IDIAP RESEARCH REPORT

Application of Information Retrieval Techniques to Single Writer Documents

Alessandro Vinciarelli [a]

IDIAP–RR 04-12

April 2004

to appear in
Pattern Recognition Letters

[a] IDIAP

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-12

Application of Information Retrieval Techniques to Single Writer Documents

Alessandro Vinciarelli

April 2004

to appear in
Pattern Recognition Letters

Abstract. This work shows Information Retrieval experiments performed over handwritten documents produced by a single writer. The same retrieval task has been performed over both manual (no errors) and automatic (Word Error Rate around 45%) transcriptions of 200 handwritten texts. The results show that the performance loss due to recognition errors is acceptable and that Information Retrieval technologies can be effectively applied to handwritten data.

1 Introduction

It is not possible to take advantage of a collection of documents without effective indexing and retrieval techniques. Handwritten documents contain textual information, but few efforts have been made, to our knowledge, to extend to them the technologies developed in the Information Retrieval (IR) domain (see below for a quick survey). This is due, in our opinion, to the fact that the errors in the automatic transcriptions of handwritten documents are expected to heavily affect the retrieval results. This work shows that IR approaches commonly applied to digital texts can be extended to handwritten documents written by a single person with an acceptable performance loss.
In our experiments, 200 documents belonging to the Reuters-21578 database (see Section 4 for more details) have been manually written by a single person and transcribed with an offline handwriting recognizer (Vinciarelli et al., 2004). This resulted into two versions of the same database: the first is composed of the original digital documents and it is referred to as *clean*, the second contains the transcriptions of the handwritten data (affected by a Word Error Rate of around 45%) and it is referred to as *noisy*. The same IR system has then been used over both clean and noisy data and the results show that the retrieval performance loss is acceptable.
The use of 200 documents written by a single person can correspond to several application domains: personal handwritten notes, letter collections (Rath and Manmatha (2003) describe applications working on the letters written by G.Washington), medical reports, etc.. On the other hand, the identity of the writer is not used in the retrieval process and the only parameter actually affecting the retrieval performance is the Word Error Rate (WER). Our results show that good retrieval results can be obtained at a WER of around 45%. If a handwriting recognition system is able to achieve such a performance on multiple writer data, equivalent retrieval results can be obtained also over them.
The only approach to handwritten document retrieval applied so far is, to our knowledge, *Word Spotting* (WS), i.e. the detection of words belonging to a query in the documents. In some cases, the words are searched after the documents have been recognized and several techniques have been proposed to make WS more robust with respect to recognition errors: Kwok et al. (2000) convert each handwritten word into a stack of scores related to the dictionary entries. Russell et al. (2002) use the N best recognizer outputs to expand the transcriptions and associate a probabilistic score to each term. In other cases, the recognition is avoided and WS is performed by matching query word images with the word images extracted from the documents (Jain and Namboodiri, 2003; Kolcz et al., 2000; Rath and Manmatha, 2003; Tomai et al., 2002; Uchiashi and Wilcox, 1999). Word Spotting has two main disadvantages: the first is that morphological variants of the same word (e.g. *start* and *starting*) are considered different even if they have the same meaning. The second is that all of the words are given the same weight even if certain terms are more representative of the document content than others.
Current IR approaches solve such problems (see Section 2) and have been shown to be more effective than simple Word Spotting (Baeza-Yates and Ribeiro-Neto, 1999). For this reason, we propose in this work to apply IR technologies to the automatic transcriptions of handwritten data. Moreover, we evaluate the effect of the recognition errors on the retrieval performance by comparing the results obtained, using the same system, over both clean and noisy data.
The rest of this paper is organized as follows: section 2 presents the IR system used in our work, section 3 illustrates the handwriting recognition approach applied, section 4 describes experiments and results, and section 5 draws some conclusions.

2 Information Retrieval

The literature presents essentially three IR models: the first is called *boolean*, the second is known as *Vector Space Model* (VSM) and the third is referred to as *probabilistic* (Baeza-Yates and Ribeiro-Neto, 1999). The boolean model is based on binary algebra: the queries are expressed as logical conditions (e.g. $keyword_1$ *and* $keyword_2$ *and not* $keyword_3$) and the systems retrieve all documents satisfying

them (in the case of the previous query, the system will retrieve the documents containing keywords 1 and 2, but not containing keyword 3). The limit of this approach is that it can be difficult to express a complex information need through a logic expression. Moreover, the system answer is binary (this does not allow partial matching with the query). The boolean model has been the first retrieval approach proposed in the literature, but it is now considered obsolete, better results can be otained with the other approaches (Baeza-Yates and Ribeiro-Neto, 1999; Van Rijsbergen, 1979).
The probabilistic approaches try to estimate the probability of a document being relevant to a certain query. The problem is that this requires a large amount of training queries and this is difficult to obtain (Baeza-Yates and Ribeiro-Neto, 1999; Van Rijsbergen, 1979). For this reason, we selected the VSM which is the most widely applied approach and it allows one to achieve state-of-the-art performances over the main benchmarks presented in the literature (Baeza-Yates and Ribeiro-Neto, 1999). In the VSM, the documents are represented as vectors and their relevance to the queries submitted by the users is measured through appropriate matching functions. The IR process has two major components: the first is the extraction of the *term by document matrix* and it is performed once for a given database. The second is the identification of the documents relevant to a query and it is performed each time a query is submitted to the system. In the next subsections, retrieval process and related performance measures are described in detail.

2.1 Term by Document Matrix Extraction

The term by document matrix A is obtained through several steps: *preprocessing*, *normalization* and *indexing*.
Preprocessing removes all elements supposed to be unuseful in a retrieval process. In our system, all non-alphabetic characters (digits, punctuation marks, etc.) are eliminated. This solution has several disadvantages (e.g. an expression like *state-of-the-art* is transformed into *state of the art* and it is no longer considered as a single term), but it is simple and it allows one to achieve, on average, good performances over all databases (Fox, 1992).
The normalization removes the variability that is not useful to the retrieval process and it is performed through two steps: *stopping* and *stemming*. During stopping, all the words expected to be poor index terms (called *stopwords*) are removed. Stopwords are typically functional words (articles, preprositions, pronouns, etc.) and words of common use (*to be*, *good*, *to do*, etc.). Stopping results, on average, in the removal of around 50% of the original document words (Fox, 1992).
Stemming is the replacement of different inflected forms of a certain word (e.g. *connection*, *connected*, *connecting*) with their *stem* (*connect*). In our system, we use the stemming algorithm proposed by Porter (1980). This technique is widely applied and represents a good trade-off between complexity and effectiveness. After the stemming, the dictionary (list of all unique terms appearing in a document collection) size is reduced, on average, by ∼30%.
After preprocessing and normalization, the original documents have been converted into streams of *terms*, but this is not a suitable form for the retrieval process. An indexing procedure is necessary in order to convert the documents into vectors. The result is the term by document matrix A where each column j corresponds to a document and each row i corresponds to a term of the dictionary. The generic element A_{ij} can be written as follows:

$$A_{ij} = L(i,j) \cdot G(i) \tag{1}$$

where $G(i)$ is a global weight taking into account information extracted from the whole database and $L(i,j)$ is a local weight based on information coming from the only document j. An extensive survey about weighting schemes has been provided by Salton and Buckley (1988). In this work we apply the $tf \cdot idf$ scheme:

$$A_{ij} = tf(i,j) \cdot \log\left(\frac{N}{N_i}\right) \tag{2}$$

where the local weight $tf(i,j)$ is the term frequency of term i in document j (i.e. the number of times term i appears in document j), N is the total number of documents in the database, and N_i

is the number of documents containing term i. The logarithm is called *inverse document frequency* (*idf*) and it has higher value for terms appearing in fewer texts. The $tf \cdot idf$ is the most applied weighting scheme (Baeza-Yates and Ribeiro-Neto, 1999; Aizawa, 2003) and it embodies the intuition that the more a term appears in a document, the more it is representative of its content and that terms appearing in few documents allow better discrimination between different texts.

2.2 Document Retrieval

The retrieval step identifies the documents relevant to a given query q, i.e. the documents answering the information need the query q expresses. In the VSM, documents and queries are matched through appropriate measures associating a Retrieval Status Value (RSV) to each query-document couple (q, d). For a given query q, the documents can be ranked according to their RSV values and the texts at the top positions of the ranking are identified as relevant.
Several matching measures have been proposed in the literature and state-of-the-art systems use typically the cosine or the Okapi (Robertson et al., 2000) measure that are the most effective and commonly applied ones (Baeza-Yates and Ribeiro-Neto, 1999). Both above measures have been used in this work in order to make our experiments more complete. The cosine of the angle between query and document vectors is calculated through their inner product:

$$RSV(q,d) = \frac{\mathbf{q} \cdot \mathbf{d}}{||\mathbf{q}|| \cdot ||\mathbf{d}||} = \frac{1}{||\mathbf{q}|| \cdot ||\mathbf{d}||} \sum_{i=1}^{T} q_i d_i \tag{3}$$

where T is the size of the dictionary (i.e. the number of unique terms appearing in the document database) and q_i and d_i are the i^{th} components of the query and document vectors respectively. The cosine is null when the vectors are orthogonal, i.e. they do not have any term in common. The Okapi measure (Robertson et al., 2000) calculates the RSV as follows:

$$RSV(q,d) = \sum_{\{l:t_l \in Q\}} \frac{(K+1) \cdot tf(l,d) \cdot idf(l)}{K \cdot [(1-b) + b \cdot NDL(d)] + tf(l,d)} \tag{4}$$

where K and b are hyperparameters[1], Q is the set of the terms belonging to the query, $idf(l)$ is the inverse document frequency of term l (see previous section) and $NDL(d)$ is the length of d divided by the average document length in the database.

2.3 Performance Evaluation

Given a query q, the set of its relevant documents is $R(q)$ and the set of the documents identified as relevant by the system is $R^*(q)$. The retrieval performance measures are based on *Precision* (the probability of a document identified as relevant by the system being actually relevant):

$$\pi(q) = \frac{|R(q) \cap R^*(q)|}{|R^*(q)|} \tag{5}$$

and *Recall* (the probability of an actually relevant document being identified as such by the system):

$$\rho(q) = \frac{|R(q) \cap R^*(q)|}{|R(q)|}. \tag{6}$$

Based on π and ρ, it is possible to obtain several measures accounting for different aspects of the system performance. In this work, we will use *Precision vs Recall curves*, *Break Even Point* (BEP),

[1] K and b must be set using queries and document corpora different from those used in the experiments. In this work, the values of K and b are those that give the highest performance over the queries proposed in the TREC conferences for the Wall Street Journal corpus (Baeza-Yates and Ribeiro-Neto, 1999).

average Precision (avgP) and *Precision at position n* (P_n).
Precision vs Recall curves are commonly applied and provide a general evaluation of the system: π is measured at predefined ρ values $(10\%, 20\%, \ldots, 100\%)$ for each query, then the average of the resulting curves is used to measure the overall performance. In order to make easier the comparison between different systems, the Precision vs Recall curves are often resumed with single figures: average Precision and Break Even Point. The first is the average π along the curves, the second is the point of the curves where $\pi = \rho$. When using P_n, the set $R^*(q)$ corresponds to the first n positions of the ranking and π is calculated with equation 5. A curve of P_n as a function of n can be obtained for each query and then an average of all curves can be used to measure the system performance over a set of queries.

3 Handwriting Recognition

The Handwriting Recognition system used in this work has been presented by Vinciarelli et al. (2004) and it is briefly described in this section.
Our recognition approach is based on Hidden Markov Models (HMM) and Statistical Language Models (SLM). The lines must be transcribed separately because of computational constraints (see Vinciarelli et al. (2004)): each line image is first preprocessed and normalized with the technique proposed by Vinciarelli and Lüttin (2001), then it is converted into a sequence of vectors $O = (o_1, o_2, \ldots, o_M)$ through a sliding window approach: a fixed width window shifts column by column from left to right and, at each position, a feature vector is extracted (the feature extraction process is described by Vinciarelli and Lüttin (2000)).
Given O, the system finds the sequence $\hat{W}$ of words belonging to the dictionary that maximizes the probability $p(W|O)$:

$$\hat{W} = \arg\max_W \frac{p(O|W) \cdot p(W)}{p(O)} \tag{7}$$

and since $p(O)$ is constant during the recognition, the last equation can be rewritten as follows:

$$\hat{W} = \arg\max_W p(O|W) \cdot p(W). \tag{8}$$

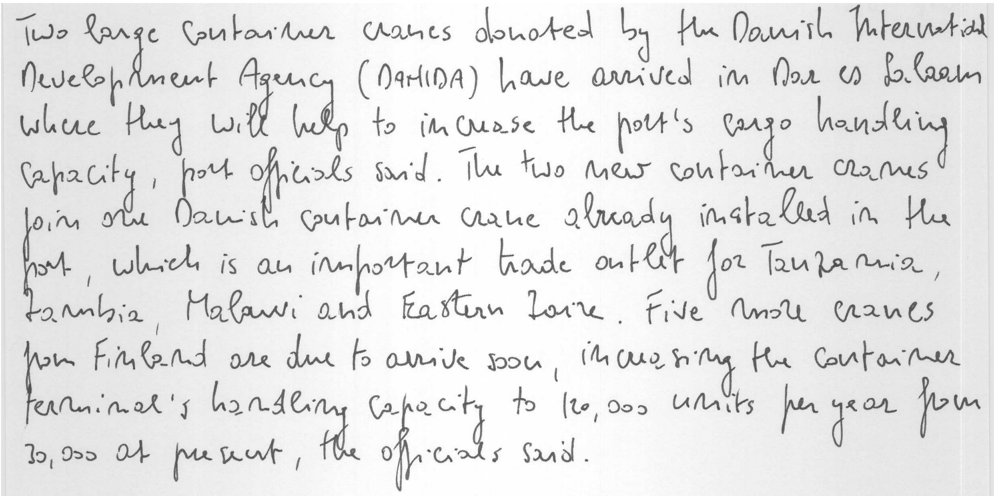
The right side of equation 8 shows the role of the two sources of information available during the recognition process. Term $p(O|W)$ is the probability of sequence O being generated given word sequence W and it is estimated with continuous density HMMs. Term $p(W)$ is the probability of word sequence W being written and it is estimated with N-grams, the most successful and widely applied Statistical Language Model (Rosenfeld, 2000).
The main advantage of this approach is that no segmentation of the line into words is required. The segmentation of the line is in fact obtained as a side product of the recognition (Vinciarelli et al., 2004). On the other hand the line by line recognition limits the performance of the language models (see Vinciarelli et al. (2004) for more details).

4 Experiments and Results

This section presents the retrieval experiments performed in this work. A set of 200 handwritten documents has been collected and the same retrieval task (20 queries) has been performed over both their manual and automatic transcriptions. The results have been compared in order to evaluate the effect of the recognition errors on the retrieval performance.
In the next subsections the data and the retrieval experiments are described in detail.

4.1 The Data

The data used in our experiments is based on the Reuters-21578 database (Lewis, 1992), a well known and widely applied benchmark publicly available on the web. The text collection has been split into



Two large container cranes donated by the Danish International Development Agency (DANIDA) have arrived in Dar es Salaam where they will help to increase the port's cargo handling capacity, port officials said. The two new container cranes join one Danish container crane already installed in the port, which is an important trade outlet for Tanzania, Zambia, Malawi and Eastern Zaire. Five more cranes from Finland are due to arrive soon, increasing the container terminal's handling capacity to 120,000 units per year from 30,000 at present, the officials said.

Figure 1: Example of handwritten document in the data set.

training and test set following the *ModApté split* (Apté et al., 1994). The training set has been used to train the Statistical Language Models (bigrams) and to extract a 20,000 words lexicon for the handwriting recognition system. A set of 250 documents (belonging to the 10 most represented categories) has been randomly selected from the test set. The use of documents belonging only to 10 categories does not affect the retrieval performance because the category of the document is not taken into account in the retrieval process. The queries used in our experiments (see below) have been created appositely for this work and have no relationship with the categories provided with the Reuters-21578 database.

The documents have been manually written by a single person (see Figure 1 for a sample) and randomly split into two subsets containing 50 and 200 documents respectively. The smaller subset has been used to train the handwriting recognition system (the details are provided by Vinciarelli et al. (2004)) used to automatically transcribe the 200 remaining documents. This experimental setup allows a rigorous separation between the data used for training and the data used for test: the 50 documents used to train the handwriting recognition system are completely independent of the 200 documents of the test set. The same applies to the linguistic knowledge: lexica and SLMs have been obtained from the training set of the Reuters-21578 database and the documents belonging to it are separated and independent from the 200 documents of the test set, thus no information extracted from the test set is used for the SLM training.

The retrieval experiments are performed over both manual (clean) and automatic (noisy) transcriptions of the above 200 documents. The WER of the noisy documents is 44.2%. This represents an overestimation of the noise because the WER takes into account the order of the words (that is not considered in the VSM) and some errors that have no effect on the retrieval process (e.g. the transcriptions of stopwords into other stopwords). For this reason we propose to use the Term Error Rate (TER), i.e. the error rate measured after stopping and stemming (see section 2) without considering the word order:

$$TER = 1 - \frac{\sum_i \min(tf(i), tf^*(i))}{\sum_k tf(k)} \tag{9}$$

where $tf(i)$ ($tf^*(i)$) is the number of times term i appears in the clean (noisy) document. The average TER is 40.7% and Figure 2 shows its variability across different documents. In some cases, TER=100% and all the information of the original document is lost. In order to verify that, although the high TER, the documents still contain enough information to effectively apply IR technologies, we performed some preliminary experiments based on artificial queries obtained with the Rocchio

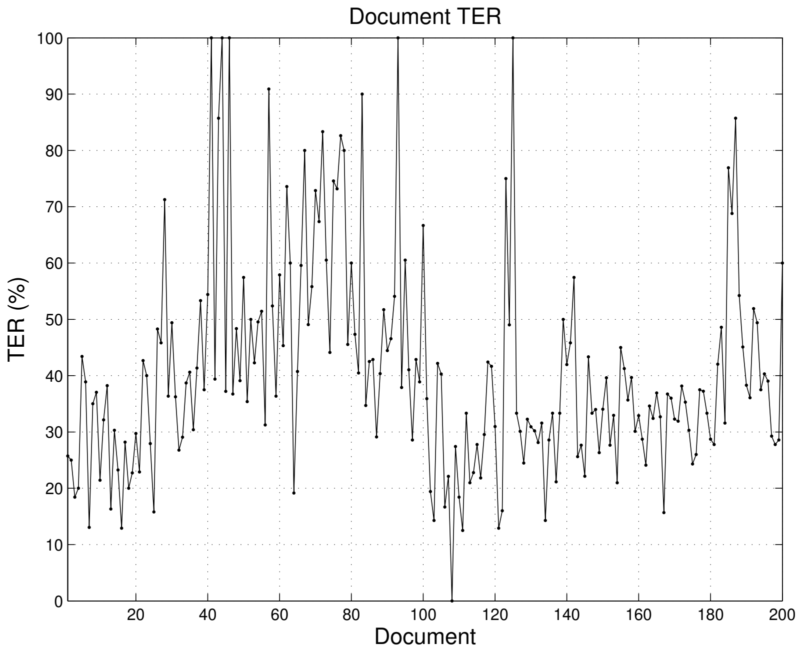


Figure 2: This plot shows the Term Error Rate per each document of the data set. When the TER is 100% the information contained in the handwritten document is completely lost.

formula (Rocchio, 1971). The results are presented by Vinciarelli (2004) and they show that it is possible to expect a reasonable performance loss (with respect to the clean texts) when using real queries. The confirmation of such hypothesis is given by the experiments performed in the next section.

4.2 Retrieval Experiments

The experiments described in this section are based on a set of 20 queries and related relevance judgements. The number of relevant documents changes significantly depending on the case (see figure 3): most of the queries have no more than four relevant documents (accounting for 2% of the database) and can thus be considered difficult. The queries are expressed in natural language (this is one of the main advantages of the VSM) and are reported in Table 1. The queries are typically very short texts and they are processed in the same way as the documents of the database (i.e. they are stopped, stemmed and indexed as described in Section 2) before the matching. The performance of a retrieval system can significantly change depending on the queries used (Salton et al., 1975), but the creation of a query set and related relevance judgements (i.e. the list of documents relevant to the queries) is difficult and time consuming. For this reason, the IR experiments never involve more than few tens of queries. The biggest benchmarks are prepared for the TREC conferences and are typically based on sets of 50 queries (Baeza-Yates and Ribeiro-Neto, 1999). However, the purpose of our work is the comparison of the system performance when passing from clean to noisy data, thus the only important aspect is that the system performs exactly the same retrieval task (i.e. uses the same query set) when dealing with both clean and noisy data.
The database statistics are collected in Table 2: the noisy documents are, on average, 23.1% shorter than their corresponding clean versions. This happens because handwriting recognizers tend to split long terms into short words that often belong to the stoplist and are thus removed. On the other hand, the dictionary of the noisy documents is 34% bigger than the clean dataset dictionary. This is

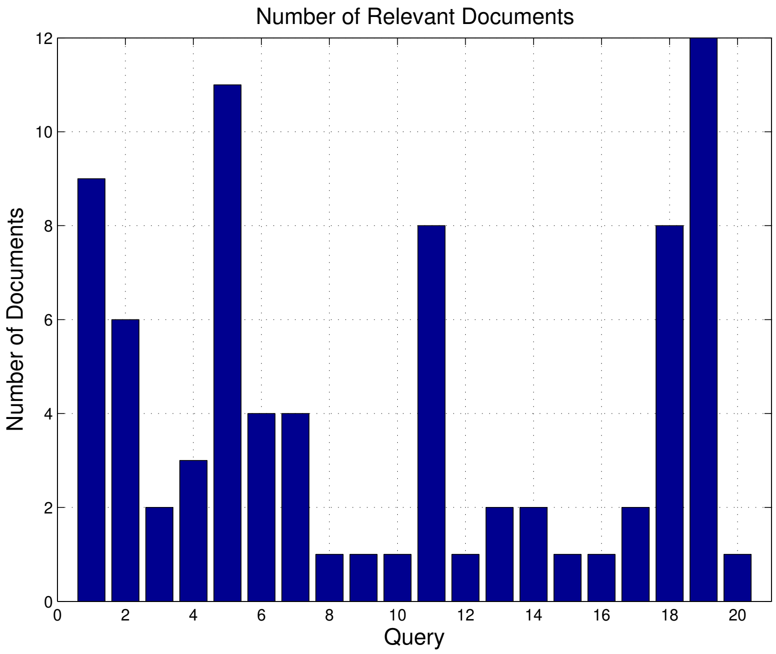


Figure 3: Number of relevant documents per query. Most of the queries have less than 4 relevant documents (accounting for 2% of the document dataset) and can thus be considered difficult.

another effect of the recognition errors: different instances of the same word are recognized differently leading to more variety in the lexicon. The query coverage is the number of query terms covered by the database dictionary: some query words are lost passing from clean to noisy data, but the difference is not significant.

The results have been evaluated with different metrics (see Section 2): Precision vs Recall curves are shown in Figure 4. The performance loss is especially high for $\rho > 50\%$ and this happens for two main reasons: the first is that the relevant documents that are significantly degraded by noise tend to fall towards lower positions of the ranking thus they make the Precision decrease. The second is that, because of the interpolation technique used to obtain the plots (Baeza-Yates and Ribeiro-Neto, 1999; Van Rijsbergen, 1979), the last part of the curve ($\rho > 50\%$) is heavily affected by the queries having few relevant documents: a change of two or three positions in the ranking can significantly lower the Precision. This can be seen by considering that the average number of relevant documents per query is four, thus to have $\pi = 75\%$ at $\rho = 100\%$ (see end of the clean data curves in Figure 4) means that all relevant documents can be found, on average, in the top five positions of the ranking. In order to lower π to 50% at the same ρ value (see Figure 4) it is sufficient that a relevant document falls to the eighth position.

The use of other performance measures (not involving interpolation) shows a more realistic difference between clean and noisy data performance. Table 3 reports avgP and BEP (see section 2) measured in the different experiments. A 10% difference in avgP corresponds, on average, to one document lost (at a given Recall value) when passing from clean to noisy documents. The BEP corresponds to the Precision at position $|R(q)|$, thus 10% difference means that the same fraction of relevant documents has been lost in the first $|R(q)|$ positions on average. Since the average number of relevant documents per query is 4, this means that the difference accounts for no more than one document.

A further estimation of the noise effect can be obtained using the P_n curves shown in Figure 5. This measure is especially important in applications where the retrieval systems return the list of the documents ordered by score. When P_n is high, it means that the user can find many relevant

ID	query
1	company merger and acquisition operations
2	trade problems for japan protectionism
3	swiss franc and swiss national bank related issues
4	agriculture and grain production in argentina
5	gulf tensions between united states and iran
6	corn businness between soviet union and united states
7	monetary policy of the federal reserve
8	allegis plans for partnerships in canada
9	privatisation performed by french government
10	surpluses in the world farm market
11	crude barrel prices
12	oil production in venezuela
13	companies dealing with the security exchange commission
14	yugoslavia economic performance
15	decisions made at the economic summit in venice
16	manhattan federal court investigatons
17	seton board of directors meeting
18	role of interest rates on financial market
19	distruction of iranian oil platform by us air force
20	natural gas and oil company purchases

Table 1: Queries submitted to the system

documents by browsing only n texts. The plot shows the highest P_n that can be achieved at each position (upper bound) and the results for clean and noisy texts using both cosine and Okapi measures. At low n, the cosine appears to be superior, but after the fifth position, the difference is no longer significant. The same can be said, given a matching measure, for the difference between clean and noisy texts. Since many interactive systems return their results in pages containing 10 retrieved items, this means that the performance loss due to recognition errors is acceptable: the number of documents to be browsed in order to find the relevant items is not significantly changed when using noisy data. The above results show that the IR system is robust with respect to a TER of around 40.7%. This means that the ranking of the documents is not significantly affected by the presence of the recognition errors. The RSV used to rank the documents is based essentially on the number of query words contained in the documents. Both cosine and Okapi measures calculate the RSV through a sum where each query term appearing in the document gives a non-zero contribution (see section 2). It is thus possible to say that the documents at the highest ranking positions are those sharing more terms with the query. The only way to loose such documents is to misrecognize all of the query words they contain, but in the following we show that the probability of such an event is low even in presence of high TER (as in our case). In our opinion, this is the main reason of the fact that the ranking is not heavily affected by the recognition errors and that the retrieval performance degradation is moderated.

dataset	avg. length	dictionary	query coverage
clean	105.2	2808	95.7%
noisy	80.8	3767	93.5%
query	4.6		

Table 2: Data statistics. The measures are obtained after stopping and stemming.

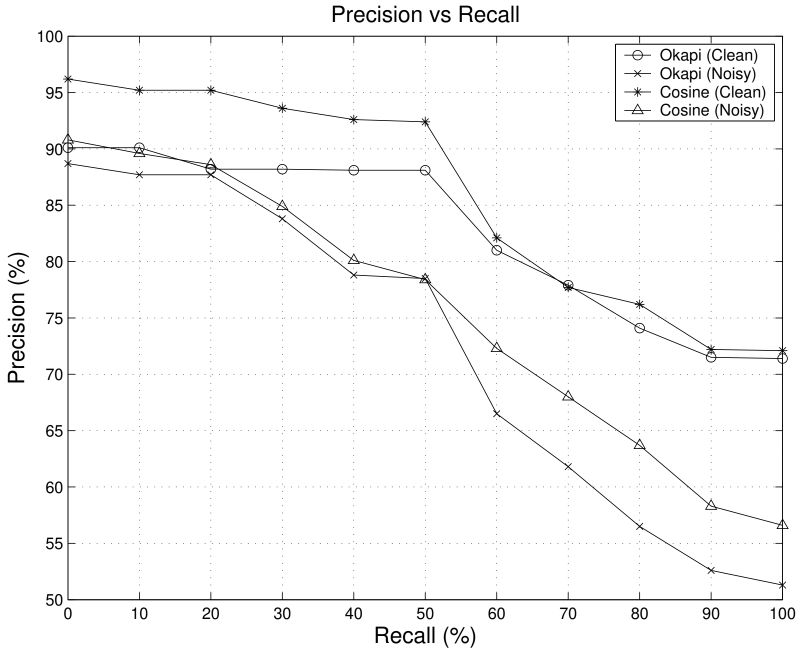


Figure 4: Precision vs Recall curves for different retrieval methods.

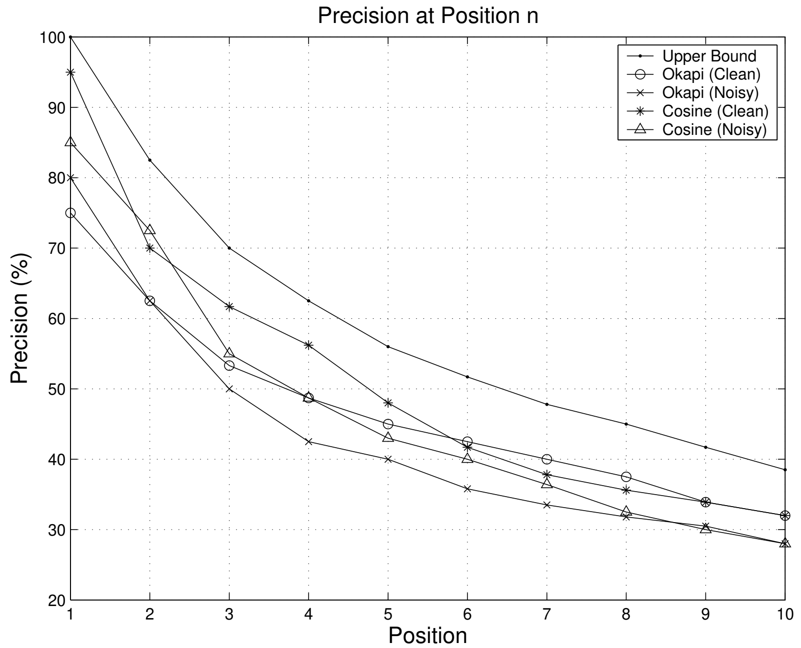


Figure 5: This plot shows the precision as a function of the ranking position. The results are obtained by averaging over all queries.

experiment	avgP (%)	BEP (%)
Okapi (clean)	81.2	74.2
Okapi (noisy)	71.2	66.2
cosine (clean)	85.4	77.1
cosine (noisy)	75.2	69.3

Table 3: Average Precision and Break Even Point.

Given a document d, the number of query terms it contains is:

$$N(q,d) = \sum_{t \in Q} tf(t,d) \tag{10}$$

where Q is the set of the query terms and $tf(t,d)$ is the term frequency. The TER can be considered as the probablity of a term being misrecognized. If T is the TER value, the probability of misrecognizing all of the query terms in d is thus $T^{N(q,d)}$. As $N(q,d)$ increases, the value of $T^{N(q,d)}$ becomes quickly low: at our TER level (40.7%), the probability of misrecognizing two or three query terms is 16.5% and 6.7% respectively. As mentioned above, documents at the top ranking positions have high $N(q,d)$. In this way, even if there are many errors, the probability of misrecognizing all of the query terms they contain is low and they can still have a score higher, on average, than other documents. This means that the documents at the top ranking positions tend to remain there even in presence of high Word Error Rates and the retrieval performance is thus degraded only moderately.

5 Conclusions

This work presented experiments on the retrieval of handwritten documents. Few works were previously presented in the literature about the same topic and they were based essentially on a WS approach (see Section 1): to our knowledge, this is the first work that applies state-of-the-art IR technologies to handwritten data. Moreover, our experiments show a comparison between the performances of the same system over both manual (no errors) and automatic (WER ∼45%) transcriptions of the same documents.

The performance has been measured with several metrics accounting for different aspects of the retrieval process. The results show that the performance loss due to noise in the data is acceptable: average Precision and Break Even Point measures suggest that, at a given Recall value, only one relevant document is lost on average. The curves of the Precision as a function of the ranking position show that the number of documents to be browsed in order to find the desired items is not significantly increased because of the recognition errors. The Precision vs Recall curves show a significant loss at high Recall, but they are negatively affected by the presence of several queries having only one relevant document. More reliable curves could be obtained by collecting queries with a higher number of relevant documents, but this would make the task too simple (a query can be considered difficult when its relevant documents account for less than 2% of the corpus they belong to).

Experiments performed on speech recording transcriptions affected by WER between 10% and 40% show that, even in such a wide range, the retrieval performance changes only slightly (Garofolo et al., 1999). The results obtained in this work show that a state-of-the-art IR system is still robust with respect to a WER of around 45%, but no results are available, to our knowledge, for higher Word Error Rates. This means that the results obtained in this work on single writer data do not allow one to guarantee that a good retrieval performance can be obtained (with our system) on multiple writer data. On the other hand, since the identity of the writer is not taken into account by the IR system, our results show that, if a handwriting recognition system achieving a WER lower or equal to 45% on multiple writer data is available, it is possible to obtain on them satisfactory retrieval results.

Acknowledgments The author wishes to thank O.Bornet and D.Grangier for technical contribution. This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). This work is dedicated to Elena Savoino.

References

Aizawa, A., 2003. An information-theoretic perspective of *tf·idf* measures. Information Processing and Management 39 (1), 45–65.

Apté, C., Damerau, F., Weiss, S., 1994. Automated learning decision rules for text categorization. ACM Transactions on Information Systems 12 (3), 233–251.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison Wesley.

Fox, C., 1992. Lexical analysis and stoplists. In: Frakes, W., Baeza-Yates, R. (Eds.), Information Retrieval. Data Structures and Algorithms. Prentice Hall, pp. 102–130.

Garofolo, J., Auzanne, C., Voorhees, E., 1999. The TREC spoken document retrieval track: a success story. In: Proceedings of 8^{th} Text REtrieval Conference. pp. 107–130.

Jain, A., Namboodiri, A., 2003. Indexing and retrieval of on-line handwritten documents. In: Proceedings of International Conference on Document Analysis and Recognition. pp. 655–659.

Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., Viorel Popescu, G., 2000. A line oriented approach to word spotting in handwritten documents. Pattern Analysis and Applications 3, 153–168.

Kwok, T., Perrone, M., Russell, G., 2000. Ink retrieval from handwritten documents. In: Proceedings of Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference. pp. 461–466.

Lewis, D., 1992. An evaluation of phrasal and clustered representations on a text categorization task. In: Proceedings of 15^{th} ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 37–50.

Porter, M., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Rath, T., Manmatha, R., 2003. Features for word spotting in historical manuscripts. In: Proceedings of International Conference on Document Analysis and Recognition. pp. 218–222.

Robertson, S., Walker, S., Beaulieu, M., 2000. Experimentation as a way of life: Okapi at TREC. Information Processing and Management 36 (1), 95–108.

Rocchio, J., 1971. Relevance feedback in information retrieval. In: Salton, G. (Ed.), The SMART retrieval system: experiments in automatic document processing. pp. 313–323.

Rosenfeld, R., 2000. Two decades of statistical language modeling: where do we go from here? Proceedings of IEEE 88 (8), 1270–1278.

Russell, G., Perrone, M., Chee, Y., 2002. Handwritten document retrieval. In: Proceedings of International Workshop on Frontiers in Handwriting Recognition. pp. 233–238.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523.

Salton, G., Wong, A., Yang, C., 1975. A vector space model for automating indexing. Communications of ACM 18 (11).

Tomai, C., Zhang, B., Govindaraju, V., 2002. Transcript mapping for historic handwritten document images. In: Proceedings of International Workshop on Frontiers in Handwriting Recognition. pp. 413–418.

Uchiashi, S., Wilcox, L., 1999. Automatic index creation for handwritten notes. In: Proceedings of International Conference on Acoustic, Speech and Signal Processing. pp. 3453–3456.

Van Rijsbergen, C., 1979. Information Retrieval. Buttersworth.

Vinciarelli, A., 2004. Effect of recognition errors on information retrieval performance. In: Proceedings of International Workshop on Frontiers in Handwriting Recognition. pp. 275–279.

Vinciarelli, A., Bengio, S., Bunke, H., 2004. Offline recognition of large vocabulary cursive handwritten text. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6), 709–720.

Vinciarelli, A., Lüttin, J., 2000. Offline cursive script recognition based on continuous density HMM. In: Proceedings of International Workshop on Frontiers in Handwriting Recognition. pp. 493–498.

Vinciarelli, A., Lüttin, J., 2001. A new normalization technique for cursive handwritten words. Pattern Recognition Letters 22 (9), 1043–1050.