Novos Algoritmos e Métodos de Avaliação Objetiva para Modelos de Rotulação em Agrupamentos Hierárquicos de Documentos

Maria Fernanda Moura, Solange Oliveira Rezende

[1]Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041- 13083-886 - Campinas, SP
Laboratório de Inteligência Computacional
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Av. Trabalhador São-carlense, 400 – Centro
Caixa Postal: 668 – CEP: 13560-970 – São Carlos – SP – Brazil

{mnanda,solange}@icmc.usp.br

Resumo. *Um novo método de rotulação para agrupamentos hierárquicos de documentos e uma proposta de validação objetiva do método são apresentados neste trabalho. O objetivo do método é produzir rótulos discriminativos dos grupos, sem repetição ao longo dos ramos, com um vocabulário variado e de qualidade, independente da língua e do algoritmo de agrupamento utilizado. Para isso o método foi baseado em uma definição formal de uma taxonomia de tópicos, também apresentada no trabalho. O método foi implementado e validado contra três outros métodos da literatura. A validação passa pelos critérios de complexidade do algoritmo, melhor discriminação dos grupos, melhor variabilidade e qualidade do vocabulário gerado; em todos esses quesitos o novo método supera estatisticamente os outros três.*

Palavras-chaves*: agrupamento hierárquico de documentos, rotulação de agrupamentos hierárquicos, taxonomia de tópicos, descritores de agrupamentos.*

1. Introdução

Encontrar tópicos em coleções de textos tem sido uma prática utilizada em aplicações voltadas para a recuperação de informações textuais, como a geração de indexadores para máquinas de busca, ou mesmo a própria apresentação de resultados de busca organizados em grupos mais significativos como os da ferramenta Vivisimo. Segundo Konchady [2006], um número considerável de propostas explora o agrupamento aglomerativo de documentos sob a perspectiva da recuperação de informação, mas um aspecto igualmente importante é o processo de rotular os grupos obtidos. O usuário final pode decidir explorar um agrupamento geralmente guiado pelo rótulo do mesmo; porém, resumir um agrupamento com base em um pequeno conjunto de termos é bastante difícil. Pois, a questão de abstração de dados em problemas de agrupamento envolve obter descrições concisas e significativas para os grupos e essas descrições chegam a ser tão importante quanto o próprio agrupamento [**?**].

Há vários trabalhos que exploram especialmente a geração de rótulos para o agrupamento de documentos textuais, que visam à identificação de palavras-chaves para indicar os possíveis tópicos aos quais os documentos agrupados se refiram. Em geral,

esses métodos dependem diretamente da forma como o agrupamento é obtido, isto é, do método utilizado para agrupar os documentos. Os agrupamentos são calculados a partir da representação matricial da coleção de textos, considerando-se cada texto um elemento a ser agrupado representado por um vetor de atributos. Cada atributo corresponde a uma palavra ou composição de palavras (por exemplo, "inteligência artificial"), para as quais é obtida alguma medida de freqüência, que pode ser qualitativa, representando a pertinência ou não da palavra no documento, ou quantitativa, como a freqüência ou a freqüência inversa de ocorrência da palavra. Os agrupamentos obtidos refletem tópicos ou sub-tópicos aos quais os documentos se referem, logo, são representados por conjuntos de atributos mais significativos no grupo, que podem ser interpretados como um conceito associado ao grupo. Um conjunto de bons descritores para um grupo consiste de um número pequeno de termos que precisamente distingam o grupo dos demais. Algumas possibilidades automáticas, segundo Feldman [2007], também são: o título do documento do medóide ou alguns títulos de documentos; as cinco ou dez palavras mais freqüentes do centróide; ou, provavelmente, uma melhor escolha seja uma frase descritiva, se for possível encontrá-la. Alguns métodos comuns de rotulação de agrupamentos baseiam-se na busca por frases comuns aos documentos sob algum grupo e que sejam de alguma forma únicas na coleção.

Dessa forma, um dos principais pontos de investigação nessa problemática é como rotular os tópicos das coleções de textos aglomerativamente organizadas. A organização hierárquica de uma coleção de textos seguida de um procedimento de rotulação tem como objetivo organizar a informação, atribuir metadados únicos a cada grupo da hierarquia e auxiliar o usuário em navegação e busca, de modo que o usuário possa melhor compreender a coleção de dados. Para isso, vem sendo desenvolvido o projeto TopTax - *Topic Taxonomy*([**?**] e *http://labic.icmc.usp.br/projects*), onde se aplica mineração de textos à organização de coleções textuais por meio de diferentes técnicas, experimental e individualmente testadas. Assim, nesse projeto, a rotulação de grupos foi separada da obtenção dos mesmos, com o objetivo de se verificar diferentes métodos, tentando não perder informação e compreensibilidade. Além disso, com o processo separado da rotulação, a técnica pode ser aplicada a qualquer hierarquia, tenha ela sido obtida de qualquer procedimento de agrupamento hierárquico, inclusive manual. Assim, este relatório trata exclusivamente da rotulação dos grupos.

Na próxima seção é descrito o desenvolvimento do trabalho, o método que o inspirou e novas propostas sobre o mesmo. Depois é discutida uma forma de avaliação objetiva do modelo proposto comparando-o a três pré-existentes. Na quarta seção discorre-se sobre os experimentos realizados, apresentando dados e resultados. E, finalmente na quinta seção são resumidas as conclusões e algumas idéias de trabalho futuro.

2. Métodos de Rotulação

Esta seção tem início com a definição de alguns termos utilizados no trabalho, a fim de evitar possíveis confusões, embora as definições formais sejam colocadas sempre que necessárias ao longo do desdobramento da seção. A seguir explica-se a idéia geral do método desenvolvido, seu embasamento na literatura e suas evoluções. Os algoritmos pré-existentes e os novos, desenvolvidos neste trabalho, são apresentados na seqüência, bem como o método utilizado para a validação objetiva.

2.1. Definição de Termos e Restrições

Segundo Aristóteles, taxonomia é a ciência de classificar coisas e engloba a noção de categorias e subcategorias, bem como conceitos e sub-conceitos; que são idéias que melhor se aproximam daquelas de sistemas orientados a objetos (veja Cimiano [2006] para detalhes). Nosso principal objetivo, neste trabalho, é chegar a uma taxonomia de tópicos, organizando uma coleção de documentos de tal forma que sua compreensão possa ser melhorada. Ou seja, a coleção de textos será agrupada em categorias e subcategorias, de modo que se obtenha descritores de cada categoria, e, para os mesmos, utilize-se a noção de herança entre os objetos, isto é, uma sub classe ou sub categoria herda os descritores da sua classe ou categoria. Para descrever o trabalho desenvolvido, a seguinte terminologia e restrições são aplicadas à coleção de documentos e seus possíveis componentes:

- **coleção de textos**: assume-se que a coleção é de um domínio específico de conhecimento. O método trata a coleção como "*bag of words*", logo ele não é capaz de tratar possíveis polissemias em diferentes contextos, dado que o contexto não é conhecido. Polissemias são termos com grafia semelhante e significados muito independentes em diferentes contextos, por exemplo, a fruta *manga* e a parte de uma camisa também com o nome *manga*. Logo, restringir o domínio diminui a probabilidade de ocorrência de polissemias na coleção.
- **termo**: o modelo de geração de termos identifica palavras simples como "*tokens*" e aplica um processo de radicalização sobre as palavras, removendo-lhes sufixos ("*stemming process*"), com o uso de uma adaptação do algoritmo de Porter para inglês, português e espanhol [**?**]. O processo pode gerar apenas palavras simples e/ou multi-palavras; aqui tratados como unigramas e n-gramas. Por exemplo, um termo pode ser "artific", "intelligenc", "conferenc", "artific intelligenc".
- **"*stopwords*"**: sempre é utilizada a lista habitual de *stopwords*, que costuma englobar os artigos, advérbios, expressões adverbiais e outras classes gramaticais que são muito comuns ou muito raras nos textos e que, conseqüentemente, possam ser estatisticamente descartadas do conjunto de atributos.
- **atributos**: a palavra se refere ao conjunto de termos selecionados da coleção de documentos após a remoção das *stopwords* e aplicação de algum filtro, na etapa de pré-processamento dos dados. O conjunto de atributos obtido é a primeira estimativa do vocabulário.
- **vocabulário**: o conjunto completo de termos que é usado na construção da taxonomia de tópicos. Logo, o vocabulário corresponde à união de todos os conjuntos de descritores de cada agrupamento. Esse vocabulário pode ser considerado como uma estimativa do vocabulário do domínio; estimativa que será melhor ou pior dependendo da representatividade[1] da coleção em relação ao domínio de conhecimento.
- **rótulo**: um conjunto de termos selecionados que é único e cujos elementos correspondem aos termos mais discriminativos para um agrupamento, segundo um dado critério estatístico.
- **taxonomia de tópicos**: uma coleção de textos, de um dado domínio de conhecimento, hierarquicamente agrupada a partir de uma função de similaridade entre seus documentos, considerado o conjunto de atributos, com cada grupo rotulado de acordo com uma estimativa de vocabulário do domínio.

[1] Neste trabalho, a representatividade da coleção no domínio não é validada.

2.2. Idéia Geral da Rotulação

Vamos usar uma coleção de seis textos, construída propositadamente para ilustrar as idéias aqui apresentadas, com os termos todos em inglês. Os textos foram cuidadosamente construídos para cobrir tópicos relacionados à piscicultura ("pisciculture"), com duas espécies de peixes do Pantanal, e tópicos sobre produção e reprodução de gado de corte ("beef catle"). Os textos foram criados para ter em comum o termo Pantanal e apenas termos do domínio foram utilizados nos textos; isto é, termos que foram obtidos de um *thesaurus*. Adicionalmente, procurou-se distribuir as freqüências dos termos mais ou menos uniformemente ao longo dos textos. Conseqüentemente, essa coleção de textos é um exemplo ideal, pois após obter a estrutura hierárquica deve-se chegar a uma hierarquia de tópicos basicamente sem erros, pois os termos selecionados supostamente são os melhores, dado que são termos de domínio e estão balanceadamente distribuídos ao longo dos textos.

Após aplicar um método de agrupamento hierárquico aglomerativo, usando similaridade de cosseno e o algoritmo "*average linkage*", espera-se que os textos sejam separados por tópicos e sub-tópicos de acordo com os respectivos assuntos e todos agrupados sob um tópico raiz que contém o termo "Pantanal" em seus descritores; como ilustrado na Figura 1. Nessa figura, para todos os nós é representada a freqüência acumulada de cada termo em cada grupo de documentos.

Dada a hierarquia da Figura 1, se os rótulos de cada grupo forem obtidos com o método de seleção dos termos mais freqüentes em cada grupo, os termos com maior freqüência acumulada nos agrupamentos são mantidos como rótulos após a determinação de quantos termos deverão compor cada conjunto de rótulos. Assim, se usarmos, no máximo, sete dos termos em cada conjunto, tem-se o resultado apresentado na Figura 2, parte A. Deve-se ressaltar que esse método é bastante usado em agrupamentos não hierárquicos e disjuntos ("*flat clusters*"), para os quais apresenta bons resultados. Porém, em uma hierarquia esse método introduz uma série de repetições desnecessárias de termos ao longo dos ramos, dado que os rótulos dos nós ascendentes também se aplicam aos seus descendentes. Logo, uma boa idéia é evitar essas repetições desnecessárias, como proposto por Moura et al [2008d] e ilustrado na parte B da Figura 2.

A proposta de Moura et al [2008-d] foi baseada nas idéias de Popescul e Ungar [2000]. Vamos observar uma hierarquia genérica como a representada na Figura 3. Supõe-se que a hierarquia foi obtida de um método de agrupamento hierárquico sobre uma coleção de documentos, $D = \{d_1, ..., d_x\}$, com x documentos, representados por medidas sobre um conjunto de atributos, $A = \{a_1, ..a_m\}$, com m atributos. Assim, cada nó n_i na hierarquia corresponde a um grupo com c filhos e é formado por uma sub-coleção de D. Deste modo, $f_i(a_k)$ é definida como a freqüência acumulada do k-ésimo atributo de A em n_i; ou similarmente, essa quantidade corresponde à freqüência acumulada de a_k nos documentos de n_i. Para tomar uma decisão sobre quais atributos melhor discriminam o i-ésimo grupo, para cada a_k, $k = 1, ..m$, em um determinado n_i, constrói-se uma tabela de contingência das freqüências dos termos em cada filho de n_i, ilustrada na Tabela 1.

Para decidir se o termo a_k discrimina somente o nó ascendente, nó pai n_i, ou somente um dos filhos ou todo o conjunto de filhos, $\{n_{i1}, .., n_{ic}\}$, um teste de independência é aplicado sobre a distribuição do termo nos filhos. Sob a hipótese de independência, isto é, o termo fixado a_k não depende dos nós filhos, cada $f_{ij}(a_k)$ depende exclusivamente das

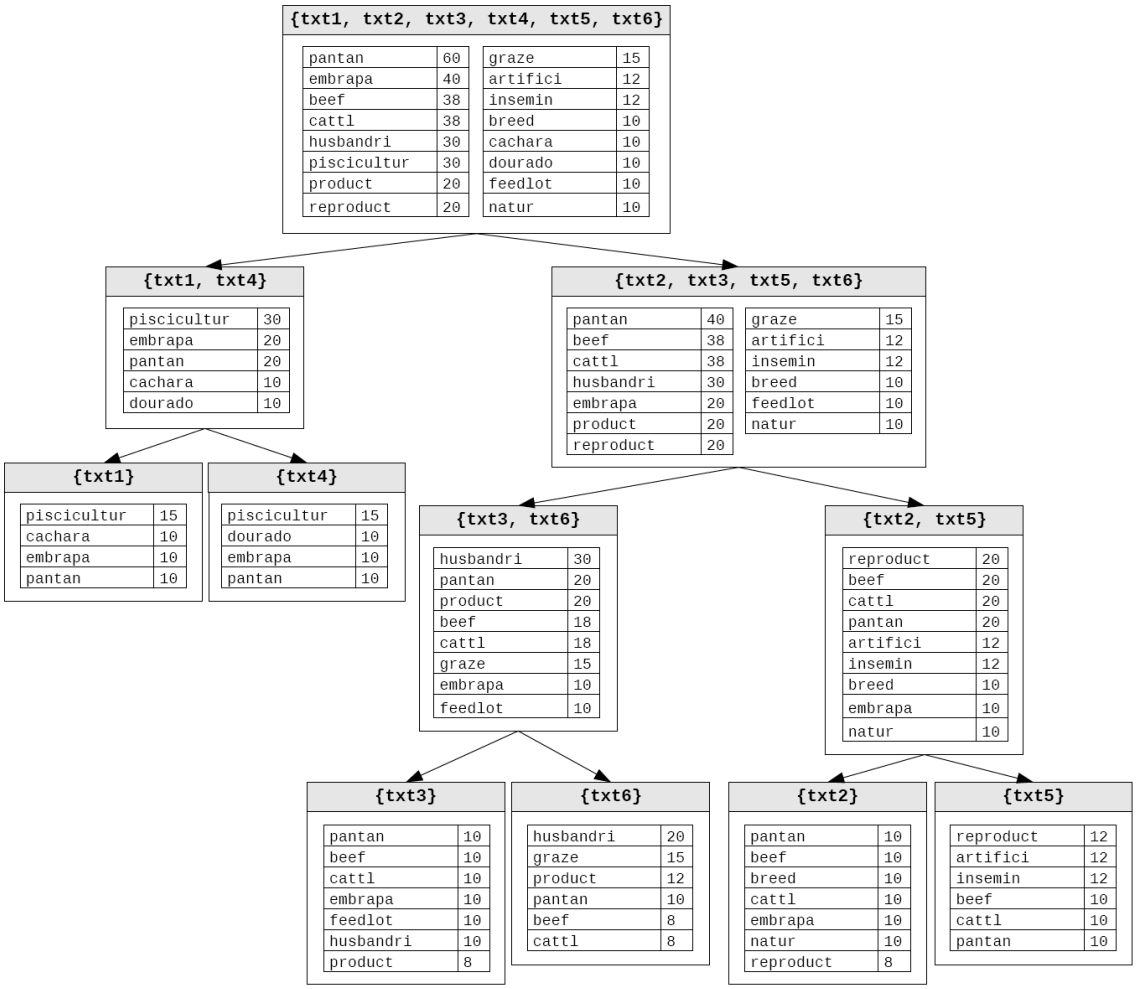

Figure 1. Hierarquia inferida sobre a coleção de textos do exemplo ideal

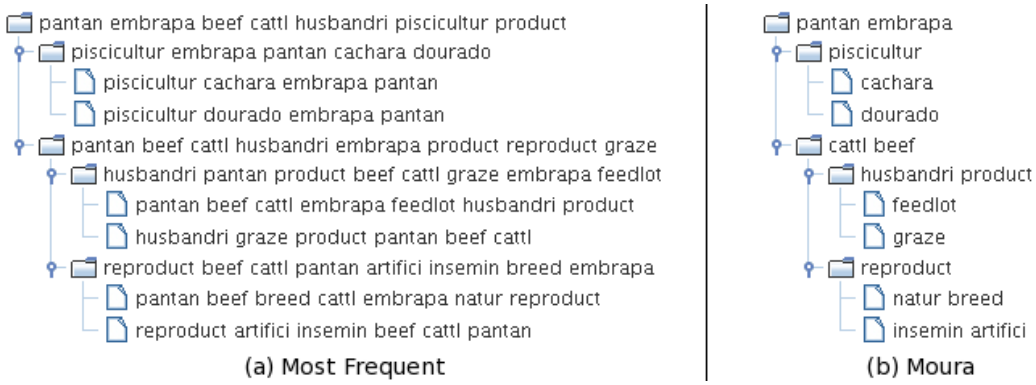

Figure 2. Ilustração dos métodos de rotulação Mais Freqüentes e Moura *et al* (2008-d)

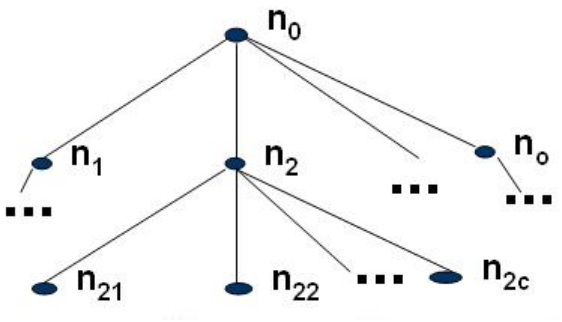

Figure 3. A general hierarchy

$filho$	a_k	$!a_k$	total
n_{i1}	$f_{i1}(a_k)$	$\sum_{t=1,t\neq k}^{m} f_{i1}(a_t)$	$\sum_{t=1}^{m} f_{i1}(a_t)$
...	...	...	...
n_{ic}	$f_{ic}(a_k)$	$\sum_{t=1,t\neq k}^{m} f_{ic}(a_t)$	$\sum_{t=1}^{m} f_{ic}(a_t)$
$total$	$f_i(a_k)$	$\sum_{t=1,t\neq k}^{m} f_i(a_t)$	$\sum_{t=1}^{m} f_i(a_t)$

Table 1. Tabela de Contingência para $f_i(a_k)$

freqüências marginais, isto é, o valor esperado para cada $f_{ij}(a_k)$ deve ser:

$$E(f_{ij}(a_k)) = f_i(a_k) \times \sum_{t=1}^{m} f_{ij}(a_t) \times \frac{1}{\sum_{t=1}^{m} f_i(a_t))} \quad (1)$$

Se o termo independe de todos os nós filhos, considera-se que ele pertence ao conjunto de rótulos genéricos do nó pai n_i, caso contrário, ele é considerado como pertencente aos conjuntos de rótulos específicos de cada nó filho com o qual a relação de dependência tenha sido verificada. Esse teste é aplicado ao longo de toda a hierarquia, para cada atributo, iniciando-se na raiz e descendo-se até as folhas.

O método de Moura et al [2008-d], a partir daqui tratado por RLDM, para ***Robust Labelling Down Method***, e o de Popescul e Ungar [2000] implementam essas idéias, tendo como principais diferenças os testes utilizados para verificar a hipótese de independência e o tratamento das restrições de aplicação dos testes. Popescul e Ungar usaram o estimador χ^2 com a restrição de todos $E(f_{ij}(a_k)) \geq 5$ bem como todas as freqüências observadas na tabela ≥ 5. Em uma série de exemplos, esse método não foi capaz de tomar decisões sobre os atributos, devido às restrições, e, conseqüentemente, acabou se aproximando muito dos resultados obtidos com o método dos Mais Freqüentes, comentado e ilustrado anteriormente. O RLDM propunha resolver esses problemas e o fez.

2.3. O RLDM

O RLDM é detalhado em Moura et al [2008-d]. Nesse método, casos onde algum $f_{ij}(a_k) \approx 0$ são tratados separadamente, como casos de associação ou dissociação completa do $k-$ésimo termo a algum grupo. Deste modo o método permite que seu algoritmo sempre tome uma decisão em relação a algum a_k em algum n_i e, conseqüentemente, os algoritmos sempre produzem um conjunto pequeno de rótulos em cada grupo e não replicam os termos ao longo da hierarquia; como pode ser observado na Figura 2, parte B.

Adicionalmente, são utilizados os estimadores $Q\ de\ Yule$ e o U^2 para testar as hipóteses de independência, por serem mais robustos que o χ^2 (para maiores detalhes veja [**?**]).

Assim, em tabelas 2×2, isto é, quando algum n_i tem somente dois filhos, $c = 1, 2$; o estimador escolhido é o $Q\ de\ Yule$. Para testar a hipótese de associação usando o $Q\ de\ Yule$, a razão do produto cruzado é calculada:

$$\alpha = (f_{i1}(a_k) * \sum_{t=1,t\neq k}^{m} f_{i2}(a_t)) \ / \ (f_{i2}(a_k) * \sum_{t=1,t\neq k}^{m} f_{i1}(a_t))$$

E, então, estima-se Q [2] [**?**]:

$$\widehat{Q} = \frac{\alpha - 1}{\alpha + 1},$$

$$\widehat{\sigma}_Q = \frac{1}{2} * (1 - \widehat{Q}^2) * \sqrt{\frac{1}{f_{i1}(a_k)} + \frac{1}{\sum_{t=1,t\neq k}^{m} f_{i2}(a_t)} + \frac{1}{f_{i2}(a_k)} + \frac{1}{\sum_{t=1,t\neq k}^{m} f_{i1}(a_t)}}$$

$$\widehat{Q} \approx N(\widehat{Q}, \widehat{\sigma}_Q) \Rightarrow \widehat{Q} \pm 2 * \widehat{\sigma}_Q$$

O valor máximo da função é alcançado quando $\widehat{\alpha} = 1$ e $\widehat{Q} = 0$ e, $\widehat{Q} = 1$ ou $\widehat{Q} = -1$ ocorrem quando algum $f_{ij} = 0$; assim, se o valor $0 \in \left[\widehat{Q} - 2 * \widehat{\sigma}, \widehat{Q} + 2 * \widehat{\sigma}\right]$ então a hipótese de independência ou desassociação é verdadeira.

A expansão do teste mais robusto para tabelas de contingência $c \times 2$, isto é, n_i pode ter qualquer número de filhos, ($c \geq 3$), usa o estimador U^2:

$$U^2 = \frac{(c-1) * BSS}{TSS},$$

$$BSS = TSS - WSS$$

$$TSS = (c/2) - (1/2c) \sum_{j=1}^{c} (f_{ij}(a_k))^2$$

$$WSS = (c/2) - (1/2) \left(f_i(a_k) + \sum_{t=1,t\neq k}^{m} f_i(a_t) \right) \times \sum_{j=1}^{c} (f_{ij}(a_k))^2$$

O valor TSS é interpretado como a variância total na tabela, ou a dispersão total entre os valores. O valor WSS é a variação dos filhos dentro da classe, sendo que a classe é positiva quando o termo está presente no nó filho. O valor BSS é a variância entre as classes. Assim, U^2 é um estimador da redução da proporção da variância explicada pelos dados, isto é, a variação da freqüência do termo e, assintoticamente, aproxima-se de uma distribuição χ^2 com $(c-1)$ graus de liberdade (veja [**?**] para detalhes), embora não dependa da probabilidade de distribuição das freqüências.

Os algoritmos 1 e 2 implementam a idéia básica de percorrer a árvore da raiz para as folhas, para cada atributo a_k, $k = 1, ..., m$, realizando os testes de hipótese e decidindo

[2] Toda estimativa é notada com um chapéu.

Input: n_i, o nó raiz precisa ser o primeiro
Output: *hierarquia atualizada*, com seus conjuntos de rótulos
for *all* a_k*, k=1,..,m* **do**
| h executa testeRLDM(n_i,a_k);
end
execute CombinaSimplifica(n_i);

Algorithm 1: Algoritmo para controlar o RLDM.

Input: n_i: i-ésimo grupo; a_k: k-ésimo atributo
for *all filhos de* n_i **do**
| verifica quantos filhos possuem a_k;
end
if *quantos filhos == 1* **then**
| executa testeRLDM(esse filho,a_k);
end
if *quantos filhos == 2* **then**
| calcula Q de Yule e testa a hipótese;
| **if** *independe dos filhos* **then**
| | fica no pai; remove dos filhos;
| **end**
| **else**
| | **if** *depende de um filho* **then**
| | | fica nesse filho; remove do pai e do irmão;
| | **end**
| | **else**
| | | remove do pai; executa testeRLDM(primeiro filho,a_k); executa
| | | testeRLDM(segundo filho,a_k);
| | **end**
| **end**
end
if *quantos filhos* ≥ 3 **then**
| calcula U^2 e testa a hipótese;
| **if** *independe dos filhos* **then**
| | fica no pai; remove dos filhos;
| **end**
| **else**
| | verifica de quais filhos independe; remove do pai e dos irmãos;
| | **for** *all filhos onde há dependência* **do**
| | | executa testeRLDM(filho,a_k);
| | **end**
| **end**
end

Algorithm 2: Algoritmo testeRLDM, testa independência

```
Input: hierarquia
Output: hierarquia atualizada
for das últimas folhas até a raiz do
    if nó tem conjunto de rótulos vazio then
        if nó tem filhos then
            seu pai recebe os seus filhos; remove o nó;
        end
    end
end
```

Algorithm 3: Algoritmo para realizar o corte semântico no RLDM

para quais nós, ou grupos, o atributo é ou não discriminativo. Deve-se notar que com esses passos sendo seguidos no processo decisório, não sobram replicações de termos ao longo de um mesmo ramo da hierarquia.

Ainda, como para cada atributo sempre existe uma decisão, em alguns casos, o conjunto de rótulos de um grupo resulta vazio. Nesses casos, não foi encontrado um atributo especificamente discriminativo para o grupo, isto é, ele apenas herda os rótulos discriminativos de seus antecessores. Isso ocorre porque a coleção de textos é limitada, bem como o conjunto de atributos. Logo, nem sempre há atributos que representam bem cada grupo, independentemente. Então, esses grupos são cortados da hierarquia e, os descendentes desses grupos passam a ser parte de seu primeiro antecessor rotulado. Para realizar esse corte semântico da hierarquia, após a aplicação dos algoritmos anteriores a toda a árvore, aplica-se o algoritmo 3.

Os métodos de Popescul e Ungar [2000] e de Moura et al [2008-d] definem algoritmos com alta complexidade computacional, da ordem de $O(m * z^2)$, considerando-se os m atributos de A e que a árvore tenha z nós. Primeiramente, esses métodos precisam acumular as freqüências de cada termo ao longo da hierarquia. Depois, percorrem a hierarquia da raiz para as folhas para cada termo. Em cada nó eles testam a discriminação do termo e decidem se o termo discrimina o nó ou ramos de seus filhos. Se o termo discrimina apenas o nó n_i ou algum n_{ij}, ele é removido de todas as listas de termos dos descendentes deste, antes de se passar ao teste de um outro termo. Além disso, o método RLDM ainda propõe o corte semântico sobre a hierarquia, e esse processo é aplicado sobre toda a hierarquia novamente, o que incrementa um pouco a sua complexidade passando a $O(m * z^2 + z)$.

Ainda, ao usar as freqüências acumuladas em cada nó e a aplicação dos testes da raiz para as folhas, um erro pode ser inserido diversas vezes no processo de decisão. Observando a Figura 2, parte B, o termo $Embrapa$ foi considerado discriminativo na raiz da hierarquia, mas de fato ele não é presente em todos os nós da hierarquia, o que pode ser facilmente verificado observando-se a Figura 1. Deve-se salientar que, como todos esses métodos são métodos estatísticos, o erro é previsto e faz parte dos resultados; pode-se, porém, tentar minimizá-lo.

2.4. O RLUM

O método aqui proposto, RLUM para ***Robust Labelling up Method***, utiliza as idéias de Popescul e Ungar e do RLDM [**?**], mas foca na redução da complexidade computacional

e em uma boa definição dos rótulos para construção da taxonomia de tópicos. Mais precisamente, o foco é uma bem definida estimativa dos rótulos da taxonomia de tópicos; pois, os rótulos identificados são conjuntos de possíveis termos de domínio, ou seja, são aproximações ou estimativas desses termos.

Assim, o conjunto de termos que define o rótulo específico L_s de um nó n_i é definido como:

$$L_s(n_i) = \{\forall\ a_k\ /\ I\ is\ True\ \wedge\ II\ is\ true\}$$

$$I : f_{ij}(a_k)\ >\ 0,\ \forall\ j = 1, ..., c$$

$$II : E(f_{ij}(a_k)) = f_i(a_k) \times \sum_{t=1}^{m} f_{ij}(a_t) \times \frac{1}{\sum_{t=1}^{m} f_i(a_t))}$$

$$\forall\ j = 1, ..., c$$

Adicionalmente,

$$L_s(n_{ij}) = \{\forall\ a_k\ /\ I\ is\ True\ \wedge\ II\ is\ false\}$$

Conseqüentemente, todo $L_s(n_i)$ é aplicado a n_i e todos os nós que dele descendem, funcionando como um rótulo genérico de todos os descendentes de n_i. Logo, um rótulo genérico pode ser recursivamente definido, a partir do nó raiz n_o, como:

$$L_s(n_0)\ =\ L_g(n_0)$$

e,

$$L_g(n_{ij})\ =\ L_s(n_{ij})\ \bigcup\ L_g(n_i)$$

Dadas essas definições, uma hierarquia de documentos (representada como uma árvore genérica), o conjunto D de documentos, o conjunto A de atributos e uma matriz de incidência das freqüências dos atributos em cada documento (também chamada de modelo espaço-vetorial ou de matriz atributo-valor), os algoritmos **4** e **5** e **6** podem ser aplicados para construir os conjuntos de rótulos de cada nó e, conseqüentemente, estimar os termos da taxonomia de tópicos dessa coleção de documentos.

Deve-se notar que, cada $L_s(n_{ij})$ produzido pelo algoritmo **5** é uma atualização dos rótulos já existentes para cada n_{ij}; pois, os que não são folhas inicialmente têm seus $L_s(n_{ij}) = \phi$ e, os que são folhas têm seus $L_s(n_{ij})$, no primeiro instante, correspondentes a todos os a_k com freqüência maior que zero nos documentos que formam o grupo n_i. Logo, neste método não é necessário totalizar as freqüências de a_k nos grupos, pois os a_k são selecionados das folhas para a raiz e suas freqüências nos grupos vão sendo atualizadas a cada passo. Ainda, os testes usados para verificar as hipóteses de independência de a_k nos n_{ij} são os mesmos utilizados no RLDM.

O algoritmo **4** percorre a árvore e controla se o nó atual é uma herança. Um nó é herdado quando seu nó pai é cortado da árvore; então ele é herdado pelo seu avô. Um nó

Input: n_i, a raiz precisa ser o primeiro nó
Output: *Hierarquia atualizada*, com os conjuntos de rótulos em cada nó
if n_i *tem pelo menos um filho* **then**
  recursivamente chama percArvore(filho de n_i);
end
if *se pai de* n_i *tem próximo filho* **then**
  if *esse filho não é herdado* **then**
    recursivamente chama percArvore(próximo filho do pai de n_i);
  end
end
else
  chama Seleciona(n_i);
end

Algorithm 4: Algoritmo percArvore, para controlar percorrer a árvore.

Input: n_i: nó corrente
Output: $L_s(n_i)$ *e* $L_s(n_{ij})$: atualizados
for *all* a_k, *k=1,..,m, e all* $f_{ij}(a_k) > 0$, *j=1,...,c* **do**
  if $c = 2$ **then**
    if $fi1 \approx 0 \ \vee \ fi2 \approx 0$ **then**
      $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$; $L_s(n_{i1}) \leftarrow L_s(n_{i1}) - \{a_k\}$;
      $L_s(n_{i2}) \leftarrow L_s(n_{i2}) - \{a_k\}$;
    end
    else
      if a_k *independente de acordo com Q de Yule* **then**
        $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$; $L_s(n_{i1}) \leftarrow L_s(n_{i1}) - \{a_k\}$;
        $L_s(n_{i2}) \leftarrow L_s(n_{i2}) - \{a_k\}$;
      end
    end
  end
  else
    if $c \geq 3$ **then**
      if a_k *independente de acordo com* U^2 **then**
        for *all filhos de* n_i, $j = 1, ..., c$ **do**
          $L_s(n_{ij}) \leftarrow L_s(n_{ij}) - \{a_k\}$;
        end
        $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$;
      end
    end
    else
      $L_s(n_i) \leftarrow L_s(n_i) \bigcup \{a_k\}$; $L_s(n_{i1}) \leftarrow L_s(n_{i1}) - \{a_k\}$;
    end
  end
end

Algorithm 5: Algoritmo Seleciona, para selecionar a_k em $L_s(ni)$ ou $L_s(n_{ij})$

Input: n_i, o nó corrente
Output: n_{ij}, com seu *status* atualizado
for *all filhos de* n_i **do**
  if *all* $f_{ij}(a_k) = 0, k = 1, ..., m$ **then**
    $n_i \leftarrow (children\ of\ this\ n_{ij})$; all desses filhos têm seu *status* mudado para "herdado"; n_{ij} é removido da hierarquia;
  end
end
if *all* $f_i(a_k) = 0,\ k = 1, ..., m$ **then**
  (seu nó pai) $\leftarrow$ (all n_i filhos); all desses filhos têm seu *status* mudado para "herdado"; n_i é removido da hierarquia;
end
Algorithm 6: Algoritmo corteSemantico, para fazer o corte semântico

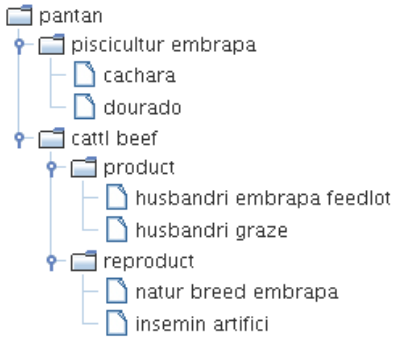

Figure 4. Resultados do RLUM para o exemplo ideal

n_i é cortado quando não há termos discriminativos em seu conjunto de rótulos, ou seja, o conjunto está vazio após a aplicação do algoritmo **5** a esse nó. Isso corresponde ao corte semântico realizado no RLDM. Conseqüentemente, no algoritmo **4**, identificar um nó herdado significa identificar que esse nó já foi tratado. O algoritmo **6** é responsável pelo corte da árvore a cada passo; e, não mais no final do processo, como no RLDM. Assim, também esse processo de corte tem sua complexidade reduzida em relação ao método anterior, que percorre novamente toda a árvore.

E, principalmente, este novo método evita o erro observado no RLDM, Figura 2, parte B. Como os testes não ocorrem primeiramente sobre as freqüências acumuladas nos nós descendentes da raiz, mas sim nos primeiros ascendentes das folhas, o erro é diminuído. O resultado fornecido pelos algoritmos do RLUM sobre os textos do exemplo ideal podem ser observados na Figura 4; onde o termo "Embrapa" é presente somente nos conjuntos de rótulos dos nós nos quais sua freqüência de ocorrência é estatisticamente mais significativa.

3. Validação do Processo de Rotulação

Embora a melhor forma de comparação deva ser uma análise subjetiva junto a especialistas do domínio de conhecimento das coleções de textos utilizadas no processo, devido à natureza dos dados e atributos, essa tarefa não é nada fácil mediante coleções de textos um pouco extensas, muito menos ainda sobre grandes coleções de textos. Mesmo em uma

pequena coleção, freqüentemente resultam grandes conjuntos de rótulos e uma grande hierarquia, se nenhum processo de poda de agrupamento for aplicado; e isso dificulta um julgamento subjetivo por parte dos especialistas. Então, propõe-se realizar uma análise objetiva dos resultados e, para obter uma validação confiável do RLUM, ele é comparado aos outros três métodos mencionados, de modo que se possam mostrar os melhoramentos esperados.

Assim os quatro métodos foram implementados, por meio de protótipos (na linguagem C): Popescul e Ungar [2000], os mais freqüentes (apenas ordenam-se as freqüências em cada nó e, estabelecendo-se um ponto de corte), o RLDM e o RLUM. Deste modo foi possível compará-los em relação aos conjuntos de rótulos para as mesmas hierarquias das mesmas coleções de documentos.

Primeiro, obtém-se o agrupamento hierárquico da coleção de textos, então uma cópia dessa hierarquia e da matriz de incidência é fornecida a cada implementação dos métodos. Como o RLDM e o RLUM podem realizar cortes na hierarquia, eles podem alterar a hierarquia final. De modo a comparar corretamente os métodos, não se devem comparar os conjuntos de rótulos obtidos para a hierarquia toda a partir dos quatro métodos, mas sim apenas os conjuntos de rótulos dos grupos (nós) mantido nas hierarquias após a aplicação dos métodos. Deste modo, as comparações são realizadas nó a nó, de acordo com a presença dos mesmos na hierarquia. Ainda, para que as comparações sejam justas, o ideal é que os conjuntos de rótulos sejam de tamanhos semelhantes. Pode-se forçar essa situação, colocando-se um corte de número máximo de rótulos em cada conjunto, como costuma ser feito para o método dos Mais Freqüentes. Utilizando essa idéia, basta aplicar o método dos Mais Freqüentes a cada resultado apresentado pelos demais métodos; ou seja, toma-se cada $L_s(n_i)$, ordenam-se os elementos por freqüência e então se limita o conjunto aos x elementos mais freqüentes. Assim, neste trabalho, em todas as comparações, serão utilizados os mesmos grupos, que são os nós mantidos pelos diferentes métodos, e conjuntos de rótulos de tamanhos semelhantes.

Decorre das definições de taxonomia de tópicos, neste trabalho, que os termos não são replicados ao longo dos mesmos ramos na hierarquia. Deste modo, o RLUM deve produzir realmente os conjuntos de rótulos mais específicos em cada nó e os mais gerais para os descendentes desse nó com melhor índice de acertos que os três outros métodos. Para testar essa primeira hipótese, a proposta é utilizar os conjuntos de rótulos selecionados como expressões de busca em um processo de recuperação de informações, como explicado na próxima sub-seção. Uma segunda hipótese é que o RLUM produz um vocabulário maior, ou no mínimo equivalente, comparado aos demais métodos. Se um vocabulário maior é obtido, espera-se que esse vocabulário tenha uma melhor cobertura do domínio em que a coleção se insere. Finalmente, se o vocabulário tem uma maior cobertura, sua qualidade deve ser medida e comparada à qualidade dos vocabulários obtidos com os três outros métodos, conforme explicado na última sub-seção.

3.1. Conjuntos de rótulos como expressões de busca

Todos os conjuntos de rótulos são tratados como expressões de busca, nos quais o operador "*and*" é usado entre os termos de um mesmo conjunto. Por exemplo, se tivermos o conjunto de rótulos $L_s(n_i) = \{husbandri, embrapa, feedlot\}$ então a expressão de busca será "*husbandri and embrapa and feedlot*". Neste caso, esperam-se obter como

resultados os documentos sob o $i-$ésimo nó, restringindo a busca à coleção de textos utilizada no exemplo ideal.

O processo de recuperação de informações foi implementado (outro protótipo em C) sobre a matriz atributo-valor utilizada no agrupamento de documentos. Na representação utilizada para essa matriz temos em suas linhas os documentos e em suas colunas os atributos utilizados, sendo que cada célula corresponde à freqüência de ocorrência do termo no documento. Para decidir se um documento foi ou não recuperado, pela expressão, todos os termos da busca devem ter freqüência maior que zero no documento. Após a aplicação do processo de recuperação, para cada grupo (nó da hierarquia) calculam-se os valores da Tabela 2.

	$recuperados$	$!recuperados$	$total$
n_i	t_p	f_n	d_c
$!(n_i)$	f_p	t_n	n_c
	r_d	n_r	o

Table 2. Número de recuperações em cada n_i

As medidas de interesse sobre os valores da Tabela 2 são:

- precisão: a proporção de documentos dentre os recuperados que está na classe correta, $p = t_p/r_d$;
- "recall": a proporção de documentos dentre os do grupo, que foram corretamente recuperados, $r = t_p/d_c$, também chamada de **cobertura de recuperação**;
- "false alarm": a proporção dentre os documentos recuperados que, de fato, não pertence ao grupo, $f_a = f_p/n_c$;
- "F measure": a média harmônica entre precisão e "recall", $F = 2 * p * r/(p + r)$. O valor ideal de F é igual a um, porque o ideal é ter $p = 1$ e $r = 1$; embora, seja considerado suficiente observar um comportamento harmônico de F ao longo de seu gráfico.

Essas medidas podem ser diretamente usadas para comparar os resultados, porque eles estão sendo computados para os mesmos grupos apenas, isto é, grupos com os mesmos documentos; a única diferença são os conjuntos de rótulos selecionados por um mesmo método, sendo estes limitados a tamanhos semelhantes, com no máximo x elementos cada.

Particularmente, a medida de maior interesse é a cobertura de recuperação ("recall"), porque ela representa bem a qualidade do conjunto de rótulos, quando usado como expressão de busca. Quando os rótulos são usados como palavras-chaves de busca, ligados pelo operador "*and*", se forem realmente representativos do grupo, o valor de "recall" deverá se aproximar de um, isto é, ter-se-ia um número muito baixo de falsos negativos (f_n) e um número de positivos verdadeiros (t_p) bastante alto, próximo ao total de documentos no grupo. Deste modo, é esperado que ao construir uma busca dessa forma, o RLUM obtenha os melhores resultados para "recall"; o que deve ser experimentalmente verificado.

3.2. Cobertura do vocabulário

Supõe-se que o RLUM produza um vocabulário maior, ou no mínimo igual ou equivalente, em cardinalidade ao dos outros métodos. Se os atributos forem escolhidos a

partir de um vocabulário controlado, como o contido em um *thesaurus*, no mínimo o vocabulário selecionado ao final do processo de rotulação deverá ser equivalente, em cardinalidade, para os quatro métodos. Por outro lado, se os atributos foram selecionados entre termos gerais da coleção e textos, retiradas as *stopwords* e aplicado algum filtro estatístico, espera-se que o número de diferentes termos produzidos pelo RLUM seja maior que por todos os outros. Como mencionado, anteriormente, se o vocabulário tiver essa propriedade, é esperado que ele seja capaz de ter uma melhor cobertura do domínio no qual a coleção de textos se insere.

Para verificar essa hipótese, as cardinalidades dos vocabulários obtidos por cada método devem ser comparadas, para diferentes coleções de textos. Experimentalmente, é aconselhável ter coleções de diferentes tamanhos, diferentes idiomas ou domínios; e, então, aplica-se um teste de comparação múltipla de médias (veja [**?**], [**?**] e [**?**], para maiores detalhes), para agrupar comparativamente as cardinalidades.

Primeiramente, vamos definir o vocabulário V como a união de todos os conjuntos de rótulos da taxonomia de tópicos obtida:

$$V = L_g(root) = \bigcup_{i=1}^{z} L_s(n_i),$$

$$z: \ número\ total\ de\ nós\ na\ hierarquia$$

Então, elabora-se a hipótese sobre as diferentes cardinalidades:

$$H_0 : card(V_R) \geq card(V_F) \geq card(V_P) \geq card(V_M)$$

$$V_R: \ vocabulário\ para\ RLUM$$

$$V_F: \ vocabulário\ para\ ``Most\ Frequent''$$

$$V_P: \ vocabulário\ para\ Popescul\ e\ Ungar$$

$$V_M: \ vocabulário\ para\ RLDM$$

Para testar essa hipótese, calculam-se os vocabulários para algumas diferentes bases, ajusta-se um modelo linear generalizado para calcular as variações de cada efeito dos método e cardinalidade do conjunto e atributos iniciais sobre as cardinalidades finais do vocabulário, da seguinte forma (veja Searle [1971] para maiores detalhes sobre modelos lineares generalizados)[3]:

$$\hat{card}_v \ = \ \hat{\mu} \ + \ < \hat{card}_A > \ + \ \hat{m} \ + \ \hat{\epsilon}$$

onde,

- $\hat{card}_v$: cardinalidade final inferida para cada vocabulário em cada método e cada base
- $\hat{\mu}$: a média geral de dispersão estimada para inferir as cardinalidades finais

[3]Toda estimativas é notada com chapéu.

- $< \hat{card}_A >$: estimativa do efeito da cardinalidade do conjunto de atributos iniciais, em cada caso, para inferência da cardinalidade final. Utilizando-se a cardinalidade de A como covariável é esperado que o efeito de diferentes número de atributos seja devidamente tratado e considerado nas inferências finais, permitindo que se possa comparar apenas os efeitos relativos a diferentes métodos de rotulação sobre a cardinalidade do vocabulário final;
- $\hat{m}$: estimativa do efeito do método de rotulação aplicado, isto é, quanto a estimativa final desvia-se da média geral de dispersão em relação a esse efeito
- $\hat{\epsilon}$: erro aleatório de ajuste do modelo linear

Então, com esse modelo é calculada a análise de variância com os dados do experimento realizado e feita a comparação múltipla de médias. Escolheu-se utilizar o teste SNK (*Student-Newman-Keuls*) para a comparação múltipla, devido a sua melhor robustez (veja [**?**] e [**?**] para maiores detalhes) e o software de domínio público MODLIN, do repositório AgroLivre (http://repositorio.agrolivre.gov.br/), para o ajuste do modelo linear e obtenção dos resultados.

3.3. Representatividade do vocabulário

Uma maior cardinalidade do vocabulário não garante a qualidade do mesmo, pois verificar essa propriedade pode ser apenas reflexo de uma série de más escolhas. Para avaliar as escolhas, que podem ser vistas como quão bem os rótulos selecionados predizem o conteúdo dos documentos em cada grupo, além do uso da medida "recall", pode-se utilizar a esperança da informação mútua entre o conjunto de atributos e o vocabulário final. Para a coleção completa de documentos após a rotulação, a Esperança da Informação Mútua (EMIM) ([**?**] e [**?**]) é definida como:

$$EMIM(V, A) = \sum_{u=1}^{card(V)} \sum_{k=1}^{m} p(v_u, a_k) * \log(\frac{p(v_u, a_k)}{p(v_u) \times p(a_k)})$$

Considerando-se cada $p(v_u, a_k)$ como a probabilidade de ocorrerem simultaneamente o vocábulo $v_u \in V$ e o atributo $a_k \in A$, e, esta estima tiva da probabilidade ser o estimador de máxima verossimilhança dado pelo número de ocorrÊncias de v_u e a_k em cada documento da coleção. Assim, como definida aqui, a EMIM reflete estatisticamente a qualidade do vocabulário selecionado. Para comparar essa qualidade entre os quatro métodos, a proposta é considerar os diferentes conjuntos de rótulos produzidos por cada método para cada grupo de documentos e calcular:

$$EMIMn_i(L_s(n_i), A) = \sum_{u=1}^{card(L_s(n_i))} \sum_{k=1}^{m} p(l_u, a_k) * \log(\frac{p(l_u, a_k)}{p(l_u) \times p(a_k)})$$

Considerando-se cada l_u como um elemento do conjunto de rótulos $L_s(n_i)$. Como esses valores são calculados para os mesmos grupos de documentos, os valores obtidos para a EMIM podem ser comparados utilizando-se comparação múltipla de médias. E, ainda, como a EMIM não tem um intervalo de valores definidos ou mesmo um máximo, os maiores valores obtidos relfetem as melhores predições. Porém, a EMIM é sensível ao tamanho dos conjuntos envolvidos; mas esse problema já foi evitado ao se considerar

apenas os nós (grupos) comuns às hierarquias e limitar o tamanho de cada conjunto de rótulos a, no máximo, x elementos cada. Assim, para testar a hipótese de que a EMIM em cada nó, para cada método, é maior para o RLUM, podendo, no mínimo, comparar-se aos demais métodos, ajusta-se o seguinte modelo linear para a análise de variância e comparação múltipla de médias:

$$\hat{emim}(n_i) = \hat{\mu} + \hat{n_i} + \hat{m} + \hat{\epsilon}$$

onde,

- $\hat{emim}(n_i)$: emim final inferida para cada nó em cada método e cada base
- $\hat{\mu}$: a média geral de dispersão estimada para inferir a $emim(n_i)$ final
- $\hat{n_i}$: estimativa do efeito de cada nó n_i em cada base na inferência da emim final. O nó funciona como efeito de categoria para cada EMIM, fazendo com que as comparações de médias dos efeitos dos métodos sejam justas por estarem sendo realizadas nó a nó;
- $\hat{m}$: estimativa do efeito do método de rotulação aplicado, isto é, quanto a média geral de dispersão desvia-se da EMIM final em relação a esse efeito;
- $\hat{\epsilon}$: erro aleatório de ajuste do modelo linear.

4. Resultados Experimentais

Quatro coleções de textos foram selecionadas para os experimentos, cada qual dividida em quatro classes, de tamanhos diferentes, que foram utilizados como bases isoladas, logo, dezesseis coleções de textos foram utilizadas nos testes. As coleções passaram pelo mesmo processo de pré-processamento, seguido por processos idênticos de agrupamento hierárquico e depois pelos quatro métodos de rotulação. As análises dos resultados foram conduzidas em relação a cada um dos fatores considerados: cobertura para a recuperação de informações a partir dos conjuntos de rótulos, cobertura de vocabulário de cada método e qualidade de vocabulário para cada método. As sub-seções a seguir, apresentam cada passo dos experimentos.

Deve-se salientar que, os resultados de comparações múltiplas de médias, apresentados neste relatório, são realizados com o teste SNK (Student-Newman-Keuls), usando o quadrado médio do erro do modelo linear utilizado na análise de variância, com o grau de liberdade desse erro e o nível de significância de 5%. O teste foi escolhido devido a sua força e sua característica de sempre mostrar a real diferença entre as médias [**?**]. Nas tabelas apresentadas, diferentes letras na coluna de grupos correspondem a diferentes grupos de médias. Em cada sub-título das tabelas tem-se os graus de liberdade, a variância do erro utilizada e o número de nós envolvidos nos testes. Pois, os testes são realizados comparando-se as medidas em cada nó, ou seja em cada grupo, mantido nas hierarquias rotuladas por diferentes métodos. Ainda, quando algum modelo de análise de variância apresentou uma matriz singular, impossibilitando a aplicação da comparação múltipla de médias a todos os métodos, isso foi indicado na tabela e, calculada a média da medida para o valor fora das comparações; exemplo para a medida F no sub-item 4.2.

4.1. Pré-processamento, agrupamento e rotulação

Na Tabela 3 é apresentado, de forma resumida, o resultado do pré-processamento das coleções utilizadas no experimento. A primeira coleção de textos corresponde a artigos

completos, em português, da segunda assembléia do Instituto Fábrica do Milênio (**IFM**[4]). No primeiro momento essa coleção tinha 614 documentos, dividida em 4 classes, com 198 documentos na maior classe; as classes correspondem a projetos de trabalho do instituto (**WP01**, **WP02**, **WP03** e **WP04**). A segunda coleção de textos corresponde a artigos em inglês, 408 documentos, sobre computação, divididos nos temas: inteligência artificial (**AI**), interface homem máquina (**HCI**), *hardware* (**Hard**) e segurança e criptografia (**Sec**). A terceira coleção corresponde a 396 artigos sobre química, divididos nos temas: orgânica (**OrgC**), inorgânica (**InoC**), analítica (**AnaC**) e ciência dos polímeros (**PolyC**). A quarta coleção são artigos sobre física, divididos em biofísica (**BioP**), geofísica (**GeoP**), mecânica (**Mech**) e física quântica (**Quan**). Essas quatro coleções de textos podem ser acessadas no $site$ do projeto TopTax[5].

Na segunda coluna da Tabela 3 encontram-se os números de documentos utilizados em cada coleção após a conversão dos textos do formato original (PDF) para texto plano. Quando esse processo é aplicado, algumas vezes, alguns documentos são descartados, devido à conversão apresentar algum tipo de problema; fica-se apenas com os textos cujo conteúdo possa ser utilizado no processo. A seguir, aplicou-se "*stemmização*" às coleções, obtendo-se unigramas (número total em cada coleção na terceira coluna da tabela), com a ferramenta PreText [**?**]. Para simplificar e padronizar o procedimento de filtro foi utilizado aproximadamente a indicação de Salton [**?**] para palavras discriminativas, que é usar as palavras que possuem freqüência de ocorrência nos documentos ("*df: document frequence*") entre um a dez por cento da coleção; isto é, com $D = \{d_1, ..., d_x\}$, com x documentos, então cada termo ($stem$ na tabela) que ficará no conjunto A precisa ter sua df no intervalo: $0.01x \leq df(termo) \leq 0.10x$. No entanto, como as coleções serão submetidas a um processo de agrupamento aglomerativo, usou-se no mínimo $df = 2$ e, para compensar a retirada de $df = 1$, a maior df foi acrescida de um ou dois pontos. Após a aplicação desse filtro, resumido na quarta coluna da tabela, tem-se a cardinalidade do conjunto A de atributos.

Ainda, para manter uma padronização de resultados, um agrupamento aglomerativo é aplicado a cada coleção, utilizando-se similaridade de cosseno e o algoritmo "*average linkage*". Após esse procedimento, são aplicados os métodos de rotulação. Como eles geram diferentes conjuntos específicos de rótulos, em cada nó, de diferentes cardinalidades, fixou-se um número máximo de rótulos por nó, em catorze (14). Para fazer isso, os rótulos obtidos em cada nó, para cada método, foram ordenados por freqüência de ocorrência, como no método dos Mais Freqüentes. Desse modo, cada método de rotulação gera, no máximo, catorze rótulos por $L_s(n_i)$ ordenados por freqüência.

4.2. Avaliando os Rótulos como Expressões de Busca

Para essa avaliação consideraram-se os conjuntos de rótulos genéricos e os conjuntos de rótulos específicos separadamente, em diferentes experimentos. A idéia é avaliar a utilização dos conjuntos específicos como expressão de busca, isto é, cada $L_s(n_{ij})$ como se o grupo fosse resultado de um "flat cluster" e, então comparar os resultados obtidos por cada método de rotulação. A seguir, utilizaram-se os rótulos genéricos $L_g(n_i)$ junto ao rótulo específico em cada filho de n_i, isto é, utilizou-se a expressão de busca

[4]O IFM é uma organização brasileira cujas ações são focadas na busca por soluções de necessidades da manufatura nas indústrias; veja: http://www.ifm.org.br/

[5]In: http://www.labic.icmc.usp.br/projects

TC	#docs	#stems	filtro com DF	$card(A)$
WP01	80	10268	$2 \geq df \leq 10$	4799
WP02	65	11630	$2 \geq df \leq 9$	5000
WP03	198	19904	$2 \geq df \leq 21$	9302
WP04	48	13870	$2 \geq df \leq 7$	5252
Hard	89	11014	$2 \geq df \leq 10$	3453
AI	94	10128	$2 \geq df \leq 11$	4445
HCI	98	8915	$2 \geq df \leq 12$	3971
Sec	127	18780	$2 \geq df \leq 14$	6253
AnaC	100	18408	$2 \geq df \leq 12$	6645
InoC	97	18436	$2 \geq df \leq 12$	8864
OrgC	99	20116	$2 \geq df \leq 12$	5238
PolyC	100	11628	$2 \geq df \leq 12$	5634
BioP	95	13179	$2 \geq df \leq 12$	6187
GeoP	97	12011	$2 \geq df \leq 12$	5100
Mech	117	12660	$2 \geq df \leq 14$	4378
Quan	82	9586	$2 \geq df \leq 10$	3798

Table 3. Resumo das Coleções de Textos Utilizadas

$L_g(n_i) \bigcup L_s(n_{ij})$ para cada n_{ij}; essa segunda expressão permite que se avalie a rotulação hierárquica propriamente dita, pois reflete o fato de que os rótulos genéricos dos antecessores discriminam o nó atual.

Na Tabela 4 é apresentado o resumo dos resultados desse primeiro experimento para a coleção de textos **WP01**. Nessa tabela estão presentes os resultados das comparações múltiplas de médias para cada medida de interesse: F, média harmônica de precisão e revocação; $prec$, precisão; r, revocação ("*recall*"); e, F_a, falso alarme. Todas as medidas discutidas na seção avaliação são colocadas na tabela e a comparação múltipla de médias utilizou o teste SNK, com probabilidade de aceitação de 95%, a partir de um modelo linear generalizado da seguinte forma:

$$\hat{m(n_{ij})} = \hat{\mu} + \hat{n_{ij}} + \hat{\epsilon}$$

onde,

- $\hat{m(n_{ij})}$: inferência da medida de interesse em cada n_{ij};
- $\hat{\mu}$: estimativa da média geral de dispersão da medida;
- $\hat{n_{ij}}$: estimativa do efeito de cada nó (grupo) na medida;
- $\hat{\epsilon}$: erro aleatório da estimativa.

Os mesmos cálculos e avaliações foram realizados para as dezesseis coleções de textos, relacionadas no Anexo A; e, o comportamento observado para as medidas nas dezesseis coleções foi semelhante.

Deve-se notar na Tabela 4 que, para algumas medidas, como a F e a precisão $prec$ não há o mesmo número de pontos calculados, para cada nó (n, na tabela); o que ocorre devido a alguma divisão por zero, por exemplo. Ainda, na tabela, para a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$, a medida F do RLUM não pode ser comparada às demais, por insuficiência de pontos nos resultados dos demais métodos. Porém, nota-se que o resultado de F para RLUM foi 1, nos trinta pontos totais, ou seja, o resultado atingiu o valor ideal. Nota-se, na Tabela, que o RLUM só perde para o RLDM na precisão, quando

utilizada a expressão de busca $L_s(n_{ij})$, mas em todos os demais casos seu desempenho é melhor que dos outros três. Quanto ao falso alarme, o comportamento do RLUM é sempre aceitável, comparável ao outros; porém o RLDM erra mais.

modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 1$

método	n	F	$grupo$
$MostFreq$	1	0.697248	a....
$PopeUngar$	1	0.697248	a....
$RLDM$	1	0.367816	..b..
$\overline{F}_{RLUM}$	30	1.000000	$singular$

$L_s(n_{ij})$, $\sqrt{e} = 0.0282, gl = 60$

método	n	F	$grupo$
$RLDM$	30	0.807975	a....
$RLUM$	30	0.701492	..b..
$PopeUngar$	17	0.401681	c
$MostFreq$	17	0.401681	c

modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0179, gl = 64$

método	n	$prec$	$grupo$
$RLUM$	30	0.590959	a....
$RLDM$	6	0.166667	..b..
$PopeUngar$	31	0.032258	c
$MostFreq$	31	0.032258	c

$L_s(n_{ij})$, $\sqrt{e} = 0.0371, gl = 89$

método	n	$prec$	$grupo$
$RLDM$	31	0.801459	a....
$RLUM$	30	0.579107	..b..
$PopeUngar$	31	0.197366	c
$MostFreq$	31	0.197366	c

modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0007, df = 89$

método	n	rec	$grupo$
$RLUM$	30	1.000000	a..
$PopeUngar$	31	0.017265	..b
$MostFreq$	31	0.017265	..b
$RLDM$	31	0.007269	..b

$L_s(n_{ij})$, $\sqrt{e} = 0.0541, df = 89$

método	n	rec	$grupo$
$RLUM$	30	1.000000	a....
$RLDM$	31	0.848667	..b..
$PopeUngar$	31	0.326405	c
$MostFreq$	31	0.326405	c

modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0042, df = 89$

método	n	F_a	$grupo$
$MostFreq$	31	0.064665	a..
$PopeUngar$	31	0.064665	a..
$RLUM$	30	0.056270	a..
$RLDM$	31	0.003314	..b

$L_s(n_{ij})$, $\sqrt{e} = 0.0041, df = 89$

método	n	F_a	$grupo$
$MostFreq$	31	0.065600	a..
$PopeUngar$	31	0.065600	a..
$RLUM$	30	0.059169	a..
$RLDM$	31	0.009922	..b

Table 4. Resultados das medidas de interesse para WP01.

Outros resultados interessantes referem-se, por exemplo, aos métodos dos Mais Freqüentes e de Popescul e Ungar, que sempre apresentam um comportamento semelhante. O método RLDM tem um melhor desempenho, no que tange à precisão e "*F measure*", quando se usam os conjuntos de rótulos específicos como expressão de busca, obtendo-se resultados melhores que os demais métodos. Já o método RLUM sempre tem um excelente desempenho no "*recall*", próximo a um, um bom índice da "*F measure*" e um falso alarme dentro do esperado. Seu desempenho é sempre superior ou, no mínimo, equivalente aos demais, quando essas medidas são boas para os demais e a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ é usada; logo, estatisticamente, ele obtém os conjuntos de rótulos mais representativos quando toda a hierarquia é considerada. Reforçando essas observações, os resultados de "*recall*" para as dezesseis coleções de textos estão resumidos nos gráficos das Figuras 5 e 6. Observando os dois gráficos, nota-se que a "*recall*" para o RLUM é sempre próxima a 1; que a "*recall*" do RLDM se aproxima de 1, quando a expressão

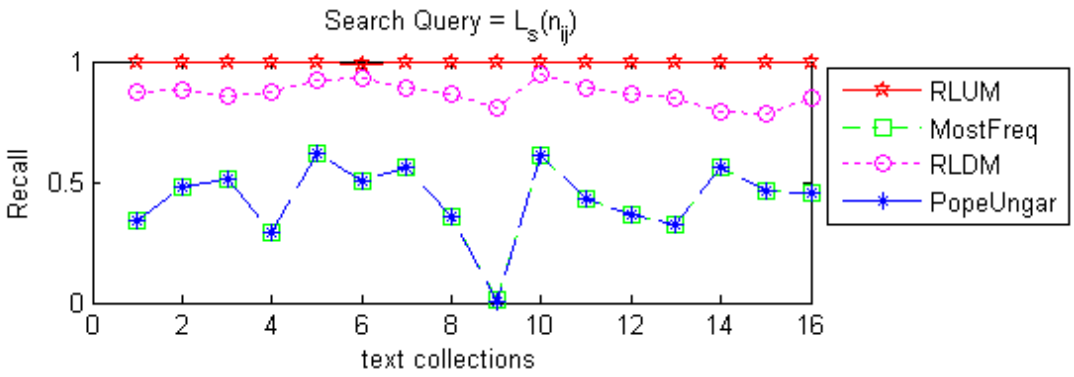

Figure 5. Gráfico de "Recall" para as expressão de busca $L_s(n_{ij})$ para as dezesseis coleções de textos.

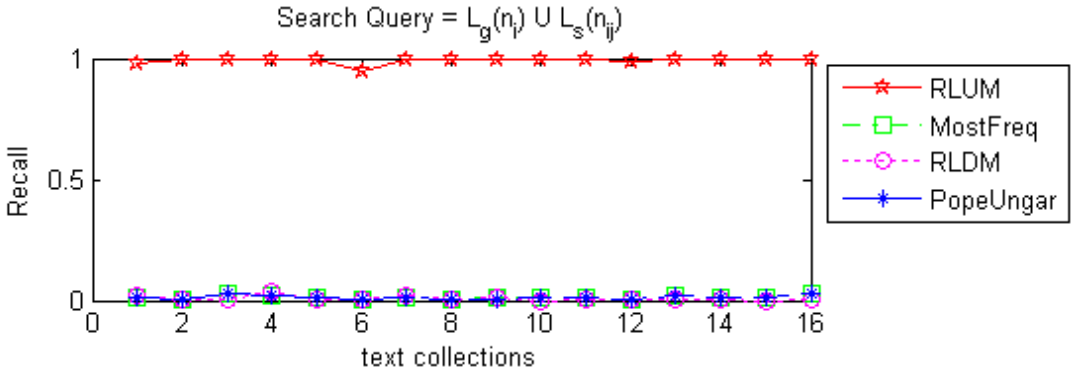

Figure 6. Gráfico de "Recall" para as expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ para as dezesseis coleções de textos.

de busca é restrita a $L_s(n_{ij})$; e, a grande similaridade dos resultados do Popescul & Ungar comparados aos do Mais Freqüentes. Logo, salienta-se o resultado de que o RLUM produz os melhores rótulos na hierarquia, dado que seus resultados de "*recall*" para a expressão de busca $L_g(n_i) \bigcup L_s(n_{ij})$ são estatisticamente iguais a um.

4.3. Avaliando a Cobertura do Vocabulário

No exemplo usado para mostrar o desenvolvimento das idéias de rotulação, foi utilizado um conjunto de atributos retirado de um vocabulário do domínio pré-existente, no caso, do Thesagro (*Thesaurus* Agrícola Nacional [**?**]). Quando isso ocorre, espera-se que os vocabulários finais tenham cardinalidades equivalentes para qualquer método de rotulação. Pois, os termos estão nos textos e, se ficaram dentro dos intervalos de corte, filtro de atributos, provavelmente farão diferença significativa nos agrupamentos e servirão para discriminar os grupos na rotulação. Quando se utiliza um conjunto de atributos com base apenas na geração de termos a partir das coleções, como feito aqui, com a ferramenta PreText [**?**], não há garantia de que todos os termos selecionados para o conjunto de atributos façam parte de algum *thesaurus*, ou mesmo que todos eles estejam presentes no vocabulário final obtido. Como há uma indicação de que o método RLUM produz um vocabulário maior que o dos outros métodos, espera-se que este método consiga melhor cobrir um vocabulário e domínio, já que se o número é maior espera-se que as chances dos termos de domínio estarem nele contidos também sejam maiores.

Assim, nesta sub-seção verifica-se a hipótese:

$$H_0 : card(V_U) \geq card(V_P) \geq card(V_M) \geq card(V_D)$$

Pois aparentemente o RLDM gera o menor vocabulário e os outros dois geram vocabulários de tamanhos semelhantes, como observado na Tabela 5. Como também se deve considerar o fato de que os conjuntos de atributos iniciais têm diferentes tamanhos, como observado na Tabela 3, faz-se necessário retirar o efeito dessa covariável do modelo, como dito anteriormente.

TC	$card(V_P)$	$card(V_M)$	$card(V_D)$	$card(V_U)$
WP01	645	645	488	927
WP02	566	566	690	867
WP03	1383	1384	1595	2245
WP04	495	495	558	739
Hard	836	824	478	1000
AI	747	746	665	928
HCI	808	801	597	937
Sec	1116	1116	766	1359
AnaC	915	914	838	1244
InoC	803	798	974	1120
OrgC	1051	1030	512	1309
PolyC	786	784	757	967
BioP	759	754	799	977
GeoP	825	824	793	1032
Mech	937	937	691	1179
Quan	719	716	607	912

Table 5. Cardinalidade dos Vocabulários Obtidos

Teste SNK para a cardinalidade final, com rejeição=0.05 *Com 59 graus de liberdade, variância média= 42482.6158*			
metodo	observações	$\overline{card(V_m)}$	grupo
RLUM	16	1108.875000	a..
PopeUngar	16	836.937500	..b
MaisFreq	16	833.375000	..b
RLDM	16	738.000000	..b

Table 6. Comparação múltipla de médias para cardinalidades dos vocabulários

Após a comparação múltipla de médias das cardinalidades finais obteve-se o resultado apresentado na Tabela 6. Pode-se notar que houve o mesmo número de observações para cada método, isto é, a cardinalidade foi observada e considerada para as dezesseis bases. A cardinalidade média para cada método, $\overline{card(V_m)}$, foi maior para o RLUM ($card(V_U)$), com 0.95 de certeza. Como esperado as cardinalidades do Popescul e Ungar ($card(V_P)$) e dos Mais Freqüentes $card(V_M)$ são estatisticamente iguais. Porém, estatisticamente, já não se observa que o RLDM ($card(V_D)$) apresente uma cardinalidade mais baixa que os demais, ou seja, ele pode ser considerado como obtendo um vocabulário tão variado quanto os métodos de Popescul e Ungar e o dos Mais Freqüentes.

4.4. Avaliação da Representatividade do Vocabulário Produzido

Como dito anteriormente, apenas a cardinalidade do método ser superior à dos demais não garante que a qualidade do vocabulário produzido seja boa, pois ele pode ser resultado de uma série de más escolhas. Embora a medida "*recall*" já indique se as escolhas foram ou não boas para a formação dos conjuntos de rótulos de cada grupo. Assim, nesta fase do experimento vamos utilizar a EMIM para testar a hipótese de que a qualidade do vocabulário produzido pelo RLUM é superior a dos outros três métodos, ou seja:

$$H_0 : EMIM(RLUM) \geq EMIM(P) \approx EMIM(M) \geq EMIM(D)$$

Para verificar essa hipótese, calcula-se a EMIM em cada nó para cada taxonomia de tópicos resultante de cada método e, então, compara-se a EMIM dos nós correspondentes em cada resultado utilizando o modelo linear generalizado, já explicado, para análise de variância e realiza-se a comparação múltipla das médias com o teste SNK, com 0.95 de certeza. A Tabela 7 foi obtida para essa comparação a partir das bases do domínio de computação; os demais resultados para as outras bases encontram-se no Anexo A.

Deve-se observar, na Tabela 7 que todos os valores são calculados, tem-se sempre o mesmo número de grupos (ou nós), n, para cada média calculada. Vemos também que o agrupamento esperado de médias se confirma nas quatro bases apresentadas (bem como se confirma nos resultados apresentados no Anexo A), ou seja, a hipótese de qualidade dos vocabulários, de acordo com a EMIM, é confirmada com 0.95 de certeza. Assim, além do RLUM apresentar a maior variabilidade de termos no vocabulário produzido, ele também garante a melhor qualidade desses termos, em relação aos três métodos comparados. Também se deve observar que os métodos de Popescul e Ungar e dos Mais Freqüentes sempre se encontram nos mesmos grupos, o que mostra que seus resultados são sempre muito próximos devido a lacuna do processo decisório no método de Popescul e Ungar, discutido anteriormente. Ainda, o RLDM tende a produzir um vocabulário mais pobre, em termos de cardinalidade, e de menor qualidade.

$Col.\ de\ Textos\ AI$ $gl = 147, var.media = 160140.0445$				$Col.\ de\ Textos\ HCI$ $gl = 108, var.media = 286839.4677$			
$metodo$	n	$emim$	$grupo$	$metodo$	n	$emim$	$grupo$
$RLUM$	50	1089.899965	$a....$	$RLUM$	37	885.627493	$a..$
$PopeUngar$	50	893.817625	$..b..$	$PopeUngar$	37	765.692200	$a..$
$MostFreq$	50	893.742870	$..b..$	$MostFreq$	37	762.961843	$a..$
$RLDM$	50	557.042183	$....c$	$RLDM$	37	373.175243	$..b$
$Col.\ de\ Textos\ Hard$ $gl = 36, var.media = 185914.7144$				$Col.\ de\ Textos\ Sec$ $gl = 159, var.media = 487370.7896$			
$metodo$	n	$emim$	$grupo$	$metodo$	n	$emim$	$grupo$
$RLUM$	13	1104.943006	$a....$	$RLUM$	54	1613.500940	$a..$
$PopeUngar$	13	909.675011	$a....$	$PopeUngar$	54	1337.548719	$a..$
$MostFreq$	13	906.329243	$a....$	$MostFreq$	54	1337.383964	$a..$
$RLDM$	13	77.376233	$..b..$	$RLDM$	54	636.411828	$..b$

Table 7. EMIM - resultados da comparação múltipla de suas médias

5. Considerações Finais

Neste relatório técnico foi apresentado um novo método de rotulação para agrupamentos hierárquicos de documentos e uma proposta de validação objetiva do método. O objetivo do método é produzir rótulos discriminativos dos grupos, sem repetição ao longo dos ramos, com um vocabulário variado e de qualidade. Para isso o método foi baseado em uma definição formal de uma taxonomia de tópicos, também apresentada no trabalho, e em seguida submetido aos testes objetivos de discriminação e qualidade.

Num primeiro momento, o novo método era uma proposta de diminuição da complexidade do RLDM, objetivo imediatamente atingido, pois o RLDM possui uma complexidade da ordem de $O(mz^2)$ e o RLUM da ordem de $O(mz)$, considerando-se m o número de atributos e z o número de nós da hierarquia. Além disso, o RLUM diminuiu a propagação de erros ao longo da tomada de decisão sobre os atributos selecionados como rótulos, porque constrói os conjuntos de rótulos das folhas para a raiz, enquanto lhes atualiza.

A melhor discriminação dos conjuntos de rótulos foi verificada utilizando-se os mesmos como expressões de busca sobre as coleções de textos e verificando a cobertura ("*recall*") dos resultados. O RLUM apresentou a melhor cobertura de recuperação, estatisticamente mais significante, 0.95 de certeza, que os demais em todos os casos testados, para as dezesseis coleções de textos.

A ordem de grandeza dos vocabulários produzidos foi testada comparando-se a cardinalidade dos vocabulários produzidos pelos quatro métodos de rotulação, em dezesseis diferentes coleções de textos. O RLUM obteve uma maior variabilidade de vocabulário em todos os testes. E, quando testada a qualidade desse vocabulário, contra a dos obtidos pelos demais métodos, também o RLUM superou os demais, ou no mínimo igualou-se a eles.

Assim, todas as hipóteses iniciais sobre o RLUM e suas relações com os métodos de Popescul e Ungar, dos Mais Freqüentes e do RLDM foram validadas como verdadeiras, ou seja, o método atingiu todos os objetivos inicialmente propostos.

Em outros trabalhos realizou-se uma validação subjetiva de métodos de rotulação junto a especialistas de domínio ([**?**] e [**?**]). Uma série de fatores interfere nesse processo

de validação, desde a forma de obtenção dos termos até o tamanho final das taxonomias, o que acaba gerando muitas limitações e não permite que se tenha muita credibilidade nos resultados das avaliações. Isto é, as avaliações não refletem exclusivamente o método de rotulação, acabam refletindo outras variáveis subjetivas, que não são fáceis de identificar. Assim, como trabalhos futuros, pretende-se utilizar o RLUM em um processo de construção de taxonomias de tópicos e validar o processo contra taxonomias *gold*, isto é, taxonomias pré-existentes e já bem aceitas, como realizado por outros autores ([**?**],[**?**]). Espera-se com esse tipo de validação conseguir simular uma validação subjetiva exclusivamente do método de rotulação.

References

BINAGRI (2009). Thesaurus agrícola nacional – thesagro. In: http:www.agricultura.gov.brportal, May 28th, 2009.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice,*. London: M.I.T. Press.

Feldman, R. and Sanger, J. (2007). *The Text Mining Hand Book - Advanced Approaches in Analysing Unstructured Data*. Cambridge University Press.

Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. (2005). Taxaminer: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266.

Krishnapuram, R. and Kummamuru, K. (2003). Automatic taxonomy generation: Issues and possibilities. In *Fuzzy Sets and Systems – IFSA 2003*, volume 2715 of *Lecture Notes in Computer Science*, pages 52–63. Springer.

Lawrie, D., Croft, W. B., and Rosenberg, A. (2001). Finding topic words for hierarchical summarization. In *SIGIR '01: Proceedings of the XXIV Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 349–357, New York, NY, EUA. ACM.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, EUA.

Matsubara, E. T., Martins, C. A., and Monard, M. C. (2003). Pre-text: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Technical Report 209, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

Moura, M. F., Marcacini, R. M., Nogueira, B. M., da Silva Conrado, M., and Rezende, S. O. (2008a). A proposal for building domain topic taxonomies. In *WTI '08: Proceedings of I Workshop on Web and Text Intelligence - SBIA '08: XIX Simpósio Brasileiro de Inteligência Artificial*, pages 83–84. São Carlos: ICMCUSP.

Moura, M. F., Marcacini, R. M., and Rezende, S. O. (2008b). Easily labelling hierarchical document clusters. In *WAAMD '08: Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados - SBBD '08: XXIII Simpósio Brasileiro de Banco de Dados*, pages 37–45. Porto Alegre : SBC.

Moura, M. F. and Rezende, S. O. (2007). Choosing a hierarchical cluster labelling method for a specific domain document collection. In Neves, J., Santos, M. F., and Machado,

J. M., editors, *New Trends in Artificial Intelligence*, chapter 11, pages 812–823. Lisboa, Portugal: APPIA - Associação Portuguesa para Inteligência Artificial. EPIA-Encontro Português de Inteligência Artificial, 2007, Guimarães, Portugal, 1 edition.

RAMSEY, P. H. (1993). Multiple comparisons of independent means. In EDWARDS, L. K., editor, *Applied analysis of variance in behavioral science*. New York: Marcel Dekker.

Salton, G., Yang, C. S., and Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Association Science*, 1(26):33–44.

Searle, S. R. (1971). *Linear models*. J. Wiley, New York, NY.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical methods, 6th ed*. Iowa State University Press, Ames, IO.

6. Anexo A - Complemento dos Resultados Experimentais

Neste anexo encontram-se todos os dados de medidas de interesse obtidas a partir de recuperação de informações e de esperança da informação mútua, para todas as coleções de textos utilizadas nos experimentos deste relatório.

6.1. Comparação Múltipla de Médias de EMIM

Para estudar o comportamento da EMIM em cada nó para cada método ajusta-se o seguinte modelo linear generalizado:

$$\hat{emim}_{n_i} = \hat{\mu} + \hat{n_i} + \hat{m} + \hat{\epsilon}$$

onde,

- $\hat{emim}_{n_i}$: emim inferida para cada nó em cada método e cada coleção de textos;
- $\hat{\mu}$: a média geral de dispersão estimada para inferir a $emim_{n_i}$;
- $\hat{n_i}$: estimativa do efeito de cada nó n_i em cada coleção de textos na inferência da $emim_{n_i}$. O nó funciona como efeito de categoria para cada EMIM, fazendo com que as comparações de médias dos efeitos dos métodos sejam justas por estarem sendo realizadas nó a nó;
- $\hat{m}$: estimativa do efeito do método de rotulação aplicado, isto é, quanto a média geral de dispersão desvia-se da EMIM final em relação a esse efeito;
- $\hat{\epsilon}$: erro aleatório de ajuste do modelo linear.

Com esse modelo, a seguinte hipótese é testada para cada coleção de textos:

$$H_0 : EMIM(RLUM) \geq EMIM(P) \approx EMIM(M) \geq EMIM(D)$$

Realizando-se a comparação múltipla de médias das estimativas de EMIM em relação aos efeitos de cada método, para cada coleção de textos é apresentada uma tabela. Cada tabela contem o nome da coleção, o número de graus de liberdade gl utilizado na comparação múltipla de médias, a variância média vm utilizada pelo teste SNK, o número de pontos n utilizado para estimar a EMIM em cada coleção, a estimativa da média de EMIM ($\bar{emim}$) para cada método em cada coleção e uma letra para $grupo$. Letras iguais representam grupos estatisticamente iguais e, a ordem dos grupos é a alfabética.

Os valores da Tabela 8 resumem as comparações múltiplas de médias para as coleções de textos de Ciência da Computação. Deve-se notar que a EMIM de RLUM é estatisticamente maior que as demais apenas para a coleção de textos de IA. Nas demais coleções de textos o resultado é estatisticamente igual ao apresentado pelos métodos de Popescul & Ungar e dos Mais Freqüentes. Já o RLDM apresenta-se sempre na última posição, exceto para a primeira coleção de textos (IA).

Na Tabela 9 encontram-se resumidas as comparações múltiplas de médias para as coleções de textos de Química. Deve-se notar que a EMIM de RLUM é indiscutivelmente maior que os demais apenas para a coleção de textos de química analítica (AnaC), nas demais coleções de textos o resultado é estatisticamente igual ao apresentado pelos métodos de Popescul & Ungar e dos Mais Freqüentes. Já o RLDM apresenta-se isoladamente

$Col.\ de\ Textos\ AI$ $gl = 147, vm = 160140.0445$				$Col.\ de\ Textos\ HCI$ $gl = 108, vm = 286839.4677$			
método	n	$\bar{emim}$	$grupo$	método	n	$\bar{emim}$	$grupo$
$RLUM$	50	1089.899965	a	$RLUM$	37	885.627493	a
$PopeUngar$	50	893.817625	b	$PopeUngar$	37	765.692200	a
$MostFreq$	50	893.742870	b	$MostFreq$	37	762.961843	a
$RLDM$	50	557.042183	c	$RLDM$	37	373.175243	b
$Col.\ de\ Textos\ Hard$ $gl = 36, vm = 185914.7144$				$Col.\ de\ Textos\ Sec$ $gl = 159, vm = 487370.7896$			
método	n	$\bar{emim}$	$grupo$	método	n	$\bar{emim}$	$grupo$
$RLUM$	13	1104.943006	a	$RLUM$	54	1613.500940	a
$PopeUngar$	13	909.675011	a	$PopeUngar$	54	1337.548719	a
$MostFreq$	13	906.329243	a	$MostFreq$	54	1337.383964	a
$RLDM$	13	77.376233	b	$RLDM$	54	636.411828	b

Table 8. EMIM - resultados para o domínio de Ciência da Computação.

$Col.\ de\ Textos\ AnaC$ $gl = 150,\ vm = 626822.0993$				$Col.\ de\ Textos\ InoC$ $gl = 153,\ vm = 1e+06$			
$method$	n	$\bar{emim}$	$group$	$method$	n	$\bar{emim}$	$group$
$RLUM$	51	1944.724002	a	$RLUM$	52	2235.525562	a
$PopeUngar$	51	1447.356021	b	$PopeUngar$	52	1664.561465	b
$MostFreq$	51	1446.846145	b	$MostFreq$	52	1661.026177	b
$RLDM$	51	747.228784	c	$RLDM$	52	1308.392552	b
$Col.\ de\ Textos\ OrgC$ $gl = 105,\ vm = 1e+06$				$Col.\ de\ Textos\ Poly$ $gl = 141,\ vm = 274807.2374$			
$method$	n	$\bar{emim}$	$group$	$method$	n	$\bar{emim}$	$group$
$RLUM$	36	2638.297039	a	$RLUM$	48	1083.452986	a
$MostFreq$	36	2129.767519	a	$PopeUngar$	48	885.448869	$a\ b$
$PopeUngar$	36	2113.688907	a	$MostFreq$	48	884.329346	$a\ b$
$RLDM$	36	303.138253	b	$RLDM$	48	656.565679	b

Table 9. EMIM - resultados para o domínio de Química.

na última posição ou, na melhor das classificações empatado com o Popescul & Ungar e Mais Freqüentes. Ainda, vale observar que, para coleção de Ciência dos Polímeros (Poly), a hipótese inicial é bastante evidente. Nessa coleção, o RLUM empata com o Popescul & Ungar e Mais Freqüentes ou pode ser considerado melhor. Isso é um grande indício de que, realmente, o RLUM provê um vocabulário representativo; dado que em avaliações subjetivas sempre foi observado que o Mais Freqüentes provê uma maior gama de vocábulos, o que auxilia a interpretação dos tópicos [**?**].

Os valores da Tabela 10 referem-se às comparações múltiplas de médias para as coleções de textos do domínio de Física. Nesse domínio, a EMIM de RLUM foi melhor que os demais em três das coleções de textos, empatando apenas em uma delas com os métodos de Popescul & Ungar e Mais Freqüentes. Já o RLDM em duas coleções empatou com os métodos de Popescul & Ungar e Mais Freqüentes, porém nas demais se apresenta na última posição.

$Col.\ de\ Textos\ BioP$ $gl = 102,\ vm = 4e+06$				$Col.\ de\ Textos\ GeoP$ $gl = 150,\ vm = 309004.6045$			
$method$	n	$\bar{emim}$	$group$	$method$	n	$\bar{emim}$	$group$
$RLUM$	35	3305.903151	a	$RLUM$	51	1644.596988	a
$RLDM$	35	2353.240417	b	$PopeUngar$	51	1209.551453	b
$PopeUngar$	35	1842.456371	b	$MostFreq$	51	1209.089450	b
$MostFreq$	35	1842.456371	b	$RLDM$	51	1069.912500	b
$Col.\ de\ Textos\ Mech$ $gl = 138,\ vm = 457452.0106$				$Col.\ de\ Textos\ Quan$ $gl = 117,\ vm = 284639.2019$			
$method$	n	$\bar{emim}$	$group$	$method$	n	$\bar{emim}$	$group$
$RLUM$	47	1335.448216	a	$RLUM$	40	1276.068038	a
$PopeUngar$	47	1030.787221	a	$PopeUngar$	40	988.725002	b
$MostFreq$	47	1030.787221	a	$MostFreq$	40	985.847125	b
$RLDM$	47	414.650337	b	$RLDM$	40	654.964093	c

Table 10. EMIM - resultados para o domínio de Física.

$Col.\ de\ Textos\ WP01$ $gl = 90,\ vm = 2e+07$			
$method$	n	$\bar{emim}$	$group$
$RLUM$	31	5190.859414	a
$PopeUngar$	31	3429.286832	$a\ b$
$MostFreq$	31	3429.286832	$a\ b$
$RLDM$	31	1897.535148	c
$Col.\ de\ Textos\ WP02$ $gl = 102,\ vm = 4e+06$			
$method$	n	$\bar{emim}$	$group$
$RLUM$	35	3305.903151	a
$RLDM$	35	2353.240417	b
$PopeUngar$	35	1842.456371	b
$MostFreq$	35	1842.456371	b
$Col.\ de\ Textos\ WP04$ $gl = 78,\ vm = 3e+06$			
$method$	n	$\bar{emim}$	$group$
$RLUM$	27	3790.885026	a
$PopeUngar$	27	2475.731493	b
$MostFreq$	27	2475.731493	b
$RLDM$	27	2094.232192	b

Table 11. EMIM - resultados para as coleções de textos do IFM

Finalmente, na Tabela 11 estão resumidas as comparações múltiplas de médias para as coleções de textos do IFM. Nessa coleção não foi possível obter o valor de EMIM para a coleção WP03; após dois dias rodando o programa ainda não havia terminado. Para implementar a EMIM necessita-se de um algoritmo de complexidade exponencial; logo, em uma coleção de textos com 198 textos, 9302 atributos e até 2300 elementos no vocabulário; tem-se uma tarefa muito custosa. Como esse dado não seria responsável por uma mudança tão grande na classificação dos métodos, foi ignorado. E, novamente, o RLUM mostrou o mesmo comportamento que nas outras coleções.

Deve-se observar que, sempre o Popescul & Ungar e Mais Freqüentes apresentam comportamento similar, dado que os vocabulários por eles selecionados são sempre muito próximos. Assumiu-se o Mais Freqüentes como uma boa estimativa de variabilidade do vocabulário devido às análises subjetivas anteriormente realizadas ([**?**] e [**?**]). E, finalmente, que a hipótese inicial ($H_0 : EMIM(RLUM) \geq EMIM(P) \approx EMIM(M) \geq EMIM(D)$) foi confirmada em todos os experimentos.

6.2. Conjuntos de Rótulos como Expressões de Busca

Para cada método, hierarquia e coleção de textos varia-se a expressão de busca, podendo-se utilizar:

- rótulo genérico: todos os conjuntos de rótulos são tratados como expressões de busca, nos quais o operador "*and*" é usado entre os termos de um mesmo conjunto. Por exemplo, o conjunto de rótulos $L_s(n_{21}) = \{husbandry, product\} \bigcup L_g(n_{21}) = \{pantanal, embrapa, catle, beef\}$ é usado como uma expressão de busca da seguinte forma: "$husbandry \wedge product \wedge pantanal \wedge embrapa \wedge catle \wedge beef$".
- rótulo específico: todos os conjuntos de rótulos são tratados como expressões de busca, nos quais o operador "*and*" é usado entre os termos de um mesmo conjunto. Por exemplo, o conjunto de rótulos $L_s(n_{21}) = \{husbandry, product\}$ é usado como uma expressão de busca da seguinte forma: "$husbandry \wedge product$".

Para verificar experimentalmente uma hipótese sobre o comportamento das medidas de interesse, sobre a Tabela 2, para cada um dos métodos de rotulação, calculam-se as medidas para cada um dos nós mantidos nas quatro taxonomias. Tabulam-se essas medidas junto ao número do nó e ao método que as gerou, e, então se ajusta um modelo linear generalizado [**?**] para analisar a variância e obter estimativas para cada medida. O modelo considera os efeitos do nó e do método de rotulação, além da média geral:

$$\hat{m_e} = \hat{\mu} + \hat{n_{ij}} + \hat{l_m} + \hat{\epsilon}$$

- $\hat{m_e}$: estimativa da medida de avaliação após ajuste do modelo para a coleção;
- $\hat{\mu}$: média geral do modelo, sem qualquer outro efeito, para a coleção;
- $\hat{n_{ij}}$: estimativa do efeito do nó n_{ij}, que é o quanto a medida $\hat{m_e}$ desvia-se da média geral devido a esse efeito, para a coleção;
- $\hat{l_m}$: estimativa do efeito do método de rotulação,que é o quanto a medida $\hat{m_e}$ desvia-se da média geral devido à aplicação de um dado método, para a coleção;
- $\hat{\epsilon}$: efeito aleatório associado a cada estimativa na coleção.

Com o uso dessas estimativas para cada medida, $\hat{m_e}$, para cada método de rotulação $\hat{l_m}$ espera-se obter comparações mais estatisticamente confiáveis. Aplicando-se os testes de comparação múltiplas de médias a cada modelo, tem-se 128 tabelas, duas para cada coleção e 4 para cada medida em cada caso. Os valores são resumidos nas Tabelas de 12 a 19. Cada pequena tabela interna às maiores contem a comparação múltipla de médias das estimativas de uma das medidas em relação aos efeitos de cada método, para cada coleção de textos. Essas tabelinhas contem o nome da coleção, o modelo generalizado para a medida, a expressão de busca utilizada, o número de graus de liberdade gl utilizado na comparação múltipla de médias, a variância média $sqrt(e)$ utilizada pelo teste

SNK, o número de pontos n utilizado para estimar a medida em cada coleção, a estimativa da medida em cada coleção e uma letra para $grupo$. Novamente, letras iguais representam grupos estatisticamente iguais e, a ordem dos grupos é a alfabética. Ainda, quando não foi possível calcular a estimativa com o modelo, é fornecida sua média e desvio padrão dp, ou na falta dessa estimativa o valor "." (*missing value*).

A expressão de busca com o rótulo genérico é tabulada na parte esquerda das Tabelas de 12 a 19. Nesse caso, observa-se a ausência de pontos calculados para a medida F para as coleções WP01, WP03, IA, HCI, Sec, AnaC, Poly, BioP, Mech e Quan. Em geral, isso ocorre quando a expressão de busca retorna conjuntos de resultados vazios, pois tanto $recall$ quanto $precision$ ficam zeradas e a F tem seu quociente zerado. E, em alguns casos o desbalanceamento entre o número de pontos é bem grande. Em qualquer desses casos, o RLUM sempre apresenta valores que estão nas melhores posições. Quando o valor de F para o RLUM não é o melhor, o número de pontos definidos para RLUM é maior que para os outros métodos, o que compensa os grandes valores dos demais. Logo, a medida F para o RLUM é sempre a melhor para o conjunto de rótulos genéricos.

Ainda com o rótulo genérico como expressão de busca, examinando os valores de $recall$ observa-se que para o RLUM ele é sempre o melhor e muito próximo a um (valor máximo). Enquanto o resultado de $recall$ para os demais métodos é ruim. Observa-se o mesmo comportamento para $precision$; e para o $falso\ alarme$ um comportamento bastante aceitável em todos os métodos, estatisticamente equivalentes. Logo, o RLUM tem um melhor desempenho ao se considerar o rótulo genérico do nó como expressão de busca.

Os resultados para os rótulos específicos como expressão de busca estão tabulados do lado direito das Tabelas de 12 a 19. Nota-se claramente a melhoria de desempenho do RLDM, enquanto os demais métodos mantêm seus desempenhos anteriores. Logo, o RLDM se aproxima do desempenho do RLUM. Assim, não há outras mudanças significativas de comportamento dos métodos e o RLUM mantém seus resultados, provando experimentalmente ter bom desempenho também quando a expressão de busca utilizada é apenas o rótulo específico do nó.

WP01 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 1$

método	n	F	$grupo$
$MostFreq$	1	0.697248	a
$PopeUngar$	1	0.697248	a
$RLDM$	1	0.367816	b
$\overline{F}_{RLUM}$	30	1.000000	$singular$

$L_s(n_{ij})$, $\sqrt{e} = 0.0282, gl = 60$

método	n	F	$grupo$
$RLDM$	30	0.807975	a
$RLUM$	30	0.701492	b
$PopeUngar$	17	0.401681	c
$MostFreq$	17	0.401681	c

WP01 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0179, gl = 64$

método	n	$prec$	$grupo$
$RLUM$	30	0.590959	a
$RLDM$	6	0.166667	b
$PopeUngar$	31	0.032258	c
$MostFreq$	31	0.032258	c

$L_s(n_{ij})$, $\sqrt{e} = 0.0371, gl = 89$

método	n	$prec$	$grupo$
$RLDM$	31	0.801459	a
$RLUM$	30	0.579107	b
$PopeUngar$	31	0.197366	c
$MostFreq$	31	0.197366	c

WP01 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0007, gl = 89$

método	n	rec	$grupo$
$RLUM$	30	1.000000	a
$PopeUngar$	31	0.017265	b
$MostFreq$	31	0.017265	b
$RLDM$	31	0.007269	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0541, gl = 89$

método	n	rec	$grupo$
$RLUM$	30	1.000000	a
$RLDM$	31	0.848667	b
$PopeUngar$	31	0.326405	c
$MostFreq$	31	0.326405	c

WP01 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0042, gl = 89$

método	n	F_a	$grupo$
$MostFreq$	31	0.064665	a
$PopeUngar$	31	0.064665	a
$RLUM$	30	0.056270	a
$RLDM$	31	0.003314	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0041, gl = 89$

método	n	F_a	$grupo$
$MostFreq$	31	0.065600	a
$PopeUngar$	31	0.065600	a
$RLUM$	30	0.059169	a
$RLDM$	31	0.009922	b

WP02 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.645833	a
$PopeUngar$	1	0.645833	a
$RLDM$	1	0.030303	b
$\overline{F}_{RLUM}$	30	1.000000	$singular$

$L_s(n_{ij})$, $\sqrt{e} = 0.0204, gl = 82$

método	n	F	$grupo$
$RLUM$	34	0.940728	a
$RLDM$	34	0.856283	b
$PopeUngar$	26	0.781311	b
$MostFreq$	26	0.781311	b

WP02 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0564, gl = 7$

método	n	$prec$	$grupo$
$RLDM$	1	1.000000	a
$RLUM$	34	0.925414	a
$PopeUngar$	5	0.200000	b
$MostFreq$	5	0.200000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0205, gl = 82$

método	n	$prec$	$grupo$
$RLDM$	34	1.000000	a
$RLUM$	34	0.925414	a
$PopeUngar$	26	0.907051	a
$MostFreq$	26	0.907051	a

WP02 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0014, gl = 101$

método	n	rec	$grupo$
$RLUM$	34	1.000000	a
$PopeUngar$	35	0.013626	b
$MostFreq$	35	0.013626	b
$RLDM$	35	0.000440	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0582, gl = 101$

método	n	rec	$grupo$
$RLUM$	34	1.000000	a
$RLDM$	35	0.788195	b
$PopeUngar$	35	0.559885	c
$MostFreq$	35	0.559885	c

WP02 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0099, gl = 101$

método	n	F_a	$grupo$
$RLUM$	34	0.051110	a
$PopeUngar$	35	0.001891	a
$MostFreq$	35	0.001891	a
$RLDM$	35	0.000000	a

$L_s(n_{ij})$, $\sqrt{e} = 0.0100, gl = 101$

método	n	F_a	$grupo$
$RLUM$	34	0.051110	a
$PopeUngar$	35	0.004620	a
$MostFreq$	35	0.004620	a
$RLDM$	35	0.000000	a

Table 12. Resultados das medidas de interesse para WP01 e WP02.

WP03 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 1$

método	n	F	$grupo$
$RLUM$	109	0.913249	a
$PopeUngar$	2	0.686275	b
$MostFreq$	2	0.686275	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0211, gl = 260$

método	n	F	$grupo$
$RLUM$	109	0.914064	a
$RLDM$	107	0.852992	b
$PopeUngar$	78	0.718027	c
$MostFreq$	79	0.718027	c

WP03 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 1$

método	n	$prec$	$grupo$
$PopeUngar$	2	1.00000	a
$MostFreq$	2	1.00000	a
$RLUM$	109	0.895565	$b.$

$L_s(n_{ij})$, $\sqrt{e} = 0.0221, gl = 260$

método	n	$prec$	$grupo$
$RLDM$	107	1.000000	a
$MostFreq$	79	0.938608	b
$PopeUngar$	79	0.937821	b
$RLUM$	109	0.893730	b

WP03 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0017, gl = 326$

método	n	rec	$grupo$
$RLUM$	109	0.994648	a
$PopeUngar$	110	0.009504	b
$MostFreq$	110	0.009504	b
$RLDM$	110	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0533, gl = 326$

método	n	rec	$grupo$
$RLUM$	109	0.997706	a
$RLDM$	110	0.780400	b
$MostFreq$	110	0.464428	c
$PopeUngar$	110	0.461398	c

WP03 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.00, gl = 326$

método	n	F_a	$grupo$
$RLUM$	109	0.031854	a
$PopeUngar$	110	0.000000	b
$RLDM$	110	0.000000	b
$MostFreq$	110	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.000, gl = 326$

método	n	F_a	$grupo$
$RLUM$	109	0.031854	a
$PopeUngar$	110	0.000000	b
$RLDM$	110	0.000000	b
$MostFreq$	110	0.000000	b

WP04 – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 1$

método	n	F	$grupo$
$RLUM$	26	0.986111	a
$PopeUngar$	2	0.476191	b
$MostFreq$	2	0.476191	b
$RLDM$	1	0.040816	c

$L_s(n_{ij})$, $\sqrt{e} = 0.00, gl = 59$

método	n	F	$grupo$
$RLUM$	26	0.960012	a
$RLDM$	27	0.890224	a
$PopeUngar$	18	0.690344	b
$MostFreq$	18	0.690344	b

WP04 – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0137, gl = 15$

método	n	$prec$	$grupo$
$RLDM$	1	1.000000	a
$RLUM$	26	0.976923	a
$PopeUngar$	9	0.166667	b
$MostFreq$	9	0.166667	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0, gl = 63$

método	n	$prec$	$grupo$
$RLDM$	27	1.000000	a
$RLUM$	26	0.938462	a
$PopeUngar$	20	0.760000	b
$MostFreq$	20	0.760000	b

WP04 – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0024, gl = 77$

método	n	rec	$grupo$
$RLUM$	26	1.000000	a
$PopeUngar$	27	0.025926	b
$MostFreq$	27	0.025926	b
$RLDM$	27	0.000772	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0621, gl = 77$

método	n	rec	$grupo$
$RLUM$	26	1.000000	a
$RLDM$	27	0.852006	b
$MostFreq$	27	0.455556	c
$PopeUngar$	27	0.455556	c

WP04 – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0004, gl = 77$

método	n	F_a	$grupo$
$MostFreq$	27	0.018075	a
$PopeUngar$	27	0.018075	a
$RLUM$	26	0.002584	b
$RLDM$	27	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0005, gl = 77$

método	n	F_a	$grupo$
$MostFreq$	27	0.018880	a
$PopeUngar$	27	0.018880	$a\ b$
$RLUM$	26	0.006801	$a\ c$
$RLDM$	27	0.000000	c

Table 13. Resultados das medidas de interesse para WP03 e WP04.

Hard – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$RLUM$	25	0.872534	a
$PopeUngar$	1	0.630769	b
$MostFreq$	1	0.630769	b
$RLDM$	2	0.521978	c

$L_s(n_{ij})$, $\sqrt{e} = 0.0181, gl = 40$

método	n	F	$grupo$
$RLUM$	26	0.897844	a
$PopeUngar$	9	0.884900	a
$MostFreq$	9	0.884900	a
$RLDM$	25	0.872534	a

Hard – modelo:$\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 0$

método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a
$MostFreq$	1	1.00000	a
$RLDM$	2	1.00000	a
$RLUM$	25	0.858595	$b.$

$L_s(n_{ij})$, $\sqrt{e} = 0.0331, gl = 40$

método	n	$prec$	$grupo$
$RLDM$	26	1.000000	a
$MostFreq$	9	1.000000	a
$PopeUngar$	9	1.000000	a
$RLUM$	25	0.858595	a

Hard – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.00117, gl = 74$

método	n	rec	$grupo$
$RLUM$	25	1.000000	a
$RLDM$	26	0.039326	b
$MostFreq$	26	0.017718	b
$PopeUngar$	26	0.017718	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0750, gl = 74$

método	n	rec	$grupo$
$RLUM$	25	1.000000	a
$RLDM$	26	0.873941	a
$MostFreq$	26	0.286949	b
$PopeUngar$	26	0.286949	b

Hard – modelo: $\hat{F_a} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0111, gl = 74$

método	n	F_a	$grupo$
$RLUM$	25	0.074451	a
$PopeUngar$	26	0.000000	a
$RLDM$	26	0.000000	a
$MostFreq$	26	0.000000	a

$L_s(n_{ij})$, $\sqrt{e} = 0.0111, gl = 74$

método	n	F_a	$grupo$
$RLUM$	25	0.074451	a
$PopeUngar$	26	0.000000	a
$RLDM$	26	0.000000	a
$MostFreq$	26	0.000000	a

IA – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.553846	a
$PopeUngar$	1	0.553846	a
$RLDM$	1	0.021053	b
$RLUM$	48	0.9178	$dp = 0.2246$

$L_s(n_{ij})$, $\sqrt{e} = 0.0079, gl = 100$

método	n	F	$grupo$
$MostFreq$	27	0.953846	a
$PopeUngar$	27	0.953846	a
$RLDM$	50	0.931746	a
$RLUM$	49	0.923607	a

IA – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0, gl = 0$

método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a
$PopeUngar$	1	1.000000	a
$RLDM$	1	1.000000	a
$RLUM$	48	0.9230	$dp = 0.2398$

$L_s(n_{ij})$, $\sqrt{e} = 0.0113, gl = 100$

método	n	$prec$	$grupo$
$MostFreq$	27	1.000000	a
$PopeUngar$	27	1.000000	a
$RLDM$	50	0.982759	a
$RLUM$	49	0.911579	b

IA – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0087, gl = 146$

método	n	rec	$grupo$
$RLUM$	49	0.945578	a
$PopeUngar$	50	0.007660	b
$MostFreq$	50	0.007660	b
$RLDM$	50	0.000213	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0752, gl = 146$

método	n	rec	$grupo$
$RLUM$	49	0.986395	a
$RLDM$	50	0.928546	a
$PopeUngar$	50	0.500993	b
$MostFreq$	50	0.500993	b

IA – modelo: $\hat{F_a} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0049, gl = 146$

método	n	F_a	$grupo$
$RLUM$	49	0.034920	a
$PopeUngar$	50	0.000000	a
$RLDM$	50	0.000000	a
$MostFreq$	50	0.000000	a

$L_s(n_{ij})$, $\sqrt{e} = 0.0046, gl = 146$

método	n	F_a	$grupo$
$RLUM$	49	0.036507	a
$RLDM$	50	0.005556	b
$PopeUngar$	50	0.000000	b
$MostFreq$	50	0.000000	b

Table 14. Resultados das medidas de interesse para Hard e IA.

HCI – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.515152	a
$PopeUngar$	1	0.492308	b
$RLDM$	25	0.115385	c
$RLUM$	36	0.9183	$dp = 0.2429$

$L_s(n_{ij})$, $\sqrt{e} = 0.0106, gl = 81$

método	n	F	$grupo$
$RLUM$	26	0.897844	a
$PopeUngar$	9	0.884900	a
$MostFreq$	9	0.884900	a
$RLDM$	25	0.872534	a

HCI – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 0$

método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a
$MostFreq$	1	1.00000	a
$RLDM$	1	1.00000	a
$RLUM$	36	0.9076	$dp = 0.2694$

$L_s(n_{ij})$, $\sqrt{e} = 0.0157, gl = 81$

método	n	$prec$	$grupo$
$MostFreq$	24	1.000000	a
$PopeUngar$	24	1.000000	a
$RLDM$	37	1.000000	a
$RLUM$	36	0.907559	b

HCI – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0005, gl = 107$

método	n	rec	$grupo$
$RLUM$	36	1.000000	a
$MostFreq$	37	0.009377	b
$PopeUngar$	37	0.008825	b
$RLDM$	37	0.001655	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0727, gl = 107$

método	n	rec	$grupo$
$RLUM$	36	1.000000	a
$RLDM$	37	0.921925	a
$MostFreq$	37	0.621989	b
$PopeUngar$	37	0.621438	b

HCI – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0049, gl = 146$

método	n	F_a	$grupo$
$RLUM$	49	0.034920	a
$PopeUngar$	50	0.000000	a
$RLDM$	50	0.000000	a
$MostFreq$	50	0.000000	a

$L_s(n_{ij})$, $\sqrt{e} = 0.0071, gl = 107$

método	n	F_a	$grupo$
$RLUM$	36	0.046655	a
$PopeUngar$	37	0.000000	a
$RLDM$	37	0.000000	a
$MostFreq$	37	0.000000	a

SEC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.540230	a
$PopeUngar$	1	0.540230	a
$RLDM$	1	0.015625	b
$RLUM$	53	0.8319	$dp = 0.3157$

$L_s(n_{ij})$, $\sqrt{e} = 0.0164, gl = 92$

método	n	F	$grupo$
$MostFreq$	21	0.952709	a
$PopeUngar$	21	0.952709	a
$RLDM$	54	0.895924	$a\ b$
$RLUM$	53	0.838208	b

SEC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0, gl = 0$

método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a
$PopeUngar$	1	1.000000	a
$RLDM$	1	1.000000	a
$RLUM$	53	0.9843	$dp = 0.0818$

$L_s(n_{ij})$, $\sqrt{e} = 0.0365, gl = 92$

método	n	$prec$	$grupo$
$MostFreq$	21	1.000000	a
$PopeUngar$	21	1.000000	a
$RLDM$	54	0.990741	a
$RLUM$	53	0.814195	b

SEC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0022, gl = 158$

método	n	rec	$grupo$
$RLUM$	53	0.984277	a
$PopeUngar$	54	0.006853	b
$MostFreq$	54	0.006853	b
$RLDM$	54	0.000146	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0750, gl = 158$

método	n	rec	$grupo$
$RLUM$	53	0.993711	a
$RLDM$	54	0.865269	b
$PopeUngar$	54	0.361792	c
$MostFreq$	54	0.361792	c

SEC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0147, gl = 158$

método	n	F_a	$grupo$
$RLUM$	53	0.089954	a
$PopeUngar$	54	0.000000	b
$RLDM$	54	0.000000	b
$MostFreq$	54	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0147, gl = 158$

método	n	F_a	$grupo$
$RLUM$	53	0.089954	a
$RLDM$	54	0.000448	b
$PopeUngar$	54	0.000000	b
$MostFreq$	54	0.000000	b

Table 15. Resultados das medidas de interesse para HCI e SEC.

AnaC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0206, gl = 86$			
método	n	F	$grupo$	método	n	F	$grupo$
$RLUM$	50	0.922434	a	$RLUM$	50	0.929100	a
$MostFreq$	1	0.611111	b	$RLDM$	50	0.914546	a
$PopeUngar$	1	0.611111	b	$PopeUngar$	20	0.912222	a
$RLDM$	2	0.519608	c	$MostFreq$	20	0.912222	a

AnaC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0000, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0188, gl = 86$			
método	n	$prec$	$grupo$	método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a	$MostFreq$	20	1.000000	a
$MostFreq$	1	1.00000	a	$PopeUngar$	20	1.000000	a
$RLDM$	2	1.00000	a	$RLDM$	50	0.988333	a
$RLUM$	50	0.921118	b	$RLUM$	50	0.921118	a

AnaC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0075, gl = 149$				$L_s(n_{ij})$ $\sqrt{e} = 0.0734, gl = 149$			
método	n	rec	$grupo$	método	n	rec	$grupo$
$RLUM$	50	0.983333	a	$RLUM$	50	0.993333	a
$RLDM$	51	0.020000	b	$RLDM$	51	0.873268	b
$MostFreq$	51	0.008627	b	$MostFreq$	51	0.341961	c
$PopeUngar$	51	0.008627	b	$PopeUngar$	51	0.341961	c

AnaC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0072, gl = 149$				$L_s(n_{ij})$ $\sqrt{e} = 0.0072, gl = 149$			
método	n	F_a	$grupo$	método	n	F_a	$grupo$
$RLUM$	50	0.041660	a	$RLUM$	50	0.041660	a
$PopeUngar$	51	0.000000	a	$PopeUngar$	51	0.000000	a
$RLDM$	51	0.000000	a	$RLDM$	51	0.000000	a
$MostFreq$	51	0.000000	a	$MostFreq$	51	0.000000	a

InoC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0101, gl = 115$			
método	n	F	$grupo$	método	n	F	$grupo$
$RLUM$	51	0.948548	a	$RLUM$	51	0.948548	a
$MostFreq$	1	0.689189	b	$RLDM$	51	0.921489	a
$PopeUngar$	1	0.689189	b	$MostFreq$	34	0.907245	a
$RLDM$	2	0.520202	c	$PopeUngar$	34	0.907245	a

InoC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0, gl = 0$				$L_s(n_{ij})$ $\sqrt{e} = 0.0103, gl = 115$			
método	n	$prec$	$grupo$	método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a	$MostFreq$	34	1.000000	a
$PopeUngar$	1	1.000000	a	$PopeUngar$	34	1.000000	a
$RLDM$	2	1.000000	a	$RLDM$	51	0.990196	a
$RLUM$	51	0.934309	b	$RLUM$	51	0.934309	b

InoC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0060, gl = 152$				$L_s(n_{ij})$ $\sqrt{e} = 0.0694, gl = 152$			
método	n	rec	$grupo$	método	n	rec	$grupo$
$RLUM$	51	1.000000	a	$RLUM$	51	1.000000	a
$PopeUngar$	52	0.019627	b	$RLDM$	52	0.889179	b
$MostFreq$	52	0.010111	b	$PopeUngar$	52	0.559791	c
$RLDM$	52	0.010111	b	$MostFreq$	52	0.559791	c

InoC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$ $\sqrt{e} = 0.0008, gl = 152$				$L_s(n_{ij})$ $\sqrt{e} = 0.0009, gl = 152$			
método	n	F_a	$grupo$	método	n	F_a	$grupo$
$RLUM$	51	0.013860	a	$RLUM$	51	0.013860	a
$PopeUngar$	52	0.000000	a	$RLDM$	52	0.000614	a
$RLDM$	52	0.000000	a	$PopeUngar$	52	0.000000	a
$MostFreq$	52	0.000000	a	$MostFreq$	52	0.000000	a

Table 16. Resultados das medidas de interesse para AnaC e InoC.

OrgC – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.544118	a
$RLDM$	1	0.476923	b
$PopeUngar$	1	0.322034	c
$RLUM$	35	0.8148	$dp = 0.3431$

$L_s(n_{ij})$, $\sqrt{e} = 0.0433, gl = 34$

método	n	F	$grupo$
$RLDM$	36	0.866026	a
$RLUM$	35	0.814839	a
$PopeUngar$	1	0.544118	a
$MostFreq$	1	0.322034	a

OrgC – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 0$

método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a
$MostFreq$	1	1.00000	a
$RLDM$	1	1.00000	a
$RLUM$	35	0.7918	$dp = 0.3729$

$L_s(n_{ij})$, $\sqrt{e} = 0.0695, gl = 34$

método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a
$PopeUngar$	1	1.000000	a
$RLDM$	36	1.000000	a
$RLUM$	35	0.791819	a

OrgC – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0002, gl = 104$

método	n	rec	$grupo$
$RLUM$	35	1.000000	a
$MostFreq$	36	0.010382	b
$RLDM$	36	0.008698	b
$PopeUngar$	36	0.005331	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0209, gl = 104$

método	n	rec	$grupo$
$RLUM$	35	1.000000	a
$RLDM$	36	0.809624	b
$MostFreq$	36	0.010382	c
$PopeUngar$	36	0.005331	c

OrgC – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0128, gl = 104$

método	n	F_a	$grupo$
$RLUM$	35	0.102751	a
$PopeUngar$	36	0.000000	b
$RLDM$	36	0.000000	b
$MostFreq$	36	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0128, gl = 104$

método	n	F_a	$grupo$
$RLUM$	35	0.102751	a
$PopeUngar$	36	0.000000	b
$RLDM$	36	0.000000	b
$MostFreq$	36	0.000000	b

Poly – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.689189	b
$PopeUngar$	1	0.689189	b
$RLDM$	.	.	.
$RLUM$	47	0.9663	$dp = 0.1515$

$L_s(n_{ij})$, $\sqrt{e} = 0.0024, gl = 104$

método	n	F	$grupo$
$RLDM$	46	0.978261	a
$RLUM$	47	0.966277	a
$MostFreq$	31	0.966157	a
$PopeUngar$	31	0.966157	a

Poly – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0, gl = 0$

método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a
$PopeUngar$	1	1.000000	a
$RLDM$	.	.	.
$RLUM$	47	0.9594	$dp = 0.1732$

$L_s(n_{ij})$, $\sqrt{e} = 0.0047, gl = 104$

método	n	$prec$	$grupo$
$MostFreq$	31	1.000000	a
$PopeUngar$	31	1.000000	a
$RLDM$	46	0.987578	a
$RLUM$	47	0.959433	a

Poly – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0013, gl = 140$

método	n	rec	$grupo$
$RLUM$	47	1.000000	a
$PopeUngar$	48	0.010833	b
$MostFreq$	48	0.010833	b
$RLDM$	48	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0785, gl = 140$

método	n	rec	$grupo$
$RLUM$	47	1.000000	a
$RLDM$	48	0.942708	a
$PopeUngar$	48	0.611528	b
$MostFreq$	48	0.611528	b

Poly – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0019, gl = 140$

método	n	F_a	$grupo$
$RLUM$	47	0.015764	a
$PopeUngar$	48	0.000000	a
$RLDM$	48	0.000000	a
$MostFreq$	48	0.000000	a

$L_s(n_{ij})$, $\sqrt{e} = 0.0019, gl = 140$

método	n	F_a	$grupo$
$RLUM$	47	0.015764	a
$RLDM$	48	0.000614	a
$PopeUngar$	48	0.000000	a
$MostFreq$	48	0.000000	a

Table 17. Resultados das medidas de interesse para OrgC e Poly.

BioP – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.560606	a
$PopeUngar$	1	0.429752	b
$RLDM$	1	0.137255	c
$RLUM$	53	0.9144	$dp = 0.2281$

$L_s(n_{ij})$, $\sqrt{e} = 0.0092, gl = 105$

método	n	F	$grupo$
$MostFreq$	28	0.953355	a
$PopeUngar$	28	0.948682	a
$RLUM$	53	0.914427	a
$RLDM$	53	0.906316	a

BioP – modelo:$\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 0$

método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a
$MostFreq$	1	1.00000	a
$RLDM$	1	1.00000	a
$RLUM$	53	1.00000	$dp = 0.000$

$L_s(n_{ij})$, $\sqrt{e} = 0.0136, gl = 105$

método	n	$prec$	$grupo$
$MostFreq$	28	1.000000	a
$PopeUngar$	28	1.000000	a
$RLDM$	53	0.982556	a
$RLUM$	53	0.896655	b

BioP – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0003, gl = 158$

método	n	rec	$grupo$
$RLUM$	53	1.000000	a
$MostFreq$	54	0.007212	b
$RLDM$	54	0.005068	b
$PopeUngar$	54	0.001365	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0798, gl = 158$

método	n	rec	$grupo$
$RLUM$	53	1.000000	a
$RLDM$	54	0.883155	b
$MostFreq$	54	0.482521	c
$PopeUngar$	54	0.480377	c

BioP – modelo: $\hat{F_a} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0039, gl = 158$

método	n	F_a	$grupo$
$RLUM$	53	0.036176	a
$PopeUngar$	54	0.000000	b
$RLDM$	54	0.000000	b
$MostFreq$	54	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0044, gl = 158$

método	n	F_a	$grupo$
$RLUM$	53	0.036176	a
$RLDM$	54	0.009972	b
$PopeUngar$	54	0.000000	b
$MostFreq$	54	0.000000	b

GeoP – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0007, gl = 1$

método	n	F	$grupo$
$RLUM$	50	0.916928	a
$MostFreq$	2	0.802159	a
$PopeUngar$	2	0.776120	a
$RLDM$	1	0.040404	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0172, gl = 107$

método	n	F	$grupo$
$RLUM$	50	0.916928	a
$MostFreq$	30	0.904588	a
$PopeUngar$	30	0.902852	a
$RLDM$	51	0.888623	a

GeoP – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0009, gl = 3$

método	n	$prec$	$grupo$
$RLDM$	1	1.000000	a
$RLUM$	50	0.907434	b
$MostFreq$	3	0.666667	c
$PopeUngar$	3	0.666667	c

$L_s(n_{ij})$, $\sqrt{e} = 0.0223, gl = 109$

método	n	$prec$	$grupo$
$RLDM$	51	1.000000	a
$PopeUngar$	31	1.956989	$a\ b$
$MostFreq$	31	1.956989	$a\ b$
$RLUM$	50	0.907434	$a\ b$

GeoP – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0073, gl = 149$

método	n	rec	$grupo$
$RLUM$	50	0.996000	a
$MostFreq$	51	0.028098	b
$PopeUngar$	51	0.027087	b
$RLDM$	51	0.000404	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0730, gl = 149$

método	n	rec	$grupo$
$RLUM$	50	0.996000	a
$RLDM$	51	0.860535	b
$MostFreq$	51	0.510124	c
$PopeUngar$	51	0.509113	c

GeoP – modelo: $\hat{F_a} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0076, gl = 149$

método	n	F_a	$grupo$
$RLUM$	50	0.051354	a
$PopeUngar$	51	0.000213	b
$MostFreq$	51	0.000213	b
$RLDM$	51	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0076, gl = 149$

método	n	F_a	$grupo$
$RLUM$	50	0.051354	a
$PopeUngar$	51	0.000420	b
$MostFreq$	51	0.000420	b
$RLDM$	51	0.000000	b

Table 18. Resultados das medidas de interesse para BioP e GeoP.

Mech – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.555556	a
$PopeUngar$	1	0.555556	a
$RLDM$	1	0.016949	b
$RLUM$	46	0.905300	$dp = 0.2562$

$L_s(n_{ij})$, $\sqrt{e} = 0.0116, gl = 81$

método	n	F	$grupo$
$MostFreq$	19	0.922222	a
$PopeUngar$	19	0.922222	a
$RLUM$	46	0.905314	a
$RLDM$	47	0.893066	a

Mech – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0000, gl = 0$

método	n	$prec$	$grupo$
$PopeUngar$	1	1.00000	a
$MostFreq$	1	1.00000	a
$RLDM$	1	1.00000	a
$RLUM$	46	0.8910	$dp = 0.2813$

$L_s(n_{ij})$, $\sqrt{e} = 0.0220, gl = 81$

método	n	$prec$	$grupo$
$MostFreq$	19	1.000000	a
$PopeUngar$	19	1.000000	a
$RLDM$	47	1.000000	a
$RLUM$	46	0.891024	b

Mech – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0007, gl = 137$

método	n	rec	$grupo$
$RLUM$	46	1.000000	a
$PopeUngar$	47	0.008183	b
$MostFreq$	47	0.008183	b
$RLDM$	47	0.000182	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0773, gl = 137$

método	n	rec	$grupo$
$RLUM$	46	1.000000	a
$RLDM$	47	0.861529	b
$MostFreq$	47	0.359247	c
$PopeUngar$	47	0.359247	c

Mech – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0095, gl = 137$

método	n	F_a	$grupo$
$RLUM$	46	0.060359	a
$PopeUngar$	47	0.000000	b
$RLDM$	47	0.000000	b
$MostFreq$	47	0.000000	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0095, gl = 137$

método	n	F_a	$grupo$
$RLUM$	46	0.060359	a
$PopeUngar$	47	0.000000	b
$RLDM$	47	0.000000	b
$MostFreq$	47	0.000000	b

Quan – modelo: $\hat{F} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.000, gl = 0$

método	n	F	$grupo$
$MostFreq$	1	0.573913	a
$PopeUngar$	1	0.573913	a
$RLDM$	1	0.047619	b
$RLUM$	39	0.967600	$dp = 0.1414$

$L_s(n_{ij})$, $\sqrt{e} = 0.0115, gl = 72$

método	n	F	$grupo$
$RLUM$	39	0.967562	a
$MostFreq$	18	0.965217	a
$PopeUngar$	18	0.965217	a
$RLDM$	40	0.909484	a

Quan – modelo: $\hat{P} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.000, gl = 0$

método	n	$prec$	$grupo$
$MostFreq$	1	1.000000	a
$PopeUngar$	1	1.000000	a
$RLDM$	1	1.000000	a
$RLUM$	50	0.960600	$dp = 0.1718$

$L_s(n_{ij})$, $\sqrt{e} = 0.0078, gl = 72$

método	n	$prec$	$grupo$
$PopeUngar$	18	1.000000	a
$MostFreq$	18	1.000000	a
$RLDM$	40	1.000000	a
$RLUM$	39	0.960570	a

Quan – modelo: $\hat{R} = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0011, gl = 116$

método	n	rec	$grupo$
$RLUM$	39	0.994872	a
$MostFreq$	40	0.010061	b
$PopeUngar$	40	0.010061	b
$RLDM$	40	0.000610	b

$L_s(n_{ij})$, $\sqrt{e} = 0.0846, gl = 116$

método	n	rec	$grupo$
$RLUM$	39	0.994872	a
$RLDM$	40	0.889360	a
$PopeUngar$	40	0.426728	b
$MostFreq$	40	0.426728	b

Quan – modelo: $\hat{F}_a = \hat{\mu} + \hat{m} + \hat{\epsilon}$

$L_g(n_i) \bigcup L_s(n_{ij})$, $\sqrt{e} = 0.0006, gl = 116$

método	n	F_a	$grupo$
$RLUM$	39	0.011072	a
$PopeUngar$	40	0.000000	a
$RLDM$	40	0.000000	a
$MostFreq$	40	0.000000	a

$L_s(n_{ij})$, $\sqrt{e} = 0.0006, gl = 116$

método	n	F_a	$grupo$
$RLUM$	39	0.011072	a
$PopeUngar$	40	0.000000	a
$RLDM$	40	0.000000	a
$MostFreq$	40	0.000000	a

Table 19. Resultados das medidas de interesse para Mech e Quan.