Identificação de Autoria de Documentos Eletrônicos

Walter R. de Oliveira Jr.[1], Luiz E. S. Oliveira[2], Edson J. R. Justino[1]

[1]PRograma de Pós-Graduação em Informática – Pontifícia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1155 – 80.215-901 – Curitiba – PR – Brasil

[2]Departamento de Informática – Universidade Federal do Paraná (UFPR)
Curitiba, PR – Brasil

woliveirajr@gmail.com, justino@ppgia.pucpr.br, lesoliveira@inf.ufpr.br

Abstract. *The identification of authorship of electronic documents lies among the demands of forensic analysis. Data compressors and the Normalized Compression Distance (NCD) are tools that can help the expert to perform this task. In this work the performance of these tools were analyzed in a dataset of 3,000 electronic documents in Brazilian Portuguese, from 100 different authors, and correct attribution was made in more than 70% of the cases, indicating that this approach might have promising results. It was also verified the influence of the number of training documents in the results.*

Resumo. *Entre as demandas das perícias forenses está a identificação da autoria de documentos eletrônicos. Compressores de dados e a Distância Normalizada de Compressão (NCD) são ferramentas que auxiliam o perito a executar esta tarefa. Estudou-se o desempenho destas ferramentas em uma base de dados com 3000 documentos eletrônicos em português, de 100 diferentes autores, e foram obtidas taxas médias de acerto superiores a 70%, indicando que esta técnica é promissora. Verificou-se, também, a influência da quantidade de documentos de treinamento no desempenho desta técnica.*

1. Introdução

Entre as tarefas de perícia desenvolvidas em relação a documentos está a determinação de autoria de documentos. Nesta atividade busca-se determinar quem produziu um documento a partir da análise de elementos deste documento. Em relação a documentos eletrônicos, entretanto, pode haver dificuldades pela ausência de determinados elementos físicos que auxiliam o perito em sua tarefa. Por exemplo, não estão presentes elementos como a tinta utilizada para a escrita, a maneira como o autor grafa determinadas letras ou mesmo desgastes físicos de peças de impressoras ou máquinas datilográficas que deixem marcas em uma folha de papel.

Poderá ocorrer, também, que elementos comuns em documentos eletrônicos (como metadados de processadores de texto) não possam ser obtidos, e desta forma o perito tenha que utilizar apenas o conteúdo do documento para realizar sua perícia. Neste caso, o perito poderá utilizar a análise de características estilísticas do autor para a tarefa de atribuição de autoria. Considera-se que estas características estilísticas são produzidas de maneira inconsciente pelo autor e são pessoais, havendo variação da maneira como cada autor expressa suas ideias e pensamentos (Stamatatos, 2009 e Pinker, 2007).

Existem diversas características estilométricas que podem ser analisadas pelo perito. Stamatatos (2009) dividiu-as em características de caracteres, léxicas, sintáticas e específicas a atividades (são as características que consideram elementos de atividades específicas, por exemplo maneira como o autor formata um determinado conteúdo de um documento).

O uso de compressão de dados para a verificação de autoria é uma análise que utiliza características de caracteres, pois os compressores de dados utilizam informações sobre os símbolos encontrados para diminuir a quantidade de símbolos necessários para representar os mesmos dados (no caso de compressores sem perdas) ou os dados degradados em um nível controlado (no caso de compressores com perdas).

No presente trabalho são utilizados os compressores de dados para a tarefa de atribuição de autoria a documentos eletrônicos. São comparadas duas abordagens que utilizam compressores de dados. A primeira utiliza a distância normalizada de compressão (*normalized compression distance* - NCD), proposta por Cilibrasi e Vitányi (2005), e a segunda abordagem utiliza a complexidade condicional de compressão (*conditional complexity of compression* – CCC), proposta por Malyutov et al. (2007). Estas abordagens são detalhadas na seção 2.

A base de dados utilizada neste trabalho é a mesma que Varela (2010) utilizou em um trabalho prévio. Esta base de dados é melhor explicada no seção 3. Naquele trabalho, Varela (2010) atribuía a autoria de documentos eletrônicos com o uso de classificadores SVM e estatísticas de palavras-função presentes nos documentos. Por exemplo, para cada autor eram geradas estatísticas sobre os pronomes utilizados em seus documentos, e os perfis de cada autor eram considerados em um classificador SVM para a atribuição de autoria.

O método proposto para o presente trabalho é apresentado no seção 4, onde detalha-se como a NCD e a CCC foram utilizados para atribuição de autoria.

No seção 5 são apresentados os resultados obtidos. Como mencionado, o fato de utilizar a mesma base de dados de outro trabalho e buscar-se o desempenho da mesma tarefa, é possível efetuar uma comparação dos resultados obtidos por Varela (2010).

Por fim, no seção 6 serão apresentadas as conclusões obtidas com este trabalho e sugestões para trabalhos futuros.

2. Distância Normalizada de Compressão e Complexidade Condicional de Compressão

O uso de compressores para a verificação da autoria de documentos decorre da teoria de Complexidade de Kolmogorov (1965). De maneira resumida, a teoria da Complexidade de Kolmogorov afirma que a complexidade de uma determinada informação pode ser medida pela quantidade de símbolos necessários para expressar esta informação em alguma linguagem universal. Logo, a complexidade de Kolmogorov $K()$ de uma informação x é representada pelo tamanho da representação desta informação em uma linguagem universal.

É possível verificar a complexidade condicional de Kolmogorov $K(x|y)$. Dada uma informação x, sua complexidade condicional será o tamanho do programa y que é processado por uma máquina de Turing e gera a saída x novamente.

Estas complexidades são incomputáveis, pois não é possível saber se o menor

programa *y* foi gerado (Li e Vitányi, 1997).

2.1. Distância Normalizada de Compressão

A partir da teoria de Kolmogorov, Li e Vitányi (1997) propuseram que é possível verificar a similaridade entre duas informações, e que era possível apresentar esta similaridade em uma medida normalizada. A Distância Normalizada de Informação (NID – *Normalized Information Distance*) é expressa na equação 1 a seguir

$$NID(x, y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}} \quad (1)$$

sendo que *x* e *y* são as informações consideradas, *K(x|y)* é a complexidade condicional de Kolmogorov, *K(x)* é a complexidade de Kolmogorov da informação *x* e *max{}* é a função que retorna o maior entre dois valores.

Como *K(x)* é incomputável, a NID também o é.

Os mesmos autores propuseram que é possível aproximar a complexidade de Kolmogorov ao se utilizar compressores de dados. Para comprimir uma informação, o compressor processa esta informação e busca, através do uso de uma codificação otimizada, gerar um programa *y* que possa ser novamente processado e que gere a informação *x* como saída, e que este programa *y* tenha um tamanho menor que a informação *x* original.

Com o uso de compressores, é possível estabelecer uma Distância Normalizada de Compressão (NCD – *Normalized Compression Distance*), expressa na equação 2 a seguir.

$$NCD(x, y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}} \quad (2)$$

sendo que *min{ }* é a função que retorna o menor valor entre dois valores dados, *C(x)* é o tamanho da informação *x* após ser comprimido por um compressor *C*, *xy* é a concatenação da informação *x* com a *y* e os demais elementos já foram definidos anteriormente.

A NCD é uma medida normalizada de similaridade entre duas informações e seu resultado, idealmente, deveria estar entre 0 e 1. Cilibrasi e Vitányi (2005) mencionam que, para determinados compressores de dados, o valor obtido poderá ser ligeiramente superior a 1, quando pouca ou nenhuma compressão for obtida e o compressor não for otimizado para esta situação, utilizando uma grande quantidade de informações como cabeçalho da compressão. Valores de NCD próximos a 0 indicam uma maior similaridade entre duas informações, e idealmente a *NCD(x,x)* deveria ser exatamente 0, pois a similaridade entre duas informações idênticas é total.

Destaca-se que a compressão de dados para cálculo da NCD é feita utilizando-se o documento completo, tal como ele se apresenta. Não é necessário realizar a extração ou pré-seleção de características específicas a serem utilizadas. Isto é um diferencial, por exemplo, em relação ao uso de SVM como classificador, pois para a utilização do SVM requer que características sejam selecionadas para a construção dos vetores de suporte.

2.2. Complexidade Condicional de Compressão

Malyutov et al. (2007), também fundamentando-se na complexidade de Kolmogorov e

na sua aproximabilidade com o uso de compressores de dados, propuseram a medida da Complexidade Condicional de Compressão (CCC – *Conditional Complexity of Compression*).

Esta CCC é expressa na equação 3 a seguir.

$$CCC(x|y) = C(yx) - C(x) \quad (3)$$

sendo que *yx* é a concatenação da informação *y* com a informação *x* e os demais elementos já foram definidos anteriormente.

Para a medida de similaridade de informações, um valor menor de CCC indica que as duas informações concatenadas puderam ser comprimidas melhor do que outras duas informações que apresentem um CCC maior, e portanto o compressor de dados conseguiu gerar um modelo melhor de compressão por ter encontrado elementos mais semelhantes entre as duas informações.

Nesta medida também não é feita a pré-seleção ou extração de características, pois os documentos a serem testados são considerados em sua integralidade.

2.3. Compressores de dados

Para aproximação da complexidade de Kolmogorov são utilizados compressores de dados. Não há, entretanto, determinação sobre qual compressor deve ser utilizado pelos métodos NCD e CCC, então foram utilizados três compressores de dados, com três métodos de compressão diferentes: método baseados em estatísticas, em dicionários e em blocos de dados.

O compressor estatístico utilizado foi o PPM-D (*prediction by partial matching D*). Este compressor gera um modelo estatístico dos símbolos encontrados no documento considerando a probabilidade da ocorrência de um símbolo em função da ocorrência anterior de outros símbolos (Shkarin, 2002). Por exemplo, na língua portuguesa, após um símbolo “ç”, há uma maior probabilidade de encontrar um símbolo “ã” ou “õ” do que um símbolo “r”, e estas probabilidades são consideradas para a geração de modelos estatísticos otimizados.

Para compressores baseados em dicionários foi escolhido o compressor Zip. Este compressor implementa o algoritmo de compressão baseado em dicionário LZ77 (Lempel-Ziv77) e uma codificação de Huffmann. De maneira simplificada, os compressores baseados em dicionário utilizam as informações visualizadas anteriormente no documento como uma tabela de dicionário, e quando esta informação é encontrada novamente, é feita apenas uma referência à tabela.

Por fim, o compressor baseado em bloco de dados utilizado é o compressor Bzip, que implementa o algoritmo de compressão de blocos de Burrows e Wheeler (Burrows e Wheeler, 1994). Este compressor caracteriza-se por separar o conteúdo a ser comprimido em blocos de dados e efetuar operações em seu conteúdo, de maneira a agrupar informações semelhantes e assim conseguir representá-las de uma maneira que possam ser codificados mais eficientemente por alguma codificação posterior. No caso do compressor Bzip, é utilizada a codificação de Huffmann.

3. Base de dados

Conforme menciona Stamatatos (2009), uma característica desejável em uma base de

dados para pesquisas de autoria de documentos é que a autoria dos documentos seja previamente conhecida, permitindo que os resultados obtidos possam ser conferidos e a taxa de acerto de um determinado método possa ser verificada. Também é desejável que testes sejam realizados sobre as mesmas bases de dados, controladas, com protocolos iguais ou semelhantes. Desta maneira os resultados de diferentes métodos podem ser comparados, permitindo uma verificação de ganhos de um determinado método em relação a métodos utilizados anteriormente.

Neste trabalho foi usada a mesma base de dados utilizada anteriormente por Varela (2010). Esta base de dados é composta por 3.000 documentos, de 100 diferentes autores, com 30 documentos de cada autor. Estes autores e documentos estão separados em 10 categorias: direito, economia, esportes, gastronomia, literatura, politica, saúde, tecnologia, turismo e assuntos gerais. O tamanho médio dos documentos é de 2989 bytes com um desvio padrão de 1531, com um valor máximo de 735 *tokens* (palavras) em um documento.

Para os testes os documentos foram separados em dois grupos. O primeiro grupo era composto pelos documentos de perfil dos autores. Cada autor era representado por 7 documentos, escolhidos aleatoriamente, em um total de 70 documentos por área. O segundo grupo compunha o conjunto de testes e era composto pelos documentos restantes, com 23 documentos sendo utilizados para cada autor em um total de 230 documentos testados por área, em um total de 2300 documentos testados.

4. Método

Para a atribuição de autoria dos documentos de teste foi utilizado o seguinte método. Inicialmente os documentos de cada área foram separados em dois grupos (perfil do autor e documentos de teste), conforme mencionado na seção anterior.

A seguir foram efetuados testes com os dois métodos de cálculo de similaridade entre documentos: NCD e CCC. Para cada um dos métodos foram utilizados os três compressores de dados mencionados: Zip, Bzip e PPM-D.

Para cada teste, o documento testado tinha sua medida NCD ou CCC efetuada em relação a todos os documentos de perfil de autor daquela área. Ou seja, para cada documento testado eram obtidas 70 medidas, sendo que cada autor era representado por 7 medidas. Para cada teste, em cada área de conhecimento, eram possíveis 10 autores, e a probabilidade de um documento ser atribuído aleatoriamente a um autor era de 10%.

Após obtidos os valores NCD e CCC de cada teste, era necessário utilizar algum mecanismo de escolha do resultado para a atribuição de autoria ao documento. Neste trabalho foram utilizados dois mecanismos de escolha: melhor valor e votação.

A escolha pelo melhor valor é feita considerando-se o melhor valor de cada teste. Por exemplo, um documento *x* era testado em relação a 7 documentos do autor A, 7 documentos do autor B, e assim por diante, gerando um total de 70 medidas. Dentre estas medidas, verificava-se qual era o melhor resultado, e a atribuição de autoria era feita ao autor cujo documento de perfil apresentou o melhor resultado.

A escolha por votação, por sua vez, é feita considerando-se o número de votos que cada autor recebeu. Determinou-se que seriam computados como votos os sete melhores resultados de medida, e a atribuição de autoria era feita ao autor mais votado. Por exemplo, supondo que um documento *x* teve as 7 melhores medidas em relação aos

autores “B, A, A, C, D, A, B”. O autor mais votado foi o autor *A*, e a autoria era atribuída a ele. Em caso de empate, a autoria foi atribuída ao autor que estivesse com o melhor resultado. Por exemplo, se os votos fossem para os autores “B, A, C, A, B, D, E”, os autores *B* e *A* teriam a mesma quantidade de votos e o autor *B* seria escolhido.

5. Resultados e Análise

São apresentados, a seguir, os resultados dos testes efetuados.

5.1. Escolha pelo melhor resultado

O primeiro teste efetuado utilizou a atribuição de autoria pelo melhor resultado dos métodos CCC e NCD, com os três compressores de dados mencionados.

Os testes eram efetuados entre os documentos de cada área, havendo 10 autores prováveis para cada documento testado. Foram obtidos os seguintes resultados, expressos na tabela 1. A coluna SVM refere-se aos resultados obtidos por Varela (2010) utilizando um classificador SVM e estatísticas de palavras-função encontradas no documento.

Tabela 1. Escolha de autor pelo melhor resultado

		NCD			CCC		
Área	SVM	Bzip	PPM-D	Zip	Bzip	PPM-D	Zip
Assuntos Variados	70,70%	79,57%	81,74%	**83,04%**	69,70%	72,16%	73,02%
Direito	**72,20%**	63,91%	68,26%	65,65%	52,75%	60,75%	57,32%
Economia	64,80%	77,83%	78,70%	**79,57%**	68,63%	67,47%	72,77%
Esportes	68,30%	82,61%	85,65%	**87,39%**	75,74%	74,10%	80,78%
Gastronomia	**75,70%**	44,78%	54,35%	53,04%	34,48%	45,80%	47,51%
Literatura	**72,20%**	59,13%	66,96%	61,74%	47,27%	56,97%	50,69%
Política	68,70%	81,74%	**83,91%**	83,04%	71,02%	72,84%	77,38%
Saúde	**72,20%**	58,26%	61,30%	63,91%	50,07%	54,72%	52,57%
Tecnologia	73,90%	74,78%	77,39%	**79,13%**	64,75%	67,04%	70,29%
Turismo	78,30%	80,00%	82,17%	**83,04%**	69,41%	72,63%	73,95%
Média	71,70%	70,26%	**74,04%**	73,96%	60,38%	64,45%	65,63%
Desvio padrão (pontos percentuais)	3,86	12,91	10,67	11,75	13,38	9,53	12,27

Em negrito estão destacados os melhores resultados de cada área.

Verifica-se que para a escolha pelo melhor resultado, o método que apresenta uma maior taxa de acerto média é o NCD com o uso do compressor PPM-D. O NCD com o compressor Zip apresenta o melhor resultado em 5 categorias, seguido pelo método SVM.

O método CCC apresentou resultados inferiores de aproximadamente 12 pontos percentuais aos obtidos pelo método NCD. Em duas áreas (Tecnologia e Turismo) o método CCC teve resultado inferior ao classificador SVM enquanto o NCD apresentava um resultado superior. A figura 1 ilustra o desempenho médio e o desvio padrão dos

métodos NCD e CCC com cada compressor.

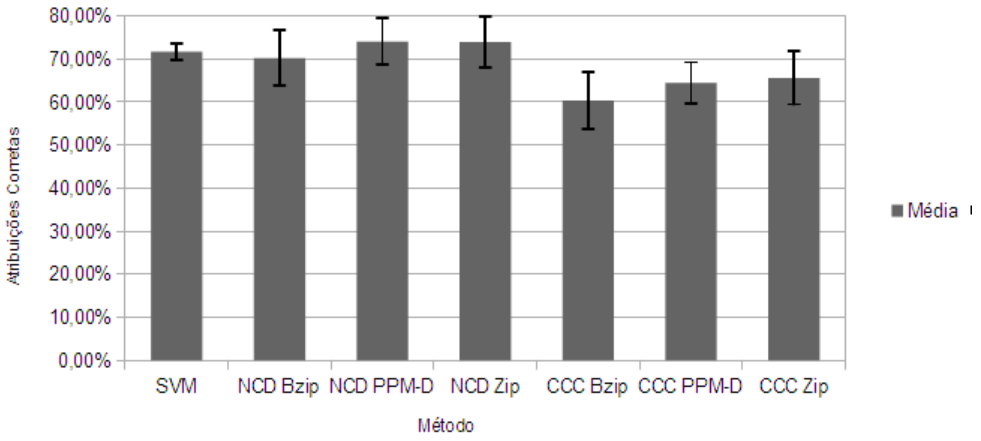

Figura 1. Escolha de autor pelo melhor resultado

A diferença de desempenho entre o método NCD com compressor PPM-D e com compressor Zip foi de aproximadamente 0,1 ponto percentual, e no gráfico esta diferença é imperceptível.

O método que utilizou o classificador SVM apresentou o menor desvio padrão, e conforme pode ser visto pelo desempenho em cada área, os seus resultados foram mais homogêneos.

5.2. Escolha por votação

O segundo teste efetuado utilizou a atribuição de autoria através de votação. Os resultados obtidos estão expressos na tabela 2 a seguir.

Tabela 2. Escolha de autor por votação

		NCD			CCC		
Tema	SVM	NCD Bzip	NCD PPM-D	NCD Zip	Bzip	PPM-D	Zip
Assuntos Variados	81,74%	**84,78%**	83,04%	79,57%	69,72%	73,64%	62,76%
Direito	**72,20%**	60,87%	64,35%	67,83%	46,59%	49,44%	55,35%
Economia	64,80%	**75,65%**	70,00%	74,78%	64,98%	59,53%	63,20%
Esportes	68,30%	80,43%	**83,91%**	83,04%	69,59%	75,95%	74,06%
Gastronomia	**75,70%**	50,87%	48,26%	51,74%	38,23%	33,35%	38,66%
Literatura	**72,20%**	61,30%	59,13%	57,83%	46,91%	43,28%	46,07%
Política	68,70%	80,87%	80,87%	**84,35%**	72,13%	69,09%	72,99%
Saúde	**72,20%**	60,87%	63,48%	66,09%	51,07%	55,32%	54,74%
Tecnologia	73,90%	75,22%	74,35%	**78,26%**	61,20%	61,25%	70,12%
Turismo	78,30%	77,39%	78,26%	**81,30%**	66,94%	64,69%	69,16%
Média	71,70%	70,83%	70,43%	**72,35%**	58,74%	58,56%	60,71%
Desvio padrão (pontos percentuais)	4,97%	11,36%	11,69%	11,19%	12,00%	13,54%	11,85%

Em negrito estão destacados os melhores resultados de cada área. Observa-se

que novamente o método CCC teve resultados inferiores ao método NCD, para qualquer dos compressores de dados considerados. Apesar do método NCD com compressor Zip ter tido a melhor média de atribuições corretas, seus resultados foram melhores que os demais métodos ou compressores apenas em 3 temas. O compressor PPM-D apresentou o melhor resultado em um tema e o compressor Bzip em dois temas.

Observa-se, também, que o método de escolha de resultado por votação apresenta um desempenho inferior à escolha pelo melhor resultado. A figura 2 ilustra o desempenho médio e o desvio padrão dos métodos NCD e CCC com cada um dos compressores.

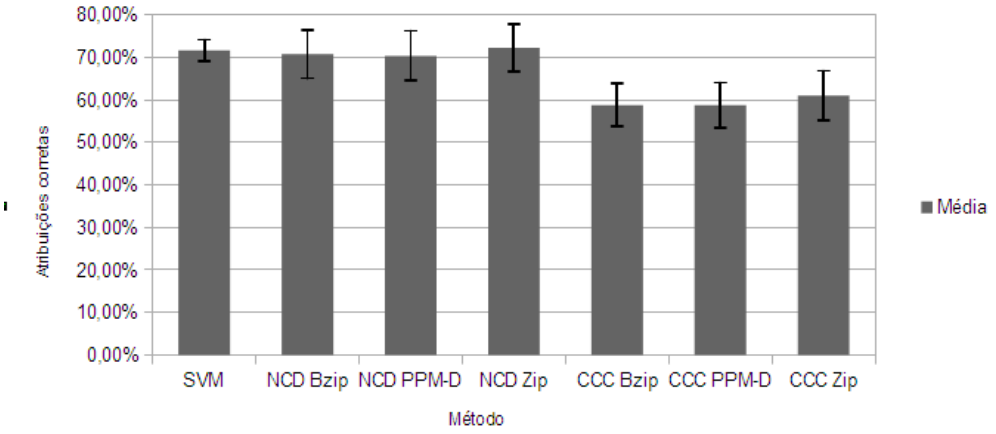

Figura 2. Escolha de autor por votação

5.3. Influência da quantidade de documentos do estilo do autor

Hipotetiza-se que com uma disponibilidade menor de documentos que representem o estilo de um autor o desempenho da atribuição de autoria apresente resultados inferiores. Para verificar a influência da diminuição da disponibilidade destes documentos de treinamento, foram efetuados testes com o método NCD, pois este apresentou os melhores resultados nos testes anteriores. Os três compressores de dados (Zip, Bzip e PPM-D) foram utilizados.

Os testes foram conduzidos nos documentos de apenas duas áreas de conhecimento: Economia e Saúde. Esta escolha é justificada pelo fato que o tema Economia teve um melhor desempenho quando o mecanismo de escolha era o melhor resultado e que o tema Saúde teve um melhor desempenho quando o mecanismo de escolha era a votação.

Os testes foram efetuados através da retirada, aleatória, dos documentos de perfil do autor, reduzindo-se a quantidade inicial de 7 documentos por autor até apenas 2 documentos por autor.

Os resultados para o tema Economia estão representados na figura 3, para o compressor Zip. Para os compressores Bzip e PPM-D os resultados obtidos são bastante semelhantes ao do compressor Zip.

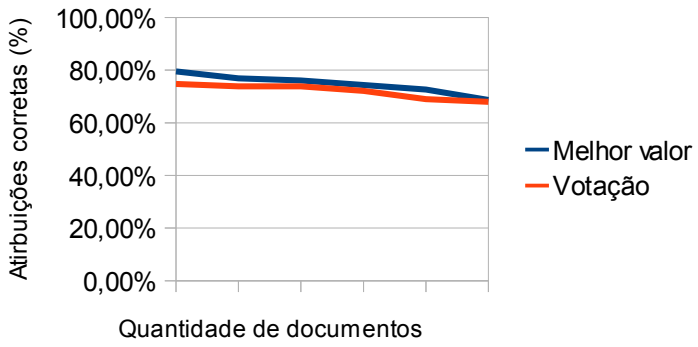

Figura 3. Tema Economia, compressor Zip

Como pode ser observado, há uma queda gradativa na taxa de atribuições corretas conforme a diminuição na quantidade de documentos que representam o estilo do autor, quando o método de atribuição é o do melhor valor de NCD.

A diminuição de atribuições corretas com a escolha sendo feita por votação é também gradativa. Quando existem apenas dois documentos de perfil do autor o resultado é idêntico à escolha pelo melhor resultado porque só restarão dois votos, e caso cada um destes votos seja dado a um autor diferente, o autor com melhor resultado é escolhido, e isto equivale a dizer que a escolha é feita pelo melhor resultado.

Os resultados para o tema Saúde estão representados na figura 4, para o compressor Zip. Para os demais compressores os resultados obtidos foram bastante semelhantes.

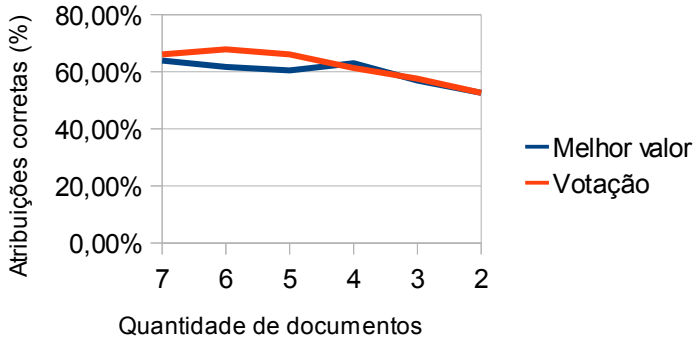

Figura 4. Tema Economia, compressor Zip

Como pode ser observado nos testes deste tema, a escolha pelo critério de votação apresentou um melhor resultado quando o perfil do autor era representado por 5 ou mais documentos. Com a diminuição da quantidade de documentos houve uma inversão de desempenho e o método de escolha pelo melhor resultado passou a apresentar um maior índice de atribuições corretas.

Uma das explicações possíveis para este comportamento é a existência de algum documento, dentro do tema, que possuísse características (por exemplo, uma grande quantidade de palavras que sejam comuns a alguns dos documentos testados) e isto fazia

com que todos os documentos fossem considerados similares a ele. Com a sua remoção, este documento deixou de influir na atribuição de autoria, e a escolha pelo melhor valor da NCD passou a ser determinante.

Verifica-se, assim, que apesar do método de escolha pelo melhor resultado NCD apresentar o melhor desempenho na tarefa de atribuição de autoria, há uma suscetibilidade a documentos que representem mal o perfil de um autor, e que por diversos fatores possíveis todos os documentos apresentem uma melhor compressão (e consequentemente uma melhor NCD) quando concatenados a ele.

Também é possível observar que mesmo com a diminuição da quantidade de documentos que fornecem o perfil do autor, há um desempenho superior ao que seria obtido por uma atribuição aleatória. No tema Economia esta atribuição foi feita corretamente em mais de 60% dos casos, mesmo com apenas 2 documentos de cada autor sendo utilizados, sendo que uma atribuição aleatória teria uma taxa de sucesso de apenas 10%. Em relação ao tema Saúde este índice, para 2 documentos, foi superior a 40%.

6. Conclusões

A tarefa de atribuição de autoria a documentos eletrônicos pode apresentar desafios pela ausência de características físicas que possam auxiliar o perito no desempenho de sua tarefa.

O uso de técnicas computacionais pode ser relevante, automatizando algumas atividades ou fornecendo resultados objetivos, reprodutíveis, sobre a semelhança entre documentos. Os métodos mostrados neste trabalho mostram que bons resultados podem ser obtidos com o uso de compressores de dados, e que a teoria da complexidade de Kolmogorov é uma fundamentação teórica importante a embasar o desempenho visto.

A utilização de uma base de documentos já utilizada em trabalho prévio permite que os resultados obtidos possam ser comparados e assim seja possível a confrontação de características de cada uma das técnicas. No trabalho presente, foi possível verificar que para estes documentos o uso da NCD possui um desempenho superior ao obtido pela CCC e pelo uso de classificadores SVM, quando estes são utilizados com determinadas características estatísticas.

Entre as vantagens do uso da compressão de dados com a NCD ou com a CCC está a desnecessidade de uma seleção prévia de características do estilo do autor e a ausência da necessidade de um treinamento prévio da ferramenta. Desta forma, caso existam outros autores prováveis incluídos posteriormente, bastará efetuar o cálculo da NCD ou CCC em relação aos novos autores e confrontar com medidas obtidas anteriormente.

A possibilidade de escolha do compressor de dados a ser utilizado, havendo em alguns casos pequenas diferenças no resultado obtido, é relevante, pois cada compressor possui métodos diferentes de extração de características dos documentos. Também existem exigências diferenciadas quanto a capacidade computacional, necessidade de memória e complexidade de processamento, e fatores externos pode levar à necessidade que apenas um determinado compressor possa ser utilizado.

A comparação entre o desempenho obtido com diferentes quantidades de documentos representando o perfil dos autores prováveis mostra que há uma influência

no desempenho da atribuição de autoria, mas que mesmo com poucos documentos ainda há uma taxa de atribuições corretas superior ao que seria obtido por uma atribuição aleatória.

Os métodos analisados neste artigo podem ser aperfeiçoados em trabalhos futuros. Sugere-se que mais pesquisas sejam feitas com diferentes compressores, inclusive considerando-se a hipótese de pré-processamento dos documentos. Por exemplo, deve ser verificado se a eliminação de alguns símbolos (como caracteres acentuados ou sinais de pontuação) permite uma melhor compressão dos documentos (gerando melhores medidas NCD) ou se esta diminuição afeta a representação do estilo de um autor.

Agradecimentos

Esta pesquisa foi parcialmente apoiada pelo Conselho Nacional para Desenvolvimento Científico e Tecnológico (CNPq) – concessão nº 301653/2011-9.

Referências

Cilibrasi, R. e Vitányi, P. M. B. (2005) "Clustering by compression", In *IEEE Transactions on Information Theory*, 51:4, pp. 1523–1545.

Kolmogorov, A. N. (1965) "Three approaches to the quantitative definition of information". In *Problems Inform. Transmission*, 1, pp. 1–7

Li, M. e Vitányi, P. M. B. (1997) "An Introduction to Kolmogorov Complexity and Its Applications", Springer, 2nd edition.

M. Burrows , D. J. Wheeler (1994) "A block-sorting lossless data compression algorithm" In *Technical Report 124*, System Research Center - Digital, Palo Alto.

Malyutov, M.B., Wickramasinghe, C.I. and Li, S. (2007): Conditional Complexity of Compression for Authorship Attribution, In *SFB 649 Discussion Paper No. 57*, Humboldt University, Berlin.

Pinker, S. (2007) "The stuff of thought: language as a window into human nature", Viking Adult, 1st edition.

Shkarin, D. (2002) "PPM: One step to practicality." In *Proceedings of the Data Compression Conference*, April 2-4, IEEE Computer Society, Washington, DC., USA., pp: 202-211.

Stamatatos, E. (2009) "A survey of modern authorship attribution methods" In *Journal of the American Society for Information Science and Technology*, Volume 60, Issue 3, pp. 538–556

Varela, P. J. (2010) “O uso de atributos estilométricos na identificação da autoria de textos”. Dissertação de Mestrado apresentada no Programa de Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica do Paraná, Brasil.

