ESPECTROMETRIA DE DOCUMENTOS ELETRÕNICOS – UM PROCESSO DE DETECÇÃO DE PLÁGIO

Joseval de Melo Santana[1]

1. INTRODUÇÃO

O estágio atual do desenvolvimento tecnológico traz à tona uma nova reorganização dos modos de produção e negócios, e conseqüentemente da economia, da sociedade e da política. Este novo paradigma toma por base a informação, contrapondo-se às revoluções tecnológicas anteriores, que tinha por base energia e matéria.

Com a consolidação da Sociedade da Informação (SI), as informações eletrônicas cresceram em projeções geométricas, hospedando-se nos mais variados *sites* e em mídias dos mais diversos tipos.

O processo acelerado de geração de informações eletrônicas decorrentes da Sociedade da Informação tem levado a uma disseminação caótica das informações. Questões sobre veracidade e autenticidade da informação eletrônica começam a ser indispensáveis, sob pena de comprometer a aceitabilidade e obstruir a utilização de documentos originais em mídia eletrônica.

Todas as áreas são afetadas por essa disseminação caótica e em particular a área acadêmica. Não é rara a dificuldade que os docentes têm em avaliar a autenticidade dos trabalhos dos seus alunos. A falta da autenticidade gera descrédito das informações e serve como barreira na disponibilidade de informações. Na realidade, sem nenhuma proteção de autoria os pesquisadores temem em colocar artigos, monografias e outros tipos de documentos à disposição em forma eletrônica, o que contrapõe, em parte, um dos pilares fundamentais do ensino-aprendizagem.

A espectrometria, denominação adotada para as técnicas de avaliação de documentos quanto à similaridade de seu conteúdo, é uma análise comparativa de espectros (conjunto de palavras) entre documentos, cujo objetivo é servir de medida para poder expressar o grau de autenticidade relativa[2] ou não. Tal métrica vem a servir como um parâmetro essencial para a análise qualitativa que norteará sobre a não autenticidade e/ou autenticidade relativa de documentos de maneira geral.

2. O DOCUMENTO ELETRÔNICO E O PLÁGIO

A evolução tecnológica e científica da humanidade tem se processado mediante a transmissão de conhecimentos de geração em geração, prática que tem a sua eficácia na produção de documentos, ou seja, toda informação contida em um suporte material que tenha a propriedade de ser comunicada. Este suporte consiste em "uma substância que permite a fixação dos signos gráficos no qual é expresso o documento" (ZAGAMI, 1996, p. 151).

O processo de documentação da informação (pensamento) tem evoluído desde os ideogramas impressos em rochas até a escrita em mídia papel e/ou digital. O documento, hoje, é a base do conhecimento colocado à disposição para tornar conhecida a expressão do pensamento, das ações e experiências de seu autor.

O autor mantém com seu documento uma relação de propriedade, mas cabe ressaltar que se trata de uma propriedade que revela não as posses do autor, e sim a intelectualidade deste. Como em toda propriedade, o Documento Eletrônico (**DE**) é alvo de ações criminosas que violam os seus direitos. Nesse caso, tem-se o plágio como um tipo específico de crime praticado contra o patrimônio intelectual.

---

[1] Professor Mestre do Departamento de Informática da Universidade Católica do Salvador – UCSal josevalms@ucsal.br

[2] A autenticidade é considerada relativa, quando visa garantir a autenticidade somente entre as amostras analisadas. Não infere sobre a originalidade do documento.

O plágio consiste na produção de um documento subsidiado na cópia, parcial ou total, de outro(s) documento(s) com intuito de se revelar como um documento autêntico. O pseudo-autor (plagiador), de forma ingênua ou intencional, tem cometido o plágio motivado, principalmente, pela idéia da não detecção do crime cometido. Detectar o plágio é condição *sine qua non* para garantir que um certo documento é inautêntico. O método de inspeção visual tem sido por muito tempo o único meio de detecção de plágio. Infelizmente, este método tem se mostrado ineficiente na detecção.

Os moinhos de documentos eletrônicos (*sites* que disponibilizam desde simples trabalhos escolares até teses de doutorado) têm sido os grandes incentivadores dos plagiadores que, de forma intencional, ou mesmo ingênua, praticam cada vez mais o crime de plágio.

Estudos recentes mostram que cerca de 30% dos estudantes devem estar plagiando documentos eletrônicos acadêmicos (PLAGIARISM, 2002, p. 1).

Mesmo diante de conseqüências severas que podem ser aplicadas em caso de comprovação de plágio pelas instituições aos plagiadores, estes não se intimidam e apostam na não detecção da violação cometida.

O plágio pode ser considerado como uma das mais sérias formas de violação da conduta acadêmica e profissional.

Métodos de detecção de plágio baseado em computadores têm surgido na última década. Contudo, tais sistemas por serem, na grande maioria, de natureza privada têm tornado difícil o acesso a essa tecnologia, principalmente devido aos custos do serviço de detecção de plágio. Por outro lado, também a falta de clareza dos parâmetros utilizados para majoração e qualificação do documento como plagiado, ou não, tem dificultado a aceitação e até mesmo a utilização de tais métodos.

A proposta apresentada adiante, denominada de Espectrometria de Documentos Eletrônicos, visa a ser uma forma de combate ao plágio diferenciando das soluções existentes pelos seus conceitos, técnicas e algoritmos.

3. ESPECTROMETRIA DE DOCUMENTOS ELETRÔNICOS

Espectrometria de Documentos Eletrônicos (EDE) consiste em um processo de detecção de plágio por computador com critérios e conceitos transparentes visando à possibilidade de padronização no deferimento de um documento eletrônico quanto à sua distinção em relação a outros documentos.

Como em qualquer outra espectrometria, o conhecimento dos elementos que a compõem é fundamental para sua compreensão. Deste modo, faz-se necessária a conceituação de seus elementos, como segue:

1. espectro: É o conjunto, finito e não vazio, de **palavras** de um documento. Segundo Paulo Menezes, em seu livro Linguagens Formais e Autômatos, "a **Palavra** (Cadeia de caracteres) é uma seqüência finita de símbolos justapostos" (MENEZES, 2001, p.21);
2. amostra: É o conjunto, finito e não vazio, de espectros não repetidos de um documento.

A espectrometria baseia-se em resultados quantitativos e qualitativos extraídos da análise. A análise quantitativa revela-se nos parâmetros espectrométricos considerados a seguir:

1. ocorrências (Oc): São espectros repetidos entre as amostras dos documentos em análise;
2. valor espectrométrico (Ve): É a razão percentual da quantidade de ocorrências (qoc) pela quantidade de espectro da amostra (qea) de um documento.

$$Ve = (qoc/qea)*100 \qquad (3.1)$$

3. Medida espectrométrica (Me): É o maior valor espectrométrico obtido entre os pares das amostras dos documentos e que também revelará o grau de distinção entre esses documentos.
4. Relação binária (Rb): É a combinação de pares das n ($2 \geq n < \infty$) amostras dos documentos analisados conforme a expressão matemática:

$$Rb = (n-1)*n/2 \qquad (3.2)$$

5. Faixa espectrométrica (Fe): É o intervalo, compreendido entre zero por cento (0%) e cem por cento (100%), em que a medida espectrométrica pode se encontrar.

A análise qualitativa identificará se um documento é ou não distinto de um outro documento eletrônico. Esta análise se baseia na análise quantitativa podendo ser subsidiada pela inspeção visual.

A espectrometria visa a minimizar as inspeções visuais dos documentos, que tenham como objetivo a detecção de plágio, e servir de diagnóstico na distinção de um documento. Ela faz a comparação dos espectros baseando-se na gramática no sentido léxico e não leva em consideração a semântica das palavras.

A espectrometria é realizada mediante um processo espectrométrico descrito a seguir.

3.1. Pré-processamento espectrométrico

O processamento espectrométrico consiste em uma análise comparativa das amostras dos documentos eletrônicos. Para tanto, tais documentos têm de passar por uma etapa inicial denominada pré-processamento espectrométrico, que consiste em:

1. arquivo texto simples: O documento é transformado em texto no formato padrão ASCII mediante conversores de tipos de documentos.
2. Normalização do documento: É o procedimento de criação dos espectros (conjunto de palavras) a partir de palavras que tenham um determinado número de caracteres e que pertençam ao arquivo texto simples. A normalização cria os espectros tomando como referência um delimitador, ou seja, um sinal de pontuação tal como: o ponto, a vírgula, dois pontos, outros.
3. Arquivo de texto *hash* (opcional)*:* O documento normalizado é submetido à função *hash*. Tal transformação é aplicada aos espectros do documento mantendo, porém, inteligíveis todos os delimitadores do documento.

4. ESPECTRÔMETRO ELETRÔNICO

O espectrômetro (Figura 4.1) é uma máquina de inferência (desenvolvido na linguagem JAVA) capaz de executar um algoritmo espectrométrico, que possibilita uma análise de vários Documentos Eletrônicos, permitindo obter a medida espectrométrica entre os referidos documentos. Tal medida servirá de revelação do grau de distinção destes.

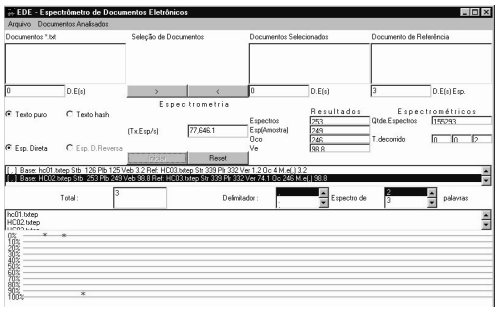

Figura 4.1 – Espectrômetro

O protótipo do espectrômetro desenvolvido apresenta algumas características básicas como segue:

1. Os espectros são constituídos de palavras de tamanho variável definido pelo usuário;
2. A ferramenta permite a análise de documentos em texto simples e em textos transformados via uma função *hash* com o algoritmo MD5;
3. Aplica espectrometria apenas direta;
4. Analisa documentos apenas em base local; e
5. A ferramenta foi desenvolvida para execução em computadores convencionais tipo *desktop* utilizando plataforma Windows.

Ensaios espectrométricos foram realizados em documentos eletrônicos de forma a demonstrar a eficiência da solução desenvolvida. Em seguida, descreve-se a experimentação e os elementos de composição dos referidos experimentos.

Experimento 01:

O experimento 01 foi baseado em documentos produzidos por alunos cujo tema abordava a área de informática, tais como Pirataria de Software, Redes de Computadores, Tecnologia Web e Sistemas Operacionais. A base dos documentos foi extraída dos trabalhos de pesquisas de 180 (cento e oitenta) alunos que produziram 60 (sessenta) documentos eletrônicos.

Experimento 02:

O experimento 02 foi baseado em buscas efetuadas sobre um mesmo tema na Internet formando uma base local de 09 (nove) documentos.

A análise espectrométrica foi realizada com as seguintes condições:

- Modo direto; e
- Documentos normalizados com espectros de duas palavras alinhados pelo delimitador vírgula (,).

Os resultados obtidos foram os seguintes:

Tabela 4.1 - Medidas espectrométricas (Experimento 01)

Faixa Espectrométrica	**DE Quantidade.**	**% (Total)**
Me = 0	4	6,67
0< Me <=5	15	25,00
5< Me <=10	26	43,33
10<Me <=15	7	11,67
15<Me <=20	0	0
20<Me <=25	0	0
25<Me <=30	0	0
30<Me <=40	0	0
40<Me <=50	2	3,33
50<Me <=60	2	3,33
60<Me <=70	0	0
70<Me <=80	0	0
80<Me <=90	0	0
90< Me <100	0	0
Me =100	4	6,67

Tabela 4.2 - Medidas espectrométricas (Experimento 02)

Faixa Espectrométrica	**DE Quantidade**	**% (Total)**
Me = 0	0	0,00
0< Me <=5	1	11,11
15< Me <=20	1	11,11
20< Me <=25	0	0,00
25< Me <=30	1	11,11
30< Me <=40	0	0,00
40< Me <=50	0	0,00
50< Me <=60	0	0,00
60< Me <=70	2	22,22
70< Me <=80	1	11,11
80< Me <=90	0	0,00
90< Me <100	2	22,22
Me =100	0	0,00

Fonte: Dados extraídos do módulo de estatística de documentos analisados – Espectrômetro (Protótipo).

Confrontando os resultados obtidos nos experimentos 01 e 02 com a inspeção visual dos documentos, conclui-se que:

1) Dos 60 documentos analisados (Tabela 4.1), constatou-se que, nos intervalos de médias espectrométricas, maiores que vinte por cento (20%) e menores e iguais a cem por cento (100%), foram encontrados oito (08) documentos, ou seja, 13,67% do total dos documentos considerados não distintos, conforme observação abaixo:

- 04 documentos (6,67%) foram classificados como cópias idênticas (**Me** = 100%).
- Os outros 04 documentos (6,67%) classificados como cópia parcial.

2) No intervalo espectrométrico, maior ou igual a zero por cento (0%) e menor e igual a dez por cento (10%), foram verificados distinções entre dezenove (19) documentos.

3) No intervalo espectrométrico maior que dez por cento (10%) e menor ou igual a vinte por cento (20%) foi verificado o seguinte:

- Existiu coincidência de espectros (referente à bibliografia, palavras-chave, etc.), que não comprometeram a distinção entre os documentos.
- Documentos com quantidade pequena de espectros (menos de 20% dos espectros do documento a ser comparado, foram considerados DE's não distintos).

  A espectrometria nestes documentos pode revelar medidas espectrométricas que comprometam a distinção dos mesmos, fato denominado de aberração espectrométrica.
- Mesmo recorrendo à inspeção visual houve incerteza na determinação de distinção de alguns documentos.

4) No intervalo espectrométrico (Tabela 4.2), maior que dez por cento (10%) e menor e igual a vinte por cento (20%), houve dúvida na garantia de distinção dos documentos, mesmo com o auxílio da inspeção visual. Este intervalo é denominado de intervalo crítico e a medida espectrométrica pertencente a este intervalo é chamada de **Me** crítica.

5) No intervalo espectrométrico (Tabela 4.2), maior que vinte por cento (20%) e menor e igual a vinte e cinco por cento (25%), 02 documentos foram classificados como não distintos pela inspeção visual.

O resultado da espectrometria de documentos eletrônicos retratou a realidade esperada e comprovada mediante confrontação com a inspeção visual. A análise espectrométrica, descartando-se os documentos que apresentaram aberrações espectrométricas e **Me** dentro do intervalo crítico, comprovou serem distintos os documentos eletrônicos com espectrometria maior ou igual a zero por cento (0%) e menor ou igual a dez por cento (10%) e não distintos, os documentos com espectrometria superior a vinte por cento (20%).

Os intervalos espectrométricos que qualificam documentos como distintos (0 a 10%), críticos (10%< **Me** <=20%), não distintos (**Me** > 20%) são intervalos convencionados tomadas como referência as análises experimentais, por isso, não devem ser tomadas como valores precisos.

Nestes experimentos, os melhores resultados espectrométricos foram obtidos na espectrometria com documentos normalizados com o delimitador vírgula (,). Contudo, a escolha do delimitador deve ser feita levando em consideração as características do idioma e/ou simulações espectrométricas que revelem as maiores medidas espectrométricas.

5. CONCLUSÃO

O trabalho desenvolvido propôs e avaliou os seguintes itens: a) conceitos e parâmetros espectrométricos; b) métrica espectrométrica; c) intervalos espectrométricos; d) classificação dos documentos em distintos, críticos e não distintos conforme intervalo espectrométrico; e) espectrômetro na forma de protótipo.

Ressalta-se que os intervalos espectrométricos foram extraídos de deduções empíricas, por isso, não devem ser tomadas como valores inflexíveis quantos aos seus limites superiores e inferiores.

Em termos de restrição e sugestão com relação ao trabalho desenvolvido têm-se: a) a espectrometria não detecta o plágio em documentos oriundos de traduções de outros idiomas; b) padronizar as métricas espectrométricas, levando em consideração as características inerentes aos documentos produzidos em cada área do conhecimento humano, mediante análise das estruturas das palavras e vários testes experimentais.

6. REFERÊNCIAS

MENEZES, Paulo Blauth. **Linguagens Formais e Autômatos**. 4. ed. Porto Alegre: Sagra Luzzatto, 2001.

PLAGIARISM.ORG. **Plagiarism**. Disponível em: http://www.plagiarism.org. Acesso em: 20 set. 2002.

ZAGAMI, Raimundo.**Firme 'Digitali' e Validità Del Documento Eletronico**.IN "II Diritto dell'informazione e dell'informatica", 1996, fasc. 1, p. 151.