Utilização de sumários humanos no modelo Cassiopeia

Caio Cesar N. Delgado[1], Helton Damascena Dias[1], Marcus Vinicius C. Guelpeli[2]
[1] Departamento de Engenharia da Computação
Centro Universitário de Barra Mansa (UBM) –Volta Redonda, RJ – Brasil
[2] Departamento de Computação
Universidade Federal dos Vales do Jequitinhonha e Mucuri – (UFVJM) –
Diamantina, MG – Brasil
{caiodelgado.new, heltondd}@ gmail.com, marcus.guelpeli@ufvjm.edu.br

Resumo. *Este trabalho tem como objetivo analisar o desempenho de sumários humanos no modelo Cassiopeia. Originalmente o modelo Cassiopeia usa no seu pré-processamento sumários automáticos, a proposta deste trabalho é avaliar o uso de sumários humanos. Desta forma seria possível comparar, qual seria o melhor sumário, humano ou automático para obter os melhores resultados nos agrupamento criados pelo modelo Cassiopeia. Serão usados para esta simulação os corpora nos idiomas portugueses e ingleses e o desempenho será medido em termos das métricas externas e internas.*

1. Introdução

Com o avanço da tecnologia, a alta velocidade de acesso e a propagação de dados na internet valorizam-se muito a área de mineração de textos - MT devido ao alto volume de dados encontrado na internet, é de suma importância o aperfeiçoamento das técnicas de agrupamento de dados para fins de refinamento de pesquisas com o intuito de trazer ao usuário resultados mais concisos e precisos para suas pesquisas

Através de estudos [Guelpeli 2012], desenvolveu um modelo de agrupamento denominado Cassiopeia. Neste modelo os agrupamentos são criados com a utilização de sumários automáticos, de forma que independente do seu domínio e idioma possa ser realizado o agrupamento, com isso trazendo um diferencial e um ganho com relação a outros métodos de agrupamento de texto encontrados na literatura.

Partindo do modelo proposto por [Guelpeli 2012], surge uma questão. Com a utilização de sumários humanos usados no pré-processamento existiria um ganho considerável com relação aos sumários automáticos para o modelo Cassiopeia? Com base nesta indagação, foram realizados testes com a utilização de sumários humanos e sumários automáticos, a fim de obter resultados dos quais possam ser analisados e comparados

2. Modelo Cassiopeia

O modelo Cassiopeia, mostrado na Figura 1, foi proposto para ser um agrupador de texto hierárquico, com novo método para definição do corte de *Luhn*. O detalhamento de cada uma das três etapas (pré-processamento, processamento e pós-processamento) que compõem este modelo pode ser encontrado em detalhes no trabalho de [Guelpeli 2010 e Guelpeli 2012]. Para um melhor entendimento deste trabalho, uma visão macro do seu funcionamento será descrita a seguir.

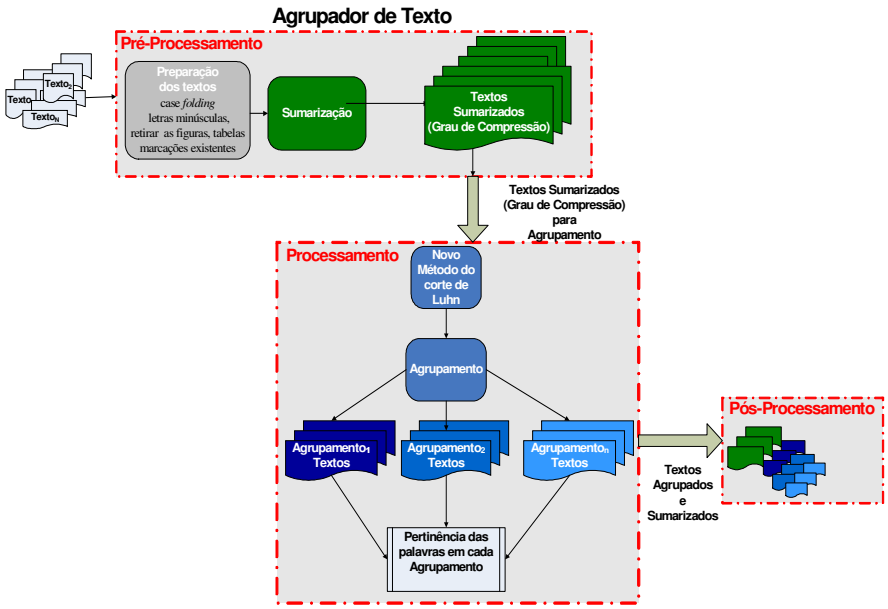


Figura 1. Modelo Cassiopeia .

O processo começa com a entrada de textos, que passam pela etapa de pré-processamento, na qual são preparados para o processo computacional, utilizando-se a técnica case *folding*, que coloca todas as letras em minúsculas, além de outros cuidados, como eliminação de todas as figuras, tabelas e marcações existentes. Os textos ficam em formato compatível para serem processados. Nesta etapa, é usado o processo de sumarização, cuja finalidade é diminuir o número de palavras, melhorando a qualidade do processamento. Com o processo de sumarização, obtém-se a parte mais importante, ou seja, a ideia principal do texto-fonte, através da criação de um sumário com as palavras mais significativas. Além de ser mais conciso do que o texto-fonte, o sumário tem um número muito menor de atributos, ou seja, palavras. Essa redução possibilita o uso de um espaço amostral que consegue atenuar a questão da alta dimensionalidade e dos dados esparsos, problema significativo na área de MT. A sumarização também consegue viabilizar a permanência das *stopwords*, o que possibilita que o modelo Cassiopeia seja independente do idioma.

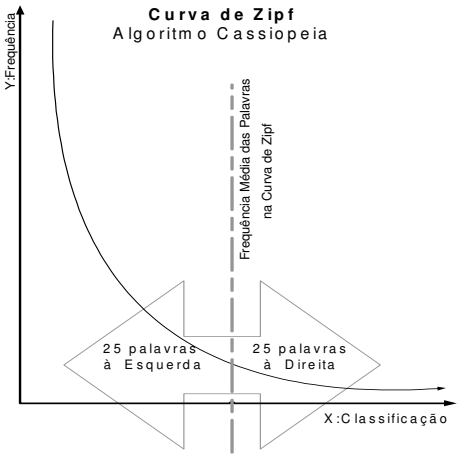
Terminada a etapa de pré-processamento, começa a de processamento onde são usadas a técnica de agrupamento de textos hierárquicos e o algoritmo *Cliques* mostrado na figura 3 para agrupar os textos com similaridade. Os agrupamentos criados nesta etapa têm um vetor de palavras de alta relevância para cada agrupamento. Os vetores são pertinentes em relação à frequência média das palavras nos textos agrupados. O modelo Cassiopeia identifica as características das palavras no documento, utilizando a frequência relativa, que define a importância de uma palavra, de acordo com a frequência com que é encontrado no documento. Quanto mais uma palavra aparecer em um documento, mais importante é, para aquele documento. A frequência relativa é calculada por meio da Equação 1, fórmula que normaliza o resultado da frequência absoluta das palavras, evitando que documentos pequenos sejam representados por vetores pequenos e documentos grandes, por vetores grandes. Com a normalização,

todos os documentos serão representados por vetores de mesmo tamanho, como mostra a Equação 1.

$$F_{\mathrm{r}}X = \frac{F_{\mathrm{abs}}X}{\mathrm{N}} \tag{1}$$

Onde $F_{\mathbf{r}}X$ é igual à frequência relativa de *X*, $F_{\mathbf{abs}}X$ é igual à frequência absoluta de *X*, ou seja, a quantidade de vezes que *X*, a palavra aparece no documento e *N* é igual ao número total de palavras no documento. Considerado um espaço-vetorial, cada palavra representa uma dimensão (existem tantas dimensões quantas palavras diferentes no documento).

À medida que novos textos são agrupados ocorre o reagrupamento, podendo surgir agrupamentos, subagrupamentos ou até mesmo a fusão destes. As palavras colocadas nos vetores são calculadas pela média da sua frequência no texto e selecionadas, de acordo com a Figura 2 e o algoritmo 1.



1 - Algoritmo do método Cassiopeia:

1. *Estabelecer frequência média do conjunto P de palavras do documento baseado na Curva de Zipf*

$$f(P,N) = \frac{\sum_{n=1}^{N} F_{\mathrm{r}}P_{\mathrm{n}}}{N}$$

2. *Escolher as 25 palavras à esquerda da média e as 25 palavras à direita da média.*

Figura 2. Seleção dos atributos no modelo Cassiopeia.

Onde: *N* é igual ao número total de palavras no documento; $F_{\mathrm{r}}P_{\mathrm{n}}$ é igual à frequência relativa de P_n; onde *P* é o conjunto de palavras no documento e P_n refere-se a quantidade de vezes que uma palavra aparece no documento e $f(P,N)$ é a frequência média das palavras na distribuição.

Os vetores de palavras ou centroides dos agrupamentos, por questão de dimensionalidade, adotam uma truncagem que, segundo [Wives 2004], é de 50 posições, não sendo necessário um valor maior. Essas palavras devem estar ordenadas da maior para a menor, com base na ocorrência da frequência média de cada palavra no conjunto de documentos.

Esses agrupamentos estão organizados de uma forma *top-down*, ou seja, seus textos são organizados em agrupamentos, que são particionados, sucessivamente, produzindo uma representação hierárquica, que facilita a visualização dos agrupamentos a cada ciclo

de processamento, bem como o grau de similaridade obtido entre documentos com uso do algoritmo *Cliques*. Este é um método que, de início, não requer definições de número de agrupamentos. O seu reagrupamento ocorre até o momento em que os centroides de cada agrupamento estejam estáveis, ou seja, não sofram mais alterações, com a inclusão de novos textos. Para determinar a similaridade desses textos nos agrupamentos é utilizado o algoritmo *Cliques* mostrado na Figura 3 descrito pelo algoritmo 2.

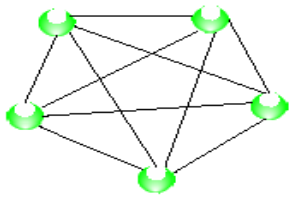


Figura 3. Algoritmo Cliques.

2 - Algoritmo Cliques:

1. *Seleciona 1º elemento e coloca em um novo cluster.*
2. *Procura o próximo objeto similar.*
3. *Se objeto é similar a todos os outros elementos do cluster, este objeto é agrupado.*
4. *Voltar ao passo 2, enquanto houver objetos;*
5. *Para os elementos não alocados, repetir o 1.*

Terminada a etapa de processamento, começa a de pós-processamento, na qual cada um dos agrupamentos ou subagrupamentos terá, por similaridade, um conjunto de textos sumarizados, que contém um número bem menor de sentenças, com alto grau de informatividade, e com as ideias principais dos textos-fonte, característica da sumarização.

O trabalho de [Guelpeli 2012] usou na fase pré-processamento apenas sumários obtidos de sumarizadores automáticos. Neste trabalho será usada também a entrada de textos sumarizados por humanos para realizar a comparação e verificar qual o melhor sumário para o modelo Cassiopeia.

3. Metodologia

A contribuição deste estudo será garantir a utilização do sumário de melhor desempenho para o modelo Cassiopeia. Para comprovar esta contribuição, foi criada a seguinte metodologia: foram usados *Corpora* que são descritos na seção 3.1 deste trabalho. Foram escolhidos os sumarizadores citados na seção 3.2. Para o cálculo do percentual compressão de cada sumário, foi utilizada a aplicação da regra de três a qual será explicada na seção 3.3. O modelo de sumarização e agrupamento desenvolvido será descrito na seção 3.4. Para a análise dos resultados foram escolhidas as métricas externas e internas que são utilizadas na mensuração do processo de agrupamento e comentadas detalhadamente na seção 3.5.

3.1 Corpus

O *coprora* usado neste trabalho é o mesmo utilizado por [Guelpeli 2012]. Serão utilizadas cem amostras de textos na língua Inglesa colhidos da *Reuters*, cem amostras de textos na língua Portuguesa do *corpus* Temário [Pardo e Rino 2004], cem amostras de sumários humanos para cada texto de cada *corpus* (*Reuters* e TeMario). Foram

usados os sumarizadores descritos na seção3.2 para cada texto do *corpus Reuters* e TeMario .

3.2 Sumarizadores

A sumarização será realizada para os textos *Reuters* com o uso de dois sumarizadores profissionais da língua inglesa (*Copernic Summarizer* e *Intellexer Summarizer Pro*). Para os textos do *corpus* TeMario foram usados sumarizadores, da literatura para língua portuguesa (SuPor 2) e um sumarizador profissional da língua inglesa (*Intellexer Summarizer Pro*), mas que tem a possibilidade de sumarizar na língua portuguesa.

3.3Calculo da compressão

Foi definido o percentual de compressão conforme a Equação 2, com base em uma aplicação matemática da regra de três, onde se relacionam quatro valores divididos entre dois pares de mesma grandeza e unidade, conforme ilustrado na Tabela1. As grandezas aplicadas são o número de palavras e o percentual de compressão, as quais são diretamente proporcionais. Para a extração do numero de palavras de cada texto fonte e cada sumário humano, para realização dos cálculos de percentual, foi utilizado o software [*FineCount* 2.6] o qual realiza a contagem de palavras.

$$C = \frac{P_{sh}\, x\, 100}{P_{tf}} \qquad \textbf{(2)}$$

Onde: C é o percentual de compressão, P_{sh} é o número de palavras do sumário humano e P_{tf} é o número de palavras do texto fonte.

Tabela 1 Regra de três simples para obtenção do percentual de compressão.

Número de Palavras	Percentual de Compressão
P_{tf}	100
P_{sh}	C

Com a aplicação da Equação 2, é obtido o percentual de compressão individual de cada sumário humano, podendo assim efetuar uma sumarização automática de cada texto fonte com o mesmo percentual de compressão do sumário humano, garantindo assim a integridade dos testes.

3.4 Modelo de Sumarização e Agrupamento

A sumarização automática será realizada baseada nos sumários humanos, ou seja, aplicando-se os percentuais de compressão obtidos pela Equação 2 em todos os textos fontes.

Desta forma a sumarização automática foi feita no *corpus* em português o TeMario, com os sumarizadores *Intellexer Summarizer Pro* e SuPor 2, e no *corpus* em inglês o *Reuters*, com os sumarizadores *Intellexer Summarizer Pro* e *Copernic Summarizer* obtendo assim, sumários automáticos com equivalência no número de palavras em relação aos sumários humanos. Após ser realizada a sumarização em todas as amostras de texto foi feito o agrupamento no modelo Cassiopeia, onde são utilizadas as métricas de coesão, acoplamento, coeficiente s*ilhouette* e *precision*, *recall, Fmeasure* que serão descritas na seção 3.5. Foram realizadas cem execuções de agrupamento em cada

conjunto de textos de cada idioma utilizado, como ilustrado nas Figura 4, para se obter os valores das medidas internas e externas e suas respectivas médias acumuladas.

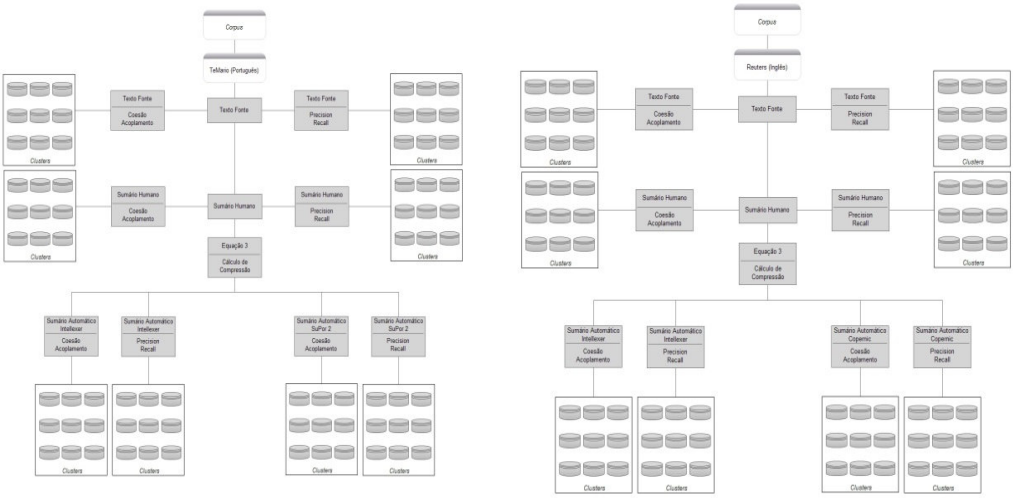


Figura 4. Modelo de Sumarização e Agrupamento (TeMario) e (Reuters).

3.5 Mensuração

A mensuração dos resultados será realizada com a aplicação das métricas internas (coesão, acoplamento e coeficiente *silhouette*) e métricas externas (*recall, precision* e *F-measure*) descritas a seguir.

3.5.1 Coesão

A coesão mede a similaridade entre os elementos do mesmo agrupamento. Quanto maior a similaridade entre eles, maior a coesão deste agrupamento [Kunz e Black 1995].

$$C=\frac{\sum_{i>j} Sim(P_i,P_j)}{\frac{n(n-1)}{2}} \tag{3}$$

Onde *Sim* (P_i,P_j) é o calculo da similaridade entre os textos *i* e *j* pertencentes ao agrupamento *P*, *n* é o número de textos *P*, e P_i e P_j são membros do agrupamento *P*.

3.5.2 Acoplamento

O acoplamento mede a similaridade média de todos os pares de elementos, sendo que um elemento pertence a um agrupamento e outro não pertence a esse mesmo agrupamento [Kunz e Black 1995].

$$A=\frac{\sum_{i>j} Sim(C_i,C_j)}{\frac{n_a(n_a-1)}{2}} \tag{4}$$

Onde *C* é o centroide de determinado agrupamento, presente em *P*, *Sim* (*Ci, Cj*) é o cálculo da similaridade do texto *i* pertencente ao agrupamento *P* e o texto *j* não pertence

a *P*, *Ci* centroide do agrupamento *P* e *Cj* é centroide do agrupamento *Pi* e na é o número de agrupamentos presentes em *P*.

3.5.3 Coeficiente Silhouette

O Coeficiente *Silhouette* baseia-se na ideia de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante de outro grupo. Assim, essa medida combina as medidas de coesão e acoplamento [Zoubi e Rawi 2008].

$$S = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{5}$$

Onde *a (i)* é a distância média entre o *i-ésimo* elemento do grupo e os outros do

$$\overline{S} = \frac{1}{N}\sum_{i=1}^{N} S$$

mesmo grupo. O *b(i)* é o valor mínimo de distância entre o *i-ésimo* elemento do grupo e qualquer outro grupo, que não contém o elemento, e *max* é a maior distância entre *a (i)* e *b(i).*

$$\overline{s} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N} S \tag{6}$$

O Coeficiente *Silhouette* de um grupo é a média aritmética dos coeficientes calculados para cada elemento pertencente ao grupo, sendo apresentado na Equação 10, o valor de *S* situa-se na faixa de 0 a 1.

3.5.4 Recall

O Recall mede a proporção de objetos corretamente alocados a um agrupamento, em relação total de objetos da classe associada a este agrupamento [Manning *et al* 2008].

$$R = \frac{n(A)}{n(A \cup D)} \tag{7}$$

Onde *n(A)* é o número de elementos do subconjunto *A* de acertos e *n(D)* é o número de elementos do subconjunto *D* de falsos negativos[1] e $n(A \cup D)$ é o número total de elementos da classe correspondente.

3.5.5 Precision

A Precision mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos deste agrupamento [Manning *et al* 2008].

$$P = \frac{n(A)}{n(A \cup B)} \tag{8}$$

[1] Falsos negativos são elementos que deveriam ter sido alocados a um grupo e que foram alocados a outros.

Onde *n(A)* é o número de elementos do subconjunto de A de acertos e n(B) é o número de elementos do subconjunto B de falsos positivos e *n(A ∪B)* é o número total de elementos do grupo.

3.5.6 F-Measure

O *F-Measure* é a medida harmônica entre o Precision e o Recall que, no F-Measure, assume valores que estão no intervalo de [0,1]. O valor zero indica que nenhum objeto foi agrupado corretamente, o valor um, que todos os objetos estão contidos corretamente agrupados. Assim, um agrupamento ideal deve retornar um valor igual a um. [Manning *et al* 2008].

$$F = 2 * \frac{Precision(P) * Recall(R)}{Precision(P) + Recall(R)} \quad \textbf{(9)}$$

Cada uma das medidas descritas é calculada para cada um dos grupos obtidos, fornecendo assim a qualidade de cada grupo. A medida de avaliação, para todo o agrupamento, é obtida através do cálculo da média entre cada uma das medidas de todos os grupos.

4. Resultados Obtidos nos Experimentos

Devido ao grande volume das medidas utilizadas para aferir os agrupamentos dos experimentos e consequentemente dos resultados obtidos, serão apresentados os gráficos das medidas harmônicas, na métrica externa que é *F-Measure* e na métrica interna que é o Coeficiente *Silhouette.*

A Figura 5 mostra os resultados dos testes no idioma português usando *corpus* TeMario, com medida externa *F-Measure*. Observa-se que o sumário automático *Intellexer* obteve valores de *F-Measure* superiores ao sumario humano.

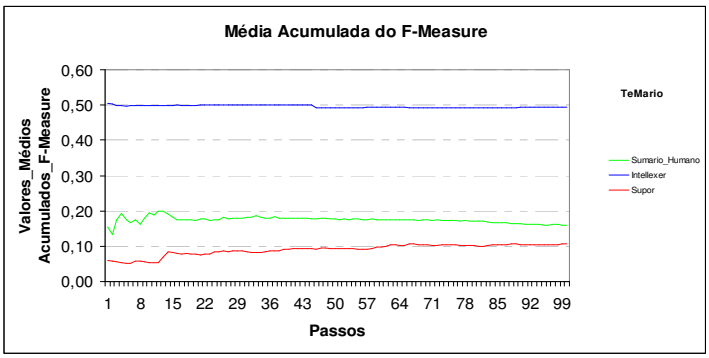


Figura 5. Resultados obtidos pelo modelo Cassiopeia usando a medida *F-Measure* no idioma português.

Na Figura 6 mostra os resultados dos testes no idioma português usando o *corpus* TeMario, com a medida interna Coeficiente *Silhouette*. Observa-se que os dois sumarizadores automáticos *Intellexer* e Supor obtiveram valores superiores ao sumario humano.

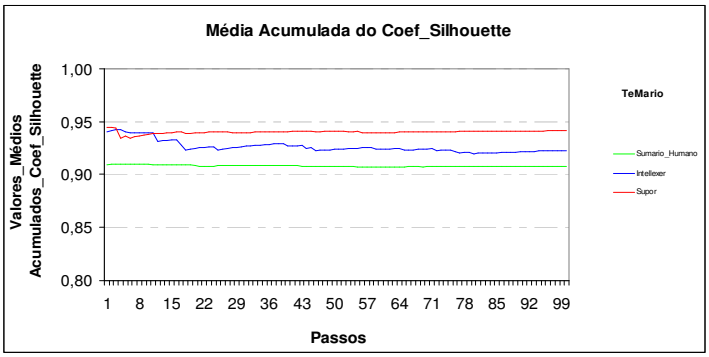


Figura 6. Resultados obtidos pelo modelo Cassiopeia usando a medida Coeficiente *Silhouette* no idioma português.

A Figura 7 mostra os resultados dos testes no idioma inglês usando *corpus Reuters*, com medida externa *F-Measure*. Observa-se que apenas nesta medida o sumário humano teve um desempenho superior aos sumários automáticos.

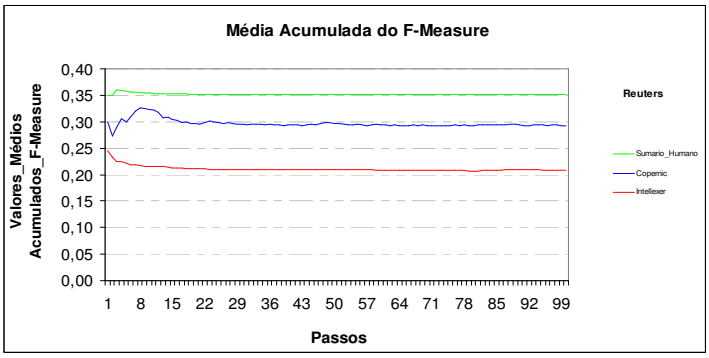


Figura 7. Resultados obtidos pelo modelo Cassiopeia usando a medida *F-Measure* no idioma inglês

Na Figura 8 mostra os resultados dos testes no idioma inglês usando o *corpus Reuters*, com a medida interna Coeficiente *Silhouette*. Observa-se que os dois sumarizadores automáticos *Intellexer* e Supor obtiveram valores superiores ao sumario humano.

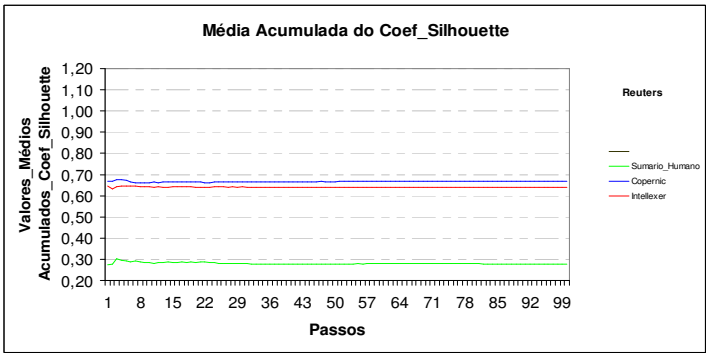


Figura 8: Resultados obtidos pelo modelo Cassiopeia usando a medida Coeficiente *Silhouette* no idioma inglês.

5. Conclusão

Observa-se que, nos resultados, a avaliação das métricas de *F-Measure* mostrados nas Figuras 5 e 7 e das métricas *Coeficiente Silhouette* nas Figuras 6 e 8, foram obtidos resultados comprobatórios que os sumários automáticos têm desempenho superior em 75% dos testes efetuados em relação ao sumário humano.

Em questão de desempenho pode-se analisar a praticidade e a velocidade superior ao desenvolver um sumário automático em comparação a de um sumário humano, pois o sumário humano necessita de alocar grupos de pessoas para realizar a sumarização, tomando tempo para análise do texto e retirada de pontos principais. Já o sumário automático não necessita de tanto tempo para a realização da sumarização. Após essa análise conclui-se que os sumários automáticos são mais eficientes que os sumários humanos, aplicados no modelo Cassiopeia nas questões de praticidade e qualidade dos agrupamentos gerados.

6. Referências

Guelpeli, M.V.C.; Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização. – Tese (doutorado) – Universidade Federal Fluminense. Programa de Pós-graduação em Computação, Niteroi, BR – RJ, Brasil, 2012.

Guelpeli, M.V.C.; e Garcia, A.C.B.; e Bernardini, F.C.; Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods. Extend an invitation to you to publish a chapter in the upcoming book on "Emergent Web Intelligence"published by Springer Verlag in the series Studies in Computational Intelligence, 2010.

Kunz, T.; and Black, J.P.; Using Automatic Process Clustering for Design Recovery and Distributed Debbuging. IEEE Trans. Software Eng.515-527, 1995.

Manning, C.D.; and Raghavan, P.; and Shutze, H.; Introduction to Information Retrieval, Cambridge University Press, 2008.

Pardo, T.A.S.; e Rino, L.H.M.; TeMário: Um Corpus para Sumarização Automática de Textos Relatórios Técnicos (NILC-TR-03-09). NILC – ICMC – USP. São Carlos, Brasil, 2004.

Wives, L.K.; Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados clustering de documentos – Tese (doutorado) – Universidade Federal do Rio Grande do Sul.Programa de Pós-graduação em Computação, Porto Alegre, Brasil, 2004.

Zoubi, M.B.; e Rawi, M.; An Efficient Approach for Computing *Silhouette* Coefficients. Journal of Computer Science Volume 4 Page No. 252-255, 2008.