Content-based text line comparison for historical document retrieval

S. Zinger, J. Nerbonne
Center for Language and Cognition Groningen, University of Groningen
Oude Kijk in 't Jatstr. 26
9700 AS Groningen
The Netherlands
s.zinger@rug.nl, j.nerbonne@rug.nl

L. Schomaker
Artificial Intelligence Department, University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
The Netherlands
schomaker@ai.rug.nl

H. van Schie
Nationaal Archief
Prins Willem Alexanderhof 20
2595 BE Den Haag
The Netherlands
henny.van.schie@nationaalarchief.nl

Abstract

In the historical handwritten document retrieval system that we are currently building, the training data set elements are the images of handwritten lines with the manually made text transcriptions. We apply sequence comparison algorithms to these text transcriptions. We explore several sequence comparison algorithms that have been applied to phonology for their usefulness in solving a problem of retrieving handwritten material. Finding an appropriate method for comparing text lines will allow us to cluster the corresponding images of handwritten lines into training sets. These training sets can then be used for pattern recognition - an important part of the historical handwritten document retrieval system. At first we study the information needs of the users of an archive where the historical documents are stored. Then we explore the longest common substring (LCS), Levenshtein and Jaccard measures for matching the text lines. Taking into account the drawbacks of these methods, we propose to weight the words in the text proportionally to their information content. This weighting is expected to provide results closer to the information needs of users. We evaluate the results in terms of the precision values for k top retrieved text lines. Using the mean precision curves we show that the performance of sequence comparisons increases up to 18% when we use the weighted sequence comparisons.

Keywords

Sequence comparison, inverse frequency weighting, transcriptions of handwritten text.

1 Introduction

Our project aims at information retrieval from handwritten documents. The project includes research in pattern recognition as well as in computational linguistics. Composing these two domains we expect to gain insight into the visual properties and linguistic structure of historical handwritten documents. We work with the archive of the Kabinet van de Koningin (Queen's Office) collection of the Nationaal Archief in the Hague (the Netherlands). To determine what kind of information has to be retrieved from historical documents, we calculate statistics on the queries that the Nationaal Archief receives. We aim at retrieving historical handwritten documents given a textual query. This will considerably facilitate access to information, making it quicker and easier. Creating a search engine for handwritten documents is a challenging problem [15]. We aim at creating such a system using a set of transcribed training data so that it will enable us to retrieve untranscribed documents from test sets [9]. Previous research on historical document retrieval systems led to difficulties in merging handwritten text with its annotations [5]. Word spotting may also be used for indexing handwritten documents [16], but it is difficult to distinguish words automatically in a handwritten text. To avoid this, we work with entire lines of handwritten text: we divide an image of a handwritten document into lines using its visual features. These lines then can be annotated and used for training a handwritten document retrieval system. We expect to benefit from matching text annotations of handwritten lines in two ways: suitable matching algorithms may be used for clustering text lines and then the images of handwritten lines to

obtain training sets for pattern recognition through machine learning; since it is easier to process words in the domain of digital text, we can explore the relevance of the results of matching algorithms on text, select the best approach and then apply it to handwritten lines. Sequence comparison applied to text lines provides a reference for matching handwritten lines and can also be applied for users' query processing in an information retrieval system. Although information retrieval is a broad field which has generated a large body of knowledge, our application has a number of particular aspects:

- text is composed of transcriptions of handwritten lines;
- relatively small amount of text;
- highly redundant context, many administrative terms;
- lack of redundancy in the dates, numerals and proper names;
- special abbreviations and terms that cannot be found in other collections.

We envisage that our engine for retrieval of images of handwritten documents will function in the following way:

1) transcriptions of lines are clustered (for example, following the algorithm presented in [2]) and the clusters are named according to the concepts present in the transcriptions in cores of the formed clusters;
2) images of handwritten lines are assigned to clusters according to their corresponding transcriptions clustered during the previous step;
3) using matching based on visual features, new untranscribed lines are assigned automatically to the formed clusters allowing to search through the untranscribed images of handwritten pages [13];
4) user gets a possibility to provide feedback on the image search results as well as to transcribe lines; this will be a source of continuous learning for the system [17].

In this article we address the problem of finding a suitable similarity measure for obtaining clusters of similar text lines. This will enable us to perform the first step in the list above.

We do not aim at an OCR-like transcription of the handwritten text because we assume that recognizing only a part of the handwritten material will provide good information retrieval results. In the following section we determine information needs of a user. It helps us to define the classes of words to be recognized in the handwritten documents.

2 Information needs

The Nationaal Archief receives queries from people and organisations that search for information. Looking for information in an archive is time consuming and often requires considerable efforts from archivists. We explore the information retrieval tasks that the archive receives in order to know what kind of information people would like to extract. There are several thousands of emails, mostly

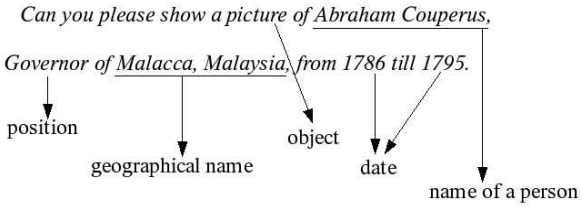


Fig. 1: *Example of a query to the Nationaal Archief.*

in English and Dutch, received by the Nationaal Archief since the year 2001 till the present time. With the help of a domain expert - an archivist, we classify these queries and calculate statistics for them. Table 1 shows our results for the queries written in English to the Nationaal Archief since 2001.

Class of a query	Number of queries	Part (%) of the total number of queries
names of people	268	28%
dates	241	25%
geographical names	212	22%
names of objects (birth certificate, ship, map, etc.)	152	16%
titles of positions, professions (king, slave, officer, etc.)	73	8%
events (visit, ship crash, arrest, etc.)	20	2%

Table 1: *Classification of the queries in English received by the Nationaal Archief during the last 6 years (2001-2006). Total number of queries is 966.*

Figure 1 illustrates that one question can contain several classes of queries. Our classification of queries is similar to the concept categories presented in [1].

Table 2 shows our results for the queries written in Dutch to the Nationaal Archief in 2001-2002. Many questions to the archive written in English concern genealogical research and often do not contain much information about the person in question. That is why the names of people are the most common class of queries. The information needs of the users who write in Dutch are somewhat broader than those of the users writing in English, therefore we included two new classes of queries: organizations and other queries. By "other queries" we mean the queries that may not be easily processed by a document retrieval system: for example, a request for some statistics over one or more centuries of the Dutch history. We notice that queries about names of people, dates and geographical names together compose approximately 70% of all information retrieval requests.

3 Text line matching

3.1 Standard methods

Here we present the sequence comparison measures that we apply to text annotations of handwritten lines. The Jaccard coefficient is defined as the number C of words that

Class of a query	Number of queries	Part (%) of the total number of queries
dates	558	25%
geographical names	491	22%
names of people	482	21%
names of objects	274	12%
organizations	173	8%
titles of positions, professions	135	6%
events	102	4%
other queries	43	2%

Table 2: *Classification of the queries in Dutch received by the Nationaal Archief during 2001-2002. Total number of queries is 2258.*

are common to the two text lines being compared (query line and match line), divided by the sum of number Q of unique words for the query, number of common words for the match and query, and number M of unique words for the match [8]:

$$Jaccard = \frac{C}{Q + C + M}. \quad (1)$$

This measure may take values from 0 to 1. It indicates the fraction of common words in the total number of words found in both lines. When the Jaccard coefficient is equal to 1, it means that the query line and the matched line have exactly the same words, but not necessarily in the same order. When the Jaccard coefficient is 0, then there is no common word between the query and the match.

The Levenshtein distance determines the minimal cost of edit operations – insertions, deletions and substitutions – that are necessary to convert one line to another [12]:

$$d(a^i, b^j) = min \begin{cases} d(a^{i-1}, b^j) + cost(a_i, \emptyset) \\ d(a^{i-1}, b^{j-1}) + cost(a_i, b_j) \\ d(a^i, b^{j-1}) + cost(\emptyset, b_j) \end{cases} \quad (2)$$

where a^i is a segment $a_1...a_i$ of the sequence a, b^j is a segment $b_1...b_j$ of the sequence b, $d(a^i, b^j)$ - Levenshtein distance between these two segments, $cost(a_i, \emptyset)$ - cost of deleting a_i, $cost(a_i, b_j)$ - cost of substitution of a^i by b_j, $cost(\emptyset, b_j)$ - cost of inserting b_j. These edit operations are weighted equally and are set to one. In this case the Levenshtein distance is the number of edit operations necessary to convert one line to another. We also use the LCS, which provides the length of the longest substring present in both the query and in the match. In our case, LCS is the number l of words in the longest common substring. Unlike the Jaccard coefficient, both the LCS and the Levenshtein distance depend on the word order in the lines that are being compared. We apply the word-based Jaccard coefficient, Levenshtein distance and LCS for measuring differences between lines of transcriptions.

We use the sequence comparison algorithms that have also been applied for phonological comparison ([3], [4]). Since some archives of dialect phonology are still handwritten, the approach described in this article may be applicable to them.

3.2 Content-based weighting

As we noticed previously, most of the queries concern dates and proper names. Such information is sparse compared with other words. Obviously, if a query line contains a date or a proper name, it is desirable to find matching lines that are as close as possible and at the same time include this date or the proper name. Nevertheless the sequence comparison methods described above do not make a distinction between the words in the text lines. The standard sequence comparisons treat a preposition and a proper name in the same way. Therefore we introduce a weighting scheme that allows us to underline the importance of words that belong to the classes in Table 2.

As we conclude from the analysis of users' queries, the information we need is mostly represented by low frequency words - proper names and dates. So the words with low frequency have to receive larger weights than the frequent ones. This can be addressed as an information encoding problem [14]. And we set the weights of words to be proportional to their information content:

$$w(i) = \alpha \times \log_2(\frac{1}{p(i)}), \quad (3)$$

where α is a parameter of the weighting, $p(i)$ is the probability of occurrence of the word i, and $i \in [1 : N]$, N is the number of words in our collection of text lines. This weighting is similar to the IDF weighting [7]. Introducing weights in the sequence comparison algorithms means calculating a sum of words' weights instead of simply counting the common or unique words for the Jaccard coefficient:

$$Jaccard = \frac{\sum_{i=1}^{C} \omega_i}{\sum_{j=1}^{Q} \omega_j + \sum_{i=1}^{C} \omega_i + \sum_{k=1}^{M} \omega_k}, \quad (4)$$

where ω_i is a weight assigned to the word i; in our case weights are positive integer numbers. Introducing weights for the longest common substring leads to the following formula:

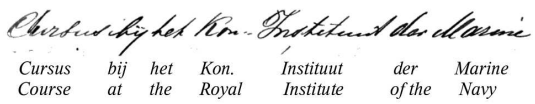
$$LCS = \sum_{i=1}^{l} \omega_i. \quad (5)$$

In the case of the Levenshtein distance we set the following costs of edit operations: $cost(a_i, \emptyset) = \omega(a_i)$; $cost(a_i, b_j) = max\{\omega(a_i), \omega(b_j)\}$; $cost(\emptyset, b_j) = \omega(b_j)$. When all weights ω are equal to 1, then we get the standard formulas for the LCS, Levenshtein distance and the Jaccard coefficient. Weights of words may also be obtained using an algorithm for learning string edit distance costs [10], but in this case a training set of pairs of strings is required.

4 Experimental results

4.1 Data description

We describe experiments on a data set from the Kabinet van de Koningin documents consisting of 5445 handwritten lines automatically segmented from scanned pages and manually transcribed. An example of an automatically segmented line and its manual text transcription is in Figure 2. The text composed of the transcriptions frequently contains dates, professions and positions and geographical names



Cursus	bij	het	Kon.	Instituut	der	Marine
Course	at	the	Royal	Institute	of the	Navy

Fig. 2: *Example of a handwritten line and its text annotation.*

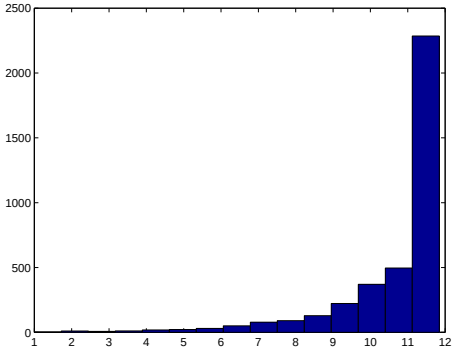


Fig. 3: *Histogram of the weights calculated according to the information content of the words in the set of transcriptions of handwritten text,* $\alpha = 1$.

while most of the last names occur only one or two times. It means that the approach described below will not lead to clusters of lines containing last names.

Before evaluating the sequence comparison algorithms, we perform the following text preprocessing: convert all characters into lower case, delete punctuation marks, delete numbers. Even though dates represent very important information, for the experiments described below we delete all numbers during the preprocessing step. The reason for it is that most of the numbers in our text do not represent dates. Instead, numbers are used as page numbers and as references to other documents in the Queen's office collection. For example, in a data sample consisting of 500 lines, only 45% of numbers indicate dates.

We prepare the test set for evaluating the precision on top k lines retrieved according to a similarity measure. Given our data set, we randomly select 25 query-lines under the condition that each of them contains at least one word belonging to a category from the Table 2: names of people, geographical names, names of object, titles of positions, professions, organizations, events. In many cases such a word is a geographical name. Then we select all good matches for every of these 25 lines. We consider a line to be a good match if one or more of its words belong to the categories mentioned above and if these words coincide with those in the query-line.

4.2 Performance comparison

We apply Jaccard, Levenshtein and LCS measures to our data before and after the information content based weighting. We can see the histogram of weights in Figure 3. For the experiments presented below we choose the coefficient α to be 0.5. We compare every line to every other and sort the results on the order of decreasing similarity to the given query-line. Reading the lists of closest matches we conclude that we can get more meaningful results if we use the weighting. Let us consider two examples of matching lines by LCS before and after the weighting. In the first example we consider the following query-line:

te Amsterdam Mr P. Scholten. _-_ Besluit fiat;
in Amsterdam Mr P. Scholten. _-_ Decision made.

In these examples, we mark the end of each line by a semicolon unless there is already a punctuation mark at the end of the line. At the end of the last line in every example there is a period. The twenty closest matches according the LCS on words are all the same: *_-_ Besluit fiat.* (In English - 'so be it decided'). These matches are not informative. The good matches have to contain either the geographical name *Amsterdam* or the name of a person *P. Scholten* or both of these words. We can see the word *Amsterdam* appearing in the two closest matches when we apply the weighting:

2e te Amsterdam, Mr J. Wiarda;
2nd in Amsterdam, Mr J. Wiarda;

hof te Amsterdam Mr J.C. Baron Band;
in Amsterdam Mr J.C. Baron.

The closest lines after the weighting contain the related words, but similar results may also be achieved by introducing stop words *besluit* and *fiat*. According to our experimental results, the information content based weighting leads to more semantically relevant results in other, more complex cases. The second example of LCS on words illustrates this. The query-line is:

nant -Kolonel, Provinciaal Adjudant in;
nant -Colonel, Provincial Adjutant in.

The main key word in this line is *Adjudant* (Adjutant). We can see that taking into account the weights of words gives matches that contain the key word more often. The results without the weighting:

tot Rechter in de arr. Rechtbank
to Judge in the district Court

het leger in Ned-Indie W. Boetje.
the army in Dutch Indies W. Boetje.

Adjudant ... en Grootmeester van H;
Adjutant ... and Grandmaster of;

And the results after applying the weighting are:

Adjudant ... en Grootmeester van H;
Adjutant ... and Grandmaster of;

Adjudant den Majoor Jhr A.S. van Feh;
Adjutant the Major Jhr A.S. van;

van der Groot en Staf, Adjudant van Z.K.H.
van der Groot and Staf, Adjutant of Z.K.H.

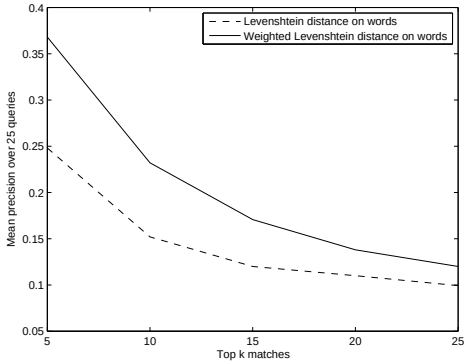


Fig. 4: *Mean precision on the top k retrieved lines for the Levenshtein distance: dashed line - without weighting, solid line - with weighting.*

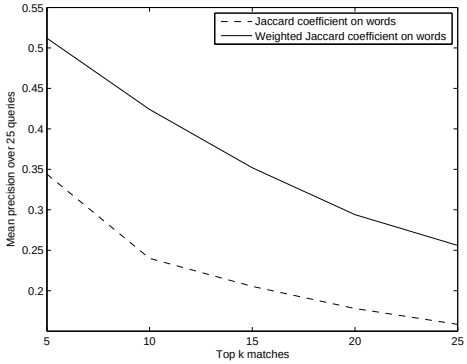


Fig. 6: *Mean precision on the top k retrieved lines for the Jaccard coefficient: dashed line - without weighting, solid line - with weighting.*

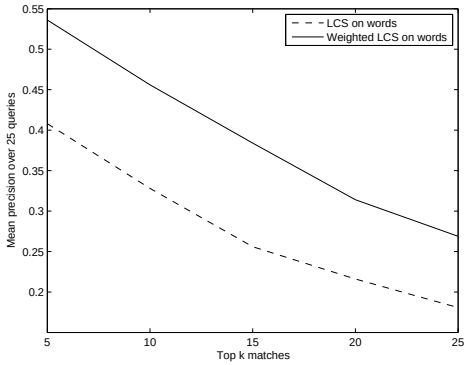


Fig. 5: *Mean precision on the top k retrieved lines for the LCS: dashed line - without weighting, solid line - with weighting.*

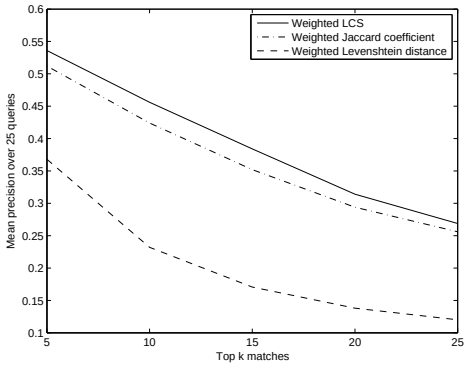


Fig. 7: *Comparison of weighted similarity measures.*

This example underlines the ability of our method to assign high weights to rare words, which often are the key words, and therefore to enable them to appear in the matching results.

In order to evaluate the ability of our approach to retrieve relevant information, we calculate its precision on top k retrieved lines. Figures 4 - 6 show the performance of our weighting method in terms of precision - the percentage of relevant text lines among the retrieved ones. These Figures show that weighting the words makes the precision higher and therefore improves our results. Since in our application we aim at grouping the text lines according to their key words, the precision is more important than recall: it is important that we extract the good matches, and not extracting some of them is less important.

5 Conclusion and future work

In order to determine the information needs for a historical document retrieval system, we study the queries to an archive. We determine the classes of queries and analyse the real-world queries of the Nationaal Archief for getting statistics. We analyse more than 3000 queries to the Nationaal Archief from people and organizations in the Netherlands and abroad. This allows us to define the content interesting to the user and adapt text line comparisons to it.

Several sequence comparison techniques are applied for matching text annotations for handwritten lines. We incorporate the word information content in the line comparison procedure and get more semantically relevant results. These results are evaluated by calculating the precision on top k retrieved lines. The best results according to the precision evaluation are provided by applying the weighted LCS as shown on Figure 7.

Some improvement for text line matching may be introduced by lemmatization. The study for finding the optimal value of the coefficient α will provide the limits for our approach and enable us to get its best performance. Our results can be applied to clustering images of automatically segmented handwritten lines. Such clusters can serve as training sets for machine learning aiming at creating handwritten document retrieval systems.

Our method is between basic keyword search and information retrieval. Lines of text contain more information than a few key words, but are much less information than is the case in full-document information retrieval. This method is most similar to passage retrieval [6], [11] in terms of the amount of text, but our lines are even smaller than a typical passage. The matching results can be used directly to train a passage retrieval system for handwritten line strips. Our information content based approach may also be applied for visual feature based matching of hand-

written lines.

References

[1] P. Constantopoulos, M. Doerr, M. Theodoridou, and M. Tzobanakis. Historical documents as monuments and as sources. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology Conference, CAA2002*, Heraklion, Greece, 2002.

[2] L. Ertoz, M. Steibach, and V. Kumar. Finding topics in collections of documents: a shared nearest neighbor approach. In *Proceedings Of Text Mine'01, Workshop on Text Mining, First SIAM International Conference on Data mining*, pages 83–104, Chicago, IL, USA, 2001.

[3] B. Kessler. Computational dialectology in Irish Gaelic. In *Proceedings of the European Association for Computational Linguistics*, pages 60–67, Dublin, Ireland, 1995.

[4] G. Kondrak. *Algorithms for Language Reconstruction*. PhD thesis, University of Toronto, 2002.

[5] E. Kornfield, R. Manmatha, and J. Allan. Text alignment with handwritten documents. In *Proceedings of Document Image Analysis for Libraries (DIAL)*, pages 195–209, Palo Alto, CA, USA, 2004.

[6] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382. ACM Press, 2002.

[7] C. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2007.

[8] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[9] T. Rath, R. Manmatha, and V. Lavrenko. A search engine for historical manuscript images. In *Proceedings of the ACM SIGIR 2004 Conference*, pages 369–376, Sheffield, UK, 2004.

[10] E. Ristad and P. Yianilos. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.

[11] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, Pittsburg, USA, 1993.

[12] D. Sankoff and J. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Publications, 1999.

[13] L. Schomaker. Retrieval of handwritten lines in historical documents. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, Curitiba, Brazil, 2007. accepted.

[14] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[15] S. Srihari, C. Huang, and H. Srinivasan. A search engine for handwritten documents. In *Document Recognition and Retrieval XII, SPIE*, pages 66–75, San Jose, CA, USA, 2005.

[16] S. Uchihashi and L. Wilcox. Automatic index creation for handwritten notes. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 3453–3456, Phoenix, AZ, USA, 1999.

[17] T. van der Zant, L. Schomaker, M. Wiering, and A. Brink. Cognitive developmental pattern recognition: Learning to learn. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 1208–1213, Taipei, Taiwan, 2006.