Use of Elliptic Curves in Term Discrimination*

Darnes Vilariño, David Pinto, Carlos Balderas,
Mireya Tovar, Beatriz Beltrán, and Sofia Paniagua

Faculty of Computer Science
Benemérita Universidad Autónoma de Puebla, Mexico
{darnes,dpinto,mtovar,bbeltran,sofia}@cs.buap.mx

Abstract. Detection of discriminant terms allow us to improve the performance of natural language processing systems. The goal is to be able to find the possible term contribution in a given corpus and, thereafter, to use the terms of high contribution for representing the corpus. In this paper we present various experiments that use elliptic curves with the purpose of discovering discriminant terms of a given textual corpus. Different experiments led us to use the mean and variance of the corpus terms for determining the parameters of a Weierstrass reduced equation (elliptic curve). We use the elliptic curves in order to graphically visualize the behavior of the corpus vocabulary. Thereafter, we use the elliptic curve parameters in order to cluster those terms that share characteristics. These clusters are then used as discriminant terms in order to represent the original document collection. Finally, we evaluated all these corpus representations in order to determine those terms that best discrimine each document.

1 Introduction

Term discrimination is a way to rank keywords of a given textual corpus [1]. The final aim of term discrimination is to support Natural Language Processing (NLP) tasks in order to improve the performance of their computational systems. Information retrieval, text classification, word sense disambiguation, summarization, are some examples of NLP tasks that may get benefit of a good term discrimination method [2].

We use the discriminant terms in order to represent the document with the hope of removing those terms that may introduce noise. Therefore, we may obtain a double benefit: on the one hand, we reduce the number of computational operations because of the corpus size reduction; on the other hand, we are expecting to increase the performance of the NLP system used in the task because we only consider to use the terms really involved in the characterization of the document [3].

* This work has been partially supported by the projects: CONACYT #106625, VIEP #VIAD-ING11-I, #PIAD-ING11-I, #BEMB-ING11-I, as well as by the PROMEP/103.5/09/4213 grant.

J.-F. Martínez-Trinidad et al. (Eds.): MCPR 2011, LNCS 6718, pp. 341–349, 2011.

Up to now, different methods for automatic term discrimination have been proposed. Perhaps one of the most successful approach is the well-known tf-idf term weighting schema which was proposed by Salton in the 1970's [4]. This model proposes a simple manner for representing documents of a collection by means of weighted vectors. Each document is represented as a vector whose entries are weights of the vocabulary terms obtained from a text collection. The problem associated with this approach is that in huge collections of documents, the dimension of the vector space can be of tens of thousands, leading to a number of computational calculations that may be prohibitive in practice.

Some other approaches for term discrimination exist in literature. For instance, in [5] it is presented a statistical analysis of some set of words without knowledge of the grammatical structure of the documents using the concept of entropy. The theory of testors is another approach that may be used for term discrimination [6]. A testor is a set of features which may be used to represent a dataset. Although this theory may be adequate for selecting terms in a collection, it lacks of algorithms for efficient calculation of the testor set. In fact, in [7] it was presented the fastest algorithm, which is not polinomial in complexity.

Even if there exist various approaches for finding discriminant terms in document collections, we consider that the problem of determining those terms that better represent the documents (with a maximum tradeoff of precision and recall) still an open problem. Therefore, we are encouraged to explore new mechanisms in the term discrimination field.

In this paper, we present diverse experiments with the purpose of investigating the usefulness of elliptic curves, a topic highly investigated in the cryptography field, in the term discrimination and document representation task.

The remaining of this paper is organized as follows. In Section 2 we present a brief description of theoretical issues of elliptic curves. Thereafter, we propose different models for document representation by stating the parameters of a reduced Weisrtrass equation that from now and forwards we will generally call as "elliptic curve". The evaluation of the different representations is given in Section 3. We use a corpus borrowed from the information retrieval field in order to perform those evaluations. Finally, in Section 4 the conclusions and findigs are given.

2 Use of Elliptic Curves in Term Discrimination

An elliptic curve is an equation $y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_5$, where x and y are variables, and $a_1, \cdots, a_5$ are constant elements of a field. Even if elliptic curves are important in mathematical areas such as number theory, they constitute a major area of current research and they find applications in some other areas such as cryptography [8].

The formal definition of an elliptic curve is fairly technical and requires some background in algebraic geometry. However, it is possible to describe some features of elliptic curves over the real numbers using only some concepts of algebra and geometry.

In this context, an elliptic curve is an smooth plane curve defined by an equation of the form:

$$y^2 = x^3 + ax + b, \tag{1}$$

where a and b are real numbers. The equation (1) is called a Weierstrass equation, and its discriminant must be different of zero in order to be non-singular; that is, its graph has no cusps or self-intersections. In Figure 1 we may see an example of an elliptic curve with parameters $a = 0.75$ and $b = 1.09$, that correspond to the mean and standard deviation of one term of the one of the eight corpus evaluated in this paper.

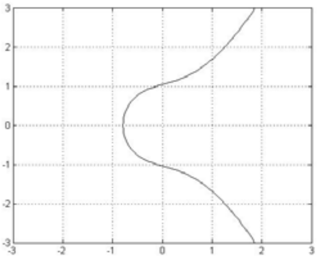


Fig. 1. An example of an elliptic curve with $a = 0.75$ and $b = 1.09$

An interesting feature of the elliptic curves is that these are parabolic curves centered on the x axis, when the parameters a and b are positive. Therefore, we may establish a distance measure between any pair of elliptic curves. In the context of NLP, we consider factible the use of elliptic curves for representing the documents. By having an appropriate set of parameters for elliptic curves would lead to have distance measures among the documents and, therefore, a similarity measure between any pair of documents. Thus, we consider important to investigate the adequate values for a and b in order to obtain an accurate representation of documents.

In this paper we propose three different approaches of values for the parameters of the elliptic curves, which we have named DR_1, DR_2 and DR_3. In the case of approaches DR_1 and DR_2, we have defined the function ascii(c_j), which is the ASCII code of the character c_j of term t ($t = \{c_1c_2c_3...c_{|t|}\}$):

DR_1 :

a is equal to $\sum_{j=1}^{|t|} \text{ascii}(c_j)$, where t is the most frequent term;

b is equal to $\sum_{j=1}^{|t|} \text{ascii}(c_j)$, where t is the less frequent term;

DR_2 :

a is equal to $\sum_{i=1}^{10} \sum_{j=1}^{|t_i|} \text{ascii}(c_{ij})$, where t_i is one of the 10 most frequent terms;

b is equal to $\sum_{i=1}^{10} \sum_{j=1}^{|t_i|} \text{ascii}(c_{ij})$, where t_i is one of the 10 less frequent terms;

DR_3 :

a_j is equal to the frequency mean of the corpus term t_j. In other words, given a corpus with n documents,

$$a_j = \bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} freq(t_j, d_i), \tag{2}$$

whereas $freq(t_j, d_i)$ is the frequency of the term t_j in the document d_i.

b_j is equal to the frequency standard deviation of the corpus term t_j. In other words, given a corpus with n documents,

$$b_j = \sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(freq(t_j, d_i) - \bar{x_j})^2}, \tag{3}$$

where $freq(t_j, d_i)$ is the frequency of the term t_j in the document d_i.

In the following section we show the obtained results after evaluating the above presented approaches in a document collection gathered for information retrieval purposes.

3 Experiments

The aim of the aforementioned document representation schemata is to detect discriminant terms. In order to visualize the appropriate representation of the documents, we present in this paper the elliptic curves of one document collection (see corpus C1 in Section 3.1). Each figure correspond to one approach proposed. The rest of the curves are also available, but due to space limitations these were not included in this paper.

In Figure 2 we may observe the DR_1 approach. As we may see, having considered only two terms for representing the documents lead us to have a very ambiguous representation schema. In this Figure is quite difficult to distinguish a clear division among the elliptic curves. The stepforward is to verify whether or not, adding more terms would improve the document representation. Figure 3 show a set of elliptic curves in which we have considered the 10 most and less frequent terms in order to represent each document. Again, we observe that the parameters do not assist for the correct representation of the document. We consider that, in particular, the second parameter (b =less frequent terms) is not helpful due to the high number of terms with frequency one in the vocabulary of the corpus.

In order to analyze the degree of discrimination that each term of the corpus of evaluation has in the representation of documents, in Figure 4 we have plotted the approach DR_3. As it may be observed in this figure, this representation schema offers (at least from the visual point of view) a set of curves that allow to study the behavior of each corpus term, in order to determine the discrimination degree of each one of them.

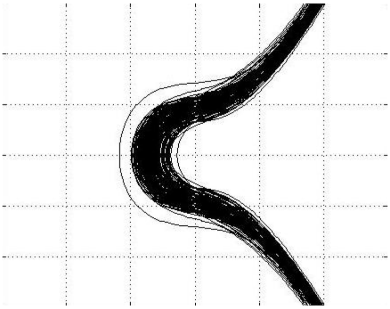


Fig. 2. Elliptic curves with approach DR_1 for corpus C1

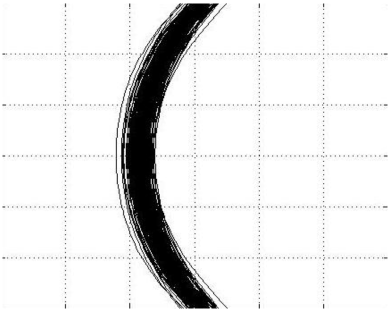


Fig. 3. Elliptic curves with approach DR_2 for corpus C1

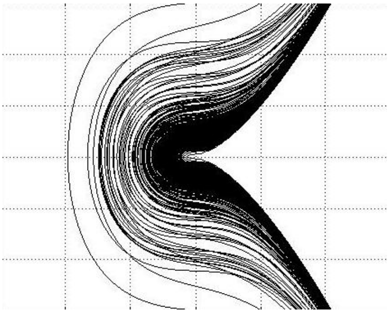


Fig. 4. Elliptic curves with approach DR_3 for corpus C1

Table 1. Corpora used in the experiments

Corpus name	Num. of docs	Vocabulary size	Maximum frequence	Max frequent term	Terms with frequence one
C1	210	15631	1299	México	7786
C3	164	12156	646	México	6160
C4	97	13533	352	México	8878
C5	256	21083	796	México	13179
C10	206	13851	686	México	6976
C11	105	8836	371	México	4676
C14	280	15751	1709	PEMEX	7630
C15	7	1357	28	México	1006

Having analyzed the three different schemata, we decided to evaluate the DR_3 approach with a greater number of documents (eight corpus). In the following subsection we describe the dataset used in these experiments. In subsection 3.2 we present the evaluation of the different document collections. Finally, we conclude this section discussing the findings of this investigation.

3.1 Dataset

In order to observe the degree of discrimination of each term, we consider a group of documents that hold some kind of similarity among them. In this case, we have selected a collection of Mexican newspaper text in the Spanish language that was used in one competition of information retrieval[1]. Each group corresponds to a set of relevant documents for a given topic/query. For instance, the first corpus is made up of documents relevant to the query "Mexican Opposition to the North American Free Trade Agreement (Oposición Mexicana al TLC)". The name we gave to each corpus, together with other features such as the vocabulary size, the total number of terms, the maximum frequency (with the associated term) and the number of terms with frequency one are shown in Table 1. We attempted to provide various features in the evaluated corpus in order to be able to obtain some conclusions of the implemented document representations.

3.2 Evaluation

The final aim of our investigation is to find an appropriate representation of each document by means of an elliptic curve. If we are able to find this curve, then we would easily determine a simple similarity measure between any pair elliptic curves and, therefore between the two corresponding documents.

In order to do so, we first need to determine the most representative terms. That is the reason because we have split the whole corpus vocabulary in various groups of terms. The different thresholds used in the partitions, together with an *ID* we have assigned to each partition, are given in Table 2.

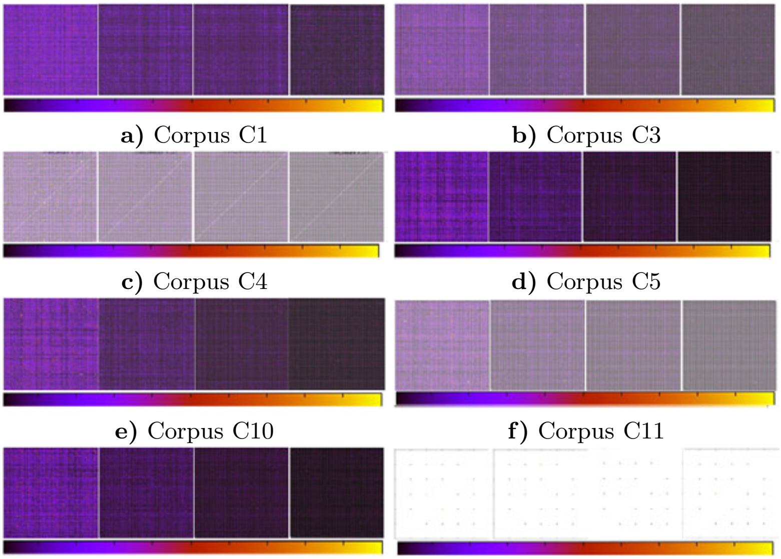
[1] `http://trec.nist.gov/`

Table 2. Thresholds used for the DR_3 representation approach

ID	Parameter thresholds
HIGH	$\bar{x_j} \in [1.0, \infty) \wedge \sigma_j \in [1.0, \infty)$
MEDIUM	$\bar{x_j} \in [0.1, 1.0) \wedge \sigma_j \in [0.1, 1.0)$
MEDIUM-LOW	$\bar{x_j} \in (0, 1.0) \wedge \sigma_j \in (0, 1.0)$
LOW	$\bar{x_j} \in (0, 0.1) \wedge \sigma_j \in (0, 0.1)$

The rationale of the aforementioned thresholds follows. HIGH was proposed with the aim of capture those terms that appear, in average, one time in each document. The standard deviation (high) in this case permits to obtain those terms whose distribution along the document collection is not uniform. We hypothesize that these thresholds would allow to obtain the best discriminant terms. MEDIUM get those terms with a lower frequency than HIGH, but the occurrence of these terms is more or less uniform through the corpus. The LOW set of thresholds bring together the terms that uniformly and nearly not appear in the corpus. Finally, MEDIUM-LOW is proposed with the goal of observing the behavior of these terms in the document representation.

In Figure 5 we may observe the behavior of each group of terms when calculating the similarity among all the documents of each corpus evaluated. Each square represents the similarity of the documents when we use only those terms that fulfill the thresholds defined in Table 2. From left to right, each square uses the HIGH, MEDIUM, MEDIUM-LOW and LOW parameters, respectively.



a) Corpus C1

b) Corpus C3

c) Corpus C4

d) Corpus C5

e) Corpus C10

f) Corpus C11

g) Corpus C14

h) Corpus C15

Fig. 5. Profile of similarity for all corpora

The lighter is one point in the square, the higher is the similarity between the two documents associated. We may observe that in all cases the HIGH representation obtains the best degree of similarity among the documents. We consider that this result is obtained due to the nature of the corpora used in the experiments. All the corpus belong to the information retrieval field and, therefore, the documents were grouped based on the frequency of their terms.

Figure 5 shows the expected behavior on document representation: the more frequent a term is, the better its degree of discrimination. Therefore, the DR_3 schema has shown to be a good representation of the corpus features.

These experiments are a first step towards the definition of proper document representation based on elliptic curves. As future work, we are considering to merge all the means and standard deviations in a vectorial representation which should be used as parameter for the elliptic curves.

4 Conclusions and Further Work

In this paper we have presented an study of the use of elliptic curves for term discrimination with the final purpose of finding an appropriate document representation. The aim is to have a simple and fast method for classifying and retrieving information from huge amount of documents.

We have evaluated three different approaches that consider the frequency of the terms in the corpus. Both, the most and less frequent terms were evaluated in order to observe their behavior in the document representation task.

In general, we have found that the most discriminant terms in the corpora used in the experiments carried out are those that appear, in average, one time in each document ($\bar{x_j} \geq 1$) with high standard deviation ($\sigma_j > 1$), i.e., those terms whose distribution along the document collection is not uniform.

However, there exist some cases in which other term frequencies allow to improve the precision of the task implemented. Therefore, it is important to further analyze a robust representation that permits to include such characteristics in a simple elliptic curve. We still need to determine a mechanism in order to integrate the characteristics of each term of a given document in a simple parameter of the elliptic curve. Further experiments will be carried out following this research line.

In conclusion, based on these preliminar results, we consider that it is possible to use the theory of elliptic curves as representation schema in order to succesfully characterize documents.

References

1. Can, F., Ozkarahan, E.A.: Computation of term/document discrimination values by use of the cover coefficient concept. Journal of the American Society for Information Science 38(3), 171–183 (1987)
2. Manning, D.C., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)

3. Pinto, D.: On Clustering and Evaluation of Narrow Domain Short-Text Corpora. Phd thesis, Department of Information Systems and Computation, UPV (2008)
4. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
5. Montemurro, M.A., Zanette, D.H.: Entropic analysis of the role of words in literary texts. Advances in Complex Systems (ACS) 05(01), 7–17 (2002)
6. Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulchloper, J.: Topic discovery based on text mining techniques. Information Processing and Management 43(3), 752–768 (2007)
7. Santiesteban, Y., Pons-Porrata, A.: LEX: a new algorithm for the calculus of typical testors. Mathematics Sciences Journal 21(1), 85–95 (2003)
8. Hankerson, D., Menezes, A.J., Vanstone, S.: Guide to Elliptic Curve Cryptography. Springer, New York (2003)