Extraction of Relevant Entities in Textual Documents. Modeling Intelligence Maps

Isnard Thomas Martins[1] **and Edgard T. Martins**[2]
[1]Administração, Universidade Estácio de Sá, Rio de Janeiro, R.J, Brazil; 55 21 98748873
[2]Coordenação Ergonomia, UFPE, Recife, Pernambuco, Brazil ; 5581 41413131

Abstract - *Police investigation activities are conducted on historical, records and occurrences reports in informations data bases structured and not structured, where are extracted knowledge to elucidate authorship, interests, crime dynamics and objects involved in criminal activities. The complexity inherent in not structured informations sources and police records, the restrictions associated time and resources available to authorship analysis assume a critical condition in the elucidation of the crimes. As a result, the automatic extraction of knowledge in criminal databases assume great importance in generating intelligence maps and research activities. Criminal reports, source of the research and criminal knowledge, usually present themselves in not structured format, inaccurate in content and difficult to analyze*

Keywords: Datamining, Shortest-path Algorithms, Law Enforcement Analysys

1 INTRODUCTION

In a criminal report are transcribed all the facts, people, circumstances and relationships that characterized the police report. Because the need to capture the occurrence with maximum accuracy of reality, investigative studies devoted to authorship analysis are based on historical police, whose transcription, aims to provide freedom for communication through narratives, often developed by people of different features, both comunicante and witnesses, as police responsible for the occurrence record

The discovery of traces who carrying the criminal to the crime scene or their activities to achieve the offense becomes one of the fundamental problems encountered in the authorship's investigations of criminal analisys [1].

The police reports are dense, complex and characterized by scattered data. The not structured police report, done on free text style, offers better support for research evidence and relationships, imposing, however complexity, time and personnel allocated in analysis.

Gradually criminal activities are growing in sophistication, technology and planning. Employing resources and technologically advanced methods, criminals are connecting in social networks and using modern communication systems such as Internet, wifi, telephone and radio. Crimes fraud related, drug trafficking, money laundering and gangs's investigation, an activity of complex and sophisticated police intelligence [1].

Analysts and researchers produce criminal knowledge from information sought as a bulky bases of police reports [2] . The biggest challenge for analysts, researchers and police intelligence departments is to achieve efficiency and accuracy in the face of growing demand for raw data found in criminal intelligence bases [3] . The potential demand of police reports without authorship identified in countries with high volumes of criminal incidents, as verified in Brazil, is the characterization of the impunity.

Xu and Chen [4] reported that the successful construction of concept maps, based on documents and police reports analyzed in the criminal investigation, depends on extensive use of techniques to automate the most of the data mining operations and identification of useful entities such as people, places, events, organizations or objects involved in the researched historical, whose identification will contribute to clarifying the facts and understanding of the relationships investigated. However, the efficiency to be obtained in the extraction of useful entities for analysis depends essentially on the cleaning of the data entered in the extraction process

Oriented systems to support the extraction of relevant entities and intelligence activities in unstructured documents are intended to increase the speed in the analysis and reduction of researchers's time in the preparation activities of relational maps extracted from criminal reports.

Expert systems, however, require many hours in the preparation of information for procedural support, such as dictionaries, references and linguistic rules [5].

According Baluja et al [5], specialized systems in data mining are specific in their goals, such as tax evasion, money laundering, census or commercial research, those systems require many hours in the preparation of information for procedural support, and its operation have restricted rules. Dictionaries are originally built in language English, Japanese, Portuguese, etc, for this reason, they are restricted, starting to understand only a particular branch of knowledge [6] .It would not be possible, for example, migrate directly rules and special specifications used by an extractor system of criminal organizations operated in the United States for the extraction of entities in criminal bulletins in Brazil [5]. An extensive list of previous procedures for preparing the extractor systems should be deployed, to meet the restrictions and formalities language as a basis for nominal entities extraction in any language. Some authors mention the use of automated tools

to accelerate the preparation of these lists in support of expert systems.

No matter the method used, the expert system development will require a training method in order to obtain the best efficiency in the analysis and recovery of entities that can meet the principles and necessary rules for entities extraction in textual documents.

2 NETWORK RELATIONSHIPS AND CLASSIFICATION OF ENTITIES

Actors, objects, events and relationships captured in the police reports can provide valuable historical evidence and provide crime patterns, usually hidden in the reports and police reports. A synthesis drawing format and simulated scenarios is known in police and criminal context map, research map or intelligence map, topological representations of the crime scene [1]. The represented social network in the intelligence maps must integrate all parts of research activities and identify possible connections between actors and potentially involved events [4]. The map or criminal relationship tree is often treated as a network [7], provides valuable evidence of extracted crime patterns in investigations, resulting in accumulated knowledge in the analysis of relationships between entities involved in the criminal offense.

The intelligence maps allow inter-relate several useful entities extracted from texts treated, establishing an association value between these various entities. These associations are of great relevance to the criminal investigation. The resulting structure aims to provide aid for research and pattern recognition in criminal offenses [8]. Systems for intelligence analysis are applied for tracking of individuals and organizations involved in criminal activities such trafficking, terrorism and fraud. [9]. The collect of entities is based on previous patterns, becoming simplified for not requiring the understanding of the text by the system entities extraction operator [10]. Extractors Systems are also employed to identify patterns such as dates, times, numeric expressions and email addresses.

The arc value associated between entities in an intelligence map expresses the intensity, on which the entities are closer or distant from each other. The value assigned to the mapped relationships helps the visibility of existing links, identifies involvement of the actors present in the scene and produces knowledge to generate conclusions and reports on the facts of the cases analyzed [1].

Various distance types are used to calculate specific measures of distance between entities in the vector space. Some measures are used applying simple Euclidean distance while others are used applying the square Euclidean distance or absolute Euclidean distance, where the distance is the sum of the square of distances, avoiding the square root calculation, which offers advantages for computational speed in applied calculations [6].

According to textual entities standardization procedures developed in the MUC-7, Seventh Message Understanding Conference and Second Multilingual Entity Task [11], nominal entities are defined as proper names, numbers, people, local references, schedules dates, percentages and monetary values . The scenario selected for extraction site is built according with the events in which the entities are participating, whose definition of domain and importance depend on the purpose of the analysis and presence of the entity on the analyzed text. Entities assist the police investigation and provide the necessary allowance for identifying patterns related to the "modus operandi" of the crime [2].

For each extracted entity must be associated attributes residing in a specific parameter table representing the properties and characteristics. This table is called Elements Table (TE), whose purpose is to qualify the identification of each entity, beyond the simple name reference.

Selecting the domain of entities and structure of the model elements table depends on the size of each entity in the specific scenario in which it is inserted. The MUC-7 provides that such definitions depend subjectively system of the author, however its accuracy is linked to the wealth of the parameters associated with the entities, serving to increase its effectiveness with the users [11].

Chen and Lynch [12] cite knowledge bases specialized on automatic creation of thematic dictionaries and algorithms for generating statistical coefficients related to frequency ratios between concepts extracted from text documents . Furthermore, the available literature provides in various academic segments developed studies in both fields, information science and cognitive studies, confirming the creation of specific areas for scientific dictionaries, such as medicine, engineering and business that resulted in the creation of efficient thesaurus, robust potentially available as basis for information retrieval applications [13] . Chen and Lynch [12] cite the specific steps for the implementation of preparatory processing for recovery of useful entities :

- Development of the list of objects and documents
- Filter the objects
- Indexing
- Analysis of co -occurrence (frequency studies)
- Recovery of associations

The crime often involves organized gangs, whose members are connected by various associations such as common interests, friendship, neighborhood or criminal association . This relationship, similarly can be treated as a network in which such criminals can perform various activities and illegal actions . Textual documents such as police reports and others are rich in information, from which you can extract entities, converting them into a topological representation connected by their criminal relationship and their criminal activities. The base of knowledge, represented by a semantic network in which nodes are words, phrases and concepts and connections represent the semantic relationship between nodes [12]. The system for capturing concepts consists of rules or procedures operated, on according the knowledge base, similar to the decisions rules from experts patterns

3- MODELING INTELLIGENCE MAPS

Martins [1] presented an expert system to capture entities from free texts and intelligence maps modelling, called Anaphora that apply as example for illustration of an automatic extractor system model. Used by some intelligence agencies in Brazil, the Anaphora system integrates the major phases of an extractor project: construction of thematic dictionary, training the useful entities and network construction. The final output provides network representation on a Graph format that examines the strongest connections between nodes of entities network, using one shortest path algorithm.

The first phase of Anaphora system involves the construction of a specialized dictionary in radicals, using policial languagem, which will serve as keywords for later extraction of knowledge.

Cognates are words derived from the same root. Is the irreducible element common to all the words of the same family [14], also called lexical family [15]. The element is irreducible when it can no longer be reduced.

Some examples of families who have the same root:

- Moon, moonlit, moonligth
- Sea, salt, sailor
- Crime, criminal, criminology
- Love, loving
- friendly, friend, friendship

Monteiro [14] mentions that the internal structure of word consists of words with associated elements which represent the minimum elements of language emissions containing individual significance. The radical in its original form is the root, the minimum element of a family of words and irreducible and common element of this family of words. The root is the element from where the first morphological operation, so their root shall be different from the radical. The radicals may have one or more affixes derivative. Thus the same word can have several radicals. The neighbor word can offer three degrees radical :

I. neig
II . neighbor
III. neighborhood

The meaning is essential in the root concept that carries the semantic word load . The suffixe particularizes the generic meaning of the root (smaller part of the word) in a series of derivatives. The more affixes (words derived by prefixes and suffixes), less general will be the meaning of the word. The roots are minimal morphological construction of a core, which may be free or attached [15]. The high degree of radical will include all derived words. We can conclude the following assertions, mutually inverse :

- A higher volume of derived words requires a higher degree of extracted radicals, as close as possible to its original form, that is, the smallest format .
- It is important to keep the meaning of the radical in its minimal primitive form, keeping the association of meaning with the derived word [14], avoiding multiple interpretations or derivations with the family of other extracted derivatives [16].

The specialist pre-processing dictionary is based on the principle of extracting radicals, derived from training sets. The extracted lists are then used as keywords, in an interactive way for obtaining derived words. The resulting structure is refined, obtaining also stop words, which are words with little meaning in the analyzed text. Dictionaries are generated from classified information, contained in documents belonging to the application domain, which are converted into specialized structures through continuous training [12]. Research in dictionaries is an important source for retrieval information systems [16]. Dictionaries include selected information in documents, databases or manually, generated by experts who provide guidance for extraction algorithms such as keywords and critical debugging routines. The resulting structures, prepared in automatic form or manually allow extracting keywords from textual documents with little or no manual interference [1].

Automatic dictionaries or semi- automatic dictionaries can be generated from processed radicals, by algorithms that serve as keywords for knowledge extraction. The initial structures are subsequently processed through specialized training performed by learning machine [18]. Dictionaries generated manually can be obtained by combining public domain words (geographic information, professions, usual acronyms, common names, titles etc). The resulting structures, prepared in automatic form or manually allow extracting keywords, from textual documents with little or no manual interference [1].

The construction of specialist dictionary is based on linguistic studies, which provides the basic guidelines for the learning algorithm of extraction model. An ordered set of phonemes is considered a word when has a meaning . The words include the names (nouns, adjectives and adverbs and verbs [14].

Table 1 shows an example of extracted key from historical words that belong to the domain of a collection of documents that were investigated.

TABLE 1.
Example of extracted keywords from police texts in portuguese language [29]

KEYWORDS	FREQUENCY
dp	56
vulgo	47
inq	44
favela	36
traficante	28
policia	25
dinamica	22
dre	22
mandado	21
traficantes	20
cv	19
prisão	16
comando vermelho	15
preso	15
policiais	15

The second stage of the extractor system involves the extraction of useful entities, modeling of relationships and calculations of co-occurrences between the extracted entities. We used the Hauck algorithm [19] adapted from the method developed by Chen and Lynch [12], an algorithm for treatment of co-occurrences on data mining routines. The algorithm sets relative levels of importance between the extracted entities on researched documents, calculating weights for relationships between each pair of extracted entity. The weights are calculated based on statistics frequencies corresponding to a value for the co- related associations. The Hauck algorithm calculates the relative weight of each entity, on each document of collection.

Originally, the co-occurrences analysis approach was devoted to the automatic generation of dictionaries based on textual documents, reflecting the frequency with which two sentences appeared together in the same document . The modern statistical approach defines co-occurrence as the frequency among entities, based on lexical statistics. Assuming that two entities appear together in a same document, there may be an association and involvement between these entities . A co-occurrence with non-zero value indicates the weight of the rapprochement between entities, so strongly associated so higher be the value represented by their co -occurrence [20] .

Statistics co-occurrence are related to the useful words found in the analyzed text. The co-occurrence concept is based on the proposition of Chen and Lynch [12] for calculating the statistics co-occurrences between extracted words on text documents. Xu and Chen [4] define co-occurrence or associative relationship as the relationship between a pair of entities, when they are found together on one document

Step 1.1

The Hauck algorithm [2] calculates the relative weight of each entity, in each document of the collection (D_{ij}; entity - document).

Equation [1] shows the calculation of the co-occurrence D_{ij} in each document of the collection, given by :

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right) \qquad (1)$$

Where:

i - represents each document of the collection

j - represents each entity found on document i

N - number of collection's documents

Df_j - Number of documents in which j is present

Tf_{ij}. number of occurrences of J entity in each document in which j entity was located

W_j - factor of importance of j entity on extraction process (relative value that can assume, greater or lesser degree, according to importance of entity on extraction process)

Step 1.2

The algorithm calculates the co-occurrence between each pair of entities found together in documents in the collection (W_{jk} and W_{kj}), using a asymmetric function, shown in (2) and (3)

$$W_{jk} = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(k) \qquad (2)$$

$$W_{kj} = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times WeightingFactor(j) \qquad (3)$$

Where:

j - represents the first entity of each examined pair in document i

k - represents the seond entity of each examined pair in document i

W_{ij} - represents the final weight among entity j and k entity

W_{kj} - represents the final weight calculated between the entity k and entity j

d_{ij} - weight of the entity j, calculated as shown in step 2 of this topic

Df_{jk} - represents the number of documents in the collection N, where the entities j and k are revealed together.

D_{ijk} - Hauck algorithm calculates the combined weight of each pair of entities found together in each document in the collection.

Equation (4) shows the calculation of the combined weight of pair jk on document i and (5) shows the calculation of the combined weight of the kj on document i. The difference between these functions is the factor of relative importance (Wi / Wi) in the calculation of the function

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right) \qquad (4)$$

$$d_{ikj} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_k\right) \qquad (5)$$

Where:

WeightingFactor j e WeightingFactor j - Influence factor that reduces the value of the very common generic instances, reducing the value of their respective influences. WeightingFactor is obtained through the calculation shown in (6) and (7):

$$WeightingFactor(j) = \frac{\log\frac{N}{df_j}}{\log N} \qquad (6)$$

$$WeightingFactor(k) = \frac{\log \frac{N}{df_k}}{\log N} \qquad (7)$$

Where:

WeightingFactor $_k$ - reduction factor for the entity k (6)

WeightingFactor $_j$ - reduction factor for the entity k j (7)

The algorithm proposed by Hauck et al. [2] produces asymmetric values for associations between entities , however penalizes with a final reduction factor the value of words most often found in the studied texts . This reduction factor is used in order to minimize the importance of extracted generic terms.

Figure 1 shows an example of useful entities extracted from the analyzed domain. The columns present the results of calculations processed by each stage of the frequency and approximation algorithm.

	Total TF i j	DF j	D i j	W j	WeightFac j
duque de caxias	22	17	60,687	0,9	5,924
comando vermel	17	7	61,979	0,9	4,779
centro	11	8	38,635	0,9	4,477
eder gonca	10	5	39,823	0,9	3,912
rio de janeiro	10	9	33,945	0,9	4,500
mangueira	10	6	38,000	0,9	4,094
ramos	9	7	32,812	0,9	4,143
naldo medeiro	9	3	40,438	0,9	3,296
acari	9	2	44,087	0,9	2,890
campo grande	8	6	30,400	0,9	3,871
rocinha	8	6	30,400	0,9	3,871
luiz costa	8	4	33,644	0,9	3,466
madureira	8	5	31,858	0,9	3,689
nova iguaçu	7	5	27,876	0,9	3,555

Fig 1. Função WeightingFactor [29]

In the third phase of construction, the entities are extracted from textual documents and organized in a structure indexed, by document , keeping data available for access of the algorithm . Each pair of extracted entities is analyzed, according to frequency computed in each document , subsequently consolidated in accordance with the totals processed throughout collection. The product obtained by this method comprises a weighted array of relationships where each array element represents a entity and the extracted weights computed represents the importance of these relationships . The depiction of a network in matrix format provides a means to describe a graph , eliminating the existence of a list of nodes and arcs to build or a representative drawing of a network [21] .

Let N be a weighted matrix with m rows and n columns, corresponding to each of the extracted entities (vertices) . Let n_{ij} the representation of the element in the ith row and column jth [21] . Each element n_{ij} of the array corresponds to an arc (i, j) and refers to a association value between entities i and j if these entities are present in the extracted relationship. The resulting structure is so called matriz of Criminal Relationships . As a result of extraction, the Anaphora system produces the following results:

Step 2.1

Construction of a temporary structure containing the totalization of frequencies and the temporary variables, such as strengthening's factor for each pair of entity extracted

Step 2.2

Construction of a temporary array containing consolidated frequencies for each pair of extracted entities calculated by the co-occurrences algorithm. The raw results processed by the algorithm will be in accordance with the frequencies computed, between each pair of entities.

Step 2.3

Constructing a normalized final results matrix containing co-occurrences for each pair of entities.
The resulting structure of the normalized matrix corresponds to a directed graph , whose vertices are represented by nominal extracted entities and their arcs are represented by the results of the entity - entity
The structure is stored in an auxiliary file (setting file) generated for further analysis , completing the cycle developed by Anaphora System [29].

The file for analysis consists of three types of information, corresponding to each pair of entities associates:

- Numeric code of connected vertices;
- Association value, calculated by Anaphora system [29];
- Reference name of connected entities.

4 - ANALYSIS OF THE STRONGEST LINKS BETWEEN ENTITIES IN THE INTELLIGENCE MAP

From associations's matrix, is then constructed a second array that will contain the reverse tracking of possible paths between pairs of entities present on the graph . This structure called Reach matrix is based on the reverse access tracking of Dijkstra algorithm, optimizes the use of the intelligence's Map because provides the pre-calculation of all possible paths between related entities , thus avoiding the time spent processing the strongest associations in research activity [1] .

Each cell of the Reach Matrix represents an entity of reference identified by column number where it is located, indicating the associations of the reverse path between pairs of entities line / column of the matrix .

The Dijkstra algorithm [22] is the classic method for minimum cost of path calculation from a source node to all other nodes of a weighted graph [4] , assuming that the graph contain no negative arcs [23] . Dijkstra lent his studies for the more efficient algorithms and solutions to shortest path, the

principles of which were based on the original structures of Dijkstra algorithm. Xu and Chen [4] mention that in a criminal network represented by a directed graph, the value of a connection, which can assume a number between zero and one, can be treated as an probability measure for approximation calculation between two directly connected entities. As a general rule, the joint probability of occurrence of a group of mutually independent events is equal to the product of the individual probabilities of occurrence of these same events. If two nodes in a graph are only connected through a sequence of intermediate connections , the association value between the two nodes is equal to the product of intermediate weights. The strongest association between a pair of nodes is represented by the largest product of the weights between the nodes.

Since the shortest path algorithms recognizes the shorter distances between graph nodes, where the value of the arcs indicates the weight of the associations , the representation of the strongest connections, after application of the shortest path algorithm will not guarantee that the strongest associations will be identified [1] .

Xu and Chen [4] proposes a heuristic search for transformation by the shortest path to the location of the strongest connections in a directed graph , using the logarithmic transformation : $l = -\ln(w) \quad 0 < w \leq 1$

Where:

l is the weight of the connection in the new transformed graph

w is the corresponding weight in the original graph

With the proposed transformation are obtained the following axioms [4] :

- All the connections in the transformed graph are non negative numbers.:
- Since: $0 < w \leq 1$, thus $\ln(w) \leq 0$, wich sugest that: $-\ln(w) \geq 0$;
- The lowest values of the arches in the transformed graph correspond to higher values in the original graph
- If $l1 < l2$, then $-\ln(w1) < -\ln(w2)$ or $\ln(w1) > \ln(w2)$.
- Since ln (w) is a monotonic increasing, is follows that $w1 > w2$;
- The shortest paths using the sum of the weights values of the transformed graph, correspond to larger Arches products using the original network.

After the modeling of a associations matrix , represented by a directed graph and constructed as a relationship between the product extracted entities from text files , the stronger links between the graph's entities are calculated. The associated weights with arcs provide calculation probabilities between each pair of entities , with the possible path and the representative value of the strongest chances of approximation between entities [1].

The calculated results are presented in a matrix of associations, as shown in Figure 2.

	luiz fernando d	comando verm	marcos marinh	marcos antoni
luiz fernando	0	100,0%	81,3%	43,8%
comando ver	86,9%	0	71,0%	38,0%
marcos mari	66,0%	66,0%	0	35,5%
marcos anto	31,2%	31,0%	31,2%	0
ederson jose	8,0%	8,0%	7,0%	4,0%
amigos dos a	22,1%	22,1%	18,0%	10,0%
celso luiz rod	34,0%	39,2%	28,0%	15,0%

Fig 2. Relationships matrix containing precalculated associations [29]

5 CONCLUSIONS AND PROBLEMS IN THE EXTRACTION OF ENTITIES FROM THE POLICE REPORT

Various errors may occur during data mining, particularly when treating extractions in textual documents, which can make inconsistent modeling, contribute for distortions or inconclusive results [1] .

Kohonen [24] quotes that are frequently occurring errors when converting text to entities, producing inaccuracies in the calculation of distances.

Goldberg & Senator [25] reported that several information bases have inconsistencies, incomplete data or multiple identifications for the same extracted references.

Han & Kamber [26] reported that many inconsistencies may occur in information bases such as violation of restrictions or cases of redundancies that can be removed by integrating of data routines. Some attributes can take different names in heterogeneous information bases. The errors and inconsistencies can be deleted manually through external references, imposing dependencies between attributes, correction parameters or creation of criticism against violation of restrictions .

May occur with some functional inadequacies applied in entities extraction model in the surveyed bases. The data handlers must identify, debug or discard incompatible documents and routines to reduce errors and deviations that would minimize the expected results in the extraction routines [27] .

Xu and Chen [28] point to problems observed in the extraction of entities related to incomplete data, incorrect or inconsistent in searched data records.

Incomplete data - criminal networks operate in stealth mode or hidden . Criminals minimize interactions, in order to not attract police attention. The data captured may become incompletes, causing the loss of connections between the nodes and loss of integrity on the the network structure.

Incorrect data - inaccuracies relating to identification, physical or addresses can result in errors in the transcription of information that are generated intentionally by the criminals themselves, aiming to confuse the police investigations . Criminals lie about their addresses and their identities when captured or investigated, which can

introduces ambiguities and inaccuracies in the bases of police records.

Inconsistency - information on criminals can come from multiple sources entries simultaneously, feeding the police records inputs, not necessarily consistently. The criminal may appear in historical police records with multiple identifications, presenting itself as different individuals, causing inaccuracies in the processed queries.

Working with criminal historic records and other intelligence policial sources used for investigation, the policial activity lacks intelligence tools to aid and elucidate crimes and discover knowledge in databases occurrences of policials records.

The shared effort between preventive police action and investigative policing is a complex scenario for operational planning decisions. Prevent and investigate deal with the same variables represented by the police force and should be shared as cooperative resources, but competitors in the search for the final results

The efficiency and effectiveness of police investigation request the use of automated tools to cover the entire research cycle, comprising the record of the occurrence, analysis and extraction of the police historical knowledge.

REFERENCES

[1] MARTINS, I. Descoberta de Conhecimento em Históricos Criminais: Algoritmos e Sistemas. Tese de Doutorado PUC-Rio Dep Engenharia Industrial, 2009

[2] HAUCK R.V., H. Atabakhsh, P. Ongvasith, H. Gupta, H. Chen, Using coplink to analyze criminal-justice data, IEEE Computer 35 (3) 30– 37, 2002.

[3] CHEN, H., Chung, W., Xu. J., Wang, G., Qin, Y., and Chau, M., Crime Data Mining: A General Framework and Some Examples, IEEE Computer, 37(4), 50-56, 2004

[4] XU Jennifer, Chen H., Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks . Decision Support System 38 (2004) 473-487

[5] BALUJA, V. Mittal, and R. Sukthankar, Applying Machine Learning for High Performance Named-Entity Extraction. Pacific Association for Computational Linguistics, 1999.

[6] VIDAL L. A. Carvalho. DataMining, a mineração de Dados no Marketink, Medicina, Economia e Administração. Editora Ciência Moderna, Rio de Janeiro, 2005.

[7] McANDREW D, The structural analysis of criminal networks, in: D. Canter, L. Alison (Eds.), The Social Psychology of Crime: Groups, Teams, and Networks, Offender Profiling Series, Aldershot, Dartmouth, vol. III, 1999.]

[8] CHAU M., J. Xu, H. Chen, Extracting meaningful entities from police narrative reports, Proceedings of the National Conference on Digital Government Research (Los Angeles, CA), 2002, pp. 271– 275.

[9] LEE R, Automatic information extraction from documents: a tool for intelligence and law enforcement analysts, Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis, AAAI Press, Menlo Park, CA, 1998.

[10] WITTEN I. H., Zane Bray, Malika Mahoui, W.J. Teahan. Using language models for generic entity extraction. Teahan Computer Science University of Waikato Hamilton, New Zealand, 1999.

[11] CHINCHOR Nancy MUC-7 Overview, seventh Message Understanding Conference and the Second Multilingual Entity Task, CA, EEUU, 1999. Search in august, 2007, http://www.muc.saic.com/proceedings/muc_7_proceedings/overview.html

[12] CHEN, H., and K. J. Lynch. Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man and Cybernetics, 22(5):885-902, September/October 1992.

[13] CHEN, H., Martinez, J., Tobun D. Ng, and Bruce R. Schatz. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. Journal of the American Society for Information Science, 1997

[14] MONTEIRO J. Lemos. Morfologia Portuguesa. Pontes Editora. São Paulo 2002

[15] LAROCA M. N. C., Manual de Morfologia do Português - 4a Edição. Editora Pontes, São Paulo, 2005.

[16] HULL D.A. Stemming Algorithms: A Case Study for Detailed Evaluation. In: Journal of the American Society for Information Science 47(1), 1996, p. 70-84.

[17] ANTIQUEIRA L, Nunes M. Oliveira Jr, Costa L. F. Modelando Textos como Redes Complexas. Encontro para o Processamento Computacional da Língua Portuguesa. PROPOR, MG, 2003.

[18] PORTER M.F. The Porter Stemming Algorithm, Computer Laboratory, Cambridge (England) 1997, revisado em Jan 2006. Disponível em http://tartarus.org/~martin/PorterStemmer/, Consulta em setembro 2007.

[19] HAUCK R.V., H. ATABAKHSH, P. ONGVASITH, H. GUPTA, H. CHEN, Using coplink to analyze criminal-justice data, IEEE Computer 35 (1.5.3) 30– 37, 2002.

[20] SCHROEDER J, J. XU, H. CHEN, M, CHAU. Automated criminal link analysis based on domain knowledge. Journal of the American Society for Information Science and Technology Volume 58, # 6 , 2007.

[21] EVANS J, E.MINIEKA. Optimization Algorithms for Networks and Graphs, Marcel Dekker, New York, 1992.

[22] DIJKSTRA E. A note on two problems in connection with graphs, Numerische Mathematik 1 269– 271, 1959.

[23] BOAVENTURA Netto. Teoria e Modelos de Grafos. Editora Edgard Blücher Ltda, São Paulo, 1979.

[24] KOHONEN, T. Self-Organization Maps, Springer-Verlag, Berlin. 1997.

[25] GOLDBERG, H.G., SENATOR, T.E. Restructuring databases for knowledge discovery by consolidation and link formation, Proceedings of the First International Conference on Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1995.

[26] HAN J., M, KAMBER, Data Mining. Concepts and Techniques. Morgan Kaufman San Francisco, USA, 2001.

[27] LIFSCHITZ S.,CÔRTES S., PORCARO R. Mineração de Dados, Funcionalidades, Técnicas e Abordagens. ISSN 0103-9741, PUC-Rio 2002

[28] XU J., CHEN H. Criminal Network Analysis and Visualization: A Data Mining Perspective . Available in http://ai.bpa.arizona.edu/COPLINK/publications/crimenet/Xu_CACM.doc Search in March , 2008

[29] Sistema ANAPHORA, Projeto para Extração e Análise em Histõricos Policiais. Isnard Martins, PUC-Rio, 2008