Geração de Impressão Digital para Recuperação de Documentos Similares na *Web*

Álvaro R. Pereira Jr[1], **Nivio Ziviani**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627 – 31270-010
Belo Horizonte – Minas Gerais

{alvaro, nivio}@dcc.ufmg.br

Abstract. *This paper presents a mechanism for the generation of the "fingerprint" of a Web document. This mechanism is part of a system for detecting and retrieving documents from the Web with a similarity relation to a suspicious document. The process is composed of three stages: a) generation of a fingerprint of the suspicious document, b) gathering candidate documents from the Web and c) comparison of each candidate document and the suspicious document. In the first stage, the fingerprint of the suspicious document is used as its identification. The fingerprint is composed of representative sentences of the document. In the second stage, the sentences composing the fingerprint are used as queries submitted to a search engine. The documents identified by the URLs returned from the search engine are collected to form a set of similarity candidate documents. In the third stage, the candidate documents are "in-place" compared to the suspicious document. The focus of this work is on the generation of the fingerprint of the suspicious document. Experiments were performed using a collection of plagiarized documents constructed specially for this work. For the best fingerprint evaluated, on average* 87.06% *of the source documents used in the composition of the plagiarized document were retrieved from the Web.*

Resumo. *Este artigo apresenta um mecanismo para geração da "impressão digital" de um documento da Web. Esse mecanismo é parte de um sistema para detectar e recuperar documentos que tenham sido plagiados da Web, sendo similares a um dado documento suspeito. O processo é composto de três etapas: a) geração de uma impressão digital do documento suspeito, b) coleta de documentos candidatos da Web e c) comparação entre cada documento candidato e o documento suspeito. Na primeira etapa, a impressão digital do documento suspeito é usada para identificá-lo. A impressão digital é constituída por um conjunto de frases mais representativas do documento. Na segunda etapa, as frases que constituem a impressão digital são usadas como consultas e submetidas para uma máquina de busca. Os documentos identificados pelas URLs da resposta da pesquisa são coletados e formam um conjunto de documentos candidatos à similaridade. Na terceira etapa, os documentos candidatos são localmente comparados com o documento suspeito. O foco deste trabalho está na geração da impressão digital do documento plagiado. Experimentos foram realizados sobre uma coleção de documentos plagiados construída especialmente para este trabalho. Para a impressão digital de melhor resultado, em média* $87,06\%$ *dos documentos usados na composição do documento plagiado foram recuperados da Web.*

1. Introdução

Com a Internet a sociedade tem praticado plágio com mais facilidade. Desde escolas de primeiro grau até cursos de pós-graduação, a facilidade de se efetuar um *download* e copiar a informação encontrada tem levado a uma epidemia de plágio digital. Talvez o problema mais alarmante desta epidemia de plágio digital seja a contribuição para que o plágio cada vez mais faça parte de nossa cultura educacional. Estudantes que estão crescendo com a Internet muitas vezes não estão percebendo que estão praticando o plágio. Passa a ser muito natural a ação de se "copiar" e "colar". Os estudantes estão se acostumando a somente repetir o que alguém já fez, sem criatividade, inovação e, principalmente, sem aprendizado, pois não foi ele quem fez.

Recuperar documentos que possuam o conteúdo desejado por um usuário é uma tarefa complexa, principalmente em grandes repositórios de documentos como a *Web*. Esta tarefa é função das *máquinas de busca*, que mantêm páginas da *Web* em sua base de documentos. Toda a base de documentos da máquina de busca fica indexada em forma de uma estrutura de dados chamada arquivo invertido, que permite a realização de consultas. O usuário entra com palavras chaves relacionadas à resposta que gostaria de obter e através de uma medida de similaridade entre os termos da consulta e cada documento indexado, os documentos de maior similaridade são retornados. Para o presente trabalho, o problema continua a ser a recuperação de documentos em um grande repositório de documentos. No entanto, a consulta não é mais por palavras chaves, mas sim por um documento inteiro.

Este trabalho apresenta um mecanismo capaz de detectar e recuperar documentos da *Web* que possuam uma relação de similaridade com um dado documento suspeito, ou seja, tenham sido plagiados da *Web*. O processo é realizado em três etapas principais. A primeira etapa compreende a retirada da impressão digital do documento. A impressão digital representa e identifica o documento suspeito. É composta de frases do texto, que são utilizadas na segunda etapa do processo. A segunda etapa tem o objetivo de coletar da *Web* documentos candidatos a apresentarem uma relação de similaridade com o documento suspeito. Cada frase da impressão digital é utilizada como consulta em um sistema de busca que retorna os documentos que compõem a base de documentos candidatos à similaridade. Na terceira etapa, cada documento candidato é comparado com o documento suspeito. O foco deste trabalho está na etapa de geração da impressão digital, que será detalhada na seção 2. As demais etapas serão apresentadas de forma sucinta, na seção 3.

A avaliação do processo desenvolvido se deu pela capacidade do sistema em recuperar os documentos usados para compor o documento suspeito. Para que a avaliação pudesse ser realizada desenvolvemos um sistema gerador de documentos plagiados, capaz de compor um documento utilizando trechos de diferentes documentos coletados da *Web*. O sistema retorna as URLs[1] dos documentos usados na composição do documento plagiado. Verificamos que, para a melhor impressão digital avaliada, em $61,53\%$ dos casos, todos os documentos da composição foram recuperados e que somente em $5,44\%$ dos casos o desempenho foi menor que 50%. Para esta impressão, em média $87,06\%$ dos documentos foram recuperados da *Web*.

Desde 1994 vários mecanismos de verificação de similaridade entre documentos foram propostos, usando diferentes modelos e com diferentes finalidades. A ferramenta SIF [Manber, 1994] foi a pioneira, e tratava o problema da similaridade não somente para documentos, mas arquivos binários em geral. A ferramenta COPS (*COpy Protection System*) [Brin et al., 1995] e as diferentes versões do SCAM (*Stanford Copy Analysis Mechanism*) [Shivakumar and Garcia-Molina, 1995, Garcia-Molina et al., 1996,

[1] URL (*Uniform Resource Locator*) é o identificador único de um documento na *Web*, o seu endereço.

Garcia-Molina et al., 1998] são resultados de um dos maiores estudos realizados sobre detecção de cópias em grandes repositórios de documentos. A primeira versão do SCAM abordou o problema considerando o repositório de documentos localmente. As últimas versões funcionavam considerando a *Web* como sendo o repositório de documentos. [Pereira-Jr, 2004] apresenta e discute estes e alguns outros mecanismos de detecção de cópias já propostos.

2. Geração da Impressão Digital

A primeira etapa do sistema consiste em gerar uma impressão digital para o documento a ser pesquisado. O problema está em definir as características dessa impressão digital, uma vez que cada frase da impressão é posteriormente usada como uma consulta na etapa de pesquisa e coleta.

Ao buscar definir as impressões digitais, devemos lembrar que o objetivo não é procurar na *Web* pelos exatos documentos que tiveram a impressão digital obtida. O objetivo neste trabalho é usar a impressão digital para buscar por vários documentos que possam ter sido usados na composição do documento suspeito. Desta forma, realizar pesquisas por uma lista de termos espalhados pelo texto, ou pelos termos de maior frequência no documento, poderia resultar em um baixo desempenho. Isto ocorreria porque as listas de termos mais frequentes dos documentos usados na composição do documento suspeito certamente não seriam as mesmas.

Seis diferentes impressões digitais foram estudadas e implementadas. A maioria das impressões digitais utilizadas são compostas por uma lista sequencial de termos do texto, que chamaremos de *frases*, mesmo que muitas vezes estas listas não tenham um sentido semântico. Em alguns casos foram usados termos específicos como âncoras no texto, e cada frase foi formada tomando o mesmo número de termos à esquerda (incluindo o próprio termo) e à direita do termo âncora.

Para cada uma das impressões digitais propostas temos opções de variar a granularidade e a resolução da mesma. A *granularidade* é medida de acordo com a quantidade de termos contidos em cada frase da impressão digital. Todas as frases de uma mesma impressão digital têm a mesma granularidade. A *resolução* é medida pela quantidade de frases a serem obtidas para compor a impressão digital.

Uma vez que cada frase da impressão digital será uma consulta no sistema de busca, a maior granularidade considerada foi de dez termos, número máximo aceito pela maioria das máquinas de busca. Pelo mesmo motivo, a resolução deve ser a menor possível, implicando em menos requisições à máquina de busca e menos páginas coletadas para compor a base de documentos candidatos. Os métodos estudados são apresentados a seguir:

1. **Termos mais frequentes – TF**
   Uma impressão digital contendo os termos que mais ocorrem no documento. Sua resolução é sempre de uma frase, podendo variar a granularidade.
2. **Frases com termo incorreto – FTI**
   A implementação desta impressão digital foi motivada pela intuição de que frases que envolvam termos com erros ortográficos representam bem o documento, uma vez que acredita-se ter maior probabilidade de não existirem outros documentos com os mesmos termos incorretos. Utilizando o programa "ispell" da GNU[2], todos os termos que não fazem parte do dicionário da língua portuguesa são gerados

[2] GNU é um projeto de gerenciamento de um ambiente para desenvolvimento de software livre – http://www.gnu.org

e ordenados do termo de maior comprimento para o de menor comprimento. Assim, é dada menor prioridade para termos curtos, que podem ser apenas siglas não encontradas no dicionário utilizado. Os termos no topo da lista funcionam como âncoras no texto, para retirada das frases que irão compor a impressão digital.

3. **Frases espalhadas constantes – FEC**
   Frases espalhadas no texto, equidistantes umas das outras, são usadas para formar a impressão digital do documento. Independente do tamanho do texto, sempre o mesmo número de frases são obtidas, mantendo a resolução constante.
4. **Frases espalhadas proporcionais – FEP**
   Como a impressão FEC, porém a resolução é proporcional à quantidade de caracteres do documento, calculada de acordo com a equação:
   $res = k \times \log(qtdCarac/10)$, onde $qtdCarac$ é a quantidade de caracteres do documento, k é uma constante e res a resolução.
5. **Frases com termos mais frequentes – FTF**
   É gerada uma lista com os termos mais frequentes, que são usados como âncoras no texto para retirada de frases.
6. **Frases com termos menos frequentes – FTMF**
   A lista utilizada é a de termos menos frequentes, que também são usados como âncoras na retirada de frases. Como na maioria dos casos existem muitos termos com frequência um, os termos de maior comprimento são escolhidos.

2.1. Exemplo de Impressão Digital

Como exemplo, vamos considerar o trecho de texto[3], da figura 1 mantido com erros ortográficos, gramaticais e frases mal elaboradas, como sendo o documento da consulta no qual queremos retirar as diferentes impressões digitais. Vamos considerar ainda a granularidade sendo de quatro termos e a resolução de duas frases.

> O movimento insurrecional de 1789 em Minas Gerais teve característica marcantes que o fizeram distinguir-se das outras tentativas de independência, ele foi mais bem elaborado preparado que a Inconfidência Baiana de 1798 e a Pernambucana de 1801. Os Mineiros que lideraram a conspiração de 1785-1789 tinham bem em vista a Independência Global do Brasil, e não uma republica em Minas Gerais. O plano mineiro era em iniciar a revolta por Minas Gerais, e estendê-la ao Rio de Janeiro e em seguida as demais Capitanias, o produto não foi produto da mente de ninguém em particular, nasceu das condições estruturais da sociedade brasileira.

Figura 1: Exemplo de texto plagiado

A tabela 1 mostra as seis impressões digitais geradas para o texto de exemplo da figura 1. Para as impressões TF, FTF e FTMF, as *stop words*[4] são retiradas. A impressão TF é composta de apenas uma frase. A resolução não se aplica a este caso.

O texto teve três termos não encontrados no dicionário utilizado: “insurrecional”, “marcantes” e “republica”. Como a resolução foi definida como sendo de duas frases, a impressão digital FTI teve frases com os termos que apareceram mais acima do texto, uma vez que dois dos três termos possuem a mesma quantidade de caracteres. Para a impressão FEP, o resultado da equação apresentada na seção 2 definiu sua resolução como sendo 4, para a constante $k = 1$. Qualquer letra maiúscula encontrada é convertida para minúscula, antes mesmo da retirada da impressão digital.

[3] Trecho de texto sobre o movimento da inconfidência mineira, retirado em 06-10-2003, de http://www.geocities.com/athens/marathon/9563

[4] *Stop words* são palavras comuns da linguagem. Por este motivo, não representam bem o documento.

Tabela 1: Exemplo de impressão digital para as seis impressões definidas

Impressão digital	Exemplo	
1. TF	gerais minas produto 1789	
2. FTI	movimento insurrecional de 1789	característica marcantes que o
3. FEC	a inconfidência baiana de	em iniciar a revolta
4. FEP	das outras tentativas de	os mineiros que lideraram
	em minas gerais o	as demais capitanias o
5. FTF	minas gerais teve característica	minas gerais o plano
6. FTMF	teve característica marcantes que	a independência global do

3. Demais Etapas do Processo

3.1. Pesquisa e Coleta de Documentos Candidatos

A pesquisa utiliza um sistema de *metabusca* para a construção da base de documentos candidatos à similaridade. Cada frase da impressão digital do documento suspeito é usada como uma consulta simples em diversas máquinas de busca. O metabuscador consiste em um programa capaz de realizar consultas em máquinas de busca, podendo utilizar diferentes serviços de busca. Sua arquitetura é bem simples, uma vez que não precisa indexar documentos da Web, apenas consultá-los através dos serviços. MetaCrawler[5] e Miner[6] são exemplos de metabuscadores disponíveis na *Web*.

Para a realização deste trabalho desenvolvemos um metabuscador com arquitetura simplificada. O metabuscador simplesmente formata a consulta de acordo com o padrão utilizado pela máquina de busca TodoBr[7] e processa o resultado retornado, de forma a obter as URLs de resposta à consulta. Os documentos identificados por suas URLs podem ser recuperados e compor a base de documentos candidatos à similaridade.

3.2. Comparação Entre os Documentos

As etapas anteriores foram importantes para a construção da base de documentos candidatos à similaridade. A terceira etapa tem a função de comparar cada documento candidato com o documento suspeito, buscando verificar a similaridade entre os pares de documentos. Dois métodos foram utilizados: árvore Patricia e *Shingles*.

O primeiro método utiliza a árvore Patricia, estrutura de dados proposta em [Morrison, 1968]. A árvore Patricia é construída sobre o documento suspeito e os documentos candidatos têm seus conteúdos pesquisados na árvore, o que permite verificar a existência de longos trechos idênticos que ocorram no documento suspeito e em cada um dos candidatos. O segundo método utiliza o conceito de *shingles* [Broder, 1998] para medir a similaridade entre o documento suspeito e cada candidato, comparados dois a dois. Maiores informações sobre os métodos e algoritmos usados nesta etapa podem ser obtidas em [Pereira-Jr and Ziviani, 2003].

4. Resultados Experimentais

4.1. Construção de Coleções de Documentos Plagiados

Para a realização dos experimentos desenvolvemos um sistema gerador de documentos plagiados, utilizando trechos de documentos *Web*. O sistema foi desenvolvido de acordo

[5] http://www.metacrawler.com, 2004.
[6] http://www.miner.com.br, 2004.
[7] http://www.todobr.com.br, 2004.

com a intuição de que o usuário que utiliza a *Web* como fonte para a composição do seu documento não realiza, de forma significativa, alterações no texto plagiado. Assim, alterações como troca de palavras por sinônimos ou troca de termos de uma frase, mantendo o sentido original, não são tratadas pelo gerador, que simula uma *composição* de um documento a partir de outros documentos.

É necessário definir a quantidade de documentos *Web* que serão usados na composição do documento plagiado, bem como o tamanho, em número de termos, que o documento da composição deverá ter em relação ao tamanho dos documentos *Web* usados. O novo documento composto é chamado de documento "plagiado". O sistema inicialmente coleta os dez primeiros documentos retornados de consultas populares[8] realizadas na máquina de busca TodoBR. Em seguida é feita a leitura termo a termo do documento HTML para que seja retirado o texto, que é então separado em trechos (chamaremos de frases), definidos através de caracteres "ponto final". Frases aleatórias de cada documento são utilizadas na composição do documento plagiado, sempre mantendo o percentual de termos do documento candidato que está presente no documento plagiado.

4.2. Metodologia Utilizada nos Experimentos

Os experimentos foram realizados com o objetivo de verificar a capacidade do sistema em recuperar da *Web* o maior número possível de documentos que foram utilizados na composição do documento plagiado. Esses documentos vão compor a base de documentos candidatos. Os experimentos foram realizados buscando minimizar os custo do sistema que são: o número de requisições geradas na máquina de busca pela impressão digital, e o número de documentos que devem ser coletados para composição da base de documentos candidatos à similaridade. Assim, uma resolução maior, ou seja, que contém um número maior de frases, representa um custo maior para o sistema, uma vez que cada frase representa uma requisição à máquina de busca. Da mesma forma, coletar todos os documentos da resposta a uma consulta teria um custo maior que coletar somente o documento do topo do *ranking*.

Buscando reduzir o custo na realização dos experimentos, foi utilizada uma coleção reduzida de documentos plagiados no primeiro experimento, onde o melhor valor de granularidade é escolhido e usado nos próximos experimentos. Pelo mesmo motivo, a impressão digital de pior desempenho é excluída nos dois primeiros experimentos, e não mais utilizadas nos experimentos seguintes.

Os três experimentos para avaliação da etapa de geração da impressão digital tinham objetivos diferentes, mas foram realizados de forma semelhante, como mostra a figura 2. Inicialmente a impressão digital do documento plagiado é obtida. Em seguida cada frase da impressão é usada como uma consulta na máquina de busca TodoBR. As páginas retornadas pela consulta têm suas URLs comparadas com as URLs dos documentos usados na composição do documento plagiado, retornando então o percentual de documentos recuperados para aquela impressão digital.

4.3. Escolha do Melhor Valor para Granularidade

O primeiro experimento foi realizado com o objetivo de filtrar as impressões digitais utilizadas, escolhendo a melhor granularidade para cada impressão e excluindo aquela de pior resultado. Uma pequena coleção de 350 documentos plagiados foi utilizada. Com excessão das impressões FEP e TF, todas foram experimentadas com resoluções de 5, 10 e 15 frases e com granularidades de 4, 6 e 10 termos, combinando cada valor de resolução

[8] Utilizamos um arquivo de histórico diário de consultas do TodoBR, onde foram consideradas consultas realizadas de cinco a dez vezes no mesmo dia.

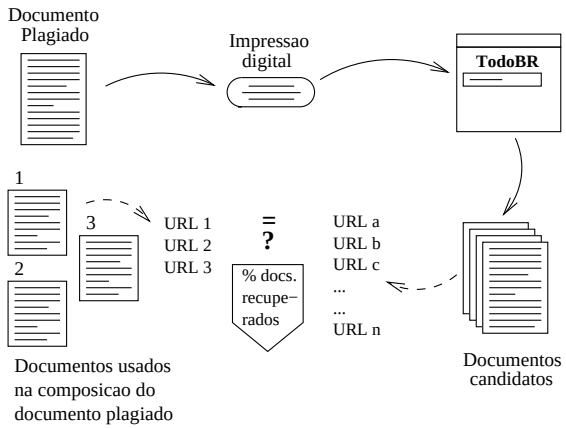

Figura 2: Modelo de experimento realizado para avaliar a etapa de geração da impressão digital

com os de granularidade. Para a impressão FEP, apenas os valores de granularidade foram variados, uma vez que a resolução é sempre definida por meio da equação apresentada na seção 2, neste caso com a constante $k = 2$. A resolução também não se aplica à impressão TF.

O gráfico da figura 3 faz a comparação entre os diferentes valores de granularidade, fazendo a média dos percentuais de documentos encontrados para as diferentes resoluções aplicadas. Percebemos que a maior granularidade experimentada, que foi de dez termos — o máximo permitido para consulta na maioria das máquinas de busca — apresentou os melhores resultados (exceto para TF), sendo este o valor de granularidade escolhido para os próximos experimentos. Para a impressão TF, formam consideradas 10, 30 e 50 páginas, sendo este último o de melhor resultado, como mostra a figura 3. Como esta impressão apresentou um baixo índice de documentos recuperados, ela será excluída dos próximos experimentos. Para as demais impressões digitais, os dez documentos do topo do *ranking* foram considerados.

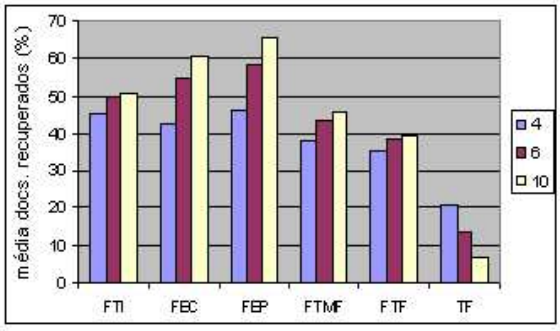

Figura 3: Comparação das diferentes granularidades para cada impressão digital

4.4. Impressões Digitais de Melhores Resultados

O experimento anterior foi útil para filtrar as possibilidades de impressões digitais para os documentos. Agora, temos o objetivo de avaliar a qualidade das impressões digitais para um número maior de documentos, tentando diminuir o custo para coleta. Foi utilizada uma coleção de 1.900 documentos plagiados para avaliar os resultados de cinco impressões digitais diferentes: FTI, FEC, FEP, FTF, FTMF, para três resoluções diferentes: 5, 10 e 15 frases. Para a FEP, a resolução é definida por meio da equação apresentada na seção 2, com dois valores para a constante k, que são $k = 1$ e $k = 2$, apresentando resoluções médias de $5,84$ e $12,15$ frases, respectivamente. A granularidade ficou fixada em dez termos, para todos as impressões digitais.

O gráfico da figura 4 faz uma comparação entre os percentuais médios de documentos recuperados, para cada impressão digital, com as diferentes resoluções. Impressões digitais de maior resolução apresentaram um melhor desempenho do que as impressões menores. Isto pode ser justificado pelo fato de que impressões maiores coletam um maior número de documentos.

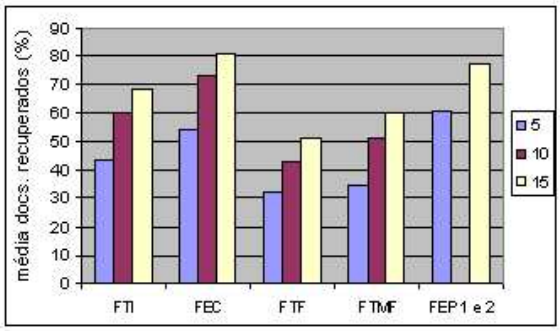

Figura 4: Comparação das diferentes resoluções para cada impressão digital

Na figura 4 vemos que a impressão de melhor resultado, a FEC com resolução 15, retornou $81,28\%$ dos documentos usados na composição do documento plagiado, seguido por FEP com $k = 2$, retornando $77,36\%$ dos documentos. A figura 5 apresenta o gráfico de pareto[9] para esta impressão, com valores acumulativos, classificando os índices de documentos recuperados de 10% em 10% (exceto para 100%). Verificamos que em $46,75\%$ dos casos, *todos* os documentos da composição foram recuperados. Somente em $8,71\%$ dos casos o desempenho ficou abaixo de 50%.

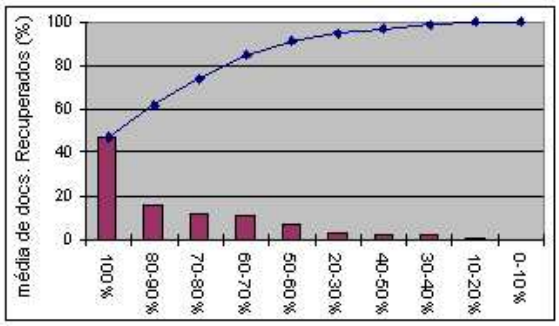

Figura 5: Gráfico de pareto para FEC com resolução 15.

Para este experimento, os *links* dos dez primeiros documentos retornados pelo sistema de busca foram analisados, em busca de algum documento que tenha sido usado na composição do documento plagiado. Coletar dez documentos de cada consulta realizada torna o processo caro em termos de coleta. Neste sentido, o experimento foi realizado de forma a também verificar qual era a posição do *ranking* do documento da composição encontrado. Verificamos que, em média, $81,66\%$ dos *documentos recuperados* estavam no topo do *ranking* e $93,12\%$ estavam ou no topo ou na segunda posição do *ranking*. Isto nos permite concluir que o desempenho do sistema, em termos de média de documentos recuperados, é pouco alterado quando se forma a base de documentos plagiados somente com os dois documentos do topo. O uso do sistema desta forma diminuiria o seu custo.

Fizemos uma análise manual buscando identificar situações específicas onde foi recuperado um número baixo de documentos usados na composição do documento plagiado, para a impressão FEC com resolução 15. Nessas situações vrificamos que os documentos usados na composição do documento plagiado eram: *home pages*, *blogs* com

[9] Gráfico de barras que enumera as categorias em ordem decrescente, da esquerda para a direita.

caracteres especiais, documentos contendo listas ou formulários. Estas verificações são indícios de que, nas situações em que um pequeno número de documentos foi recuperado, os documentos usados na composição do documento plagiado não eram boas representações de textos que normalmente são plagiados da *Web*. Assim, em uma situação real o sistema poderia apresentar melhor performance do que a verificada nos experimentos.

4.5. Combinação de Impressões Digitais

O experimento anterior buscou medir o desempenho do sistema para as diferentes impressões digitais de forma isolada. O objetivo agora é combinar as impressões, a fim de formar uma nova impressão com maior capacidade de recuperação de documentos similares. A mesma coleção do experimento anterior foi utilizada. A impressão de pior resultado do experimento anterior, FTF, não foi considerada. A resolução máxima considerada para as combinações foi de 30 frases. Desta forma, foi possível combinar todas as quatro impressões de resolução 5, ou combinar três a três as impressões com resolução de tamanho 10, ou ainda combinar duas a duas as impressões com resolução 15.

A nova impressão de melhor desempenho foi "FTI-FEC-FEP-10" (combinação das impressões FTI, FEC e FEP, com resolução 10 cada uma), seguida de "FTI-FEC-15", recuperando em média, respectivamente, $87,06\%$ e $86,63\%$ dos documentos usados na composição do documento plagiado. A figura 6 mostra o gráfico de pareto para a combinação "FTI-FEC-FEP-10". A análise do gráfico nos permite verificar um aumento significativo do desempenho para a nova impressão digital: em $61,53\%$ dos casos, *todos* os documentos da composição foram recuperados, contra $46,75\%$ da melhor impressão isolada, FEC, apresentada na figura 5. Isso representa um aumento de mais de 30% nas execuções que retornaram todos os documentos da composição, relacionado à impressão FEC. Para a mesma combinação, somente em $5,44\%$ dos casos o desempenho foi menor que 50%.

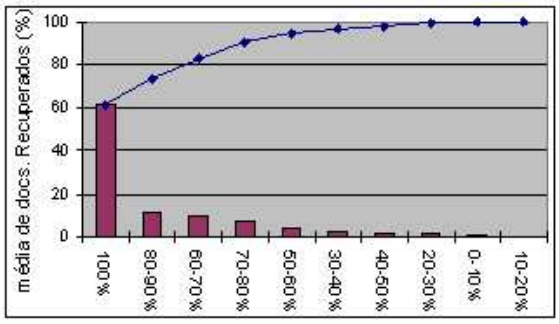

Figura 6: Gráfico de pareto para a combinação de impressões digitais "FTI-FEC-FEP-10".

5. Conclusões e Trabalhos Futuros

Um processo para detecção e recuperação de documentos similares na *Web* foi proposto e implementado. Através da construção de uma coleção de documentos plagiados, onde cada documento continha trechos de documentos da *Web*, foi possível medir e analisar o desempenho do processo.

O trabalho apresenta experimentos para medir o desempenho dos métodos utilizados na etapa de geração da impressão digital. Os experimentos foram realizados sobre uma coleção de documentos plagiados construída especialmente para este trabalho. Para a melhor impressão digital avaliada, em média 87% dos documentos usados na composição do documento suspeito são recuperados da *Web* e passam a compor a base de documentos

candidatos. Para a combinação de impressão digital "FTI-FEC-FEP-10", em quase 62% das execuções foi possível recuperar *todos* os documentos usados na composição do documento plagiado. Em média 93% destes documentos recuperados estavam entre os dois documentos do topo do *ranking*.

Como contribuições do trabalho, destacamos a proposta de um modelo eficaz para recuperação de documentos similares na *Web* e, ainda, um processo para avaliação do desempenho do modelo proposto, que pode ser utilizado para avaliar outros sistemas similares.

Uma sugestão de trabalho futuro é a construção de uma coleção de documentos plagiados a partir de documentos da *Web*, para ser disponibilizada para pesquisas em tópicos relacionados. As coleções utilizadas para este trabalho possuem tamanhos limitados (máximo de 1.900 documentos plagiados) e não estão estruturadas de forma a serem utilizadas com eficácia por terceiros. Para a construção desta coleção seria importante o levantamento estatístico do perfil de um documento plagiado. Com uma base de documentos que tenham sido manualmente alterados para fins de plágio, deve-se analisar os tipos de alterações que normalmente são feitas para, a partir daí, construir a coleção de documentos plagiados.

Referências

Brin, S., Davis, J., and Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. In *ACM SIGMOD Annual Conference*, pages 398–409, San Francisco.

Broder, A. (1998). On the resemblance and containment of documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society.

Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1996). dscam : Finding document copies across multiple databases. In *4th International Conference on Parallel and Distributed Systems (PDIS'96)*, Miami Beach.

Garcia-Molina, H., Ketchpel, S. P., and Shivakumar, N. (1998). Safeguarding and charging for information on the internet. In *International Conference on Data Engineering (ICDE'98)*.

Manber, U. (1994). Finding similar files in a large file system. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 1–10, San Fransisco, CA, USA.

Morrison, D. R. (1968). Practical algorithm to retrieve information coded in alphanumeric. *ACM*, 15(4):514–534.

Pereira-Jr, A. R. (2004). Recuperação de documentos similares na web. Master's thesis, Departamento de Ciência da Computação da Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

Pereira-Jr, A. R. and Ziviani, N. (2003). Syntactic similarity of web documents. In *First Latin American Web Congress*, pages 194–200, Santiago, Chile.

Shivakumar, N. and Garcia-Molina, H. (1995). Scam: A copy detection mechanism for digital documents. In *2nd International Conference in Theory and Practice of Digital Libraries (DL'95)*, Austin, Texas.

Stricherz, M. (2001). Many teachers ignore cheating, survey finds. *Education Week*. http://www.edweek.org/ew/ewstory.cfm?slug=34cheat.h20.