5355	0.4236	99:13
20	0.33	0.0	0.33	0.33	4	2099	5.9410	0.5283	0.4182	96:41

Já a Tabela 11 apresenta os resultados dos testes efetuados utilizando o algoritmo UPGMA para a base de dados Reuters-21578. Para este algoritmo apenas uma configuração da abordagem 1 superou os resultados sem a utilização de entidades. O tempo de execução neste caso é bem similar em todas as execuções.

Tabela 11 - Testes do algoritmo UPGMA executados para a base de dados Reuters-21578 utilizando a Abordagem 1 de uso de entidades nomeadas

Configurações					Resultados					Tempo
Num.	**T**	**EN**	**ET**	**ETN**	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
1	0.8	0.0	0.1	0.1	106	10375	2.0000	0.8166	0.2697	99:25
					120	10375	2.0000	0.8128	0.3211	98:00
2	0.8	0.1	0.1	0.0	102	10375	2.0000	0.8116	0.3734	93:36
3	0.8	0.1	0.0	0.1	100	10375	2.0000	0.7890	0.3931	98:56
4	0.7	0.1	0.1	0.1	103	10375	2.0000	0.7793	0.2654	96:23
6	0.6	0.2	0.0	0.2	105	10375	2.0000	0.7392	0.2761	95:45
7	0.6	0.0	0.2	0.2	104	10375	2.0000	0.7391	0.2692	99:15
5	0.6	0.2	0.2	0.0	104	10375	2.0000	0.7273	0.3666	93:09
9	0.5	0.0	0.0	0.5	114	10375	2.0000	0.7154	0.2728	93:43
10	0.5	0.5	0.0	0.0	112	10375	2.0000	0.7055	0.3291	100:28
14	0.4	0.2	0.0	0.4	117	10375	2.0000	0.7019	0.2746	95:39
18	0.33	0.33	0.0	0.33	111	10375	2.0000	0.7014	0.2765	100:16
8	0.5	0.0	0.5	0.0	113	10375	2.0000	0.6964	0.3741	98:08
12	0.4	0.4	0.0	0.2	113	10375	2.0000	0.6960	0.2691	95:57
13	0.4	0.2	0.4	0.0	102	10375	2.0000	0.6866	0.2703	93:07
15	0.4	0.0	0.4	0.2	100	10375	2.0000	0.6768	0.2655	93:08
16	0.4	0.2	0.2	0.2	103	10375	2.0000	0.6762	0.2689	93:10
11	0.4	0.4	0.2	0.0	104	10375	2.0000	0.6732	0.3619	94:00
17	0.4	0.0	0.2	0.4	107	10375	2.0000	0.6730	0.2688	98:52
20	0.33	0.0	0.33	0.33	101	10375	2.0000	0.6546	0.2734	95:14
19	0.33	0.33	0.33	0.0	104	10375	2.0000	0.6525	0.2691	94:05
21	0.25	0.25	0.25	0.25	107	10375	2.0000	0.6190	0.2708	93:50

6.1.2 Testes executados para a base de dados Ohsumed

A Tabela 12 mostra os resultados das execuções do algoritmo DHC para a base Ohsumed. Observa-se que nenhuma das configurações da abordagem 1 contribuiu para a melhora dos valores de *F-measure* e *F1-travel*. O tempo de execução do algoritmo chega a ser quatro vezes maior quando utiliza a abordagem 1 para esta bases de dados, em comparação ao não uso das entidades.

Tabela 12 - Testes do algoritmo DHC executados para a base de dados Ohsumed utilizando a Abordagem 1 de uso de entidades nomeadas

Configurações					Resultados					Tempo
Num.	**T**	**EN**	**ET**	**ETN**	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
					5	2303	5.2604	0.3270	0.3126	19:56
4	0.7	0.1	0.1	0.1	4	2271	5.3261	0.3094	0.3095	84:43
2	0.8	0.1	0.1	0.0	4	2295	5.2809	0.3039	0.3052	90:52
1	0.8	0.0	0.1	0.1	4	2294	5.2828	0.3039	0.3052	89:53
21	0.25	0.25	0.25	0.25	5	2272	5.3242	0.2846	0.2804	86:12
13	0.4	0.2	0.4	0.0	5	2212	5.4415	0.2809	0.2782	81:43
20	0.33	0.0	0.33	0.33	4	2245	5.3762	0.2805	0.2863	81:08
18	0.33	0.33	0.0	0.33	5	2309	5.2550	0.2790	0.2684	86:13
3	0.8	0.1	0.0	0.1	4	2317	5.2403	0.2775	0.2725	81:08
19	0.33	0.33	0.33	0.0	4	2243	5.3801	0.2766	0.2851	81:04
9	0.5	0.0	0.0	0.5	4	2300	5.2716	0.2679	0.2692	86:46
5	0.6	0.2	0.2	0.0	5	2250	5.3665	0.2671	0.2662	81:44
7	0.6	0.0	0.2	0.2	5	2249	5.3684	0.2663	0.2659	81:55
15	0.4	0.0	0.4	0.2	4	2205	5.4556	0.2656	0.2775	80:41
11	0.4	0.4	0.2	0.0	5	2266	5.3357	0.2634	0.2594	80:52
16	0.4	0.2	0.2	0.2	5	2266	5.3357	0.2633	0.2592	83:25
17	0.4	0.0	0.2	0.4	4	2262	5.3433	0.2590	0.2643	80:21
6	0.6	0.2	0.0	0.2	4	2316	5.2421	0.2589	0.2628	82:37
12	0.4	0.4	0.0	0.2	4	2298	5.2753	0.2547	0.2535	80:15
14	0.4	0.2	0.0	0.4	4	2295	5.2809	0.2539	0.2525	81:54
8	0.5	0.0	0.5	0.0	4	2167	5.5337	0.2478	0.2567	84:16
10	0.5	0.5	0.0	0.0	4	2299	5.2735	0.2458	0.2390	83:14

Para esta base de dados o algoritmo UPGMA obteve melhora no valor de *F-measure* em três configurações dentre as testadas para a abordagem 1, sendo que o valor de *F1-travel* é bastante similar em todas as execuções. Também houve pouca variação entre os testes realizados com relação ao tempo gasto. Os resultados são apresentados na Tabela 13.

Tabela 13 - Testes do algoritmo UPGMA executados para a base de dados Ohsumed utilizando a Abordagem 1 de uso de entidades nomeadas

Configurações					Resultados					Tempo
Num.	**T**	**EN**	**ET**	**ETN**	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
2	0.8	0.1	0.1	0.0	185	9828	2.0000	0.3535	0.2146	86:31
1	0.8	0.0	0.1	0.1	183	9828	2.0000	0.3528	0.2147	88:15
3	0.8	0.1	0.0	0.1	199	9828	2.0000	0.3482	0.2148	85:55
					190	9828	2.0000	0.3461	0.2148	76:19
4	0.7	0.1	0.1	0.1	183	9828	2.0000	0.3439	0.2148	76:31
21	0.25	0.25	0.25	0.25	185	9828	2.0000	0.3309	0.2147	75:22
8	0.5	0.0	0.5	0.0	182	9828	2.0000	0.3258	0.2146	82:08
13	0.4	0.2	0.4	0.0	179	9828	2.0000	0.3227	0.2146	82:10
14	0.4	0.2	0.0	0.4	202	9828	2.0000	0.3194	0.2147	83:23
6	0.6	0.2	0.0	0.2	200	9828	2.0000	0.3166	0.2158	83:57
12	0.4	0.4	0.0	0.2	204	9828	2.0000	0.3163	0.2147	83:18
15	0.4	0.0	0.4	0.2	178	9828	2.0000	0.3137	0.2146	83:41
18	0.33	0.33	0.0	0.33	201	9828	2.0000	0.3134	0.2150	82:55
20	0.33	0.0	0.33	0.33	179	9828	2.0000	0.3128	0.2146	82:28
19	0.33	0.33	0.33	0.0	183	9828	2.0000	0.3117	0.2146	82:23
9	0.5	0.0	0.0	0.5	199	9828	2.0000	0.3114	0.2149	83:53
11	0.4	0.4	0.2	0.0	180	9828	2.0000	0.3113	0.2150	83:32
16	0.4	0.2	0.2	0.2	181	9828	2.0000	0.3113	0.2149	76:31
5	0.6	0.2	0.2	0.0	172	9828	2.0000	0.3095	0.2148	84:52
17	0.4	0.0	0.2	0.4	177	9828	2.0000	0.3095	0.2150	83:22
7	0.6	0.0	0.2	0.2	174	9828	2.0000	0.3082	0.2148	83:16
10	0.5	0.5	0.0	0.0	207	9828	2.0000	0.3039	0.2154	82:18

6.1.3 Testes executados para a base de dados 20 Newsgroups

Os resultados obtidos em execuções do algoritmo DHC para a base de dados 20 newsgroups são apresentados na Tabela 14. Diversas configurações da abordagem 1 superaram os resultados obtidos sem o uso de entidades nas métricas *F-measure* e *F1-travel*. Não houve diferença significativa entre os tempos de execução das configurações.

Tabela 14 - Testes do algoritmo DHC executados para a base de dados 20 newsgroups utilizando a Abordagem 1 de uso de entidades nomeadas

Configurações					Resultados					Tempo
Num.	**T**	**EN**	**ET**	**ETN**	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
18	0.33	0.33	0.0	0.33	5	2289	5.3266	0.1942	0.1911	82:20
16	0.4	0.2	0.2	0.2	4	2177	5.5491	0.1836	0.1821	91:37
17	0.4	0.0	0.2	0.4	4	2181	5.5408	0.1819	0.1820	82:46
19	0.33	0.33	0.33	0.0	4	2035	5.8664	0.1818	0.1854	83:54
20	0.33	0.0	0.33	0.33	4	2029	5.8808	0.1817	0.1832	82:52
13	0.4	0.2	0.4	0.0	3	1992	5.9714	0.1813	0.1821	89:49
11	0.4	0.4	0.2	0.0	4	2178	5.5470	0.1803	0.1842	83:20
9	0.5	0.0	0.0	0.5	5	2490	4.9775	0.1797	0.1829	86:40
21	0.25	0.25	0.25	0.25	3	2014	5.9171	0.1794	0.1870	84:51
14	0.4	0.2	0.0	0.4	4	2363	5.1912	0.1784	0.1871	83:59
8	0.5	0.0	0.5	0.0	3	1848	6.3586	0.1767	0.1822	85:29
1	0.8	0.0	0.1	0.1	5	2631	4.7644	0.1755	0.1818	84:14
12	0.4	0.4	0.0	0.2	4	2360	5.1965	0.1749	0.1863	89:30
15	0.4	0.0	0.4	0.2	3	1982	5.9965	0.1749	0.1833	85:55
4	0.7	0.1	0.1	0.1	5	2598	4.8122	0.1729	0.1818	89:23
2	0.8	0.1	0.1	0.0	5	2625	4.7730	0.1712	0.1818	86:54
3	0.8	0.1	0.0	0.1	5	2684	4.6901	0.1693	0.1818	85:52
7	0.6	0.0	0.2	0.2	4	2374	5.1718	0.1685	0.1818	83:55
10	0.5	0.5	0.0	0.0	4	2489	4.9791	0.1639	0.1859	90:50
5	0.6	0.2	0.2	0.0	4	2378	5.1648	0.1637	0.1818	85:09
6	0.6	0.2	0.0	0.2	4	2567	4.8583	0.1502	0.1822	86:09
					5	2601	4.6314	0.1491	0.1810	13:18

Já nos testes efetuados com o algoritmo UPGMA, apresentados na Tabela 15, todas as configurações da abordagem 1 superaram os resultados obtidos sem o uso de entidades. Neste, a melhora obtida com o algoritmo UPGMA foi mais significativa do que aquela alcançada com o algoritmo DHC.

Tabela 15 - Testes do algoritmo UPGMA executados para a base de dados 20 newsgroups utilizando a Abordagem 1 de uso de entidades nomeadas

Configurações						Resultados				Tempo
Num.	**T**	**EN**	**ET**	**ETN**	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
18	0.33	0.33	0.0	0.33	101	9907	2.0000	0.2388	0.1820	82:38
12	0.4	0.4	0.0	0.2	100	9907	2.0000	0.2219	0.1820	80:26
14	0.4	0.2	0.0	0.4	107	9907	2.0000	0.2206	0.1820	86:32
21	0.25	0.25	0.25	0.25	215	9907	2.0000	0.1995	0.1819	83:26
10	0.5	0.5	0.0	0.0	106	9907	2.0000	0.1958	0.1819	81:47
8	0.5	0.0	0.5	0.0	306	9907	2.0000	0.1945	0.1819	86:45
9	0.5	0.0	0.0	0.5	106	9907	2.0000	0.1944	0.1819	82:57
13	0.4	0.2	0.4	0.0	287	9907	2.0000	0.1936	0.1819	84:09
17	0.4	0.0	0.2	0.4	207	9907	2.0000	0.1924	0.1820	85:17
15	0.4	0.0	0.4	0.2	285	9907	2.0000	0.1923	0.1820	83:08
20	0.33	0.0	0.33	0.33	262	9907	2.0000	0.1920	0.1820	84:40
16	0.4	0.2	0.2	0.2	201	9907	2.0000	0.1914	0.1820	83:53
19	0.33	0.33	0.33	0.0	243	9907	2.0000	0.1913	0.1820	80:18
11	0.4	0.4	0.2	0.0	211	9907	2.0000	0.1910	0.1820	83:12
7	0.6	0.0	0.2	0.2	210	9907	2.0000	0.1856	0.1819	88:13
5	0.6	0.2	0.2	0.0	212	9907	2.0000	0.1854	0.1819	86:22
6	0.6	0.2	0.0	0.2	114	9907	2.0000	0.1840	0.1820	84:42
4	0.7	0.1	0.1	0.1	126	9907	2.0000	0.1833	0.1819	73:30
1	0.8	0.0	0.1	0.1	131	9907	2.0000	0.1832	0.1819	85:40
2	0.8	0.1	0.1	0.0	134	9907	2.0000	0.1830	0.1819	80:57
3	0.8	0.1	0.0	0.1	99	9907	2.0000	0.1824	0.1819	85:02
					107	9907	2.0000	0.1821	0.1819	83:17

6.1.4 Análise

A Figura 9 apresenta gráficos com os valores de *F-measure* alcançados utilizando a abordagem 1. Para a base de dados Reuters-21578, é possível observar que o uso da abordagem 1 não alcançou grande melhora para os

resultados. Poucas configurações obtiveram melhora nos resultados, e tal diferença foi pouco significativa. Para a base de dados Ohsumed também não foi obtida melhora expressiva, sendo que os resultados foram bem próximos àqueles obtidos sem o uso de entidades, ainda que sendo inferiores. Para a base de dados 20 newsgroups os resultados obtidos com o uso da abordagem 1 mostram uma melhora. Nesta base de dados, quase todas as configurações obtiveram resultados superiores, sendo que três das configurações para o algoritmo UPGMA alcançaram os melhores resultados.

A Figura 10 exibe os valores de *F1-travel* obtidos nos testes da abordagem 1. Observa-se que para nenhuma das bases de dados obteve-se melhora expressiva para os valores desta métrica. Para a base de dados Reuters-21578 houve maior influência do uso de entidades no resultados, porém, para as demais bases de dados tal influência foi mínima.

Nos testes utilizando a abordagem 1 o algoritmo UPGMA obteve melhores valores de *F-measure* que o algoritmo DHC, ou seja, foi mais eficaz no que diz respeito a qualidade dos agrupamentos gerados. Tal resultado é mais perceptível na base de dados Reuters-21578, porém, para as demais bases os resultados foram bem próximo. Já no quesito navegabilidade da hierarquia gerada, calculada através da métrica *F1-travel*, o algoritmo DHC apresentou resultados superiores, também acentuados na base de dados Reuters-21578.

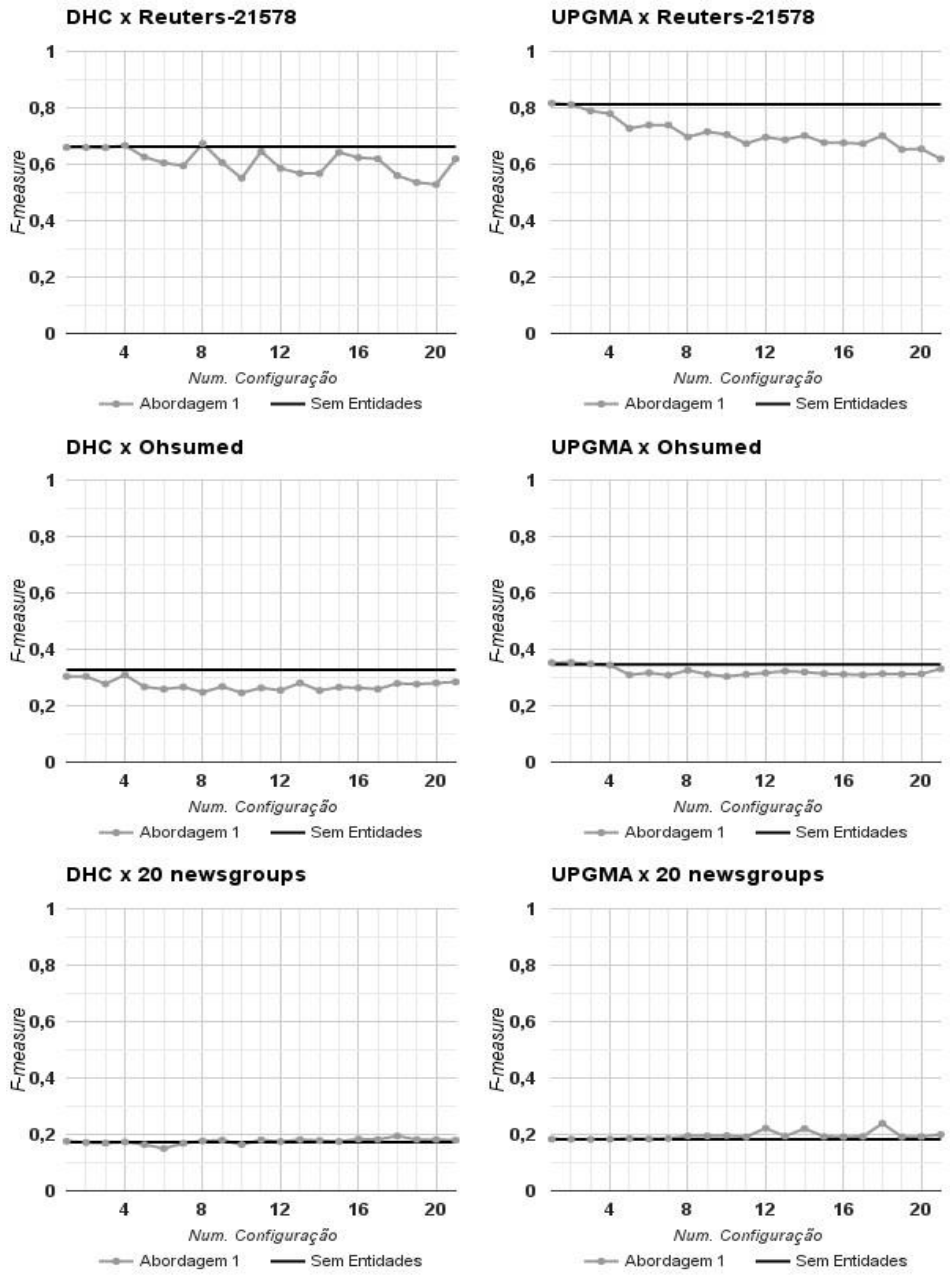


Figura 9 - Valores de *F-measure* alcançados nos testes da abordagem 1

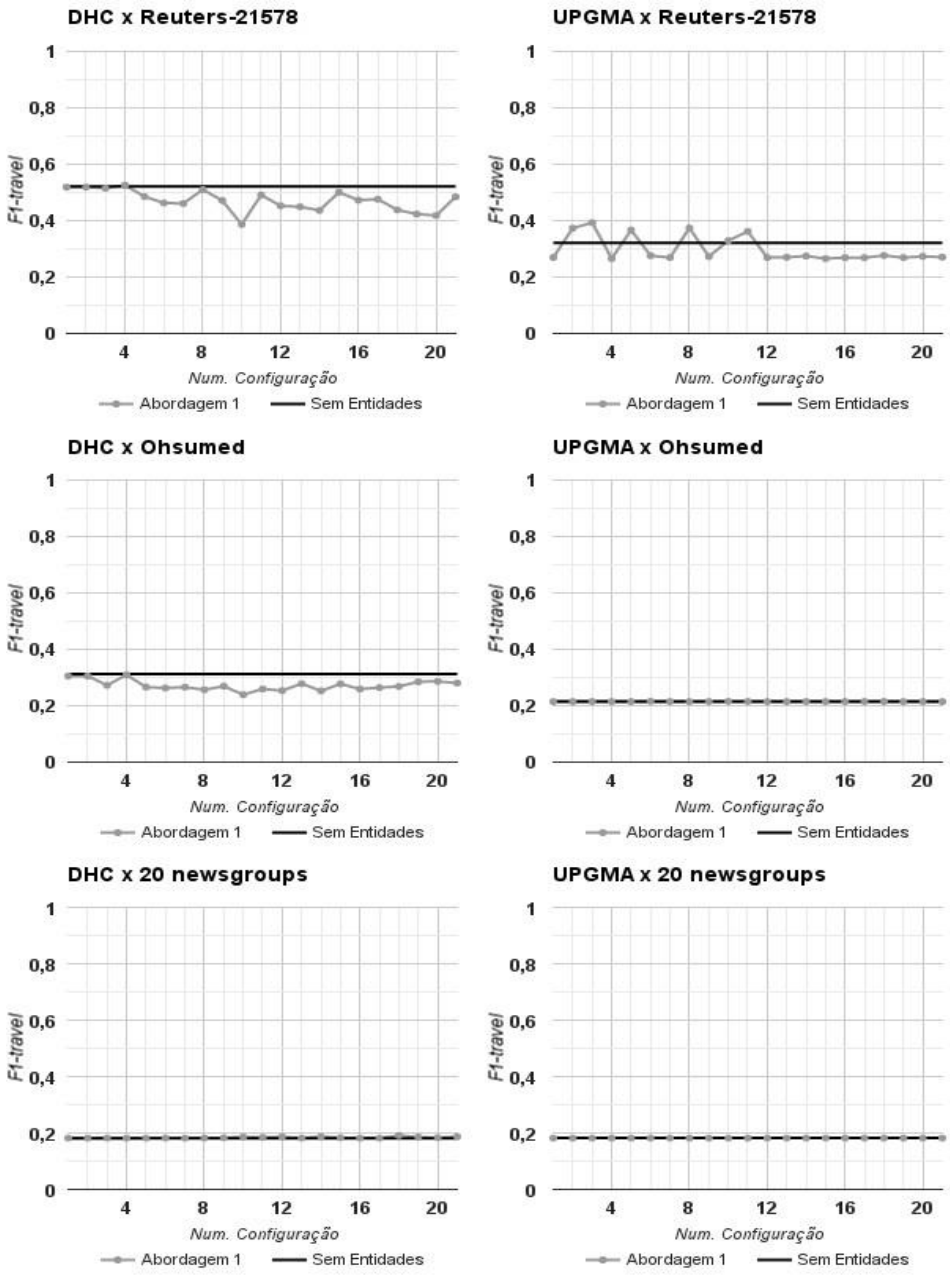


Figura 10 - Valores de *F1-travel* alcançados nos testes da abordagem 1

6.2 Avaliação da Abordagem 2

Nesta seção são apresentados os resultados obtidos utilizando a Abordagem 2 (entidades como termos de maior peso) para uso de entidades. A Seção 6.2.1 mostra os resultados obtidos para a bases de dados Reuters-21578, a Seção 6.2.2 para a base de dados Ohsumed e a Seção 6.2.3 para a base 20 newsgroups. Nas tabelas contidas nesta seção, o registro cuja configuração está vazia representa a execução sem o uso de entidades.

6.2.1 Testes executados para a base de dados Reuters-21578

A Tabela 16 exibe os resultados obtidos na execução do algoritmo DHC para a base de dados Reuters-21578 utilizando a abordagem 2. Pode-se observar que apenas uma configuração desta abordagem superou os resultados obtidos pelo não uso de entidades. Porém, a melhora obtida com esta configuração (W = 1) já é bem mais expressiva que aquelas alcançadas pela abordagem 1, sendo que o tempo de execução se aproxima do tempo gasto sem o uso de entidades.

Tabela 16 - Testes do algoritmo DHC executados para a base de dados Reuters-21578 utilizando a Abordagem 2 de uso de entidades nomeadas

Configurações			**Resultados**			**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
1	5	2172	5.7750	0.7344	0.5575	39:11
	5	2162	5.7970	0.6629	0.5214	38:42
2	6	2189	5.7379	0.6677	0.4919	35:44
3	5	2214	5.6844	0.6394	0.4995	33:11
4	5	2268	5.5729	0.6175	0.4217	31:04
5	5	2269	5.5709	0.5421	0.3894	29:29
9	5	2205	5.7035	0.5302	0.3976	22:52
6	5	2231	5.6487	0.5175	0.3606	27:34
10	5	2213	5.6865	0.5138	0.3527	21:59
8	5	2215	5.6823	0.4980	0.3289	24:21
7	5	2217	5.6781	0.4923	0.3480	25:40

Já para as execuções utilizando o algoritmo UPGMA (TABELA 17) não foi obtida melhora no valor de *F-measure*, apesar de a configuração $W = 1$ se aproximar bastante do resultado sem o uso de entidades. Esta configuração obteve melhora no quesito navegabilidade da hierarquia, representado pela métrica *F1-travel*.

Tabela 17 - Testes do algoritmo UPGMA executados para a base de dados Reuters-21578 utilizando a Abordagem 2 de uso de entidades nomeadas

Configurações			**Resultados**			**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
	120	10375	2.0000	0.8128	0.3211	98:00
1	123	10375	2.0000	0.8097	0.3637	101:50
2	108	10375	2.0000	0.7790	0.3740	101:26
3	121	10375	2.0000	0.7595	0.3842	101:24
4	130	10375	2.0000	0.7304	0.3445	102:25
6	141	10375	2.0000	0.6944	0.2652	101:53
5	134	10375	2.0000	0.6828	0.2735	103:41
7	147	10375	2.0000	0.6595	0.2626	102:21
9	163	10375	2.0000	0.6508	0.2589	103:08
8	161	10375	2.0000	0.6500	0.2583	102:33
10	177	10375	2.0000	0.6327	0.2602	102:37

6.2.2 Testes executados para a base de dados Ohsumed

A Tabela 18 exibe os resultados dos testes efetuados com o algoritmo DHC e a base dados Ohsumed. Observa-se que a abordagem 2 não obteve sucesso nestes testes, visto que nenhuma das configurações superou os resultados sem o uso de entidades nomeadas.

Tabela 18 - Testes do algoritmo DHC executados para a base de dados Ohsumed utilizando a Abordagem 2 de uso de entidades nomeadas

Configurações		**Resultados**				**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
	5	2303	5.2604	0.3270	0.3126	19:56
1	4	2283	5.3034	0.3205	0.3152	19:48
5	4	2341	5.1968	0.3099	0.3085	19:00
2	5	2303	5.2604	0.3069	0.2980	19:32
3	4	2297	5.2772	0.3026	0.3067	19:28
4	4	2318	5.2385	0.2999	0.3003	19:25
10	5	2089	5.2344	0.2788	0.2734	17:42
8	4	2377	5.1333	0.2761	0.2816	18:05
7	4	2372	5.1420	0.2749	0.2781	18:35
6	4	2357	5.1684	0.2670	0.2762	18:39
9	4	2369	5.1473	0.2651	0.2668	17:58

Também para os testes utilizando algoritmo UPGMA (TABELA 19) não foram alcançados melhoras significativas nos resultados. Apenas uma configuração da abordagem 2 superou a execução sem uso de entidades, sendo pequena a melhora nos resultados.

Tabela 19 - Testes do algoritmo UPGMA executados para a base de dados Ohsumed utilizando a Abordagem 2 de uso de entidades nomeadas

Configurações		Resultados				Tempo
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
2	189	9828	2.0000	0.3503	0.2148	80:06
	190	9828	2.0000	0.3461	0.2148	76:19
7	220	9828	2.0000	0.3443	0.2146	78:24
4	194	9828	2.0000	0.3398	0.2148	80:04
1	201	9828	2.0000	0.3387	0.2162	79:32
3	196	9828	2.0000	0.3363	0.2146	80:06
6	213	9828	2.0000	0.3347	0.2147	78:37
5	203	9828	2.0000	0.3337	0.2150	78:50
8	218	9828	2.0000	0.3314	0.2146	78:26
10	230	9828	2.0000	0.3277	0.2147	78:21
9	229	9828	2.0000	0.3236	0.2146	78:26

6.2.3 Testes executados para a base de dados 20 Newsgroups

A Tabela 20 apresenta os resultados obtidos em testes do algoritmo DHC para a base de dados 20 newsgroups. Foi obtida melhora nos resultados em metade das configurações, sendo elas as que dão mais peso às entidades nomeadas. O tempo gasto para a execução com tais configurações chega a ser metade do tempo gasto sem o uso de entidades.

Tabela 20 - Testes do algoritmo DHC executados para a base de dados 20 newsgroups utilizando a Abordagem 2 de uso de entidades nomeadas

Configurações			**Resultados**			**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
7	4	2116	5.5867	0.1800	0.2036	6:24
9	4	2058	5.7135	0.1769	0.2041	6:10
8	4	2172	5.5596	0.1747	0.1974	6:17
10	4	1408	5.9099	0.1720	0.1905	6:03
6	4	2263	5.3763	0.1698	0.1964	6:43
	5	2601	4.6314	0.1491	0.1810	13:18
5	4	2406	5.1163	0.1474	0.1838	7:00
2	4	2727	4.6320	0.1456	0.1818	9:20
4	4	2548	4.8870	0.1275	0.1883	7:18
3	4	2673	4.7053	0.1215	0.1818	8:15
1	5	2561	4.6710	0.1175	0.1809	11:12

Já para os resultados alcançados com o algoritmo UPGMA, apresentados na Tabela 21, todas as configurações da abordagem 2 superaram os valores de *F-measure* obtidos sem o uso de entidades. Os valores de *F1-travel* foram bastante similares e foi alcançada melhora no tempo de execução.

Tabela 21 - Testes do algoritmo UPGMA executados para a base de dados 20 newsgroups utilizando a Abordagem 2 de uso de entidades nomeadas

Configurações			Resultados			Tempo
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
10	145	9907	2.0000	0.2825	0.1820	77:40
9	145	9907	2.0000	0.2817	0.1819	77:30
8	145	9907	2.0000	0.2742	0.1819	78:21
7	140	9907	2.0000	0.2550	0.1820	79:09
6	136	9907	2.0000	0.2311	0.1820	79:49
5	131	9907	2.0000	0.2259	0.1819	84:41
4	133	9907	2.0000	0.2220	0.1820	83:17
3	119	9907	2.0000	0.2062	0.1820	82:57
2	125	9907	2.0000	0.1844	0.1819	83:06
1	110	9907	2.0000	0.1823	0.1819	86:02
	107	9907	2.0000	0.1821	0.1819	90:06

6.2.4 Análise

Os gráficos apresentados na Figura 11 mostram que, para as bases de dados Reuters-21578 e Ohsumed, não se alcançou grande melhora no valor de *F-measure*. Para a base Reuters-21578 os melhores resultados foram obtidos para as configurações que dão menor peso às entidades nomeadas. Já para a base de dados Ohsumed, o uso de entidades não teve grande influência sobre os resultados, sendo os resultados pouco inferiores àqueles alcançados sem o uso de entidades. Para a base de dados 20 newsgroups, obteve-se melhora na maioria das configurações, sendo que aqui o comportamento foi inverso àquele apresentado na base Reuters-21578: melhores resultados foram obtidos com maior peso atribuído às entidades.

Os valores de *F1-travel*, apresentados nos gráficos da Figura 12 não apresentaram melhora significativa. Para a base de dados Reuters-21578 houve pequena melhora para configurações com menor peso atribuído às entidades e resultados piores com o aumento de tal peso. Para as bases de dados Ohsumed e

20 newsgroups o uso de entidades não influenciou significativamente os valores de *F1-travel*.

Para esta abordagem, na base de dados Reuters-21578, o algoritmo UPGMA apresentou melhores resultados para *F-measure*, enquanto o algoritmo DHC obteve maiores valores de *F1-travel*. Nas demais bases os algoritmos alcançaram resultados próximos. É possível observar nos gráficos que ambos os algoritmos apresentaram influência similar do uso de entidades nas diferentes bases de dados.

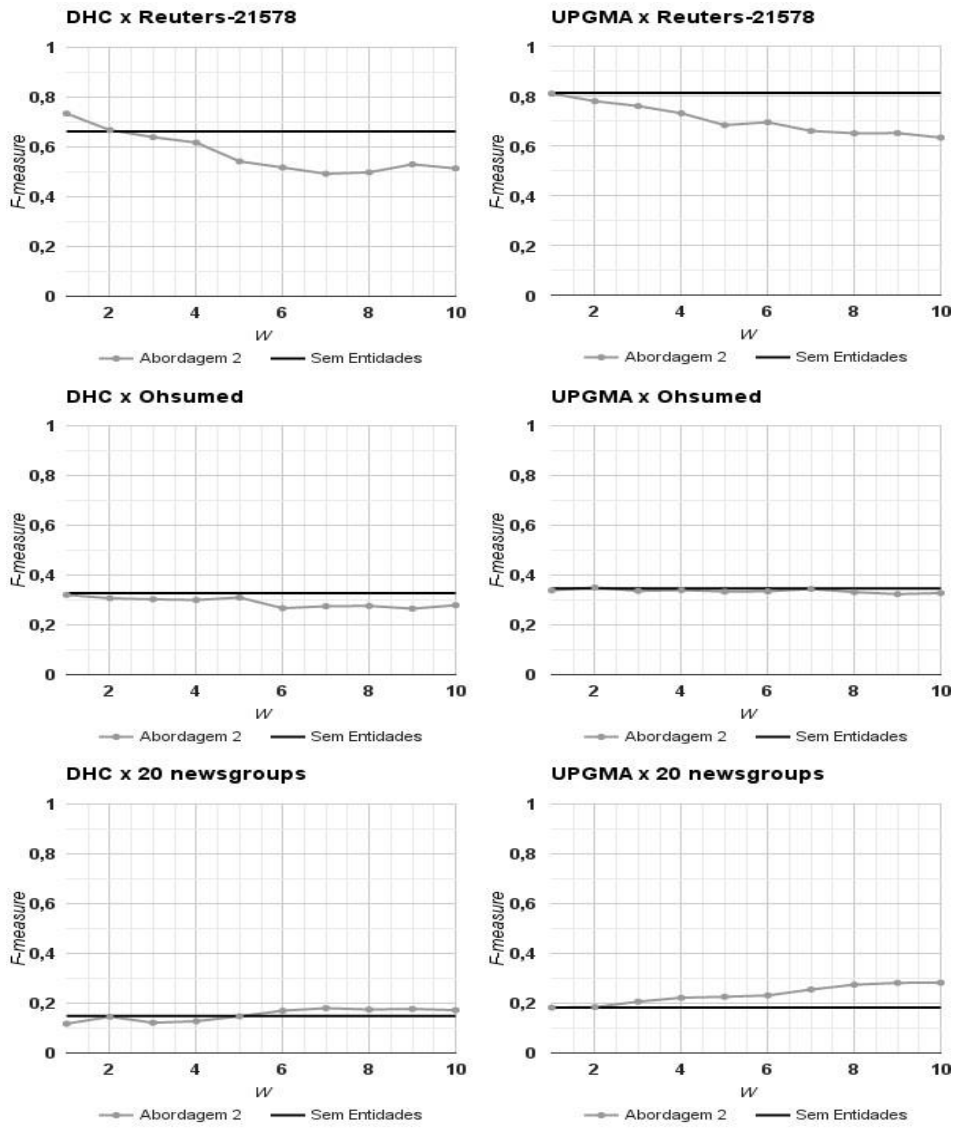


Figura 11 - Valores de *F-measure* alcançados nos testes da abordagem 2

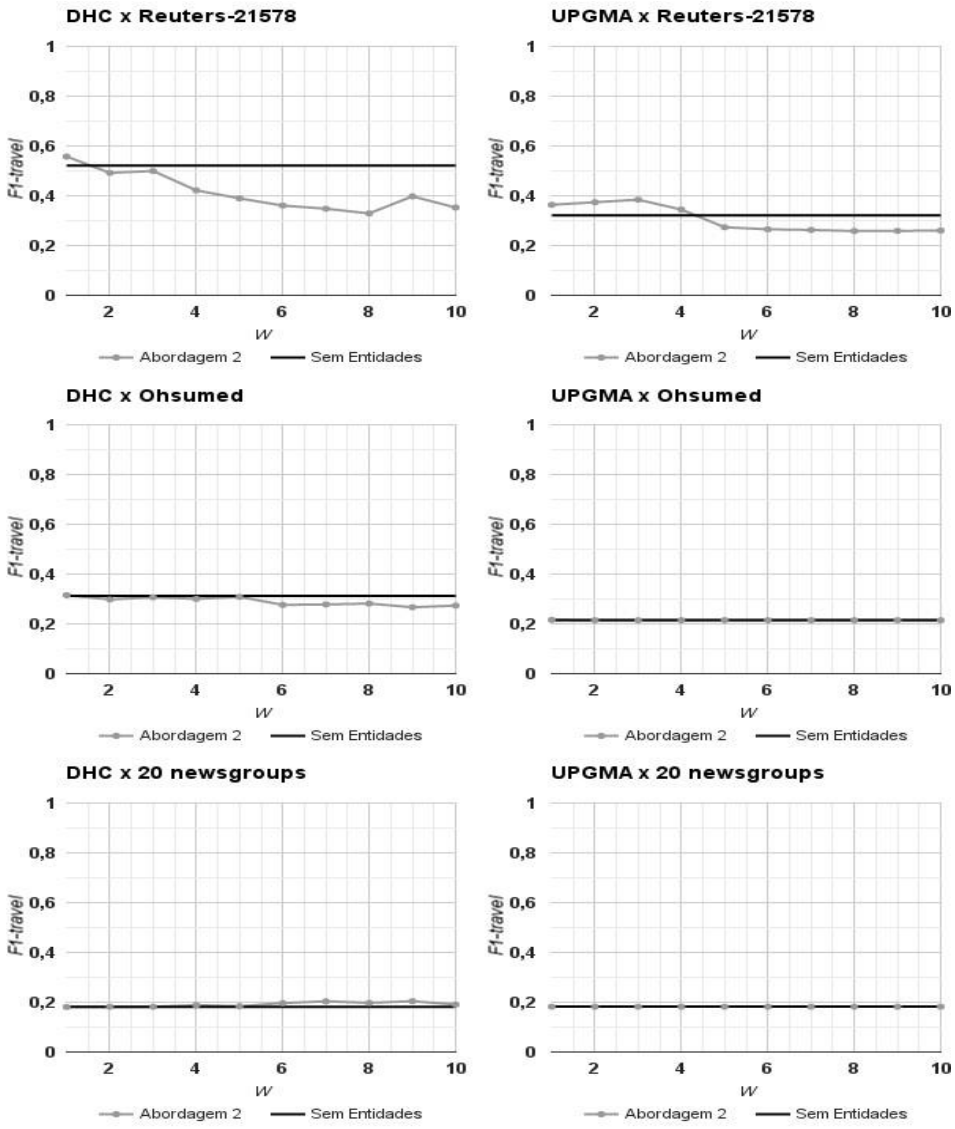


Figura 12 - Valores de *F1-travel* alcançados nos testes da abordagem 2

6.3 Avaliação da Abordagem 3

Esta seção apresenta os testes efetuados utilizando a Abordagem 3 (entidades como termos com peso em função do vetor de termos). Os resultados obtidos a partir das execuções dos algoritmos para as bases de dados Reuters-21578, Ohsumed e 20 newsgroups são apresentados nas seções 6.3.1, 6.3.2 e 6.3.3, respectivamente.

6.3.1 Testes executados para a base de dados Reuters-21578

A Tabela 22 apresenta os resultados obtidos a partir dos testes efetuados com o algoritmo DHC e a base de dados Reuters-21578 utilizando a abordagem 3 de uso de entidades. Duas das configurações para esta abordagem superaram os valores de *F-measure* e *F1-travel* dos resultados obtidos sem o uso de entidades, inclusive com tempos de execução um pouco inferiores.

Tabela 22 - Testes do algoritmo DHC executados para a base de dados Reuters utilizando a Abordagem 3 de uso de entidades nomeadas

Configurações			**Resultados**			**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.2	5	2172	5.7750	0.7051	0.5531	36:22
0.1	5	2179	5.7596	0.7033	0.5469	37:33
	5	2162	5.7970	0.6629	0.5214	38:42
0.3	5	2203	5.7078	0.6520	0.5094	34:07
0.4	4	2283	5.5429	0.5880	0.4648	28:30
0.6	5	2364	5.3873	0.5716	0.4288	24:33
0.5	5	2313	5.4840	0.5468	0.4011	27:35
1	5	2404	5.3143	0.5359	0.3779	18:12
0.7	5	2343	5.4185	0.5231	0.3648	23:37
0.8	5	2370	5.3665	0.5134	0.3603	20:51
1.2	5	2404	5.3143	0.5092	0.3450	16:21
1.4	5	2378	5.3506	0.5092	0.3546	15:30
1.1	5	2378	5.3615	0.5080	0.3349	17:33
1.6	5	2397	5.3253	0.5010	0.3519	15:23
1.3	5	2392	5.3360	0.5008	0.3372	16:19
1.9	5	2391	5.3094	0.4994	0.3551	13:43
2	5	2396	5.3004	0.4981	0.3548	13:20
0.9	5	2381	5.3560	0.4962	0.3529	19:47
1.5	5	2392	5.3360	0.4959	0.3530	15:00
1.8	5	2411	5.3018	0.4903	0.3426	14:14
1.7	5	2354	5.3737	0.4893	0.3328	14:22

Para o algoritmo UPGMA (TABELA 23), apenas a configuração W = 0,2 superou os resultados sem o uso de entidades para as métricas *F-measure* e *F1travel*. Porém, destaca-se também a configuração W = 0,1, que obteve valor de *F-measure* muito próximo e superou todas as demais no valor de *F1-travel*.

Tabela 23 - Testes do algoritmo UPGMA executados para a base de dados Reuters utilizando a Abordagem 3 de uso de entidades nomeadas

Configurações		Resultados				Tempo
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.2	122	10375	2.0000	0.8217	0.3775	94:40
	120	10375	2.0000	0.8128	0.3211	98:00
0.1	114	10375	2.0000	0.8024	0.4017	93:27
0.3	118	10375	2.0000	0.7861	0.3911	93:42
0.4	119	10375	2.0000	0.6955	0.2805	91:43
0.5	117	10375	2.0000	0.6646	0.3077	92:34
0.6	134	10375	2.0000	0.6422	0.3396	93:57
0.7	137	10375	2.0000	0.6206	0.3568	94:24
0.8	136	10375	2.0000	0.6149	0.3548	96:07
1.1	176	10375	2.0000	0.6110	0.3198	98:14
1.4	194	10375	2.0000	0.6098	0.3227	97:40
1.3	194	10375	2.0000	0.6092	0.3015	98:15
0.9	154	10375	2.0000	0.6048	0.2863	97:45
1.5	199	10375	2.0000	0.6045	0.3238	97:30
1	167	10375	2.0000	0.6038	0.3315	97:09
1.7	221	10375	2.0000	0.6034	0.3266	97:22
1.6	214	10375	2.0000	0.6028	0.3310	97:40
1.8	224	10375	2.0000	0.5991	0.3293	97:55
1.2	179	10375	2.0000	0.5891	0.3278	96:22
1.9	227	10375	2.0000	0.5871	0.3217	99:13
2	230	10375	2.0000	0.5847	0.3229	99:41

6.3.2 Testes executados para a base de dados Ohsumed

Os resultados obtidos nos testes do algoritmo DHC para a base de dados Ohsumed utilizando a abordagem 3 são apresentados na Tabela 24. Novamente destaca-se a configuração $W = 0.2$ que foi a única que superou os resultados obtidos sem o uso de entidades. Os tempos de execução foram bem próximos para as configurações que obtiveram os melhores resultados.

Tabela 24 - Testes do algoritmo DHC executados para a base de dados Ohsumed utilizando a Abordagem 3 de uso de entidades nomeadas

Configurações		**Resultados**				**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.2	5	2308	5.2477	0.3289	0.3169	21:37
	5	2303	5.2604	0.3270	0.3126	19:56
0.1	4	2283	5.3034	0.3073	0.2983	20:16
0.3	5	2169	5.2986	0.3042	0.3018	21:12
0.5	4	2376	5.1350	0.2842	0.2834	19:09
0.9	5	2303	5.0738	0.2794	0.2873	17:15
0.8	4	2439	5.0283	0.2736	0.2728	17:39
0.4	4	2342	5.1950	0.2734	0.2726	19:43
1.2	5	2314	5.0497	0.2703	0.2663	16:25
1	4	2479	4.9633	0.2687	0.2687	16:59
0.6	4	2412	5.0734	0.2668	0.2661	18:44
1.1	4	2484	4.9553	0.2640	0.2640	18:51
1.7	4	2541	4.8666	0.2578	0.2577	15:19
1.8	4	2553	4.8485	0.2574	0.2573	15:20
2	4	2576	4.8133	0.2525	0.2530	14:48
1.9	4	2557	4.8417	0.2521	0.2526	15:13
0.7	4	2435	5.0349	0.2499	0.2505	18:05
1.4	4	2521	4.8973	0.2477	0.2429	15:40
1.5	4	2534	4.8773	0.2472	0.2426	15:41
1.6	4	2537	4.8727	0.2472	0.2437	15:32
1.3	5	2120	5.1537	0.2469	0.2430	15:55

Nos testes efetuados com o algoritmo UPGMA (TABELA 25) apenas uma configuração melhorou os resultados com relação ao não uso de entidades. Nestes testes todas as execuções obtiveram resultados muito próximos, tanto para *F-measure* quanto para *F1-travel*.

Tabela 25 - Testes do algoritmo UPGMA executados para a base de dados Ohsumed utilizando a Abordagem 3 de uso de entidades nomeadas

Configurações		Resultados				Tempo
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.3	183	9828	2.0000	0.3557	0.2146	79:57
	190	9828	2.0000	0.3461	0.2148	76:19
0.2	189	9828	2.0000	0.3413	0.2149	79:09
0.9	246	9828	2.0000	0.3358	0.2146	78:22
0.4	201	9828	2.0000	0.3309	0.2147	79:16
0.7	228	9828	2.0000	0.3300	0.2146	79:16
0.1	191	9828	2.0000	0.3295	0.2149	82:03
1.6	308	9828	2.0000	0.3260	0.2146	83:53
1.4	289	9828	2.0000	0.3259	0.2146	75:29
1.5	298	9828	2.0000	0.3259	0.2146	77:48
1.7	313	9828	2.0000	0.3251	0.2146	75:22
1.3	285	9828	2.0000	0.3232	0.2146	81:51
2	335	9828	2.0000	0.3231	0.2146	76:36
1.8	321	9828	2.0000	0.3226	0.2146	75:55
1.9	331	9828	2.0000	0.3212	0.2146	77:19
1.1	266	9828	2.0000	0.3202	0.2146	79:46
1.2	287	9828	2.0000	0.3192	0.2146	78:16
1	256	9828	2.0000	0.3171	0.2146	78:09
0.6	221	9828	2.0000	0.3137	0.2146	81:14
0.5	213	9828	2.0000	0.3130	0.2148	79:05
0.8	233	9828	2.0000	0.3109	0.2146	78:48

6.3.3 Testes executados para a base de dados 20 Newsgroups

Os resultados obtidos com o algoritmo DHC para a base de dados 20 newsgroups são apresentados na Tabela 26. Nestes testes, o uso de entidades com a abordagem 3 proporcionou melhora significativa nos resultados na maioria das configurações utilizadas. Pode-se observar também que o tempo de execução da maioria das configurações foi consideravelmente inferior ao tempo sem o uso de entidades.

Tabela 26 - Testes do algoritmo DHC executados para a base de dados 20 newsgroups utilizando a Abordagem 3 de uso de entidades nomeadas

Configurações		**Resultados**				**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
1.2	4	2549	4.8839	0.2038	0.2167	4:49
1	4	2486	4.9682	0.1983	0.2106	4:55
1.1	4	2512	4.9375	0.1975	0.2107	4:59
1.9	4	2624	4.7684	0.1890	0.2118	4:45
1.4	4	2488	4.9237	0.1862	0.2064	4:41
2	4	2596	4.7955	0.1852	0.2055	4:46
0.6	4	2084	5.1376	0.1838	0.2064	6:20
1.8	4	2614	4.7859	0.1812	0.2056	5:03
1.7	4	2370	4.9582	0.1812	0.2010	4:59
1.5	4	2169	5.0530	0.1797	0.2000	4:48
0.8	4	2272	5.1012	0.1740	0.1992	5:41
0.7	3	2371	5.1771	0.1711	0.1962	5:58
1.6	4	2356	4.9724	0.1711	0.1970	5:00
0.9	4	1651	5.2536	0.1687	0.1878	5:10
	5	2601	4.6314	0.1491	0.1810	13:18
1.3	4	2586	4.8288	0.1670	0.1908	4:46
0.5	4	2532	4.8800	0.1417	0.1910	6:56
0.1	5	2375	4.6225	0.1290	0.1794	11:40
0.4	4	2655	4.7176	0.1115	0.1818	7:57
0.2	4	2706	4.6598	0.1093	0.1818	10:54
0.3	4	2668	4.7119	0.1075	0.1818	9:12

Como mostra a Tabela 27, para o algoritmo UPGMA a melhora foi ainda mais significativa com o uso da abordagem 4. Todas as configurações obtiveram melhores resultados que o não uso de entidades tanto nas métricas *F-measure* e *F1-travel*, quanto no tempo de execução.

Tabela 27 - Testes do algoritmo UPGMA executados para a base de dados 20 newsgroups utilizando a Abordagem 3 de uso de entidades nomeadas

Configurações		Resultados				Tempo
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
1.8	183	9907	2.0000	0.3087	0.1820	80:19
1.5	169	9907	2.0000	0.3083	0.1820	79:23
1.9	192	9907	2.0000	0.3069	0.1819	77:58
1.7	184	9907	2.0000	0.3069	0.1820	77:19
1.2	168	9907	2.0000	0.3054	0.1820	77:39
1.1	156	9907	2.0000	0.3050	0.1820	79:24
1.6	176	9907	2.0000	0.3046	0.1820	77:24
1.4	167	9907	2.0000	0.3033	0.1820	79:43
2	192	9907	2.0000	0.3030	0.1820	77:21
1.3	163	9907	2.0000	0.2996	0.1821	77:37
1	146	9907	2.0000	0.2903	0.1821	80:19
0.9	146	9907	2.0000	0.2899	0.1819	80:30
0.8	140	9907	2.0000	0.2818	0.1822	78:13
0.7	134	9907	2.0000	0.2553	0.1819	78:50
0.6	121	9907	2.0000	0.2460	0.1820	80:16
0.5	117	9907	2.0000	0.2413	0.1821	81:01
0.4	118	9907	2.0000	0.2131	0.1819	81:25
0.3	109	9907	2.0000	0.1882	0.1820	82:35
0.2	115	9907	2.0000	0.1823	0.1819	82:58
0.1	108	9907	2.0000	0.1822	0.1819	84:06
	107	9907	2.0000	0.1821	0.1819	90:06

6.3.4 Análise

A Figura 13 apresenta os gráficos dos valores de *F-measure* obtidos utilizando a abordagem 3. Para a base de dados Reuters-21578, observa-se melhora nos resultados apenas nas primeiras configurações, onde o peso dado às entidades é menor. Nas demais configurações há uma queda considerável nos resultados. Já para a base de dados Ohsumed os resultados são um pouco inferiores, mas ainda próximos, aos obtidos sem o uso de entidades. Para a base de dados 20 newsgroups houve melhora nos dados na maioria das configurações, com destaque para a melhora obtida com o algoritmo UPGMA.

Os resultados obtidos para a métrica *F1-travel* são apresentados nos gráficos da Figura 14. É possível observar que o algoritmo UPGMA obteve

melhora em algumas configurações para a base de dados Reuters-21578, e que o algoritmo DHC obteve resultados piores na maioria das configurações para a mesma base. Para as demais bases de dados, os algoritmos tiveram pouca influência das entidades neste quesito, estando os resultados próximos daqueles obtidos sem o uso de entidades.

Assim como nas demais abordagens, aqui o algoritmo UPGMA obteve melhores resultados que o DHC para a métrica *F-measure*, enquanto o algoritmo DHC alcançou resultados levemente superiores para *F1-travel*.

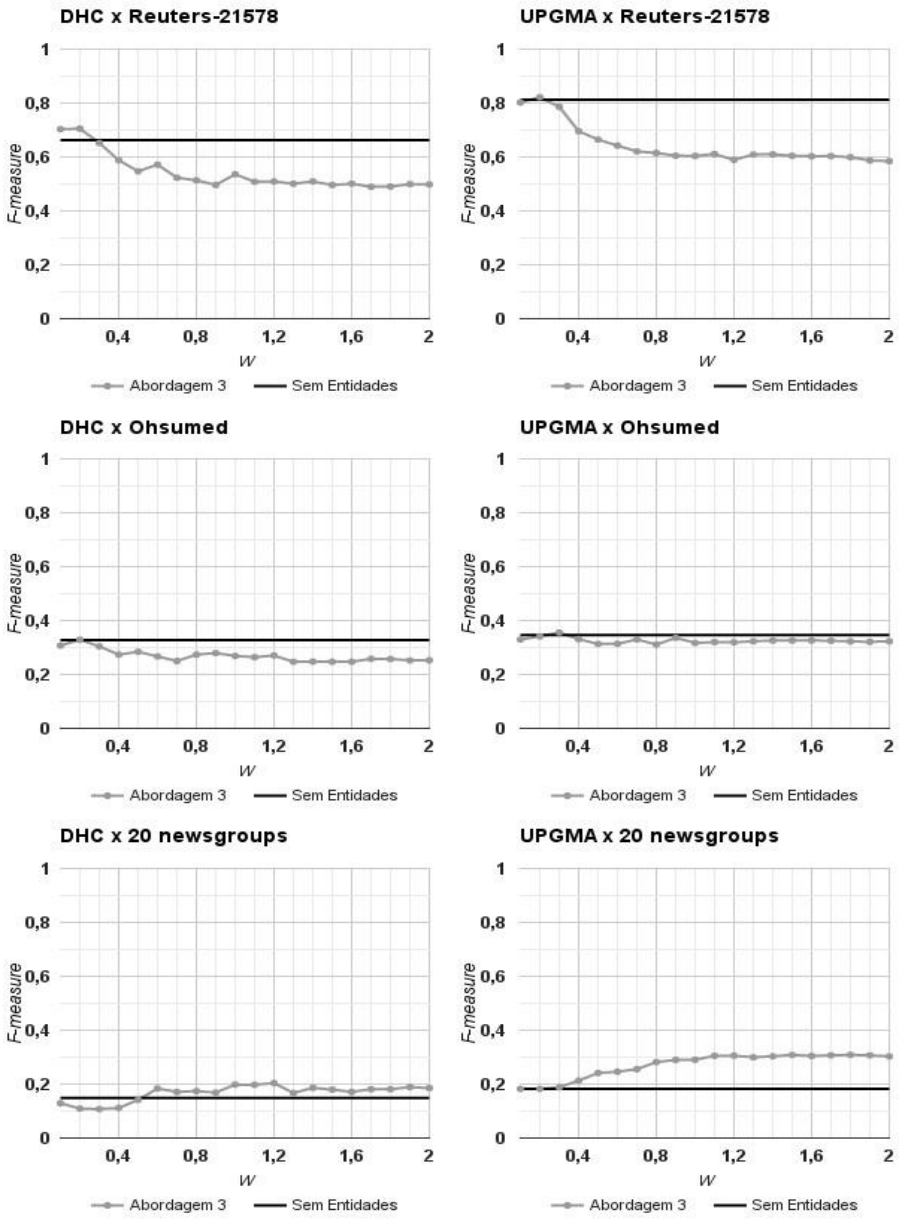


Figura 13 - Valores de *F-measure* alcançados nos testes da abordagem 3

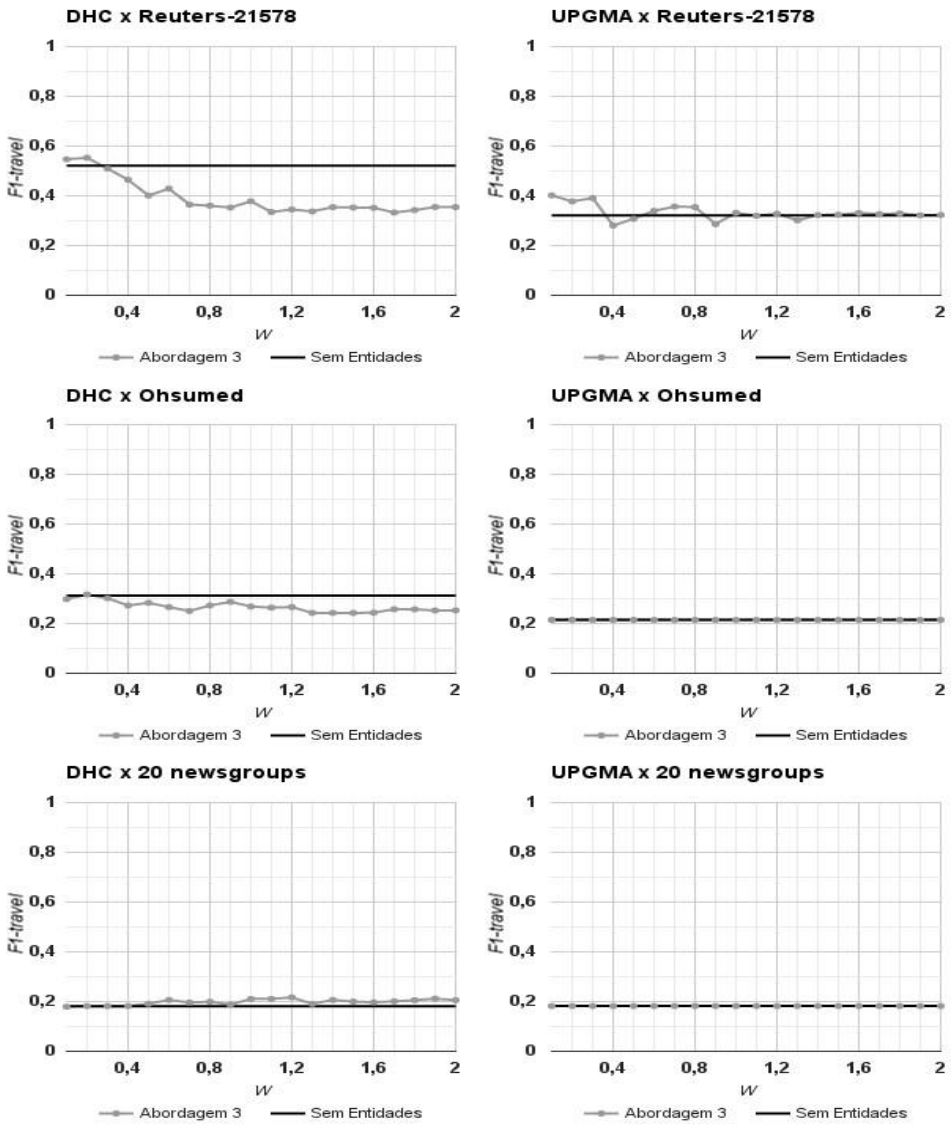


Figura 14 - Valores de *F1-travel* alcançados nos testes da abordagem 3

6.4 Avaliação da Abordagem 4

Esta seção apresenta os testes efetuados utilizando a Abordagem 4 (entidades como termos com peso em função do vetor de termos e número de

entidades). Os resultados obtidos a partir das execuções dos algoritmos para as bases de dados Reuters-21578, Ohsumed e 20 newsgroups são apresentados nas seções 6.4.1, 6.4.2 e 6.4.3, respectivamente.

6.4.1 Testes executados para a base de dados Reuters-21578

A Tabela 28 apresenta os resultados obtidos a partir dos testes efetuados com o algoritmo DHC e a base de dados Reuters-21578 utilizando a abordagem 4 de uso de entidades. A Tabela 29 apresenta os resultados obtidos para esta mesma abordagem utilizando o algoritmo UPGMA.

Tabela 28 - Testes do algoritmo DHC executados para a base de dados Reuters utilizando a Abordagem 4 de uso de entidades nomeadas

Configurações		**Resultados**				**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.2	5	2173	5.7728	0.7267	0.5689	38:9
0.1	5	2173	5.7728	0.7252	0.5511	37:30
0.3	5	2172	5.7750	0.7203	0.5483	41:37
0.4	5	2176	5.7662	0.6933	0.5460	37:0
0.5	6	2178	5.7618	0.6808	0.4845	37:20
0.7	5	2182	5.7531	0.6563	0.5135	36:38
0.8	5	2224	5.6634	0.6538	0.5139	34:51
0.6	5	2177	5.7640	0.6534	0.5108	37:30
1.2	6	2282	5.5449	0.6255	0.4506	33:19
1	5	2240	5.6301	0.6166	0.4657	33:3
1.3	5	2287	5.5350	0.6096	0.4593	31:11
1.1	5	2263	5.5830	0.6077	0.4282	32:27
1.5	5	2295	5.5192	0.5910	0.4325	29:6
1.7	5	2280	5.5489	0.5827	0.4279	27:29
1.6	4	2276	5.5569	0.5702	0.4518	28:37
1.4	4	2281	5.5469	0.5619	0.4371	29:36
1.8	5	2291	5.5271	0.5429	0.3971	26:48
1.9	5	2301	5.5074	0.5427	0.4046	26:40
2	5	2282	5.5449	0.5262	0.3761	25:22
	5	2162	5.7970	0.6629	0.5214	38:42

Tabela 29 - Testes do algoritmo UPGMA executados para a base de dados Reuters utilizando a Abordagem 4 de uso de entidades nomeadas

Configurações			**Resultados**			**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.5	113	10375	2.0000	0.8387	0.4464	95:43
0.6	110	10375	2.0000	0.8385	0.3669	95:4
0.2	117	10375	2.0000	0.8345	0.3068	94:46
0.9	107	10375	2.0000	0.8194	0.3106	96:56
0.4	112	10375	2.0000	0.8134	0.3954	95:33
0.1	123	10375	2.0000	0.8097	0.3638	93:21
0.3	113	10375	2.0000	0.8095	0.3878	96:36
0.8	116	10375	2.0000	0.8093	0.4061	97:0
0.7	109	10375	2.0000	0.8025	0.3163	94:8
1	119	10375	2.0000	0.7734	0.3793	96:46
1.1	123	10375	2.0000	0.7604	0.2702	97:17
1.2	122	10375	2.0000	0.7434	0.2670	96:28
1.3	128	10375	2.0000	0.7396	0.2766	98:0
1.5	147	10375	2.0000	0.7332	0.2769	96:52
1.4	142	10375	2.0000	0.7300	0.2810	96:24
1.6	159	10375	2.0000	0.7193	0.2642	95:50
1.7	169	10375	2.0000	0.7128	0.2602	94:10
1.8	175	10375	2.0000	0.7066	0.2599	98:56
1.9	178	10375	2.0000	0.6797	0.2666	94:12
2	189	10375	2.0000	0.6743	0.2582	96:38
	120	10375	2.0000	0.8128	0.3211	98:0

6.4.2 Testes executados para a base de dados Ohsumed

Esta seção apresenta os resultados dos testes efetuados utilizando a abordagem 4 de uso de entidades nomeadas e a base de dados Ohsumed. A Tabela 30 apresenta os resultados obtidos utilizando o algoritmo DHC e a Tabela 31 apresenta aqueles obtidos utilizando o algoritmo UPGMA.

Tabela 30 - Testes do algoritmo DHC executados para a base de dados Ohsumed utilizando a Abordagem 4 de uso de entidades nomeadas

Configurações		**Resultados**				**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.6	5	2206	5.3303	0.3294	0.3138	20:06
0.4	4	2273	5.3223	0.3255	0.3268	20:33
0.3	5	2196	5.3291	0.3234	0.3193	20:39
0.2	4	2281	5.3072	0.3194	0.3142	21:46
0.1	5	2283	5.2977	0.3147	0.3092	20:24
0.5	5	2288	5.2883	0.3126	0.3067	20:15
1.2	5	2340	5.1401	0.3032	0.3019	19:15
1	4	2359	5.1648	0.2996	0.2996	19:51
0.8	5	2203	5.2813	0.2974	0.2492	19:46
1.3	4	2385	5.1194	0.2931	0.2974	18:14
1.1	4	2362	5.1595	0.2920	0.2974	19:24
1.7	5	2244	5.1091	0.2893	0.2867	17:49
0.9	4	2336	5.2058	0.2864	0.2800	19:50
1.6	5	2326	5.1092	0.2840	0.2821	17:56
1.9	4	2427	5.0482	0.2764	0.2801	17:28
2	4	2436	5.0332	0.2751	0.2794	17:31
0.7	5	1491	5.3954	0.2738	0.2364	20:02
1.5	5	2334	5.1096	0.2682	0.2683	17:57
1.8	4	2430	5.0432	0.2620	0.2684	17:40
1.4	4	2395	5.1023	0.2527	0.2580	18:23
	5	2303	5.2604	0.3270	0.3126	19:56

Tabela 31 - Testes do algoritmo UPGMA executados para a base de dados Ohsumed utilizando a Abordagem 4 de uso de entidades nomeadas

Configurações			**Resultados**			**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
0.8	196	9828	2.0000	0.3654	0.2147	80:03
1.1	212	9828	2.0000	0.3497	0.2149	79:47
0.9	198	9828	2.0000	0.3459	0.2147	80:34
0.5	189	9828	2.0000	0.3443	0.2146	80:06
0.6	196	9828	2.0000	0.3439	0.2147	80:08
0.4	198	9828	2.0000	0.3437	0.2173	80:44
0.1	200	9828	2.0000	0.3418	0.2162	79:37
0.3	188	9828	2.0000	0.3412	0.2437	79:36
0.2	195	9828	2.0000	0.3408	0.2149	79:20
1.7	252	9828	2.0000	0.3401	0.2146	76:51
1	212	9828	2.0000	0.3379	0.2148	80:03
1.6	245	9828	2.0000	0.3378	0.2146	78:12
0.7	201	9828	2.0000	0.3311	0.2147	81:59
1.5	238	9828	2.0000	0.3308	0.2147	77:49
1.3	222	9828	2.0000	0.3211	0.2148	77:58
1.2	213	9828	2.0000	0.3164	0.2147	79:43
2	258	9828	2.0000	0.3138	0.2146	77:05
1.8	250	9828	2.0000	0.3109	0.2146	77:39
1.9	255	9828	2.0000	0.3084	0.2147	77:23
1.4	231	9828	2.0000	0.3072	0.2146	77:26
	190	9828	2.0000	0.3461	0.2148	76:19

6.4.3 Testes executados para a base de dados 20 Newsgroups

A seguir são apresentados os resultados obtidos utilizando a abordagem 4 de uso de entidades nomeadas e a base de dados 20 newsgroups. A Tabela 32 apresenta os resultados obtidos com a utilização do algoritmo DHC e a Tabela 33 os resultados obtidos com o algoritmo UPGMA.

Tabela 32 - Testes do algoritmo DHC executados para a base de dados 20 newsgroups utilizando a Abordagem 4 de uso de entidades nomeadas

Configurações		**Resultados**				**Tempo**
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
	5	2601	4.6314	0.1491	0.1810	13:18
2	4	2400	4.9917	0.1323	0.1850	6:28
0.3	4	2697	4.6723	0.1260	0.1818	11:46
1.2	4	2459	4.8703	0.1232	0.1811	8:03
1.9	4	2093	5.0874	0.1227	0.1803	6:46
0.8	4	2666	4.7150	0.1189	0.1818	9:29
0.5	5	2653	4.6534	0.1180	0.1815	11:10
0.1	5	2564	4.6616	0.1176	0.1809	11:44
0.4	4	2709	4.6561	0.1146	0.1818	11:21
0.2	5	2671	4.6385	0.1142	0.1815	11:50
1.7	4	2322	4.9359	0.1127	0.1805	7:20
1.1	4	2676	4.7008	0.1096	0.1818	8:18
1	5	2451	4.7847	0.1079	0.1807	8:47
0.7	5	2245	4.7244	0.1077	0.1788	10:21
0.9	4	2677	4.6998	0.1069	0.1818	9:03
0.6	4	2686	4.6874	0.1043	0.1818	10:55
1.3	4	2389	4.8264	0.1042	0.1804	7:41
1.6	4	2508	4.8230	0.1028	0.1813	7:32
1.4	4	2248	4.9102	0.0998	0.1797	7:19
1.8	4	889	5.1292	0.0932	0.1575	7:15
1.5	4	676	4.9276	0.0771	0.1457	7:45

Tabela 33 - Testes do algoritmo UPGMA executados para a base de dados 20 newsgroups utilizando a Abordagem 4 de uso de entidades nomeadas

Configurações		Resultados				Tempo
W	**Níveis**	**Grupos**	**Subgr.**	**F1**	**F1Travel**	**mm:ss**
2	201	9907	2.0000	0.2010	0.1820	79:04
1.9	189	9907	2.0000	0.2001	0.1820	80:54
1.6	181	9907	2.0000	0.1916	0.1820	83:42
1.8	198	9907	2.0000	0.1909	0.1820	82:08
1.5	164	9907	2.0000	0.1897	0.1820	84:48
1.7	183	9907	2.0000	0.1895	0.1820	82:31
1.4	170	9907	2.0000	0.1874	0.1820	81:43
1.3	164	9907	2.0000	0.1873	0.1819	82:46
1.2	157	9907	2.0000	0.1851	0.1820	82:32
1.1	152	9907	2.0000	0.1850	0.1820	85:05
1	143	9907	2.0000	0.1849	0.1820	84:11
0.9	130	9907	2.0000	0.1844	0.1820	84:21
0.8	122	9907	2.0000	0.1830	0.1820	81:53
0.7	111	9907	2.0000	0.1826	0.1819	80:59
0.6	111	9907	2.0000	0.1824	0.1820	81:45
0.1	112	9907	2.0000	0.1823	0.1819	82:48
0.2	104	9907	2.0000	0.1823	0.1819	82:32
0.3	112	9907	2.0000	0.1823	0.1820	81:40
0.5	112	9907	2.0000	0.1822	0.1819	82:18
0.4	112	9907	2.0000	0.1822	0.1820	81:55
	107	9907	2.0000	0.1821	0.1819	90:06

6.4.4 Análise

A Figura 15 apresenta os gráficos dos valores de *F-measure* alcançados utilizando a abordagem 4. Para a base de dados Reuters-21578 algumas configurações superaram os resultados sem o uso de entidades. Também houve menor perda na qualidade dos resultados com o aumento do peso das entidades. Para as bases de dados Ohsumed e 20 newsgroups os resultados obtidos foram mais próximos daqueles alcançados sem entidades.

Para a métrica *F1-travel* (FIGURA 16), os testes para a base Reuters-21578 apresentaram melhora em algumas configurações e piora para as demais. Já para as demais bases de dados não houve influência significante nos

resultados, visto que ficaram muito próximos àqueles alcançados sem o uso de entidades.

O comportamento dos algoritmos se manteve similar às demais abordagens, tendo o algoritmo UPGMA obtido melhores resultados para *F-measure* e o DHC melhores resultados para *F1-travel*.

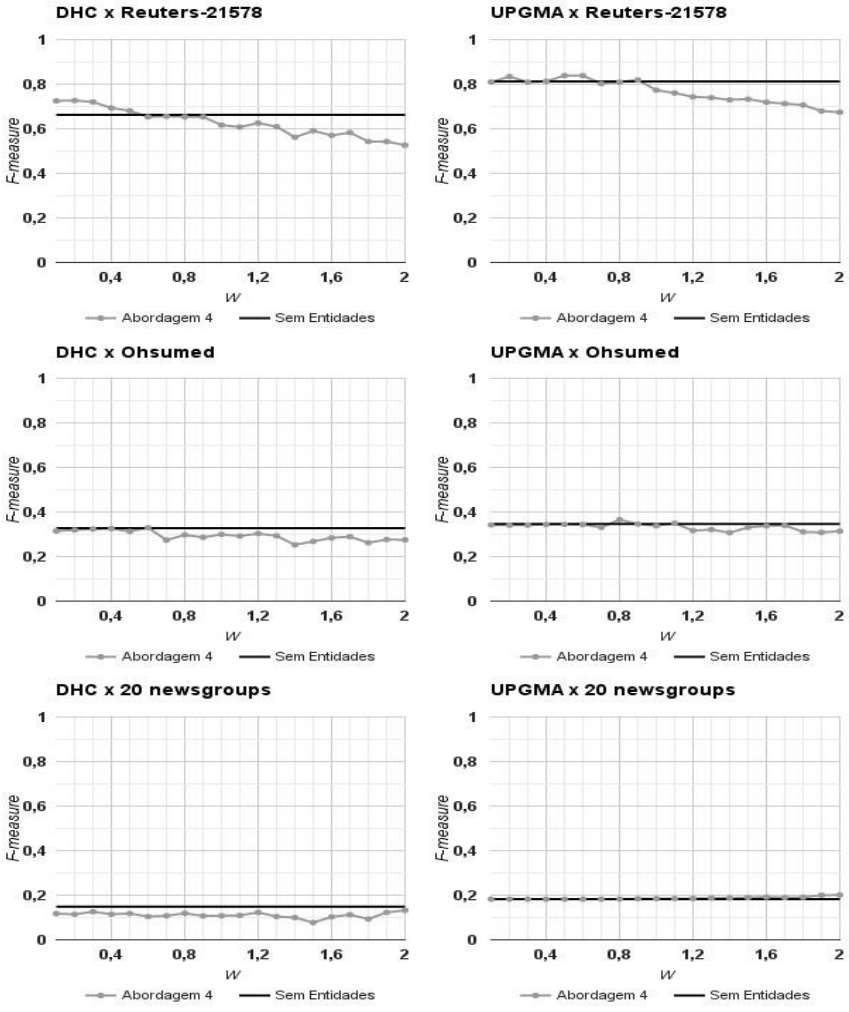


Figura 15 - Valores de *F-measure* alcançados nos testes da abordagem 4

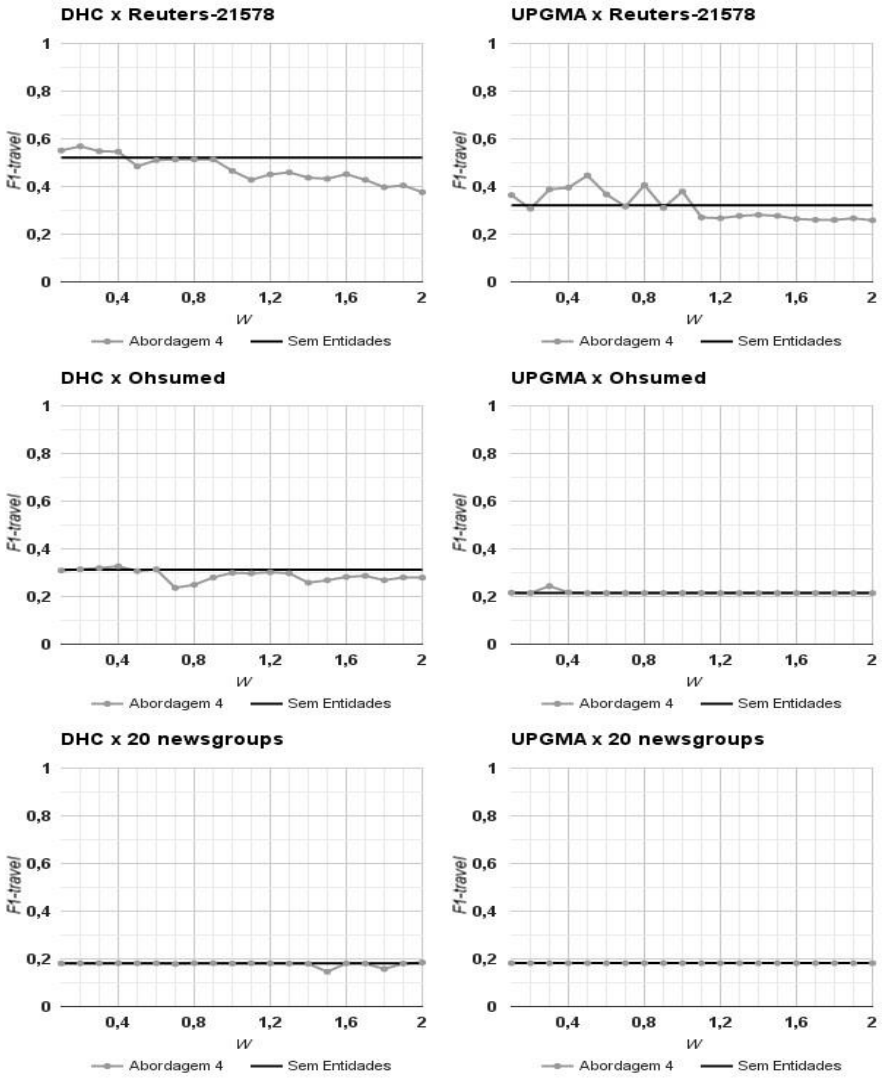


Figura 16 - Valores de *F1-travel* alcançados nos testes da abordagem 4

6.4.5 Comparação entre as abordagens

Analisando os resultados já apresentados observa-se que, em algumas situações, o uso de entidades pode contribuir para a eficácia de algoritmos de agrupamento hierárquico. Porém, encontrar a configuração mais adequada nas

abordagens apresentadas pode ser um grande desafio. Com o objetivo de comparar a eficácia das abordagens, nesta seção os resultados apresentados equivalem à média dos valores alcançados com as três melhores configurações de cada abordagem. As três melhores configurações são aquelas que obtiveram melhores valores de *F-measure* para uma determinada base de dados.

A Figura 17 apresenta um comparativo dos valores de F-measure obtidos com o algoritmo DHC em cada uma das abordagens. É possível observar que, considerando as melhores configurações, destaca-se a abordagem 4 com os melhores resultados para as bases de dados Reuters-21578 e Ohsumed. Porém para a base de dados 20 newsgroups esta foi superada pelas demais abordagens.

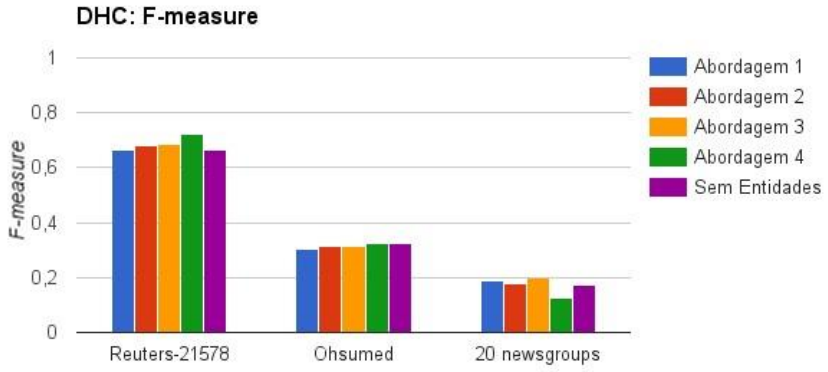


Figura 17- Comparativo dos valores de *F-measure* obtidos com o algoritmo DHC nas diferentes abordagens

A Figura 18 apresenta o mesmo comparativo para o algoritmo UPGMA.

Observa-se que, apesar de o algoritmo UPGMA alcançar maiores valores de *Fmeasure*, principalmente para a base de dados Reuters-21578, o comparativo é bem similar.

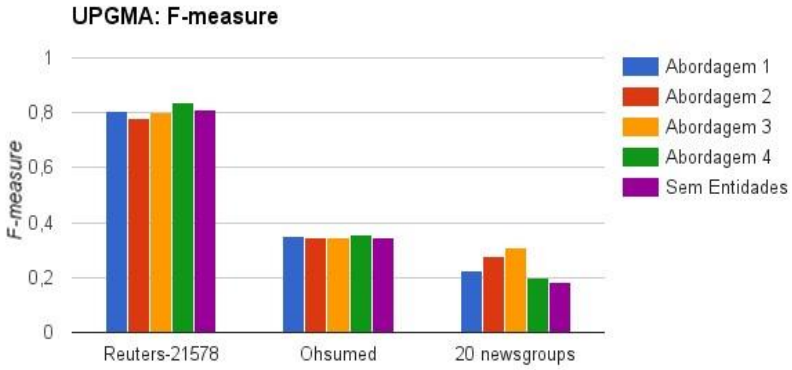


Figura 18 - Comparativo dos valores de *F-measure* obtidos com o algoritmo UPGMA nas diferentes abordagens

Para o quesito navegabilidade da hierarquia gerada, a Figura 19 apresenta um comparativo dos valores de *F1-travel* para o algoritmo DHC nas diferentes abordagens. É possível observar que o comparativo é bem similar ao obtido para *F-measure*.

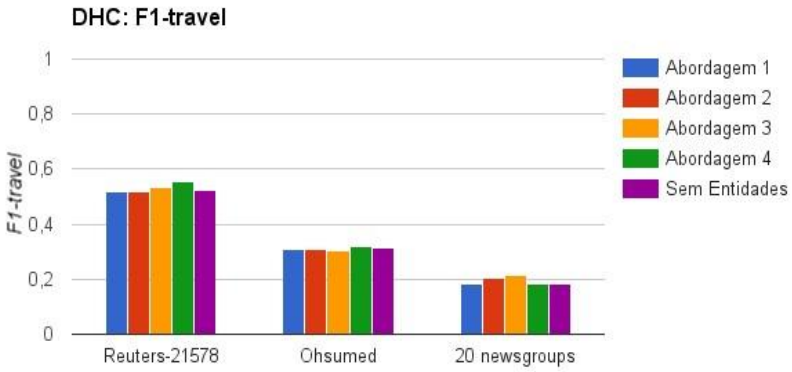


Figura 19 - Comparativo dos valores de *F1-travel* obtidos com o algoritmo DHC nas diferentes abordagens

Já para o algoritmo UPGMA (FIGURA 20) a abordagem 3 se destaca para a base de dados Reuters-21578, porém, para as demais bases de dados o valor de *F1-travel* praticamente não é influenciado pela abordagem utilizada.

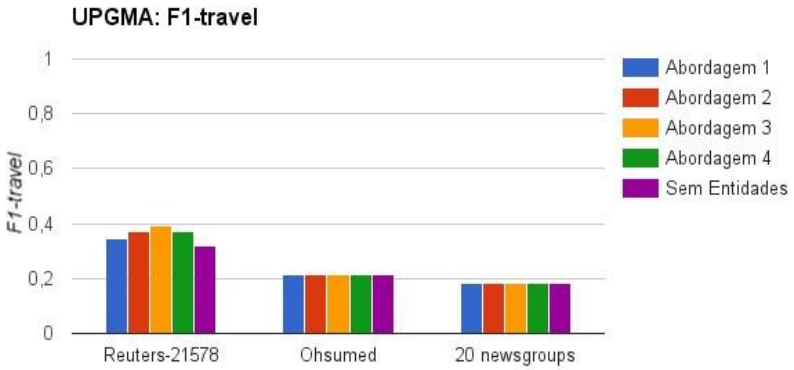


Figura 20 - Comparativo dos valores de *F1-travel* obtidos com o algoritmo UPGMA nas diferentes abordagens

Com relação ao tempo de execução dos testes realizados, a Figura 21 apresenta o comparativo para o algoritmo DHC. O tempo para todas as abordagens é bem próximo, com exceção da abordagem 1, que gastou quantidade consideravelmente maior de tempo para gerar as hierarquias. Pelo fato de este algoritmo ser incremental, a forma como a similaridade entre os documentos é calculada pode influenciar as adaptações a serem feitas nas estruturas de agrupamento ao longo das iterações. Isto pode explicar esta variação no tempo de execução entre as abordagens.

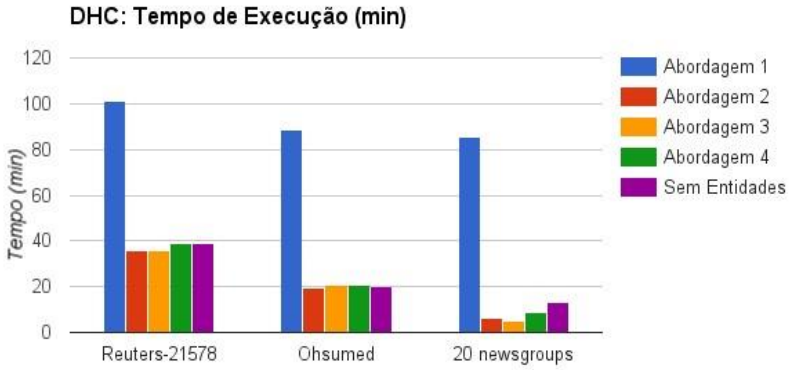


Figura 21 - Comparativo dos tempos de execução do algoritmo DHC nas diferentes abordagens

Para o algoritmo UPGMA houve maior equilíbrio no que diz respeito ao tempo de execução (FIGURA 22). Tal resultado pode ser explicado pelo fato de o algoritmo se comportar de forma similar em todas as suas iterações e gerar sempre o mesmo número de grupos.

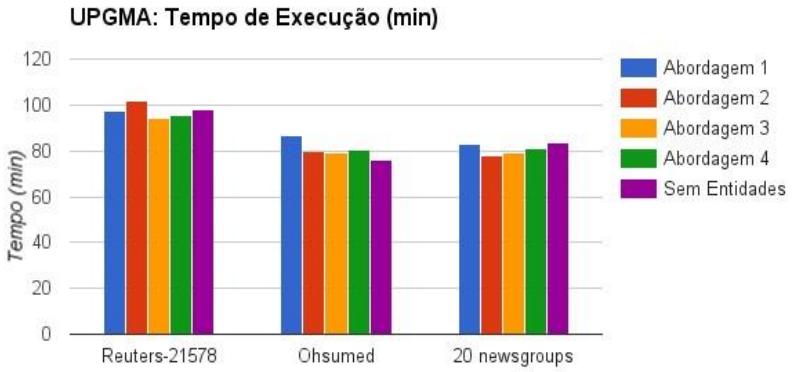


Figura 22 - Comparativo dos tempos de execução do algoritmo UPGMA nas diferentes abordagens

Com os resultados apresentados neste trabalho, pode-se observar que as entidades nomeadas podem auxiliar na melhora de resultados em técnicas de agrupamento hierárquico. As quatro abordagens obtiveram resultados muito próximos, o que dificulta uma definição sobre qual a melhor dentre elas. Porém, nos experimentos aqui realizados destacam-se as abordagens 3 e 4, que obtiveram melhoras nos resultados em diversas situações e conseguiram manter o tempo de execução próximo, e em alguns casos inferior, ao tempo gasto nos testes sem o uso de entidades.

7 CONCLUSÕES E TRABALHOS FUTUROS

7.1 Conclusões

Este trabalho teve como foco o estudo de técnicas de agrupamento hierárquico incremental de textos. Inicialmente foram apresentados os conceitos básicos sobre o problema e técnicas de agrupamento de textos. Visando o desafio de melhor organização de documentos em diferentes níveis de granularidade, e também tendo em mente a dinamicidade das informações no cenário atual, este trabalho teve como foco mais específico os algoritmos de agrupamento hierárquico incremental. Foi efetuado um levantamento sistemático dos trabalhos desenvolvidos nos últimos anos relacionados a este tema e tais trabalhos foram estudados e compilados com o intuito de facilitar a compreensão sobre o estado da arte para este tópico.

Observando os trabalhos estudados, pode-se concluir que agrupamentos hierárquicos e incrementais podem ser aplicados para diversos fins, o que enfatiza a relevância do estudo e desenvolvimento destas técnicas. É possivel constatar que este ainda não é um problema completamente resolvido e não há uma estratégia definitiva para resolvê-lo.

Assim, tendo em vista o desafio de buscar melhores soluções para este problema, este trabalho investigou o uso de entidades nomeadas, extraídas dos documentos, como informação adicional a auxiliar o processo de agrupamento. Para tal, foram estudadas duas abordagens de uso das entidades e proposta mais duas.

Para validar as abordagens de uso de entidades, foram selecionados dois algoritmos descritos nos trabalhos estudados, o *Dynamic Hierarchical Compact* (DHC) e o *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA). Tais algoritmo foram implementados e adaptados para utilizar cada uma das

abordagens. Assim, diversos testes foram efetuados em diferentes bases de dados e com diferentes configurações para os usos de entidades.

Observando os resultados obtidos, conclui-se que as entidades nomeadas são informações que podem auxiliar os algoritmos a obter melhores hierarquias de documentos, visto que as abordagens de uso de entidades superaram em diversos testes os resultados obtidos sem o seu uso. Porém, os resultados mostram que encontrar a configuração correta para utilizar cada uma das abordagens pode ser desafiador, sendo que apenas algumas configurações alcançaram melhoras na qualidade das hierarquias de grupos geradas. Conclui-se também que as entidades nomeadas podem ser utilizadas nos algoritmos de forma a não prejudicar significativamente seu desempenho, tendo como gasto adicional apenas o processamento necessário para extração das mesmas.

Outro ponto a ser observado é que as diferentes abordagens e configurações obtiveram resultados divergentes para cada uma das bases de dados, não sendo possível determinar uma única abordagem e configuração que seja a mais adequada. Para tal, surge a necessidade de estudos mais profundos que busquem relacionar as características de uma base de dados com o melhor uso suas entidades nomeadas.

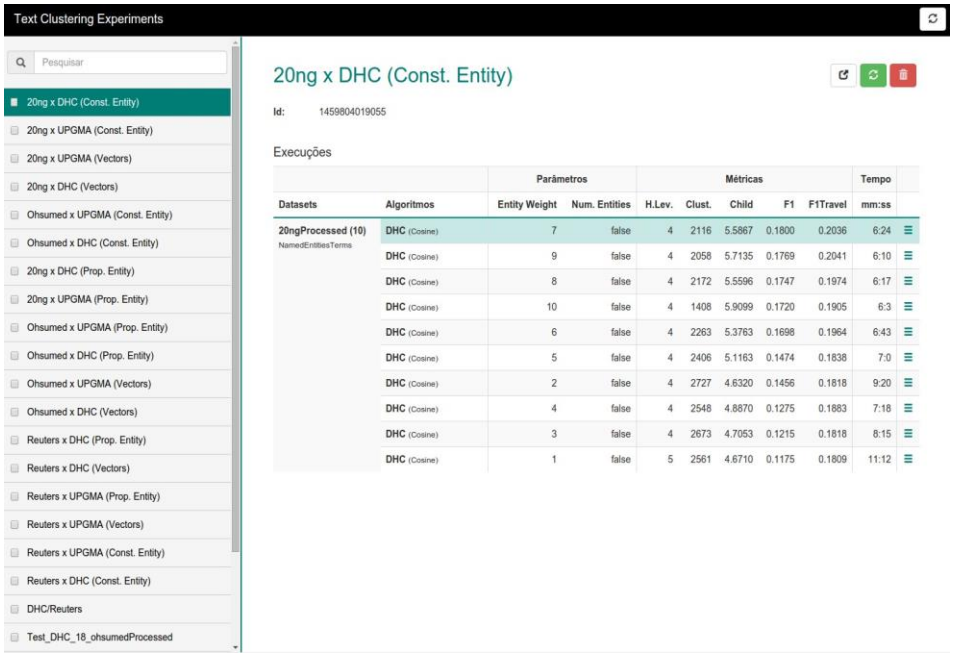
7.2 Contribuições

Este trabalho apresentou três principais contribuições. A primeira delas consiste no levantamento e comparativo dos trabalhos relacionados ao problema de agrupamento hierárquico e incremental de documentos, apresentado do Capítulo 4.

A segunda contribuição consiste na proposta de duas novas abordagens de incorporação de entidades nomeadas no processo de agrupamento de documentos (Seções 3.3 e 3.4), assim como o comparativo destas com outras abordagens já existentes na literatura (Capítulo 6).

A terceira contribuição consiste na construção de um ambiente de execução de experimentos (Seção 5.2), que viabilizou a execução dos algoritmos e a análise dos resultados. A Figura 34 apresenta a tela de visualização dos experimentos, onde à esquerda são listados diversos experimentos executados, e à direita a tabela comparativa de um experimento selecionado. Já a Figura A Figura 35 apresenta a visualização de uma das hierarquias geradas nos experimentos. Tais implementações podem ser incrementadas e contribuirão para o desenvolvimento de trabalhos futuros.

Tabela 34 - Visualização dos experimentos executados



7.3 Trabalhos futuros

Com o intuito de melhor validar o uso de entidades, tem-se como trabalho futuro a implementação de outros algoritmos dentre os recentemente propostos e sua adaptação para o uso de entidades nomeadas. Testes com outros

algoritmos pode validar se o uso de entidades tem o mesmo impacto em diferentes algoritmos, assim como no algoritmo DHC e UPGMA.

Outro trabalho futuro é um estudo focado nas características das bases de dados, com o intuito de compreender as diferentes influências do uso de entidades nas diferentes bases de dados. Como exemplo, observa-se neste trabalho a diferença de resultados entre as bases de dados Reuters-21578 e 20 newgroups. Estudos com um maior número de bases de dados pode auxiliar a compreender tal comportamento.

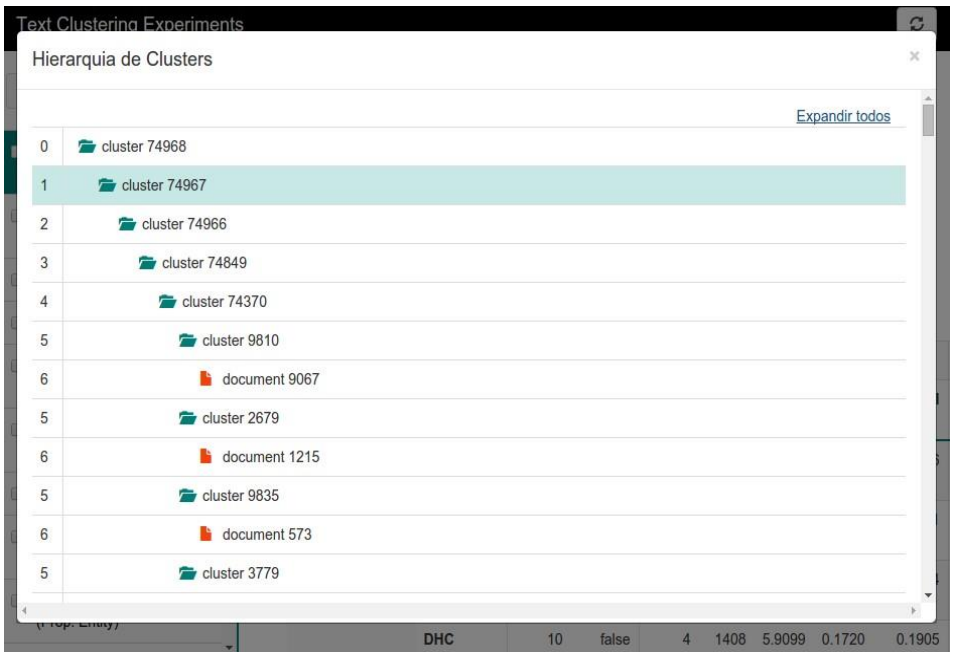


Figura 35 - Visualização da hierarquia gerada por um algoritmo de agrupamento

Mais um campo a ser explorado é a forma de cálculo de similaridade entre documentos considerando as entidades nomeadas. Neste trabalho foram testadas quatro abordagens. Porém, inúmeras outras abordagens podem ser

estudadas, como o uso de diferentes medidas de similaridade entre os termos do texto e as entidades nomeadas.

REFERÊNCIAS

AGRAWAL, R. et al. Automatic subspace clustering of high dimensional data for data mining applications. **ACM SIGMOD Record**, New York, v. 27, n. 2, p. 94-105, June 1998.

AL-ONAIZAN, Y.; KNIGHT, K. Translating named entities using monolingual and bilingual resources. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., 2002, Stroudsburg. **Proceedings...** Stroudsburg: Association for Computational Linguistics, 2002. p. 400-408.

ALONSO, A.; SUÁREZ, A.; PAGOLA, J. Acons: a new algorithm for clustering documents. In: RUEDA, L.; MERY, D.; KITTLER, J. (Ed.). **Progress in pattern recognition, image analysis and applications**. Berlin: Springer, 2007. p. 664-673. (Lecture Notes in Computer Science, 4756).

AMIGÓ, E. et al. A comparison of extrinsic clustering evaluation metrics based on formal constraints. **Information Retrieval**, Wageningen, v. 12, n. 4, p. 461-486, 2009.

APACHE Stanbol. Disponível em: <https://stanbol.apache.org>. Acesso em: 10 mar. 2016.

ASLAM, J. A.; PELEKHOV, E.; RUS, D. The star clustering algorithm for static and dynamic information organization. **Journal of Graph Algorithms and Applications**, Perugia, v. 8, n. 1, p. 95-129, 2004.

BERKHIN, P. A survey of clustering data mining techniques. In: KOGAN, J.; NICHOLAS, C.; TEBOULLE, M. (Ed.). **Grouping multidimensional data**. Berlin: Springer, 2006. p. 25-71.

CAI, R. et al. A general framework of hierarchical clustering and its applications. **Information Sciences**, New York, v. 272, p. 29-48, July 2014.

CAO, T. H.; TANG, T. M.; CHAU, C. K. Text clustering with named entities: a model, experimentation and realization. In: _____. **Data Mining:** foundations and intelligent paradigms: volume 1, clustering, association and classification. Berlin: Springer, 2012. p. 267-287.

CHERKASSKY, V.; MULIER, F. M. **Learning from data:** concepts, theory, and methods. New York: J. Wiley, 2007. 538 p.

CORREA-MORRIS, J. et al. An incremental nested partition method for data clustering. **Pattern Recognition**, Amsterdam, v. 43, n. 7, p. 2439-2455, 2010.

CUTTING, D. R. et al. Scatter/-gather: a cluster-based approach to browsing large document collections. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 15., 1992, New York. **Proceedings...** New York: ACM, 1992. p. 318-329.

DAI, X. Y. et al. Online topic detection and tracking of financial news based on hierarchical clustering. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS, 6., 2010, Qingdao. **Proceedings…** Qingdao: IEEE, 2010. v. 6, p. 3341-3346.

DEFAYS, D. An efficient algorithm for a complete link method. **The Computer Journal**, London, v. 20, n. 4, p. 364-366, 1977.

FELDMAN, R.; SANGER, J. **Text mining handbook:** advanced approaches in analyzing unstructured data. New York: Cambridge University Press, 2006. 424 p.

GAO, Z. et al. Tracking and connecting topics via incremental hierarchical dirichlet processes. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 11., 2011, Vancouver. **Proceedings…** Vancouver: IEEE, 2011. p. 1056-1061.

GIL-GARCÍA, R.; BADÍA-CONTELLES, J.; PONS-PORRATA, A. Extended star clustering algorithm. In: SANFELIU, A.; RUIZ-SHULCLOPER, J. (Ed.). **Progress in pattern recognition, speech and image analysis**. Berlin: Springer, 2003. p. 480-487. (Lecture Notes in Computer Science, 2905).

GIL-GARCÍA, R.; PONS-PORRATA, A. Dynamic hierarchical algorithms for document clustering. **Pattern Recognition Letters**, Amsterdam, v. 31, n. 6, p. 469-477, 2010a.

GIL-GARCÍA, R.; PONS-PORRATA, A. Improving the dynamic hierarchical compact clustering algorithm by using feature selection. In: IBEROAMERICAN CONGRESS CONFERENCE ON PROGRESS IN PATTERN RECOGNITION, IMAGE ANALYSIS, COMPUTER VISION, AND APPLICATIONS, 15., 2010, Berlin. **Proceedings…** Berlin: Springer Verlag, 2010b. p. 113-120.

GUHA, S. et al. Clustering data streams: theory and practice. **IEEE Transactions on Knowledge and Data Engineering**, Piscataway, v. 15, n. 3, p. 515-528, Mar. 2003.

HAN, X. W.; ZHAO, T. J. An evaluation method for clustering quality and its application. **Journal of Harbin Institute of Technology**, Ontario, v. 41, n. 11, p. 52, Dec. 2009.

HANSEN, P.; JAUMARD, B. Cluster analysis and mathematical programming. **Mathematical Programming**, Berlin, v. 79, n. 1/3, p. 191-215, 1997.

HERSH, W. et al. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In: CROFT, B.; RIJSBERGEN, C. van (Ed.). **SIGIR' 94**. London: Springer, 1994. p. 192-201.

HOFFART, J. et al. Robust disambiguation of named entities in text. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2011, Stroudsburg. **Proceedings…** Stroudsburg: Association for Computational Linguistics, 2011. p. 782-792.

HUANG, A. Similarity measures for text document clustering. In: NEW ZEALAND COMPUTER SCIENCE RESEARCH STUDENT CONFERENCE, 6., 2008, Christchurch. **Proceedings…** Christchurch, 2008. p. 49-56.

HUANG, S. et al. News topic detection based on hierarchical clustering and named entity. In: INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND KNOWLEDGE ENGINEERING, 7., 2011, Tokushima. **Proceedings…** Tokushima: IEEE, 2011. p. 280-284.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, Amsterdam, v. 31, n. 8, p. 651-666, 2010.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Upper Saddle River: Prentice-Hall, 1988. 304 p.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, New York, v. 31, n. 3, p. 264-323, 1999.

KARYPIS, M. S. G.; KUMAR, V.; STEINBACH, M. A comparison of document clustering techniques. In: KDD WORKSHOP ON TEXT MINING, 2000, Minneapolis. **Proceedings…** Minneapolis: University of Minnesota, 2000. p. 1-2.

LARSEN, B.; AONE, C. Fast and effective text mining using linear-time document clustering. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1999, New York. **Proceedings…** New York: ACM, 1999. p. 16-22.

LEWIS, D. D. **Reuters-21578**. 2004. Disponível em: <http://www.dcs.gla.ac.uk-/Keith/Preface.html>. Acesso em: 10 fev. 2015.

LIN, F. R.; HUANG, Y. T.; LIAO, D. Incrementally clustering legislative interpellation documents. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCE, 45., 2012, Maui. **Proceedings…** Maui: IEEE, 2012. p. 2521-2530.

LIN, G. et al. A general approach for incremental approximation and hierarchical clustering. In: ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHM, 70., 2006, Philadelphia. **Proceedings…** Philadelphia: Society for Industrial and Applied Mathematics, 2006. p. 1147-1156.

LU, P. et al. Hspknn: an effective and practical framework for hot topic detection of internet news. In: INTERNATIONAL CONFERENCE ON COMPUTING AND CONVERGENCE TECHNOLOGY, 7., 2012, Nanjing. **Proceedings…** Nanjing, 2012. p. 888-893.

LUO, Z.; YETISGEN-YILDIZ, M.; WENG, C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. **Journal of Biomedical Informatics**, New York, v. 44, n. 6, p. 927-935, 2011.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Oakland. **Proceedings…** Oakland, 1967, p. 281-297.

MARCACINI, R. M.; HRUSCHKA, E. R.; REZENDE, S. O. On the use of consensus clustering for incremental learning of topic hierarchies. In: BRAZILIAN CONFERENCE ON ADVANCES IN ARTIFICIAL INTELLIGENCE, 12., 2012, Berlin. **Proceedings…** Berlin: Springer-Verlag, 2012. p. 112-121.

MARCACINI, R. M.; REZENDE, S. O. Incremental hierarchical text clustering with privileged information. In: ACM SYMPOSIUM ON DOCUMENT ENGINEERING, 2013, New York. **Proceedings...** New York: ACM, 2013. p. 231-232.

PENG, T.; LIU, L. A novel incremental conceptual hierarchical text clustering method using cfu-tree. **Applied Soft Computing**, New York, v. 27, p. 269-278, Feb. 2015.

PHRIDVIRAJ, M. S. B.; SRINIVAS, C.; GURURAO, C. Clustering text data streams: a tree based approach with ternary function and ternary feature vector. **Procedia Computer Science**, New York, v. 31, p. 976-984, 2014.

RENNIE, J. **20 newsgroups dataset**. 2008. Disponível em: <http://qwone.com-/~jason/20Newsgroups/>. Acesso em: 10 mar. 2015.

RIJSBERGEN, C. J. V. **Information retrieval**. 2nd ed. Newton: Butterworth-Heinemann, 1979. 224 p.

SAHOO, N. et al. Incremental hierarchical clustering of text documents. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 15., 2006, New York. **Proceedings…** New York: ACM, 2006. p. 357-366.

SCHOLKOPF, B.; SMOLA, A. J. **Learning with kernels:** support vector machines, regularization, optimization, and beyond. Cambridge: MIT Press, 2001. 644 p.

SHAWE-TAYLOR, J.; CRISTIANINI, N. **Kernel methods for pattern analysis**. New York: Cambridge University Press, 2004. 478 p.

SIBSON, R. Slink: an optimally efficient algorithm for the single-link cluster method. **The Computer Journal**, London, v. 16, n. 1, p. 30-34, 1973.

SILVA, J. A. et al. Data stream clustering: a survey. **ACM Computing Surveys**, New York, v. 46, n. 1, p. 13:1-13:31, July 2013.

SINOARA, R. A. et al. Named entities as privileged information for hierarchical text clustering. In: INTERNATIONAL DATABASE ENGINEERING & APPLICATIONS SYMPOSIUM, 18., 2014, New York. **Proceedings…** New York: ACM, 2014. p. 57-66.

WANG, D.; LI, T. Document update summarization using incremental hierarchical clustering. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 19., 2010, New York. **Proceedings…** New York: ACM, 2010. p. 279-288.

WANG, L.; SONG, H.; LIU, X. Incremental document clustering using multirepresentation indexing tree. In: INTERNATIONAL CONFERENCE ON INFORMATION SCIENCE AND ENGINEERING, 2., 2010, Hangzhou. **Proceedings…** Hangzhou: IEEE, 2010. p. 3778-3781.

XU, R.; WUNSCH, D. **Clustering**. New York: Wiley-IEEE Press, 2009. 368 p.

XU, R. et al. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, New York, v. 16, n. 3, p. 645-678, 2005.

ZHANG, K.; ZI, J.; WU, L. G. New event detection based on indexing-tree and named entity. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 30., 2007, New York. **Proceedings…**
New York: ACM, 2007. p. 215-222.

ZHAO, Y.; KARYPIS, G.; FAYYAD, U. Hierarchical clustering algorithms for document datasets. **Data Mining and Knowledge Discovery**, New York, v. 10, n. 2, p. 141-168, 2005.

APÊNDICE A - PESQUISA SISTEMÁTICA

Para o levantamento dos trabalhos desenvolvidos nos últimos anos sobre o tema tratado neste trabalho, foi utilizada uma pesquisa sistematizada, que é descrita neste capítulo. A pesquisa teve como objetivo obter trabalhos sobre técnicas de agrupamento que possuíssem algumas características específicas, sendo listados, para cada uma delas, os termos chave que as representam:

a) Técnicas de agrupamento: cluster, clustering;
b) Agrupamento hierárquico: hierarchical, hierarchy;
c) Agrupamento de documentos textuais: document, text, textual, news, topic;
d) Agrupamento Incremental: incremental, dynamic, datastream, data stream, dataflow, data flow.

A partir destes termos, foi criada uma consulta, apresentada na Figura 1, que foi aplicada às bibliotecas digitais ACM, Science Direct e IEEE Xplore. Foram utilizados como campos de busca os atributos “Título", “Resumo" e “Palavras Chaves". Foram considerados trabalhos dos anos de 2010 a 2015.

(“hierarchical"OU “hierarchy")
E
(“document"OU “text"OU “textual"OU “news"OU “topic")
E
(“cluster"OU “clustering")
E
(“incremental"OU “dynamic"OU “datastream"OU “data stream" OU “dataflow” OU “data flow”)

Figura 1 - Critérios utilizados na pesquisa sistematizada.

O objetivo desta pesquisa foi buscar trabalhos que desenvolveram novas técnicas ou abordagens de agrupamento hierárquico de textos em ambientes

dinâmicos. Porém, ao longo da busca, também foram encontrados alguns trabalhos que consistem em aplicações destas técnicas e, assim, estes também foram descritos aqui.

Efetuada a pesquisa, foram obtidos como resultado um total de 58 trabalhos, 6 da biblioteca Science Direct, 30 da ACM e 22 da IEEE Xplore. Dentre estes trabalhos, 10 foram trabalhos repetidos, encontrados em mais de uma biblioteca. Após uma breve análise dos trabalhos, um subconjunto de 15 foi selecionado, contendo aqueles que estão relacionados ao tema, sendo 10 deles trabalhos relacionados ao desenvolvimento de novas técnicas de agrupamento e 5 deles aplicações das técnicas em algum contexto. A Tabela 1 apresenta o número de trabalhos obtidos por biblioteca digital e a Tabela 2 apresenta o número de trabalhos selecionados.

Tabela 1 - Trabalhos obtidos através da pesquisa sistematizada.

Biblioteca Digital	Trabalhos
Science Direct	6
ACM	30
IEEE Xplore	22
Total de trabalhos	58

Tabela 2 - Trabalhos selecionados.

Tipo de trabalho	Trabalhos
Técnicas de Agrupamento	10
Aplicações	5
Total de Trabalhos	15

A Figura 2 apresenta um gráfico com a contabilização total dos trabalhos selecionados para cada ano do período considerado. É importante

ressaltar que esta pesquisa foi aplicada às bibliotecas digitais no mês de janeiro de 2015 e, caso seja repetida futuramente, outros trabalhos podem vir a ser encontrados.

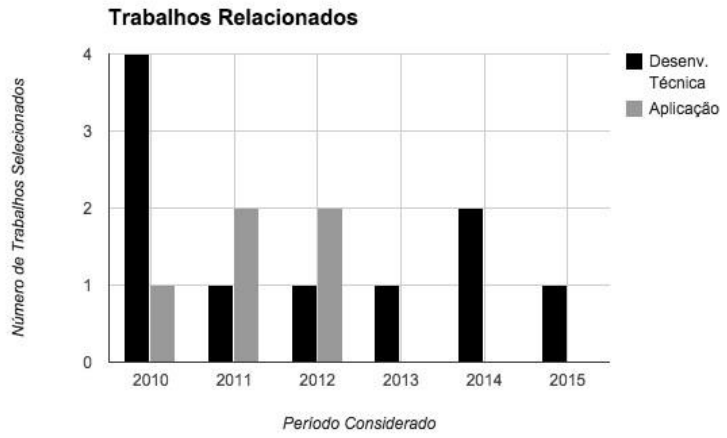


Figura 2 - Trabalhos relacionados por ano.

Após a obtenção e seleção dos trabalhos, algumas informações foram extraídas dos mesmos e comparadas. Estas comparações foram efetuadas apenas para trabalhos cujo objetivo é o desenvolvimento de novas técnicas de agrupamento, sendo as seguintes informações coletadas:

a) Medidas de similaridade e distância utilizadas;
b) Bases de dados utilizadas nos experimentos;
c) Métricas utilizadas para a avaliação dos algoritmos;
d) Algoritmos utilizados em comparações.

Detalhes sobre os trabalhos encontrados através deste processo são apresentados na Seção 4.

APÊNDICE B - OUTRAS MÉTRICAS DE AVALIAÇÃO DOS AGRUPAMENTOS

Este capítulo apresenta duas métricas de avaliação dos agrupamentos: *FCubed* e *HF*1. Tais métricas foram estudadas ao longo da execução do projeto, porém, optou-se por não utilizá-las nos experimentos.

B.1 FCubed

A métrica *FCubed* (GIL-GARCÍA; PONS-PORRATA, 2010a) é apropriada para situações onde pode haver intersecções entre os grupos gerados. Se dois documentos pertencem a duas classes em comum, em um agrupamento ideal espera-se que eles também estejam contidos juntos em dois grupos. Assim, esta métrica busca avaliar o quanto os grupos gerados atendem a esta expectativa. Ela consiste uma adaptação da métrica *F-measure* utilizando a *Precision BCubed* e *Recall BCubed* (AMIGÓ et al., 2009). Primeiramente é calculado o valor de *Multiplicity Precision* (EQUAÇÃO B.1) e *Multiplicity Recall* (EQUAÇÃO B.2). Nas equações seguintes devese considerar que, para dois documentos d_i e d_j, L_{ij} é o conjunto de classes que contém ambos os documentos e C_{ij} é o conjunto de grupos que contém ambos.

$$MP(d_i, d_j) = \frac{min(|C_{ij}|,|L_{ij}|)}{|C_{ij}|} \quad \text{(B.1)}$$

$$MR(d_i, d_j) = \frac{min(|C_{ij}|,|L_{ij}|)}{|L_{ij}|} \quad \text{(B.2)}$$

Assim, o valor de *Precision BCubed* (EQUAÇÃO B.3) é dado pela média de $MP(d_i,d_j)$ para todos os documentos d_i e d_j, sendo *Recall BCubed* (EQUAÇÃO B.4) calculado de forma similar utilizando $MR(d_i,d_j)$. O valor de de *FCubed* é dado pela equação B.5.

$$Precision\ BCubed(d_i, d_j) = \text{Avg}_{d_i}\left[\text{Avg}_{d_j, C_{ij} \neq \emptyset}\left[MP(d_i, d_j)\right]\right] \quad \text{(B.3)}$$

$$Recall\ BCubed(d_i, d_j) = \text{Avg}_{d_i}\left[\text{Avg}_{d_j, C_{ij} \neq \emptyset}\left[MR(d_i, d_j)\right]\right] \quad \text{(B.4)}$$

$$FCubed = \frac{2 \times Precision\ BCubed \times Recall\ BCubed}{Precision\ BCubed + Recall\ BCubed} \quad \text{(B.5)}$$

Pode-se observar nesta métrica que, quando maior o valor de *FCubed*, mais o algoritmo está conseguindo capturar a relação entre documentos que compartilham as mesmas classes, inserindo-os em nos mesmos grupo.

B.2 HF1

HF1 é uma métrica proposta por Gil-García e Pons-Porrata (2010a) que é baseada na *F-measure* e é apropriada para agrupamentos hierárquicos, pois considera a hierarquia gerada em seus cálculos.

Considerando a hierarquia composta por N tópicos manualmente classificados, sua estrutura pode ser representada de acordo com ancestrais de cada tópico, na forma $TH = \{(t_1, a_1^1), \ldots, (t_1, a_{k_1}^1), \ldots, (t_N, a_1^N), \ldots, (t_N, a_{k_n}^N)\}$ sendo que $A(t_i) = \{a_1^i, \ldots, a_{k_i}^i\}$ é o conjunto dos K_i ancestrais do tópico t_i. De forma similar, a hierarquia de grupos gerada por um algoritmo pode ser representada por $CH = \{(c_1, a_1^1), \ldots, (c_1, a_{m_1}^1), \ldots, (c_M, a_1^M), \ldots, (c_M, a_{m_M}^M)\}$ sendo c_i um dentre os M grupos gerados e $A(t_i) = \{a_1^i, \ldots, a_{m_i}^i\}$ o conjunto de seus ancestrais na hierarquia.

Pode-se definir então o valor de $\sigma(t_i)$ (EQUAÇÃO B.6), que consiste em um grupo $c_i \in C$ que retorna o maior valor de $F(c_i,t_i)$. É importante observar que $\sigma(c_i)$ é um grupo e não um valor.

$$\sigma(t_i) = \arg\max_{c_i \in C} F(c_i, t_i) \quad \text{(B.6)}$$

É ainda necessário definir o valor de *CP*, utilizado nas equações seguintes, que consiste no número de pares da hierarquia de tópicos *TH* corretamente identificados na hierarquia de grupos em *CH*. Dado um par de tópicos $(t_i,t_j) \in TH$, este par é considerado corretamente identificado se existe o par de grupos $(\sigma(t_i),\sigma(t_j)) \in CH$. Em outras palavras, se dois tópicos possuem uma relação de ancestralidade, seus grupos correspondentes também devem possuir esta relação.

Assim, o valor de *HF*1 (EQUAÇÃO B.9) é definido em função de *HPrecision* (EQUAÇÃO B.7) e *HRecall* (EQUAÇÃO B.8).

$$HPrecision = \frac{CP}{|CH|} \tag{B.7}$$

$$HRecall = \frac{CP}{|TH|} \tag{B.8}$$

$$HF1 = \frac{2 \times HPrecision \times HRecall}{HPrecision + HRecall} \tag{B.9}$$

Esta métrica difere das demais por ser específica para agrupamentos hierárquicos, e buscar avaliar a relação de ancestralidade entre os grupos na hierarquia, e não apenas se os documentos foram corretamente agrupados. Um ponto importante a ser observado é que, para o uso desta métrica, é necessário que os dados previamente classificados utilizados para teste estejam organizados de forma hierárquica (hierarquia de tópicos).