Text Retrieval from Document Images based on N-Gram Algorithm

Chew Lim Tan, Sam Yuan Sung, Zhaohui Yu and Yi Xu
School of Computing, National University of Singapore
Kent Ridge, Singapore 117543

Abstract

In this paper, we propose a method of text retrieval from document images using a similarity measure based on an N-Gram algorithm. We directly extract image features instead of using optical character recognition. Character image objects are extracted from document images based on connected components first and then an unsupervised classifier is used to classify these objects. All objects are encoded according to one unified class set and each document image is represented by one stream of object codes. Next, we retrieve N-Gram slices from these streams and build document vectors. Lastly, we obtain the pair-wise similarity of document images by means of the scalar product of the document vectors. Four copora of news articles were used to test the validity of our method. During the test, the similarity of document images using this method was compared with the result of ASCII version of those documents based on the N-Gram algorithm for text documents.

Keywords: *Document Image, Information Retrieval, Similarity Measure, N-Gram Algorithm*

1. Introduction

The Singapore National Library archives the entire set of past issues of major newspapers in Singapore. All issues of the newspaper are in microfilms carefully preserved in the National Library [1]. Many researchers have frequented the microfilm section of the Singapore National Library with a wide variety of interests. It is thus proposed that the microfilm images be digitized to facilitate retrieval of relevant news articles based on text similarity.

There are many ways to measure text similarity of documents. One way is to analyze the similarity of the documents' contents based on semantics but this needs a large amount of processing time and is dependent on the specific language used. Another way is to use a statistical method to gauge the text similarity directly without the need to understand the meaning of documents. A common statistical method is N-Gram algorithm. This method is easy to implement without too much processing time. Many researchers have used it to classify document texts.

There are two methods to retrieve information from document images. One is the retrieval based on optical character recognition (OCR) followed by the usual text retrieval techniques. However, OCR systems are not perfect and they require significantly more processing time than gauging similarity. Another approach is the retrieval based on the image content. This does not require language identification. This paper adopts the latter approach by gauging the similarity of document images using an N-Gram algorithm without the recognition of the characters.

The remainder of this paper is organized as follows. Section 2 surveys related works in text retrieval of electronic texts as well as document images. Section 3 describes the feature extraction process to detect and classify character objects from the document images. Section 4 presents the N-Gram algorithm that measures the text similarity based on the character objects extracted. Section 5 discusses experimental results that confirm the validity of the proposed model. Finally, conclusions and future work are given in Section 6.

2. Related Works

Over the past few decades, methods of categorisation and retrieval of machine-readable texts [2-4] had been proposed. They have relied on self-evident utility of words, sentences, and paragraphs for sorting, categorising, and retrieving texts. Furthermore, various means of suppressing uninformative words, removing prefixes, suffixes, and endings, interpreting inflected forms, etc. have been developed. Depending on the application, these methods share a number of potential drawbacks: they require a linguist or a polyglot for initial set-up and subsequent tuning, they are vulnerable to variant spellings, misspellings, and random character errors, and they tend to be both language-specific and domain-specific.

The purely statistical characterisation of text in terms of its constituent N-Grams (sequences of N consecutive characters) [5] has been applied to text analysis and document processing, including spelling and error correction [6-12], text compression [13], language identification [14-15], and text search and retrieval [16-17]. Basing on this statistical characterisation, M. Damashek [18] has proposed a simple but novel vector-space technique that makes sorting, clustering and retrieval feasible in a large multilingual collection of documents.

Damashek's method does not rely on words to achieve its goal, and no prior information about the document content or language is required. It only collects the frequency of each N-Gram to build a vector for each document and the processes of sorting, clustering and retrieval can be implemented by measuring the similarity of the document vectors. It is language-independent. A little random error only influences a small quantity of N-Grams and will not change the total result. This method thus provides a high degreee of robustness.

Text in document images is a more complicated matter for text retrieval. One common method is to convert it to machine readable text using optical character recognition (OCR) first and then use the usual text retrieval techniques. However, character recognition systems are not perfect and they require a significant amount of processing time. Furthermore, OCR is language-dependent. A typical system can only recognize one or several languages. We need to know the specific language in the document beforehand.

Another approach is to retrieve information based on the image content directly. This does not require language identification. Recently, several researchers have made such an attempt in a number of applications. For example, F. R. Chen and D. S. Bloomberg [19-20] have described a method for automatically selecting sentences for creating a summary from a document image without recognition of the characters in each word. They build word equivalence classes by using a rank blur hit-miss transform to compare word images and use a statistical classifier to determine the likelihood of each sentence being a summary sentence. Hull and Cullen [21] have proposed a method to detect equivalent document images by matching the pass codes of document. They create a feature vector that counts the numbers of pass codes in each cell of a fixed grid in the image and equivalent images are located by applying the Hausdorff distance to the feature vectors.

Other researchers have also proposed methods to retrieve text directly from non-English document images. For instance, Y. He et al [22] have proposed an index and retrieval method for Chinese document images based on stroke density code. Language classification of multilingual documents is another field having been researched. A. L. Spitz et al [23-24], C. Y. Suen et al [25] and C. L. Tan et al [26] have developed systems to identify Latin-based languages, Han-based languages and other languages using the character shape coding [27].

3. Feature Extraction

Figure 1 outlines the steps in gauging the similarity of document images based on content. To identify the features, character objects in the predominant font are extracted from the document images and then character object equivalent classes are identified based on shape similarity. From several sets of classes, one unified class set can be estimated. The objects, which belong to the same class, are assumed to represent the same identity. Layout analysis is performed to determine the reading order of character objects. An object sequence can be obtained from each document image. This information is used to calculate the similarity of document images using the N-Gram algorithm.

3.1 Character Object

In document images, there are three kinds of character objects. The first is the isolated characters, that have each only one connected component. The second is also isolated characters, but they each have more than one connected component, such as lower characters “i” and “j”. The third kind is the characters that are connected to each other, such as “ft” and “ff”. Character objects can be extracted by measuring the connected components in the image and comparing the relative positions of adjacent components. Thus a character object contains only one connected component or several connected components, which have unambiguous relative positions.

We divide the document image into many rectangle zones and each zone contains one character line. The comparison of the relative positions of connected components is restricted to the interior of each zone. The components, which are in different zones, belong to different character objects. So, these components can be expressed as $C_{i,j}$, where i is the number of zone and j is the order number of connected components in each zone from left to right. If the horizontal overlapping extent in $C_{i,j}$ and $C_{i,j+1}$ is larger than a threshold, they belong to same character object. Otherwise, they belong to different objects.

Punctuation does not have special meaning in the N-Gram algorithm. It is wasteful to spend too much time processing punctuation marks. In general, the height of a punctuation mark is less than that

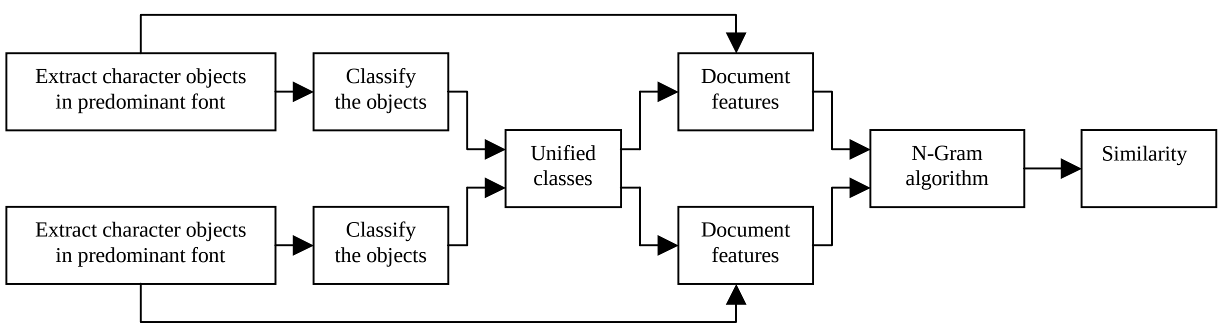
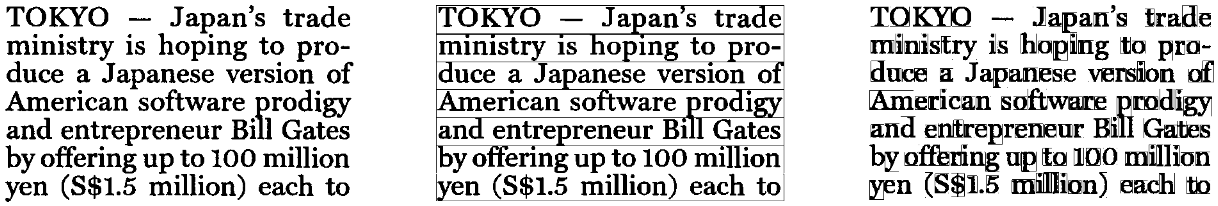


Figure 1. Gauging the similarity of document images based on content

of a character. So, when character objects have been retrieved, we do an additional operation to filter small objects whose width or height is less than a pre-determined threshold. Most of the punctuation marks and noise will be removed in this manner.

Figure 2 shows the result of character object retrieval. Item (a) is the original document image. Item (b) outlines the zones of character lines and item (c) shows the extracted character objects.



a. Original image b. Separated line zone c. Character objects

Figure 2. Extraction of character objects

3.2 Character Object Class

In newspapers, the text may be printed in different font sizes and font styles. The main body of text is usually printed in one font, which is generally the predominant font, whereas headings and captions may appear in a variety of fonts. For gauging similarity of document images, only text in the predominant fonts is considered. To identify characters, an unsupervised classifier is used to place each character object into a set of classes. Each class is regarded as representing a unique identity.

For each character object, we use two vectors to store the object features: Vertical Traverse Density (VTD) Vector and Horizontal Traverse Density (HTD) Vector. Samples of HTD and VTD are shown in Figure 3.

For two character objects *i* and *j*, their distance d_{ij} will be calculated by the following function:

$$d_{ij} = diff\,(HTD_i, HTD_j) + diff\,(VTD_i, VTD_j) \tag{1}$$

where, $diff\,(V_i, V_j)$ is a function to calculate the distance between the two vectors V_i and V_j. We assume that n_i and n_j are the dimensions of vectors V_i and V_j, respectively, and $V_i = v_{i0}v_{i1}v_{i2}\cdots v_{in_i-1}$, $V_j = v_{j0}v_{j1}v_{j2}\cdots v_{jn_j-1}$. The function $diff\,(V_i, V_j)$ is defined as follow:

$$diff\,(V_i, V_j) = \min_{-c \le k \le c}(distance(U_i^k, U_j^k)) \tag{2}$$

where, *c* is a positive integer constant. The vector U_i^k and U_j^k have the same dimension, which is

$$n_{ij}^k = \begin{cases} \max(n_i + k, n_j) & if\ k \ge 0 \\ \max(n_i, n_j - k) & if\ k < 0 \end{cases} \tag{3}$$

and their elements are

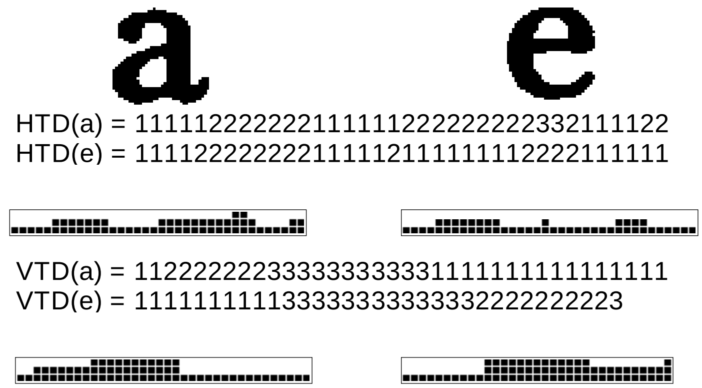


Figure 3: The illustration of HTD and VTD of character 'a' and 'e'

$$u_{il}^{k} = \begin{cases} v_{il} - \max(k,0) & if \ \max(k,0) \le l < \max(k,0) + n_i \\ 0 & otherwise \end{cases} \tag{4}$$

and

$$u_{jl}^{k} = \begin{cases} v_{jl} - \max(-k,0) & if \ \max(-k,0) \le l < \max(-k,0) + n_j \\ 0 & otherwise \end{cases}, \tag{5}$$

respectively.

The initial value of $d(U_i^k, U_j^k)$ is the maximum of *m* and *n*. For each $l \in (0, \min(n,m)]$, if

$$u_{il}^{k} = u_{jl}^{k}$$

or

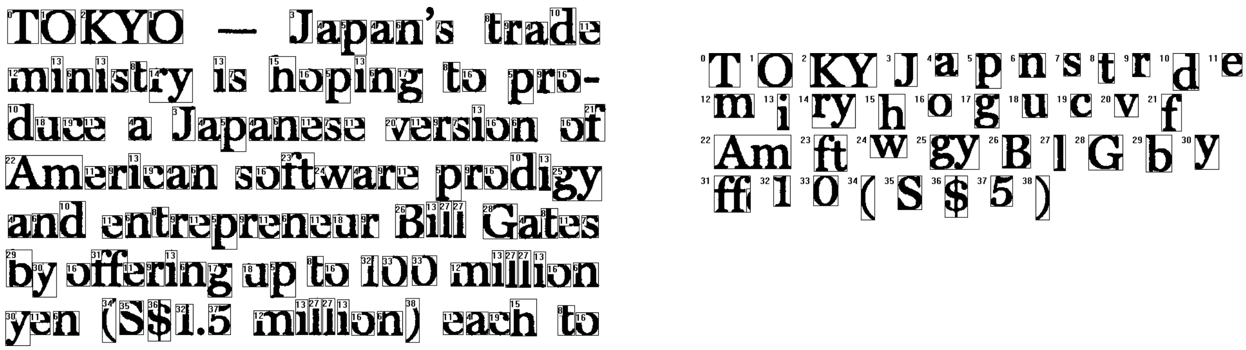
$$u_{il}^{k} = u_{il-1}^{k} \ \& \ u_{il+1}^{k} = u_{jl}^{k}$$

or

$$u_{jl}^{k} = u_{jl-1}^{k} \ \& \ u_{jl+1}^{k} = u_{il}^{k}$$

the value of $d(U_i^k, U_j^k)$ is decreased by one.

The result of classification is shown in Figure 4. In item (a), each rectangle outlines a character object extracted and the number expresses the sequence number of the class that the object belongs to. The objects that have the same number represent the same class. Item (b) lists the total class set created from this image.



(a) (b)

Figure 4. Character object classification

3.3 Unifying Character Object Class

One set of character object classes can be obtained from each document image. The predominant font of different images may have different font sizes. And the numbers of different sets of classes may be also different. To find a unified way to express the features of the document images in question, we must build a unified object classes among these document images.

First, the VTD vector and HTD vector of all character object classes are normalised so that the vectors of object classes with different sizes will have the same dimension number. The number of permitted dimension of vectors is set sufficiently large. All the features of VTD and HTD vectors will be preserved. Next, all elements of the normalised class sets are unified. They are classified again to create a set of unified classes. The equivalent classes of different sets are merged to one class. As a result, the equivalent objects in different document images will be denoted by one same object class. Finally, a look-up table from the original class set to the unified class set is built for each document image. Using these tables, all character objects in these document images can be mapped to the unified character object classes. Objects corresponding to the same class will be regarded as having the same identity.

3.4 The Class Number List of Character Objects

After all character objects have been expressed by a set of classes, a layout analysis is performed to determine the reading order of character objects and the space between two adjacent objects. One list is built for each document. Each item in the list is the class identification number that the character object belongs to. When the interval between two adjacent objects is very large or they belong to different lines, one element of blank class will be added as a word separator. The list so constructed will be used to measure similarity with other document images.

4. N-Gram Algorithm

The use of N-Gram algorithm for text similarity was proposed by M. Damashek [18]. The approach basically uses frequency statistics to calculate the similarity between the two vectors representing two documents. The use of N-Gram algorithm in other text processing has also been reported in [5-17]. These methods are based on the ASCII values of the electronic texts. Instead of relying on ASCII values, the use of image contents has been attempted by researchers [19-28] for summary extraction, similarity measurement and document retrieval. The present method attempts to adapt the N-Gram algorithm for image-based similarity measure.

4.1 N-Gram slice

The N-Gram slice is the basic unit to be sampled in the N-Gram algorithm. The class identification number list is converted to a set of N-Gram slices. An N-Gram is a sequence of N consecutive items of a stream. Using a window of N-item length, which is moved over the list one item forward at a time, N-Grams are copied out of the list.

4.2 Document Vector

First, every possible N-Gram is given a number, so called the hash key. How the N-Grams are numbered is not important, as long as each instance of a certain N-Gram is always given the same number, and that two different N-Grams are always assigned different numbers.

Next, a hash table is created to keep track of the frequency of occurrence in the list being studied. Each hash table can be treated as a vector, so called the document vector. Every time an N-Gram is picked, the element of the document vector given to the N-Gram is increased by one. The hash key of N-Gram determines the corresponding position of this N-Gram in the vector.

The occurrence frequency of each N-Gram is normalised by dividing it by the total number of the extracted N-Grams. This means that the absolute number of occurrence will be replaced with the relative frequencies of corresponding N-Grams. The reason for doing this is that similar texts of different lengths after this normalisation will have similar document vectors.

4.3 Similarity Measure

Document vectors for similar documents generally point in the same direction. The similarity score between two document vectors is defined as their scalar product divided by their lengths. A scalar product is calculated through summing up the products of the corresponding elements. This is equivalent to the cosine of the angle between two document vectors seen from the origin. So, the similarity between document images *m* and *n* will be

$$Similarity\ (X_m, X_n) = \frac{\sum_{j=1}^{J} x_{mj} x_{nj}}{\sqrt{\sum_{j=1}^{J} x_{mj}^2 \sum_{j=1}^{J} x_{nj}^2}} \quad (6)$$

where, X_m and X_n are the document vectors of image *m* and *n* respectively, *J* is the dimension number of document vector, and $X_i = x_{i1} x_{i2} \cdots x_{iJ}$.

5 Experimental results

Experiments were carried out to test the effectiveness of our image-based similarity measure in comparison with the text-based similarity N-Gram algorithm. All document images in the experiments were obtained by scanning at 600 pixels/inch (ppi). To make the process simple, some preprocessing is done by de-skewing the images [29] and removing noise such as small dirty spots. In the case where there are headlines and pictures or photographs, they are removed from the images. Four different corpora of documents were used in the following tests.

Corpus One (E01 - E26) is made up of articles that were extracted from the Internet and were already electronically available. The news articles were printed using MS-Word in 10-point Times New Roman font. The printed documents were then scanned as images. These articles address four different kinds of topic, respectively. E01-12 talk about economic crises in Brazil, E13-17 refer to personal computer, E18-E21 tell of scholarship and E22-E26 describe the news of a nuclear spy in US. For each topic, we picked the first one of each group as the reference article and thus E01, E13, E18 and E22 were selected. Similarity measures of all the articles in this corpus with the respective four reference articles were made using the image-based and text-based methods. The results are summarized in Table 1 and Figure 4.

The above images came from paper printed by a printer. We next used two corpora that came from newspapers directly and scanned them to get the images. To create the ASCII versions of these articles as a means of benchmarking, an OCR system was used to extract the text from the images. The extracted texts were corrected by hand for any error from the OCR.

Corpus Two (N1 - N8) contains eight news articles in *The Straits Times,* a local English news daily in Singapore. In this corpus, four articles talk about Indonesia, while the other fours contain news about Japan, Cambodia, Thailand and Russia, respectively. The eight articles are shown in figure 5 and pair-wise comparisons among these articles are summarized in Table 2 for both the image-based and text-based similarity. Article N1 was chosen as a reference article to compare with the rest. The result is shown in figure 6.

Corpus Three (C1 - C7) comprises recent news in *LianHe ZaoBao,* a local Chinese news daily in Singapore. In this corpus, articles C1 and C2 talk about the relationship between Singapore and Malaysia, articles C3 and C4 are about the economy of Malaysia, and articles C5 to C7 contains news about the relationship between Mainland China and Taiwan. The articles are shown in figure 7 with codes C1 to C7 indicated to help the non-Chinese readers. Image-based and text-based similarity measures among articles in corpus three are shown in table 3. Article C1 was used as a reference article to compare with the rest. The result is shown is figure 8.

From the above results, we can see that the similarities of documents measured from text-mode articles and image-based articles share some resemblance though not entirely equivalent to each other. The result of the text version of documents provides more distinguishable similarity measures. This is because the character objects extracted from the document images are not equivalent to characters and objects corresponding to the same character may be classified into different object classes. Nevertheless, the image-based similarity provides an adequate means to retrieve similar news articles with respect to a reference article. Furthermore, the results from corpus three show equally convincing similarity measures for Chinese news articles, thus confirming the language independence of our approach.

From the testing with the three corpora, it can be seen that a threshold may be set to decide whether a text is similar to a reference article. The threshold lies somewhere in the region of 0.1 to 0.2. A fourth corpus is thus chosen to see the effect of the choice of threshold. This corpus contains a total of 159 English news articles which may be roughly grouped into nine major topics depending on their contents. These articles are different from those in the above three corpora. Corpus four is a mixture of articles downloaded from the Internet and articles taken from newspaper cuttings. An article from each of the nine topics is chosen as the reference article for that topic to retrieve articles from the corpus. Knowing the number of articles in topic i (let it be n_i), we first allowed the system to retreive n_i topmost similar articles and determined how many of these n_i articles are about topic i.

Let this number of correctly retrieved articles be m_i. We define *accuracy* of this retrieval process as m_i/ n_i. We next retrieved articles based on the threshold instead of a pre-determined number of articles. We set threshold at 0.1, 0.15 and 0.2 in the next three experiments respectively, and find the values of precision[1] and recall[2] based on the usual definitions adopted in information retrieval. We carried out the above experiments for all articles, but taking one article in turn as a reference each time. The average accuracies, precisions and recalls were then obtained for each class. They are tabulated in Table 4. It can be seen that if the number of relevant articles are known beforehand, then retrieving that number of articles for a topic in question can achieve an average accuracy of 87.7%. Using a threshold as a basis of retrieval, one can see a trade-off between precision and recall. At the threshold of 0.2, the average precision and recall are 100% and 44%, respectively, whereas choosing 0.1 as the threshold will give an average precision and recall of 73.9% and 85.9%, respectively. Thus, setting a higher threshold gives a better precision but poorer recall, and the reverse is true for a lower threshold. If the emphasis is on retreiving only relevant articles, then a 0.2 threshold should be used. On the other hand, if the intent to retrieve as many as possible news articles, then a threshold of 0.1 may be adopted. The 0.15 threshold appears to be a good compromise.

6 Conclusion and Future Work

A new model of document image text retrieval based on an image-based similarity measurement without the use of OCR is proposed in this paper. We extract the features of document images by obtaining and classifying the character objects. Then, a N-Grams algorithm is used to measure their similarity. Experiments using four copora of news articles have confirmed the validity of the model with an average of precision ranging from 73.9% to 100% and an average recall ranging from 44% to 85.7%, depending on the similarity threshold.

The method is suited for gauging the similarity of document images that have the same font style. One of our future research directions is to examine documents of different font sizes and styles. The final object of our method is to use it in the retrieval of news articles from microfilm images [1]. Microfilm images are noisier than the images used in the present study. So, how to deal with the noise will also be our future research.

[1] Precision is defined as percentage of the number of correctly retrieved articles over the number of all retrieved articles.
[2] Recall is defined as percentage of the number of correctly retrieved articles over the number of articles in the category.

Table 1: Image-based and Text-based Similarity for Corpus One

		E01		E13		E18		E22	
		Image based	Text based	Image based	Text based	Image based	Text based	Image based	Text based
Group 1	E01	1.0000	1.0000	0.1357	0.0543	0.0842	0.0217	0.1294	0.0589
	E02	0.3271	0.4299	0.1053	0.0477	0.0577	0.0131	0.0854	0.0507
	E03	0.5062	0.5554	0.1226	0.0318	0.0782	0.0227	0.1304	0.0649
	E04	0.3886	0.4410	0.1017	0.0281	0.0563	0.0135	0.0872	0.0636
	E05	0.7967	0.8459	0.1407	0.0443	0.0824	0.0209	0.1342	0.0505
	E06	0.3425	0.4112	0.1059	0.0681	0.0445	0.0087	0.0991	0.0633
	E07	0.3047	0.3993	0.0761	0.0297	0.0474	0.0264	0.0976	0.0301
	E08	0.3401	0.4303	0.1025	0.0313	0.0645	0.0194	0.1198	0.0723
	E09	0.5030	0.6296	0.1268	0.0397	0.0740	0.0395	0.1191	0.0524
	E10	0.1883	0.2463	0.1919	0.1290	0.1094	0.0652	0.1296	0.0572
	E11	0.2892	0.4186	0.0971	0.0348	0.0572	0.0225	0.0992	0.0424
	E12	0.5458	0.6635	0.1287	0.0532	0.0618	0.0191	0.1187	0.0414
Group 2	E13	0.1357	0.0543	1.0000	1.0000	0.1018	0.0474	0.0970	0.0191
	E14	0.0986	0.0348	0.2782	0.3491	0.0721	0.0299	0.0701	0.0336
	E15	0.1001	0.0333	0.2043	0.2959	0.0744	0.0454	0.0852	0.0580
	E16	0.0853	0.0209	0.2364	0.3644	0.0757	0.0773	0.0625	0.0313
	E17	0.0947	0.0347	0.2897	0.3559	0.0804	0.0441	0.0633	0.0141
Group 3	E18	0.0842	0.0217	0.1018	0.0474	1.0000	1.0000	0.0773	0.0594
	E19	0.1397	0.0562	0.1091	0.0263	0.1891	0.2917	0.0834	0.0437
	E20	0.1142	0.0326	0.0818	0.0344	0.1576	0.2055	0.0719	0.0380
	E21	0.0889	0.0257	0.0750	0.0351	0.1424	0.2102	0.0675	0.0526
Group 4	E22	0.1294	0.0589	0.0970	0.0191	0.0773	0.0594	1.0000	1.0000
	E23	0.1941	0.0573	0.1556	0.0217	0.1145	0.0607	0.2288	0.2808
	E24	0.1691	0.0448	0.1255	0.0177	0.0832	0.0412	0.3660	0.4700
	E25	0.1910	0.1073	0.1294	0.0299	0.1010	0.0370	0.1540	0.1834
	E26	0.1149	0.0648	0.0856	0.0139	0.0547	0.0323	0.1281	0.1761

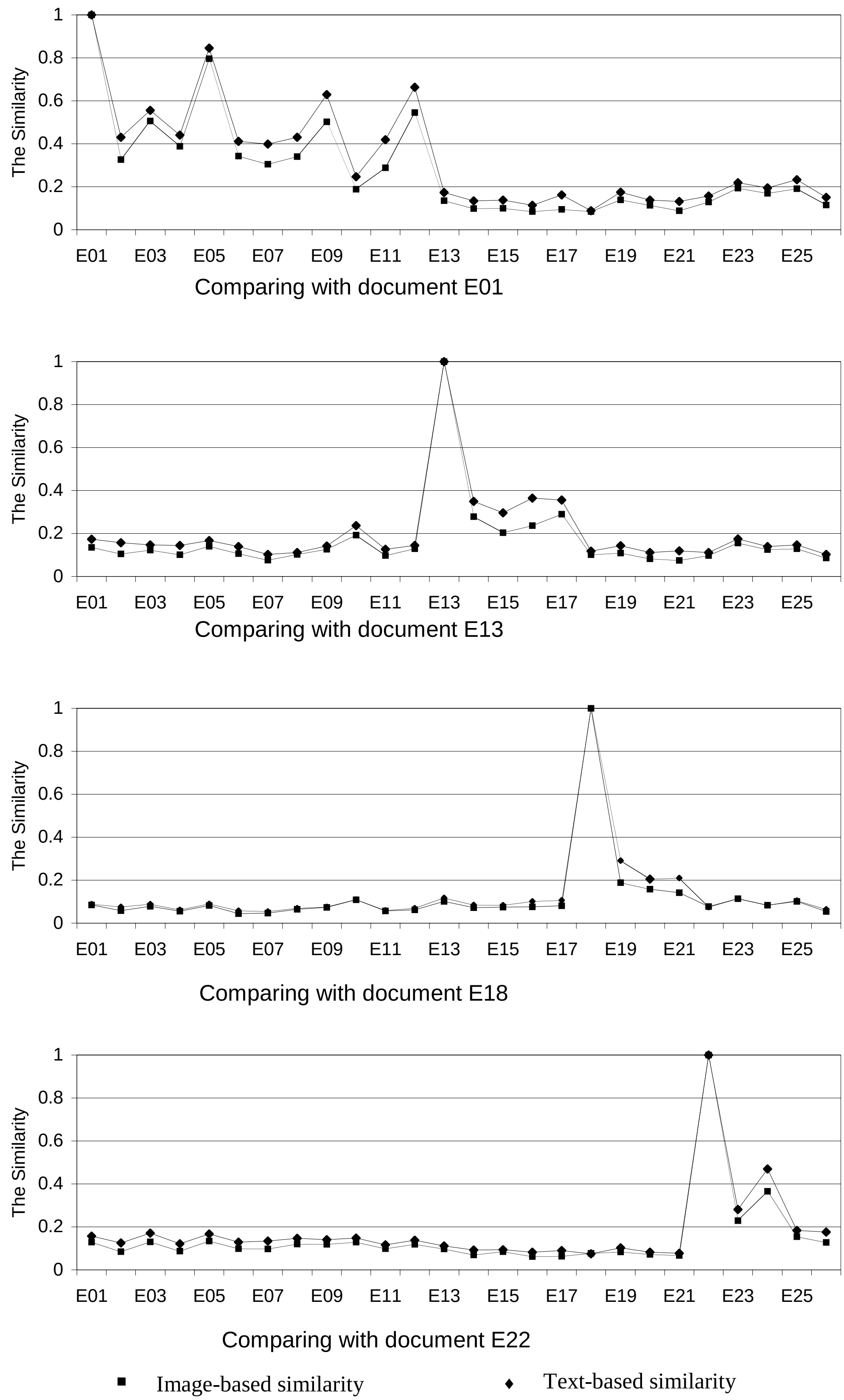


Figure 4. Comparison of Image-based and Text-based Similarity for Corpus One

Wanted: A Japanese Bill Gates

By KWAN WENG KIN
JAPAN CORRESPONDENT

THE REWARD: UP TO $1.5M EACH FOR TOP PROGRAMMERS

TOKYO — Japan's trade ministry is hoping to produce a Japanese version of American software prodigy and entrepreneur Bill Gates by offering up to 100 million yen (S$1.5 million) each to 100 genius-class programmers over the next five years.

In the past, the ministry has provided some 20 billion yen annually in assistance to the Japanese software industry, but the money was limited to corporations or research institutes.

According to a report yesterday by the influential Asahi Shimbun daily, the new policy to extend financial aid directly to promising individual programmers was the decision of Trade Minister Kaoru Yosano.

Mr Yosano was reportedly convinced that great ideas and software could spring only from the creativity of individuals and not through lengthy consultations or majority decisions by committees.

His ministry's initiative is, in fact, a challenge to Japanese programmers to emulate the example of Mr Gates — who developed the MS-Basic programming language at the age of 19 in just eight weeks — and come up with outstanding software with universal acceptance.

Twenty awards in 10 software areas, including operating systems, picture processing and cipher technology, are to be given out in the first year.

To minimise the risk of sifting out unusual candidates, the ministry will abandon the traditional method of selection by committee.

Instead, 10 professors, each an expert in one of the 10 software areas, will be given full authority to pick the candidates.

Candidates will be chosen purely on the basis of their ideas and programming abilities and may even be high school or university students, the ministry was quoted as saying.

The successful programmers will receive money to cover equipment and research staff necessary for software development.

If these software geniuses require management back-up as well, they will be paired up with venture capitalists.

Japan is not short of programmers but many are employed by companies that specialise in churning out customised, run-of-the-mill management software.

Operating systems, the basic software that runs computers, remain the virtual monopoly of US companies.

The Japanese are also behind the Americans in Internet technology. But Japanese programmers have shown themselves to be capable of turning out excellent game software.

Bangkok wants Thais to holiday at home

Thais spent S$1.3 billion abroad in the first five months of this year and Bangkok is trying to stop the worrying trend by promoting domestic tourism

By TIPAYA POONUM
IN BANGKOK

EFFORTS to boost tourism — a vital sector for Thailand's economic recovery — has been undermined by high-spending Thais holidaying abroad who have reportedly spent close to S$1.3 billion in just five months.

According to the Tourism A TAT official told The Straits Times: "Thais are taking out precious foreign exchange, which can be used to revive the economy."

He said that if the authorities did not act soon, there could be a severe drawdown of foreign reserves, to the tune of 62 billion baht by the end of the year.

The total number of Thais going abroad this year is estimated at 1.7 million, up from last year's 1.4 million.

Tourism revenue last year amounted to 242 billion baht, but 59 billion baht were drained out of the system as a result of the Thai people's overseas spending.

Tourism is a major foreign-exchange earner, representing more than 21 per cent of the Gross Domestic Product last year.

The government is fighting back to reduce the number of outbound travellers, by authorising 185 million baht to promote domestic tourism.

The year-long "Thais tour Thailand" campaign is expected to cut the number of outbound travellers by 20 per cent and save 11 billion baht in foreign exchange.

A similar campaign last year succeeded in reducing outbound travel by 15 per cent.

Spurred by recent reports of economic recovery, Thais,

Hun Sen set to meet Annan over tribunal

Cambodia is warned that the international community will not accept Khmer Rouge trials that exclude the United Nations

By KAY JOHNSON
IN PHNOM PENH

THE UNITED Nations Secretary-General will meet Cambodian Prime Minister Hun Sen later this month to break the deadlock on establishing a mixed tribunal to try leaders of the 70s Khmer Rouge regime for crimes against humanity, a top UN official said yesterday.

The international community will not accept Khmer Rouge trials that do not include the UN, warned Mr Ralph Zacklin, leader of a mission, which is heading back to New York today without an agreement on a first-of-its-kind mixed tribunal.

"The government... would like to have this process recognised as legitimate by the international community, and only the UN can provide that kind of legitimacy," he said yesterday.

With the recent talks deadlocked, the ball is effectively in Mr Hun Sen's court.

He is due to travel to New York on Sept 20 to address the UN General Assembly.

Mr Zacklin, the UN assistant secretary-general for legal affairs, said that the world body would pull out of the process rather than involve itself in any "show trial" that might exclude allies of the Cambodian Premier.

"We are not interested in taking part in a process which will just judge one or two people in order to put aside crimes that have been committed," he said.

"I think that justice in Cambodia is long overdue."

> *'I think that justice in Cambodia is long overdue.'*
> — Mr Zacklin, the UN assistant secretary-general for legal affairs

The UN tough talk was the latest move in stumbling progress this year towards addressing finally atrocities of the Khmer Rouge's horrific 1975-79 rule over Cambodia.

The end of Cambodia's long civil war has finally put leaders of the regime within reach, but Mr Hun Sen has rejected an international tribunal similar to those for Rwanda and Yugoslavia.

Instead, his government insists that trials should take place in Cambodian courts with international financial and legal assistance.

Human-rights groups and the UN fear that such a trial would be politically in-

Plan to send more police to Dili

JAKARTA — Indonesian police plan to send reinforcements to East Timor to boost shaky security before the announcement of the result of the UN-sponsored vote, the Antara state news agency said yesterday.

East Timor police chief Colonel Timbul Silaen was quoted as saying the reinforcements were partly in anticipation of a possible violent reaction to the result announcement scheduled for around next Tuesday.

"Within a short time we will add five companies of men to the existing force and to anticipate the worst possibility after the result announcement," said Colonel Silaen. One company comprises 125 to 140 men.

"There are many rumours in Dili but we don't expect any problems," he added.

Meanwhile, armed forces chief General Wiranto said Indonesia would pull out its troops from East Timor if the territory's landmark vote went against autonomy.

The government "will definitely recall the troops if the result of the ballot is won by the pro-independence side", he told the state Antara news agency in Malang, East Java, yesterday.

"However, the recall process has to go through a procedure... and the People's Consultative Assembly's decree on East Timor's integration with Indonesia has to be revoked first," Gen Wiranto said.

The People's Consultative Assembly (MPR), Indonesia's supreme legislative body, must ratify the decision East Timorese made on Monday on whether to accept Jakarta's offer of autonomy.

Gen Wiranto did not give a time-frame for the troop pullout, but Indonesia said last week that it planned a gradual withdrawal of its forces over three to six months if autonomy was rejected. — AFP

Tensions rise after vote in East Timor

Pro-integration militias renew campaign of intimidation while their leaders threaten to derail peace talks

By SUSAN SIM
INDONESIA CORRESPONDENT

DILI — Tensions began rising again as East Timor hovered on the brink of independence following Monday's peaceful vote, which saw an almost 99 per cent turnout.

The prospect of separation from Indonesia led pro-integration militias to renew their campaign of intimidation while their political leaders threatened to derail reconciliation talks unless there was a repeat ballot — charging irregularities by allegedly biased United Nations staff on Monday.

But even Jakarta seemed to dismiss their allegations. Foreign Minister Ali Alatas described the vote as constituting "a free and peaceful and therefore fair execution of the consultation".

On the irregularities, which were also highlighted by government representatives in Dili, he said after meeting President B.J. Habibie: "I would much rather concentrate on the total picture, which is on the whole a good picture ... despite all the doomsday predictions about violence."

The ground situation here deteriorated after polls closed, when suspected militiamen stabbed to death a local Unamet staff in the Emera sub-district, UN officials said.

They also confirmed that 150 Unamet staff were prevented from leaving the town of Gleno for several hours by

But pro-integration spokesman Basilio Dias Araujo charged bias on the part of the Unamet for not allowing their agents to observe the vote and claimed that pro-independence supporters among the electoral staff were allowed to exercise undue influence on voters.

He announced yesterday that until their complaints were looked into, his group was "withdrawing temporarily from any further negotiations and arrangements in relation to the establishment of a consultative board, the counting of the ballot and futher arrangements after the ballot".

He later told The Straits Times that his United Front for East Timor Autonomy wanted a second ballot, this time with their representatives present.

Denying that his stance was a negotiating tactic in the face of likely defeat, he said: "We want a fair win or a fair defeat, but not defeat by

Russian graft probe moves to Switzerland

Money-laundering scandal, which may involve aid funds, prompts US congressman to say that the West should re-examine its ties with Russia

ZURICH — A top Russian investigator has arrived in Switzerland to look into allegations of corruption among top Russian officials.

A United States congressman, meanwhile, said that the money-laundering scandal, which might have involved funds from the International Monetary Fund (IMF) intended as aid, meant the West should re-examine its ties with Russia.

he added, but gave no details.

He said Mr Volkov was probing whether "several high-placed Russian officials" were corrupt or had abused their office.

Swiss prosecutors are already investigating possible corruption cases involving various firms with Russian links, including the awarding of contracts to Swiss construction firm Mabetex to renovate Kremlin buildings.

probe, which has been billed as the biggest money-laundering scandal ever.

Bank of New York chairman Thomas Renyi told employees in an internal memo on Monday that the bank was examining its controls on cash and funds transfers.

Some newspaper reports said the alleged laundering could include money from the IMF, although its head of external affairs said on Monday he had no proof of this.

Mr Jim Leach, chairman of the House Banking Committee, said on Monday the West needed to re-examine its ties with Russia in the light of the new probes.

"I know of very few issues in international affairs that

Vigilantes taking no chances

Young fighters armed with knives are back on the streets guarding their neighbourhoods as ballot boxes are brought into Dili from the regions

DILI — The young men with the long hair and the camouflage jackets were back on duty yesterday in their east Dili neighbourhood, taking no chances against a return of the feared pro-Indonesian militia.

"We're still afraid because there's still terror," said Mr Cancio Pareira, 25, one of a self-appointed force of young men guarding streets in the Becora area where six people died in militia carnage last Thursday.

Mr Pareira and his comrades were gathered beside the road at the quiet Becora bus terminal as East Timorese began the nervous wait for the results of Monday's autonomy ballot to be announced, probably in a week's time.

ly-rebuilt remains of his village that was attacked by militia in April, he said he and his friends wore military fatigues because their other clothes were lost in the April violence.

In the hills outside Dili, pro-independence youths armed themselves with knives.

"We have no guns," said one 18-year-old.

A large group of Aitarak militiamen was seen burning material in front of the ruined office of the pro-independence National Resistance Council of East Timor (CNRT) yesterday morning.

The militiamen, many armed with guns, had ransacked the office last Thursday.

Journalists also saw sev-

Armed with machetes, pro-independent East Timc district against the rumoured return of the feared

where the vote count is to Aitarak members, who

Jakarta rushes troops to E. Timor

Elite soldiers dispatched after UN request for protection; result of autonomy referendum out today

By SUSAN SIM
INDONESIA CORRESPONDENT

DILI — With pro-integration militias already in control of some towns across East Timor, the authorities here are bracing themselves for more violence when the result of an autonomy referendum is announced today.

Fear that the militias — until now allowed to intimidate and terrorise those who support independence and United Nations staff with impunity — will attempt to

Australian Foreign Minister Alexander Downer that planning was underway for such a force if the 430,000 East Timorese voted for independence.

Unamet spokesman David Wimhurst reported yesterday that militias in Maliana "rampaged through the town all night" and torched houses, forcing UN staff to take refuge in a police station. Two more of its East Timorese staff were killed, bringing the death toll of local UN staff to four, with another six "unaccounted for".

area in the morning. Taking advantage of a lull in tensions, residents went about their business again.

Police patrols appeared to have been stepped up, especially in the main shopping street, where shopkeepers said they felt more at ease.

But a UN military observer told The Straits Times that the lack of incidents did not mean a change in the security assessment: "I think the situation the day before was better. Things will deteriorate, especially with the results expected soon."

The counting of votes from Monday's ballot, which saw a 99 per cent turnout, started on Thursday.

The final tally was expected to be finished last night

Figure 5. Corpus Two : Articles from English newspapers

Table 2. Comparison of Image-based and Text-based similarity for Corpus Two

		N1	N2	N3	N4	N5	N6	N7	N8	News Title
N1	*	1.000	0.284	0.162	0.210	0.099	0.119	0.142	0.159	Tensions rise after vote in East Timor
	**	1.000	0.362	0.221	0.260	0.118	0.143	0.179	0.185	
N2	*	0.284	1.000	0.192	0.235	0.104	0.112	0.121	0.148	Jakarta rushes troops to E. Timor
	**	0.362	1.000	0.276	0.276	0.119	0.127	0.164	0.168	
N3	*	0.162	0.192	1.000	0.193	0.064	0.086	0.125	0.082	Plan to send more police to Dili
	**	0.221	0.276	1.000	0.224	0.084	0.095	0.146	0.104	
N4	*	0.210	0.234	0.193	1.000	0.097	0.098	0.115	0.134	Vigilantes taking no chances
	**	0.260	0.276	0.224	1.000	0.108	0.114	0.146	0.165	
N5	*	0.099	0.104	0.064	0.097	1.000	0.089	0.081	0.091	Wanted: A Japanese Bill Gates
	**	0.118	0.119	0.084	0.108	1.000	0.117	0.109	0.107	
N6	*	0.119	0.112	0.086	0.098	0.089	1.000	0.120	0.092	Bangkok wants Thais to holiday at home
	**	0.143	0.127	0.095	0.114	0.117	1.000	0.137	0.103	
N7	*	0.142	0.121	0.125	0.115	0.081	0.120	1.000	0.164	Hun Sen set to meet Annan over tribunal
	**	0.179	0.164	0.146	0.146	0.109	0.137	1.000	0.201	
N8	*	0.159	0.148	0.082	0.134	0.091	0.092	0.164	1.000	Russian graft probe moves to Switzerland
	**	0.185	0.168	0.104	0.165	0.107	0.103	0.201	1.000	

Notes:

*: Image-based Similarity

**: Text-based Similarity

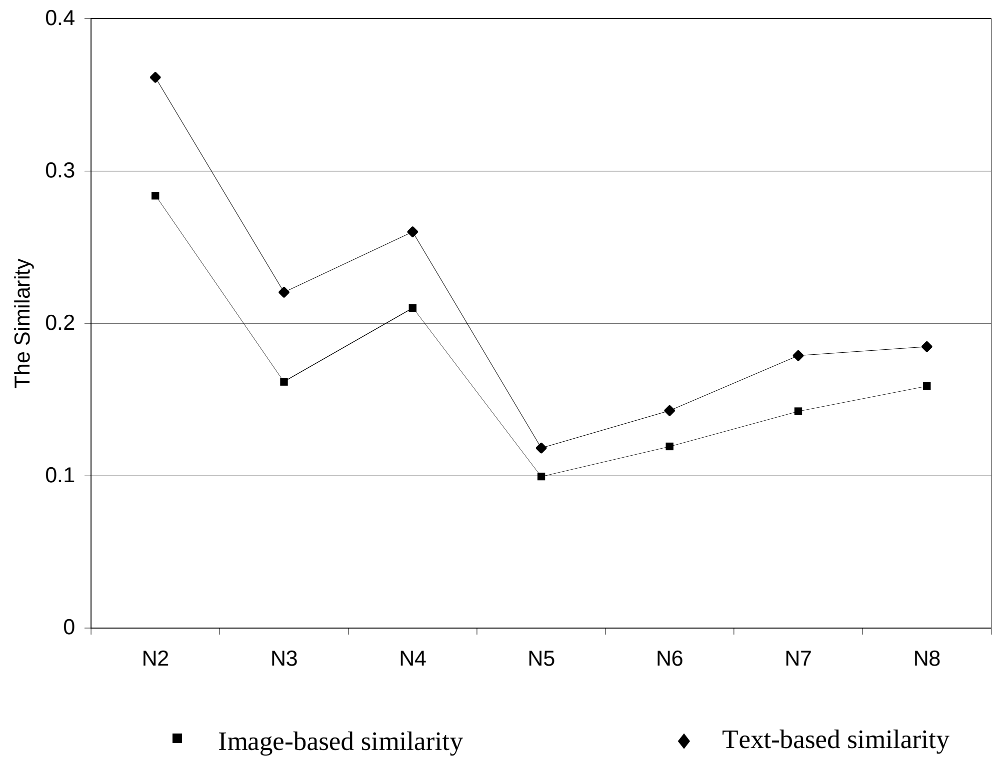


Figure 6 Comparison of image-based similarity and text-based similarity between N1 and other articles

马哈迪：如果新加坡也准备妥协

马愿在双边课题妥协与新寻求解决方案

"如果我们每一次都是妥协的一方，这些课题将难以解决。"
新加坡应该同意，丹戎巴葛马来亚铁道公司土地的课题应该在同一配套中解决，而不是分开解决。
——马哈迪

外长否认巫统党选影响新马问题谈判

旅游部长：若各领域全力配合

马三年内可成购物天堂

马哈迪：带动柔南经济成长

柔第二港口是马经济发展催化剂

每年处理380万集装箱

北京再警告台北 统一要有时间表

正视中国人民的统一意愿

——评蔡玮对两岸问题的分析

（加州）陈阿蕊

台湾纵容分裂势力

成为反华的马前卒

谈"'台独'即意味战争"

充分估计美国动手的可能性

中国军力不容低估

国民党的政权在台湾能延续长达50年，在很大程度上是靠美国人撑腰。

表面上，美国在台湾这场险

Figure 7. Corpus Three – Articles from Chinese newspapers

Table 3. Comparison of Image-based and Text-based similarity for Corpus Three

		C1	C2	C3	C4	C5	C6	C7	Rough Translation of News Title
C1	*	1.000	0.444	0.127	0.120	0.145	0.121	0.191	Malaysia willing to compromise to resolve issues on bilateral ties with Singapore
	**	1.000	0.417	0.231	0.177	0.094	0.130	0.166	
C2	*	0.444	1.000	0.107	0.127	0.122	0.102	0.200	Foreign Minister denies impact of Umno election on negotiation with Singapore
	**	0.417	1.000	0.191	0.194	0.116	0.140	0.197	
C3	*	0.127	0.107	1.000	0.147	0.059	0.045	0.058	Malaysia to become shoppers' paradise in three years
	**	0.231	0.191	1.000	0.227	0.102	0.135	0.089	
C4	*	0.120	0.127	0.147	1.000	0.069	0.067	0.066	The second port in Johore: a catalyst for Malaysia economic development
	**	0.177	0.194	0.227	1.000	0.084	0.123	0.085	
C5	*	0.145	0.122	0.059	0.069	1.000	0.378	0.345	Beijing warns Taipei again on the need for a schedule for reunification
	**	0.094	0.116	0.102	0.084	1.000	0.442	0.297	
C6	*	0.121	0.102	0.045	0.067	0.378	1.000	0.350	Analysis of the wish for reunification of the people in China
	**	0.130	0.140	0.135	0.123	0.442	1.000	0.285	
C7	*	0.191	0.200	0.058	0.066	0.345	0.350	1.000	Talking about "Taiwan Independence" means war
	**	0.166	0.197	0.089	0.085	0.297	0.285	1.000	

Notes:

*: Image-based Similarity

**: Text-based Similarity

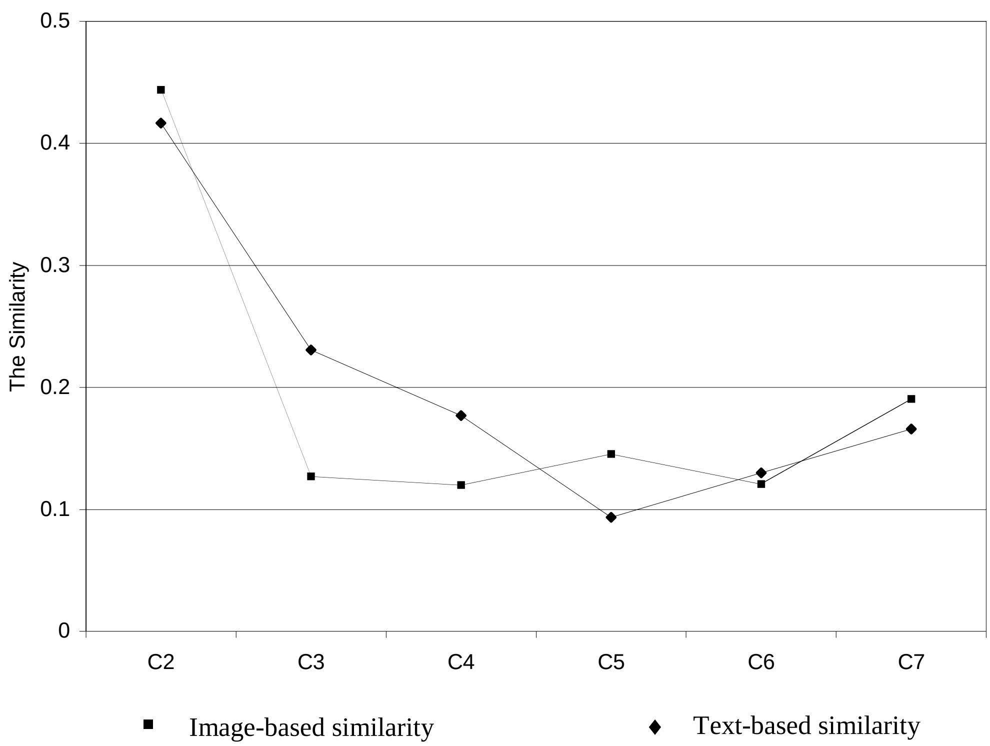


Figure 8 Comparison of image-based similarity and text-based similarity between C1 and other articles

Table 4. Accuracy, Precision and Recall of image-based document text retrieval for Corpus Four

Topic id. no. i	No. of articles on topic i (n_i)	Average similarity among n_i articles retrieved	Accuracy %	Threshold = 0.2			Threshold = 015			Threshold = 0.10		
				Average similarity	Precision %	Recall %	Average similarity	Precision %	Recall %	Average similarity	Precision %	Recall %
1	26	0.1292	72.9	0.2646	100	13.2	0.2097	95.3	22.1	0.1482	83.8	56.9
2	22	0.1995	94.8	0.2451	100	47.1	0.2153	98.2	75.3	0.1730	73.8	93.7
3	18	0.1981	94.0	0.2316	100	52.0	0.2068	96.9	82.1	0.1606	57.1	97.3
4	16	0.1940	88.4	0.2565	100	49.2	0.2252	96.7	68.0	0.1730	76.7	87.9
5	22	0.2013	82.9	0.2566	100	53.8	0.2303	96.5	67.9	0.1722	66.0	86.2
6	19	0.1697	86.2	0.2399	100	35.9	0.2099	98.2	59.9	0.1742	82.8	85.8
7	18	0.1578	85.0	0.2619	100	18.4	0.2006	98.0	54.2	0.1674	92.9	78.1
8	18	0.2764	97.7	0.2921	100	82.1	0.2763	97.2	96.0	0.2082	58.1	100
Average		0.1908	87.7	0.2560	100	44.0	0.2218	97.1	65.7	0.1721	73.9	85.7

Acknowledgements:

The authors would like to thank Mr. Lim Seng Ping and Mr. Johnson Paul of National Library Board of Singapore for project discussion and assistance in preparation of newspaper microfilm images. This project is supported by the research grant RP3992713 from the National Science and Technology Board and Ministry of Education of Singapore.

References

[1] C.L. Tan, S.Y. Sung, D. Shi, B. Yuan, Y.T. Lim, Y. Xu, “News articles retrieval from microfilm images”, IJCAI'99 Workshop: Text Mining: Foundations, Techniques and Application, Stockholm, Sweden, pp. 110-116, August 2, 1999.

[2] G. Salton, “Developments in Automatic Text Retrieval,” Science 253, pp.974-980, 1991

[3] G. Salton and C. Buckley, “Global Text Matching for Information Retrieval,” Science 253, pp.1012-1015, 1991.

[4] G. Salton, J. Allan, C. Buckley, and A. Singhal, “Automatic Analysis, Theme Generation, and Summarization of Machine-readable Text,” Science 264, pp.1421-1426, 1994

[5] C. E. Shannon, “The Mathematical Theory of Communication,” University of Illinois Press, Urbana, 1949.

[6] C. Y. Suen, “N-Gram Statistics for Natural Language Understanding and Text Processing,” IEEE Trans. on Pattern Analysis & Machine Intelligence. PAMI, 1(2), pp.164-172, April 1979.

[7] A. Zamora, “Automatic Detection and Correcting of Spelling Errors in A Large Data Base,” J. Amer. Soc. Inf. Sci. 31, 51, 1980.

[8] J. L. Peterson, “Computer Programs for Detecting and Correcting Spelling Errors,” Comm. ACM 23, 676, 1980.

[9] E. M. Zamora, J. J. Pollock, and Antonio Zamora, “The Use of Trigram Analysis for Spelling Error Detection,” Inf. Proc. Mgt. 17, 305, 1981.

[10] J. J. Hull and S. N. Srihari, “Experiments in Text Recognition with Binary N-Gram and Viterbi Algorithms,” IEEE Trans. Pattern Analysis & Machine Intelligence, PAMI-4, 520, 1980.

[11] J. J. Pollock, “Spelling Error Detection and Correction by Computer: Some Notes and A Bibliography,” J. Doc. 38, 282, 1982.

[12] R. C. Angell, G. E. Freund, and P. Willette, “Automatic Spelling Correction Using Trigram Similarity Measure,” Inf. Proc. Mgt. 18, 255, 1983.

[13] E. J. Yannakoudakis, P. Goyal, and J. A. Huggill, “The Generation and Use of Text Fragments for Data Compression,” Inf. Proc. Mgt. 18, 15, 1982.

[14] J. C. Schmitt, “Trigram-based Method of Language Identification,” U.S. Patent No. 5,062,143, 1990.

[15] W. B. Cavnar and J. M. Trenkle, “N-Gram-based Text Categorization,” Proceeding of the Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas, 1994.

[16] P. Willett, “Document Retrieval Experiments Using Indexing Vocabularies of Varying Size. II. Hashing, Truncation. Digram and Trigram Encoding of Index Terms.” J. Doc. 35, 296, 1979.

[17] W. B. Cavnar, “N-Gram-based Text Filtering for TREC-2,” The Second Text Retrieval Conference (TREC-2), NIST Special Publication 500-215, National Institute of Standards and Technology, Gaitherburg, Maryland, 1994.

[18] Marc Damashek, “Gauging Similarity via N-Grams: Language-independent Sorting, Categorization, and Retrieval of Text,” Science, 267, pp.843-848, 1995.

[19] F. R. Chen and D. S. Bloomberg, “Extraction of Thematically Relevant Text from Images”, Proceedings of the Symposium on Document Analysis and Information Retrieval, pp.163-178, 1996.

[20] F. R. Chen, D.S. Bloomberg, "Extraction of Indicative Summary Sentences from Imaged Documents," Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97), Volume 1, pp.227-232, 1997.

[21] J. J. Hull; J. F. Cullen, "Document Image Similarity and Equivalence Detection," Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97), Volume 1, pp. 308-312, 1997.

[22] Yaodong He; Zao Jiang; Bing Liu; Hong Zhao, "Content-based Indexing and Retrieval Method of Chinese Document Images", Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99), pp.685-688, 1999.

[23] P. Sibun, and A. L. Splitz, "Language Determination: National Language Processing from Scanned Document Images," Proceedings of the fourth Conference on Applied Natural Language Processing, pp. 423-433, Las Vegas, April 1995.

[24] A. L. Spitz, "Determination of the Script and Language Content of Document Images," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, no. 3, pp.235-245, 1997.

[25] C. Y. Suen, S. Bergler, N. Nobile, B. Waked, C. P. Nadal, and A. Bloch, "Categorizing Document Images into Script and Language Classes," Proceedings of the International Conference on Advances in Pattern Recognition, Plymouth, UK, pp.297-306, 23-25 Nov 1998.

[26] C. L. Tan, P. Y. Leong, and S. He, "Language Identification in Multilingual Documents," International Symposium on Intelligent Multimedia and Distance Education, 1999

[27] A. F. Smeaton, A. L. Spitz, "Using Character Shape Coding for Information Retrieval", Proceeding of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97), Volume 2, pp.974-978, 1997.

[28] D. Doermann, Li Huiping, O. Kia, "The detection of duplicates in document image databases", Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97), Volume 1, pages 314-318, 1997.

[29] D.S. Bloomberg, G.E. Kopec and L. Dasari, "Measuring Document Image Skew and Orientation," SPIE Conf. 2422, Document Recognition II, San Jose, CA, pp.302-316, Feb 6-7, 1995.