DOI: http://dx.doi.org/10.5769/C2011001

Uma análise do reconhecimento textual de nomes de pessoas e organizações na computação forense

Osvaldo Dalben Junior[1,2], and Daniela Barreiro Claro[3]
(1) SETEC/SR/DPF/BA, Departamento de Polícia Federal, Salvador, Brasil
(2) ENE/PG/FT/DEE, Universidade de Brasília, Brasília, Brasil, dalben.odj@dpf.gov.br
(3) FORMAS/LASID/IM/DCC, Universidade Federal da Bahia, Salvador-Bahia, Brasil, dclaro@ufba.br

Resumo — O Reconhecimento de Entidades Mencionadas (REM) é uma tarefa da área de Extração da Informação (EI) que visa identificar e classificar entidades (nomes de pessoas, organizações, locais, etc.) contidas em textos não estruturados. Este artigo propõe uma análise à aplicação do REM no campo da computação forense, em especial nas tarefas associadas ao exame de mídias apreendidas. Os resultados dos experimentos mostram que a utilização do REM contribui para a redução do tempo investido na etapa de análise de conteúdo das mídias apreendidas e para a revelação de informações latentes de nomes de pessoas e organizações contidos nessas mídias.

Palavras chaves — Reconhecimento de entidades mencionadas; REM; computação forense;

Abstract — Named entity recognition (NER) is a task in Information Extraction (IE) domain that aims to identify and classify entities (names, organizations, locations, etc.) inside unstructured text. This paper proposes to analyze NER applications in computer forensics area, especially in tasks related to seized media. Our experiments show that NER algorithms can minimize the time spent on analyzing of seized media and also revelate informations about names and organizations contained in these media.

Keywords — Named entity recognition; NER; computer forensics;

1. Introdução

O campo da computação forense é baseado na realização de exames periciais que objetivam o esclarecimento de ilícitos relacionados à área de informática. Segundo o autor de [1], cerca de 90% dos exames forenses computacionais no Brasil estão associados ao uso de equipamentos computacionais como ferramenta de apoio aos chamados crimes convencionais[1], como os fazendários e previdenciários, e não como um meio para a realização do crime.

O crescente processo de inclusão digital da sociedade brasileira ocorrido nas últimas duas décadas popularizou os equipamentos computacionais. Nestes equipamentos são armazenados diversos tipos de arquivos, tais como cópias de contratos, licitações, alvarás, guias de aposentadoria, cédulas falsas, trocas de *emails*, mensagens instantâneas, dentre outros.

As investigações policiais, em geral, envolvem a apreensão de materiais que possam conter informações sobre pessoas suspeitas, para que sejam periciados quanto à existência ou não dessas informações. Constata-se então que, independente do tipo criminal associado à investigação, poucos são os casos em que não ocorre a apreensão de equipamentos ou mídias computacionais, o que justifica o elevado percentual de exames periciais de informática associados a crimes convencionais.

Os exames periciais realizados em mídias computacionais são compostos por 4 etapas: a preservação das mídias originais, a extração do conteúdo das mídias, a análise do conteúdo extraído e a formalização dos exames. Em particular, nos exames periciais de informática associados a crimes convencionais, a etapa de análise do conteúdo se refere à análise manual dos arquivos contidos nas mídias e é normalmente realizada pela mesma equipe que analisa outros materiais apreendidos, como agendas e cadernos, dada a sua capacitação específica para esta tarefa. Neste caso, a etapa anterior (extração do conteúdo das mídias) pode sofrer a aplicação de filtros com o objetivo de reduzir a quantidade de arquivos que serão manualmente analisados.

Dentre os filtros automatizados que podem ser utilizados na etapa de extração do conteúdo das mídias, o mais comum é o descarte de arquivos conhecidos baseado em listas de *hash*[2]. Essas listas são geralmente compostas por arquivos típicos de sistemas operacionais e de instalação de aplicativos, portanto arquivos irrelevantes às investigações. Entretanto, mesmo com a utilização de listas de *hash* atualizadas, há ainda uma grande quantidade de arquivos irrelevantes que passam por esse filtro e, consequentemente, integram o conteúdo a ser analisado manualmente. Diante deste contexto, a equipe de análise tem normalmente duas opções: (i) investe tempo na análise de cada arquivo; (ii) parte do princípio de que o arquivo é irrelevante, devido às características do seu nome ou da pasta onde está armazenado, e o ignora. A opção (i) resulta no investimento desnecessário de tempo para a análise manual de arquivos irrelevantes, enquanto que a opção (ii) implica no risco de serem ignorados arquivos relevantes à investigação.

[1] Crimes convencionais, nesse contexto, são aqueles tipificados pelo Código Penal Brasileiro, cuja existência independe da informática

[2] *Hashes* são funções unidirecionais aplicadas a qualquer sequência de *bits* (como partições ou arquivos), com o objetivo de identificar unicamente esta sequência

O presente trabalho se utiliza de dados reais resultantes da análise humana de arquivos contidos em mídias apreendidas em operações da Polícia Federal do Brasil para demonstrar que o reconhecimento automatizado de nomes de pessoas e organizações contidos nesses arquivos pode representar um avanço no processo do exame pericial de mídias apreendidas. Para isso, um sistema de Reconhecimento de Entidades Mencionadas (REM) baseado em regras gramaticais e na Wikipédia foi aplicado sobre o conteúdo de mídias apreendidas em duas grandes operações policiais associadas a crimes de desvio de verbas federais, fraudes em licitações e exploração de jogos de azar. Sem intervenção manual, o sistema de REM identificou, nessas mídias, o conjunto dos arquivos cujos textos contêm nomes de pessoas ou organizações. Foi realizado então um estudo comparativo entre esse conjunto e o conjunto dos arquivos manualmente analisados pelas equipes de análise durante o andamento das operações. A análise dos resultados mostra que a automatização do REM pode contribuir para a revelação de informações latentes sobre pessoas e organizações suspeitas, além de funcionar como um filtro que elimina mais de 91% dos arquivos comumente analisados manualmente, com risco de descarte de arquivos relevantes inferior a 1%.

Este artigo está dividido da seguinte forma: as seções 2 e 3 apresentam o embasamento teórico da tarefa de REM e o Rembrandt, que é o sistema determinístico de REM utilizado no presente trabalho; a seção 4 descreve os experimentos e a seção 5 analisa os seus resultados; a seção 6 apresenta a conclusão e direcionamentos futuros da pesquisa.

2. Fundamentação Teórica

A. Reconhecimento de Entidades Mencionadas (REM)

O REM é uma das tarefas do campo da Extração da Informação (EI) que, por sua vez, está associada à Mineração de Textos (MT) e tem sido amplamente utilizada em várias áreas do Processamento de Linguagem Natural (PLN).

Um dos principais objetivos do REM é identificar e classificar os nomes das entidades contidas em textos não estruturados ou escritos em linguagem natural[3]. As principais entidades encontradas nos textos são: *pessoa*, *organização*, *local e tempo.* Além dessas, é comum também a existência de trabalhos de REM focados em expressões temporais e numéricas, como percentagem e valor monetário. A Tabela 1 exemplifica a tarefa de REM, através da etiquetação de um texto que contém nomes das entidades pessoa (PES), organização (ORG), local (LOC), tempo (TPO) e valor monetário (VAL).

TABELA 1: Exemplo de um texto etiquetado

As <ORG>Organizações Pedras Preciosas LTDA</ORG> foram vendidas para o <PES>Sr. Fulano dos Santos Jr.</PES> por <VAL>R$200.000,00</VAL>, o que levou à mudança da sua sede de <LOC>São Paulo</LOC> para o <LOC>Rio de Janeiro</LOC> em <TPO>2011</TPO>.

O contexto de uma entidade mencionada (EM), formado pelas palavras a ela relacionadas (geralmente por proximidade no texto), interfere no significado da EM e, consequentemente, na sua etiquetação [2]. Nesse sentido, um sistema de REM deve ser capaz de resolver ambigüidades e identificar, por exemplo, no texto da Tabela 1, que as palavras "Pedras" e "Preciosas" compõem a entidade organização e não são substantivos; a palavra "Santos" se refere a uma pessoa e não a um time de futebol ou um substantivo; e "Paulo" deve ser interpretado como um local e não uma pessoa.

O REM pode ser caracterizado como um problema de classificação cujo objetivo é atribuir para cada valor de entrada uma classe, identificada por um nome de EM. Na forma clássica de REM, os valores de entrada são representados pelas palavras ou *tokens* de um texto e a entidade representa a classe ou rótulo do *token* associado. Por exemplo, no texto da Tabela 1 a sequência de *tokens* "São Paulo" está associada à classe "LOC" ou está classificada como "LOC". Outra forma de entendimento é analisar o REM como um problema de predição de sequência de estados [3]. Neste caso, dada uma sequência X de *n tokens* de entrada, o objetivo é inferir a sequência Y de *n* estados de saída correspondente, onde y_i é classe de x_i, $0 < i \leq n$. A Tabela 2 representa as sequências X e Y associadas ao texto da Tabela 1, sendo que os elementos da sequência Y são representados pelas iniciais B, I e O, referentes à identificação de *tokens* situados na primeira posição de uma EM (*Begin*), situados em qualquer outra posição de uma EM (*Inside*) e não pertencentes a uma EM (*Outside*). As identificações B e I são acompanhadas da classificação da EM (PES, ORG, LOC, TPO ou VAL).

Considerando a Tabela 2, observa-se, por exemplo, que X contém o valor *Organizações* que está associado ao valor *B-ORG* de Y, indicando que o *token Organizações* se encontra na primeira posição de uma entidade mencionada no texto do tipo ORG, ou organização.

TABELA 2 – Sequências X e Y representando a tarefa de REM associada ao exemplo da Tabela 1

X={	As	;Organizações	;Pedras	;Preciosas	;LTDA	;foram	;vendidas	;para	;o	;Sr.				
	;Fulano	;dos	;Santos	;Jr.	;[...]	;Janeiro	;em	;2011	;.	}				
Y={	O	;B-ORG	;I-ORG	;I-ORG	;I-ORG	;O	;O	;O	;O	;B-PES				
	;I-PES	;I-PES	;I-PES	;I-PES	;[...]	;I-LOC	;O	;B-TPO	;O	}				

[3] Linguagem natural pode ser entendida como a linguagem humana ou linguagem ordinária.

A grande maioria dos sistemas já propostos para resolver o problema do REM se divide em duas categorias: baseados em regras manuais ou determinísticos; e baseados em aprendizado de máquina ou probabilísticos [4]. Ambas requerem alto grau de conhecimento lingüístico, seja para descrever as regras manuais ou para modelar os algoritmos de aprendizado.

Os modelos baseados em regras manuais formam a base dos primeiros sistemas de REM [5]. A sua concepção, em geral, é baseada na utilização de expressões regulares criadas manualmente que representam regras linguísticas associadas às palavras, como características gramaticais, ortográficas ou de vocabulário. A etiquetação é realizada de forma direta a cada associação existente entre palavras e regras. Na sentença "*O senhor Júlio está na praia.*", por exemplo, um modelo que contenha a regra "*se a palavra é precedida pelo pronome 'senhor(a)' e é iniciada com letra maiúscula, então é uma entidade do tipo 'pessoa'*" etiquetaria a palavra "*Júlio*" com a entidade "*pessoa*". Esses modelos possuem a vantagem de não necessitar de coleções de dados etiquetados para treinamento, pois não há qualquer aprendizado de máquina, entretanto requerem maior esforço de desenvolvimento e manutenção das regras, pela sua forte dependência das propriedades linguísticas associadas ao idioma dos textos.

Já os algoritmos probabilísticos de aprendizado de máquina baseiam-se no estudo quantitativo dos exemplos positivos e negativos contidos em coleções textuais de treinamento (etiquetadas) para modelar um sistema estocástico que objetiva inferir a identificação e classificação das entidades mencionadas contidas em um texto-alvo [6]. A precisão desses modelos está diretamente relacionada à quantidade de palavras, à qualidade da etiquetação, ao idioma e ao domínio das suas coleções de treinamento, que são os conjuntos de textos cujas EMs são previamente etiquetadas e preferencialmente revisadas, usados para o treinamento do modelo.

Quanto ao aprendizado de máquina, os modelos probabilísticos podem ser supervisionados, quando dependem de grandes volumes de textos etiquetados, semi-supervisionados, quando pouca informação etiquetada é suficiente para iniciar o modelo, ou não supervisionados, quanto independem de qualquer etiquetação prévia [4]. Para alcançar níveis de precisão próximos aos dos modelos supervisionados, os demais modelos utilizam-se de métodos complementares que objetivam o reconhecimento de padrões, como a exploração do contexto associado às entidades etiquetadas [7], a generalização de palavras através de classes semânticas pré-estabelecidas [8], a identificação de padrões de repetição de EMs em certos domínios textuais [9] e a similaridade de contexto entre grupos usando técnicas de agrupamento (*clustering*) em textos não etiquetados [10].

Dentre os principais algoritmos probabilísticos de REM, destacam-se o HMM [11], o MEM [12] e o CRF [13]. O HMM é um modelo generativo baseado na probabilidade de junção entre X (*token*) e Y (rótulo), que usa o teorema de Bayes para resolver o problema da probabilidade condicional, o MEM e o CRF são discriminativos, modelam p(Y|X) de forma direta e permitem a inclusão de um grande número de *features*[4] associadas aos *tokens*.

B. REM em textos escritos na língua portuguesa

Poucos são os trabalhos que propõem algoritmos de REM focados em textos da língua portuguesa, sejam eles baseados em regras manuais ou associados a modelos probabilísticos, assim como poucas são as coleções de dados etiquetadas para o treinamento dos sistemas de REM disponíveis publicamente neste idioma.

Dentre as pesquisas realizadas utilizando o REM para a lingua portuguesa, o Rembrandt [14] foi o sistema que obteve um melhor desempenho na etiquetação do corpus *Coleção Dourada do segundo HAREM* [15]. O Rembrandt é baseado em regras manuais e pode utilizar a Wikipédia [16] como uma fonte adicional de informação para a etiquetação das entidades.

C. REM independente do domínio textual

A utilização de um sistema de REM independente de domínio ainda requer uma analise aprofundada. Segundo os autores de [17], as coleções públicas, etiquetadas ou não, são extremamente limitadas ao domínio e ao idioma dos seus textos, e isso contribui para o estado incipiente no qual os trabalhos na área de EI independente do domínio se encontram. Atualmente, as ferramentas de análise de linguagem em geral requerem bastante interação de usuário quando há necessidade de mudanças associadas ao domínio. Alguns trabalhos estão sendo desenvolvidos na área de adaptação de domínio, como [18] e [19], cujo objetivo é, dado um modelo de REM treinado para etiquetar textos-alvo pertencentes a determinado domínio, adaptá-lo para a etiquetação em outro domínio. Estas propostas são, portanto, aplicáveis a textos pertencentes a domínios específicos, por isso não resolvem o problema da etiquetação independente do domínio.

Autores em [20] propuseram um sistema de REM baseado no modelo probabilístico MEM utilizando um algoritmo genético para a otimização das *features* locais e globais de acordo com os diferentes domínios dos textos-alvo. O sistema foi avaliado usando o *corpus* de notícias disponível na tarefa compartilhada do ConLL'03 [21] para treinamento e um corpus jurídico para testes, e vice-versa. Os resultados mostraram que a utilização das *features* otimizadas representou um ganho da ordem de 1% a 2% em relação à sua não utilização, porém o sistema alcançou somente 70% de precisão nos testes treinados com o corpus jurídico, enquanto que sistemas participantes do CoNLL'03 alcançaram precisão superior a 88% [21]. Esses números comprovam que a

[4] *Features*, nesse contexto, são características associadas às palavras, como o fato de ser numeral ou de ser iniciada por letra maiúscula

etiquetação de REM independente de domínio é um desafio ainda em aberto para a área de EI.

3. Rembrandt: Um Sistema de REM Basado em Regras Manuais

O Rembrandt [14] é um sistema de REM determinístico, baseado em regras gramaticais manuais e em informações extraídas da Wikipédia. Foi o sistema que obteve os melhores resultados para a etiquetação das entidades pessoa e organização na tarefa de avaliação conjunta de REM do segundo HAREM [15].

O presente trabalho depende da análise das ocorrências de EM em arquivos contidos em mídias apreendidas. Com o intuito de minimizar o esforço no reconhecimento dessas entidades, foi optado pela utilização do Rembrandt, devido principalmente à sua adaptação ao idioma português, precisão nos resultados do segundo HAREM, cujo corpus de teste contém textos pertencentes a diferentes domínios, e desenvolvimento através de código aberto, o que facilita possíveis intervenções adaptativas.

No Rembrandt, a Wikipédia é a base do conhecimento para a classificação das EMs. Para interagir com a Wikipédia, é utilizada uma interface que o autor denominou Saskia, que facilita a navegação pelas estruturas de categorias, ligações e redirecionamentos, possibilitando a extração do conhecimento.

Cada documento lido pelo sistema é submetido a uma sequência de processos de etiquetação sucessivos até alcançar a versão final. O funcionamento do Rembrandt pode ser dividido em três grupos:

(1) Divisão dos textos em sentenças e palavras; reconhecimento de expressões numéricas; identificação de palavras candidatas a EM; e geração de entidades candidatas aos tipos expressões temporais e valores. Na terceira sub-etapa, a regra de identificação de palavras candidatas a EM busca por sequências de palavras contendo pelo menos uma letra maiúscula e/ou algarismo, com ocorrência facultativa dos termos de, da, do, das, dos e e (da expressão regular "d[aeo]s?|e", também conhecida por "termos daeose"), exceto no início ou no final da sequência.

(2) Classificação das entidades candidatas resultantes da etapa anterior. Este processo é realizado primeiro pela Wikipédia, que relaciona todos os significados que a EM pode ter, e depois pelas regras gramaticais, que se utilizam das características internas e externas da EM para tentar a desambiguação. Em seguida, considerando as classificações obtidas, ocorre uma segunda aplicação de regras gramaticais com o objetivo de classificação das EM compostas, com ou sem os termos *daeose*, utilizando-se novamente da Saskia e de regras de classificação.

(3) Repescagem das EMs sem classificação. Nessa etapa, são aplicadas regras específicas para a detecção de relações entre EMs, com o objetivo de identificar relações entre EMs com e sem classificação e assim classificar as últimas. Em seguida, uma última tentativa de classificação é realizada através da comparação de EMs com uma lista de nomes comuns e, por fim, as EMs não classificadas são eliminadas. A Figura 1 ilustra essas três etapas descritas.

A estratégia de classificação de EM da Saskia, em linhas gerais, é dividida em três etapas [14]: (1) o emparelhamento das EMs, ou seja, cada entidade mencionada no texto deve estar associada a pelo menos uma página na Wikipédia, seja diretamente ou através da utilização de âncoras e redirecionamentos, (2) a memorização das categorias às quais a página da Wikipédia está associada e (3) a classificação dessas categorias através da aplicação de regras gramaticais específicas. Por exemplo, a EM "Porto" está associada às seguintes páginas da Wikipédia portuguesa: "a segunda maior cidade portuguesa", "cidade no Piauí, Brasil", "cidade em Zamora, Espanha", "aldeia na freguesia de Troviscal", "área localizada à beira d'água destinada à atracação de embarcações", "Futebol Clube do Porto", etc.. A etapa (1) faz

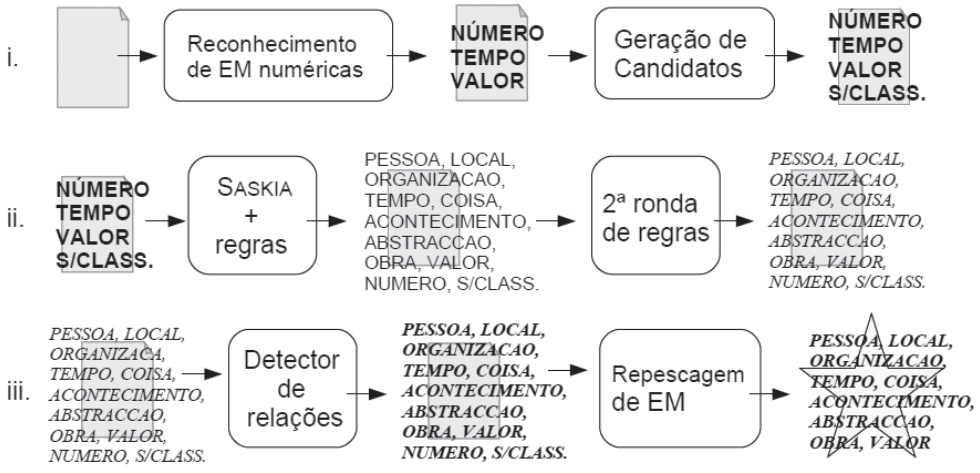

Figura 1 – O funcionamento do Rembrandt (extraído de [14])

esta associação entre a EM e as páginas, a etapa (2) memoriza as categorias associadas às páginas encontradas, como "Cidades de Portugal", "Municípios de Portugal", "Clubes de futebol de Portugal", etc., e a etapa (3) faz a associação entre essas categorias e as classes de EMs, como "local", "organização", etc. Cabe às regras gramaticais a desambiguação dessas classificações.

As regras gramaticais são desenhadas manualmente e buscam por padrões que revelem a existência de EMs nas sentenças e a execução de determinadas ações quando EMs forem encontradas. A sua atuação ocorre através de cláusulas aplicadas ordenadamente, de modo que a regra só é considerada bem sucedida no caso de todas as cláusulas a ela associadas retornarem o valor verdade. Além da geração de novas EMs, as regras podem disparar também outras duas ações: a detecção de relações entre entidades e a geração de mais de uma EM associada a uma mesma palavra.

A aplicação das regras é seqüencial, da esquerda para a direita, tanto para as palavras de uma sentença, quanto para as sentenças de um texto, e as regras bem sucedidas são imediatamente executadas, de modo que as novas EMs identificadas ou classificadas passam a ser consideradas pelas próximas regras aplicadas.

4. Experimentos

Os experimentos realizados no presente trabalho buscaram validar a utilização de sistemas de REM no processo de análise de mídias apreendidas em operações policiais. Através destes experimentos buscou-se analisar se a utilização do REM no contexto forense minimiza o tempo de análise manual do conteúdo das mídias apreendidas, sem prejuízo à qualidade das informações analisadas.

Para a realização dos testes, foi selecionado um cenário real que generalizasse ao máximo a dinâmica envolvida nas duas etapas intermediárias do processo do exame de mídias apreendidas, quais sejam, a extração do conteúdo das mídias e a análise do conteúdo extraído. O cenário foi composto por duas grandes operações policiais associadas a crimes de desvio de verba pública, fraude em licitações e exploração de jogos de azar. Esses tipos de operações, em geral, resultam na apreensão de grande quantidade de mídias computacionais, devido ao seu uso comum para armazenamento de dados pessoais e profissionais dos suspeitos. Como conseqüência, é bastante comum também a necessidade da análise manual de um número muito grande de arquivos contidos nessas mídias, tarefa essa muitas vezes realizada por poucos profissionais e sem dedicação exclusiva, variáveis que são inversamente proporcionais ao tempo necessário para a conclusão do processo de análise.

A Tabela 3 apresenta o quantitativo de mídias e arquivos envolvidos nas duas operações avaliadas, bem como o tempo que foi necessário para o cumprimento de algumas tarefas.

Na Tabela 3, a quantidade de mídias analisadas (1) corresponde às mídias apreendidas que contêm informações relevantes para serem analisadas. Mídias que apresentaram erro de leitura ou que não continham arquivos do tipo documento não estão representadas nesses números. Nota-se que houve considerável quantidade superior de mídias na *operação A*. O tempo investido no processamento das mídias (2) envolve as duas primeiras etapas de exame, que são a sua duplicação (para preservação do conteúdo original) e a extração dos arquivos que serão analisados manualmente (nessa etapa ocorre o filtro de arquivos conhecidos, que foi detalhado na introdução deste artigo), além do tempo necessário para a disponibilização dos arquivos para a equipe de análise. A quantidade de arquivos analisados manualmente (3) corresponde ao número de arquivos do tipo documento que passaram pelo filtro de arquivos conhecidos. Os arquivos do tipo documento englobam os seguintes tipos: 'XML', 'Hypertext Document', 'Plain Text Document', 'Microsoft Word 97 Document', 'Unicode Text Document', 'Acrobat Portable Document Format (PDF)', 'Microsoft Word 2000 Document'. Os experimentos são restritos a esses tipos devido ao fato da aplicação da tarefa de REM estar associada à linguagem natural, ou seja, a textos não estruturados. Ressalta-se que esses arquivos correspondem a 94% do total de arquivos extraídos das mídias, cujos tipos incluem também planilhas eletrônicas, *emails* e bases de dados, que não estão representados na Tabela 3. O item (4) corresponde à quantidade de arquivos selecionados manualmente pelas equipes de análise das operações, ou seja, são os arquivos que,

TABELA 3: Cenário Utilizado nos Experimentos

	Operação A	*Operação B*	**Totais**
1. Quantidade de mídias analisadas	258	50	308
2. Tempo investido no processamento das mídias	2 meses	1 mês	-
3. Quantidade de arquivos analisados	1.259.093	658.465	1.917.558
4. Quantidade de arquivos selecionados	3.473(0,276%)	243(0,037%)	3.716(0,194%)
5. Tempo investido na análise	6 meses	4 meses	-

TABELA 4: Resultado dos Experimentos (Sistema Rembrandt)

		Operação A	*Operação B*	Totais
Arquivos analisados	1. Avaliados pelo Rembrandt	1.259.093	658.465	1.917.558
	2. Etiquetados pelo Rembrandt	108.520(8,6%)	55792(8,5%)	164312(8,6%)
Arquivos selecionados	3. Avaliados pelo Rembrandt	3.473	243	3716
	4. Etiquetados pelo Rembrandt	3.411(98%)	210(91%)	3.621(97%)

depois de analisados, foram marcados como "relevantes". E o item (5) reflete o tempo investido na análise manual destes arquivos.

Os experimentos propostos se resumem à obtenção de dois resultados: o percentual de arquivos **analisados** manualmente que contêm EM e o percentual de arquivos **selecionados** manualmente que contêm EM.

No presente trabalho, o sistema Rembrandt foi utilizado para a tarefa de REM. O Rembrandt foi instalado em um microcomputador com processador Intel Core 2 de 2,33GHz e 2GB de memória de acesso randômico. Foi necessária, entretanto, a realização das seguintes alterações neste aplicativo: (1) desativação da consulta remota via web à Wikipédia e ativação da consulta local (DBPEDIA), (2) criação de estrutura de *loop* para possibilitar a etiquetação de vários arquivos contidos em um diretório através de uma única execução do programa e (3) desativação da etiquetação de vários tipos de EMs suportados pelo Rembrandt, mantendo somente os tipos pessoa e organização. Essas adaptações melhoraram o tempo de execução do aplicativo e não interferiram na interpretação dos resultados.

D. Etapas dos Experimentos

A primeira etapa dos experimentos se concentrou na aplicação do Rembrandt sobre os 1.917.558 arquivos do tipo documento contidos nas mídias apreendidas, ou seja, sobre todos os arquivos do tipo documento que foram analisados manualmente nas operações, com o objetivo de identificar, dentre eles, o percentual de arquivos que contêm pelo menos uma EM do tipo pessoa ou organização no seu texto. Ao término essa etapa, identificou-se que somente 164.312 (8,6%) arquivos foram etiquetados com pelo menos uma entidade do tipo pessoa ou organização.

A segunda etapa se caracterizou pela execução da mesma tarefa, porém aplicada sobre o conjunto dos 3.716 arquivos de documento selecionados manualmente pelas equipes de análise das operações, ou seja, arquivos considerados relevantes para a investigação. A aplicação do Rembrandt sobre este conjunto resultou em 3.621 arquivos (97%) etiquetados com a entidade pessoa ou organização.

A Tabela 4 apresenta os resultados obtidos nos experimentos após aplicação do sistema Rembrandt.

Ainda associado à segunda etapa dos experimentos, analisou-se individualmente os 95 (3.716 menos 3.621) arquivos que foram selecionados manualmente pelas equipes de análise e não foram etiquetados pelo Rembrandt. Desses arquivos, 86 (90,5%) contêm a entidade pessoa ou organização, sendo que a sua não etiquetação pelo Rembrandt está associada às características dos seus conteúdos, quais sejam: 50 arquivos contendo texto estruturado ou semi-estruturado, compostos por tabelas, listagens ou cópias de *email*; 21 arquivos contendo imagem de textos digitalizados, que podem ser resolvidos por OCR; e 15 arquivos protegidos por senha, o que impede a extração textual do seu conteúdo[5]. Por fim, a análise dos 9 arquivos (9,5% de 95) que não contêm a entidade pessoa ou organização mostrou que 4 desses arquivos estavam corrompidos e os outros 5 arquivos continham 2 listas de produtos de informática, 2 manuais de instalação de máquinas e 1 planilha de medicamentos, o que permite a conclusão de que somente 5 dos 95 arquivos não etiquetados não contêm entidades mencionadas do tipo pessoa ou organização, o que representa 0,13% de todos os 3.716 arquivos selecionados manualmente pelas equipes de análise das investigações.

Os resultados associados à análise individual dos 95 arquivos selecionados manualmente pelas equipes de análise e não etiquetados pelo Rembrandt são apresentados na Tabela 5.

5. Análise dos Resultados

Os resultados obtidos nos experimentos permitiram duas constatações relevantes (vide Tabela 4):

TABELA 5: Resultado da Análise dos 95 Arquivos Relevantes Não Etiquetados Pelo Sistema Rembrandt

95 arquivos selecionados pela equipe de análise e não etiquetados pelo Rembrandt	contêm EM: 86(90,5%)	texto estruturado ou semi: 50(53%)
		imagem digitalizada: 21(22%)
		protegido por senha: 15(16%)
	não contém EM: 9(9,5%)	corrompido: 4(4%)
		texto não contém EM: 5(5%)

5 Através de ferramentas forenses, foram recuperadas as senhas dos arquivos protegidos, o que permitiu a identificação da existência de entidades nos seus conteúdos

(i) Do conjunto de arquivos analisados manualmente, somente cerca de 8,6% contêm nomes de pessoa ou organização no seu texto;
(ii) Do conjunto de arquivos julgados relevantes pelas equipes de análise das investigações, cerca de 97% contêm nomes de pessoa ou organização no seu texto.

Esta análise revela que pesquisas relacionadas à tarefa de REM são aplicáveis ao contexto forense. A automatização do REM contribui para a otimização do processo de análise de mídias apreendidas em operações policiais, e isso foi comprovado nos experimentos ao reduzir o quantitativo de arquivos suspeitos que necessitam de análise manual em 91,4%. Além disso, os números revelam também que essa redução implica em menos de 3% de falsos-negativos, ou seja, de arquivos ignorados (pois não contêm a EM pessoa ou organização no seu texto) que seriam julgados relevantes caso fossem analisados manualmente.

Entende-se então que, com a aplicação do REM automatizado na etapa da extração do conteúdo das mídias, o conjunto (filtrado) dos arquivos a serem analisados manualmente, que contém cerca de 97% dos arquivos relevantes para a investigação, é suficiente para a formação da convicção da autoridade policial que chefia a investigação, quanto às informações que se pretende obter das mídias. Entretanto, como foi identificado o risco de 3% de perda de arquivos relevantes durante a aplicação do sistema de REM, propõe-se a utilização do REM como um mecanismo de priorização na etapa de análise de arquivos, ou seja, que não seja descartada a análise dos arquivos sem o filtro do REM nos casos em que esse filtro resultar em informações insuficientes para a autoridade policial. Esta proposta está representada no fluxograma da Figura 2.

É importante ressaltar que esses percentuais refletem somente as etiquetações resultantes da aplicação do sistema Rembrandt. Conforme mostrado no final do capítulo 4, dos 95 arquivos relevantes que não foram etiquetados pelo Rembrandt, 86 contêm EM do tipo pessoa ou organização, 5 não contêm e 4 são arquivos corrompidos. Entretanto, dos 86 arquivos que contêm EM, 50 possuem textos estruturados ou semi-estruturados no seu conteúdo, o que os torna incompatíveis com os sistemas clássicos de REM. Incorporando-se então os resultados da Tabela 5 da Seção 4 à análise dos arquivos relevantes e excluindo-se desta os 50 arquivos incompatíveis retrocitados e os 4 arquivos corrompidos, obtém-se que 99,9% dos arquivos relevantes contêm EM do tipo pessoa ou organização (vide Tabela 6), o que robustece ainda mais a idéia da aplicação do REM automatizado como forma de otimização do processo de análise de mídias.

Outra contribuição positiva do REM automatizado é a revelação dos nomes de pessoas e organizações contidos nos arquivos das mídias apreendidas. Esta informação, representada na caixa "*Relação de EM reconhecidas nas mídias*" da Figura 2, é de grande relevância para as equipes de análise e coordenação da investigação, pois pode ratificar levantamentos prévios de nomes, facilitar a identificação de novos nomes suspeitos e revelar vínculos com nomes envolvidos em outras investigações, além de permitir direcionar o processo de análise manual das mídias quanto à priorização dos alvos.

Também foram analisados eventuais problemas relacionados à velocidade de execução do sistema Rembrandt, que examinou, em média, 3.000 arquivos por hora, o que corresponde a cerca de 500.000 arquivos por semana com processamento dedicado. Constata-se que, em relação ao tempo investido nas etapas de extração dos arquivos e análise manual dos mesmos, os resultados obtidos favorecem à utilização desta ferramenta. Considerando-se, por exemplo, que a *operação A*, utilizada nos experimentos, demandou 2 meses para as etapas de duplicação das mídias,

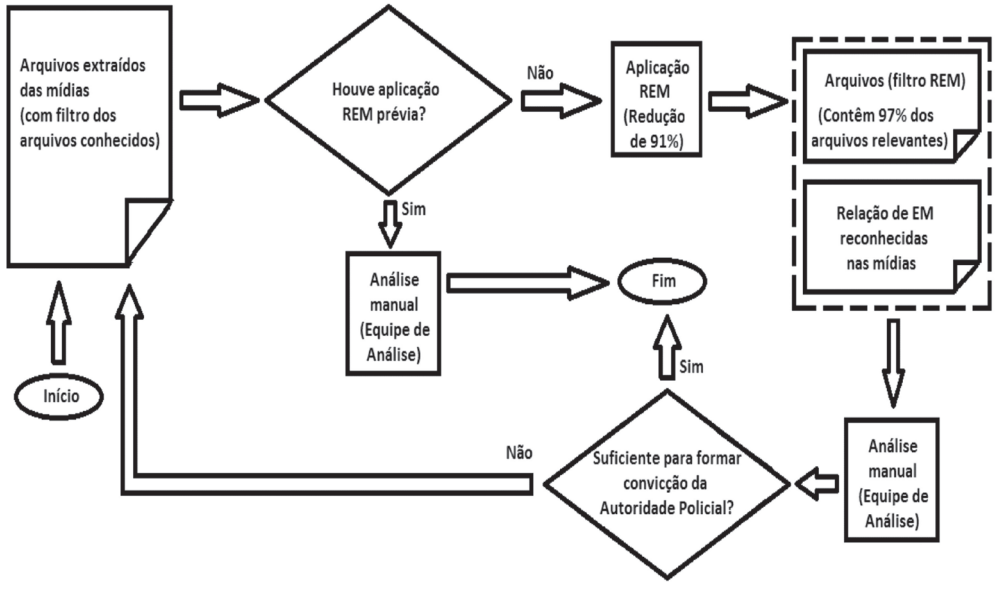

Figura 2 - Fluxograma das etapas de extração e análise do conteúdo

extração do seu conteúdo e disponibilização dos 1.259.093 arquivos, e mais 6 meses para a etapa de análise manual dos arquivos, projeta-se que, com a utilização do REM automatizado, o tempo investido na etapa de extração dos arquivos aumentaria em cerca de 17 dias (1.917.558 arquivos avaliados à velocidade de 3.000/hora), o que leva à conclusão de que qualquer redução maior ou igual a 9,4% no tempo de análise manual, o que equivale a 17 dias, resulta em ganho de tempo no processo de análise como um todo. Apesar de não ter sido possível mensurar esse tempo, espera-se que o tempo necessário para a análise manual de cerca de 108.282 arquivos (8,6% de 1.259.093) seja bastante inferior aos 6 meses que foram investidos na análise dos 1.259.093 arquivos da *operação A* (ver Tabela 3 da Seção 4). Se há uma redução de 91,4% no número de arquivos analisados, então espera-se haver redução superior a 9,4% no tempo de análise.

TABELA 6: Resultado da Análise dos Arquivos Relevantes

Total	3.716 – 54 = 3.662
Contêm EM pessoa ou organização	3.621 + 21 + 15 = 3.657 (99,9%)
Não contêm EM pessoa ou organização	5 (0,1%)

6. Conclusão e Trabalhos Futuros

A inserção da tarefa de automatização do REM no processo de exame de mídias apreendidas em operações policiais contribui para a redução da quantidade de arquivos pendentes de análise manual, a redução do tempo investido na análise dessas mídias e para a revelação de informações relevantes para a equipe de investigação.

Os experimentos se basearam na utilização do Rembrandt, um sistema de REM baseado em regras manuais, e mostraram que, em média, somente 8,6% dos arquivos do tipo documento contidos nas mídias apreendidas fazem referência a nomes de pessoas ou organizações e que 99,9% dos arquivos julgados relevantes no processo de análise manual estão contidos nesse conjunto de arquivos, ou seja, é possível utilizar a automatização do REM como filtro de arquivos suspeitos com risco de falsos-negativos na ordem de 0,1%. Esse risco representa a possibilidade de serem ignorados arquivos relevantes para a investigação. O fato desse filtro representar redução quantitativa de 91,4% dos arquivos a serem analisados manualmente sugere considerável ganho de tempo no processo de análise. Além disso, como resultado do REM, são revelados automaticamente à equipe de investigação os nomes das pessoas e organizações mencionados nas mídias, informação essa de grande relevância para o investigador.

Como complemento ao presente trabalho, propõe-se um estudo sobre a viabilidade de criação de um modelo de REM híbrido, composto por um algoritmo baseado em regras manuais e um algoritmo baseado em um modelo probabilístico, com o objetivo de utilizar o aprendizado de máquina para treinar o reconhecimento de entidades em diferentes tipos de estruturas textuais, com desempenho superior e velocidade de execução não inferior ao sistema Rembrandt.

Além disso, é válida também a adaptação do sistema Rembrandt através da inserção de *gazetteers* de pessoas e organizações, com o objetivo de identificar o impacto no desempenho, em especial quanto ao reconhecimento de entidades em textos estruturados e semi-estruturados que, apesar de serem minoria quando comparados aos não estruturados, estão frequentemente presentes em mídias apreendidas pelas polícias.

Outra proposta é a criação de uma ontologia capaz de identificar vínculos entre nomes de pessoas e organizações reconhecidos em diferentes mídias apreendidas.

Por fim, propõe-se que o presente estudo seja repetido com base em operações policiais associadas a outros tipos penais, como crimes previdenciários, fazendários ou ambientais, que envolvam apreensão de grande quantidade de mídias computacionais, que estudos sejam feitos no sentido de se analisar o desempenho qualitativo da etiquetação do sistema Rembrandt em textos forenses, uma vez que o presente trabalho concentrou-se primordialmente na sua análise quantitativa, e que seja realizada uma análise para identificar a relação de proporção entre a quantidade de mídias analisadas manualmente e o tempo médio investido nessa análise em operações policiais.

Referências

[1] P.M.S. Eleutério e M.P. Machado, "Desvendando a Computação Forense." São Paulo/SP : Novatec Editora, 2011. ISBN: 978-85-7522-260-7

[2] B.T. Todorovic, et al. "Named entity recognition and classification using context Hidden Markov Model." In the 9th Symposium on Neural Network Applications in Electrical Engineering (NEUREL 2008), 2008, pp. 43-46

[3] S.M. Weiss, et al. "Text Mining: predictive methods for analyzing unstructured information." New York : Springer Science+Business Media Inc., 2005. ISBN 0-387-95433-3

[4] D. Nadeau, S. Sekine, "A survey of named entity recognition and classification." Lingvisticae Investigationes. 1, 2007, Vol. 30, pp. 3-26.

[5] L.F. Rau, "Extracting Company Names from Text." In IEEE Conference on Artificial Intelligence Applications. 1991.

[6] R. Feldman, J. Sanger, "The text mining handbook: advanced approaches analyzing advanced unstructured data." New York : CAMBRIDGE UNIVERSITY PRESS, 2007. ISBN-13 978-0-521-83657-9.

[7] E. Riloff e R. Jones. "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping." Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99). 1999, pp. 474-479.

[8] M. Pasca, et al. "Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge." In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06). 21, 2006, Vol. 2.

[9] Y. Shinyama e S. Sekine, "Named Entity Discovery Using Comparable News Articles." In Proceedings of the 20th international conference on Computational (COLING '04). 20, 2004.

[10] S. Miller, J. Guinness e A. Zamanian. "Name tagging with word clusters and discriminative training." In Proceedings of HLT. 2004, pp. 337-342.

[11] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition." In *Readings in speech recognition*, Alex Waibel and Kai-Fu Lee (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 267-296..

[12] A. McCallum, D. Freitag, e F. Pereira, "Maximum entropy Markov models for information extraction and segmentation." Proceedings of the 7th International Conference on Machine Learning (ICML 2000). 2000, pp. 591–598.

[13] J. Lafferty, A. McCallum e F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001). 18, 2001.

[14] N. Cardoso, "REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto." In Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Cristina Mota & Diana Santos (eds.), Linguateca, 2008, pp. 195-211.

[15] C.M. Santos e D. Santos, "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. " Linguateca, 2008. ISBN: 978-989-20-1656-6.

[16] "The DBpedia Knowledge Base"; Disponível em http://dbpedia.org. Acessado em 15/08/2011

[17] A.L. Louis e A.P. Engelbrecht, "Unsupervised discovery of relations for analysis of textual data." Digital Investigation. 2011, Vol. 7, pp. 154-171.

[18] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, S. Vaithyanathan, "Domain Adaptation of Rule-based Annotators for Named-Entity Recognition Tasks." EMNLP. 2010.

[19] H. Guo, et al. "Domain adaptation with latent semantic association for named entity recognition." In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09). 2009.

[20] K.F. Edward, V. Baryamureeba, G.D. Pauw, "Towards Domain Independent Named Entity Recognition." International Journal of Computing and ICT Research. 2, 2008, Vol. 2, pp. 84-95.

[21] E.F. Meulder, T.K. Sang e F. De, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." Proceedings of CoNLL-2003. 2003, pp. 142-147.