Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 357–363
ISBN 978-83-60810-27-9
ISSN 1896-7094

Using Self Organizing Map to Cluster Arabic Crime Documents

Meshrif Alruily, Aladdin Ayesh, Abdulsamad Al-Marghilani
Software Technology Research Laboratory
De Montfort University
The Gateway, Leicester, LE1 9BH UK
Email: meshrif,aayesh,abduls@dmu.ac.uk

***Abstract*—This paper presents a system that combines two text mining techniques; information extraction and clustering. A rule-based approach is used to perform the information extraction task, based on the dependency relation between some intransitive verbs and prepositions. This relationship helps in extracting types of crime from documents within the crime domain. With regard to the clustering task, the Self Organizing Map (SOM) is used to cluster Arabic crime documents based on crime types. This work is then validated through experiments, the results of which show that the techniques developed here are promising.**

I. Introduction

ONE OF the most important motivations for creating this system is because of the lack of Arabic systems in general, and the crime domain in particular. Furthermore, the crime domain has been chosen as an application area because of its social importance. Information extraction aims to extract specific, predefined entities from text. In this current research, one of our aims is to develop a system that is able to recognize crime phrases in a given document in order to extract types of crime. Feldman and Sanger [1] have stated that entities, such as peoples names, organizations names, locations, attributes (e.g. age of a person), and crime type can all be extracted. According to Toral and Munoz [2] and Collins and Singer [3], there are two types of evidence that help in identifying entities. Internal evidence can be deduced from the sentence that contains the entity by noticing a particular sequence of words. On the other hand, external evidence is gained from the context. In the first stage, the rule based approach (based on syntactical analysis) is adopted. In the second stage, the extracted types of crime are then used to by Self Organizing Map (SOM) in order to perform clustering. So instead of processing the whole content of each document, the rule based approach is used to guide the SOM to cluster the data by extracting important or meaningful patterns. According to Flexer [4] SOM is a very common tool for clustering and visualizing high dimensional data spaces. More details about SOM will be presented in one of the following sections.

The rest of the paper is organized as follows. In section II, a background and a review of the related work are given. The crime domain is described in section III. Domain analysis is presented in section IV. Section V presents the proposed clustering system. Section VI provides the results of the experiments and an evaluation of their performance. Finally, the conclusion of this work is presented in section VII.

II. Background

According to Michailidis [5], the Message Understanding Conference (MUC-6) introduced Named Entity Recognition (NER) in 1995, and it has been used in many different text-based applications, such as information extraction, question and answering, information retrieval and text classification [5], [6]. The approaches that are used to identify named entities are as follows [5]:

- Hand-craft rules, known as linguistic approaches.
- Machine learning approaches.
- Hybrid, which combines hand-craft recognition grammar with machine learning methods.

Alruily et al. [7] have developed a software package to extract crime information from texts. Their approach is based on a dictionary that is created manually. On the other hand, the task of automatically constructing lists of entities has been studied by many researchers. Riloff [8] has developed a program called AutoSlog that automatically constructs a domain-specific dictionary for information extraction. Nadeau et al. [9] used an approach that has two aspects: retrieve pages with seed, and a web page wrapper in order to build or generate gazetteers. Toral and Munoz [2] have proposed an approach to automatically build and maintain dictionaries of proper nouns using a noun hierarchy and a POS tagger. Also, Chau et al. [10] used named entity extraction techniques to identify meaningful entities from police narrative reports. To the authors' knowledge, there is no work available regarding the SOM technique applied to Arabic texts within the crime context for crime analysis. On the other hand, in the English language, Chen et al. [11] have developed a system based on SOM to cluster and visualize crime-related data.

III. Crime Domain

As far as we know, no information systems have been applied to the crime domain in the Arabic language. Therefore, the major problem we faced was the lack of data. This issue has been solved by compiling news articles on crime incidents, published by some Arabic newspapers. The reason for exploiting newspapers is that it is difficult to obtain official reports or narrative reports from police stations, especially in

978-83-60810-27-9/09/$25.00 © 2010 IEEE

Arab countries. The news articles contain the information that the police reports would normally include. So, collecting these data was an important step in gaining a better understanding of the crime domain and the nature of the data that our system will deal with.

A. Types of Crime

The crime domain includes several types of crime, starting from civic crimes, such as drinking and driving, to international crimes, for instance, homicide by terrorists [12]. In this research, types of crime have been categorized into six main types, as in Table I.

TABLE I
TYPES OF CRIME.

English	Arabic	Pronunciation
Theft	السرقة	Alsareqah
Fraud	الاحتيال	alehtial
Drug and alcohol smuggling	تهريب المخدرات	Tahreeb almokhdrat
Magic and sorcery	السحر والشعوذه	Alseher walshaawathah
Sex crime	الجنس	Aljens
Violent crime	العنف	Alonf

As previously mentioned, the aim is to extract crime types from Arabic news articles for work that will be presented later. This extracted information is considered as "keywords"; which are significant words that are able to give clues about the main idea of the document or article. Consequently, keywords play an important role in many text mining tasks, such as clustering, summarization and document retrieval. In this research, the extracted words will be treated as keywords to guide the Self Organizing Map (SOM) in order to perform clustering. So, the previous task of keyword extraction is an important process that must be considered carefully. Accordingly, the crime domain must be studied and its characteristics explored in order to find the appropriate extraction algorithm.

B. Event Description

Most Arabic newspaper reports have the same structure, with respect to writing style, in the crime domain. Most journalists or reporters start with a sentence containing the name of the police station that has investigated the crime, followed by a description of that crime. The crime description is about the type of crime committed and the type of criminal. Following these, details of any victims and other information are described. The reason for this formulaic approach is because they are dealing with a specialist domain that has its own language, and which can be called here the language of crime. According to Almas and Ahmad [13], each special language has a limited vocabulary and idiosyncratic syntactic structures. The journalists who work with these restricted languages seem to share the same words and same sentence structures. In other words, the usage of their words has the same behaviour. Fig. 1 shows a description for a theft crime

أعلنت شرطة الشارقة أن عناصر قسم التحريات بإدارة البحث الجنائي تمكنت من إلقاء القبض على ثلاثة أشخاص من جنسية عربية تورطوا في عملية سرقة 100 ألف درهم.

Alsharqa police have announced that its detectives have arrested three Arab people involved in stolen 100,000 Dirham

Fig. 1. Article Excerpt from Alriyadh Newspaper.

published by Alriyadh newspaper [14]. The location of the committed crime can be deduced from the following pattern: "شرطة الشَارقة / shurtat alsharqa / Alsharga police" . Furthermore, other related crime information can be extracted from the text, such as the nationalities of the people involved in the crime, e.g. "جنسية عربية / jnsyt arabiat / Arabic nationality". The type of crime can also be extracted from this pattern: "تورطوَا في سرقة / tawaratwo fi sareekat / involved in stealing". Also, the following is another example that has been taken from Aljazirah newspaper articles [15]:

أطاحت شرطة جنوب الدمام بلص بنغالي (33) سنة **متخصص في سرقة** اللوحات الدعائية من الشوارع العامة بالدمام.

South Dammam police arrested Bengali thief ,(33) years, **specialized in** the **theft** of advertisements boards from public streets in Dammam.

Fig. 2. Article Excerpt from Aljazirah Newspaper.

It can be seen from the above example in Fig. 2 that the same writing style is followed. The name of the police station, which carries the name of the location, is stated first. The nationality of the criminal is also mentioned, and the type of offense is described. However, the crime type information is what we want to concentrate on in this research.

IV. DOMAIN ANALYSIS

A. Arabic Language

In this research, the Arabic language is studied; it is one of the Semitic languages and it is used in over 21 Arab countries. This language consists of 29 letters that can be used to form a word. Moreover, other languages, such as Farsi and Urdu use mostly Arabic characters [16]. From the sentence construction point of view, Arabic words can be divided into three classes: nouns, verbs and particles [17], [18], [19]. When working with the Arabic language, some other important characteristics need to be taken into account [20]:

1) A character may have up to three different forms, each form corresponds to the position of that character in the word (beginning, middle or end), such as letter "ع / Ayn " in Table II.

TABLE II
POSITION OF THE CHARACTER IN THE WORD.

End	Middle	Beginning
ـع	ـعـ	ع

2) Arabic does not have capital letters; this characteristic represents a considerable obstacle to the NER task because in other languages capital letters represent a very important feature.
3) Finally, it is a language with a very complex morphology because it is highly inflectional.

A linguistic study of Arabic words and grammatical structures will be required before extracting the most appropriate structures for common Arabic sentence forms within the crime domain. Hence, the use of the linguistic internal structures of Arabic sentences will allow us to identify logical sequences of words. As previously mentioned, the structure of Arabic can comprise of three categories : noun (اسم), verb (فعل) and particle (حرف).

- Noun
  This category in Arabic comprehends any word that describes a thing, idea or person. It can be divided into two types: primitive and derivative. Primitive nouns are nouns that are not derived. Derivative nouns are nouns that are derived from verbs, other nouns, and particles. The Arabic nouns are inflected for gender (masculine and feminine) and number (singular, dual and plural). Also Nouns are either definite, which starts with the article "ال / al" or indefinite, which has no "ال" article at the beginning of the noun. Moreover affixes and clitics, such as some prepositions, conjunctions and possessive pronouns, can be attached to them. The clitic is subdivided into proclitic (located at the beginning of a stem) and enclitics (located at the end of a stem). For example, Table III shows the different morphological segments for the word "وبدرجَاتهم" which means "and by their grades".

TABLE III
EXAMPLE FOR MORPHOLOGICAL SEGMENTS.

	enclitic	affix	stem	proclitic	proclitic
Arabic	هم	ات	درج	ب	و
pronunciation	hm	at	drg	be	wa
Gloss	their	s	grade	by	and

- Verb
  This word type points out an event or action. Verbs are also inflected in terms of number (singular, plural, dual), gender (masculine, feminine), person (1st, 2nd, 3rd), voice (active and passive) and mood (subjunctive, indicative and jussive). Furthermore, from the tense point of view, the verb can be in the past, present or future.
- Particle
  This class includes prepositions, conjunctions, interrogative particles, exceptions, and interjections. In other words, it includes the words that are not nouns or verbs, and sometimes these words are called function words.

In the Arabic language, sentences are divided into two types, as follows:

- A nominal sentence, according to Hadj et al [19], a nominal sentence can start with a noun or a particle.
- A verbal sentence
  A verbal sentence can start with a verb or a particle. The verb is divided into two types: transitive and intransitive. With respect to transitive verbs, a sentence, e.g., "قطفت التفَاحة / qtft altfaht / I picked up the apple" contains one object or more as well as the subject. In other words, it takes more than one argument. On the other hand, a sentence that contains an intransitive verb has no object, and is composed of a verb and only one argument, e.g., "مرض خَالد / mrd Khaled / Khaled became ill". In some cases, some intransitive verbs can be converted into transitive verbs, for example, "ذهبت إِلَى دبي / thhbt ela dubai / I went to Dubai". In this example, the verb has a subject and a quasi sentence (the prepositional phrase) "إِلَى دبي / ela dubai / to Dubai" which is the complement of the sentence "ذهبت / thhbt / I went"; this help in determining the meaning of the whole sentence. Most Arab linguists state that most of the intransitive verbs can not refer to the object of the sentence but they can be strengthened by some prepositions, which are called transitive prepositions, such as "البَاء / bi / because", "الَّام / allam / for", "من / mn / from", "علَى / ala / on", "في / fi / in", "إِلَى / ela / to", in order to refer to the object. These verbs are called "transitive verbs by preposition" [21], and they play an important role in achieving our goal.

B. Intransitive Verbs and Prepositions in the Crime Domain

As shown in the above section, our study of the crime domain corpus has led us to identify the characteristics of the language used. The first feature is that the past tense is used when describing crime incidents, whether the verbs describe the action of the crime itself or indicate a phrase that carries information about the crime type. Moreover, the modifier, sometimes called the qualifier, such as prepositional phrases and adjectives, are used for describing the type of offence. However, in this research, we concentrate on using the characteristics of some intransitive verbs and their propositions in order to recognize the type of crime. In other words, the correlation or dependency relationship between some intransitive verbs and some prepositions will be exploited. The Arabic language has approximately fourteen prepositions, most of which are short. Most of them are formed from three letters, such as "علَى / ala" or from two, such as "في / fi", but they can be formed with only one Arabic letter, such as "ل / li" or "ب / bi". The structure of a prepositional phrase is usually

composed of two parts: preposition and noun-phrase [22]. For example, ”الولد في البيت / alwalad fi albyt” means ”the boy in the house”. In this example, the preposition is ”في / fi” and the noun is ”البيت / albyt”. Table IV shows a list of Arabic prepositions with their English translations. Only the most frequently used prepositions in crime domain texts are presented but more illustration regarding the listed prepositions is given in order to clarify them semantically.

TABLE IV
LIST OF PREPOSITION IN ARABIC LANGUAGE.

Arabic Preposition	Pronunciation	English Translation
علَى	Ala	on
في	Fi	in
إِلَى	Ela	to
ب	Bi	because
ل	Li	for

The preposition ”ب / bi” has different meanings, so based on a whole sentence, its meaning can be inferred. In this research context, the preposition ”ب / bi” means ”because”. In other words, it answers the question ”why”. Also, the preposition ”في / fi” represents the preposition ”in” in the English language, for example, ”تورط في قتل / twrat fi qtl / involved in killing” and ”متخصّص في قتل / motakasys fi qtl / specialized in killing”. With regard to the preposition ”ل / li”, it has many meanings but in the current domain being studied it again means ’because’; put simply, it justifies the reason, for example, ” هو ابلغ الشرطة لتعرضه لّلسرقة / howa ablqa alshortat lita’arrudh lilsareekat / he reported to the police because he was robbed”. Thus, in this context ”ب / bi” and ”ل / li” carry very similar meanings. Sometimes the preposition ”إِلَى / ela” takes the place of the preposition ”ل” (e.g. ”تعرض إِلَى السرقة / ta’arrada ela lilsareekat / s/he was exposed to theft”). Thus, these prepositions can be utilized in many different ways. Additionally, they can work as a link between some verbs and nouns, whether they are adjacent or not. Moreover, the meanings of sentences that contain some specific verbs cannot be identified without prepositional phrases. Based on this, the main forms where a type of crime is located in the text can be identified. Fig. 3 depicts this case. Thus, in order to mark the crime phrases in the text, the system looks for the verbs in the text and (their prepositions), and the type of crime should not be more than three words away from the preposition. Therefore, only three words are extracted after the preposition. As a result, in order to recognize and extract types of crime, the list of these intransitive verbs and their prepositions (that indicate the patterns of crime type) must be defined beforehand.

From the above illustrations, it can be seen that there are strong correlations between some verbs and prepositions.

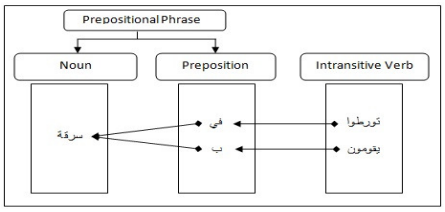

Fig. 3. Types of crime in Prepositional Phrases.

These relationships are considered as key elements, and form an important part in accomplishing this work. Accordingly, these can lead to discovering a local grammar for crime type patterns, which will help in their extraction. Fig. 4 shows the various frequent patterns used in news articles for describing the type of crime committed.

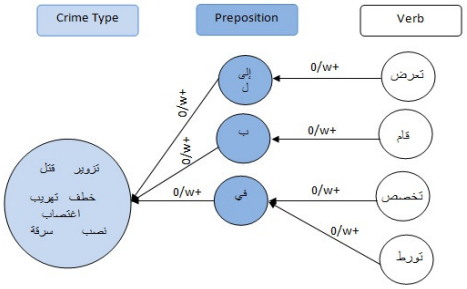

Fig. 4. Crime types Local Grammar.

V. CLUSTERING SYSTEM

Fig. 5 shows the proposed framework, which consists of stages, divided as follows:

- Normalization
  In this stage, some letters that perform the same function, and that can be written in different forms in a word, are normalized. For example, the letters ”أ” , ”إ” and ”آ” are converted into the letter ا.
- Information Extraction
  The system here is focused on prepositional phrases. Hence, the dependency relationship between prepositions and intransitive verbs is exploited. The advantage of this method is that there is no need for an annotated corpus, whether manual or automatic. In other words, the system has no linguistic components, such as PoS taggers or shallow parsers. Instead, lists of intransitive verbs and their prepositions are provided to the system in order to extract the desired patterns. Fig. 4 shows the list of verbs, with their appropriate prepositions, that are used in this research. In other words, it describes the local grammar for extracting types of crime. To sum up, in the processing phase, the system looks for words in a text that match the words in the verb list. When a match occurs, the next step is to search for a preposition that always comes with the verb. After that, the three words

that immediately follow the preposition are extracted to represent the whole content of the document. Moreover, the crime type should appear within these three words.

- Stemming
  Through this process, the stem of a word can be obtained by eliminating the word's affixes. After stemming, the content of file is automatically converted into numbers by giving each word of interest a unique number. Also, during this process the stopwords are removed.
- Clustering and Visualization
  For the clustering process, the Self Organizing Map (SOM) has been chosen to cluster documents that were generated by the information extraction process, based on their similarity. The SOM technique is popular and widely used for clustering and visualizing high dimensional data spaces. According to Eyassu and Gamback [23] SOM has many different structures but the most popular architecture is composed of two layers of processing units; the input layer and the output layer. These two layers are fully interconnected. The idea behind SOM is that it performs mapping for similar input vectors to similar areas on the output grid. The following is the SOM algorithm:
  1) Initialize weight randomly
  2) Initialize neighbourhood ratio
  3) Set input pattern
  4) Calculate Euclidean distance
  5) Find the winner neuron (smaller distance)
  6) Update winner and neighbour weight neurons
  7) Repeat Steps 3 to 7 until the convergence criterion is satisfied

  As can be seen, this algorithm is iterative. The first step is to randomly initialize the weight vectors of the output map. At each iteration (training), a sample vector is randomly chosen from the input data. This phase is called the learning process or competitive learning. Through competitive learning, the Euclidean distance is calculated for choosing the Best Matching Unit (BMU). The wining neuron or BMU is the one most similar to the input pattern. That is, its weight is close to the input pattern. As a result, all neurons on the output layer enter into a competition with each other. The neuron on the output layer that has the smallest distance to the input pattern is the winner. Once the winning neuron has been selected, its weight and the weights of its neighbour are both updated in order to make them more similar to the input pattern. This process is repeated with other documents until accurate results are found or the maximum number of iterations (epochs) are reached.

VI. EXPERIMENTAL RESULT AND EVALUATION

A. Corpus

Text mining research relies on the availability of a suitable corpus. As a result, many corpora have been created for specific purposes. For this research, two corpora have been collected from different Arabic newspapers published in different Arabic countries, such as Alriyadh, Aljazeera and Okaz from Saudi Arabia, Elkhabar newspaper issues from Algeria, Addustour from Jordan and Ahram from Egypt. The first corpus contains 26 documents and the other includes 24 documents. The reason for compiling corpora from different resources is to avoid the problem of bias, which could occur if the system is tested on documents that were collected from only one country.

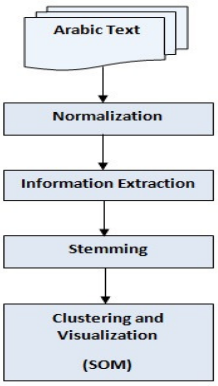

Fig. 5. System Architecture.

B. Experiments

Two experiments have been carried out on 26 documents in order to show how the information extraction process guides the Self Organizing Map (SOM) to gain accurate results. The size of this corpus is 25.3 KB. The SOM has been trained on the same documents, obtaining good results; the best learning rate, radius and iteration are 0.5, 8 and 1000, respectively.

- First Experiment
  In the first experiment, the information extraction process is used. So, instead of processing the whole of each document's content, the extracted patterns from each document are used by SOM to perform clustering. As a result, the size of the corpus has become 1.56 KB. Only 21 unique words represent the 26 text documents and they comprise 57 tokens. Fig. 6 shows the results of the first experiment after achieving the clustering process.

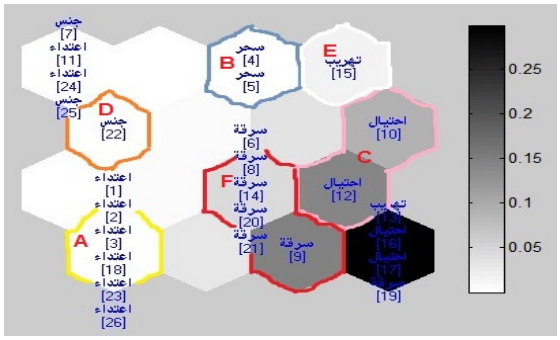

Fig. 6. Clustering Result for First Experiment (A: Violante, B: Magic, C: Fraud, D: Sex, E: Smuggling, F: Theft).

- Second Experiment
  This experiment does not rely on the information extraction process. So the whole content of each file is stemmed and used for the clustering process. The 26 textual files are represented by 32 unique words. These 32 words form 228 tokens. Also, the results of this experiment can be seen in Fig. 7.

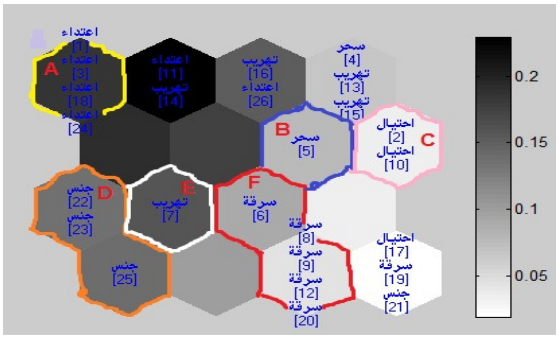

Fig. 7. Clustering Result for Second Experiment (A: Violante, B: Magic, C: Fraud, D: Sex, E: Smuggling, F: Theft).

In order to examine the proposed system and the SOM (at the same learning rate, radius and iteration value) on a new and untouched corpus other experiments have been also carried out. This new corpus contains 24 documents, and its size is 28.5 KB. The third and fourth experiments are as follows:

- Third experiment
  This is exactly like the first experiment because the information extraction process is used. The size of the whole corpus after the extraction became 1.83 KB. The number of unique words is 13 which, form 44 tokens. These 44 tokens are then used by SOM to cluster the documents. The clustering results can be seen in Fig. 8.

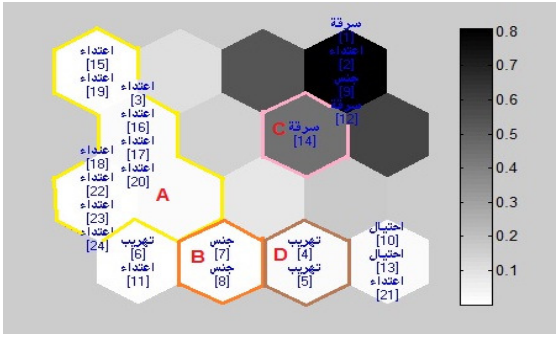

Fig. 8. Clustering Result for Third Experiment (A: Violante, B: Sex, C: Theft, D: Smuggling).

- Fourth experiment
  This experiment is as the second experiment. It uses the whole content of each document because no information extraction technique is used. 23 unique words represent the tested documents, and they form 235 tokens. As a result, 235 tokens are used by SOM to cluster the documents in this experiment. Fig. 9 shows the clustering outcome of SOM.

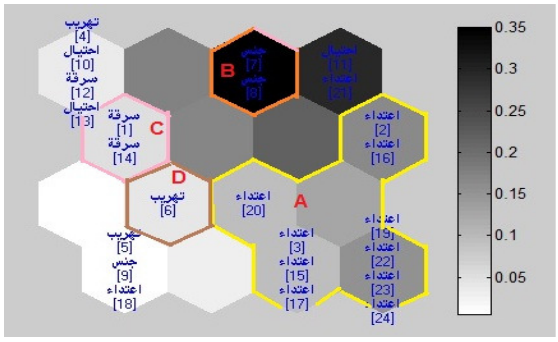

Fig. 9. Clustering Result for Fourth Experiment (A: Violante, B: Sex, C: Theft, D: Smuggling).

C. Result Analysis

The average distance between each data vector and its BMU (quantization error) in the first experiment is 0.686, and in the second experiment it is 0.949. So the performance of the SOM is better when using the information extraction process. As is well-known, SOM clusters documents based on their similarity. Moreover, each document is treated as a vector of words. So the number of a word's frequency that occurs in a file affects its clustering and sometimes this leads to a wrong cluster. For example, File 23 from Group A in Fig. 6 is clustered as a violent crime, which is true, but in Fig. 7 it is clustered as a sex crime. The reason for this clustering mistake in Experiment 2 is because of the phrase "سعودي الجنسية / saudi aljensyat", which means "Saudi nationality", occurred many times when talking about the criminal. So the word "الجنسية / aljensyat" means "nationality" in English, but after stemming this word by removing its affixes (the article "ال" and the letter "ة") the word becomes "جنس / jens", which means "sex"; this affected its clustering in the second experiment. Also, another clustering mistake occurred in Experiment 2 for File 14. This file belongs to Group F in the first experiment; i.e. it is clustered as a theft crime, but in the second experiment it has been labelled as "smuggling", and is far from its Group (F) in Fig. 7. The reason for this being wrongly clustered is because of the word "هرب / hrb", which appears many times in the file; it has two meanings in the Arabic language, and it means "flee" or "smuggle" in English. The file is totally about a theft crime but in the crime description the phrase "thieves fled / هرب اللّصوص " is stated many times in different ways, and "هرب / hrb" in our context means "smuggle". As a result, Group F in Experiment 2 does not contain the file number 14. Also, Fig. 10 shows that the letters attached to the vertical axis represent categories of crime types, as in Table V, and the numbers that are underneath the horizontal axis refer to the files. This graph clarifies how our system labels files with the right category names using the extracted patterns, and it worked better than

when using the full content. In Experiment 1, the file numbers 12, 16 and 17 have incorrect crime type category names. On the other hand, in Experiment 2 file numbers 2, 7, 14, 17, 21 and 23 also have the wrong category names. With regards to Experiments 3 and 4, the quantization errors are 0.6 and 0.888, respectively.

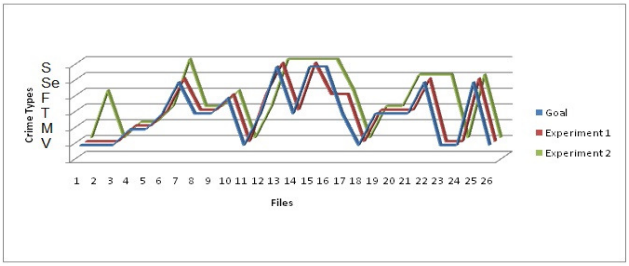

Fig. 10. Result of Labeling Crime Documents.

TABLE V
MEANING OF VERTICAL AXIS

Letter	V	M	T	F	Se	S
Crime Type	Violent	Magic	Theft	Fraud	Sex	Smuggling

VII. CONCLUSION

In this paper, we have developed a system for clustering textual documents containing information about different types of crimes. The Self Organizing Map (SOM) technique was chosen to perform the clustering. Moreover, the rule-based approach, based on intransitive verbs and propositions, was used to help in obtaining good clustering results. Also, a comparison study was carried out through four experiments in order to show the effects of the rule-based method on the Self Organizing Map. The results show that although SOM used fewer tokens in Experiments 1 and 3, it was able to achieve clustering that was as good as or better than in Experiments 2 and 4. Therefore, it can be confirmed that the SOM technique has been improved in terms of its performance. The reason behind this remarkable achievement is because of the system's ability to extract keywords based on syntactic principles, and this led directly to the improved clustering results.

REFERENCES

[1] R. Feldman and J. Sanger, "Information extraction," in *The Text Mining Handbook: Advanced Approches in Analyzing Unstructured Data: Cambridge university*, no. 94-130, 2006.

[2] A. Toral and R. Munoz, "A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia," 2006.

[3] M. Collins and Y. Singer, "Unsupervised models for named entity classification," in *In Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 100–110.

[4] A. Flexer, "On the use of self-organizing maps for clustering and vi," *Intelligent Da*, vol. 5, no. 5, pp. 371–384, 2001.

[5] I. Michailidis, K. Diamantaras, S. Vasileiadis, and Y. Frre, "Greek named entity recognition using support vector machines, maximum entropy and onetime," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006, pp. 47–52.

[6] P. Srikanth and K. N. Murthy, "Named entity recognition for telugu," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, 2008, pp. 41–50.

[7] M. Alruily, A. Ayesh, and H. Zedan, "Crime type document classification from arabic corpus," in *Second International Conference on Developments in eSystems Engineering*. Los Alamitos, CA, USA: IEEE Computer Society, 2009, pp. 153–159.

[8] E. Riloff, "Automatic constructing a dictionary for information extraction tasks," in *The Eleventh National Conference on Artifical Inteligence*, 1993, pp. 811–816.

[9] D. Nadeau, P. D. Turney, and S. Matwin, "Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity," 2006, pp. 266–277.

[10] M. Chau, J. J. Xu, and H. Chen, "Extracting meaningful entities from police narrative reports," in *dg.o '02: Proceedings of the 2002 annual national conference on Digital government research*. Digital Government Society of North America, 2002, pp. 1–5.

[11] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. OToole, M. Chau, T. Cushna, D. Casey, and Z. Huang, "Coplink: Visualization for crime analysis," in *dg.o 03: proceedings of the 2003 annual national conference on Digital government research*, 2003.

[12] P. Thongtae and S. Srisuk, "An analysis of data mining applications in crime domain," in *Proc. IEEE 8th International Conference on Computer and Information Technology Workshops CIT Workshops 2008*, 2008, pp. 122–126.

[13] Y. Almas and A. Kurshid, "Lolo: a system based on terminology for multilingual extraction," in *IEBeyondDoc 06: Proceeding of the Workdhop om Information Extraction Byeond The Document*, 2006, pp. 56–65.

[14] Alriyadh. Crimes articles. [Online]. Available: http://www.alriyadh.com/

[15] Aljazirah, "Crimes articles." [Online]. Available: http://www.al-jazirah.com/

[16] A. M. AL-SHATNAWI and K. OMAR, "Methods of arabic language baseline detection the state of art," *Arab Research Institute in Sciences and Engineering (ARISER)*, vol. 4, pp. 158–193, 2008.

[17] R. Al-Shalabi, G. Kanaan, B. Al-Sarayreh, K. Khanfer, A. Al-Ghonmein, H. Talhouni, and S. Al-Azazmeh, "Proper noun extracting algorithm for arabic language," in *International Conference on IT, Thailand*, 2009.

[18] S. KHOJA, "Apt: Arabic part-of-speech tagger," in *Proc. of the Student Workshop at NAACL*, 2001.

[19] Y. M. E. Hadj, I. Al-Sughayeir, and A. Al-Ansari, "Arabic part-of-speech tagging using the sentence structure," in *Proceeding of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt*, 2009, pp. 241–245.

[20] Y. Benajiba, P. Rosso, and J. Ruiz, "Anersys: An arabic named entity recognition system based on maximum entropy," in *CICLing*, 2007, pp. 143–153.

[21] معجم الأفعال المتعدية بحرف, موسى بن محمد الميّاني الأحمدي, بيروت, دار العلم للملايين, 1986.

[22] A. C. Satterthwait, "Computational research in arabic," *Mechanical Translation*, vol. 7, pp. 62–70, 1963.

[23] S. Eyassu and B. Gamback, "Classifying Amharic news text using self-organizing maps," *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pp. 71–78, 2005.