WALTER RIBEIRO DE OLIVEIRA JUNIOR

ATRIBUIÇÃO DE AUTORIA DE DOCUMENTOS EM LÍNGUA PORTUGUESA UTILIZANDO A DISTÂNCIA NORMALIZADA DE COMPRESSÃO

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

CURITIBA
2011

WALTER RIBEIRO DE OLIVEIRA JUNIOR

ATRIBUIÇÃO DE AUTORIA DE DOCUMENTOS EM LÍNGUA PORTUGUESA UTILIZANDO A DISTÂNCIA NORMALIZADA DE COMPRESSÃO

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: Metodologias e Técnicas de Computação

Orientador: Prof. Dr. Edson José Rodrigues Justino
Co-orientador: Prof. Dr. Luiz Eduardo S. Oliveira

CURITIBA
2011

Oliveira Jr., Walter Ribeiro

Atribuição de autoria de documentos em língua portuguesa utilizando a distância normalizada de compressão. Curitiba, 2011. 151 p.

Dissertação – Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática Aplicada.

1. Autoria 2. Forense 3. Compressão de dados 4. Palavra-chave. I. Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e de Tecnologia. Programa de Pós-Graduação em Informática Aplicada

Esta página deve ser reservada à ata de defesa e termo de aprovação que serão fornecidos pela secretaria após a defesa da dissertação e efetuadas as correções solicitadas.

À esposa e família, principalmente ao vô Brigola...

Agradecimentos

Agradeço a todos que apoiaram e auxiliaram para que este trabalho tivesse êxito: esposa, família, amigos, professores e colegas de aulas.

À família por todo o apoio recebido desde o primeiro momento que pensei em entrar no mundo das pesquisas. Sem o auxílio e incentivo de pais e irmã, nada teria sido feito.

Aos amigos por entenderem algumas ausências e por cobrarem, constantemente, que a vida não pode ser feita apenas de ausências.

Aos professores, em especial ao prof. Justino, por todo incentivo, correções e cobranças. Todo o conhecimento que consegui extrair ainda é pequeno diante do que poderia ter aprendido e aproveitado, e que espero ter oportunidade de ainda aproveitar.

Aos colegas de pesquisa, pela troca de experiências, ideias e incentivos ao longo de todo o trabalho.

À Cheila, que tanto auxilia e entende os alunos ao longo de todo o caminho, em especial nos momentos de desespero.

E, principalmente, à minha esposa pelo carinho e compreensão ao longo de todo o trabalho. Agradeço por ter estado ao meu lado quando eu precisava, sendo que muitas vezes eu nem sabia o que eu precisava. Sem sua ajuda, tudo seria mais difícil e não haveria nenhum objetivo.

A todos,

muito obrigado!

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

H(.)	Entropia da distribuição de uma mensagem
K(.)	função de complexidade de Kolmogorov
Max{ . }	função que retorna o maior valor entre os verificados
Min{ . }	função que retorna o menor valor entre os verificados
K(. | .)	função de complexidade condicional de Kolmogorov
C(.)	tamanho do arquivo comprimido

Lista de Abreviaturas

AMDL	*Approximate Minimum Description Length*
ASCII	*American Standard Code for Information Interchange*
BCN	*Best-Compression Neighbor*
BIT	*binary digit*
CFG	*Context-free grammar*
HTML	*Hyper text markup language*
HTTP	*Hyper text transfer protocol*
JPEG	*Joint photographic experts group*
K-NN	*k--nearest-neighbors*
NCD	*Normalized compression distance*
NID	*Normalized information distance*
PPM	*Prediction by partial matching*
PPMC	*Prediction by partial matching – escape C*
RLE	*Run-lenght encoding*
SMDL	*Standard Minimum Description Length*
SVM	*Support vector machine*

Resumo

A atribuição de autoria de documentos em língua portuguesa tem sido objeto de estudos recentes. Este trabalho propõe o uso de compressores de dados, com o uso da medida normalizada de compressão, para a tarefa de atribuição de autoria. São utilizados mecanismos de escolha juntamente com a medida de distância de documentos para a atribuição de autoria do documento questionado a um dos autores considerados candidatos. O uso de compressores de dados faz com que a tarefa de atribuição de autoria independa da escolha prévia de características. As bases de dados utilizadas são as mesmas de outros trabalhos de Daniel Pavelec e Paulo Junior Varela, permitindo que seja feita a comparação entre os resultados obtidos com o uso de compressores de dados e os obtidos por outros métodos, como classificadores SVM. Os resultados obtidos foram promissores, havendo o igualamento ou superação do desempenho dos resultados obtidos anteriormente. Em uma das bases de dados a média de atribuição de autorias corretas foi de 97,17%, em outra base foi de 74,96%. A quantidade de autores e documentos disponíveis em uma das bases permitiu que fosse verificado a influência da quantidade de documentos de treinamento e de autores possíveis no desempenho da atribuição de autoria.

Palavras-chave: Atribuição de autoria, forense, compressão, NCD

Abstract

Authorship attribution of documents written in Portuguese has been object of recent researches. This research proposes the use of data compressors, using the normalized compression distance, to the authorship attribution task. The questioned document has the authorship attributed to one of the candidate authors with the use of the document distance measure and mechanisms of choice. The use of data compressors make the task independent of previous choice of stylistic characteristics. The document data set used are the same of other researches of Daniel Pavelec and Paulo Junior Varela, allowing the comparison of the results with other researches that used data compressors and other methods, like SVM classifiers. The results were promising, being equal or overcoming the previous results. In one database, the average of correct attributions was 97,17%, in the other database it was 74,96%.The amount of authors and documents in the data set allowed the verification of the influence of the number of documents in the training set and the number of possible authors in the authorship attribution performance.

Keywords: authorship attribution, forensic, compressor, NCD

Capítulo 1

Introdução

A atribuição de autoria a documentos questionados é uma atividade que, em regra, requer a opinião de um perito.

O uso de características que pudessem auxiliar no trabalho do perito, por exemplo o uso de medidas estatísticas, remonta a 1887, quando Mendenhall buscou estabelecer a autoria de peças de Shakespeare (Willians, C. B., 1975). Mas, provavelmente, o primeiro trabalho mais significativo sobre atribuição de autoria foram as pesquisas de Mosteller e Wallace, em 1964 (Mosteller, F. E Wallace, D. L., 1964), sobre uma série de ensaios políticos conhecidos como *The Federalist Papers* ("Os papéis federalistas"), composto por 134 documentos de autoria conhecida e 12 documentos de autoria questionada entre dois prováveis autores.

Em sua pesquisa, Mosteller e Wallace utilizaram uma análise estatística Bayesiana da frequência de ocorrência de pequenas palavras como "e", "para", "então". Os resultados obtidos foram significativos, permitindo estabelecer uma autoria provável a cada um dos documentos. Este trabalho pioneiro serviu de orientação a inúmeros trabalhos posteriores, nos quais foram (e são) pesquisadas características que permitam a definição de características de escrita (estilometria) que sejam discriminantes o suficiente para serem utilizadas na atribuição de autoria de documentos questionados. E, assim, foram pesquisadas diversas características: a escolha de atributos como o tamanho de frases e palavras, ou a frequência de ocorrência de palavras ou letras, em um total de mais de 1000 características identificadas até a pesquisa de Rudman ter sido publicada (Rudman, J., 1998). Conforme observa Stamatatos, estas pesquisas utilizavam ferramentas computacionais muito mais para auxiliar o trabalho de cálculo de

estatísticas do que para criar um sistema automatizado de extração de características. (Stamatatos, E., 2009).

A identificação de autoria é uma atividade que poder ser dividida em duas abordagens principais: atribuição de autoria e verificação de autoria.

A tarefa de atribuição de autoria ocorre quando se está diante de uma situação onde há um documento cuja autoria é desconhecida e existem diversos autores prováveis para o documento. Neste caso, busca-se uma classificação do documento questionado entre as diversas categorias possíveis, cada categoria correspondendo a cada um dos autores prováveis, sendo que a categorização será feita com a escolha de uma única categoria. Ou seja, a utilização de um classificador multiclasse com uma única categorização sendo feita.

A verificação de autoria é feita quando dado um documento de autoria questionada é verificado se este documento foi elaborado por um autor determinado ou não. Trata-se, neste caso, de uma classificação binária, onde o resultado é uma resposta positiva ou negativa de autoria.

Um dos problemas dos primeiros trabalhos de identificação de autoria foi a falta de avaliação objetiva dos métodos e dos resultados propostos. Alguns estudos eram feitos em documentos de autoria questionada, o que impedia a medição do desempenho de cada método, pois não havia certeza da autoria para que pudesse ser feita confirmação da correção ou não do resultado obtido. Trabalhos posteriores, nos quais a presente dissertação se enquadra, buscaram corrigir esta ausência de objetividade, através dos seguintes pontos:

- a utilização de um maior número de autores candidatos;
- a separação ou controle dos tópicos dos documentos utilizados;
- homogeneização do tamanho dos documentos da base de dados;
- utilização de bases de dados que já tenham sido utilizadas por outros trabalhos, permitindo um confronto de desempenhos e análise de resultados comparativa.

Trabalhos vem sendo publicados sobre o uso de ferramentas computacionais para a atribuição de autoria em documentos em língua portuguesa (Pavelec, D. F., 2007; Varela, P. J. 2010; Justino, E. J. R. , 2002; Coutinho, B. C. *et al*., 2005]. Estas pesquisas tem utilizado atributos estilométricos com classificadores e a compressão de dados com o compressor

PPMc. Uma medida possível de distância entre documentos, denominada de distância normalizada de compressão, é estudada para verificar se a sua utilização na atribuição de autoria apresenta resultados satisfatórios.

1.1. Desafio

O desafio deste trabalho está na verificação da utilidade da NCD na atribuição de autoria de documentos de língua portuguesa. Além desta verificação, o seu desempenho será comparado com as diversas propostas de compressão de dados já utilizadas em trabalhos que utilizaram bases de dados em língua inglesa, e será utilizada uma base de dados única para a comparação entre as diversas abordagens propostas, permitindo assim a comparação de seu desempenho. A utilização de uma base de dados em língua portuguesa também permite que os seus resultados sejam mais significativos para eventuais utilizações que venham a ser feitas em processos judiciais no Brasil.

1.2. Motivação

A atribuição de autoria a documentos digitais, em abordagens que considerem apenas as características estilométricas do autor, é o grande fator motivacional deste trabalho. Nestas abordagens, características específicas do documento (por exemplo, o software que o produziu ou a determinação do computador onde o documento foi redigido) são dispensadas e apenas as características de estilo de escrita do autor são relevantes. A motivação desta pesquisa é o uso de abordagens de compressão de dados para a identificação de autoria de documentos, comparando-se os resultados obtidos com outros trabalhos publicados que utilizaram abordagens de estatísticas de atributos estilométricos (Pavelec, D. F., 2007, Varela, P. J. 2010, Justino, E. J. R., 2002) . Esta motivação pode ser detalhada nos seguintes itens:

- implementação de outras abordagens baseadas em compressão de dados, permitindo a comparação de resultados entre estas diversas abordagens;
- utilização uma base única de documentos em língua portuguesa, permitindo assim que os resultados obtidos possam ser comparados;

- utilização a mesma abordagem para a separação dos documentos em grupos de treinamento e de testes, para que o procedimento seja único em todas as abordagens que serão utilizadas, de forma que o resultado obtido possa ser comparado;
- fomentação da pesquisa da utilização de ferramentas computacionais para que provas possam ser produzidas em processos judiciais brasileiros, fornecendo embasamento científico para os resultados obtidos

1.3. Objetivos

O objetivo geral do presente trabalho é verificar o desempenho da NCD na atribuição de autoria de documentos de língua portuguesa. Para tanto, o presente trabalho apresentará:

- uma descrição da teoria da complexidade da informação e sua importância para as abordagens de atribuição de autoria com o uso de compressores de dados;
- uma descrição das abordagens de compressão de dados propostas em outros trabalhos;
- uma descrição do método NCD de medição de similaridade entre documentos;
- uma implementação de atribuição de autoria com o uso da NCD;
- uma comparação do desempenho destas abordagens por meio do uso de uma base de dados única;
- contribuições para o trabalho de perícias realizados por peritos e linguistas em procedimentos judiciais brasileiros

Esta pesquisa não tem como objetivo a proposição de novas abordagens que utilizem a compressão de dados para a atribuição de autoria. Serão exploradas as abordagens já propostas em outros trabalhos de maneira sistematizada e organizada. Será feita a comparação de resultados obtidos em outros trabalhos.

1.4. Contribuições

A principal contribuição deste trabalho é fomentar as pesquisas da perícia de documentos digitais escritos em língua portuguesa, contribuindo para que os peritos possam dispor de procedimentos que possuem fundamentação científica para a elaboração de laudos

em processos judiciais brasileiros. As contribuições indiretas seguintes também podem ser mencionadas:

- o aprofundamento no conhecimento das abordagens baseadas em compressão de dados para a atribuição de autoria;
- o aumento das pesquisas realizadas em documentos digitais de língua portuguesa;
- o auxílio ao trabalho de peritos com abordagens que fornecem resultados objetivos às análises efetuadas

1.5. Organização

O presente trabalho está organizado em 6 capítulos. O primeiro capítulo refere-se à introdução. O segundo capítulo trata da fundamentação teórica necessária ao desenvolvimento do presente trabalho. O terceiro capítulo apresenta o estado da arte do uso de compressores para a tarefa de atribuição de autoria a documentos. O quarto capítulo apresenta o método proposto para a realização de testes. O quinto capítulo apresenta os testes executados, os resultados obtidos e a discussão sobre os resultados. O sexto capítulo apresenta a conclusão obtida após a pesquisa realizada.

Capítulo 2

Fundamentação Teórica

Este capítulo contém a fundamentação teórica necessária para o desenvolvimento da pesquisa. São abordados a identificação de autoria e sua relevância para perícias forenses, as características estilométricas e o uso de compressores de dados para a tarefa de atribuição de autoria.

2.1. Identificação de Autoria

A tarefa de identificação de autoria desperta interesse há diversos anos. A identificação de autoria divide-se em duas abordagens principais: atribuição de autoria e verificação de autoria.

A atribuição de autoria busca, através de diversas técnicas, verificar quem é o autor provável de um documento questionado, quando existem diversos autores prováveis para o documento. Nesta abordagem os testes são executados contra diversos autores possíveis e, ao final, um autor é escolhido como o autor provável do documento.

A verificação de autoria, por sua vez, efetua testes do documento questionado contra um único autor, obtendo ao final uma resposta positiva ou negativa se o documento é de autoria do autor testado.

É possível transformar o problema de atribuição de autoria em uma escolha feita por diversos classificadores binários, desde que sejam estabelecidas regras para a escolha da classe vencedora entre as diversas escolhas binárias (Sebastiani, F., 2002). Ao mesmo tempo, isto não significa que um classificador multiclasse (ou protocolos de atribuição de autoria,

essencialmente fornecedores de resultados multiclasse) poderá ser utilizado para a tarefa de verificação de autoria.

Somando-se o fato que a verificação de autoria requer dois exemplos de treinamento (exemplo positivo e exemplo negativo), sabe-se que é impossível construir uma base de treinamento que represente adequadamente todos os exemplos negativos possíveis de autoria. Como no exemplo mencionado por (Koppel, M. e Schler, J., 2004), a verificação de autoria de um documento, quanto ao seu autor ter sido Shakespeare, significa que é necessário construir uma base de treinamento de exemplos positivos, com as diversas obras de Shakespeare, abrangendo todos os estilos que foram utilizados pelo autor durante toda a sua vida, em todas as suas formas de expressão. E a construção da base de treinamento de exemplos negativos implica em utilizar todos (ou pelo menos uma quantidade representativa) de documentos que não tenham sido produzidos por Shakespeare. E o resultado da verificação da autoria dependerá da qualidade da base de treinamento formada, pois se houver um autor que tenha um estilo bastante semelhante ao de Shakespeare, seja o autor do documento questionado, mas não tenha sido incluído entre os exemplos negativos de autoria, o mecanismo de verificação de autoria poderá considerar que o estilo de Shakespeare e do documento questionado são semelhantes o suficiente, resultando em uma verificação positiva de autoria, pelo simples fato de não ter havido treinamento negativo suficiente em um estilo parecido.

No restante deste trabalho, considerando que os métodos a serem analisados envolvem uma classificação multiclasse com atribuição de uma única classe como resultado final, será feita apenas a atribuição de autoria.

As pesquisas conduzidas a respeito da atribuição de autoria assumem, em muitos casos, que é possível estabelecer um perfil único para cada autor, já que cada pessoa aprendeu e desenvolveu de maneira diferente suas habilidades de comunicação e de escrita. Ao mesmo tempo, imagina-se que o perfil estilométrico de cada pessoa não pode ser resumido a características simples, tais como o tamanho médio das palavras utilizadas, pois o processo de aprendizagem e até mesmo o vocabulário disponível em cada língua não permitem uma grande variação no tamanho das palavras mais comuns a cada idioma, por exemplo (Juola, P., 2008).

Conforme mencionado na introdução, um dos trabalhos de pesquisa mais mencionados a respeito da atribuição de autoria é a pesquisa dos papéis federativos, ensaios políticos que

foram publicados nos anos de 1787 e 1788 sob a forma de pseudônimo. Da série de ensaios publicados, doze tiveram a sua autoria questionada entre os três autores que publicaram os demais ensaios. Diversas pesquisas foram efetuadas para a determinação da autoria, cabendo a Mosteller e Wallace a pesquisa pioneira de utilizar dados estatísticos sobre a utilização de palavras funcionais (por exemplo, conjunções, artigos e preposições) por cada um dos prováveis autores para determinar quem seria o autor provável de cada um dos doze ensaios questionados(Juola, P., 2008; Stamatatos, E., 2009).

2.2. Características estilométricas

As características estilométricas utilizadas na atribuição de autoria, conforme classificação sugerida por Stamatatos, estão elencadas a seguir. A figura 2.1 apresenta, esquematicamente, as características estilométricas.

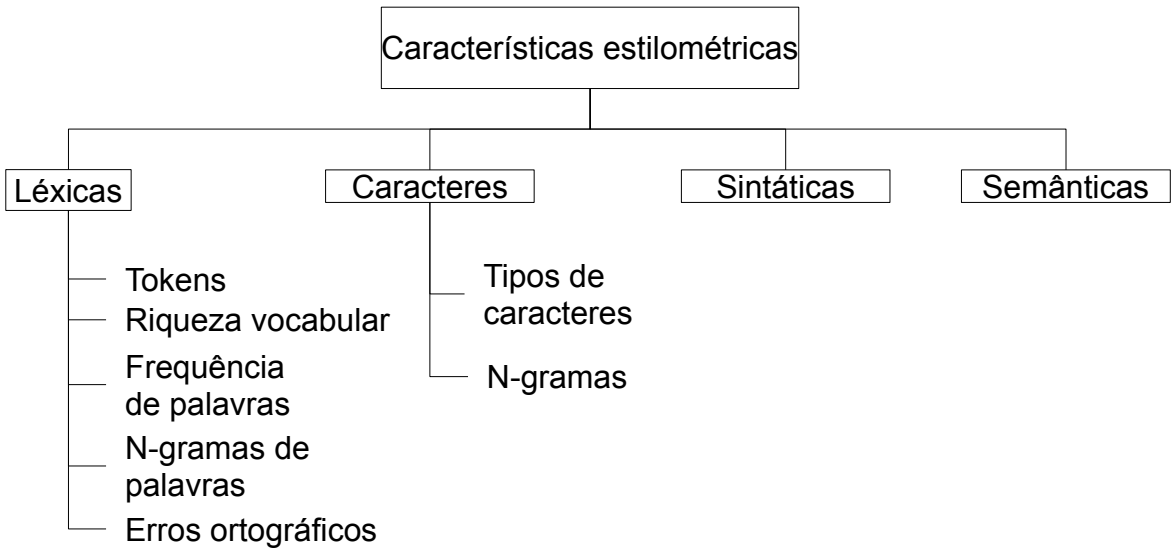


Figura 2.1: Características estilométricas

A seguir explicamos como se apresentam cada uma destas características estilométricas.

2.2.1. Características Léxicas

As características léxicas são as que consideram as características do conjunto de vocábulos de um idioma. As abordagens léxicas podem ser:

- baseadas em *tokens*
- baseadas em riqueza vocabular
- baseadas em frequência de palavras
- baseadas em n-gramas de palavras
- baseadas em erros ortográficos

Na abordagem baseada em *tokens*, um documento é considerado como uma coleção de *tokens* agrupados em sentenças, sendo que os *tokens* correspondem a letras, palavras, símbolos, sinais de pontuação ou algarismos. Nesta abordagem, as características estilométricas de um autor são definidas a partir da contagem da frequência absoluta ou da proporção de aparição de um ou mais *tokens*. Por exemplo, a contagem da quantidade média de palavras em cada sentença ou a quantidade média de palavras por parágrafo.

Na abordagem baseada em riqueza vocabular é verificada a diversidade de vocábulos empregados pelo autor em seu documento. Por exemplo, pode ser medida a proporção entre a quantidade de vocábulos diferentes utilizados no documento e a quantidade total de vocábulos (tamanho do documento) ou então a quantidade de vocábulos que são utilizados apenas uma vez no documento.

Na abordagem baseada em frequência de palavras são gerados vetores indicando a frequência com que cada palavra é utilizada no documento, sendo que podem ser consideradas todas as palavras empregadas ou então um subgrupo com palavras selecionadas previamente. Por exemplo, o estilo de um autor pode ser determinado pela frequência que ele utiliza determinadas palavras pertencentes a uma determinada categoria sintática, tais como artigos e preposições. Estas categorias sintáticas são denominadas de “palavras funcionais” pois não carregam, por si só, informações semânticas, sendo utilizadas para a estruturação lógica de um documento.

Conforme (Stamatatos, E., 2009), a maior parte dos estudos de atribuição de autoria utiliza esta abordagem, principalmente porque as palavras funcionais são utilizadas de forma

inconsciente pelos autores e são independentes de assunto . A escolha das palavras funcionais que serão utilizadas dependem do conhecimento do idioma e são escolhidas, em geral, de forma arbitrária. Em alguns trabalhos são utilizados um número limitado de palavras funcionais, sendo encontrados trabalhos que utilizaram as 100 palavras mais comuns (Burrows, J. F., 1987), ou as 250 palavras mais comuns (Koppel, M.; Schler, J. e Bonchek-Dokow, E., 2007) ou até mesmo todas as palavras que tenham sido utilizadas pelo menos duas vezes no documento (Madigan, D. *et al*, 2005).

Como estas abordagens mencionadas desconsideram que um autor pode utilizar determinadas palavras de maneira agrupada com uma frequência diferente de outro autor, é utilizada a abordagem que considera a frequência de *n*-gramas de palavras. Desta forma, dado um determinado valor *n*, as palavras são consideradas como agrupamentos. Assim, ao invés da verificação da frequência da utilização da palavra "então" em um documento, é verificada a frequência da utilização do *n*-grama "e então".

Por fim, outra abordagem léxica possível é a verificação da quantidade de erros cometidos pelo autor para a verificação de seu estilo. É feita a escolha de quais estilos de erros que serão verificados e as medidas de frequência são calculadas conforme o autor comete estes erros. Por exemplo, pode se medir a frequência de palavras que aparecem com letras invertidas ou com grafia incorreta.

Como menciona (Juola, P., 2008), um cuidado que deve ser tomado em relação a características de vocabulário utilizado é a possibilidade de se medir características de assuntos abordados ao invés de características de autores. Em um bom exemplo, o autor menciona que diversas pessoas, se solicitadas a reescrevem a estória infantil "Chapeuzinho Vermelho" certamente utilizarão vários vocábulos em comum, como "cestinha", "floresta", "lobo", "boca grande", e estas características identificam muito mais a estória do que cada um dos diferentes autores.

2.2.2. Características de caracteres

Para esta abordagem a única informação relevante a ser extraída de um documento são os caracteres que o compõe. As abordagens que consideram as características de caracteres são:

- Baseadas em tipos de caracteres (letras, símbolos)
- Baseadas em *n*-gramas de caracteres

Na abordagem baseada em tipos de caracteres são feitas medições considerando os caracteres utilizados pelo autor. Por exemplo, a frequência de utilização de um determinado sinal de pontuação ou a frequência da ocorrência de um caractere específico em relação à frequência de utilização daquele caractere no idioma considerado, ou mesmo a utilização de letras em maiúsculas ao longo do documento.

Na abordagem baseada em n-gramas de caracteres, os caracteres são considerados em agrupamentos de *n* caracteres. Por exemplo, a palavra "exemplo", se considerarmos *n=3*, resultaria nos trigramas "exe", "xem", emp", "mpl" e "plo". Esta abordagem tende a reduzir a influência de erros do autor ao longo do documento, diminuindo assim a presença de ruídos. Por exemplo, as palavras "elefante" e "elefanti" produziriam uma grande quantidade de trigramas idênticos, e apenas um trigrama seria afetado pelo erro do autor.

(Stamatatos, E., 2009) observa que nas abordagens de n-gramas de palavras e de caracteres a característica estilística mais importante é o conjunto de n-gramas mais frequentes, permitindo a captura do estilo do autor pelas combinações de palavras ou de letras mais utilizadas por ele.

Uma abordagem derivada da baseada em n-gramas de caracteres é feita com a utilização de compressores de dados. Esta abordagem utiliza compressores de dados para capturar a frequência de repetição de caracteres ou de palavras. Nesta abordagem, diversas medidas podem ser extraídas a partir da compressão dos documentos questionados e dos documentos de autoria conhecida. Por ser a abordagem que será utilizada neste trabalho, seu funcionamento será detalhado adiante.

2.2.3. Características sintáticas

A abordagem baseada em características sintáticas utilizam os padrões sintáticos que um autor utiliza, em geral de maneira inconsciente, para a produção de seus documentos. É necessário o uso de uma ferramenta de extração de regras sintáticas para a construção do padrão utilizado pelo autor, já que não é suficiente extrair as palavras utilizadas, devendo ser atribuída a classe sintática a que a palavra pertence (ou é utilizada). Por exemplo, que um

autor costuma utilizar uma frase na forma “sujeito verbo objeto direto” com uma frequência diferente de outros autores.

Esta abordagem é bastante dependente do idioma e requer que a ferramenta de extração de regras sintáticas que não produza ruídos em níveis inaceitáveis quando submetidas a construções sintáticas que fujam de regras estabelecidas pela norma culta, sob pena de não poder ser utilizada em documentos informais.

2.2.4. Características semânticas

Esta abordagem considera as características semânticas das palavras e frases utilizadas. Por exemplo, pode ser verificado as construções que o autor utiliza que servem para qualificar a sentença imediatamente anterior, ou como o autor faz uso de sentenças em oposição ao longo do documento.

As ferramentas para a análise de características semânticas são dependentes do idioma considerado.

2.3. Compressão de dados para a extração de características estilísticas

O uso de características estilísticas baseadas em caracteres faz uso de medidas extraídas a partir dos caracteres encontrados no documento, podendo considerá-los de maneira isolada (cada caractere é independente dos demais caracteres) ou em função de seu contexto (os caracteres que estão ao redor são significativos).

Os compressores de dados, de maneira geral, fazem uso da informação dos caracteres presentes em um documento para extrair elementos que possam permitir que o documento seja armazenado com um tamanho menor mas preservando todas informações do documento original ou as informações mais relevantes.

Desta forma, torna-se possível o uso de compressores de dados como uma abordagem baseada em caracteres para a extração de informações estilísticas de um determinado autor.

Esta abordagem é melhor detalhada nos tópicos a seguir.

2.3.1. Teoria da informação

Pesquisas sobre a atribuição de autoria utilizam as pesquisas feitas sobre a teoria da informação para verificar a quantidade de informação que é transmitida em uma determinada

comunicação, ou seja, para medir quanta informação é transmitida sobre um fenômeno a partir da observação deste fenômeno. Existem duas grandes teorias para medir esta quantidade de informação: a teoria da informação de Shannon e a teoria da complexidade de Kolmogorov (Juola, P., 2008; Grünwald, P. D. e Vitányi, P. M.B., 2003; Schmidhuber, J. 1995; Hammer, D. *et al.,* 2000).

2.3.2. Teoria da informação de Shannon

A teoria da informação de Shannon é formalizada a partir das conclusões de sua pesquisa, publicadas em 1948 (Shannon, C. E., 1948). Shannon formulou uma maneira de calcular quanta informação um observador recebe de um fenômeno *F* após este fenômeno ter ocorrido, através da entropia de uma comunicação, definida pela equação

$$H(P) = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{1}$$

sendo que *P* é uma variável aleatória, representando a mensagem a ser transmitida, com uma distribuição finita *i=1,...,n.* A função *H(P)* é a entropia da distribuição desta mensagem *P* e p_i é a probabilidade da mensagem *P* ter o valor *i*, ou seja, a probabilidade *i* da mensagem ser enviada.

Através da equação de Shannon é possível estimar qual será o limite inferior da compressão de dados sem perda. Sabendo-se a frequência de ocorrência dos símbolos em um arquivo, é possível calcular qual será a quantidade mínima de *bits* necessários para representar cada um destes símbolos. Em um experimento, Shannon utilizou a capacidade humana de prever a ocorrência de determinadas letras do alfabeto, no idioma inglês, e estimou que são necessários, em média, entre 0,6 e 1,3 *bits* por caractere para a representação de textos.

Em trabalhos separados, (Shannon, C. E., 1948, Fano, R. M., 1949) estabeleceram o que é conhecido como código Shannon-Fano: através de uma codificação, é possível codificar símbolos que apareçam em uma frequência maior com uma quantidade menor de *bits*, enquanto que símbolos com uma probabilidade menor de aparição são mapeados em sequências maiores de *bits*. Desta forma, uma informação pode ser codificada com uma quantidade menor de *bits* que a informação original, sendo possível a recuperação da informação original sem que ocorram perdas.

Uma crítica feita à teoria de Shannon é que ela considera que todas as mensagens possíveis possuem a mesma probabilidade de ocorrência, ou seja, considera que todas as mensagens possíveis em um determinado conjunto de mensagens poderão ser transmitidas. Conforme criticam (Grünwald, P. D. e Vitányi, P. M.B., 2003) isto desconsidera que algumas mensagens possuem regularidades que permitem que a sua compressão ocorra em taxas muito maiores, requerendo uma quantidade muito menor de *bits* para sua representação. E, como mencionam os autores, uma maneira de se representar a quantidade de informação existente em uma mensagem, sem depender de probabilidades de ocorrência de seus símbolos, é através da teoria de complexidade de Kolmogorov.

2.3.3. Teoria de complexidade de Kolmogorov

A teoria de complexidade de Kolmogorov (Schmidhuber, J. 1995) é a única teoria existente, para alguns autores, que fornece um critério objetivo para o conceito de simplicidade de uma informação .

Desenvolvida de maneira independente por Solomonoff e Kolmogorov, a teoria de complexidade de Kolmogorov estabelece que a complexidade de uma informação é dada pelo tamanho mínimo de descrição que é possível fazer por meio do uso de uma linguagem descritora universal. Desta forma, dada uma mensagem *x* que possua diversas descrições possíveis, sendo possível reconstruir *x* a partir de qualquer uma destas descrições, a complexidade de Kolmogorov é determinada pelo tamanho da menor descrição possível. Ou, dada uma máquina Turing completa, a complexidade de Kolmogorov é o programa de menor medida capaz de produzir uma informação especificada. (Cilibrasi, R., 2006; Grünwald, P. D. e Vitányi, P. M.B., 2003).

É representado por meio da equação $K_{\Phi}(x)$ sendo que *K* representa a função de complexidade de Kolmogorov, ϕ é a função que representa uma máquina de Turing completa e *x* é a mensagem a ser representada.

Pela complexidade de Kolmogorov é possível comparar a complexidade da descrição de dois objetos considerando cada objeto por si só, ou seja, a complexidade de um objeto *x* depende apenas da descrição de sua complexidade e não da probabilidade de ocorrência de um outro objeto *y*.

Para mensagens em geral, cujo conteúdo seja uma sequência aleatória de dados, a complexidade de Kolmogorov será aproximadamente equivalente ao tamanho da própria mensagem, pois não haverá alguma regularidade que permita que a mensagem seja representada por uma descrição com menor tamanho. Entretanto, existirão mensagens que possuirão regularidades que permitirão que a informação seja representada por um conjunto menor de instruções, o que resultará em uma complexidade de Kolmogorov de tamanho menor que o tamanho da mensagem.

A complexidade de Kolmogorov não é computável. Não é possível haver um programa que, rodando em uma máquina de Turing, receba uma sequência de dados como entrada e emita como saída a complexidade de Kolmogorov desta sequência. (Lee, T. J., 2006 , Grünwald, P. D. e Vitányi, P. M.B., 2003).

2.3.4. Compressão de dados

Compressão de dados é o processo de converter uma sequência de dados de entrada em uma outra sequência de dados, de saída, que possua um tamanho menor. A compressão é feita pela eliminação de dados redundantes ou pela redução da quantidade de dados que são necessários para representar corretamente uma informação. (Salomon, D., 2004)

Sua fundamentação é a teoria da informação: conforme demonstrado por Shannon, há um limite de quanto uma informação pode ser comprimida sem que haja perda. Este limite é dado pela entropia da informação. Em alguns casos poderá ser aceitável que existam perdas na informação, possibilitando que a informação seja comprimida ainda mais através do descarte de fragmentos da informação que sejam considerados dispensáveis. Estas compressões são conhecidas como compressão com perda porque não é possível, a partir da informação comprimida, se recuperar integralmente a informação original.

2.3.5. Processos simples de compressão de dados

Um dos processos mais simples de compressão de dados é feita através da *Run-lenght encoding (RLE)*: nesta codificação, as informações que se repetem dentro de um documento são substituídas por uma indicação desta repetição. Por exemplo, supondo que a informação a ser comprimida é a sequência de caractere “aaaabbbbaaaaabbbbb”. Em uma codificação RLE, é feita a indicação do símbolo que se repete e quantas vezes o símbolo é repetido. Por

exemplo, considerando que só existam letras na sequência de caracteres, a codificação poderia ser “a4b4a5b5”, indicando que há 4 repetições da letra “a” seguida por 4 repetições da letra “b” e assim por diante.

Outra maneira de obter a compressão de dados é através da codificação relativa. Nesta codificação, ao invés de serem armazenados todos os valores absolutos dos dados, é feito o cálculo de quanto uma informação difere da imediatamente anterior. Por exemplo, dada a sequências de números *A = {100, 95, 100, 98, 98, 100}*, é possível representar esta informação através da sequência *B = { 100, -5, 5, -2, 0, 2}*, sendo que o primeiro elemento da sequência numérica é mantido e os demais valores são substituídos pela diferença em relação ao elemento anterior, ao invés de seu valor absoluto.

Apesar de serem codificações simples, há grande aplicabilidade destes métodos de compressão. Por exemplo, na representação digital de uma desenho qualquer, é esperado que existam grandes áreas onde a mesma informação é repetida. Ao invés de representar esta área por seus valores absolutos, é possível que a codificação RLE represente a mesma informação com o uso de uma quantidade menor de dados.

Estas codificações permitem a compressão de dados sem perda. Se a perda de informação for aceitável, é possível obter-se uma compressão ainda maior dos dados. Por exemplo, considerando o mesmo exemplo da sequência numérica *A = {100, 95, 100, 98, 98, 100}*. Se for possível admitir que a informação seja representada apenas pelos valores 100, 95 e 105, é possível atribuir símbolos que indiquem que houve a perda ou ganho de 5 unidades de valor. Assim, se considerarmos que *-1* indica a perda de 5 unidades de valor e *1* indica o acréscimo destas mesmas 5 unidades, a sequência poderia ser representada por *C = { 100, -1, 1, 0, 0 ,0}*, sendo que a reconstrução da informação resultaria em *A' = {100, 95, 100, 100, 100, 100}* e a informação referente ao número *98* seria perdida.

2.3.6. Processos estatísticos de compressão de dados

Uma das maneiras de se obter uma compressão de dados é através do uso de uma codificação diferente para a representação da informação. E a teoria de Shannon auxilia nesta codificação. Por exemplo, pode-se imaginar que a informação mais frequente em uma sequência de dados é representada com o uso de uma quantidade menor de símbolos, sendo que as informações mais infrequentes utilização uma quantidade maior de símbolos para sua

representação. Assim, na tabela de caracteres ASCII, todos os símbolos utilizam 7 *bits* para sua representação, independente da frequência de sua ocorrência. Através da teoria de Shannon, em um determinado documento, os caracteres mais frequentes poderiam ser representados utilizando-se menos *bits*, enquanto os caracteres mais infrequentes poderiam utilizar mais de 7 *bits* para serem representados.

A codificação de Shannon-Fano decorre da teoria de Shannon, buscando definir uma codificação otimizada com uma quantidade de *bits* variáveis para representar cada símbolo conforme a probabilidade de sua ocorrência. Outra codificação que utiliza uma quantidade variável de bits é a codificação de Huffmann, sendo que a diferença mais marcante entre elas é a ordem de construção da codificação, o que faz com que a codificação Huffmann seja mais eficiente.

As codificações de Shannon-Fano e Huffmann consideram que a frequência de ocorrência de cada símbolo é conhecida previamente, ou seja, que primeiro o documento é inteiro analisado para se extrair a frequência de ocorrência de cada símbolo. Isto, evidentemente, leva a uma demora na compressão: antes de iniciar a codificação, é necessário analisar o documento inteiro para a obtenção da frequência de ocorrência de cada símbolo.

Uma solução para isto é a utilização de um método de codificação de Huffmann adaptativa. Desta maneira a informação é obtida e, para cada *byte* de informação processado, a árvore de codificação é modificada.

A codificação de Huffmann é satisfatória quando os símbolos apresentam uma probabilidade de ocorrência que sejam potências negativas de 2, ou seja, probabilidades de ocorrência iguais a ½ , ¼ e assim por diante. Para probabilidades distintas, há uma perda de desempenho. Uma alternativa é a utilização de codificações aritméticas, onde não é feita a codificação separa para cada símbolo e sim a codificação de toda a mensagem (ou de fragmentos de tamanho preestabelecido) é feita em um único número n, sendo que $0 \leq n \leq 1$.

Por exemplo, o formato de arquivo de imagens JPEG utiliza uma modalidade de codificação numérica denominada de *QM-coder* e a codificação Huffmann, conforme o resultado obtido com cada codificação seja mais eficiente.

Um dos métodos de compressão estatístico mais eficiente para a codificação de documentos de texto é a *prediction by partial matching* (PPM). Este método foi proposto originalmente por Cleary e Witten em 1984. Este método é baseado em um codificador que

mantém um modelo estatístico do documento e, para cada símbolo *S* de entrada, atribuí uma probabilidade *P* e envia *S* para um processo de codificação aritmética. (Salomon, D., 2004)

Este método de compressão é altamente eficiente para documentos de texto. Em alguns de seus modelos, o modelo estatístico gerado considera o contexto onde a informação apareceu, estimando com uma probabilidade melhor qual será o próximos símbolo a ser visto. Por exemplo, na língua portuguesa, há uma determinada probabilidade do símbolo *ã* ser encontrado, mas é bastante provável que após o símbolo *ç* seja encontrado um símbolo *ã*. Desta forma, ao considerar o contexto, a probabilidade de cada símbolo é atribuída com maior precisão. Desta forma o compressor PPM utiliza conhecimentos prévios para estimar a probabilidade.

O método PPM utiliza a modelagem de ordem *N* para estimar probabilidade do próximo símbolo. Isto é, para cada símbolo *S* que é lido, o modelador de contexto considera *N* símbolos que precedam *S* para estimar a probabilidade do símbolo *S* ser encontrado. Em geral, compressores PPM utilizam contextos de ordem 2 a 10. O PPM poderá ser adaptativo e alterar a ordem que efetua esta busca, aprimorando assim as probabilidades estimadas. Por exemplo, ao verificar que não houve a aparição de um determinado símbolo em um contexto *N*, poderá reiniciar a busca em um contexto *N-1* (Salomon, D., 2004)

Existem diversas variações do método PPM devido à necessidade de saber como lidar com símbolos que não tenham sido vistos anteriormente. Ao iniciar a compressão de um documento,por exemplo, é certo que o primeiro símbolo não terá sido visto anteriormente e é bastante provável que os símbolos seguintes também não tenham sido vistos anteriormente. E o decodificador PPM, ao receber a informação que deverá decodificar, precisa de métodos que sinalizem que houve a alteração do contexto. Isto é feito através de um símbolo de escape. Cada variante do método PPM trata de forma diferente como a sinalização de escape é gerada.

2.3.7. Compressão de dados baseada em dicionário

Este método de compressão gera um dicionário com símbolos que sejam encontrados no documento e, para cada vez que um símbolo é encontrado, ele é substituído por um *token* que indica qual entrada do dicionário que deverá ser utilizada para recuperar a informação.

Em um modelo de dicionário estático, um dicionário é predeterminado (a partir de outra fonte de dados, ou pela leitura integral do documento) e é utilizado sem alterações em um documento inteiro. Por exemplo, uma lista de palavras é utilizada como dicionário e cada vez que uma palavra é encontrada no documento, é substituída pelo índice que indica a sua posição no dicionário. Pode ser bastante útil, por exemplo, para documentos que possuam sempre uma repetição de palavras, como um código HTML.

Para compressores que sejam de propósito geral poderá ser difícil preestabelecer um dicionário que seja adequado. Neste caso, existem diversas abordagens onde o dicionário é adaptativo ao documento que está sendo comprimido.

Um dos métodos de compressão por dicionário adaptativo mais utilizados é o LZ77, proposto em 1977 por Lempel e Ziv. Neste método, é determinado uma janela de busca (denominada de janela deslizante) que indica onde será feita a busca de informações que já tenham ocorrido anteriormente no documento.

Por exemplo, supondo que a informação a ser comprimida seja a frase:

"Existem três tipos de mentiras: mentiras, mentiras sujas e estatísticas" (fonte: atribuída a Benjamin Disraeli)

De maneira aproximada, em uma codificação do tipo LZ77, supondo que a janela deslizante tenha um tamanho de 10 caracteres, o processo de compressão seria o seguinte. Para cada um dos caracteres de entrada, o compressor verificaria se este caractere ocorreu anteriormente, dentro da janela deslizante. Caso tenha ocorrido, ele verifica se o próximo caractere da posição atual também já foi observado anteriormente, na mesma sequência, e assim por diante, até o máximo de correspondências encontradas. Caso isto tenha ocorrido e o tamanho da repetição seja superior ao tamanho que será ocupado para fazer a referência da localização desta repetição, é feita a substituição da ocorrência atual por uma indicação de onde houve a ocorrência anterior. O algoritmo sempre buscará a ocorrência mais longa, ou , caso todas as ocorrências anteriores tenham o mesmo tamanho, qual foi a ocorrência mais distante. Assim, por exemplo, o algoritmo poderá observar que a sequência "s " do final da palavra "tipos " já ocorreu ao final da palavra "três ", mas a indicação da repetição ocupará mais espaço que a repetição em si. Mas, quando encontrar a palavra "mentiras", observará que

ela já ocorreu anteriormente no fragmento “mentiras:” e substituirá esta ocorrência por uma indicação da ocorrência anterior.

Supondo que esta indicação seja feita pelo símbolo *@k,n* sendo que @ é o símbolo que indica a repetição de informações, *k* é a quantidade de caracteres existentes entre a posição atual e a informação original que foi repetida e *n* é a quantidade de caracteres repetidos, a frase do exemplo poderia ser comprimida resultando em:

“Existem três tipos de mentiras: @10,8, @17,8 sujas e estatísticas”

Desta forma, @10,8 representa que 10 caracteres antes da aparição do símbolo encontra-se a informação que deve ser repetida, e que esta informação é composta por uma sequência de 8 caracteres. Ao realizar esta substituição, obtém-se o texto original:

“Existem três tipos de mentiras: mentiras, @17,8 sujas e estatísticas”

Desta forma, a informação original poderia ser armazenada de maneira a ocupar um menor espaço com a possibilidade de ser reconstruída sem perdas quando fosse necessário.

Existem diversas variações do método de compressão LZ77. Cada variação busca aprimorar a taxa de compressão ao lidar de forma diferente com a janela deslizante, com a codificação utilizada para indicar que a informação é uma repetição de uma informação existente em outro segmento do documento, ou a maneira como estas informações são armazenadas, ou mesmo como é gerado um dicionário.

Uma das variações bastante conhecidas e utilizadas é a *Deflate*. Este método de compressão utiliza uma variação do método LZ77 e a codificação de Huffmann para comprimir documentos. O método de compactação *deflate* é utilizado no formato de arquivo ZIP e na compressão de dados transmitidos através do protocolo HTTP. (Salomon, D., 2004)

2.3.8. Compressão de dados baseada em blocos

Existem diversos outros métodos de compressão que não são puramente (ou principalmente) estatísticos ou baseados em dicionário. Entre estes métodos, um que se destaca é o método de compressão baseado em blocos, denominado de método Burrows-Wheeler. Neste método, os dados são processados em blocos pelo compressor e, através de

um processo de rotação e ordenação deste bloco, os símbolos existentes no bloco são codificados de maneira mais eficiente por métodos de codificação como o RLE ou Huffmann.

O formato de arquivo *bzip* utiliza o método de Burrows-Wheeler para comprimir documentos.

2.4. Classificação de documentos com uso da compressão de dados

Exemplos da utilização da classificação de documentos com o uso de compressão podem ser citados:

- a classificação de documentos anônimos, cuja autor provável era de Antonio Gramsci, para a composição de uma obra com a coletânea de seus trabalhos, na Itália (Basile, C., 2010);
- o uso medida de similaridade entre documentos para a construção de árvores genealógicas de espécimes biológicos a partir da classificação da sequência de seu genoma (Merivuori, T. e Roos, T., 2009);
- a classificação de gênero musical de músicas a partir da linha melódica do baixo (Şimşekli, U., 2010);
- o uso da medida de similaridade entre fragmentos de códigos para verificação de vulnerabilidades em código-fonte sensível (Mahmood, W. e Akhtar, M. F., 2009).

Conforme mencionado anteriormente, a tarefa de atribuição de autoria pode ser entendida como uma tarefa de classificação de documentos. Nesta tarefa, os compressores são utilizados para a verificação de similaridade entre documentos, com a atribuição sendo feita em função da similaridade encontrada. Estes métodos são detalhados a seguir.

2.4.1. Atribuição de autoria baseadas em compressores de dados

A atribuição de autoria através do uso de compressores de dados apresenta algumas vantagens em relação a outros métodos. A classificação de documentos com o uso de compressores é de fácil aplicação e não requer, como regra, que os documentos sejam pré-tratados, mas também apresentam desvantagens como um maior tempo de processamento. (Marton, Y., Wu, N., e Hellerstein, L. 2005).

Como exemplo desta facilidade de aplicação, os autores mencionados citam o fato que alguns métodos podem utilizar compressores disponíveis comercialmente, não requerendo o desenvolvimento ou implementação de um método específico de compressão de dados.

A discussão sobre o tempo de processamento é bastante relativa. Alguns métodos que utilizam outras abordagens estatísticas, por exemplo, podem requerer um grande esforço para o estabelecimento de quais características estatísticas são relevantes para um determinado idioma, autor ou categoria de documento, apresentando um tempo de execução reduzido apenas após este conhecimento estar disponível.

Cilibrasi (2006) propôs, em sua tese, o uso de uma aproximação da complexidade de Kolmogorov para a medida da similaridade existente entre documentos. O presente trabalho utilizará esta medida, denominada NCD, para a atribuição de autoria em documentos de língua portuguesa.

2.4.2. Distância Normalizada de Compressão

O método da distância normalizada de compressão foi proposta por M. Li *et al,* 2004. Seu fundamento é a complexidade de Kolmogorov.

Conforme mencionado anteriormente, a complexidade de Kolmogorov afirma que a complexidade de uma informação pode ser medida pelo tamanho do atributo que descreva a informação de maneira completa. Esta complexidade, entretanto, é incalculável.

Entretanto, como verificado ao tratarmos de compressores, um compressor recebe um documento (mensagem) e busca, através de mecanismos, representar este documento com um número menor de símbolos. Desta forma, um compressor pode ser utilizado como uma aproximação da implementação da complexidade de Kolmogorov.

Uma definição do conceito de métrica do grau de similaridade entre dois documentos, proposta por (Cilibrasi, R., 2006), deve associar dois documentos *x* e *y* de forma que, caso os documentos sejam similares, a distância entre eles seja próxima a 0 e, caso sejam dissimilares, a distância seja próxima a 1.

Para isto, inicia com a proposta de M. Li *et al,* 2004 do método de distância normalizada de informação (NID), definido pela equação

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

sendo *x* e *y* dois documentos considerados, *max{ }* a função que retorna o valor máximo entre dois valores, *K(a)* a complexidade de Kolmogorov para um documento *a* e *K(a | b)* a complexidade condicional de Kolmogorov de um documento *a* considerando-se um documento *b*.

Considerando que a complexidade *K()* é incomputável mas pode ser aproximada pelo uso de compressores, Cilibrasi propõe a distância normalizada de compressão, representada na equação simplificada

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

sendo que *x* e *y* são os documentos para os quais se deseja medir a distância normalizada de compressão, *C(xy)* é o tamanho do arquivo *x* e *y* concatenados comprimido, *C(x)* é o tamanho do arquivo *x* comprimido, *max* é a função que retorna o maior entre dois valores e *min* é a função que retorna o valor mínimo entre dois valores.

Idealmente esta equação deveria resultar em valores compreendidos entre 0 e 1 (inclusive), mas (Cilibrasi, R., 2006) relata que na prática é possível encontrar valores acima de 1 por imperfeições existentes em compressores, já que em vários compressores disponíveis comercialmente é possível que informações sejam acrescidas ao começo do resultado da compactação como informações de cabeçalho.

A NCD entre duas cópias de um mesmo documento deveria resultar em 0, já que há um grau máximo de semelhança entre dois documentos idênticos. Em compressores de dados, esta propriedade é denominada idempotência: idealmente, a compressão de um arquivo formado pela concatenação de dois conjuntos de dados idênticos deveria resultar em um arquivo comprimido que tivesse o mesmo tamanho de um arquivo formado por um único conjunto de dados. Ou seja, o tamanho da compressão de um arquivo *A* de conteúdo *A'* deveria ser igual ao tamanho da compressão de um arquivo *A* de conteúdo *A'A'*. Entretanto, pela aproximação que é feita pelos compressores, a NCD entre documentos idênticos é sempre superior a 0 em um pequeno valor. Isto explica-se porque qualquer compressor necessita utilizar algum espaço para indicar que a informação que se apresenta na concatenação é a mesma que já foi vista na primeira parte do arquivo.

2.4.3. Complexidade Condicional de Compressão

O método da complexidade condicional de compressão é apresentada por (Malyutov, M.B. Wickramasinghe,C. I. e Li, S., 2007), embasando-se, também, na complexidade de Kolmogorov.

A CCC de um documento *y* em função de um documento *x* é expressa na equação

$$CCC(y|x)=C(xy)-C(x) \tag{3}$$

sendo que *x* e *y* são os documentos considerados, *xy* é a concatenação do documento *y* com o documento *x*, e *C(x)* é o tamanho da compressão de um documento *x*.

A forma a complexidade de Kolmogorov seria melhor representada medindo-se de que forma um compressor aproveita as informações de um documento de treinamento para utilizar no documento que está sendo testado.

Também é proposta a complexidade condicional de compressão relativa (CCCr), expressa na equação

$$CCCr(y|x)=\frac{C(xy)-C(x)}{C(y)} \tag{4}$$

sendo que *x* e *y* são os documentos considerados, *xy* é a concatenação do documento *y* com o documento *x*, e *C(x)* é o tamanho da compressão de um documento *x*.

Malyutov, Wickramasinghe e Li também propõe outras medidas em seu trabalho, fazendo a observação que os resultados obtidos com o uso delas não foi satisfatório. (Malyutov, M.B. Wickramasinghe,C. I. e Li, S., 2007)

A distância relativa de complexidade (RDC), representada pela equação

$$RDC(y|x) = C(y)-CCC(y|x) = C(y)-C(xy)-C(x) \tag{5}$$

sendo que *x* e *y* são os documentos considerados, *xy* é a concatenação do documento *y* com o documento *x*, e *C(x)* é o tamanho da compressão de um documento *x*.

A razão da distância relativa de complexidade (RRDC), representada pela equação

$$RRDC(y|x) = \frac{RDC(y|x)}{C(y)} = \frac{C(y)-C(xy)-C(x)}{C(y)} \tag{6}$$

sendo que *x* e *y* são os documentos considerados, *xy* é a concatenação do documento *y* com o documento *x*, e *C(x)* é o tamanho da compressão de um documento *x*.

2.4.4. Método de Coutinho *et al*

Em 2005, Coutinho *et al.* publicaram trabalho onde compressores PPM-C foram utilizados para a atribuição de autoria.

Conforme mencionado anteriormente, compressores PPM são compressores onde é gerada uma probabilidade condicional para o símbolo que está sendo tratado tendo-se por referência um contexto de *n* símbolos anteriores. No trabalho o compressor PPM utilizou ordens de Markov 4, 5 e 6.

A atribuição de autoria foi feita através da geração de um modelo estatístico a partir dos documentos de treinamento e, em seguida, este modelo foi aplicado ao documento de teste. Em seguida, foi atribuída a autoria ao modelo estatístico que permitiu a maior taxa de compressão do documento de teste. Isto pode ser representado pela equação a seguir

$$TC(x) = \frac{C(x)}{x*} \tag{7}$$

sendo que *x* é o documento de teste considerado, *C(x)* é o tamanho da compressão de um documento *x* com um determinado modelo estatístico gerado a partir dos documentos de treinamento de cada autor provável e *x** é o tamanho do documento *x.*

2.4.5. Trabalho de Benedetto

Em um trabalho de Benedetto, para cada documento de treinamento *x* é calculado o seu tamanho comprimido, *C(x)*. Em seguida, cada documento de treinamento *x* é concatenado com o documento de teste *y*, gerando um documento *xy*, sendo então calculado o seu tamanho comprimido *C(xy)*. Calcula-se, então, a diferença dos tamanhos comprimidos de *C(x)* e *C(xy)*, conforme a equação a seguir (Benedetto, D., Caglioti, E. e Loreto, V., 2002).

$$B(x|y) = C(xy) - C(x) \tag{8}$$

Verifica-se que a equação de Benedetto é a mesma da CCC de (Malyutov, M.B. Wickramasinghe,C. I. e Li, S., 2007). Doravante denominaremos de CCC toda vez que for feita referência a esta equação.

2.5. Problemas e cuidados em relação a compressores de dados

Como estes procedimentos de classificação de documentos utilizam compressores, algumas observações importantes merecem ser feitas. Alguns algoritmos de compressão

possuem peculiaridades que podem desaconselhar o seu uso como classificadores ou, no mínimo, requerer que alguns cuidados especiais sejam tomados em relação às bases de dados utilizadas.

2.5.1. Problemas com desbalanceamento do tamanho da base de treinamento

As bases de treinamento devem possuir tamanhos semelhantes (com aproximadamente a mesma quantidade de documentos) e os documentos que a compõem também devem possuir tamanhos próximos. É recomendável equalizar o tamanho das bases de treinamento, descartando dados das bases maiores, após ser feita a concatenação dos arquivos, ou mesmo concatenar apenas fragmentos de mesmo tamanho, resultando assim em um arquivo concatenado de tamanho igual para todas as categorias (Marton, Y., Wu, N., e Hellerstein, L. 2005).

2.5.2. Problemas com o tamanho dos documentos utilizados

Alguns algoritmos de compressão possuem características em seu funcionamento em relação ao tamanho dos documentos das bases de dados que podem alterar o seu desempenho como classificadores (Cébrian, M., Alfonseca, M. e Ortega, A., 2005).

Uma das assunções feitas em relação ao NCD é a existência da identidade entre dois documentos idênticos, ou seja, a distância normalizada de compressão entre dois documentos idênticos deveria ser 0.

Após fundamentar matematicamente que esta identidade decorre da idempotência de um compressor, (Cébrian, M., Alfonseca, M. e Ortega, A., 2005) analisaram o desempenho de alguns compressores em relação a esta afirmação. Como resultado, obtiveram que alguns compressores (*gzip*, representando compressores derivados de Lempel-Ziv, e *bzip2*, representando compressores de ordenação de blocos) possuem problemas em relação à identidade quando o tamanho dos documentos ultrapassa determinados limites.

Especificamente, o compressor *gzip* utiliza uma janela deslizante de 32k bytes para verificar os dados que já ocorreram e que podem ser representados através de alguma codificação. Quando o tamanho dos documentos (considerados individualmente, ou em etapas de concatenação) ultrapassa o tamanho da janela deslizante, o compressor deixa de utilizar a informação dos documentos anteriores (que estão ao início da concatenação) para comprimir

o documento que vem ao final. Desta forma o desempenho do NCD é afetado, pois as informações "vistas" ao início de um documento não são utilizadas em documentos seguintes.

O compressor *bzip2* apresenta um problema semelhante. A ordenação de blocos é feita com blocos de 900k bytes, no máximo, o que significa que blocos de tamanhos maiores serão fracionados antes de seu processamento. Desta forma, um baixo desempenho do compressor quando utilizado para classificação de documentos acontecerá quando os documentos (considerados individualmente ou concatenados) ultrapassarem o tamanho de 900k bytes.

Esta característica de limitação de tamanho de documentos a serem processados, provavelmente, não é restrita ao NCD. Os demais procedimentos que utilizem compressores que possuam limitações no tamanho de janelas deslizantes, tamanho de dicionário ou de ordenação de blocos também apresentarão um desempenho inferior quando as suas respectivas limitações de tamanho forem ultrapassadas.

2.6. Considerações finais

Neste capítulo foi apresentada a fundamentação teórica necessária para a elaboração e compreensão deste trabalho. Foram apresentados os conhecimentos mínimos sobre os compressores de dados, explicitando-se como a complexidade de Kolmogorov pode ser aproximada com o seu uso. A distância normalizada de compressão é proposta como suficiente para permitir a atribuição de autoria e seu desempenho será comparado ao de outros métodos de cálculo de semelhança ou de distância entre documentos. No próximo capítulo são descritos brevemente os trabalhos já publicados na área de identificação de autoria de documentos.

Capítulo 3

Estado da Arte

A atribuição de autoria de documentos tem sido pesquisada em uma abordagem científica apenas recentemente. Os primeiros experimentos realizados buscavam identificar a autoria de documentos de autoria desconhecida, o que não permitia a confirmação que a identificação do autor teve sucesso.

O uso de métodos científicos de verificação de resultados e utilização de bases de testes controladas permitiu que os resultados obtidos pudessem ser reproduzidos e confirmados. Isto só é possível com o uso de um ambiente de teste controlado, onde se sabe qual o resultado correto esperado, sendo assim possível verificar o desempenho obtido e se as identificações de autoria foram efetuadas corretamente.

3.1. Histórico

A atribuição de autoria possui uma utilização remota na história, sendo impossível determinar quando houve o primeiro questionamento sobre a autoria de algum documento e quais critérios foram utilizados para a tarefa de atribuição.

De maneira documentada, tem-se que o primeiro trabalho publicado sobre o questionamento de autoria de um documento foi feito em 1787 por Edmond Malone, questionando a autoria de peças de Shakespeare (King, E. G. C., 2010).

A partir do século XX houveram trabalhos importantes sobre a atribuição de autoria, entre os quais podem ser destacados:

- no final do século XX, Mendenhall estudou como a representação gráfica de frequência de palavras e seu tamanho poderiam identificar o perfil do autor de um documento (Williams, C. B., 1975);
- Zipf, no início dos anos 30, estudou se a frequência da aparição de vocábulos diferentes era suficiente para estabelecer o perfil estilístico de um autor (Zipf, G. K., 1975);
- ao final dos anos 30, Yule publicou trabalho onde era estudada o comprimento das frases como característica do perfil de cada autor. Após concluir que o tamanho da frase apresentava pouca variabilidade média, Yule aprofundou os estudos de Zipf em relação à frequência de aparição de palavras (Yule. G. U., 1938);
- a análise de Mosteller e Wallace, no início dos anos 60, utilizando o teorema Bayesiano, mostrou que novas abordagens estatísticas eram possíveis. O sucesso relativo obtido pelos autores fez com que diversos métodos propostos posteriormente fizessem a análise dos mesmos documentos (os papéis federalistas), permitindo desta forma uma comparação de resultados (Mosteller, F. E Wallace, D. L., 1964);
- em 1985, Holmes publicou uma análise sobre fatores discriminantes possíveis de interesse para a atribuição de autoria, identificando diversas medidas estatísticas de palavras (Holmes. D. I., 1985);
- Shakespeare teve novamente seus trabalhos analisados em 1987, quando Thisted e Efron atribuíram a ele um poema de autoria até então questionada (Thisted, R. e Efron, B., 1987);
- em 1996, Merriam novamente analisou os trabalhos de Shakespeare, comparando o seu estilo ao de Christopher Marlowe (Merriam, T., 1996);
- neste mesmo ano, Foster também estudou os trabalhos de Shakespeare, atribuindo a ele o poema “Uma Elegia Fúnebre” (Foster, D., 1996);

Mais recentemente, diversos trabalhos foram publicados onde o objeto de análise era a atribuição de autoria (e não apenas uma verificação de autoria). Podem ser citados os seguintes trabalhos:

- Stamatatos, Fakotakis, e Kokkinakis, em 2001, propuseram um método automatizado de extração de características estilísticas através do uso de ferramentas de processamento de linguagem natural, extraindo características como frequência de sinais de pontuação, tamanho de palavras, frequência de frases verbais e nominais (Stamatatos, E., Fakotakis, N. e Kokkinakis, G.);
- Kešelj *et al.*, em 2003, propuseram um novo método de atribuição de autoria baseado na frequência de n-gramas de caracteres, criando perfis baseados em uma pequena quantidade de n-gramas; (Keselj, V. *et al*., 2003)
- Zheng *et al.*, em 2006, propuseram um método para a atribuição de autoria de documentos mensagens online por meio do uso de características léxicas, sintáticas, estruturais e específicas ao domínio do objeto de estudo (Zheng, R. *et al, 2006)*;
- Cilibrasi e Vitanyi (R. Cilibrasi, R. e Vitányi, P. M. B., 2005) agruparam corretamente os documentos de 4 autores russos.

No Brasil, destacam-se as seguintes pesquisas sobre a atribuição de autoria:

- Coutinho *et al.* utilizou o compressor de dados PPM-C para a atribuição de autoria em documentos de língua portuguesa, método este que foi abordado no tópico 2.4.4. Foram obtidos resultados de aproximadamente 82% de atribuições corretas de autoria (Coutinho, B. C. *et al*., 2005);
- em 2007, Pavelec (Pavelec, D. F., 2007) estudou a identificação da autoria de documentos com uso de classificadores do tipo SVM, sendo que uma das abordagens utilizadas permitia a atribuição de autoria a documentos, e que será utilizada comparativamente no presente trabalho, no tópico 4.1.1
- em 2010, Varela estudou o uso de atributos estilométricos na identificação de autoria de textos, também com classificadores do tipo SVM, sendo que uma das abordagens utilizadas também é semelhante à atribuição de autoria a documentos e também será utilizada comparativamente no presente trabalho, conforme o tópico 4.1.2

3.2. Abordagens para extração de conhecimento da base de treinamento

Marton, Wu e Hellerstein (Marton, Y., Wu, N., e Hellerstein, L. 2005) vislumbram três abordagens para a classificação de documentos baseados em compressão. Estas três abordagens são denominadas de SMDL (*Standard Minimum Description Length* - tamanho mínimo de descrição padrão), AMDL (*Approximate Minimum Description Length* – tamanho mínimo de descrição aproximado) e BCN (*Best-Compression Neighbor* – vizinhança de melhor compressão).

3.2.1. Procedimento SMDL

O procedimento SMDL consiste na geração de um modelo de compressão a partir de documentos de treinamento e a utilização deste modelo para a compressão dos documentos de teste. Ou seja, dado um conjunto de categorias $C_1,...,C_n$, todos os documentos $D_1,...,D_n$ de cada uma das categorias são concatenados em um único documento A_i, onde i é a categoria que será representada. A concatenação é feita através do acréscimo dos dados de um documento a outro documento de texto, sem que seja feita qualquer inserção de símbolo neste processo. Assim, procede-se à cópia do conteúdo do primeiro arquivo e, ao final deste, inicia-se a cópia do conteúdo do segundo arquivo, sem interrupção nem inserção de qualquer símbolo entre estes conteúdos.

Em seguida, é executado o algoritmo de compressão neste documento A_i, obtendo-se um modelo ou um dicionário de compressão denominado M_i. Desta forma, a partir de cada categoria de treinamento, é obtido um modelo de compressão.

A seguir o documento de teste T é submetido ao algoritmo de compressão, uma vez para cada categoria, utilizando-se o modelo de compressão M_i obtido anteriormente para cada categoria. O modelo de compressão é utilizado de maneira estática, ou seja, não é feita uma atualização do modelo de compressão conforme o documento de teste é processado.

Desta forma, para o documento testado, é obtida uma série de valores de taxas de compressão para cada uma das categorias. A categoria escolhida para o documento testado é aquela em que o documento T obteve a maior taxa de compressão.

Este processo é ilustrado na figura 3.1 abaixo.

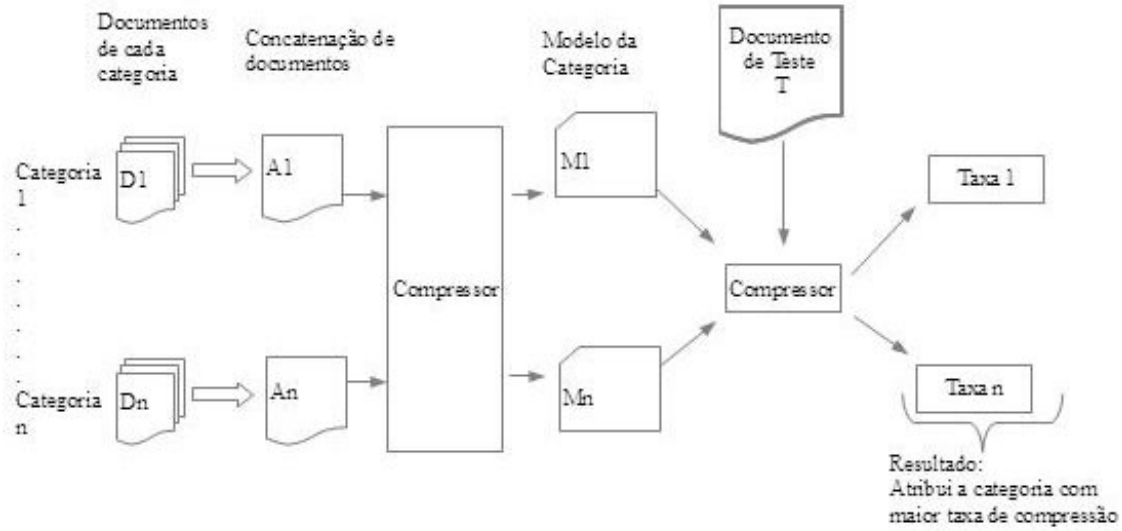


Figura 3.1: Procedimento SMDL

É necessário que o compressor gere um modelo de compressão para que, posteriormente, o algoritmo de compressão seja alimentado com o modelo de compressão e com o documento de teste. Isto impede, em geral, que compressores disponíveis comercialmente / publicamente possam ser utilizados.

3.2.2. Procedimento AMDL

O procedimento AMDL é bastante semelhante ao procedimento anterior, sendo sua principal diferença o fato que não é gerado um modelo estático de compressão para cada categoria de treinamento.

Dado um conjunto de categorias $C_1,...,C_n$, o conjunto de todos documentos $D_1,...,D_n$ de cada categoria são concatenados em um único documento A_i. Cada documento A_i, representativo de cada categoria, é submetido ao compressor, gerando um documento A_i* que possui um tamanho $C(A_i)$. A seguir, o documento de teste T é concatenado com cada um dos documentos A_i, gerando um documento A_iT. Este documento resultante é então submetido ao compressor, gerando um novo documento A_iT* que possui tamanho $C(A_iT)$. O documento testado T é atribuído à categoria que tiver minimizado a diferença entre os tamanhos dos arquivos comprimidos V_i, representado pela equação

$$V_i = C(A_i T) - C(A_i) \tag{11}$$

Este processo é ilustrado na figura 3.2abaixo.

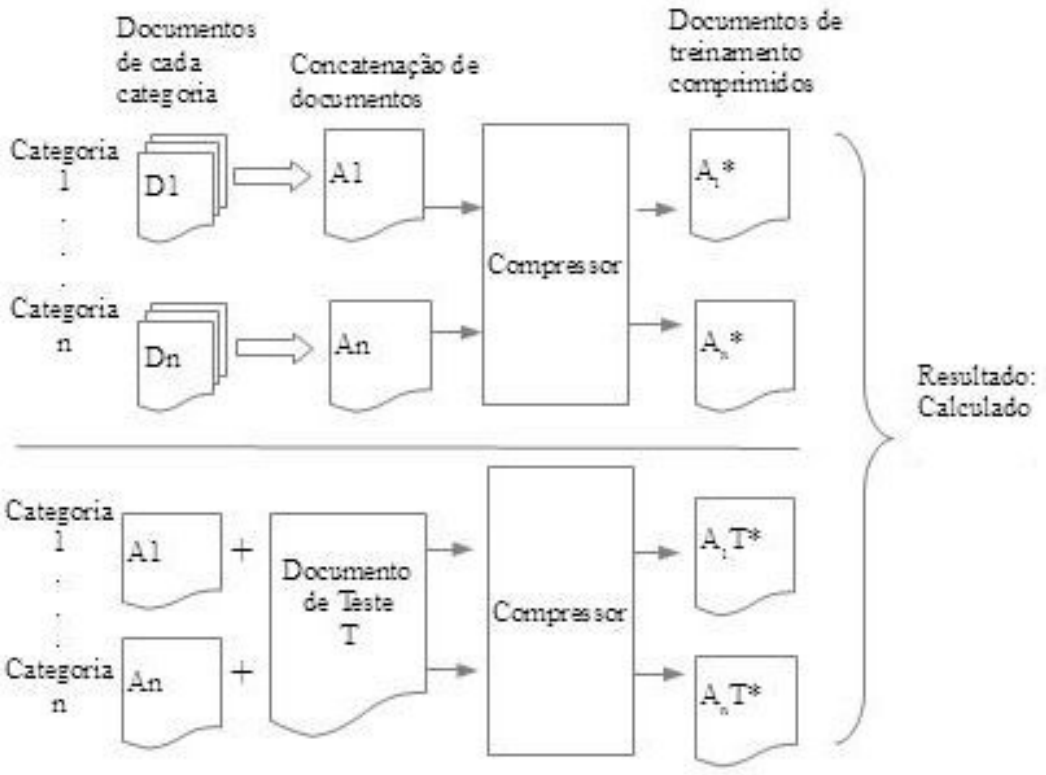


Figura 3.2: Procedimento AMDL

A principal diferença deste procedimento em relação ao SMDL é que é possível a utilização de compressores disponíveis comercialmente (*off-the-shelf*). No procedimento SMDL, os documentos utilizados para o treinamento geram um modelo estatístico de compressão, e este modelo não é mais atualizado. No procedimento AMDL, o modelo estatístico é gerado pelos documentos de treinamento, mas este modelo sofre alterações quando os documentos de teste são processados.

3.2.3. Procedimento BCN

Este procedimento foi desenvolvido por (Benedetto, D., Caglioti, E. e Loreto, V., 2002). Bastante semelhante ao procedimento AMDL, sua principal diferença é o fato que os documentos de treinamento, de cada categoria, não são concatenados.

Para cada documento de treinamento *D* é calculado o seu tamanho comprimido, $|D^*|$. Em seguida, cada documento de treinamento *D* é concatenado com o documento de teste *T*, gerando um documento *DT,* sendo então calculado o seu tamanho comprimido $|DT^*|$. Calcula-se, então, a diferença dos tamanhos comprimidos de *D* e *DT*, conforme a equação a seguir.

$$V_{DT} = |DT^*| - |D^*| \tag{9}$$

O documento de teste T é atribuído à classe do documento de treinamento *D* que resultar na menor valor de diferença, conforme calculado acima. Este processo é ilustrado na figura 3.3.

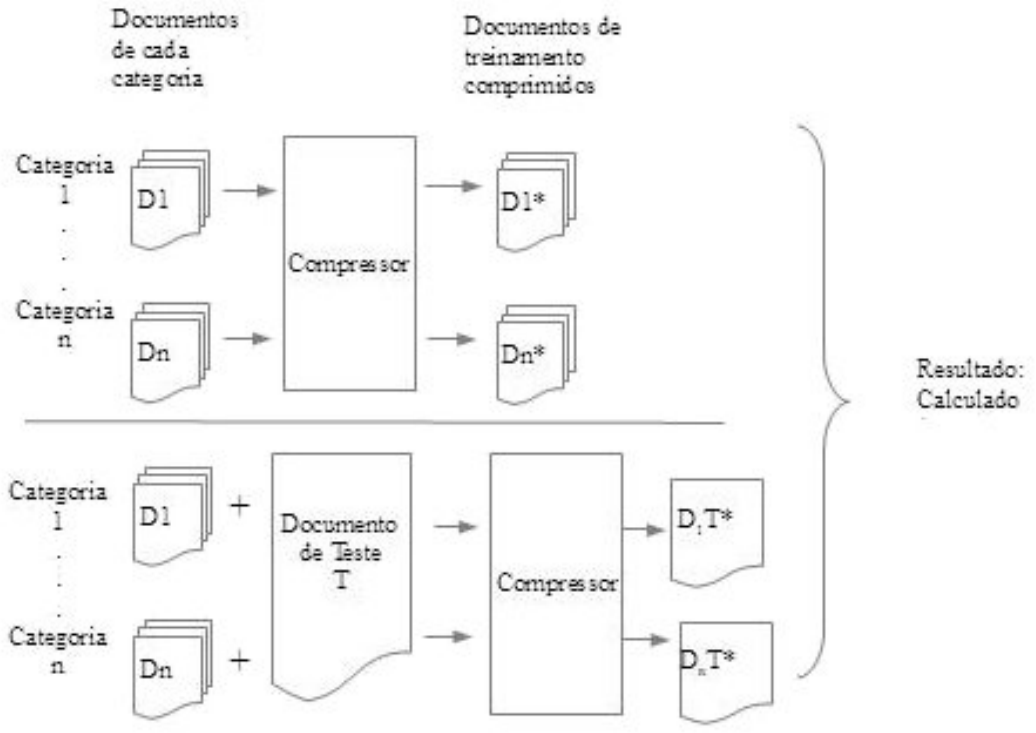


Figura 3.3: Procedimento BCN

O procedimento BCN é um método de "1-vizinho mais próximo" e é altamente suscetível a ruídos. Abordagens de "*k*-vizinhos mais próximos" poderiam trazer resultados melhores. (Marton, Y., Wu, N., e Hellerstein, L. 2005)

Como os documentos de treinamento não são concatenados, o tempo de execução deste procedimento pode ser significativamente maior do que o procedimento AMDL e aumenta conforme o número de documentos de treinamento utilizados. Se, por exemplo, houverem 5 categorias com 5 documentos cada para treinamento, no procedimento AMDL cada documento testado terá exigido um total de (*5 + 5 =*) 10 execuções do compressor, enquanto o procedimento BCN necessitará de (*5*5 + 5*5 =*) 50 execuções.

Assim como no procedimento AMDL, também é possível a utilização de compressores disponíveis comercialmente.

3.3. Análise de métodos

Conforme verificado anteriormente, por vezes diversos autores utilizam a mesma equação para o cálculo da similaridade ou da distância entre documentos, sendo entretanto variável o procedimento utilizado até se chegar à etapa de aplicação da equação de cálculo.

A figura 3.4 a seguir sintetiza os procedimentos e as etapas utilizadas nos trabalhos revisados sobre atribuição de autoria de documentos eletrônicos com uso de compressores de dados.

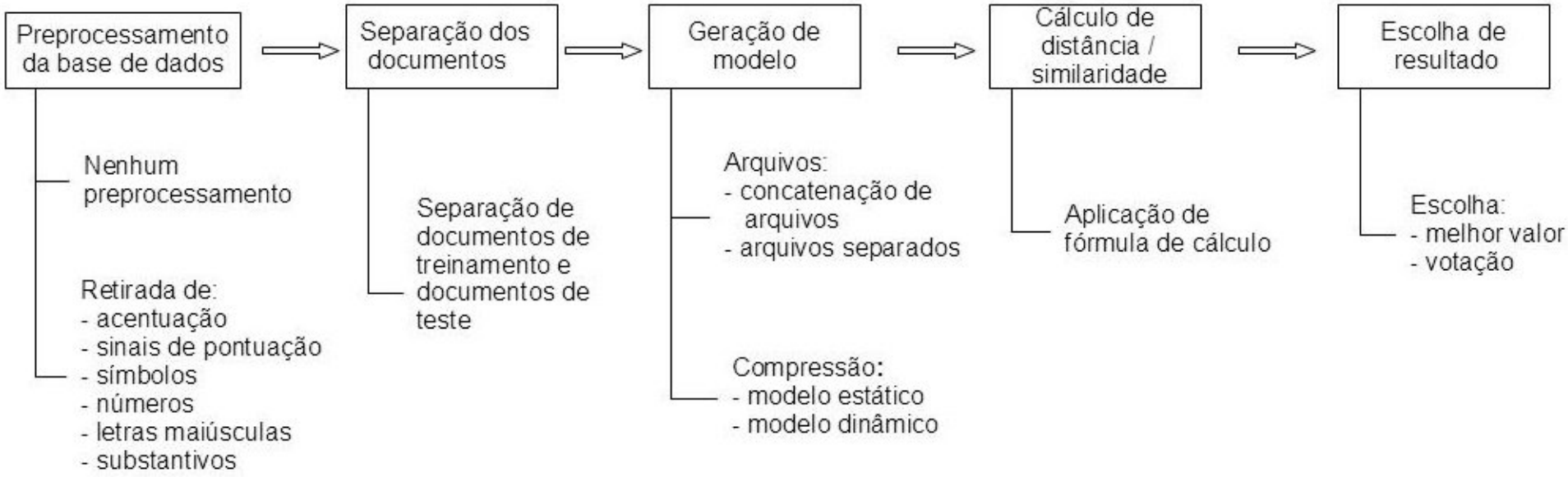


Figura 3.4: Análise de métodos

As diversas etapas são explicadas a seguir.

3.3.1. Preprocessamento da base de dados

Em uma primeira etapa podem ser feitos preprocessamentos da base de dados. Estes preprocessamentos visam simplificar a quantidade de símbolos existentes nos documentos (visando aumentar, por exemplo, o desempenho de compressores estatísticos), remover informações que possam prejudicar o teste de identificação de autoria, ou outros motivos. Este preprocessamento poderá ser importante, por exemplo, se cada documento possuir uma linha com o nome do autor ou outra informação que seja equivalente a uma assinatura. Esta informação poderá facilitará a identificação da autoria e ao mesmo tempo será desaconselhável o seu uso para a produção de um método robusto de verificação de estilometria, já que seria uma informação facilmente forjável.

Por outro lado um preprocessamento poderá eliminar símbolos que sejam importantes para a elaboração de um perfil estilométrico do autor. Como mencionado anteriormente, existem abordagens baseadas na frequência de ocorrência de sinais de pontuação, a simplificação dos sinais existentes em um documento fará com que esta característica não seja aproveitada.

3.3.2. Separação dos documentos

A etapa seguinte é a separação dos documentos em documentos que serão utilizados para o treinamento do modelo, ou seja, que comporão a base de conhecimento, e os documentos que serão submetidos a teste, ou seja, os documentos questionados.

Conforme mencionado na introdução deste trabalho, um método científico de estudo de atribuição de autoria deve considerar que os experimentos devem ser controlados. Se os documentos testados tiverem uma autoria realmente desconhecida, o resultado produzido nunca poderá ser avaliado com certeza.

Há necessidade de equilíbrio entre a quantidade de documentos separados para o treinamento e de documentos que serão utilizados para teste. Em alguns outros métodos de obtenção de estilometria de autor pode haver o problema de sobretreinamento e nestes casos a quantidade de documentos separados para treinamento deve ser selecionada mais criteriosamente. Em alguns experimentos, a quantidade de documentos para cada autor é

variável (por exemplo, para (Kukushkina, O. V., Polikarpov A. A. e Khmelev, D. V., 2001) são utilizados de 2 a 30 documentos por autor), e nestes casos o percentual de documentos separados para treinamento e para testes é bastante variável.

No trabalho de Pavelec (Pavelec, D. F., 2007) foram separados 33% dos documentos para treinamento e 67% dos documentos para testes. No trabalho de Varela (Varela, P. J. 2010) foram separados 23% dos documentos para treinamento e 77% de documentos para testes. Conforme será explicado adiante, estes dois autores são importantes para o presente trabalho porque suas bases de dados serão utilizadas para a análise comparativa.

3.3.3. Geração do modelo

A etapa de geração do modelo compreende as tarefas que são desempenhadas para o treinamento e testes dos documentos. Esta etapa pode ser subdividida em duas abordagens, conforme explicado a seguir.

3.3.4. Arquivo

Em relação aos arquivos, a etapa de geração do modelo compreende a maneira como os arquivos de treinamento serão considerados para os testes.

Em uma abordagem que busque os *1-* ou *k-nearest-neighbors*, os documentos de treinamento são considerados individualmente e os testes são feitos em relação a cada um destes documentos. Para ver como os arquivos são tratados nesta abordagem, referimos a explicação fornecida no tópico 3.2.3 - Procedimento BCN.

Em abordagens onde se deseja que o compressor extraia as características (ou modelos estatísticos) do autor a partir do conjunto de seus documentos, os documentos de treinamento são concatenados em um ou mais arquivos. Para ver como os arquivos são tratados nesta abordagem, referimos a explicação fornecida nos tópicos 3.2.1 - Procedimento SMDL e 3.2.2 - Procedimento AMDL.

Em alguns trabalhos o arquivo (ou arquivos) de treinamento é separado em diversos fragmentos e estes fragmentos são utilizados para o treinamento (Malyutov, M.B. Wickramasinghe,C. I. e Li, S., 2007). Este procedimento é equivalente a se considerar que existem diversos documentos de treinamento, então a abordagem utilizada também considerará os arquivos das mesma maneira que nos procedimentos mencionados acima.

3.3.5. Compressão

Em relação à compressão, conforme visto nos tópicos 3.2.1 e 3.2.2, existem dois procedimentos.

No primeiro procedimento é gerado um modelo estatístico a partir dos documentos de treinamento e em seguida este modelo é utilizado para os documentos de testes, permanecendo estático ao longo deste processo.

No segundo procedimento o modelo estatístico é gerado a partir do documento de treinamento e em seguida é atualizado conforme o documento de teste é processado. Desta forma o modelo estatístico do compressor é alterado dinamicamente.

Conforme pode ser verificado no tópico 3.2.3, o procedimento BCN não apresenta uma abordagem do modelo estatístico que não esteja compreendida entre as mencionadas logo acima. Desta forma, não é necessário especificar um novo procedimento em relação à compressão.

3.3.6. Cálculo da distância / similaridade

Nesta etapa é aplicada a equação de cálculo da distância ou similaridade conforme a abordagem que esteja sendo estudada. A partir das etapas anteriores os documentos foram preprocessados, separados em documentos de treinamento e de teste, os documentos de treinamento foram utilizados para a geração do modelo de cada autor e o uso de compressores forneceu informações sobre os documentos. A partir destas informações, são aplicados os métodos de cálculo desejados para a geração de uma medida de distância ou similaridade entre cada documento testado e o modelo de treinamento.

3.3.7. Escolha do resultado

Nesta última etapa é feita a atribuição de autoria conforme o resultado escolhido a partir das etapas anteriores.

Existem diversos mecanismos possíveis para a escolha do resultado.

No mecanismo de melhor valor, a atribuição de autoria é feita ao autor que obteve o melhor resultado após as etapas anteriores. O melhor valor é uma característica de cada equação de cálculo. Por exemplo, no procedimento NCD, o melhor valor é a menor distância NCD, pois esta distância é mais próxima a 0 quanto mais semelhantes forem os documentos.

No mecanismo de votação é feita a atribuição de autoria ao autor que obteve uma maior quantidade de indicações a partir de um número *n* de votos. A quantidade de votos a ser considerada é empírica, sendo em regra igual ou inferior à quantidade de documentos de treinamento que foram utilizados.

A abordagem de votação deve prever o mecanismo de escolha quando há um empate na votação, ou seja, se uma mesma quantidade de votos foi atribuída a autores diversos. E, nos casos onde a geração do modelo de arquivo foi feita através da concatenação de documentos e apenas um documento passou a ser utilizado para o treinamento, não é possível a utilização da escolha através da votação, pois cada autor poderá receber, no máximo, um único voto.

Em ambas abordagens é possível a utilização de cálculos estatísticos para a aplicação do mecanismo de escolha. Por exemplo, pode-se considerar que o melhor valor será considerado a partir da média de valores obtidos para um determinado autor. Assim, se para um autor foram utilizados 7 documentos para treinamento, a escolha seria feita primeiro calculando-se a média dos valores obtidos para o autor, e em seguida feita a atribuição para o autor que apresentou a melhor média.

3.4. Exemplos de trabalhos com compressão de dados

3.4.1. Marton, Wu e Hellerstein

Marton, Wu e Hellerstein (Marton, Y., Wu, N., e Hellerstein, L. 2005) realizaram testes com 7 bases de dados, cujas características são representadas na tabela a seguir.

Tabela 3.1: Características das bases de dados testadas por Marton, Wu e Hellerstein (Marton, Y., Wu, N., e Hellerstein, L. 2005)

Nome	Categorias	Documentos por categoria	Tamanho médio dos documentos (bytes)	Percentual de documentos de teste	Fonte dos documentos
20news	20	940 *	-	20,00%	Grupos de discussão
10news	10	899 *	4k *	20,00%	Grupos de discussão
Industry Sector	105	60 *	15k *	20,00%	Páginas da internet
Reuters 10	10	544 *	7k *	29,00%	Artigos separados por assunto
Reuters 9	9	12	2,5k *	17,00%	Artigos separados por autor
Gutemberg 10	10	4	600k *	25,00%	Documentos Literários
Federalist Papers	2	14 *	14k *	55,00%	Documentos históricos

* - dados estimados a partir da descrição feita pelo autor

Este autor testou três algoritmos/programas de compressão: RAR, gzip e LZW. O algoritmo gzip foi testado com a melhor compressão possível (*best compression*). O algoritmo LZW testado era modificado para gerar um dicionário de 16 bits. O programa RAR foi testado em seu modo padrão.

A primeira comparação utilizou o procedimento AMDL e comparou os três algoritmos/programas. A taxa de atribuição correta de documentos é representados na tabela a seguir.

Tabela 3.2: Resultados obtidos por compressor e por base de dados

Nome	RAR	LZW	GZIP
20news	90,00%	-	47,00%
10news	96,00%	66,00%	56,00%
Industry Sector	90,00%	61,00%	19,00%
Reuters 10	87,00%	84,00%	83,00%
Reuters 9	78,00%	66,00%	79,00%
Gutemberg 10	82,00%	65,00%	62,00%
Federalist Papers	94,00%	83,00%	67,00%

O autor concluiu que o formato de compressão utilizado pelo programa RAR possui um desempenho superior aos algoritmos de compressão testados em quase todos os testes executados e possui um desempenho superior ou comparável a outros trabalhos que utilizaram as mesmas base de documentos e utilizaram outras técnicas como Naïve Bayes, Naïves Bayes estendidos ou SVM.

Comparando os procedimentos BNC e AMDL, o autor verificou que o procedimento AMDL apresentou resultados superiores quando utilizados os compressores RAR e LZW e inferiores quando utilizado o compressor gzip.

3.4.2. Kukushkina, Polikarpov e Khmelev

(Kukushkina, O. V., Polikarpov A. A. e Khmelev, D. V., 2001) utilizaram 82 categorias (autores), com uma média de 4,7 documentos por autor. Para cada teste, foi utilizado o procedimento AMDL com um documento de cada autor sendo utilizado para teste e os demais documentos de cada autor sendo utilizados para treinamento (criação do arquivo concatenado). Foram testados 16 programas compressores disponíveis comercialmente / publicamente. A atribuição da autoria era feita pelo uso do melhor resultado.

Os programas compressores que apresentaram melhor resultado estão elencados na tabela abaixo.

Tabela 3.3: Resultados obtidos por compressor

Programa	algoritmo	Nº acertos	% nº acertos
rarw	Variante de LZ77 com codificação Huffmann	71	86,59%
rar	Variante de LZ77 com codificação Huffmann	58	70,73%
rk	PPMZ	52	63,41%
gzip	Shannon-Fano com codificação Huffmann	50	60,98%
ha	Dicionário + codificação aritmética	47	57,32%

3.4.3. Coutinho *et al.*

(Coutinho, B. C. *et al*., 2005) utilizaram documentos de 12 autores. Para cada autor foram utilizados 4 documentos, sendo que 3 documentos foram separados para treinamento e o documento restante foi utilizado para teste.

O procedimento utilizado foi o SMDL, descrito no tópico 3.2.1. Os documentos de treinamento foram concatenados e, em seguida, submetidos a um compressor PPM-C, gerando um modelo estatístico de compressão. Em seguida, este modelo estatístico era utilizado para comprimir o documento de teste.

A autoria do documento de teste foi atribuída conforme o método estatístico que gerou a maior taxa de compressão do documento de teste.

Foi utilizada uma única base de testes, variando-se o tamanho dos documentos de treinamento e de testes (ou seja, utilizando-se fragmentos dos documentos da base de documentos). O compressor PPM-C foi utilizado para ordem de Markov 4, 5 e 6.

Os resultados obtidos estão representados no gráfico 3.5 a seguir.

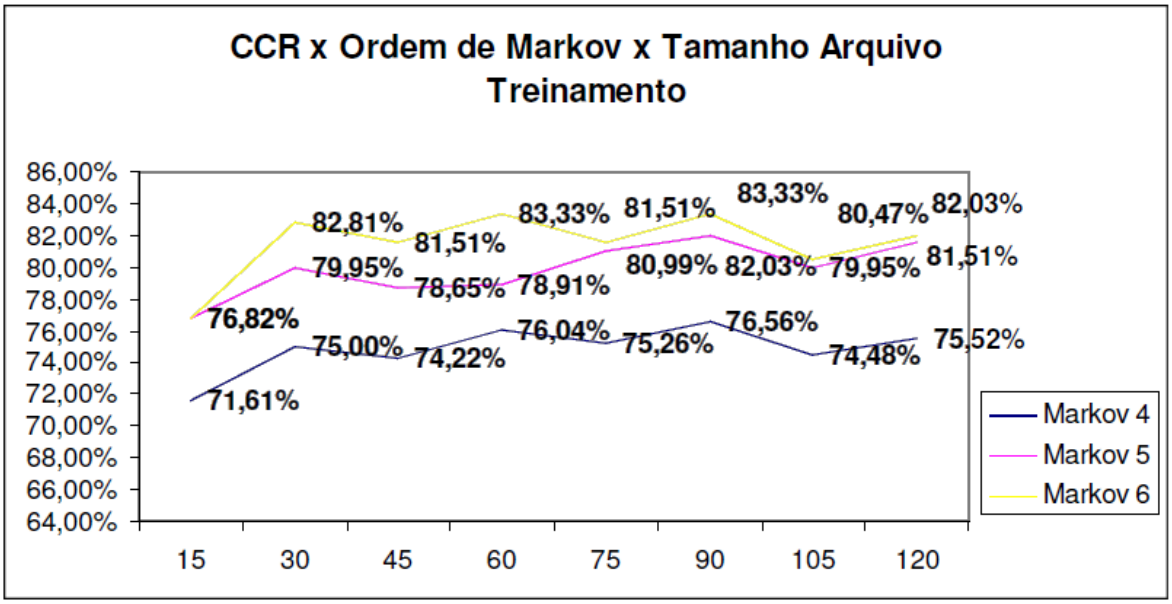


Figura 3.5: Atribuição de autoria com compressor PPM-C (Coutinho, B. C. *et al*., 2005)

Uma vantagem de se utilizar uma abordagem em que o modelo estatístico do compressor de dados seja gerado nos documentos de treinamento e permaneça estático durante os testes é que características intrínsecas do documento de teste não afetarão o modelo gerado anteriormente. Uma vez gerado o modelo não é necessário refazer todo o processo de compressão dos documentos de treinamento novamente, resultando em um menor processamento computacional.

3.5. Considerações finais

Neste capítulo foram apresentados, brevemente, as pesquisas já efetuadas sobre a atribuição de autoria, principalmente com o uso de compressores de dados. Foi também apresentada uma proposta de como entender os diversos trabalhos realizados na área, através da análise de métodos já utilizados e seu resumo em um diagrama explicativo, com um detalhamento de cada uma de suas fases.

No capítulo seguinte será apresentado o método proposto para a condução dos testes da pesquisa efetuada.

Capítulo 4

Método proposto

Neste capítulo serão apresentadas as etapas do método proposto para a atribuição de autoria através da estilística do autor com o uso de compressores de dados.

Como foram apresentadas diversas abordagens possíveis e o presente trabalho tem como um de seus objetivos a comparabilidade com resultados anteriores, será detalhado quais bases de dados foram utilizadas e em seguida serão explicadas as etapas a serem seguidas para obtenção dos resultados.

A figura 3.4 do tópico 3.3 - Análise de métodos será utilizada na explicação de cada uma das etapas.

4.1. Base de dados

Esta etapa de coleta de base de dados é uma etapa anterior às identificadas no diagrama proposto no tópico 3.3 - Análise de métodos mas é de extrema importância para o desenvolvimento de um trabalho científico. A correta identificação da base de dados permite a reprodutibilidade dos resultados e o entendimento de alguns resultados obtidos, conforme a peculiaridades dos dados utilizados.

4.1.1. Base de dados “Pavelec”

A base de dados utilizada originalmente no trabalho de Pavelec (Pavelec, D. F., 2007) era composta por 30 autores. Como esta base de dados foi utilizada apenas parcialmente em um dos experimentos realizados em seu trabalho, descreveremos apenas este subconjunto que foi utilizado em nossos experimentos .

Em situações onde uma prova pericial de documentos é exigida, muitas vezes a quantidade de informação destes documentos é pouco extensa. Por exemplo, bilhetes de sequestro, cartas demissionárias, propostas de empregos, comunicações trocadas entre empresas ou dentro do ambiente corporativo; todos estes documentos costumam ter uma extensão reduzida.

Para tentar reproduzir esta situação foram escolhidos documentos que apresentassem uma quantidade pequena de informação. Foram selecionados 20 autores de colunas de jornais, disponíveis na internet. Estes autores foram separados em dois grupos:

Tabela 4.1: Autores do grupo A-J

Autor	Fonte	Tema	Código
Celso Nascimento	Gazeta do Povo	Economia	A
Antônio Delfim Netto	Gazeta do Povo	Economia e Política	B
Carneiro Neto	Gazeta do Povo	Esportes (Futebol)	C
Carlos Brickmann	Diário do ABC	Economia e Política	D
Dom Moacyr Vitti	Gazeta do Povo	Arcebispo de Curitiba (2007) Temas sobre religião e ética	E
Francisco Giovanni D. Vieira	Jornal Eletrônico	Marketing	F
Gilberto Dimenstein	Folha de São Paulo	Jornalismo Comunitário	G
Leone Farias Grande	Diário do ABC	Economia e Análise de Mercado Financeiro	H
Reinaldo Bessa	Gazeta do Povo	Cultura e Cotidiano	I
Reginaldo Aparecido Carneiro	Jornal Eletrônico	Administração	J

Tabela 4.2: Autores do grupo P-Y

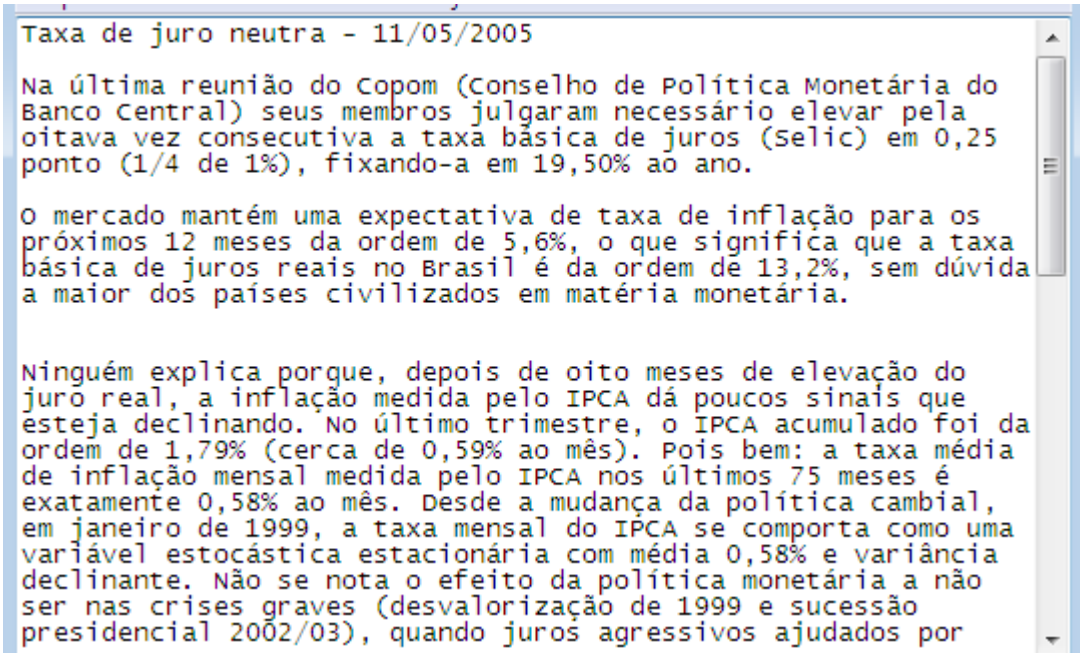
Autor	Fonte	Tema	Código
Alberto Dines	Correio Popular	Política	P
Albert Zeutoni	Correio Popular	Assuntos Gerais e Cotidiano	Q
Arnaldo Jabor	Correio Popular	Cotidiano, Economia e Política com humor	R
Cecílio Elias Netto	Correio Popular	Reflexão e Cotidiano	S
Carlos Alberto Di Franco	Correio Popular	Consultoria em Estratégia de Mídia	T
Flávio Gomes	Correio Popular	Esportes (Automobilismo)	U
Jose Pedro Martins	Correio Popular	Política	V
Manuel Carlos Cardoso	Correio Popular	Direito	W
Paulo R. Castro	Correio Popular	Psicanálise e psiquiatria	X
Rogério Verzignasse	Correio Popular	Temas Gerais	Y

Os sites correspondentes a cada uma das fontes é:

- Correio Popular: www.cpopular.com.br;
- Diário do Grande ABC: home.dgabc.com.br;
- Folha de São Paulo: www.folha.uol.com.br
- Gazeta do Povo: www.gazetadopovo.com.br;
- Jornal Eletrônico: www.wnet.com.br;
- Tribuna do Paraná: www.parana-online.com.br;

Para cada um dos autores foram escolhidos 15 documentos, com tamanho médio de 3kB. Os documentos foram todos salvos em formato ASCII preservando-se acentuação, sinais de pontuação e demais elementos, sendo retirados apenas a hifenização de palavras (presentes em alguns documentos). A figura 4.1 ilustra um dos documentos de um dos autores escolhidos.

Figura 4.1: Exemplo de documento da base de dados Pavelec (Pavelec, D. F., 2007)


Taxa de juro neutra - 11/05/2005

Na última reunião do Copom (Conselho de Política Monetária do Banco Central) seus membros julgaram necessário elevar pela oitava vez consecutiva a taxa básica de juros (Selic) em 0,25 ponto (1/4 de 1%), fixando-a em 19,50% ao ano.

O mercado mantém uma expectativa de taxa de inflação para os próximos 12 meses da ordem de 5,6%, o que significa que a taxa básica de juros reais no Brasil é da ordem de 13,2%, sem dúvida a maior dos países civilizados em matéria monetária.

Ninguém explica porque, depois de oito meses de elevação do juro real, a inflação medida pelo IPCA dá poucos sinais que esteja declinando. No último trimestre, o IPCA acumulado foi da ordem de 1,79% (cerca de 0,59% ao mês). Pois bem: a taxa média de inflação mensal medida pelo IPCA nos últimos 75 meses é exatamente 0,58% ao mês. Desde a mudança da política cambial, em janeiro de 1999, a taxa mensal do IPCA se comporta como uma variável estocástica estacionária com média 0,58% e variância declinante. Não se nota o efeito da política monetária a não ser nas crises graves (desvalorização de 1999 e sucessão presidencial 2002/03), quando juros agressivos ajudados por

Buscou-se escolher autores que expressem opiniões através de seus artigos, reduzindo-se a influência de uma linha editorial do veículo de comunicação sobre o conteúdo dos documentos. Muitos dos autores selecionados tem o seu material distribuído através de vários jornais, mantendo desta forma o conteúdo idêntico, e nestes casos não houve preocupação em se buscar o documento em um jornal determinando, sendo irrelevante a escolha do veículo de divulgação.

Os autores tiveram os temas sobre os quais escrevem estabelecidos conforme indicado e, para cada autor, também foi atribuído um código, que será utilizado no restante do documento juntamente com a indicação da base de dados a que pertencem.

A maneira como esta base de dados foi utilizada é detalhada no Capítulo 5 - Experimentos realizados e análise dos resultados.

4.1.2. Base de dados “Varela”

A base de dados utilizada originalmente no trabalho de Varela (Varela, P. J. 2010) foi utilizada integralmente em nosso experimento.

Os autores foram selecionados entre jornalistas e colunistas de jornais ou blogs. Foram utilizados as seguintes fontes de documentos, através de seus sites na internet:

- A Gazeta do Acre
- A Gazeta do Povo
- A Notícia
- Colunistas IG
- Diário do Grande ABC
- Folha UOL Online
- Jornal de Beltrão
- Jornal de Brasília
- O Estado do Paraná
- O Extra
- O Gerente
- O Povo
- O Tempo
- Paraná On-Line
- Zero Hora

Foram escolhidas 10 temas sobre os quais os autores poderiam ser classificados. Para cada um dos temas foi atribuído um código, conforme mostrado abaixo:

Tabela 4.3: Temas utilizados e códigos atribuídos

Tema	Código
Assuntos Variados	Q
Direito	R
Economia	S
Esportes	T
Gastronomia	U
Literatura	V
Política	W

Saúde	X
Tecnologia	Y
Turismo	Z

Para cada um dos temas foram escolhidos 10 autores, resultando em um total de 100 autores escolhidos. Para cada autor foi também atribuído um código, sendo possível identificar o autor a partir da combinação do código do tema com o seu código individual. Os autores escolhidos possuem relevância nacional, tendo seus artigos por vezes publicados em mais de um jornal. Desta forma, o documento foi extraído de qualquer um dos meios que tivesse reproduzido o artigo em seu conteúdo integral, sem edições, tornando-se irrelevante a escolha de um jornal ou outro.

Por exemplo, para o tema “Esporte”, foram selecionados os seguintes autores, com os respectivos códigos:

Tabela 4.4: Autores do tema "Esporte" e códigos atribuídos

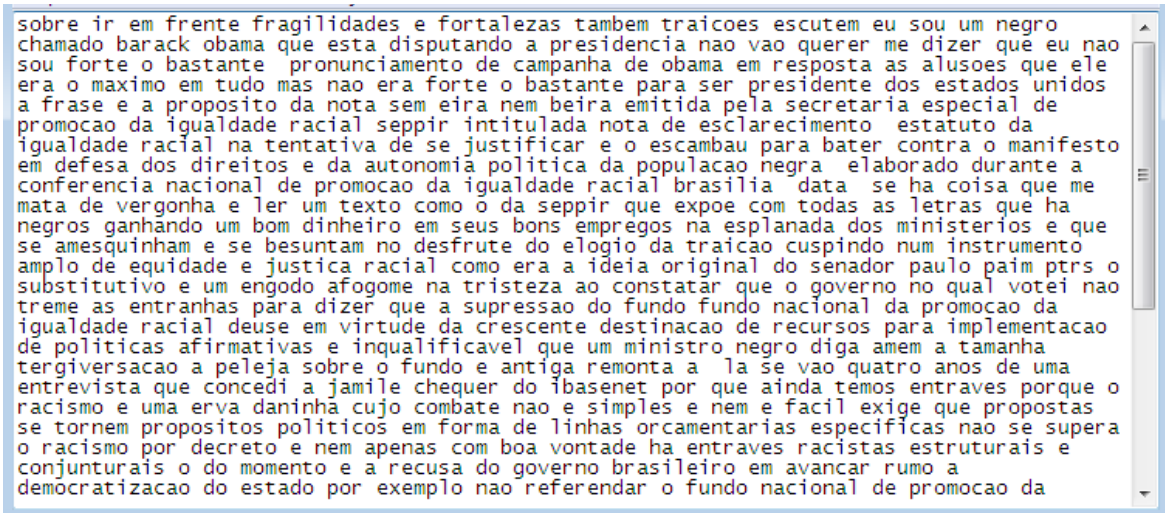
Tema	Autor	Fonte	Código
Esporte	André Ribeiro	Diário do Grande ABC	Ta
	Augusto Mafuz	O Estado do Paraná	Tb
	Diogo Olivier	Zero Hora	Tc
	Marcelo Senna	O Extra	Td
	Marcio Bernardes	Diário do Grande ABC	Te
	Sérgio Redes	O Povo	Tf
	Tostão	Gazeta do Povo	Tg
	Valdir Bicudo	A Gazeta do Paraná	Th
	Vicente Datolli	Jornal de Brasilia	Ti
	Wianey Carlet	A Notícia	Tj

Os demais autores estão elencados no Apêndice A do presente trabalho.

Para cada um dos autores foram selecionados 30 documentos, em um total de 300 documentos para cada tema e 3000 documentos no total. Estes documentos foram armazenados no formato ASCII, sem qualquer retirada de acentuação ou sinais de pontuação,

sendo apenas retirada a hifenização de palavras quando fosse o caso (em palavras separadas ao final da coluna). A figura 4.2 ilustra um documento de um dos autores selecionados.

Figura 4.2: Exemplo de documento da base de dados Varela



A maneira como esta base de dados foi utilizada é detalhada no Capítulo 5 - Experimentos realizados e análise dos resultados.

4.2. Preprocessamento da base de dados

Os dados extraídos das duas bases de dados tiveram apenas as hifenizações de palavras corrigidas para evitar que a formatação dos documentos nos sites, com a sua divisão em colunas de texto, pudessem alterar as palavras utilizadas.

Nenhum outro preprocessamento foi feito nos documentos, sendo utilizados os documentos tais como obtidos.

4.3. Separação de documentos

Conforme mencionado anteriormente, para que seja possível a comparação do resultado deste trabalho com os trabalhos anteriores de Pavelec e Varela foram utilizadas as mesmas bases de dados.

A separação dos documentos em documentos de treinamento e documentos de teste também foi a mesma utilizada pelos autores mencionados. Como cada autor teve pequenas peculiaridades em seus métodos, passaremos a analisar a seguir como esta separação foi efetuada.

4.3.1. Separação de documentos na base de dados "Pavelec"

O trabalho de Pavelec tratava da identificação de autoria e atribuição de autoria. Conforme mencionado no tópico 2.1 - Identificação de Autoria, com compressores de dados apenas a atribuição de autoria é possível de ser aplicada. Por isto, apenas reproduziremos o método de separação de documentos aplicável ao nosso caso.

Para esta base de dados, os autores foram separados em dois grupos, conforme mencionado no tópico 4.1.1. Para os testes, esta base de dados foi utilizada de três maneiras distintas.

A primeira utilizou apenas os 10 autores de códigos A a J. Os documentos foram separados com 5 documentos sendo separados para comporem a base de treinamento e os 10 documentos restantes foram separados para comporem a base de testes. Como haviam 15 documentos, foi possível repetir este procedimento em um total de 3 vezes, sem repetição de documentos de treinamento.

A figura 4.3 ilustra estas 3 separações que foram efetuadas, sendo que os números (de 1 a 15) representam cada um dos documentos de cada autor.

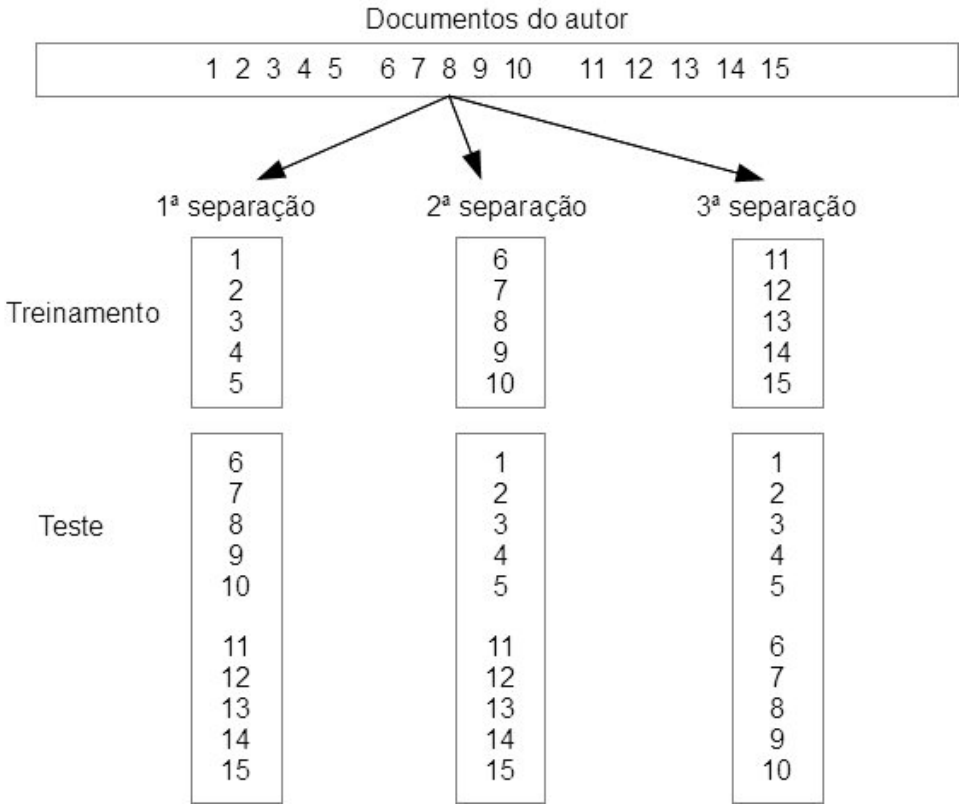


Figura 4.3: Separação de documentos de treinamento

A segunda maneira de utilização desta base de dados foi feita com os autores de código P a Y. Os procedimentos adotados foram os mesmos da primeira abordagem, permitindo assim a geração de 3 separações distintas sem repetição de documentos de treinamento.

A terceira maneira foi com o uso de todos os 20 autores selecionados. Novamente os mesmos procedimentos foram adotados, permitindo assim a geração de 3 separações distintas.

Desta forma, na base de dados "Pavelec", os documentos foram separados em 33% dos documentos para treinamento e 67% dos documentos para testes.

Como a base de dados era dividida em dois grupos de autores, e esta separação foi utilizada de três formas (grupo de autores A-J, grupo de autores P-Y e grupo de autores A-Y), e cada uma destas formas permitiu 3 separações, no total a base de dados "Pavelec" pode ser utilizada 9 vezes sem que quaisquer dos documentos de treinamento tenham sido repetidos.

Como foi possível ter acesso aos arquivos utilizados por Pavelec em seus testes foi possível determinar quais documentos foram utilizados a cada vez como treinamento e como

teste, sendo possível assim a realização de testes exatamente nos mesmos documentos que o autor havia utilizado em seus experimentos.

4.3.2. Separação de documentos na base de dados “Varela”

No trabalho de Varela foram efetuados dois testes de atribuição de autoria, sendo que para ambos a separação dos documentos foi feita da mesma maneira.

Existem 30 documentos por autor. Estes documentos foram separados com 7 documentos sendo selecionados para o treinamento e os 23 documentos restantes utilizados como teste.

Varela menciona que os documentos de treinamento foram selecionados aleatoriamente e sem repetição (Varela, P. J. 2010). Desta forma não é possível determinar quais 7 documentos foram utilizados como treinamento e quais documentos foram utilizados para teste, para cada autor. Assim, optou-se por também selecionar, de maneira aleatória, 7 documentos para comporem os arquivos de treinamento e os 23 restantes como arquivos de teste.

4.4. Geração de modelo

Conforme mencionado no tópico 3.3 - Análise de métodos, a geração de modelo compreende duas abordagens: uma abordagem em relação aos arquivos e outra em relação aos modelos gerados pelos compressores. Estas duas abordagens serão detalhadas abaixo.

4.4.1. Modelo de arquivos

Conforme visto no tópico 3.2 - Abordagens para extração de conhecimento da base de treinamento, existem duas abordagens que são utilizados para a extração de conhecimento dos documentos de treinamento.

Na primeira abordagem os arquivos são concatenados e o modelo é gerado a partir das informações disponíveis a partir desta concatenação. Esta abordagem mostra-se mais adequada quando se deseja gerar um modelo estatístico com o uso de compressores, pois presume-se que uma quantidade maior de informações do estilo de um autor serão encontradas para a geração do modelo quando uma maior quantidade de informações estiverem disponíveis.

Em outra abordagem todos os arquivos de treinamento são considerados individualmente. Nesta abordagem o estilo de um autor estará disponível de maneira esparsa, sendo que métodos de busca de conhecimento que considerem os *k-nearest-neighbors* para a atribuição do resultado serão mais promissores.

Buscando explorar a atribuição de autoria através da NCD e a comparação com outros métodos que envolvem a compressão de dados, as duas abordagens em relação aos arquivos serão utilizadas.

Assim, nas duas bases de dados (Pavelec com suas subvariações e Varela) os documentos de treinamento serão submetidos às duas abordagens. Em uma delas, os documentos de treinamento de cada autor (5 documentos de treinamento por autor na base de dados "Pavelec" e 7 documentos de treinamento por autor na base de dados "Varela") serão concatenados, produzindo um único documento de treinamento para extração de características. Na outra abordagem, cada um dos documentos de cada autor permanecerá individualizado e assim será utilizado para a extração de características.

4.4.2. Modelo de compressão

Os modelos de compressão possíveis são os mencionados nos tópicos 3.2.1 - Procedimento SMDL e 3.2.2 - Procedimento AMDL.

O procedimento SMDL gera um modelo estatístico a partir dos documentos de treinamento e mantém este modelo estático durante a compressão do documento de teste. Para isto, é necessário que o compressor possua a capacidade de gerar um modelo estatístico que possa ser armazenado (fase de treinamento) e que o compressor possa ser inicializado com um determinado modelo estatístico (fase de teste), devendo ainda manter este modelo estatístico inalterado durante a fase de teste.

Isto não é possível com os compressores de dados disponíveis comercialmente. É necessário que o programa compressor seja construído ou adaptado para possuir estas características. E entre os métodos de compressão, abordados no tópico 2.3.4 - Compressão de dados, a compressão de dados baseada em blocos não é compatível com este procedimento.

Decidiu-se, então, que apenas o procedimento AMDL será utilizado no presente trabalho. O procedimento AMDL é aquele no qual os modelos estatísticos gerados nos documentos de treinamento podem ser atualizados com os dados do documento de teste que

está sendo processado. Como este procedimento é compatível com todos métodos de compressão de dados abordados no tópico 2.3.4, será o procedimento adotado.

O modelo SMDL, entretanto, estará representado na comparação dos resultados do presente trabalho com os trabalhos anteriores de Pavelec e Varela, pois o trabalho de Pavelec utilizou este procedimento com o compressor PPM-C.

4.5. Cálculo de distância ou similaridade entre os documentos

Os documentos de treinamento e de teste são submetidos a processamento, com uso de compressores de dados, para que a distância NCD possa ser calculada.

As equações apresentadas no tópico 2.4.1 - Atribuição de autoria baseadas em compressores de dados serão utilizadas, permitindo assim uma comparação do desempenho da abordagem que utiliza a medida da distância NCD com outras abordagens sugeridas na literatura. São as equações mostradas anteriormente, respectivamente da CCC, complexidade condicional de compressão relativa, distância relativa de complexidade, razão da distância relativa de complexidade e taxa de compressão.

4.6. Escolha do resultado

A última etapa da atribuição de autoria de um documento questionado é a escolha do resultado. Como mencionado no tópico 3.3.7 - Escolha do resultado, existem dois mecanismos principais de escolha do resultado: a escolha do melhor resultado ou uma votação entre *n* melhores valores.

No presente trabalho utilizamos os dois mecanismos, quando possível, para a escolha do resultado. O uso de uma votação entre *n* melhores valores pressupõe que existam diversas medidas de distância ou similaridade para cada autor. Nos casos onde os documentos de treinamento eram concatenados gerando um único resultado por autor, não há como se pensar em escolher o resultado através de votação.

A escolha do melhor resultado foi feita através da tomada de decisão que apresentasse:

- o menor valor: por exemplo, utilizando o cálculo da distância NCD, foi escolhido como autor o que apresentou o menor valor NCD entre todos os documentos de treinamento;

- o menor valor médio: foi calculada a média entre todas as distâncias NCD entre o documento de teste e os documentos de treinamento, sendo gerado um valor médio para cada autor, e em seguida o menor valor médio foi escolhido para a atribuição de autoria

A figura 4.4 destaca quais foram os métodos utilizados no presente trabalho. Em negrito estão destacados os métodos que foram selecionados para sua utilização. Os métodos que não estão destacados (modelo estático e processamento da base de dados) não foram feitos pelos seguintes motivos:

- modelo estático: os resultados obtidos em trabalhos anteriores já contemplaram o uso do modelo estático de compressão em relação à mesma base de dados, utilizando o compressor PPM-C, e este modelo estático não é compatível com alguns compressores de documentos que serão utilizados no presente trabalho;
- processamento da base de dados com retirada de alguns símbolos: desejou-se utilizar a complexidade máxima dos documentos para que todas as informações disponíveis pudessem ser utilizadas pelos compressores de dados.

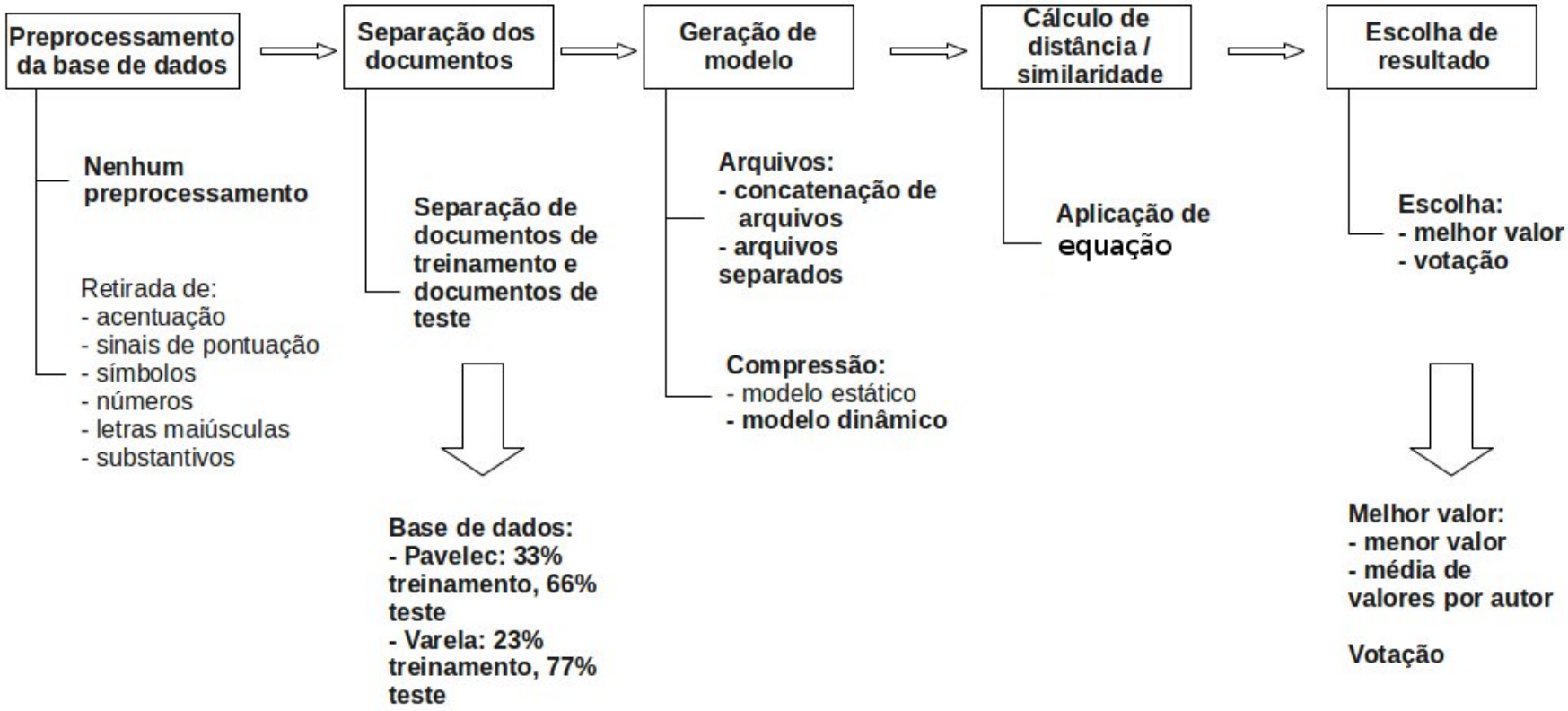


Figura 4.4: Método proposto

4.7. Análise dos resultados

Na atribuição de autoria costuma ser utilizada apenas a medida do índice de acertos, ou seja, a quantidade de documentos que foram atribuídos ao autor corretamente, expressa em percentual em relação à quantidade total de documentos. Isto é expresso pela equação

$$Índice\ de\ acertos = \frac{A}{total\ de\ documentos} \tag{10}$$

sendo que A representa os documentos que foram atribuídos corretamente.

Como os classificadores serão utilizados no modelo multiclasse, cada um dos documentos de teste será atribuído a uma categoria (autor), sendo significante a quantidade de atribuições corretas efetuadas. Em alguns testes também será verificado se houve confusão entre os autores e temas e será indicado a quantidade de atribuições feitas corretamente e incorretamente.

Também é possível utilizar curvas *Receiver operating characteristic* (ROC) para a análise de resultados. Neste caso, os valores obtidos de cada medida de semelhança entre documentos são re-normalizados (ou seja, é efetuada uma função *minmax* para cada documento questionado, atribuindo o valor 0 para o melhor resultado para cada documento de teste e o valor 1 para o pior resultado) e assinalados quais são os resultados positivos e negativos esperados.

4.8. Compressores utilizados

Para os experimentos foram utilizados três compressores diferentes, cada um representando os métodos de compressão descritos no tópico 2.3.4 - Compressão de dados.

O compressor de dados baseado no processamento de blocos utilizado foi o compressor bzip2, versão 1.0.5, disponível em http://www.bzip.org. Foram utilizadas as opções padrão de taxa de compressão com o tamanho do bloco a ser processado de 900kB. No restante do trabalho o método de compressão BZIP refere-se ao uso do compressor bzip2 aqui mencionado.

O compressor de dados baseado em dicionário utilizado foi o compressor gzip, versão 1.3.12, disponível em http://www.gzip.org. Foram utilizadas as opções padrão da taxa de compressão e o tamanho da janela deslizante desta implementação é o padrão de 32kB. No restante do trabalho o método de compressão ZIP refere-se ao uso do compressor gzip aqui mencionado.

O compressor de dados estatístico utilizado foi o PPMD, implementado no programa 7-zip, versão 9.04, disponível em http://www.7-zip.org. Foram utilizadas as opções padrão do compressor PPMD implementado, com a ordem de Markov 6. No restante do trabalho o método de compressão PPMD refere-se ao uso do compressor 7-zip aqui mencionado.

4.9. Considerações finais

Neste capítulo são propostos os métodos que serão utilizados para os testes de atribuição de autoria nas bases de dados consideradas. Após a caracterização das bases de dados, são apresentadas as etapas de separação de documentos de treinamento e de testes, como será gerado o modelo de arquivos de treinamento (documentos separados ou concatenados), os métodos de compressão que serão utilizados, equações de distância entre os documentos e como será feita a escolha do resultado.

No capítulo a seguir são apresentados os experimentos realizados e a análise dos resultados.

Capítulo 5

Experimentos realizados e análise dos resultados

Neste capítulo são detalhados os experimentos realizados e os resultados obtidos são analisados logo após cada experimento.

5.1. Idempotência na medida NCD

Em um primeiro experimento verificou-se a idempotência da medida NCD. Como mencionado no tópico 2.4.2 - Distância Normalizada de Compressão, a distância entre dois documentos idênticos é idealmente 0, dada a característica de idempotência dos compressores ideais.

Na prática esta medida afasta-se de 0 por não haverem compressores ideais, capazes de atender à equação

$$C(xx)=C(x) \quad e \quad C()=0 \tag{11}$$

sendo *C(x)* a definição já apresentada anteriormente e *C()* o tamanho da compressão de um documento vazio.

Para verificar como a medida de idempotência se apresentava em relação aos documentos que serão utilizados no presente trabalho, utilizou-se a base de dados Pavelec, com os autores A-J e todos os documentos de cada autor, em um total de 150 documentos.

A medida da idempotência foi feita através do método NCD mencionada anteriormente.

Os resultados obtidos estão representados na tabela 5.1 a seguir.

Tabela 5.1: Idempotência dos documentos da base de dados "Pavelec"

	BZIP	ZIP	PPMD
média	0,2477396636	0,0276302668	0,1915372152
desvio padrão	0,010066932	0,0016271607	0,0246612727

O gráfico que representa a idempotência dos arquivos está representado na figura 5.1.

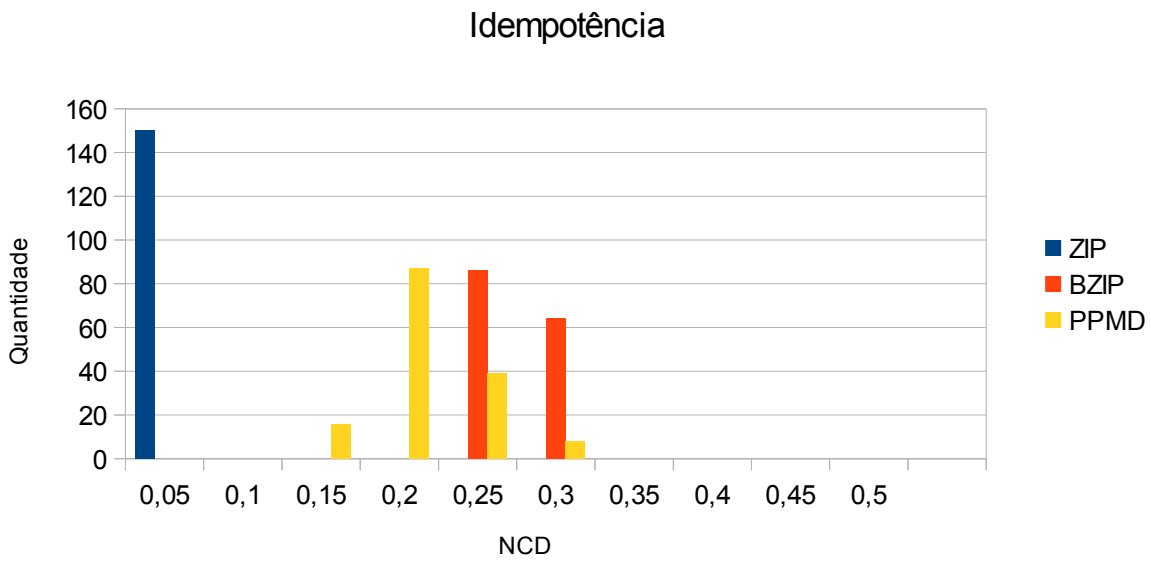


Figura 5.1: Idempotência

Como pode ser observado o compressor ZIP produz uma NCD bastante próximo a 0 enquanto os compressores BZIP e PPMD produzem um valor elevado de NCD para documentos idênticos.

Para verificar se estes valores eram devidos ao conteúdo dos documentos, foram gerados 5 documentos com 3kB de tamanho, semelhante aos encontrados na base de dados Varela, com conteúdo totalmente aleatório. Em seguida foi calculada a NCD destes arquivos, obtendo-se os resultados da tabela 5.2 e do gráfico 5.2 abaixo.

Tabela 5.2: Distância NCD de documentos de conteúdo aleatório

	BZIP	ZIP	PPMD
média	0,16523	0,02513	0,00750
desvio padrão	0,00517	0,00047	0,00000

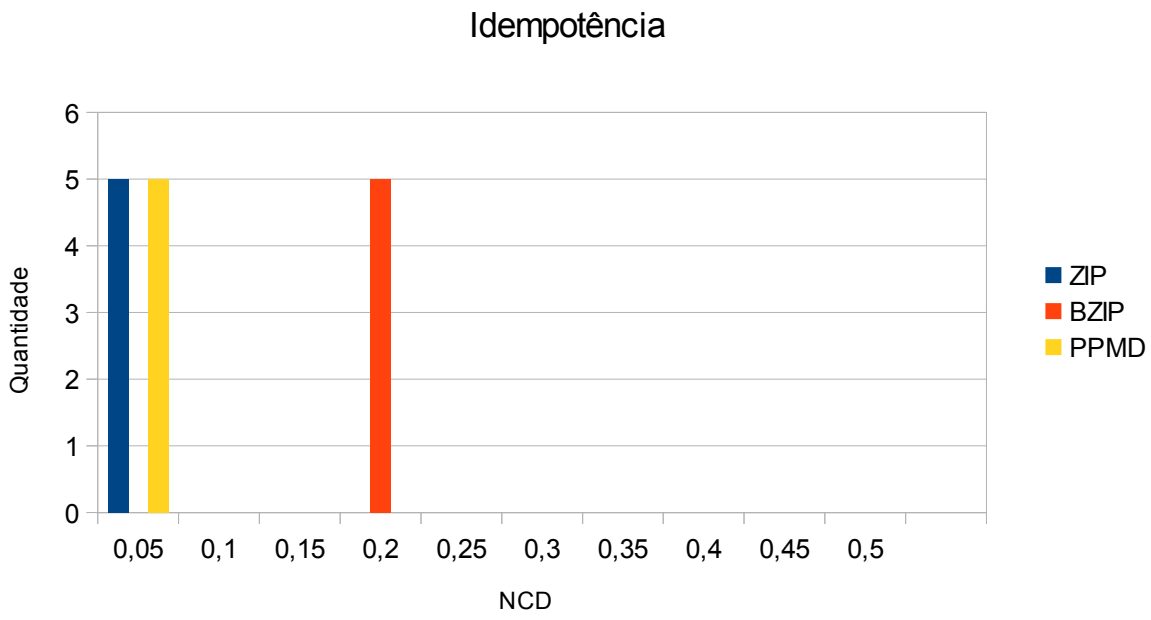


Figura 5.2: Distância NCD de documentos de conteúdo aleatório

Verifica-se que quando o conteúdo do documento é totalmente aleatório, não sendo obtida nenhuma compressão do documento com o uso de compressores, o compressor ZIP mantém os resultados obtidos, apresentando uma pequena distância NCD de idempotência. O compressor PPMD apresenta uma melhora em seu desempenho, mostrando que as informações vistas previamente passam a ser melhor aproveitadas, enquanto que o compressor BZIP mantém um resultado aproximadamente igual ao obtido com os documentos que serão utilizados nos demais testes.

O compressor ZIP beneficia-se das informações vistas anteriormente para comprimir as informações subsequentes, e como o tamanho dos arquivos está dentro de sua janela deslizante de 32kB, todo o conteúdo da repetição do arquivo (a segunda metade do arquivo concatenado é idêntica à primeira metade) é representada por poucos símbolos, havendo uma grande taxa de compactação do arquivo concatenado. O compressor PPMD também

beneficia-se do fato da informação vista anteriormente ser repetida, sendo possível utilizar um contexto de Markov elevado para representar esta duplicação de informações. O compressor BZIP, por sua vez, obtêm sua compressão pela codificação eficiente onde símbolos apareçam em uma determinada ordem ao longo de todo o documento. O fato do documento estar duplicado faz com que, no mínimo, toda sequência de dois símbolos seja vista no mínimo duas vezes, mas este ganho é proporcional ao acréscimo do tamanho do arquivo (já que dois documentos iguais, ao serem concatenados, resultarão em um documento com o dobro do tamanho). Desta forma, a medida de idempotência através do compressor BZIP tende a apresentar um desempenho inferior que os outros compressores utilizados.

Apesar dos resultados serem superiores a 0, verifica-se que no compressor ZIP a medida da distância NCD é bastante próxima a 0 enquanto nos outros compressores é mais elevada.

Entretanto, conforme será observado no resultado dos demais testes, esta diferença entre o esperado e o verificado na medida da idempotência não inviabilizou o uso da distância NCD na atribuição de autoria de documentos.

5.2. Base de dados Pavelec: documentos separados

Conforme mencionado no tópico 4.1.1 - Base de dados “Pavelec” (Pavelec, D. F., 2007), esta base de dados foi dividida em 3 subgrupos: um contendo autores cujos códigos vão de A a J, outro contendo os autores cujos códigos vão de J a Y e outro subgrupo contendo todos os 20 autores.

Os experimentos conduzidos são explicados a seguir.

5.2.1. Autores A - J

No primeiro experimento foi utilizado um subgrupo da base de dados “Pavelec” (Pavelec, D. F., 2007) contendo os documentos dos autores A – J.

O procedimento utilizado baseou-se no descrito no tópico 3.2.3 - Procedimento BCN com características representadas na figura 5.3 e na descrição a seguir.

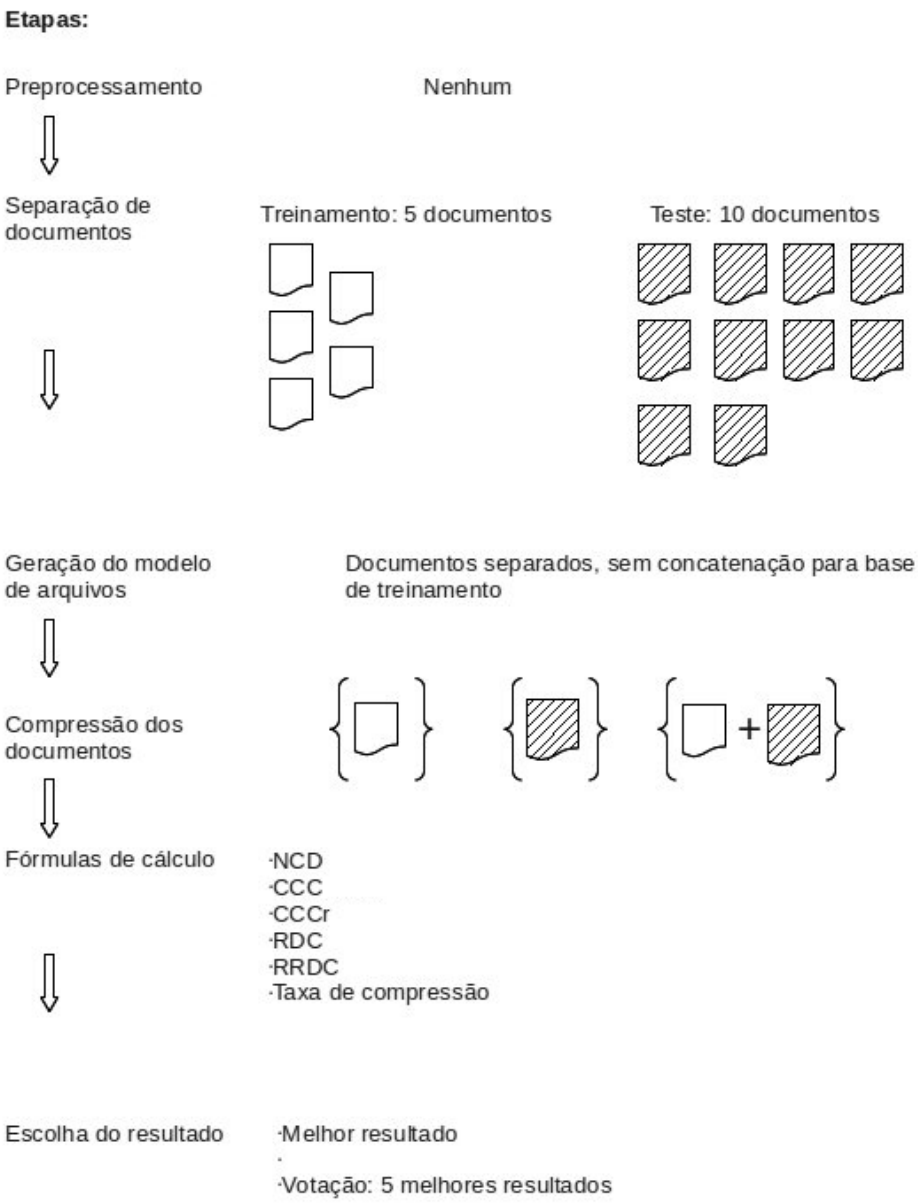


Figura 5.3: Procedimento de teste com documentos de treinamento individuais

Os documentos não sofreram nenhum preprocessamento, mantendo todos os seus símbolos (como caracteres acentuados e sinais de pontuação).

Cada autor é representado por 15 documentos. Como mencionado anteriormente, para permitir a comparação de resultados, adotou-se o mesmo protocolo que (Pavelec, D. F., 2007) para a separação de documentos de treinamento e de teste.

Foram separados 5 documentos de cada autor para a base de treinamento e os 10 documentos restantes foram utilizados como teste. Este procedimento foi repetido 3 vezes, de

maneira que os documentos de treinamento não sofressem repetição. Assim, cada teste pode ser repetido 3 vezes. A figura 4.3 (p. 53) ilustra como foi feita esta separação.

Os arquivos de treinamento foram considerados de maneira individual, ou seja, não houve concatenação dos documentos de treinamento.

A seguir procedeu-se à compressão dos documentos. Para as equações que foram adotadas, são relevantes as informações do tamanho original dos documentos de treinamento e de teste, o tamanho do documento resultante da concatenação entre o documento de treinamento e teste, e os seus respectivos tamanhos após o processo de compressão.

Após obtenção dos dados, foram efetuados os cálculos conforme as equações elencadas nos tópicos anteriores.

O primeiro mecanismo de escolha adotado foi o do melhor resultado. Os resultados são apresentados a seguir, na tabelas 5.3, 5.4 e 5.5. Para permitir desde logo comparações, são mostrados os resultados obtidos por (Pavelec, D. F., 2007) em seu trabalho utilizando classificadores SVM e o método de compressão PPM-C com uma abordagem estática de geração de modelos de compressão

Tabela 5.3: Desempenho do compressor Bzip

Documentos Treinamento	**PPM-C**	**SVM**	CCC	CCCr	RDC	RRDC	NCD
1-5	77,00%	80,00%	**95,00%**	26,00%	4,00%	4,00%	**97,00%**
6-10	80,00%	80,00%	**88,00%**	30,00%	10,00%	10,00%	**100,00%**
11-15	79,00%	72,00%	**92,00%**	25,00%	10,00%	10,00%	**94,00%**
Média	78,67%	77,33%	91,67%	27,00%	8,00%	8,00%	97,00%

Tabela 5.4: Desempenho do compressor PPMD

Docs Treinamento	**PPM-C**	**SVM**	CCC	CCCr	RDC	RRDC	NCD
1-5	77,00%	80,00%	**90,00%**	28,00%	1,00%	1,00%	**98,00%**
6-10	80,00%	80,00%	**87,00%**	34,00%	10,00%	10,00%	**99,00%**
11-15	79,00%	72,00%	**95,00%**	29,00%	10,00%	10,00%	**95,00%**
Média	78,67%	77,33%	90,67%	30,33%	7,00%	7,00%	97,33%

Tabela 5.5: Desempenho do compressor Zip

Docs Treinamento	**PPM-C**	**SVM**	CCC	CCCr	RDC	RRDC	NCD
1-5	77,00%	80,00%	**94,00%**	25,00%	0,00%	0,00%	**100,00%**
6-10	80,00%	80,00%	**90,00%**	29,00%	9,00%	9,00%	**99,00%**
11-15	79,00%	72,00%	**93,00%**	29,00%	10,00%	10,00%	**98,00%**
Média	78,67%	77,33%	92,33%	27,67%	6,33%	6,33%	99,00%

Como pode ser observado houve um bom desempenho dos métodos de cálculo de distância ou semelhança entre os documentos pelas equações CCC e NCD. Estes resultados foram destacados em negrito para uma visualização mais fácil. E os resultados NCD foram superiores a CCC, em média, em 6 ponto percentuais.

Nos três compressores considerados os valores obtidos foram superiores às demais equações propostas por (Malyutov, M.B. Wickramasinghe,C. I. e Li, S., 2007), conforme os autores haviam constatado em seu trabalho. Por este motivo, estas equações deixarão de ser consideradas nos próximos testes.

Os resultados obtidos foram, também, superiores aos alcançados por Pavelec (Pavelec, D. F., 2007) em seu trabalho. Os resultados são comparáveis por ter sido usada a mesma base de dados com o mesmo protocolo de separação de documentos de treinamento e de teste.

Apesar de também ter sido usado um compressor, o resultado PPM-C utilizou os documentos de treinamento de uma maneira diferente, extraindo um modelo estatístico após a concatenação dos documentos e não com os documentos considerados de maneira individual, como neste teste.

Ao se observar a média do resultado obtido constata-se que o compressor utilizado foi pouco relevante, havendo apenas uma pequena diferença percentual entre os resultados obtidos pelo compressor ZIP (que obteve o melhor resultado médio) e o compressor BZIP.

Uma segunda maneira de escolha para a atribuição de autoria foi feita através de votação. Neste método, os 5 melhores resultados para cada documento de teste foram selecionados, sendo escolhido o resultado mais votado. Em caso de empate na votação, escolheu-se o autor que tenha obtido a posição mais elevada entre os votos, ou seja, o autor que teve o primeiro voto atribuído a ele.

Os resultados são mostrados na tabela 5.6 a seguir. Como algumas equações de cálculo foram dispensadas, é possível exibir o resultado de cada compressor na mesma tabela.

Tabela 5.6: Comparativo de desempenho com escolha por votação

			Bzip		PPMd		Zip	
Documento Treinamento	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	80,00%	77,00%	86,00%	93,00%	86,00%	96,00%	90,00%	**97,00%**
6-10	80,00%	80,00%	89,00%	97,00%	90,00%	98,00%	90,00%	**99,00%**
11-15	72,00%	79,00%	90,00%	92,00%	90,00%	91,00%	88,00%	**95,00%**
Média	77,33%	78,67%	88,33%	94,00%	88,67%	95,00%	89,33%	**97,00%**

Observa-se que há uma piora dos resultados ao se efetuar a atribuição de autoria por meio de votação. Apesar dos resultados serem bastante satisfatórios, houve uma perda em relação à atribuição da autoria pelo melhor resultado. Uma justificativa para isto é o fato que o método proposta por Cilibrasi busca a distância entre dois documentos e, ao fazer a comparação com vários documentos, a escolha do melhor resultado é feita com o documento que apresentar a menor distância, existindo *n* chances que um dos documentos do autor do documento questionado seja bastante semelhante. Ao se fazer um processo de votação, é necessário que mais documentos do autor apresentem uma distância menor. Desta forma, um único documento semelhante do autor torna-se menos significativo.

Para verificar esta hipótese, procedeu-se à atribuição de autoria pela média das distâncias NCD. Para isto, foi calculada a média da distância NCD entre o documento questionado e os 5 documentos de treinamento de cada autor. A atribuição de autoria foi feita para o autor que apresentou o melhor resultado médio. Os resultados são mostrados na tabela 5.7 a seguir.

Tabela 5.7: Comparativo de desempenho com escolha da melhor média de resultados

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	80,00%	77,00%	90,00%	91,00%	86,00%	92,00%	89,00%	**97,00%**
6-10	80,00%	80,00%	86,00%	97,00%	84,00%	97,00%	85,00%	**98,00%**
11-15	72,00%	79,00%	89,00%	91,00%	89,00%	90,00%	90,00%	**97,00%**
Média	77,33%	78,67%	88,33%	93,00%	86,33%	93,00%	88,00%	**97,33%**

Os resultados apresentados são consistentes com os apresentados anteriormente, com o método CCC e NCD possuindo resultados superiores aos obtidos por Pavelec (Pavelec, D. F., 2007) e os resultados NCD sendo superiores aos obtidos pelo método CCC. Observa-se, entretanto, que os valores são inferiores aos obtidos pela escolha do melhor valor ou pela votação. Considerando-se que os resultados obtidos são elevados independente mecanismo de escolha para atribuição de autoria, vislumbra-se que o uso de uma base de dados maior, com uma quantidade maior de autores possíveis, poderá apresentar diferenças mais significativas entre os mecanismos de escolha.

5.2.2. Autores P-Y

O segundo experimento utilizou os autores P-Y da base de dados "Pavelec" (Pavelec, D. F., 2007).

Foram utilizados os mesmos métodos utilizados no tópico anterior para a preparação dos documentos, seu processamento e extração de resultado.

São apresentados resultados apenas das equações de cálculo CCC e NCD. No mecanismo de escolha para atribuição de autoria que considera apenas o melhor resultado, foram obtidos os seguintes valores, conforme tabela 5.8.

Tabela 5.8: Comparativo de desempenho de escolha do melhor resultado

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	87,00%	89,00%	83,00%	**98,00%**	90,00%	97,00%	92,00%	97,00%
6-10	88,00%	91,00%	82,00%	95,00%	93,00%	96,00%	**98,00%**	**98,00%**
11-15	91,00%	93,00%	73,00%	85,00%	91,00%	**99,00%**	92,00%	96,00%
Média	88,67%	91,00%	79,33%	92,67%	91,33%	**97,33%**	94,00%	97,00%

Verifica-se que novamente o método NCD apresenta um desempenho superior ao obtido com o método CCC e que os resultados são, em geral, superiores aos obtidos por Pavelec (Pavelec, D. F., 2007) em seu trabalho.

Observa-se que o compressor BZIP apresentou desempenho inferior no método CCC em relação aos demais compressores, com valores de atribuição inferiores aos resultados de Pavelec (Pavelec, D. F., 2007) para quaisquer subgrupos de teste. Analisando-se os

documentos desta base de dados, não se encontrou nenhuma característica intrínseca aos documentos, ao compressor ou à equação de cálculo que pudesse justificar este resultado.

No mecanismo de escolha através de votação, foram obtidos os resultados expressos na tabela 5.9.

Tabela 5.9: Comparativo de desempenho de escolha por votação

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	80,00%	77,00%	82,00%	97,00%	90,00%	97,00%	93,00%	**98,00%**
6-10	80,00%	80,00%	79,00%	**98,00%**	88,00%	**98,00%**	**98,00%**	**98,00%**
11-15	72,00%	79,00%	79,00%	90,00%	90,00%	**97,00%**	89,00%	**97,00%**
Média	77,33%	78,67%	80,00%	95,00%	89,33%	97,33%	93,33%	**97,67%**

Neste teste o compressor BZIP voltou a apresentar resultados satisfatórios, sendo iguais ou melhores aos obtidos por (Pavelec, D. F., 2007). Como observado anteriormente, o método NCD apresenta resultados superiores, em média, em 6 pontos percentuais em relação ao método CCC.

O mecanismo de escolha baseado na média dos valores para os documentos de treinamento do autor são apresentados na tabela 5.10 a seguir.

Tabela 5.10: Comparativo de desempenho de escolha do melhor resultado médio

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	80,00%	77,00%	82,00%	98,00%	92,00%	96,00%	95,00%	**99,00%**
6-10	80,00%	80,00%	80,00%	**98,00%**	95,00%	96,00%	98,00%	97,00%
11-15	72,00%	79,00%	80,00%	87,00%	92,00%	**95,00%**	93,00%	**95,00%**
Média	77,33%	78,67%	80,67%	94,33%	93,00%	95,67%	95,33%	**97,00%**

É, novamente, verificado um desempenho inferior do método CCC com o compressor BZIP, sendo apenas igual ou ligeiramente superior aos resultados obtidos por Pavelec (Pavelec, D. F., 2007).

Os resultados entre este mecanismo de escolha e os anteriores são bastante semelhantes, tornando-se difícil justificar a superioridade de um método em relação a outro.

A diferença entre os resultados das equações CCC e NCD, para os demais compressores, tornou-se menor, sendo de aproximadamente 3 pontos percentuais. Nota-se, entretanto, que este desempenho ligeiramente superior do método NCD e do compressor ZIP é constante em todos os testes efetuados nestes duas variações de base de dados. A proximidade dos resultados, entretanto, dificulta qualquer análise, pois a variação de um ponto percentual significa que apenas em um dos documentos testados houve uma escolha diferente entre cada mecanismo testado.

5.2.3. Autores A-Y

Esta outra variação da base de dados "Pavelec" possui uma quantidade maior de autores, sendo considerados de maneira conjunta os 20 autores tratados anteriormente.

Todos os procedimentos de teste são os mesmos efetuados anteriormente, com a única diferença residindo no fato que existirão 20 autores possíveis para cada documento de teste. Há também uma maior repetição de temas abordados para cada um dos autores. Desta forma, há um aumento da complexidade na realização da tarefa.

Para o primeiro teste, com a atribuição de autoria sendo feita através da escolha do melhor resultado, temos os resultados representados na tabela 5.11.

Tabela 5.11: Comparativo de desempenho de escolha do melhor resultado

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	83,00%	84,00%	83,00%	97,00%	86,50%	97,00%	89,50%	**98,00%**
6-10	83,00%	83,00%	83,00%	96,50%	89,00%	96,00%	91,50%	**97,00%**
11-15	85,00%	86,00%	79,00%	88,50%	91,00%	94,50%	89,50%	**95,50%**
Média	83,67%	84,33%	81,67%	94,00%	88,83%	95,83%	90,17%	**96,83%**

Podemos verificar que, apesar da maior complexidade da base de dados, o desempenho da abordagem adotada neste teste permaneceu compatível com os observados anteriormente. o método de cálculo NCD permaneceu com um desempenho superior ao método CCC, em média, em 11 pontos percentuais.

O desempenho do compressor BZIP com o método CCC foi semelhante aos resultados obtidos por Pavelec (Pavelec, D. F., 2007). Para os demais compressores o método CCC apresenta um resultado melhor.

O fato de haver uma quantidade maior de possíveis autores para o documento de teste influenciou muito pouco o resultado obtido. Isto pode ser melhor visualizado no gráfico da figura 5.4, onde é representado o resultado para as três divisões de documentos utilizadas: uma composta por 20 autores e duas compostas por 10 autores (diferentes entre si) cada.

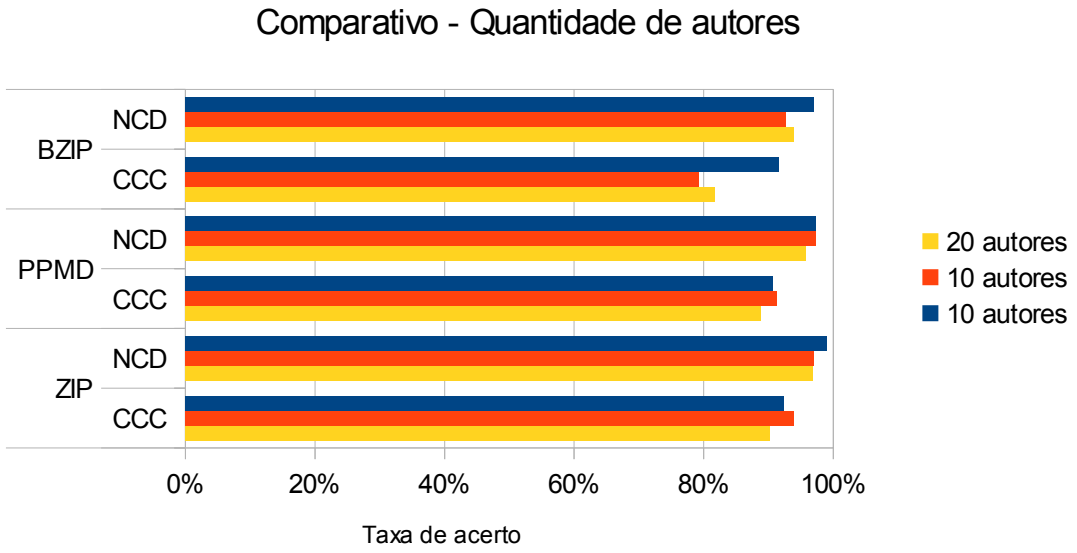


Figura 5.4: Comparativo de desempenho

O segundo teste é feito com a atribuição de autoria sendo feita a partir de votação. Os resultados são mostrado na tabela 5.12.

Tabela 5.12: Comparativo de desempenho de escolha por votação

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	83,00%	84,00%	79,00%	95,00%	83,00%	96,50%	87,00%	**97,00%**
6-10	83,00%	83,00%	81,50%	97,00%	87,50%	96,50%	90,50%	**98,50%**
11-15	85,00%	86,00%	79,50%	89,50%	84,50%	94,00%	84,00%	**95,50%**
Média	83,67%	84,33%	80,00%	93,83%	85,00%	95,67%	87,17%	**97,00%**

Verifica-se que também quase não houve alteração nos resultados obtidos com o método NCD apesar da maior quantidade de autores possíveis. O método CCC, por sua vez, comparado com os resultados obtidos por (Pavelec, D. F., 2007), apresenta valores inferiores para o compressor BZIP e valores semelhantes para o compressor PPMD. A diferença havida entre o método NCD e CCC é de aproximadamente 10 pontos percentuais.

O gráfico da figura 5.5 ilustra a pequena influência no desempenho deste método pelo aumento da quantidade de autores possíveis.

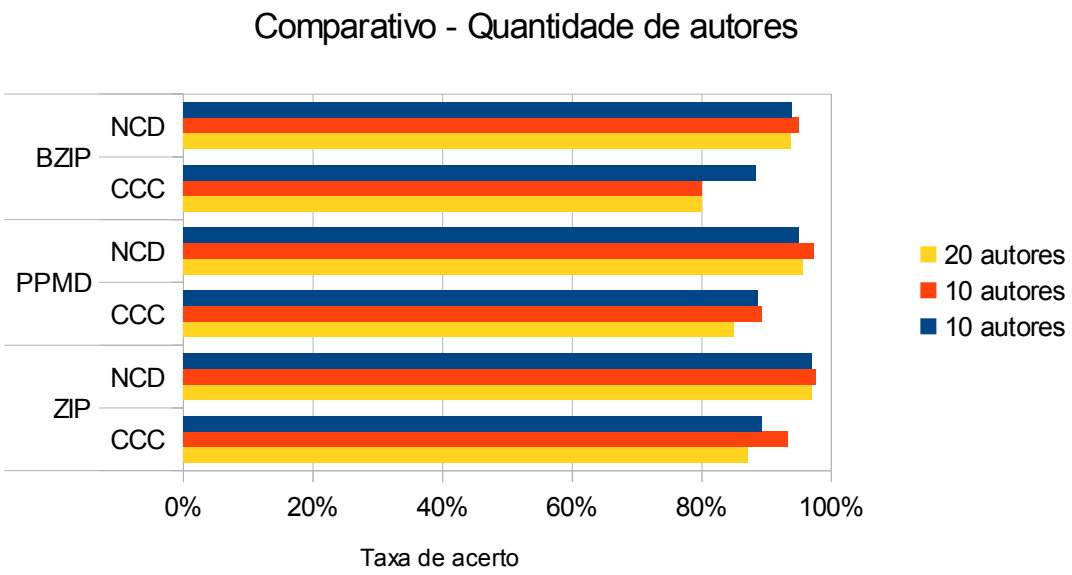


Figura 5.5: Comparativo da taxa de acerto

O terceiro teste, com a atribuição de autoria sendo feita considerando a média de resultados dos documentos de treinamento para cada autor é apresentada na tabela 5.13 a seguir.

Tabela 5.13: Comparativo de desempenho de escolha pelo melhor resultado médio

			Bzip		PPMd		Zip	
Testes	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	83,00%	84,00%	84,00%	95,00%	87,50%	94,00%	91,00%	**97,50%**
6-10	83,00%	83,00%	81,00%	97,00%	89,50%	95,50%	91,00%	**98,00%**
11-15	85,00%	86,00%	82,00%	88,00%	87,00%	92,50%	90,00%	**96,00%**
Média	83,67%	84,33%	82,33%	93,33%	88,00%	94,00%	90,67%	**97,17%**

Os resultados são semelhantes aos obtidos anteriormente. o método NCD apresenta resultados melhores que o método CCC, com o compressor ZIP apresentando o melhor resultado que os compressores PPMD e BZIP. O gráfico 5.6 ilustra a pequena influência do aumento de quantidade de autores em relação ao índice de acertos.

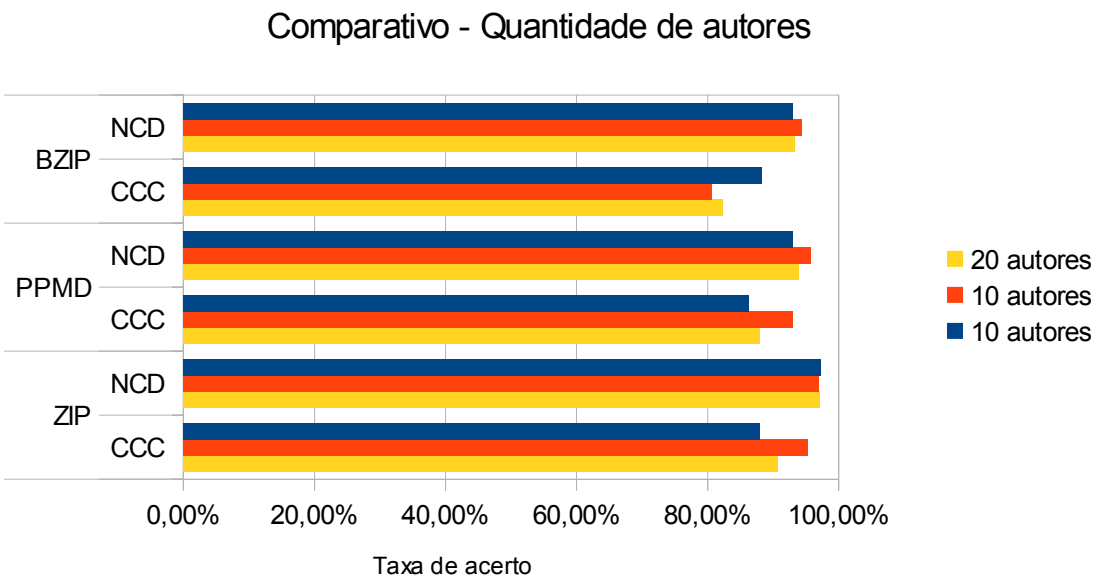


Figura 5.6: Comparativo da taxa de acerto

5.2.4. Conclusões dos experimentos na base de dados Pavelec com documentos de treinamento separados

Nesta primeira série de experimentos foram executados diversos testes com a base de dados “Pavelec”, com o uso dos documentos de treinamento de maneira individual, isto é, sem que haja uma concatenação dos documentos de treinamento para sua utilização com compressores de dados.

Os resultados apresentados demonstram que o método de cálculo NCD para a atribuição de autoria apresenta os melhores resultados independentemente do compressor considerado. O método CCC possui resultados inferiores ao NCD, e em alguns casos é igual ou inferior aos resultados obtidos por (Pavelec, D. F., 2007).

O compressor ZIP apresentou os melhores resultados, sendo seguido pelo compressor PPMD. O compressor BZIP apresentou o pior resultado entre os três compressores. Isto parece ser coerente com os resultados obtidos em relação à distância medida de um documento em relação a ele mesmo, verificando-se a idempotência do compressor e a distância NCD obtida com isto.

O gráfico 5.7 ilustra os resultados médios obtidos para o primeiro método de escolha de resultado para atribuição de autoria, que é o método de escolha a partir do melhor resultado obtido por cado método.

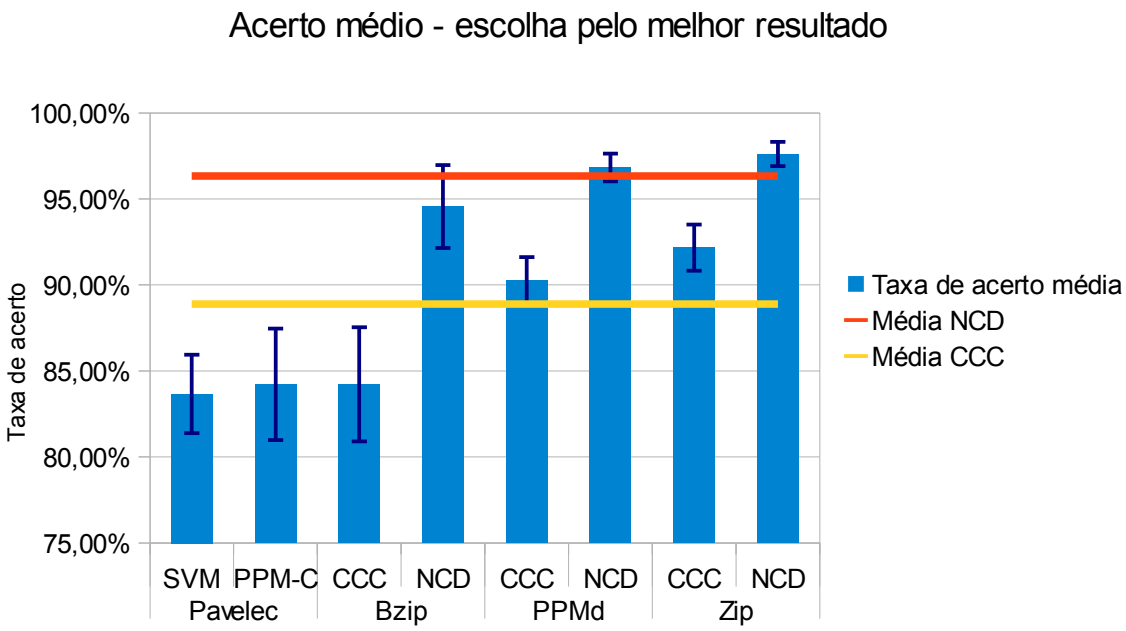


Figura 5.7: Comparativo da taxa de acerto com escolha pelo melhor resultado

Neste gráfico estão representados as taxas de acerto médios da atribuição de autoria para cada um dos compressores e os métodos considerados, bem como os resultados obtidos por (Pavelec, D. F., 2007).

Além dos valores médios da taxa de acerto são indicados:

- o desvio padrão, acima de cada resultado, indicado por uma barra de cor azul escura;
- a taxa média de acerto do método NCD para os três compressores analisados;
- a taxa média de acerto do método CCC para os três compressores.

Os valores obtidos por Pavelec são apenas indicados para comparação, não fazendo parte das taxas médias de acerto.

Verifica-se que os compressores PPMD e ZIP apresentam os resultados mais homogêneos, possuindo um desvio padrão sensivelmente menor que o compressor BZIP. E, considerando-se os compressores de melhor resultado, o desvio padrão obtido no método de cálculo NCD é menor que a CCC.

O gráfico 5.8, a seguir, ilustra o resultado médio de atribuições obtidos em relação ao segundo mecanismo de escolha, que era o mecanismo de votação dos 5 melhores resultados.

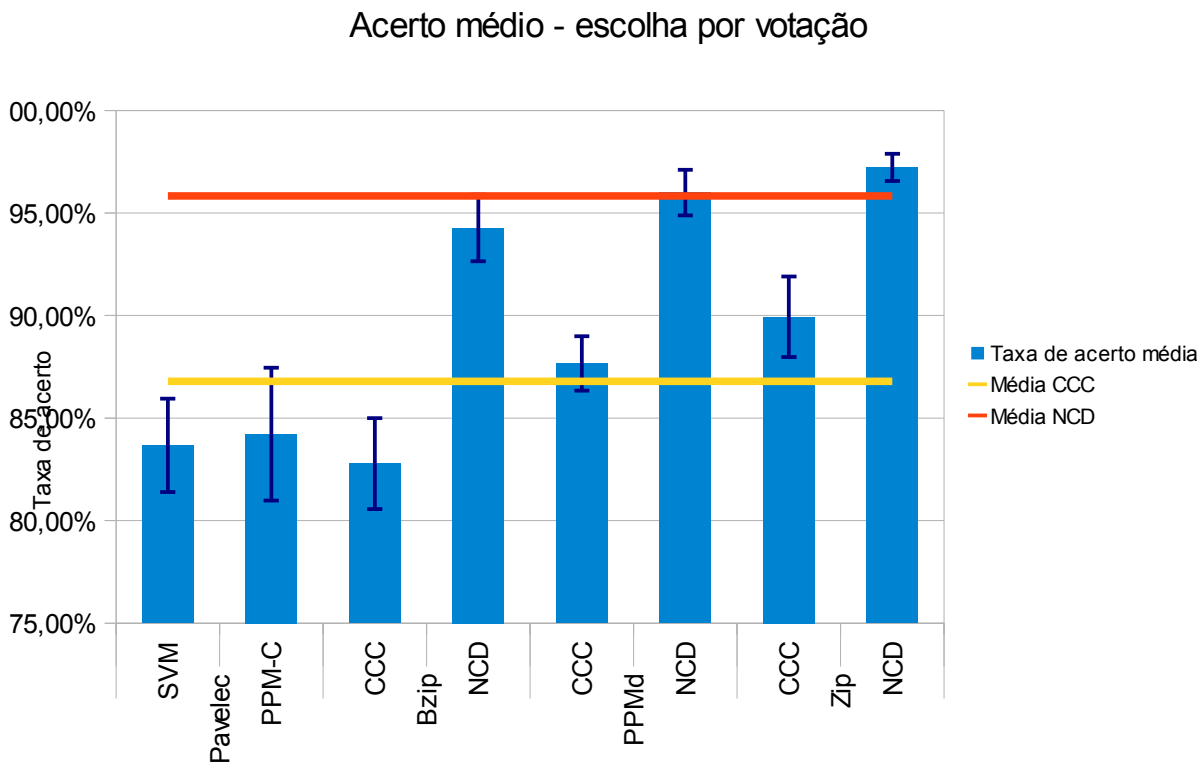


Figura 5.8: Comparativo da taxa de acerto com escolha por votação

Neste mecanismo de escolha os resultados obtidos pela NCD apresentaram pouca diferença, mantendo aproximadamente a mesma média de acerto e o mesmo desvio padrão que o uso do melhor resultado como mecanismo de escolha para atribuição de autoria.

O método CCC apresentou uma maior consistência de resultados (um desvio padrão menor), mas o resultado médio diminuiu para o compressor BZIP, com resultados inferiores aos obtidos por Pavelec (Pavelec, D. F., 2007).

O gráfico 5.9 a seguir, mostra os valores médios de acerto na atribuição de autoria utilizando o terceiro mecanismo de escolha, feito com a escolha da melhor média dos valores NCD ou CCC.

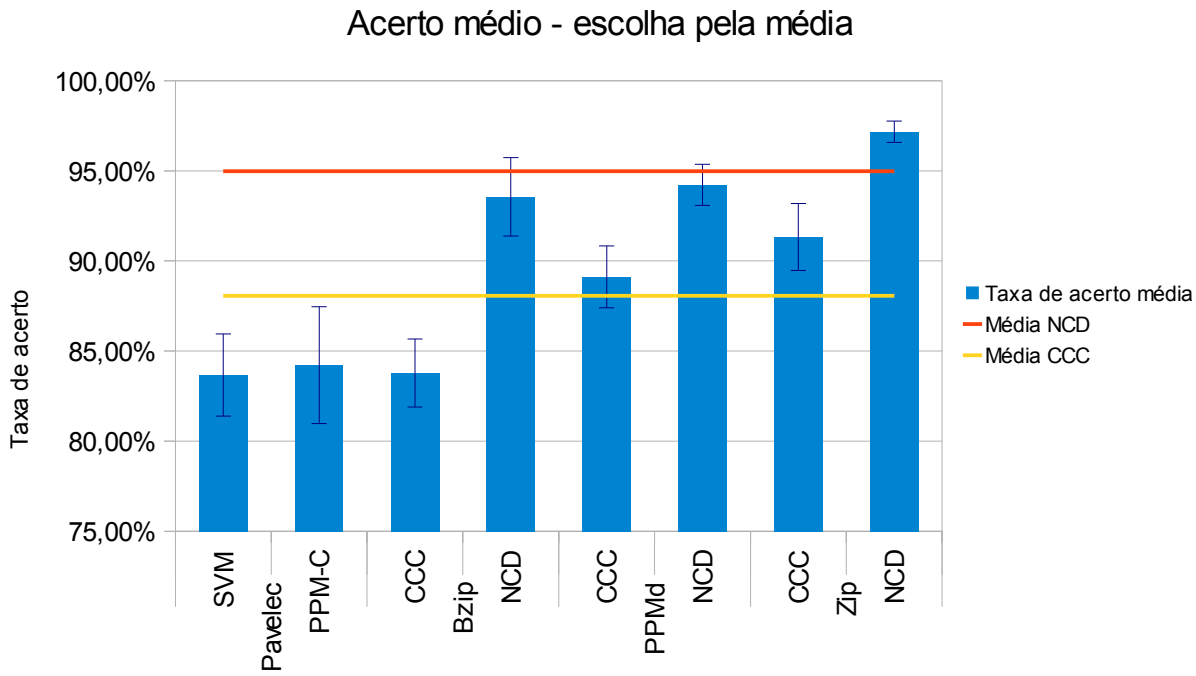


Figura 5.9: Comparativo da taxa de acerto com escolha pela média de resultados

Neste terceiro método de escolha os resultados obtidos são semelhantes aos anteriores, com o compressor BZIP apresentando o pior desempenho entre os compressores e o método NCD apresentando um resultado superior ao CCC.

É possível verificar que o método NCD apresenta resultados semelhantes independentemente do método de escolha de atribuição de autoria utilizado, como é possível visualizar no gráfico 5.10 a seguir, com o compressor ZIP (que apresentou os melhores resultados).

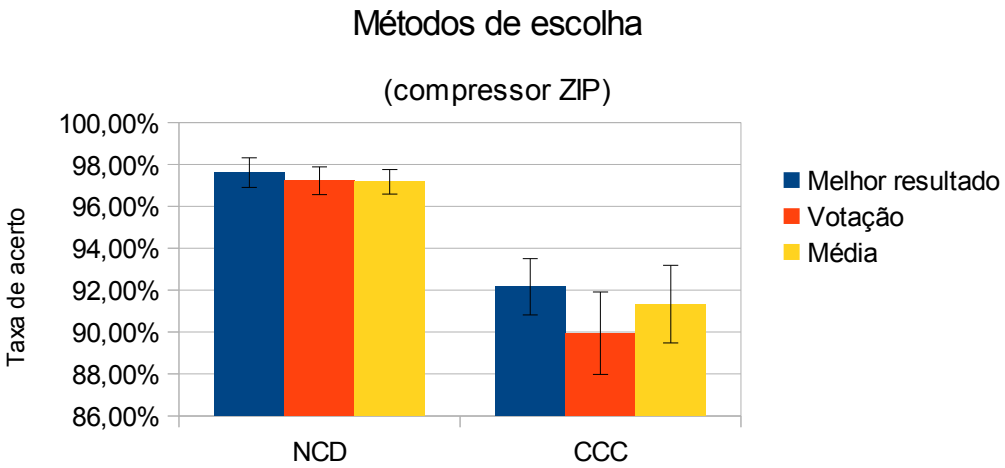


Figura 5.10: Comparativo da taxa de acerto dos métodos de escolha

O uso do mecanismo de escolha de votação e de melhor média de resultados obtidos tendem a "normalizar" o resultado obtido, fazendo com que sejam necessários mais resultados bons de atribuição a um determinado autor para que ele seja escolhido. O uso do melhor resultado, por sua vez, necessita apenas do melhor resultado, sendo irrelevante os resultados dos demais documentos. Há maior benefício para o método NCD quando há um número suficiente de documentos de treinamento e um deles apresenta um resultado superior aos demais.

5.3. Base de dados Pavelec: documentos concatenados

Nesta série de experimentos serão utilizadas as mesmas bases de dados do tópico anterior, com uma diferença importante em relação aos documentos de treinamento.

Os documentos de treinamento de cada autor serão concatenados em um único documento. Desta forma, para cada autor, os testes serão realizados utilizando-se um documento de teste e um documento-modelo de treinamento, conforme é ilustrado na figura 5.11.

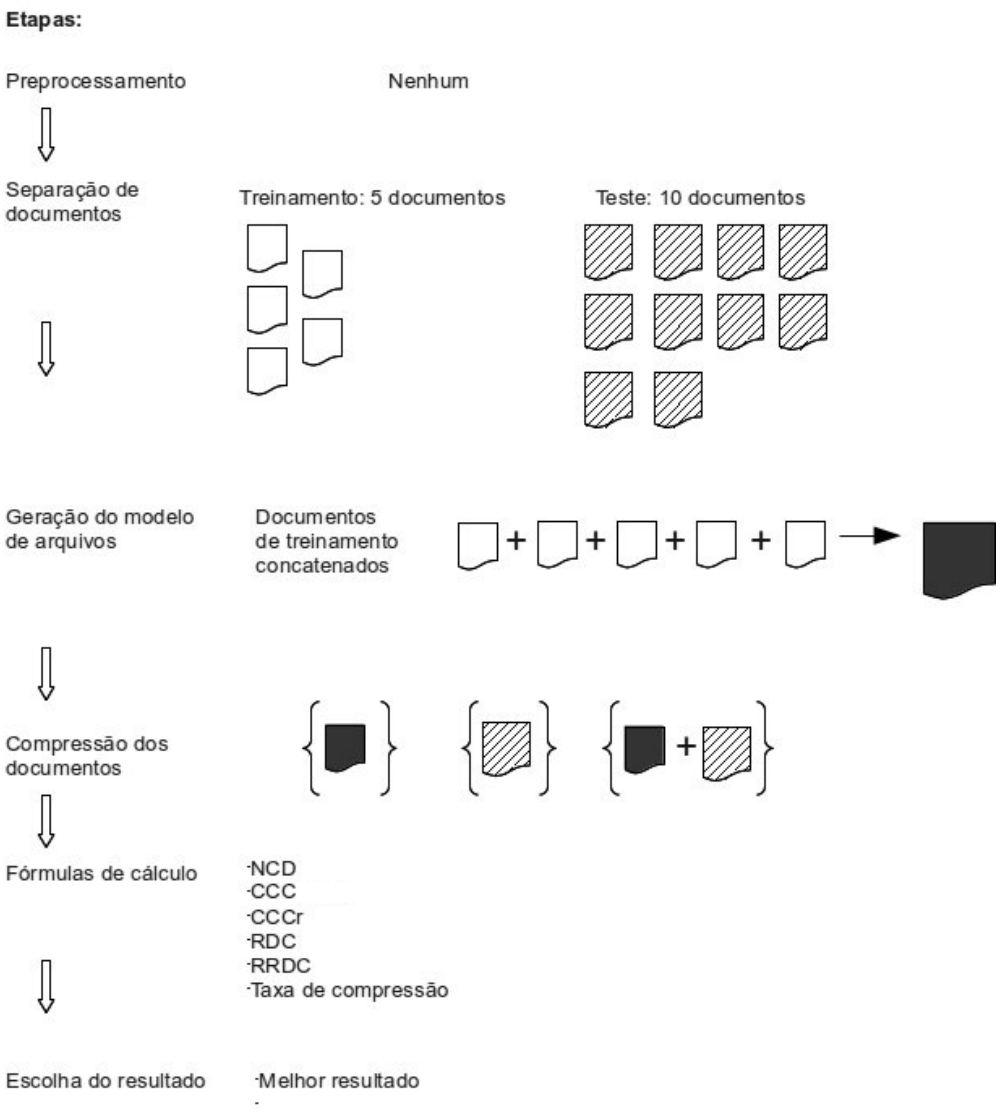


Figura 5.11: Procedimento de testes com documentos de treinamento concatenados

Este procedimento foi descrito no tópico 3.2.2 - Procedimento AMDL. Este procedimento requer menos processamento computacional que o método utilizado nos testes anteriores, pois para cada documento questionado são efetuados menos processos de compressão (equivalente a termos apenas um documento de treinamento, contra os 5 documentos de treinamento utilizados anteriormente).

Foram utilizados os mesmos subgrupos da base de dados “Pavelec”: um subgrupo contendo os autores cujos códigos vão de A-J, outro contendo os autores cujos códigos vão de J-Y e um terceiro subgrupo contendo todos os 20 autores.

Os experimentos conduzidos são explicados a seguir.

5.3.1. Autores A - J

Com exceção do fato que os documentos de treinamento de cada autor foram concatenados para a geração do modelo de treinamento, todos os demais procedimentos utilizados foram os mesmos. Por este motivo apenas apresentaremos detalhes que forem diferentes em relação ao já mencionado.

O primeiro mecanismo de escolha adotado foi o do melhor resultado. A tabela 5.14 mostra os resultados obtidos

Tabela 5.14: Comparativo de desempenho de escolha pelo melhor resultado

			Bzip		PPMd		Zip	
Teste	**SVM**	**PPM-C**	**CCC**	**NCD**	**CCC**	**NCD**	**CCC**	**NCD**
1-5	80,00%	77,00%	98,00%	57,00%	97,00%	59,00%	96,00%	72,00%
6-10	80,00%	80,00%	98,00%	73,00%	95,00%	70,00%	93,00%	83,00%
11-15	72,00%	79,00%	98,00%	57,00%	96,00%	56,00%	95,00%	76,00%
Média	77,33%	78,67%	98,00%	62,33%	96,00%	61,67%	94,67%	77,00%

Verifica-se que, ao se utilizarem os documentos de treinamento concatenados, o desempenho do método CCC foi muito superior aos resultados obtidos com a NCD. Quase todos os resultados NCD são inferiores aos obtidos por (Pavelec, D. F., 2007), enquanto todos os resultados CCC são superiores.

Para a NCD o compressor ZIP continua sendo o que apresenta melhor resultado enquanto o compressor BZIP apresenta os melhores resultados para o método CCC.

Estes mesmos resultados são apresentados no gráfico 5.12.

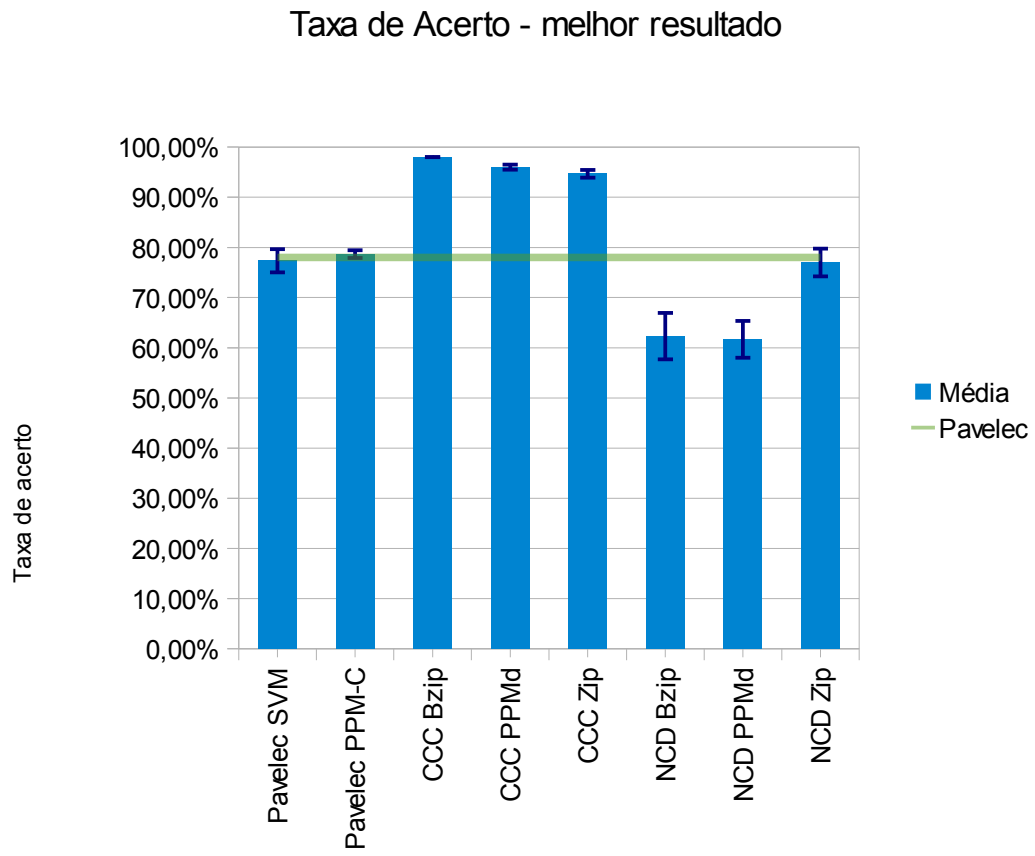


Figura 5.12: Taxa de acerto com escolha pelo melhor resultado

A explicação para a piora sensível dos resultados do método NCD é a busca pela distância relativa de similaridade entre os documentos. Ao serem concatenados os documentos de treinamento são aproximadamente 5 vezes maiores que os documentos de teste, gerando uma grande diferença entre eles. Quando os documentos de treinamento estão separados, e se utiliza o mecanismo de escolha de melhor resultado, basta que haja um único documento de treinamento que apresente um bom resultado que a atribuição será feita a este autor. Com os documentos concatenados, a distância NCD será maior e o bom resultado de um arquivo de treinamento estará prejudicado pela diferença de tamanho entre o documento questionado e o documento de treinamento.

Nesta abordagem, com a concatenação dos documentos, não é possível utilizar como mecanismo de escolha o processo de votação ou média de resultados pois é obtido apenas um resultado por autor para documento de teste.

Cabe destacar que apesar de utilizarem a concatenação de arquivos e usarem compressores estatísticos, as abordagens utilizadas por Pavelec (Pavelec, D. F., 2007) no seu resultado "PPM-C" e o do método CCC com compressor PPMD são bastante diferentes. A primeira considera a taxa de compressão obtida entre o documento de treinamento e o documento de teste e utiliza uma variação do compressor PPM enquanto a abordagem "CCC PPMD" utiliza a diferença de tamanho de compressão entre a concatenação dos arquivos e o arquivo de treinamento e utiliza outra variação do compressor PPM.

5.3.2. Autores P - Y

Neste outro subgrupo de documentos da base de dados "Pavelec" foram executados os mesmos testes do tópico anterior.

Os resultados obtidos estão representados na tabela 5.15 e no gráfico 5.13.

Tabela 5.15: Comparativo de desempenho de escolha pelo melhor resultado

				CCC			NCD	
Teste	**SVM**	**PPM-C**	**Bzip**	**PPMd**	**Zip**	**Bzip**	**PPMd**	**Zip**
1-5	87,00%	89,00%	94,00%	95,00%	98,00%	45,00%	33,00%	42,00%
6-10	88,00%	91,00%	98,00%	99,00%	98,00%	71,00%	72,00%	56,00%
11-15	91,00%	93,00%	95,00%	93,00%	95,00%	39,00%	36,00%	50,00%
Média	88,67%	91,00%	95,67%	95,67%	97,00%	51,67%	47,00%	49,33%

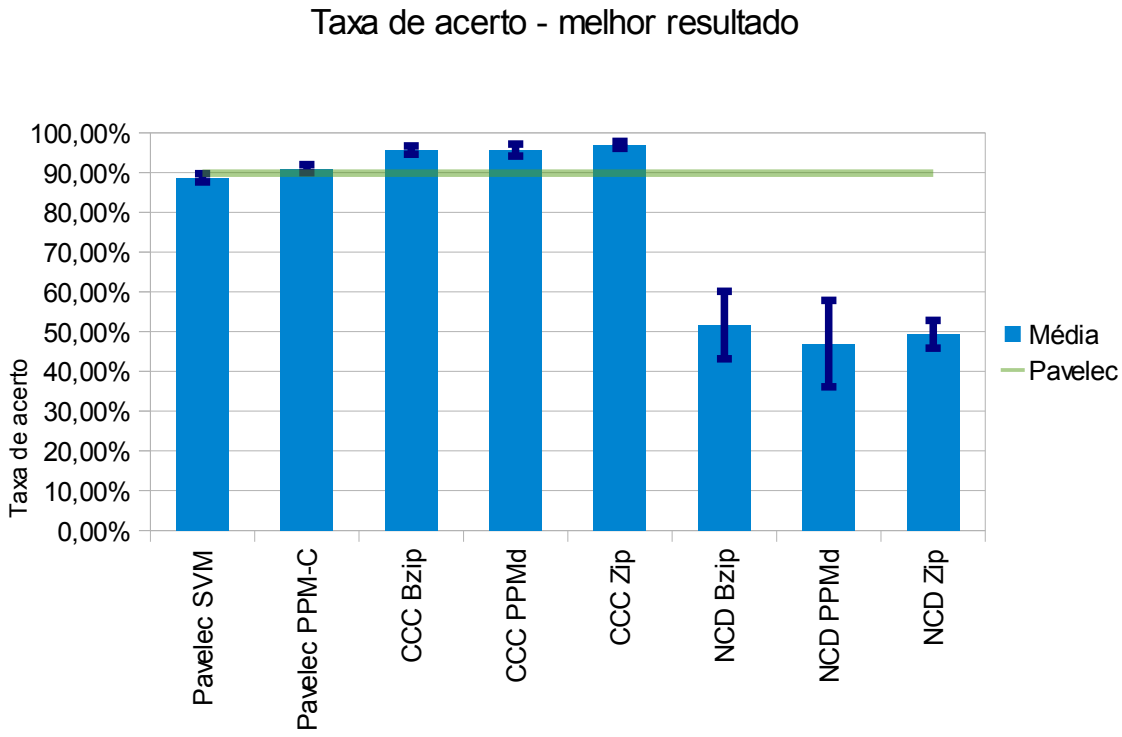


Figura 5.13: Taxa de acerto com escolha pelo melhor resultado

Verifica-se, novamente, que o desempenho do método NCD é bastante inferior aos resultados obtidos por Pavelec (Pavelec, D. F., 2007) e pelo método CCC, enquanto o método CCC apresenta um resultado superior aos de (Pavelec, D. F., 2007) de aproximadamente 5 pontos percentuais, em média. O resultado do método é bastante semelhante para qualquer um dos 3 compressores utilizados.

5.3.3. Autores A - Y

Para o terceiro subgrupos de documentos da base de dados são considerados todos os 20 autores utilizados nos testes executados logo acima.

A tabela 5.16 e o gráfico 5.14 apresentam os resultados obtidos.

Tabela 5.16: Comparativo de desempenho de escolha pelo melhor resultado

	Pavelec		CCC			NCD		
Teste	**SVM**	**PPM-C**	**Bzip**	**PPMd**	**Zip**	**Bzip**	**PPMd**	**Zip**
1-5	83,00%	84,00%	95,00%	95,00%	95,50%	49,00%	43,50%	52,50%
6-10	83,00%	83,00%	95,50%	94,00%	94,00%	53,50%	52,00%	64,50%
11-15	85,00%	86,00%	95,00%	92,50%	94,50%	45,50%	44,50%	61,00%
Média	83,67%	84,33%	95,17%	93,83%	94,67%	49,33%	46,67%	59,33%

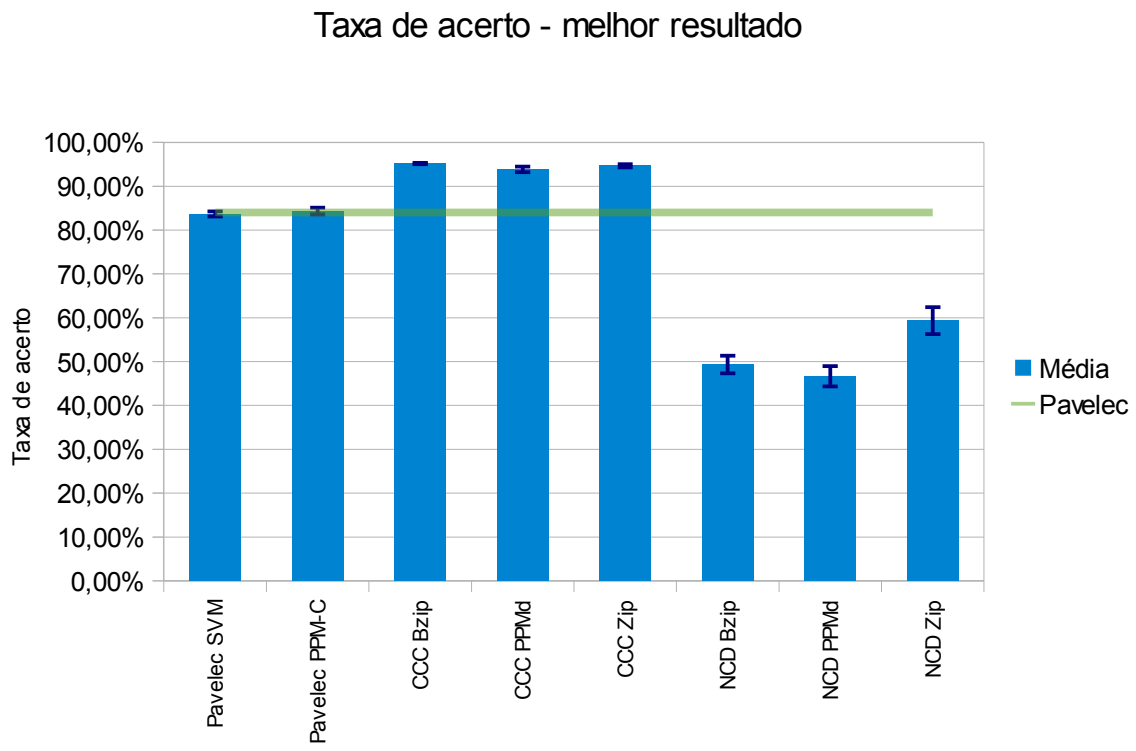


Figura 5.14: Taxa de acerto com escolha pelo melhor resultado

Nota-se que com o aumento da quantidade de autores possíveis para a atribuição de autoria dos documentos de teste há uma variação muito pequena nos resultados obtidos para o método CCC. Estes resultados continuam sendo superiores aos obtidos por (Pavelec, D. F., 2007) e são resultados bastante estáveis, com um desvio padrão pequeno.

Os resultados obtidos pelo método NCD continuam bastante inferiores aos obtidos pelo método CCC e por Pavelec (Pavelec, D. F., 2007).

5.3.4. Conclusões da base de dados Pavelec com documentos de treinamento concatenados

O processamento computacional requerido pelo método de concatenação é inferior ao necessário quando os documentos de treinamento são utilizados de maneira individual.

A equação (23) ilustra quantas compressões de documentos são efetuados quando os arquivos são considerados de maneira concatenada.

$$Número\ de\ compressões = nAutores + nTestes + (nAutores * nTestes) \quad (12)$$

sendo *nAutores* o número de autores possíveis para a atribuição de autoria e *nTestes* o número de documentos a serem testados.

Para a base de testes composta por 20 autores e 200 documentos de testes (10 documentos por autor), são realizadas 4220 compressões de documentos no total. Há também o esforço computacional de concatenar os documentos de treinamento em um único documento, que é desprezível com o processamento necessário para a compressão dos documentos.

Quando os documentos de treinamento são considerados de maneira individual, a quantidade de compressões necessárias é expressa na equação (24).

$$Número\ de\ compressões = nAutores + nTestes + (nAutores * nTestes * nTreinamento) \quad (13)$$

sendo *nAutores* e *nTestes* definidos anteriormente e *nTreinamento* a quantidade de documentos de treinamento por autor.

Para a base de testes compostas por 20 autores, 200 documentos de testes e 5 documentos de treinamento por autor, são realizadas 20220 compressões de documentos no total.

O método CCC apresenta bons resultados, superiores aos de (Pavelec, D. F., 2007), quando os documentos de treinamento são concatenados. O método NCD, por sua vez, apresenta uma piora considerável em seu desempenho. O fato do tamanho dos arquivos de treinamento (concatenados) ser superior ao tamanho do documento testado faz com que a distância normalizada de compressão tenha um desempenho pior para a atribuição de autoria.

O melhor resultado médio, com concatenação de documentos de treinamento, é obtido pelo método CCC com o compressor BZIP. O melhor resultado médio, sem a concatenação de documentos de treinamento, é obtida pelo método NCD com o compressor ZIP. A comparação

entre o melhor resultado com a concatenação de arquivos de treinamento é apresentada no gráfico 5.15.

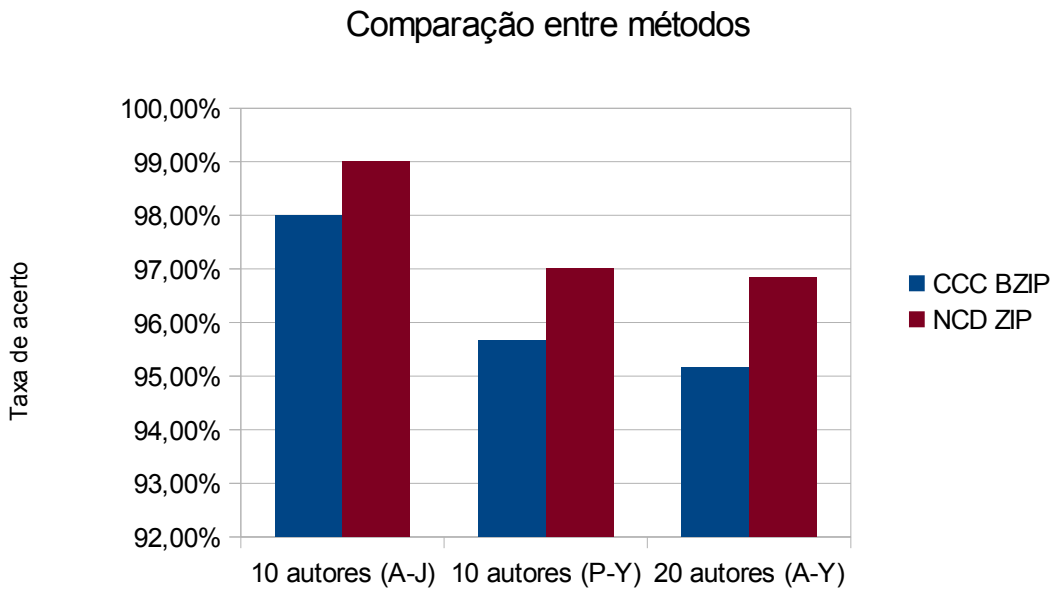


Figura 5.15: Comparação de resultados

Verifica-se que apesar dos resultados serem bastante próximos, com uma diferença aproximada de 1,5 pontos percentuais, a combinação "NCD ZIP" apresenta um resultado superior, o que pode justificar o seu uso quando os recursos computacionais disponíveis forem suficientes.

5.4. Base de dados Varela

A base de dados Varela (Varela, P. J. 2010) é formada por 3000 documentos, com 100 autores diferentes possuindo 30 documentos cada. Estes 100 autores são divididos conforme temas que escrevem, existindo 10 temas diferentes na base de dados.

O protocolo adotado por Varela, nos testes utilizando classificadores SVM para a atribuição de autoria, consiste na separação de 7 documentos de cada autor para compor a base de treinamento daquele autor e na utilização dos 23 documentos restantes como teste.

Como o autor separou os documentos de treinamento de maneira aleatória e não-repetitiva, não é possível estabelecer quais documentos foram separados. Desta forma,

também separamos 7 documentos de maneira aleatória para compor a base de treinamento e utilizamos os demais documentos para testes.

Foram realizados duas séries de experimentos com estas bases de documentos, detalhadas nos tópicos a seguir.

5.4.1. Documentos de treinamento separados

Assim como feito na base de dados Pavelec, os primeiros experimentos foram conduzidos considerando cada documento de treinamento individualmente, sem haver a concatenação dos mesmos para a geração de um modelo de treinamento.

A figura 5.16 ilustra o procedimento utilizado.

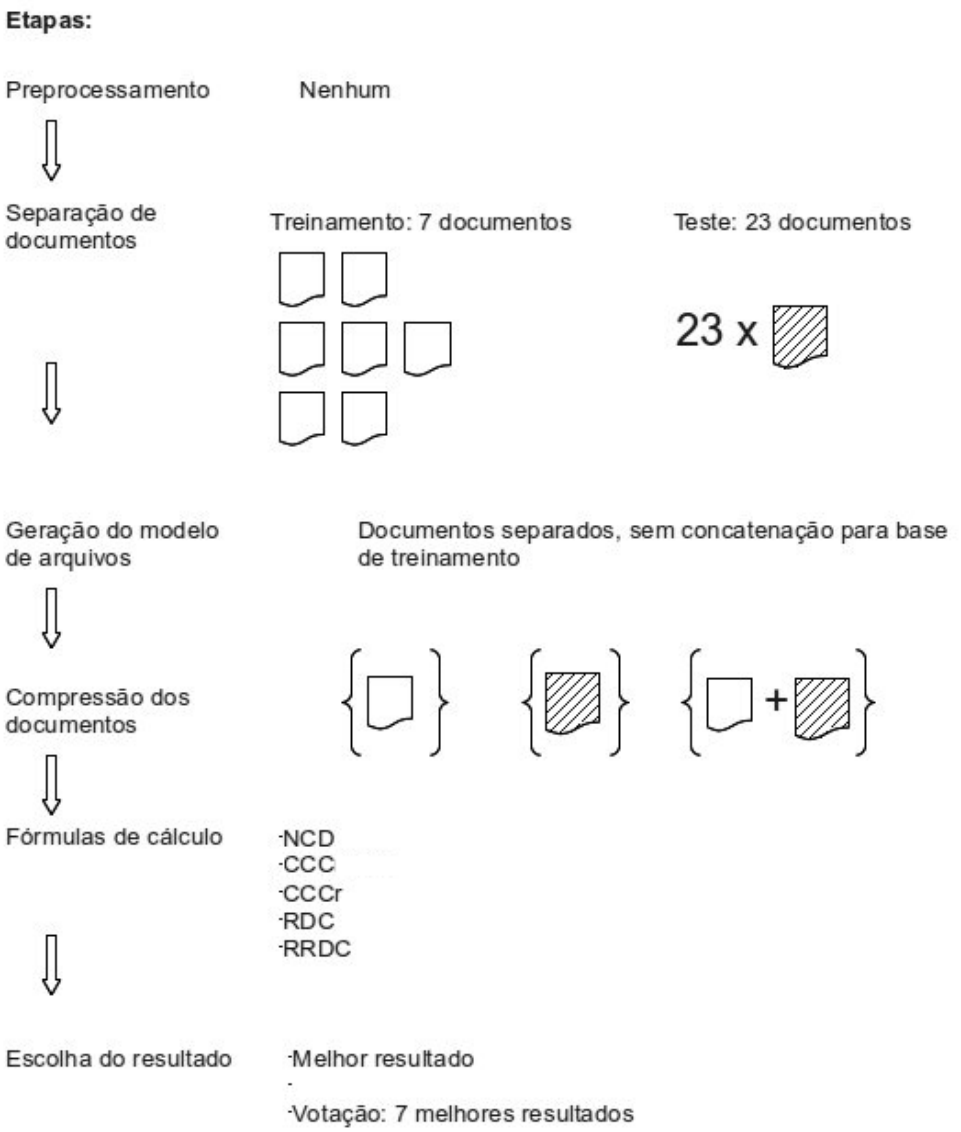


Figura 5.16: Procedimento de teste com documentos de treinamento individuais

Os demais procedimentos foram os mesmos adotados nos testes com as outras bases de dados: calculou-se o tamanho da compressão do arquivos de conhecimento e de teste e, em seguida, concatenou-se o documento de treinamento com o documento de teste e calculou-se o tamanho da compressão deste documento concatenado.

Como primeiro mecanismo de escolha adotou-se a escolha do melhor resultado. Os resultados obtidos, para cada uma dos temas em que os autores foram separados, é exibido na tabela 5.17.

Tabela 5.17: Comparativo de desempenho de escolha pelo melhor resultado

	Paulo Varela		BZIP	PPMD	ZIP	
Tema	todos	Selecionados				*Selecionados*
Assuntos Variados	70,70%	72,20%	79,57%	81,74%	**83,04%**	*83,04%*
Direito	72,20%	**74,40%**	63,91%	68,26%	65,65%	*68,26%*
Economia	64,80%	69,10%	77,83%	78,70%	**79,57%**	*79,57%*
Esportes	68,30%	69,60%	82,61%	85,65%	**87,39%**	*87,39%*
Gastronomia	**75,70%**	73,50%	44,78%	54,35%	53,04%	*54,35%*
Literatura	72,20%	**76,10%**	59,13%	66,96%	61,74%	*66,96%*
Política	68,70%	75,70%	81,74%	**83,91%**	83,04%	*83,91%*
Saúde	72,20%	**74,40%**	58,26%	61,30%	63,91%	*63,91%*
Tecnologia	73,90%	78,70%	74,78%	77,39%	**79,13%**	*79,13%*
Turismo	78,30%	81,70%	80,00%	82,17%	**83,04%**	*83,04%*
Média	71,70%	**74,54%**	70,26%	74,04%	73,96%	***74,96%***

São mostrados apenas os resultados do método NCD, pois os resultados obtidos pelo método CCC foram sempre inferiores, com uma taxa de acerto, em média, inferior em 12 pontos percentuais.

Verifica-se que em 4 temas o método proposto por Varela (Varela, P. J. 2010), de utilizar um classificador SVM com características estilométricas selecionadas a partir de palavras-função (tais como verbos ou pronomes) apresentou um desempenho superior, sendo que para três temas a escolha da categoria das palavras-função apresentou um resultado melhor e para um dos temas o uso de todas as categorias das palavras-função selecionadas apresentou o melhor resultado.

Nas outras 6 classes o desempenho do uso de compressores de dados com o método NCD apresentou um resultado melhor, sendo que na maioria dos casos o compressos ZIP apresentou o melhor resultado, e em apenas um caso o compressor PPMD teve resultado superior.

Apenas nas categorias tecnologia e turismo o compressor ZIP foi relevante para que o método testado neste tópico (NCD, documentos de treinamento separados, escolha pelo melhor resultado) apresentasse um desempenho superior ao obtido por Varela (Varela, P. J.

2010). Não foi possível verificar, entretanto, a razão desta superioridade quando utilizado o ZIP e inferioridade quando utilizados os compressores BZIP ou PPMD.

A média dos resultados obtidos indica que o método de (Varela, P. J. 2010), utilizando características selecionadas para cada uma das bases de dados, apresenta um desempenho superior ao método NCD com escolha do melhor resultado, se considerarmos apenas os resultados de cada compressor de dados individualmente. Se, no entanto, selecionarmos o melhor resultado dos compressores, veremos que a média de atribuições corretas pelo método NCD é superior.

O gráfico 5.17 ilustra este resultado, onde mostra-se apenas o resultado do compressor ZIP (que foi o que apresentou uma quantidade maior de temas onde a atribuição de autoria foi feita corretamente).

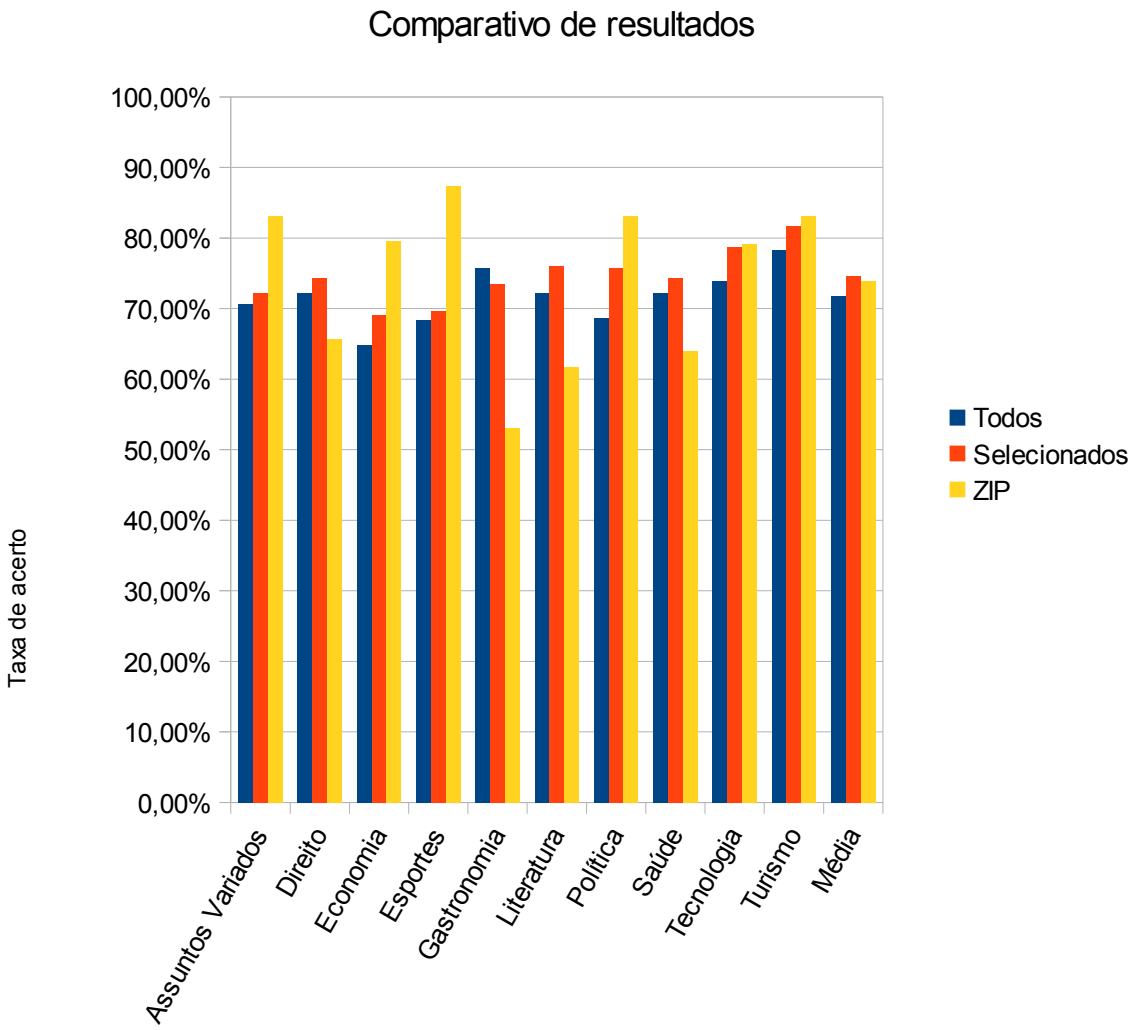


Figura 5.17: Taxa de acerto por temas e resultado médio

Apesar de cada tema possuir o resultado de 10 autores, o resultado não é comparável aos obtidos com a base de dados Pavelec. A quantidade de autores de cada tema e da base de Pavelec é a mesma mas a complexidade é diferente. Na base de dados Pavelec, os autores abordavam temas diversos, enquanto na base de dados Varela os temas abordados são os mesmos. Há uma maior probabilidade que os autores utilizem palavras ou expressões repetidas ao abordarem o mesmo tema, o que faz com que a base de dados Varela apresente uma maior complexidade para análise.

Apesar do desempenho médio inferior e de ter um resultado melhor na quantidade de temas onde a atribuição de autoria teve um desempenho superior, deve-se destacar que o

método que utiliza a compressão de documentos não depende de nenhum processamento ou seleção prévia de características, podendo ser utilizado em qualquer base de dados de documentos de texto de qualquer idioma.

Outro teste foi executado com a mesma base de dados e mesmo método de escolha para atribuição de autoria, mas considerando-se os documentos de todas as classes simultaneamente. Desta forma, para documento testado, haveriam 100 autores prováveis. O resultado obtido é apresentado na tabela 5.18.

Tema	BZIP	PPMD	ZIP
Assuntos Variados	72,61%	72,61%	**73,48%**
Direito	53,91%	**60,43%**	59,13%
Economia	63,91%	**64,78%**	63,91%
Esportes	80,87%	84,78%	**85,65%**
Gastronomia	36,09%	**46,09%**	41,30%
Literatura	42,17%	**49,57%**	47,39%
Política	69,13%	**73,48%**	73,04%
Saúde	47,83%	53,91%	**58,26%**
Tecnologia	70,00%	73,04%	**77,83%**
Turismo	56,96%	**68,26%**	**68,26%**
Total	59,35%	64,70%	**64,83%**

São apresentados os resultados por cada tema e o total geral. O uso do compressor ZIP permitiu que fosse feita a atribuição correta de autoria dos documentos em aproximadamente 65% dos testes. Ao analisarmos cada tema individualmente, verificamos que o compressor PPMD apresentou o melhor resultado em 5 temas e teve o mesmo resultado que o compressor ZIP em um 1 tema. O compressor BZIP não apresentou, para nenhum dos temas, resultado superior aos demais compressores.

É possível comparar o resultado de cada tema em função da quantidade de autores possíveis. Ao considerar que todos os autores da base de dados eram possíveis, a atribuição de autoria de cada documento é feita em relação a um total de 100 autores. Ao considerar-se apenas os autores por tema, apenas 10 autores eram possíveis para cada documento de teste. A dificuldade de atribuição quando todos os autores são considerados é maior porque há uma maior quantidade de autores. Este comparativo é mostrado na tabela 5.19.

Tabela 5.19: Comparativo de desempenho em função de autores possíveis

	Todos autores		Autores por tema	
Tema	PPMD	ZIP	PPMD	ZIP
Assuntos Variados	72,61%	**73,48%**	81,74%	**83,04%**
Direito	**60,43%**	59,13%	**68,26%**	65,65%
Economia	**64,78%**	63,91%	78,70%	**79,57%**
Esportes	84,78%	**85,65%**	85,65%	**87,39%**
Gastronomia	**46,09%**	41,30%	**54,35%**	53,04%
Literatura	**49,57%**	47,39%	**66,96%**	61,74%
Política	**73,48%**	73,04%	**83,91%**	83,04%
Saúde	53,91%	**58,26%**	61,30%	**63,91%**
Tecnologia	73,04%	**77,83%**	77,39%	**79,13%**
Turismo	**68,26%**	**68,26%**	82,17%	**83,04%**

Verifica-se que para todos os temas o aumento da quantidade de autores possíveis fez com que o resultado obtido diminuísse. Em alguns temas (esporte e tecnologia) a queda do desempenho foi menor, ficando aproximadamente 2 pontos percentuais inferior apesar de haverem 10 vezes mais autores possíveis. Em média, o desempenho foi inferior em 9 pontos percentuais quando a quantidade de autores possíveis aumentou para 100 autores.

Estes resultados apresentados não podem ser comparados com os resultados obtidos por Varela (Varela, P. J. 2010) porque a quantidade de autores possíveis é maior, aumentando a complexidade da tarefa de atribuição de autoria.

O segundo mecanismo de escolha para a atribuição de autoria é o de votação. Neste mecanismo, são selecionados *n* melhores resultados do método considerado e é feita uma votação, sendo a autoria atribuída ao autor mais votado. Como cada autor possui 7 documentos de treinamento, foram considerados 7 votos.

A tabela 5.20 apresenta os resultados por tema, ou seja, para cada tema foram considerados apenas os autores daquele tema, em um total de 10 autores possíveis por documento testado. Para permitir comparação com o teste anterior também é apresentado o resultado do mecanismo de escolha “melhor resultado”.

Tabela 5.20: Comparativo de desempenho de escolha por votação

	Varela		PPMD		ZIP		BZIP	
Tema	todos	Selecionados	Melhor	**Voto**	Melhor	**Voto**	Melhor	**Voto**
Assuntos Variados	70,70%	72,20%	81,74%	81,74%	83,04%	78,26%	79,57%	**84,78%**
Direito	72,20%	**74,40%**	68,26%	64,35%	65,65%	67,83%	63,91%	60,87%
Economia	64,80%	69,10%	78,70%	70,00%	79,57%	74,78%	77,83%	**75,65%**
Esportes	68,30%	69,60%	85,65%	**83,91%**	87,39%	83,04%	82,61%	80,43%
Gastronomia	**75,70%**	73,50%	54,35%	48,26%	53,04%	51,74%	44,78%	50,87%
Literatura	72,20%	**76,10%**	66,96%	59,13%	61,74%	57,83%	59,13%	61,30%
Política	68,70%	75,70%	83,91%	80,87%	83,04%	**84,35%**	81,74%	80,87%
Saúde	72,20%	**74,40%**	61,30%	63,48%	63,91%	66,09%	58,26%	60,87%
Tecnologia	73,90%	**78,70%**	77,39%	74,35%	79,13%	78,26%	74,78%	75,22%
Turismo	78,30%	**81,70%**	82,17%	78,26%	83,04%	81,30%	80,00%	77,39%
Média	71,70%	**74,54%**	74,04%	70,43%	73,96%	72,35%	70,26%	70,83%

O resultado do método CCC deixa de ser apresentada por ter tido resultados, em média, inferiores em 15 pontos percentuais aos obtidos com o método NCD.

Como pode ser observado, considerando-se apenas o mecanismo de escolha por votação, o método proposto por Varela (Varela, P. J. 2010) apresenta resultados melhores em 6 classes e apresenta a melhor média de resultados.

O método “NCD – votação” apresentou piora de resultado, superando os resultados de (Varela, P. J. 2010) apenas em 4 temas.

O compressor ZIP, que apresentava os melhores resultados, só possui um desempenho melhor em um único tema, com os compressores PPMD e BZIP passando a apresentar resultados melhores em 3 temas.

O mecanismo de escolha “votação” melhorou os resultados obtidos, em relação ao mecanismo de escolha “melhor resultado”, para o compressor PPMD (no tema Saúde), para o compressor ZIP (nos temas Direito, Política e Saúde) e para o compressor BZIP (Assuntos Variados, Gastronomia, Literatura, Saúde e Tecnologia)

5.4.2. Quantidade de documentos de treinamento

Para verificar se a quantidade de documentos de treinamento de cada autor influencia na taxa de acerto da atribuição de autoria, foram feitos novos testes variando-se o número de documentos de treinamento disponíveis por autor.

Primeiramente, foram selecionados dois temas (Economia e Saúde) para serem testados com o compressor ZIP. O compressor ZIP foi escolhido pelo seu desempenho no primeiro teste (onde a escolha é feita pelo melhor resultado obtido), onde superou os demais compressores na atribuição de autoria. O tema Economia foi escolhido por ter sido um tema que apresentou melhores resultados para a atribuição de autoria pelo métodos de escolha do melhor resultado que o método de votação para os três compressores utilizados. O tema Saúde foi escolhido por ter sido o tema que apresentou melhores resultados do método de escolha votação do que o método do melhor resultado.

A base de dados Varela é composta por 30 documentos por autor. Ao início de sua utilização, quando feita a geração de um código para cada tema e nome de autor, foi também atribuído um número ao documento, de maneira sequencial. Desta forma, os documentos de um determinado autor, por exemplo, são identificados por seu arquivo possuir um nome no padrão Xa03.txt, indicando que trata-se do tema Saúde, do autor "a" dentro do tema Saúde, e que é o terceiro arquivo considerado para o autor.

Para todos os autores foram utilizados os documentos de número 01 a 07 para o treinamento, restando os documentos 08 a 30 como documentos de teste. Para a diminuição dos documentos da base de treinamento, os documentos 08 a 30 foram mantidos como documentos de teste e foi-se retirando sempre o último documento de treinamento. Então, no teste com 6 documentos de treinamento, foram utilizados para treinamento os documentos 01 a 06; no teste com 5 documentos, os documentos 01 a 05, e assim por diante.

Foram feitos testes para a atribuição considerando como possíveis apenas os autores do tema e considerando todos os autores, utilizando-se o compressor ZIP. Na tabela 5.21 é mostrado o resultado para o tema Economia e autores possíveis apenas os do tema Economia (ou seja, apenas os 10 autores do tema. No gráfico 5.18 é apresentado o resultado deste mesmo teste.

Tabela 5.21: Comparativo de desempenho em função da quantidade de documentos de treinamento – apenas autores do tema

	Quantidade de documentos de treinamento					
	7	6	5	4	3	2
Melhor	79,57%	76,96%	76,09%	74,35%	72,61%	68,70%
Voto	74,78%	73,91%	73,91%	72,17%	67,39%	43,48%

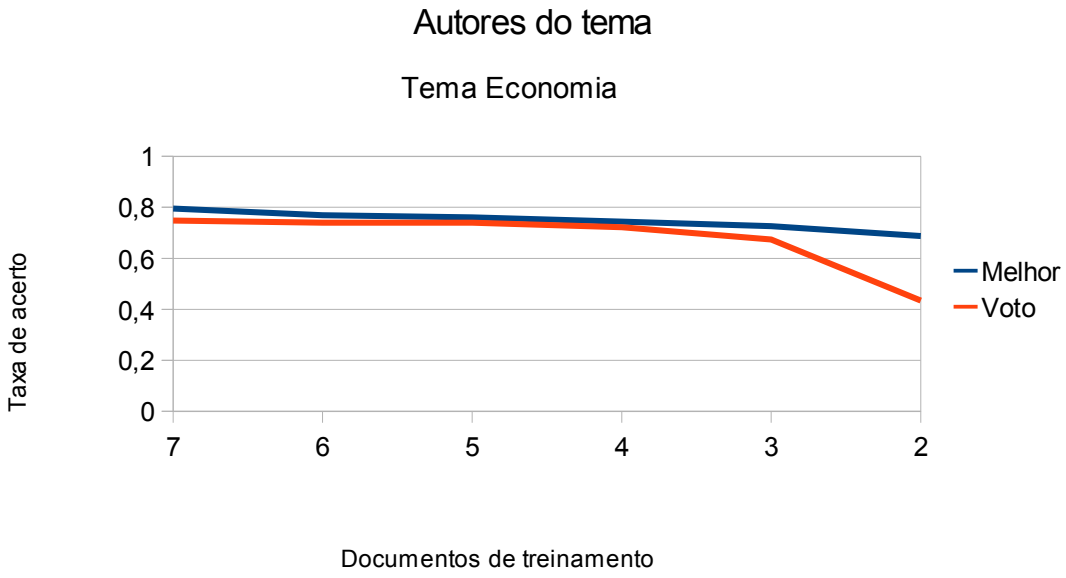


Figura 5.18: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor ZIP

Na tabela 5.22 e no gráfico 5.19 são apresentados os resultados para o teste efetuado no tema Economia mas com possibilidade de atribuição de autoria a qualquer um dos 100 autores existentes na base de documentos.

Tabela 5.22: Comparativo de desempenho em função da quantidade de documentos de treinamento – todos autores

	Todos					
	7	6	5	4	3	2
Melhor	63.91%	60.00%	61.30%	60.00%	56.52%	51.74%
Voto	63.91%	58.26%	56.96%	56.09%	45.65%	22.61%

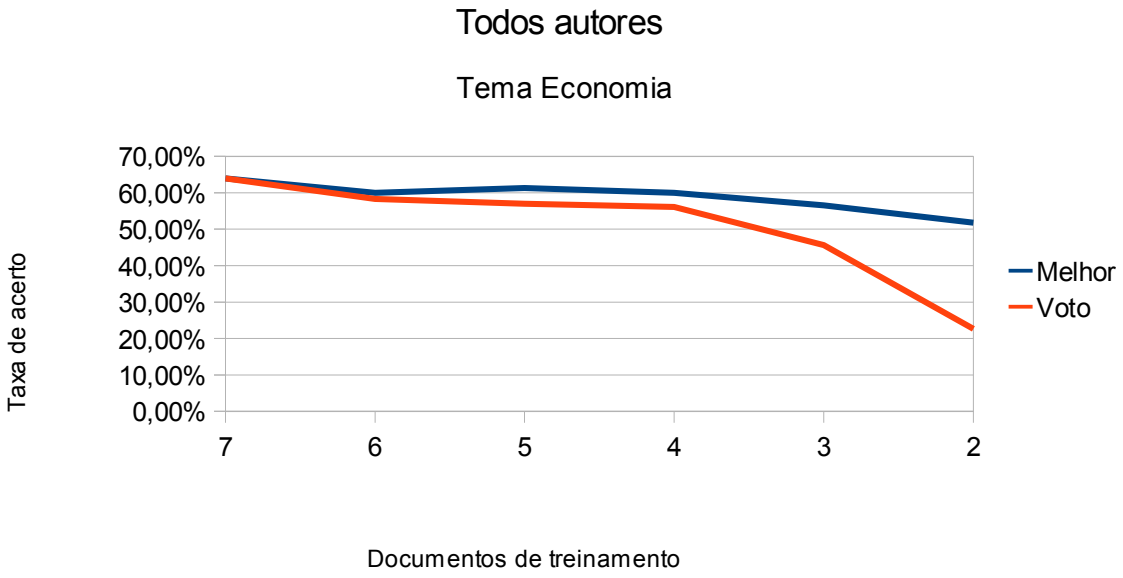


Figura 5.19: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor ZIP

Como pode ser observado, houve uma diminuição no desempenho do compressor ZIP, nos documentos do tema Economia, com a diminuição da quantidade de documentos de treinamento.

Foi realizado um novo teste utilizando o compressor ZIP e os documentos do tema Saúde. Os resultados obtidos estão representados na tabela 5.23 e no gráfico 5.20 para a atribuição de autoria feita tendo como autores possíveis apenas os do tema (10 autores possíveis).

Tabela 5.23: Comparativo de desempenho em função da quantidade de documentos de treinamento – apenas autores do tema

	Documentos de treinamento					
	7	6	5	4	3	2
Melhor	63.91%	61.74%	60.43%	63.04%	56.96%	52.61%
Voto	66.09%	67.83%	66.09%	61.30%	40.00%	22.17%

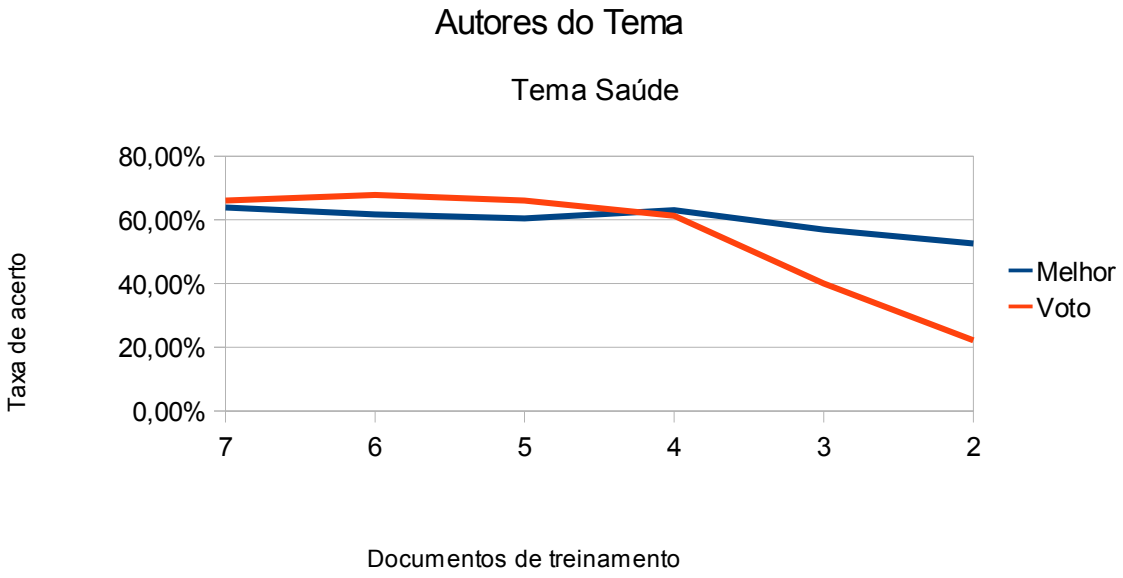


Figura 5.20: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor ZIP

Na tabela 5.24 e no gráfico 5.21 são apresentados os resultados para o teste efetuado no tema Saúde mas com possibilidade de atribuição de autoria a qualquer um dos 100 autores existentes na base de documentos.

Tabela 5.24: Comparativo de desempenho em função da quantidade de documentos de treinamento – todos autores

	Documentos de treinamento					
	7	6	5	4	3	2
Melhor	58.26%	55.65%	51.74%	51.74%	45.22%	45.65%
Voto	55.65%	54.78%	53.48%	48.70%	31.30%	15.22%

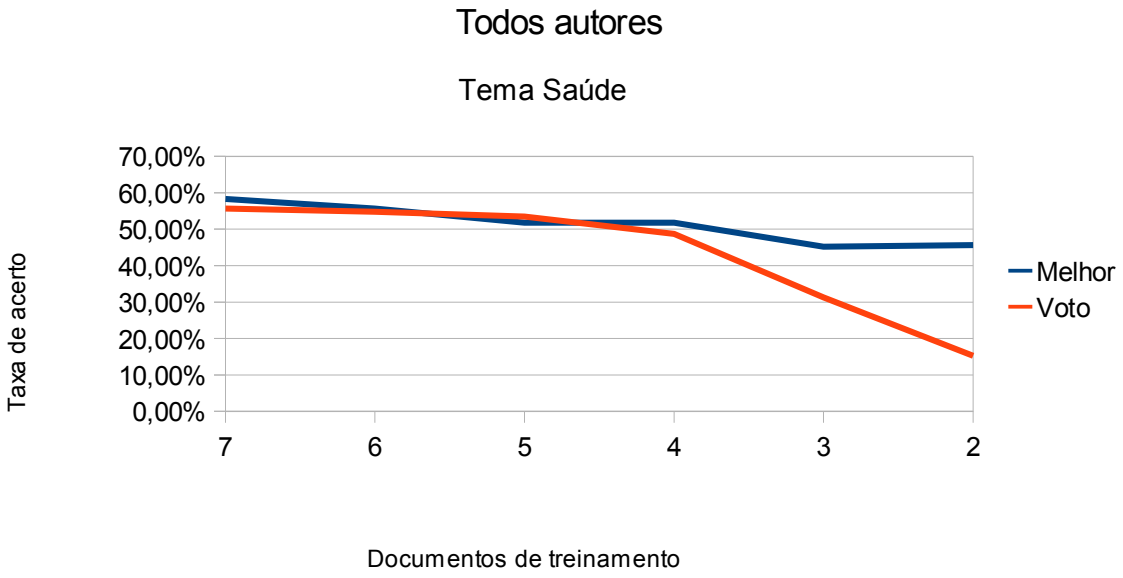


Figura 5.21: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor ZIP

É possível observar que, em geral, a diminuição da quantidade de documentos de treinamento disponíveis por autor diminui o desempenho da atribuição de autoria independentemente da quantidade de autores possíveis considerados. Em alguns pontos específicos é possível observar que a diminuição da quantidade de documentos disponíveis não afetou o desempenho, ou mesmo tornou o desempenho melhor. Isto ocorre, provavelmente, devido à retirada de algum documento que estivesse agindo como ruído, prejudicando a escolha feita.

Por exemplo, conforme pode ser visto na tabela 5.24 e no gráfico 5.21 acima, os resultados de votação apresentavam um desempenho melhor que a escolha do melhor resultado no início. Com a retirada do documento 05 da base de treinamento de cada autor, o resultado de votação deixou de ter o melhor desempenho, e a escolha pelo melhor resultado passou a apresentar uma maior taxa de acerto na atribuição de autoria. Possivelmente isto se deve ao fato que o documento 05 de um autor estava apresentando uma distância NCD menor em relação aos documentos de teste que os documentos do próprio autor. Como a escolha do melhor valor é sempre pelo resultado que apresentar a menor distância NCD, este arquivo servia como ruído a constantemente prejudicar a atribuição de autoria. Ao ser retirado da base

de treinamento, os documentos do verdadeiro autor passaram a ser mais relevantes e, assim, permitiram uma atribuição correta.

Não é possível, entretanto, deixar de considerar o resultado com o melhor valor para a atribuição de autoria. Por exemplo, na mesma base de dados e no mesmo tema Saúde, a retirada do melhor resultado dos testes faz com que o resultado diminua de 63,91% para 57,83% (quando o critério de escolha é o melhor resultado) e de 66,09% para 43,91% (quando o critério de escolha é a votação), o que indica que o melhor valor desempenha um papel importante para a atribuição de autoria e não pode ser descartado.

Para os compressores PPMD e BZIP foram efetuados os mesmos testes, reduzindo-se a quantidade de documentos de treinamento. Apresentamos apenas os gráficos obtidos, pois os mesmos já permitem uma análise do resultado obtido.

Nos gráfico 5.22 e 5.23 temos o desempenho do compressor PPMD no tema Economia, respectivamente considerando como possíveis apenas os autores do tema ou todos os autores.

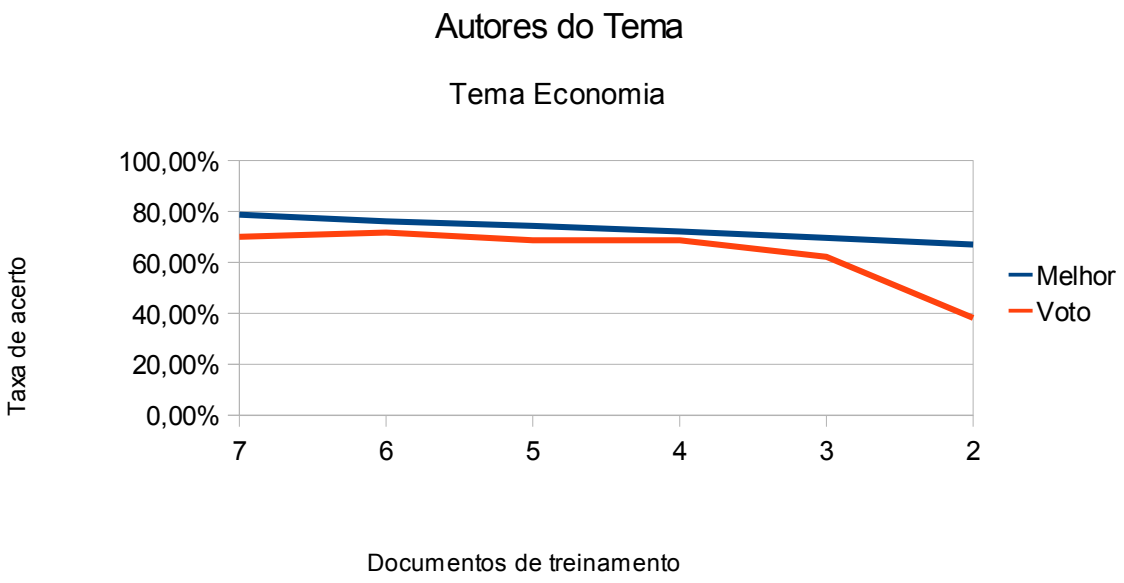


Figura 5.22: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor PPMD

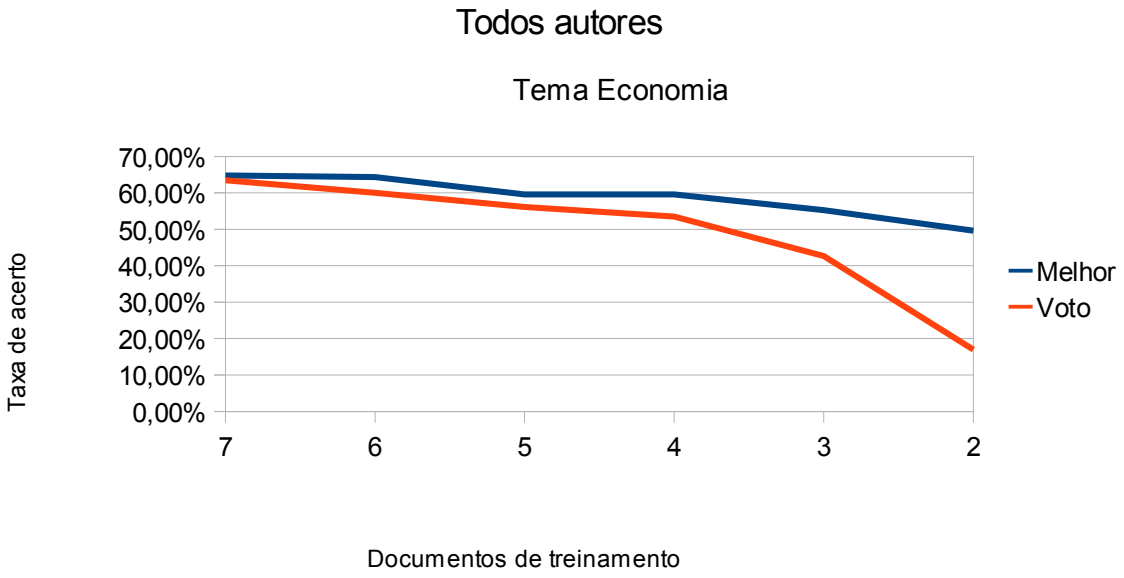


Figura 5.23: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor PPMD

Nos gráfico 5.24 e 5.25 temos o desempenho do compressor PPMD no tema Saúde, respectivamente considerando como possíveis apenas os autores do tema ou todos os autores.

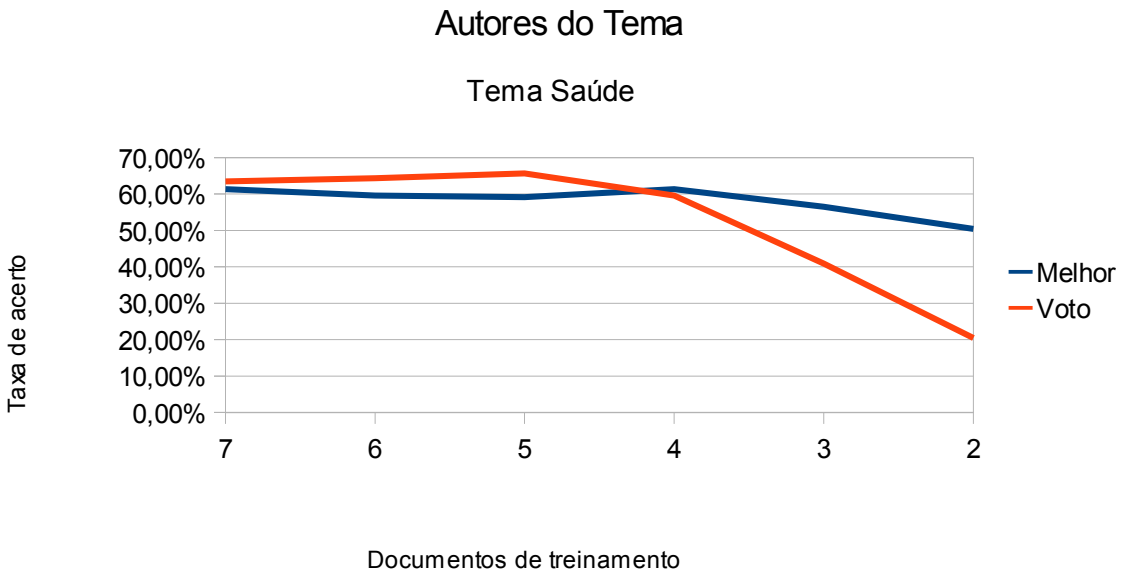


Figura 5.24: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor PPMD

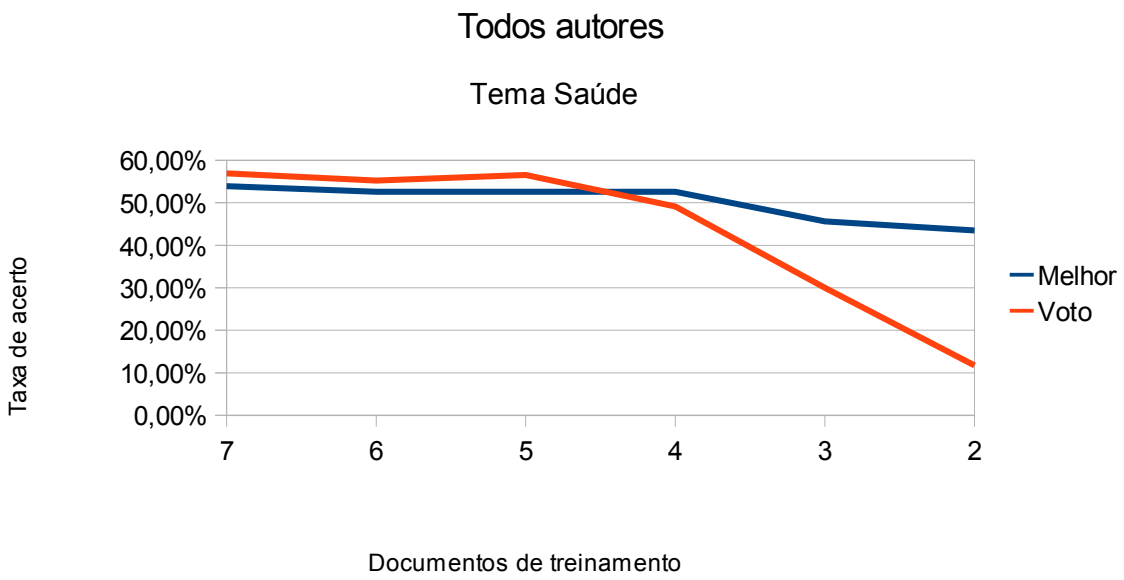


Figura 5.25: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor PPMD

Nos gráfico 5.26 e 5.27 temos o desempenho do compressor BZIP no tema Economia, respectivamente considerando como possíveis apenas os autores do tema ou todos os autores.

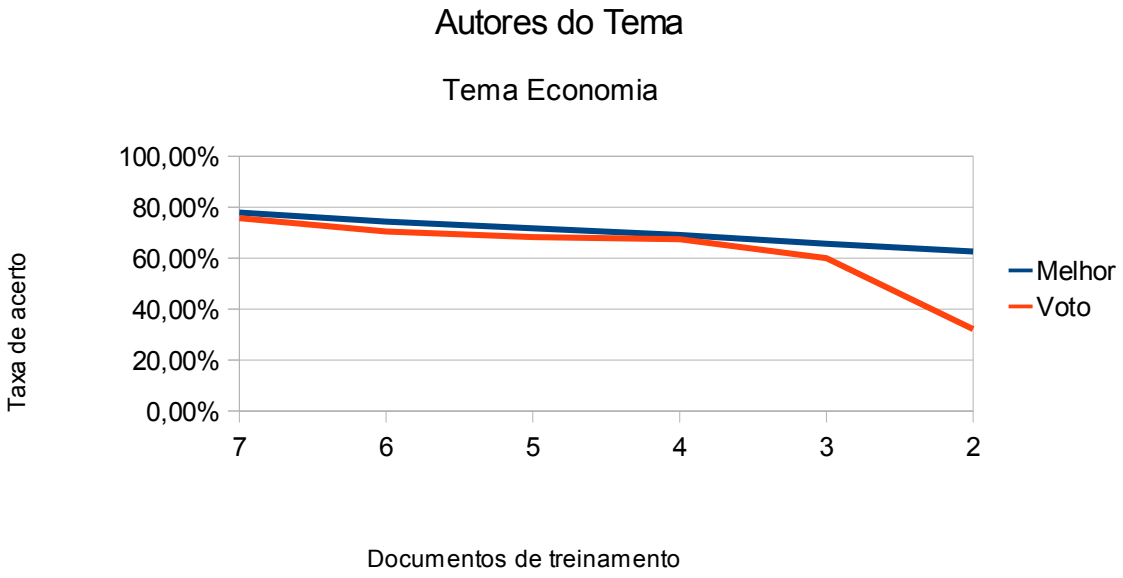


Figura 5.26: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor BZIP

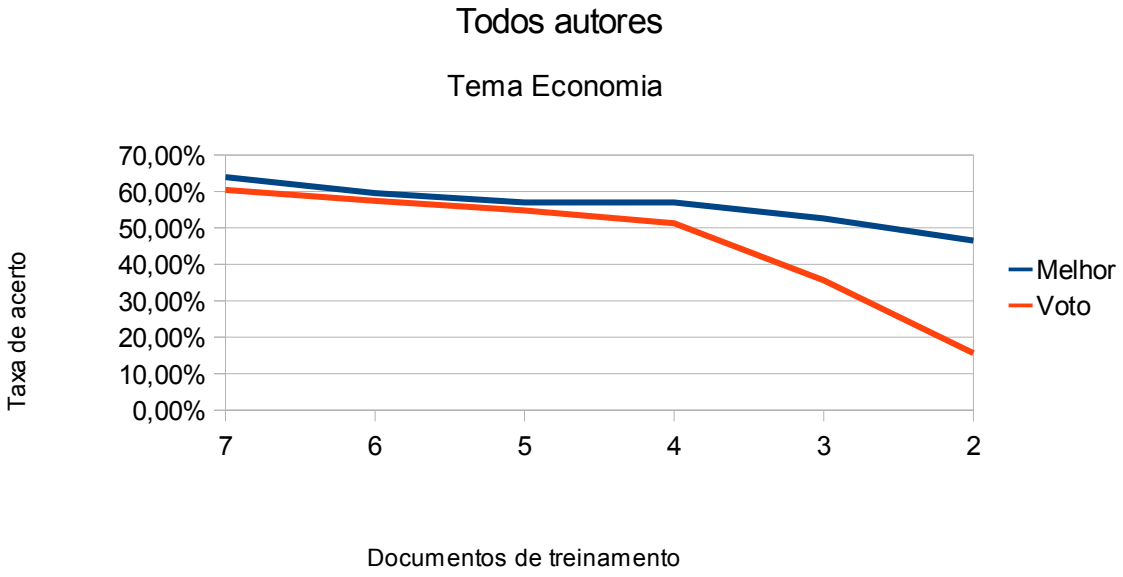


Figura 5.27: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor BZIP

Nos gráfico 5.28 e 5.29 temos o desempenho do compressor BZIP no tema Saúde, respectivamente considerando como possíveis apenas os autores do tema ou todos os autores.

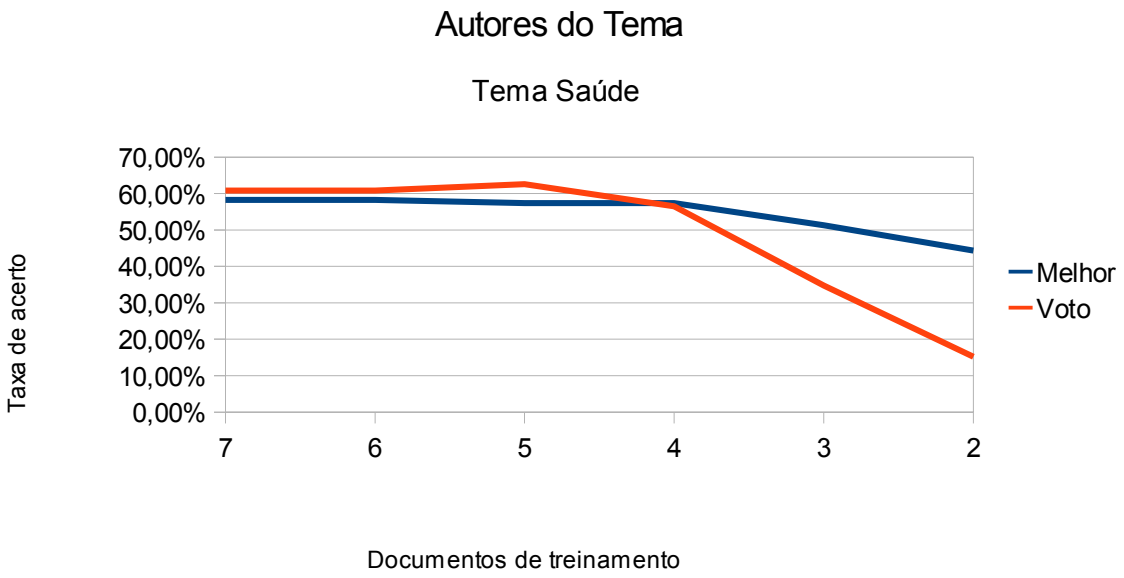


Figura 5.28: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor BZIP

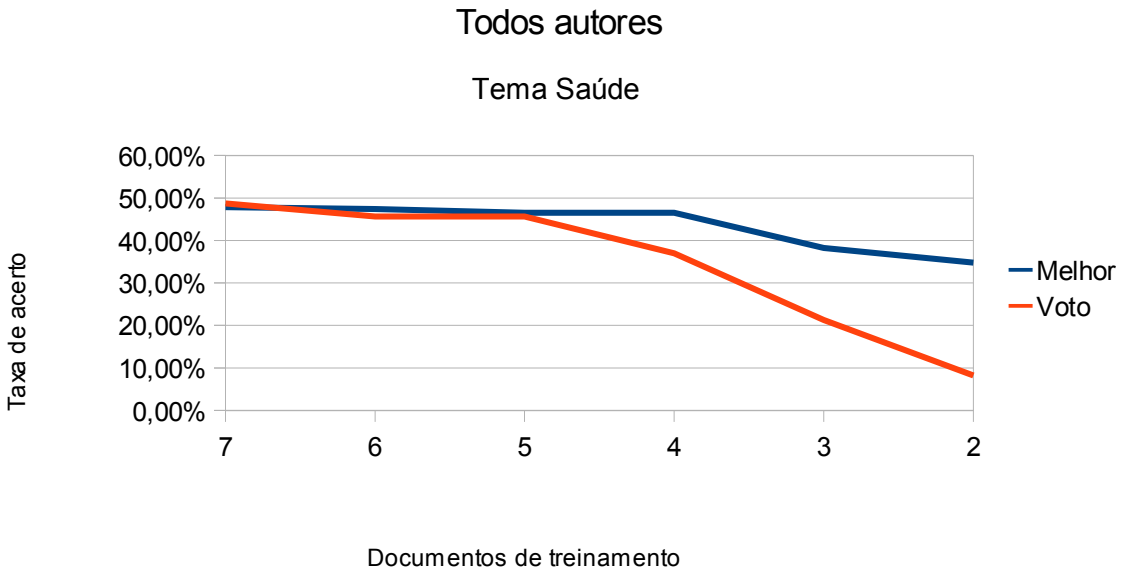


Figura 5.29: Taxa de acerto com diferentes quantidades de documentos de treinamento, compressor BZIP

Nota-se que há uma piora no desempenho de atribuição de autoria quando a quantidade de documentos de treinamento de cada autor é reduzida.

No tema Saúde, onde haviam sido obtidos melhores resultados através da escolha por votação, observa-se que há uma queda acentuada nos resultados obtidos por este mecanismo de escolha quando a quantidade de documentos de treinamento é reduzida. Em média, o mecanismo de escolha por votação apresenta uma queda suave no desempenho com até 5 documentos de treinamento por autor, havendo uma queda de desempenho abrupta a partir da redução desta quantidade de documentos de treinamento para 4 documentos ou menos.

5.4.3. Documentos de treinamento concatenados

Como na base de dados Pavelec, foram realizados testes com uma configuração da base de dados onde os documentos de treinamento são concatenados antes de serem utilizados pelo compressor de dados para a medida de distância ou de similaridade entre os documentos.

A figura 5.30 ilustra o protocolo utilizado para estes testes.

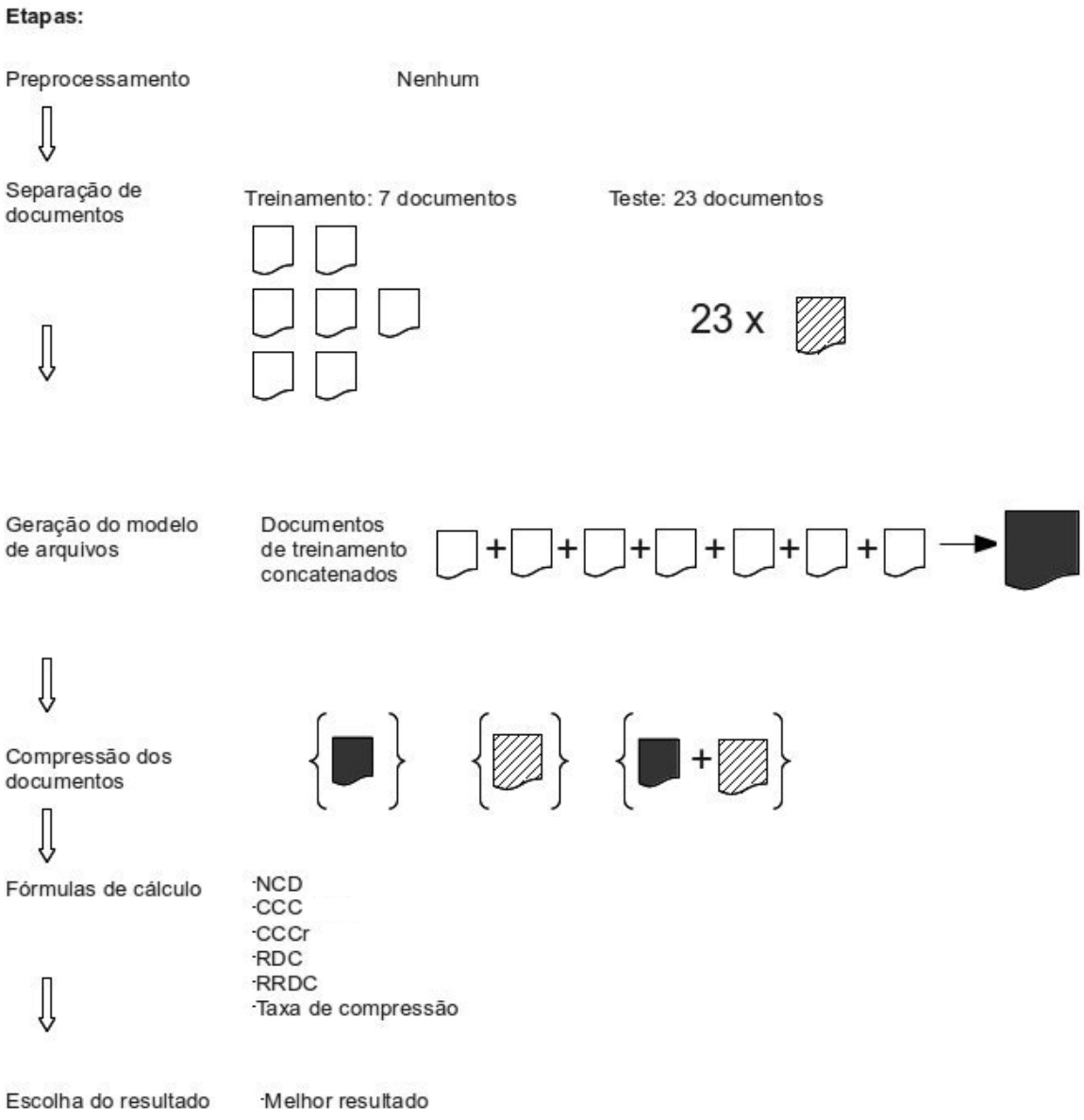


Figura 5.30: Procedimento de teste com documentos de treinamento concatenados

O mecanismo de escolha para atribuição é o do melhor resultado, pois não existem resultados diversos por autor que possam ser analisados por algum método estatístico (como a média) ou por votação.

Os resultados obtidos para o compressor ZIP é exibido na tabela 5.25. São mostrados os resultados obtidos quando todos os autores eram considerados como prováveis (em um total de 100 autores) e os resultados obtidos em cada tema, quando se considerou que apenas

os autores de cada tema seriam os autores possíveis para os documentos de teste. São mostrados os resultados dos metódos CCC e NCD.

Tabela 5.25: Resultados obtidos com documentos de treinamento concatenados – compressor Zip

	Benedetto	NCD
Todos	59,96%	6,61%
Assuntos Variados	80,87%	27,83%
Direito	57,39%	16,96%
Economia	56,96%	10,00%
Esportes	77,39%	18,70%
Gastronomia	57,83%	20,43%
Literatura	54,78%	17,39%
Política	74,35%	33,91%
Saúde	62,17%	20,87%
Tecnologia	83,04%	28,70%
Turismo	75,65%	12,17%

As tabelas 5.26 e 5.27 mostram o resultado dos mesmos testes, respectivamente, com os compressores PPMD e BZIP.

Tabela 5.26: Resultados obtidos com documentos de treinamento concatenados – compressor PPMD

	CCC	NCD
Todos	60,00%	3,65%
Assuntos Variados	83,91%	20,43%
Direito	56,96%	16,09%
Economia	53,48%	10,00%
Esportes	78,70%	10,87%
Gastronomia	59,57%	20,00%
Literatura	52,61%	16,09%
Política	78,26%	28,26%
Saúde	52,17%	12,61%
Tecnologia	83,04%	18,26%
Turismo	76,96%	10,00%

Tabela 5.27: Resultados obtidos com documentos de treinamento concatenados – compressor Bzip

	CCC	NCD
Todos	52,52%	5,78%
Assuntos Variados	79,13%	25,65%
Direito	54,35%	16,96%
Economia	56,52%	10,00%
Esportes	80,43%	13,04%
Gastronomia	58,70%	20,87%
Literatura	53,48%	17,39%
Política	62,17%	33,04%
Saúde	55,65%	14,35%
Tecnologia	79,57%	22,17%

É possível observar que, como aconteceu nos experimentos com a base de dados Pavelec, o resultado do método NCD apresenta uma taxa de acerto de atribuição de autoria bastante inferior à obtida pelo método CCC, para qualquer compressor de dados que seja considerado. O gráfico 5.31 ilustra esta diferença para o compressor ZIP.

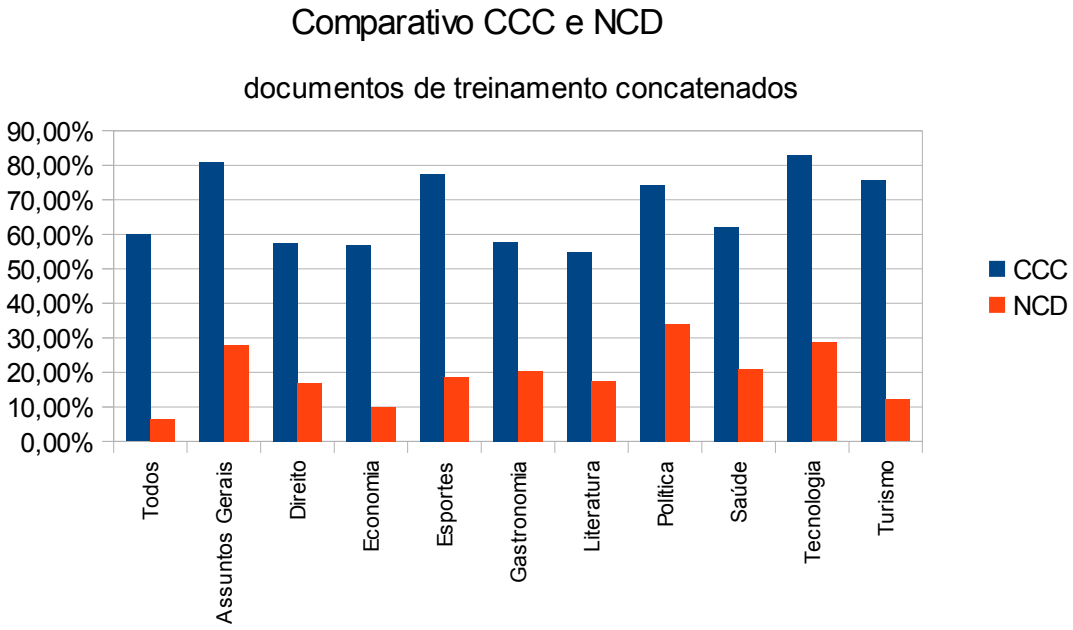


Figura 5.31: Comparativo de taxas de acerto por temas entre equações, compressor ZIP

Como observado no experimento conduzido com a base de Pavelec, a concatenação de documentos de treinamento prejudica o resultado do método NCD. A diferença do tamanho dos documentos de treinamento (concatenados) e dos documentos de teste faz com que a distância NCD torne-se muito elevada, deixando de ser útil para a tarefa de atribuição de autoria.

5.4.4. Conclusões dos testes com a base de dados Varela

A base de dados Varela é mais complexa, possuindo uma quantidade maior de temas abordados e de autores.

O método NCD apresenta resultados melhores quando os documentos de treinamento são considerados individualmente, fornecendo uma quantidade maior de medidas de distância entre o documento de teste e o conjunto de treinamento do autor. A concatenação de documentos de treinamento é prejudicial a esta abordagem, levando a resultados piores que os obtidos pelo método CCC.

A comparação entre os melhores resultados obtidos pela concatenação de documentos (método CCC, compressor ZIP) com os melhores resultados obtidos considerando-se os documentos de treinamento separados (método NCD, compressor ZIP) é apresentada na

tabela 5.28. São mostrados os resultados obtidos considerando-se todos os 100 autores como autores possíveis dos documentos e o resultado em cada tema, considerando-se que apenas os autores daquele tema eram possíveis.

Tabela 5.28: Comparativos de desempenho - documentos de treinamento individuais e concatenados

	Concatenados – CCC – ZIP	Separados – NCD – ZIP
Todos	59,96%	64,83%
Assuntos Variados	80,87%	83,04%
Direito	57,39%	65,65%
Economia	56,96%	79,57%
Esportes	77,39%	87,39%
Gastronomia	57,83%	53,04%
Literatura	54,78%	61,74%
Política	74,35%	83,04%
Saúde	62,17%	63,91%
Tecnologia	83,04%	79,13%
Turismo	75,65%	83,04%
Média dos temas	68,04%	73,96%

Verifica-se que o conjunto “Separados – NCD – ZIP” apresenta um resultado melhor quando são considerados todos os autores como prováveis e é superior em 8 dos 10 temas, quando considerados apenas os autores de cada tema. Seu resultado também é superior ao se considerar a média da taxa de acerto dos temas (excluindo-se o resultado que considerava todos os autores).

5.5. Influência da quantidade de autores prováveis

Ainda sendo utilizada a base de dados Varela, passou-se a analisar a influência da quantidade de autores prováveis na taxa de acerto do método NCD.

Pelos resultados anteriores na base de dados Varela, verificou-se que, quando são considerados todos os autores como prováveis, a método NCD apresentou o melhor

desempenho com os documentos de treinamento considerados individualmente com o compressor de dados ZIP.

Para verificar a influência da quantidade de autores prováveis no desempenho deste mecanismo de atribuição de autoria, realizou-se o seguinte teste: foram escolhidos 10 autores, sendo um autor de cada tema. Para cada autor foram escolhidos 10 entre os 23 documentos de testes utilizados, resultando em uma base de teste de 10 autores e 100 documentos. A seguir foram utilizados todos os 100 autores para o treinamento, cada autor com 7 documentos, em um total de 700 documentos de treinamento. Passou-se, então, a diminuir a quantidade de autores prováveis, retirando-se a cada vez 10 autores (um de cada tema) e todos os seus 7 documentos de treinamento.

Desta forma, foram realizados 10 testes: o primeiro teste contendo 100 autores possíveis, o segundo teste com 90 autores possíveis (9 em cada tema), o próximo teste com 80 autores possíveis (8 autores em cada tema), até ao final serem testados 10 autores possíveis, um de cada tema.

Cuidou-se para que o autor do documento de teste sempre estivesse na base de treinamento, evitando-se desta forma que algum documento de teste não possuísse resposta correta.

Os resultados obtidos para o mecanismo de escolha “melhor resultado” estão representados na tabela 5.29 e no gráfico 5.32. Na tabela, são apresentados os resultados do total de documentos testados e separados pelos temas.

Tabela 5.29: Comparativo do desempenho em função da quantidade de autores possíveis - escolha pelo melhor resultado – compressor ZIP

	Quantidade de autores									
	100	90	80	70	60	50	40	30	20	10
Todos autores	70%	71%	71%	71%	72%	73%	74%	75%	77%	81%
Assuntos Variados	30%	30%	30%	30%	30%	30%	30%	40%	60%	70%
Direito	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Economia	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Esportes	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Gastronomia	40%	40%	40%	40%	40%	40%	40%	40%	40%	50%
Literatura	80%	80%	80%	80%	80%	80%	80%	80%	80%	100%
Política	70%	70%	70%	70%	70%	70%	70%	70%	70%	70%
Saúde	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%
Tecnologia	70%	70%	70%	70%	70%	80%	80%	80%	80%	80%
Turismo	60%	70%	70%	70%	70%	70%	80%	80%	80%	80%

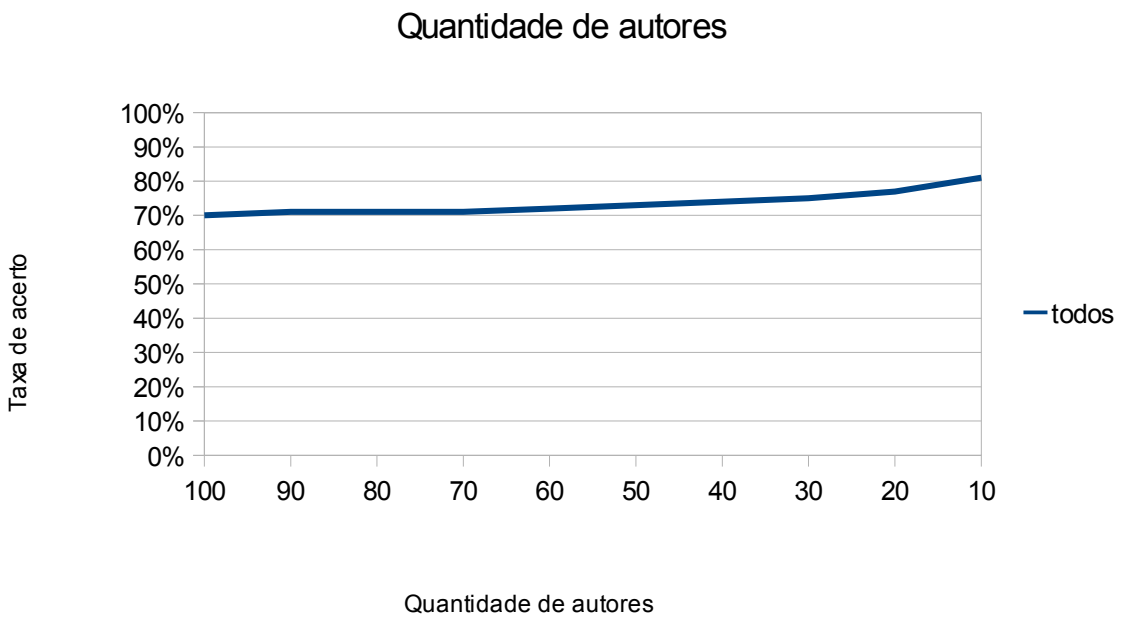


Figura 5.32: Taxa de acerto com diferentes quantidades de autores e escolha do melhor resultado – compressor ZIP

No mecanismo de escolha por votação e pela média os resultados são bastante semelhantes, motivo pelo qual exibimos apenas os resultados da escolha pela votação (tabela 5.30 e gráfico 5.33)

Tabela 5.30: Comparativo do desempenho em função da quantidade de autores possíveis - escolha por votação – compressor ZIP

	Quantidade de autores									
	100	90	80	70	60	50	40	30	20	10
todos	70%	72%	73%	73%	73%	75%	75%	77%	83%	83%
Assuntos Variados	30%	30%	30%	30%	30%	30%	30%	40%	80%	80%
Direito	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Economia	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Esportes	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Gastronomia	20%	20%	20%	20%	20%	30%	30%	30%	30%	30%
Literatura	80%	80%	80%	80%	80%	80%	80%	80%	90%	90%
Política	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
Saúde	70%	80%	80%	80%	80%	80%	80%	80%	80%	80%
Tecnologia	60%	60%	60%	60%	60%	70%	70%	80%	80%	80%
Turismo	50%	60%	70%	70%	70%	70%	70%	70%	80%	80%

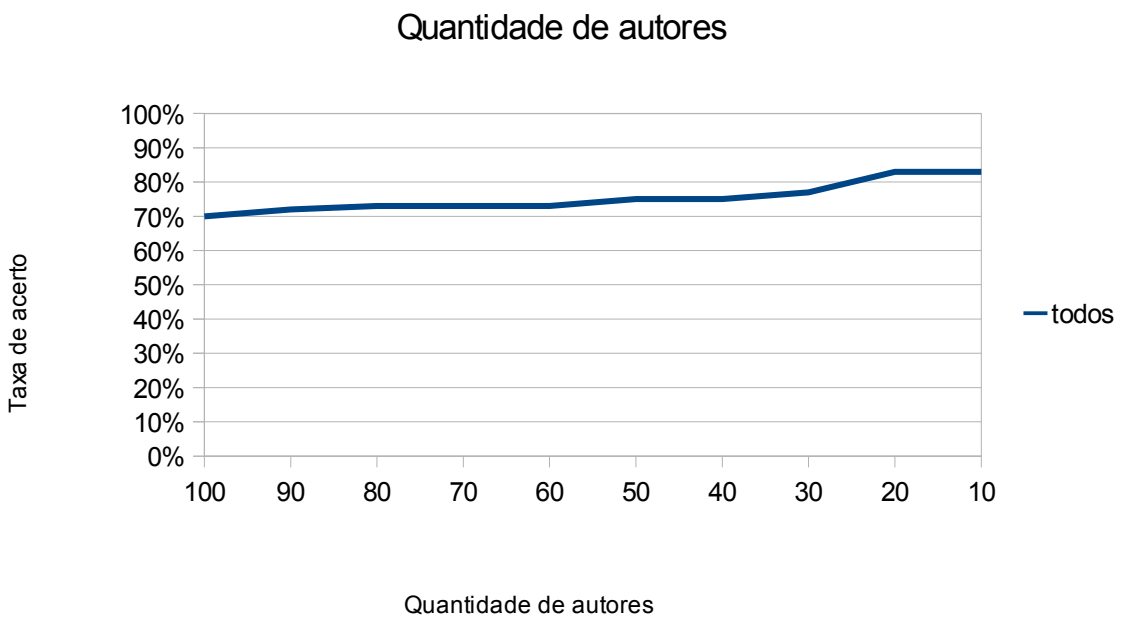


Figura 5.33: Taxa de acerto com diferentes quantidades de autores e escolha por votação – compressor ZIP

Como pode ser verificado, a quantidade de autores possíveis para a atribuição de autoria apresenta um impacto relativo na taxa de acerto de atribuições de autoria corretas. O

aumento na taxa de acerto varia discretamente entre 100 autores possíveis e 40 autores possíveis, passando a ter um impacto maior com a diminuição do número de autores.

Mesmo nos temas onde a atribuição de autoria teve um resultado pequeno quando eram possíveis 100 autores (por exemplo, nos temas Assuntos Variados, Gastronomia e Saúde na escolha pelo melhor resultado e nos temas Assuntos Variados e Gastronomia na escolha por votação), a diminuição da quantidade de autores prováveis foi significativa para aumentar o desempenho da escolha apenas no tema Assuntos Variados. No tema Gastronomia a diminuição de 100 autores possíveis para 10 autores fez com que apenas um documento a mais fosse atribuído corretamente. Sendo que, neste caso, apenas um autor era utilizado por tema, significando que todos os documentos do tema Gastronomia foram testados contra o seu verdadeiro autor no tema Gastronomia e contra autores de temas diversos.

Conforme foi possível verificar no tópico 5.4.1 - Documentos de treinamento separados, o tema Gastronomia apresentou os piores resultados quando eram considerados apenas os documentos do tema e quando eram considerados todos os autores possíveis, e a diminuição da quantidade de autores fez com que o desempenho deste mecanismo de escolha (método NCD, documentos concatenados e compressor ZIP) tivesse pouco ganho. Os temas que já possuíam um bom resultado com 100 autores possíveis mantiveram o bom resultado, havendo um pequeno ganho apenas quando a quantidade de autores possíveis foi reduzida para 40 ou 30 autores.

5.6. Matriz de confusão dos resultados obtidos

É possível verificar as dificuldades encontradas pelo método ao se analisar a matriz de confusão das atribuições de autoria. Nesta matriz são mostradas quais foram as atribuições feitas para cada um dos temas verificados, exibindo quantas atribuições foram feitas a autores de outros temas. A tabela 5.31 apresenta os resultados para a base de dados Varela, com compressor ZIP, documentos de treinamento considerados individualmente, método NCD, escolha do resultado feita pelo melhor resultado obtido, considerando-se como autores possíveis todos os 100 autores constantes da base de treinamento.

Tabela 5.31: Matriz de confusão entre temas

Temas	Assuntos Variados	Direito	Economia	Esportes	Gastronomia	Literatura	Política	Saúde	Tecnologia	Turismo
Assuntos Variados	86.96%	1.74%	3.04%	1.30%	0.00%	3.91%	1.30%	0.87%	0.43%	0.43%
Direito	5.22%	83.48%	3.91%	0.87%	2.17%	0.87%	0.43%	0.43%	1.74%	0.87%
Economia	4.78%	7.83%	74.35%	0.00%	1.30%	0.00%	6.09%	2.17%	2.17%	1.30%
Esportes	0.43%	0.43%	0.43%	96.96%	0.43%	0.00%	0.43%	0.43%	0.00%	0.43%
Gastronomia	1.74%	6.96%	6.96%	2.17%	65.65%	2.17%	0.00%	8.26%	4.35%	1.74%
Literatura	6.09%	9.57%	4.35%	2.17%	3.04%	57.39%	4.35%	4.78%	6.09%	2.17%
Política	2.17%	0.87%	7.39%	0.00%	0.00%	0.43%	83.91%	0.87%	4.35%	0.00%
Saúde	1.74%	4.35%	6.96%	0.87%	2.17%	0.00%	0.00%	81.74%	2.17%	0.00%
Tecnologia	0.43%	1.30%	3.48%	0.87%	0.00%	0.43%	0.00%	0.87%	91.74%	0.87%
Turismo	3.48%	2.17%	7.39%	1.74%	0.43%	0.87%	2.17%	3.48%	4.35%	73.91%

Observa-se que os temas “Gastronomia” e “Literatura” apresentaram os piores resultados, com menos de 70% dos documentos de teste sendo atribuídos a qualquer um dos autores do tema.

Os documentos do tema Gastronomia foram atribuídos incorretamente, principalmente, aos temas Direito (16 documentos), Economia (16 documentos) e Saúde (19 documentos). No trabalho de (Varela, P. J. 2010) os documentos do tema Gastronomia também apresentaram os maiores índices de confusão em relação a estes mesmos temas, mas em valores inferiores (4 documentos a Direito, 3 documentos a Economia e 5 documentos a Saúde).

Analisando-se os documentos de cada autor do tema Gastronomia não foi possível estabelecer um autor em específico que tenha apresentado confusão com o tema Direito ou Economia, havendo no máximo três documentos por autor que foram confundidos. Em relação ao tema Saúde foi possível verificar que foram os documentos da autora Marcia Daskal que apresentaram o maior número de confusões: 7 documentos da autora foram atribuídos incorretamente ao tema Saúde. Verificando-se o perfil da autora e o conteúdo dos documentos, constata-se que a autora é uma nutricionista e que seus documentos tratam, principalmente, de temas como alimentação saudável, alimentos e a prática de exercícios físicos e alimentos que auxiliem a prevenir ou combater problemas de saúde. Desta forma,

explica-se o baixo desempenho obtido pela atribuição de autoria com o uso de compressores em relação aos documentos desta autora: muitas das palavras ou frases utilizadas por ela em seus documentos são bastante semelhantes às encontradas em documentos de autores que escrevam sobre o tema Saúde.

Os documentos do tema Literatura foram atribuídos incorretamente, principalmente, aos temas Assuntos Variados (14 documentos), Tecnologia (14 documentos) e Direito (22 documentos). No trabalho de (Varela, P. J. 2010) os documentos deste tema apresentaram confusão com Assuntos Variados (8 documentos) e Política (6 documentos).

O autor Marcelo Coelho teve 8 documentos atribuídos incorretamente ao tema Direito, dos quais 4 foram confundidos com documentos do autor Jorge Alberto Araújo. Entretanto não há nenhuma característica aparente de palavras utilizadas ou conteúdos abordados que se sobressaiam a justificar a confusão.

É possível verificar quais foram os temas que mais receberam atribuições de documentos. O resultado é apresentado na tabela 5.32.

Tabela 5.32: Atribuições feitas a cada tema

	Documentos atribuídos	Atribuições oriundas de outros temas
Assuntos Variados	113.04%	26.09%
Direito	118.70%	35.22%
Economia	118.26%	43.91%
Esportes	106.96%	10.00%
Gastronomia	75.22%	9.57%
Literatura	66.09%	8.70%
Política	98.70%	14.78%
Saúde	103.91%	22.17%
Tecnologia	117.39%	25.65%
Turismo	81.74%	7.83%

Verifica-se que os temas que apresentaram o pior resultado de atribuições de documentos dentro do tema estão entre os temas que menos receberam atribuições errôneas a partir de outros temas.

Como no teste realizado no tópico 5.4.1 - Documentos de treinamento separados (p. 88) os temas Direito (R), Gastronomia (U), Literatura (V) e Saúde (X) apresentaram um desempenho inferior ao obtido por (Varela, P. J. 2010), verificou-se a matriz de confusão para cada um destes temas.

Primeiramente foi avaliado o tema Direito, apresentado na tabela 5.33.

Tabela 5.33: Matriz de Confusão - Direito

	BoleslauSliviany	CarlosZamithJunior	FabioTokars	FernandoCesarFaria	FredericoVasconcelos	IgorFonsecaRodrigues	JorgeAlbertoAraujo	MarialnesDolci	OscarIvanPrux	ReneArielDotti
BoleslauSliviany	23	0	0	0	0	0	0	0	0	0
CarlosZamithJunior	0	4	0	0	16	2	0	0	0	1
FabioTokars	0	0	21	0	0	0	0	0	2	0
FernandoCesarFaria	1	0	0	12	4	2	1	0	0	3
FredericoVasconcelos	0	4	0	1	10	1	2	2	1	2
IgorFonsecaRodrigues	1	0	0	1	6	9	2	1	0	3
JorgeAlbertoAraujo	0	1	0	0	3	1	16	0	1	1
MarialnesDolci	0	0	0	0	1	8	1	13	0	0
OscarIvanPrux	0	0	0	0	0	0	0	0	23	0
ReneArielDotti	0	0	2	0	0	1	0	0	0	20
Total de atribuições	25	9	23	14	40	24	22	16	27	30

Como pode ser verificado vários autores tiveram menos de 16 atribuições corretas. O autor Carlos Zamith Junior teve apenas 4 de seus documentos atribuídos corretamente, apresentando 16 documentos atribuídos erroneamente a Frederico Vasconcelos. Este último autor foi o que recebeu o maior número de atribuições (40 documentos), dos quais apenas 10 eram de sua autoria.

No trabalho de (Varela, P. J. 2010), neste tema, não houve nenhum autor que tenha recebido atribuição de autoria em todos os seus documentos (como pode ser verificado em relação a Boleslau Sliviany e Oscar Ivan Prux), mas nenhum dos autores havia apresentado menos que 16 atribuições corretas. No método utilizado no presente trabalho houveram 5 autores que tiveram menos que 16 atribuições corretas.

O tema Gastronomia também foi avaliado, apresentando o resultado que é mostrado na tabela 5.34.

Tabela 5.34: Matriz de confusão - Gastronomia

	AlessandraBlanco	AndreaKaufmann	CarlosBertolazzi	CilmaraCastilho	MarciaDaskal	MarthaStewart	NeideRigo	NigellaLawson	RicardoCastilho	TatianaDamberg
AlessandraBlanco	9	0	2	1	4	0	0	6	1	0
AndreaKaufmann	1	13	1	0	4	1	1	1	0	1
CarlosBertolazzi	0	1	5	2	1	1	1	1	6	5
CilmaraCastilho	0	0	0	15	0	2	0	1	2	3
MarciaDaskal	0	2	1	0	20	0	0	0	0	0
MarthaStewart	0	0	1	0	0	17	0	4	1	0
NeideRigo	3	1	0	3	6	2	3	2	2	1
NigellaLawson	1	0	0	0	0	0	0	22	0	0
RicardoCastilho	4	3	2	1	2	1	1	0	9	0
TatianaDamberg	1	1	0	5	0	2	0	0	5	9
Total de atribuições	19	21	12	27	37	26	6	37	26	19

Neste tema também verifica-se que apenas dois autores tiveram mais que 20 atribuições efetuadas corretamente. Os autores Carlos Bertolazzi e Neide Rigo tiveram mais documentos atribuídos erroneamente a um único outro autor do que a si mesmo. No caso dos autores Alessandra Blanco, Carlos Bertolazzi e Neide Rigo, no máximo 6 de seus documentos foram atribuídos erroneamente a um único outro autor, sendo as demais atribuições errôneas feitas a diversos autores.

No trabalho de (Varela, P. J. 2010) todos os autores possuíam, no mínimo, 17 atribuições corretas.

No tema Literatura também houveram poucos resultados com mais do que 20 atribuições corretas, conforme é mostrado na tabela 5.35.

Tabela 5.35: Matriz de confusão - Literatura

	ArnaldoJabor	CeciliaGiannetti	FernandoMonteiro	LauraMedioli	LuizBras	ManoelLobato	MarceloCoelho	NelsondeOliveira	PauloCoelho	SergioRodrigues
ArnaldoJabor	22	0	0	1	0	0	0	0	0	0
CeciliaGiannetti	0	5	0	0	3	0	5	1	2	7
FernandoMonteiro	0	1	9	1	1	0	3	1	3	4
LauraMedioli	0	1	0	20	0	0	0	0	2	0
LuizBras	0	3	1	0	11	0	1	1	2	4
ManoelLobato	0	0	0	0	0	23	0	0	0	0
MarceloCoelho	0	0	1	1	1	1	8	3	5	3
NelsondeOliveira	0	1	0	1	6	0	2	7	2	4
PauloCoelho	0	0	0	0	0	0	1	0	22	0
SergioRodrigues	0	0	0	0	0	1	2	5	0	15
Total de atribuições	22	11	11	24	22	25	22	18	38	37

A autora Ceilia Giannetti teve menos atribuições corretas do que a quantidade de documentos atribuídos erroneamente ao autor Sergio Rodrigues. O autor Paulo Coelho recebeu erroneamente a atribuição de autoria de documentos de diversos autores, sendo atribuído incorretamente mais vezes que os demais autores. No trabalho de (Varela, P. J. 2010) todos os autores possuíam, no mínimo, 17 atribuições corretas.

Por fim, o tema Saúde tem a sua matriz de confusão apresentada na tabela 5.36.

Tabela 5.36: Matriz de confusão - Saúde

	ClaudioLima	DrauzioVarela	FabioCesardosSanto	FernandaAranda	FlavioSettanni	JohnCookLane	LeandroPerché	LeoKahn	LilianeFerrari	LoirCarlosCosta
ClaudioLima	16	0	0	0	0	3	0	1	0	3
DrauzioVarela	0	17	0	0	0	0	0	5	0	1
FabioCesardosSantos	0	0	15	1	0	0	0	4	0	3
FernandaAranda	1	1	3	11	0	0	2	4	0	1
FlavioSettanni	0	1	1	0	13	0	3	0	0	5
JohnCookLane	0	0	2	0	0	10	5	0	2	4
LeandroPerché	1	1	2	3	0	0	12	1	1	2
LeoKahn	0	3	0	0	0	0	0	18	1	1
LilianeFerrari	0	0	0	0	1	1	6	0	14	1
LoirCarlosCosta	0	0	2	0	0	0	0	0	0	21
Total de atribuições	18	23	25	15	14	14	28	33	18	42

A quantidade de atribuições corretas foi sempre igual ou superior a 10 documentos e nenhum autor teve mais documentos atribuídos erroneamente a outro autor do que atribuídos a si próprio.

No trabalho de (Varela, P. J. 2010) houveram sempre, no mínimo, 17 atribuições corretas de documentos de cada autor, enquanto que é verificado que apenas os autores Drauzio Varela, Leo Kahn e Loir Carlos Costa tiveram resultado igual ou superior a 17 documentos, mostrando porque o desempenho do método NCD foi inferior na mesma base de dados.

Conclusão

A atribuição de autoria de documentos digitais apresenta um grande desafio para peritos e profissionais da área jurídica e de ciências da computação. A ausência de características como a grafoscopia ou elementos identificadores de equipamentos utilizados para a elaboração dos documentos (tais como endereços IP de computadores ou vestígios de programas utilizados para a elaboração dos documentos) representa uma majoração deste desafio, restando muitas vezes apenas características do estilo de escrita do autor para a atribuição de autoria a um documento questionado.

Diversas abordagens são propostas em pesquisas para a atribuição de autoria: o uso de características sintáticas, semânticas, medidas estatísticas extraídas a partir de frases, palavras e caracteres utilizados, riqueza vocabular, e outros elementos. Dentre estas propostas, uma abordagem que é pouco estudada em relação a documentos de língua portuguesa é o uso de características que possam ser extraídas com o uso de compressores de dados.

Uma das abordagens propostas para o uso de compressores de dados é a medida da distância normalizada de compressão de dois documentos. Através desta medida busca-se aproximar a complexidade de Kolmogorov, incomputável, a valores de medida de distância entre dois documentos, indicando o quanto eles são semelhantes.

O uso desta medida pode ser importante para a tarefa de atribuição de autoria, motivo pelo qual foi escolhida e estudada no presente trabalho.

Buscou-se, desta forma, a realização de uma série de testes, conforme descrito no capítulo 5, onde foram utilizadas duas bases de dados elaboradas em trabalhos anteriores para que o resultados dos testes feitos com a medida NCD pudessem ser comparados ao desempenho de outras abordagens já utilizadas.

Após este trabalho, conclusões importantes puderam ser obtidas:

- O uso da medida NCD é importante para a tarefa de atribuição de autoria, produzindo resultado relevantes para esta tarefa;
- A medida NCD apresenta melhores resultados quando os documentos de treinamento são utilizados de forma independente, sem que seja feita a sua concatenação;

- Os resultados do método NCD podem ser processados por diversos mecanismos para a decisão da atribuição de autoria, sendo que a utilização do melhor resultado (ou seja, da menor distância NCD) apresenta um melhor desempenho;
- Existem algumas outras medidas estatísticas propostas, por exemplo, por (Malyutov, M.B. Wickramasinghe,C. I. e Li, S., 2007), que apresentaram um resultado inferior ao obtido pela distância NCD;
- Existem outras medidas que produzem bons resultados quando os documentos de treinamento são concatenados, por exemplo os resultados obtidos pela método CCC;
- O processamento computacional exigido pela distância NCD é influenciado pela quantidade de documentos de treinamento utilizados, sendo diretamente proporcional a estes;
- A quantidade de autores possíveis influencia o desempenho da atribuição de autoria, mas de uma maneira pouco intensa, já que a redução de 90% do número de autores disponíveis significou um aumento médio de 20% a mais de acertos.

Diversas melhorias são possíveis em trabalhos futuros a respeito do tema abordado, podendo ser destacadas:

- a verificação de outros temas que possam ser utilizados para a composição da base de documentos utilizadas nos testes;
- a verificação de outros compressores, especialmente compressores especializados na compressão de documentos de texto de pequeno tamanho;
- a verificação do desempenho da medida NCD para a atribuição de autoria de textos menores, por exemplo de documentos que possuam no máximo 140 caracteres, tais como utilizados em mensagens de celular;
- a pesquisa de modelos estatísticos que permitam caracterizar um documento de treinamento como relevante ou dispensável para a medida da distância NCD e posterior atribuição de autoria.

Os objetivos do presente trabalho foram alcançados, pois a medida de distância NCD pode ser estudada e utilizada para a atribuição de autoria de documentos de texto de língua portuguesa, alcançando resultados importantes; teste foram feitos comparando duas maneiras de se utilizar os documentos de treinamento para a posterior utilização do cálculo da medida

NCD; alguns mecanismos de escolha do melhor resultado puderam ser comparados, indicando uma pequena vantagem na utilização da menor distância NCD obtida como mecanismo de atribuição de autoria; a verificação da influência da quantidade de documentos de treinamento para cada autor e da quantidade de autores possíveis; e a obtenção de resultados que indicam que este trabalho pode contribuir para o aumento do conhecimento sobre o uso de compressores de dados para a atribuição de autoria e para o auxílio de peritos e outros profissionais que tenham que verificar a autoria de um documento.

Referências Bibliográficas

Basile, C. **Entropy and Semantics: Textual Information Extraction Through Statistical Methods.** Università di Bologna, 2010. Tese de doutorado.

Benedetto, D., Caglioti, E. e Loreto, V. **Language Trees and Zipping.** Physical Review Letters (88), 2002.

Burrows, J. F. **Word patterns and story shapes: The statistical analysis of narrative style.** Literary and Linguistic Computing (2), p61–70, 1987.

Cébrian, M., Alfonseca, M. e Ortega, A. **Common Pitfalls Using The Normalized Compression Distance: What To Watch Out For In A Compressor.** Communications In Information And Systems (5), p. 367-384, 2005.

Cilibrasi, R. e Vitányi, P. M. B. **Clustering by Compression.** IEEE Transactions On Information Theory (51) p. 1523–1545, 2005

Cilibrasi, R. **Statistical Inference Through Data Compression**. Institute for Logic, Language and Computation . Amsterdan, 2006. Tese de doutorado.

Coutinho, B. C. *et al*. **Atribuição de Autoria usando PPM**. XXV Congresso da Sociedade Brasileira de Computação. III TIL, p. 2208-2217, 2005.

Fano, R. M. **The transmission of information.** Technical Report No. 65: Research Laboratory of Eletronics - MIT, 1949.

Foster, D. **A funeral elegy: William Shakespeare's "best-speakingwitnesses"**. Publications of the Modern Language Association of America (111), p.1080, 1996.

Grünwald, P. D. e Vitányi, P. M.B. **Kolmogorov Complexity and Information Theory (With an Interpretation in Terms of Questions and Answers)**.Journal of Logic, Language and Information (12), p. 497-529, 2003.

Hammer, D. *et al.* **Inequalities for Shannon Entropy and Kolmogorov Complexity**. Journal of Computer and System Sciences (60), p. 442-464, 2000.

Holmes. D. I. **The analysis of literary style — a review.** Journal of the Royal Statistical Society: Series A (148), p. 328–341, 1985.

Juola, P. **Authorship Attribution. Foundations and Trends in Information Retrieval** (3), p. 233-334, 2008.

Justino, E. J. R. **Análise de Documentos Questionados.** Tese de Doutorado. Pontifícia Universidade Católica do Paraná: Curitiba, 2002.

Keselj, V. *et al*. **N-gram-based author profiles for authorship attribution.** In Proceedings of the Pacific Association for Computational Linguistics, p. 255–264, 2003.

King, E. G. C. (2010). **Fragmenting authorship in the eighteenth-century Shakespeare edition.** Shakespeare (6), p 1-19, 2010.

Koppel, M. e Schler, J. **Authorship Verification as a One-Class Classification Problem**. Proceedings of the 21st International Conference on Machine Learning. New York, 2004.

Koppel, M., Schler, J. e Bonchek-Dokow, E. **Measuring differentiability: Unmasking pseudonymous authors.** Journal of Machine Learning Research (8), p. 1261-1276, 2007.

Kukushkina, O. V., Polikarpov A. A. e Khmelev, D. V. **Using Literal and Grammatical Statistics for Authorship Attribution.** Problems of Information Transmission (37), p. 172-184, 2001.

Lee. T. J. **Kolmogorov Complexity and Formula Size Lower Bounds**. Institute for Logic, Language and Computation - Amsterdan, 2006. Tese de doutorado.

Li, M. *et al*. **The similarity metric.** IEEE Trans. Information Theory, (50), p.3250–3264, 2004.

Madigan, D. *et al*. **Author identification on the large scale**. In Proceedings of CSNA-05, 2005.

Mahmood, W. e Akhtar, M. F. **Validation of Machine Learning and Visualization based Static Code Analysis Technique**. Department of Interaction and System Design - Ronneby , 2009. Dissertação de mestrado.

Malyutov, M. B., Wickramasinghe, C. I. e Li, S. **Conditional Complexity of Compression for Authorship Attribution**. SFB 649 Discussion Paper No. 57, Humboldt University, Berlim, 2007.

Marton, Y., Wu, N., & Hellerstein, L. **On compression-based text classification.** In Proceedings of the European Conference on Information Retrieval (pp. 300–314). Berlin, Germany: Springer. 2005

Merivuori, T. e Roos, T. **Some Observations on the Applicability of Normalized Compression Distance to Stemmatology.** Proceedings of the Second Workshop on Information Theoretic Methods in Science and Engineering, 2009. Tampere, Finland.

Merriam, T. **Tamburlaine Stalks in Henry VI**. Computers and the Humanities (30), p. 267-280, 1996.

Mosteller, F. e Wallace, D. L. **Inference and disputed authorship: The federalist.** Addison-Wesley. Reading, Massachusetts: 1964.

Pavelec, D. F. **Identificação da Autoria de Documentos: Análise Estilométrica da Língua Portuguesa usando SVM.** Pontifícia Universidade Católica do Paraná, 2007. Dissertação de Mestrado.

Rudman, J. **The state of authorship attribution studies: Some problems and solutions.** Computers and the Humanities (31), p. 351-365, 1998.

Salomon, D. **Data Compression: The Complete Reference.** 3rd Edition. Springer, New York, 2004.

Schmidhuber, J. **Discovering Solutions with Low Kolmogorov Complexity and High Generalization Capability.** In Machine Learning: Proceedings of the 12th International Conference., San Francisco, p. 488-496, 1995.

Sebastiani, F. **Machine Learning in Automated Text Categorization**. ACM Computing Surveys (34), p. 1-47, 2002.

Shannon, C. E. **A Mathematical Theory of Communication**. Bell System Technical Journal (27), p. 379-423, 1948.

Şimşekli, U. **Automatic Music Genre Classification Using Bass Lines**. 2010 International Conference on Pattern Recognition , Washington DC, 2010.

Stamatatos, E. **A survey of modern authorship attribution methods.** Journal of the American Society for Information Science and Technology (60), p. 538-556, 2009.

Stamatatos, E., Fakotakis, N. e Kokkinakis, G. **Automatic text categorization in terms of genre and author**. Computational Linguistics (26), p. 471-495, 2001.

Thisted, R. e Efron, B. **Did shakespeare write a newly-discovered poem?**. Biometrika, (74), p. 445–455, 1987

Varela, P. J. **O uso de atributos estilométricos na identificação da autoria de textos**. Pontifícia Universidade Católica do Paraná, 2010. Dissertação de Mestrado.

Williams, C. B. **Mendenhall's Studies of Word-length Distribution in the Works of Shakespeare and Bacon.** Biometrika (62), p.207–212, 1975.

Yule. G. U. **On sentence-length as a statistical characteristic of style in prose, with applications to two cases of disputed authorship.** Biometrika (30), p. 363–390, 1938.

Zheng, R. *et al.* **A framework for authorship analysis of online messages: Writing-style features and techniques.** Journal of the American Society for Information Science and Technology (57), p.378–393, 2006.

Zipf, G. K. **Selected studies of the principle of relative frequency in language**. Harvard University Press. Cambridge: 1975.

Apêndice A

Área	Código Área	Código Autor	Código Área e autor	Autor
Assuntos Variados	Q	a	Qa	FatimaOliveira
		b	Qb	GilbertoDimenstein
		c	Qc	GildaDeCastro
		d	Qd	GracePassô
		e	Qe	LuizFlavioSapori
		f	Qf	MarceloRossi
		g	Qg	OswaldoBraga
		h	Qh	SebastiaoNunes
		i	Qi	SilvanaMascagna
		j	Qj	Trigueirinho
Direito	R	a	Ra	BoleslauSliviany
		b	Rb	CarlosZamithJunior
		c	Rc	FabioTokars
		d	Rd	FernandoCesarFaria
		e	Re	FredericoVasconcelos
		f	Rf	IgorFonsecaRodrigues
		g	Rg	JorgeAlbertoAraujo
		h	Rh	MariaInesDolci
		i	Ri	OscarIvanPrux
		j	Rj	ReneArielDotti
Economia	S	a	Sa	AnaCristinaCavalcante
		b	Sb	AntonioPietrobelli
		c	Sc	BenedictoDutra
		d	Sd	ClaudioGradilone
		e	Se	FernandoCanzian
		f	Sf	GuilhermeBarros
		g	Sg	KarlonAredes

Área	Código Área	Código Autor	Código Área e autor	Autor
		h	Sh	LuisNassif
		i	Si	ValdoCruz
		j	Sj	ViniciusTorresFreitas
Esportes	T	a	Ta	AndreRibeiro
		b	Tb	AugustoMafuz
		c	Tc	DiogoOlivier
		d	Td	MarceloSenna
		e	Te	SergioRedes
		f	Tf	MarcioBernardes
		g	Tg	Tostão
		h	Th	ValdirBicudo
		i	Ti	VicenteDatolli
		j	Tj	WianeyCarlet
Gastronomia	U	a	Ua	AlessandraBlanco
		b	Ub	AndreaKaufmann
		c	Uc	CarlosBertolazzi
		d	Ud	CilmaraCastilho
		e	Ue	MarciaDaskal
		f	Uf	MarthaStewart
		g	Ug	NeideRigo
		h	Uh	NigellaLawson
		i	Ui	RicardoCastilho
		j	Uj	TatianaDamberg
Literatura	V	a	Va	ArnaldoJabor
		b	Vb	CeciliaGiannetti
		c	Vc	FernandoMonteiro
		d	Vd	LauraMedioli
		e	Ve	LuizBras
		f	Vf	ManoelLobato
		g	Vg	MarceloCoelho
		h	Vh	NelsondeOliveira

Área	Código Área	Código Autor	Código Área e autor	Autor
		i	Vi	PauloCoelho
		j	Vj	SergioRodrigues
Politica	W	a	Wa	AcílioLaraRezende
		b	Wb	BadgerVicari
		c	Wc	CarlaKreeft
		d	Wd	CarlosBrickmann
		e	We	ClaudioHumberto
		f	Wf	ClaudioSchamis
		g	Wg	FabioCampana
		h	Wh	FabioCampos
		i	Wi	MargritSchimidt
		j	Wj	VittorioMedioli
Saude	X	a	Xa	ClaudioLima
		b	Xb	DrauzioVarela
		c	Xc	FabioCesardosSantos
		d	Xd	FernandaAranda
		e	Xe	FlavioSettanni
		f	Xf	JohnCookLane
		g	Xg	LeandroPerché
		h	Xh	LeoKahn
		i	Xi	LilianeFerrari
		j	Xj	LoirCarlosCosta
Tecnologia	Y	a	Ya	Alexandre Magalhaes
		b	Yb	Cezar Taurion
		c	Yc	Denny Roger
		d	Yd	Eduardo Tude
		e	Ye	Ewandro Schenkel
		f	Yf	Fernando Birman
		g	Yg	Julio Preuss
		h	Yh	Marcelo Coutinho
		i	Yi	Marcelo Minutti

Área	Código Área	Código Autor	Código Área e autor	Autor
		j	Yj	Patricia Peck
Turismo	Z	a	Za	Adriano Gambarini
		b	Zb	Carlos Sarli
		c	Zc	Fabio Zanini
		d	Zd	Ivonildo Lavor
		e	Ze	Jose Pinto
		f	Zf	Lucia Malla
		g	Zg	Raul Lores
		h	Zh	Roberto Couto
		i	Zi	Roberto Linsker
		j	Zj	Rodrigo Baleia