Text Document Classification and Pattern Recognition

Qin Wu, Eddie Fuller, Cun-Quan Zhang

Department of Mathematics,

West Virginia University,

Morgantown, WV 26506-6310

USA

Emails: qinwu@~, ed@~, cqzhang@math.wvu.edu

Abstract

In this extended abstract, a novel approach is proposed for text pattern recognition. Instead of the traditional models which are mainly based on the frequency of keywords for text document classification, we introduce a new graph theory model which is constructed based on both information about frequency and position of keywords. We applied this new idea to the detection of fraudulent emails written by the same person, and plagiarized publications. The results on these case studies show that this new method performs much better than traditional methods.

I. Introduction

Among the huge number of Internet documents, chatting log files, and archives of other digitized documents, how can we classify them into smaller categories in terms of their subject? Keywords frequency has long been used as a tool for estimating the probabilistic distribution of features in a document. A number of applications have been developed including language modeling [1], feature selection [2,3] and term weighting [4,5]. Based on the term frequency information, documents can be classified by several clustering methods such as decision trees [6], neural networks [7,8], Bayesian methods [9,10], or support vector machines [11,12,13].

Although keyword frequency is one of the most popular approaches for such classification processing, extensive studies and experiments have shown that such classification is not detailed enough to support decision makes due to the fact that the similarity between documents based on keywords frequency is too rough. This popularly used traditional method is evidently very effective for separating documents from very different categories, such as, physics, biology, social science, etc. However, it is not able to further cluster them into smaller groups. In order to provide an effective tool for decision support, this traditional method should be improved so that it can be further applied for intelligence

analysis, cyber crime detection, internet monitoring, chatting log surveillance, detection of plagiarism, and many data mining problems related to security, ethical issues ([1] – [13]).

In this extended abstract, we will introduce a novel approach. By using some graph theory tools, we are able to further classify documents from the same category into smaller groups based on their writing patterns. A weighted directed graph is created for each document, it records the information not only the keyword frequency, but also their location in the document. The cosine similarity between the adjacency matrices of graphs for two documents is considered as the measurement of similarity for further clustering.

A set of algorithms for the estimation of signature vectors and clustering will be briefly presented in this paper. This algorithm has been implemented and applied to two sets of sample documents: Nigerian Fraud Emails, each of which has the same topic: to transfer money into some bank accounts in order to receive larger sum of payback [14]; Papers in academic journals of mathematics, some of which are well-known cases as plagiarized publications [15,16].

Each group is in the SAME category. The testing results show that the new method is able to detect patterns of documents written by the same author, and therefore, clearly separate samples into meaningful and accurate subgroups. However, the testing result of the traditional method shows a significantly large amount of false positive outputs.

In next section, we describe the schema for representing a document as a weighted directed multigraph. Computational complexity of the algorithm will also be presented (with no mathematical detail in this extended abstract). In section IV, we will present the outputs of this new method and the comparison with the traditional method.

II. Graph Model for Pattern Recognition

The overall approach of this algorithm begins with the identification of a set of relevant keywords. Once these are selected, we then aggregate the relative distances of the keywords with a document. This in turn is used to construct a weighted directed multigraph that generates representing vectors for each document in a high dimensional feature space. These vectors can then be used to determine similarity values for any pair of documents.

II-1. Summary of our Method

Step 1: Using weighted directed multigraph to find a signature vector for each document.

Step 2: Calculate the similarities between any two documents via their signature vectors.

Step 3: Using Quasi-Clique Merge clustering method [17] to classify all documents.

Details of each step will be described step by step with a simple example.

II-2 Details of the Step 1.

To have a clear view of the algorithm, we will use the following example (see Figure 1, a sample from an open data-set source [14]) to explain the procedure.

(i) Record the keyword information appeared in the document.

For a given document, the following steps are applied to it. Suppose we have already chosen a set of words as keywords, say, $K=\{K_1, K_2, \dots \dots K_m\}$. Record every keyword and its position in the document, and the frequency of each keyword as well. For this example (Figure 1), we use the keyword set: $\{bank, fund, account, transfer\}$. Thus we have the tables (Table 1 and Table 2).

From: Mr. Davies Manga
Tel: 234 - 80- 33345356
E-mail: daviesmanga@qrio.com
Dear Friend,

This mail might be a surprise since we have not met, but this can be remedied as we proceed in this transaction.

I am the personal assistance to the Nigerian late Minister of Sports Mr. Ishaya Mark Aku who died on a plane crash which took place on the 4th April 2002. Prior to his death, I assisted him in depositing the sum of US\$18.5 million (Eighteen Million Five Hundred Thousand U. S Dollars) in a local bank in my country (Nigeria).

I solicit your assistance in receiving these funds on our behalf into your account for further investment into a profitable venture in your country. The details of the transfer will be forwarded to you once I receive the assurance of your full cooperation.

You will be entitled to 15% of the total sum as soon as this fund is transferred into your account.

The nature of your business does not necessarily matter in this transaction. Awaiting your urgent response through my above telephone and e-mail address

Yours truly,
Mr. Davies Manga.

Figure 1. A sample email

<u>Keyword appears in the document</u>	<u>Position in the document</u>
bank	91
fund	103
account	109
transfer	124
fund	153
transfer	155
account	158

Table 1. Position of keywords in the sample document

<u>Keyword</u>	<u>Frequency</u>
bank	1
fund	2
account	2
transfer	2

Table 2. Frequency of keywords in the sample document

(ii) Construct a weighted directed multigraph

A weighted directed multigraph G_m with the vertex set $K=\{K_1, K_2, \dots \dots K_m\}$ is constructed as follows. For each

pair of keywords K_i, K_j , if K_i appears at the position p_i and K_j appears at the position p_j with $p_j > p_i$, then add an arc from the vertex K_i to the vertex K_j with the weight $p_j - p_i$. (Note that the resulted weighted directed multigraph may contain not only parallel arcs but also loops.) For the given sample document (Figure 1), a weighted directed multigraph G_m is constructed (see Figure 2).

Figure 2. A weighted multi-graph G_m representing the document in Fig. 1

(iii) Simplification of representing graphs

The weighted directed multigraph G_m constructed in the previous step is simplified as follows: a directed graph G_s is constructed from G_m , in which, parallel arcs are combined (see Figure 3).

Figure 3. A (simplified) weighted graph G_s

II-2. Details of the Step 2

(i) Signature Vector

An input document is represented by a 20-dimensional vector, in which the first 4 components are the frequencies of keywords, while the last 16 components is the (4×4) -adjacency matrix of G_s . For the set of 500 input documents, a (500×20) -matrix is created where each row represents a document. The matrix is further

normalized for every column of the matrix as a pre-processing. Each row v_i of the resulting matrix is called the signature vector of the corresponding document D_i .

(ii) Similarity

The similarity s_{ij} between any two documents D_i and D_j is determined by the cosine similarity of the signature vectors. That is, $s_{ij} = \cos \alpha$ where α is the angle between signature vectors.

II-3. Details of the Step 3

After the calculation of the similarities for every pair of documents, we can classify these documents into different groups. Here we applied Quasi-Clique Merge (QCM) method to cluster the documents due to its special feature of multi-membership clustering [17].

III. Computational complexity

In this extended abstract, we omit the mathematical detail of the algorithm and complexity analysis. The worst-case complexity is no more than $O(n^2 m^2)$ where m is the size of selected keywords set, n is the number input documents for comparison.

IV. Testing Results

In order to evaluate the effectiveness of our algorithm, we compare the results of our method with other method. We calculate the similarity between every pair of documents by two different ways. *KF*: only use keyword frequency information; *KFP*: use keyword frequency and pattern information, which is based on the weighted directed multigraph model described in this paper.

IV-1. Fraudulent emails

The traditional method (*KF*) does produce large amount of false positives. For example, the emails displayed in Figures 4 and 5 are obviously written in very different styles. The similarities estimated by *KF* and *KFP* methods are 1 and 0.43177, respectively.

IV-2. Plagiarism Papers

Date description: a well-known plagiarism paper [16] (named Paper-1A) on independence number of a graph and its corresponding original paper (named Paper-1B). In order to have a comparison, a set of another 35 papers from the internet (named Paper-2, Paper-3, ..., Paper-36) was downloaded randomly, which are all related to the

same subject: independence numbers of graphs. The following pictures are the first pair of pages: Paper-1A and Paper-1B (Figure 6 and Figure 7).

2002-02-20a.html

Calvary Greetings.
I am a preacher with the Seed Harvest Ministry, and I will like to get advice from you. I believe the advice I need is secular to some extent, but do have the patience to understand my intention.

A few year's ago I was in Liberia where I had established a little congregation where I preached regularly, but the civil war escalated and the church was converted to a hospital of sought. On one faithful day three Nigerian Soldiers came to me and left trunks of money with me and swore to come back for it.

When the war in Liberia intensified I deposited this box in a security vault in The Netherlands (Holland), I was escorted to Holland, by the ECOMOG soldiers who felt the content of the trunks were documents and never cared to check the trunks, I was air lifted to England with other refugees then, I stayed in England and when I was getting frustrated I returned to Africa only to find out that two of the three Soldiers died in Battle while the last was declared MIA (Missing in Action).

Now I have the fund, which the soldiers said was over thirty million dollars (US\$30 million), but it lies in a security vault in a Security Company in Holland. I was contacted in a letter dated 12th of September by the security company asking me to indicate if I was in receipt of the letter as a sign that I am living and still subscribed for their service.

I am interested in using a small fraction of this money, much less than one percent for a re-organization of the work of God, but I do not need the rest and do not want to have any direct dealing with it, but I need someone who will be able to use the fund better maybe for charity or something universally profitable, I have thought of doing it myself but,

my ministry is the apocalypse and I believe and preach the soon coming of the Lord which make me not indulgent in reliance on money or wealth in any form.

I also do not have the money that has accrued on the service and safe deposit of the company that holds the fund, I will be glad if you have any interest in the direction of maybe managing or transmitting the fund as you prefer in your capacity and understanding.

I will be glad to get a response from you ASAP.

Regards.
Rev. Sunday Nana

Figure 4. email 2002-02-20a.html

2002-07-04b.html

Pardon the abruptness and the liberty of this letter, it is due to its exigency.

I am a government functionary in my country. I represent a caucus in need of a foreign partner to assist in the transfer and subsequent investment of a confidential sum in confidence.

We are ready to offer you a percentage for your assistance in the transfer and a negotiable fee if you can put together a portfolio for investment.

All we need is your trustworthiness and your capability to protect our interest.

Please note that we shall go through the legal procedures entailed in our laws and international laws in transferring the funds to you in trust. The source of the money is authoritative.

I wish that you would take the essence of this Letter in strict confidence.

You can reach me immediately for us to commence dialogue on the mod operandi entailed first, by email in earnest. Thank you for your courtesy.

Best regard,
Franklin Numa.

Figure 5. email 2002-07-04b.html

J. Indian Inst. Sci., May–Aug. 2003, 83, 95–97
©Indian Institute of Science.

Short Communication

Some results on the independence number of a graph

Abstract

In this paper, we give new lower bounds for the independence number $\alpha(G)$ of a finite and simple graph G .

Keywords: Graphs, independence number, lower bounds.

Graphs, considered here, are finite and simple (without loops or multiple edges), and [1], [2] are followed for terminology and notation. Let $G = (V, E)$ be an undirected graph, with the set of vertices $V = \{v_1, v_2, \dots, v_n\}$ and the set of edges E , such that $|E| = m$.

We denote by $d(v)$ the degree of a vertex v in G . It is well known (e.g. see [2]) that $\sigma(G) = d(v_1) + d(v_2) + \dots + d(v_n) = 2m$.

Let $\delta_i(v)$ be the number of vertices having the distance i from a vertex v of G and let $\alpha(G)$ be the independence number of G .

LEMMA 1. *If G is a triangle-free graph, then*

$$\alpha(G) \geq \alpha'(G) = \sum_{v \in V} \delta_1(v) / (1 + \delta_2(v) + \delta_3(v)).$$

Proof. We randomly label the vertices of G with a permutation of the integers from 1 to n . Let $S \subseteq V$ be the set of vertices v for which the minimum label on vertices at distance 0, 1 or 2 from v is on a vertex at distance 1. Obviously, the probability that S contains a vertex v is given by $\delta_1(v) / (1 + \delta_2(v) + \delta_3(v))$ and, therefore, the expected size of S is equal to $\alpha'(G)$. Moreover, S must be an independent set of G , since, otherwise, if S contains an edge it is easy to see that it must lie in a triangle of G , contradicting the hypothesis. Thus, the lemma is proved.

THEOREM 1. *If G is a triangle- and pentagon-free graph with m edges, then $\alpha(G) \geq \sqrt{m}$.*

Proof. Let $d(G)$ be the average degree of vertices of G . Since G is a triangle- and pentagon-free graph, then we have $\alpha(G) \geq \delta_1(v)$, by considering the neighbours of v , and $\alpha(G) \geq 1 + \delta_2(v)$, by considering v and the vertices at distance 2 from v , for any vertex v of G . Thus, by the above lemma, $\alpha(G) \geq \alpha'(G) \geq \sum_{v \in V} \delta_1(v) / 2\alpha(G)$, that is,

$$\alpha(G)^2 \geq nd(G)/2 \text{ or } \alpha(G) \geq \sqrt{nd(G)/2}.$$

But, $d(G) \geq \sigma(G)/n = 2m/n$ and, therefore, $\alpha(G) \geq \sqrt{m}$, the theorem being proved.

LEMMA 2. *If G is a graph with an odd girth $2k+3$ ($k \geq 2$) or greater, then*

$$\alpha(G) \geq \sum_{v \in V} \left(\frac{1}{2} (1 + \delta_1(v) + \dots + \delta_{k-1}(v)) / (1 + \delta_1(v) + \dots + \delta_k(v)) \right).$$

Figure 6. Paper-1A

THE INDEPENDENCE NUMBER OF DENSE GRAPHS WITH LARGE ODD GIRTH

James B. Shearer
Department of Mathematics
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
JBS at WATSON.IBM.COM

Submitted: January 31, 1995; Accepted: February 14, 1995

Abstract. Let G be a graph with n vertices and odd girth $2k+3$. Let the degree of a vertex v of G be $d_1(v)$. Let $\alpha(G)$ be the independence number of G . Then we show $\alpha(G) \geq 2^{-\binom{k+1}{2}} \left[\sum_{v \in V} d_1(v)^{\frac{1}{2^{k-1}}} \right]^{\frac{1}{k-1}}$. This improves and simplifies results proven by Denley [1].

AMS Subject Classification. 05C35

Let G be a graph with n vertices and odd girth $2k+3$. Let $d_i(v)$ be the number of points of degree i from a vertex v . Let $\alpha(G)$ be the independence number of G . We will prove lower bounds for $\alpha(G)$ which improve and simplify the results proven by Denley [1].

We will consider first the case $k=1$. We need the following lemma.

Lemma 1: Let G be a triangle-free graph. Then

$$\alpha(G) \geq \sum_{v \in V} d_1(v) / [1 + d_1(v) + d_2(v)].$$

Proof. Randomly label the vertices of G with a permutation of the integers from 1 to n . Let A be the set of vertices v such that the minimum label on vertices at distance 0, 1 or 2 from v is on a vertex at distance 1. Clearly the probability that A contains a vertex v is $d_1(v) / [1 + d_1(v) + d_2(v)]$. Hence the expected size of A is $\sum_{v \in V} d_1(v) / [1 + d_1(v) + d_2(v)]$. Furthermore, A must be an independent set since if A contains an edge it is easy to see that it must lie in a triangle of G a contradiction. The result follows at once.

Typeset by $\mathcal{A}_{\mathcal{M}\mathcal{S}}\mathcal{T}_{\mathcal{E}}\mathcal{X}$

1

Figure 7. Paper-1B

A set of keywords was selected which consists of 23 frequently used standard terminologies in popular graph theory textbooks [18,19]. Tables 3 and 4 indicate the significant difference in the applications of both methods: *KF* and *KFP*.

		<u>Similarity by KFP Method</u>	<u>Similarity by KF Method</u>
<u>Paper 1A</u>	<u>Paper 1B</u>	0.77857	0.97074
<u>Paper 11</u>	<u>Paper 13</u>	0.34563	0.99500
<u>Paper 13</u>	<u>Paper 28</u>	0.20377	0.98567
<u>Paper 11</u>	<u>Paper 28</u>	0.09859	0.98011
<u>Paper 31</u>	<u>Paper 6</u>	0.05503	0.97190

Table 3. Some false positive of *KF* method

	<u>KFP method</u>	<u>KF method</u>
<u>Similarity between Paper 1A and Paper 1B</u>	0.78	0.97
<u>Similarities between all other pairs</u>	all < 0.6, and average < 0.2	6 pairs > 0.97

Table 4. Similarity for all sample pairs

From Table 4, estimated by *KFP* method, the similarity between the Paper-1A (the plagiarism paper) and Paper-1B (the original paper) is 0.78, which is the highest among all pairs, and the similarities between all other pairs are less than 0.6, most of them are far less than 0.2. However, when *KF* method is applied, the similarity between the 1A and 1B is 0.97 which is not the highest, and there are other 6 pairs of papers have similarities greater than 0.97 (some of those pairs are listed in Table 3).

Acknowledgements.

The authors of this paper are partially supported by a WV EPSCoR grant.

References

- 1) J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval, in Research and Development in Information Retrieval, pages 275-281, 1998.
- 2) Y. Yang and J. O. Pedersen, A comparative study on feature selection in text categorization, in Proceedings of the 14th International Conference on Machine Learning, Nashville, US, 1997.
- 3) H. Schutze, D. A. Hull, and J. O. Pedersen, A comparison of classifiers and document representations for the routing problem, in Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, 1995.
- 4) M. Lan, C. Tan, H. Low and S. Sungy, A comprehensive comparative study on term weighting schemes for text categorization with support vector machines, in Proceedings of the 14th international conference on World Wide Web, pages 1032-1033, 2005.
- 5) R. Robertson and K. Sparck-Jones, Simple, proven approaches to text retrieval, Technical Report, 1997.
- 6) C. Apte, F. Damerau, and S. Weiss, Text mining with decision rules and decision trees, in Workshop on Learning from text and the Web, Conference on Automated Learning and Discovery, 1998.
- 7) M. E. Ruiz and P. Srinivasan, Hierarchical text categorization using neural networks. Information Retrieval, 5(1), pp. 87-118, 2002.
- 8) H. Ng, W. Goh, and K. Low, Feature selection, perceptron learning, and a usability case study for text categorization, in Proc. 20th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR97), pages 67-73, 1997.
- 9) T. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.
- 10) Y. Yang and X. Liu, A re-examination of text categorisation methods, in Proc. 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR99), pages 67-73, 1999.
- 11) S. Tong and D. Koller, Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research, Volume 2, pages 45-66, 2001.
- 12) T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in Proc. European Conf. Machine Learning (ECML'98), pages 137-142, 1998.
- 13) O. de Vel, A. Anderson, M. Corney and G. Mohay, Mining E-Mail Content for Author Identification Forensics. SIGMOD Record, Vol. 30, No. 4, pages 55-64, 2001.
- 14) An open data-set source: Nigerian Fraud Email Gallery, <http://potifos.com/fraud/>
- 15) Wikipedia, the free encyclopedia: <http://en.wikipedia.org/wiki/Plagiarism>

- 16) Wikipedia, the free encyclopedia:
http://en.wikipedia.org/wiki/D%C4%83nu%C5%A3_Marcu
- 17) Y. Ou and C.-Q. Zhang, A new multimembership clustering method, Journal of Industrial and Management Optimization, Vol. 3, No. 4 (2007), 619-624.9
- 18) J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, Macmillan, London, (1976).
- 19) D. B. West, *Introduction to Graph Theory*, Prentice Hall, Upper Saddle River, NJ. (1996).