Consolidating the results of automatic search in large scale digital collections.

Noga Levy[1], Adiel Ben-Shalom[2], Itai Ben-Shalom[3], Lior Wolf[1], Nachum Dershowitz[1], Roni Shweka[2], Yaacov Choueka[2]

[1] The Blavatnik School of Computer Science, Tel-Aviv University

[2] The Friedberg Genizah Project

[3] The School of Electrical Engineering, Tel-Aviv University

Abstract

We present a tool for grouping the results obtained from image-based document search engines. The obtained groups provide structure to the search engine's output and allow immediate accessibility to the most prominent results. The tool is based on employing graphical Bayesian models and it reinforces retrieved items that are strongly linked to other retrievals. The utility of the tool is demonstrated within the context of visual search of documents from the Cairo Genizah.

Introduction

Searching digital collections as part of ongoing research is inherently different from everyday use of Internet search engines. A researcher is often interested in gathering all results relevant to her work and is not satisfied with just the most relevant one that best matches the intent of the query. Our focus is on large scale digital collections, where a query can retrieve thousands of results. Many of these results might be irrelevant, but many might require a careful consideration. In order to provide practical tools for researchers using such a system, we develop tools that combine the various retrieved documents into coherent groups. This serves two major purposes: (1)it reduces the exploration time needed to examine the query results, and (2) the group elements reinforce each other, pointing the researcher's attention to results that are more likely to match the query in a meaningful way.

The collection that we consider in this work is the digital collection of the Cairo Genizah manuscripts, which is collected and maintained by the Friedberg Genizah Project [Glickman10]. The Cairo Genizah is a large collection of discarded codices, scrolls, and documents, written in the 10th to 15th centuries, and which is now distributed in over fifty libraries and collections around the world. The texts are written mainly in Hebrew, Aramaic, and Judeo-Arabic (in Hebrew characters), but also in many other languages. Genizah documents have had an enormous impact on 20th century scholarship in a multitude of fields, including Bible, rabbinics, liturgy, history, and philology. Most of the material recovered from the Cairo Genizah has been digitized and catalogued. Unfortunately, most of the leaves that were found were not found bound together. Pages and fragments from the same work may have found their way to disparate collections around the world and some fragments are very difficult to read. Scholars have therefore expended a great deal of time and effort on manually rejoining leaves of the same original book or pamphlet.

Previously, we developed a visual similarity measure that is used to find pages that are likely to have originated from the same original manuscript, before the vicissitudes of the Genizah separated them. Such groups of pages are called *joins* and are of great importance in the study of the Cairo Genizah.

This visual similarity is used for searching joins in the following manner. A researcher points to a fragment or a shelfmark of interest in a digital Genizah manuscript (www.genizah.org) and the system returns the shelfmarks of Genizah fragments that are the most similar to the query. Currently, the results are presented fragment after fragment, and the researcher can explore the list for as long as she wishes. It is our goal to help her to explore more efficiently, focusing on the more relevant results. This is

based on the assumption that if several fragments are similar to the query fragment and are similar to each other, then this group as a whole is more likely to be of interest than a random set of visually-different retrieval results.

Results

To illustrate our method of finding fragments that are most likely to be joins of a fragment of interest, we chose the shelfmark 'Paris, AIU: II.B.79' from the Alliance Israélite Universelle library in Paris. According to the catalogue, this shelfmark contains a fragment from the exegesis by Sa'adiah ben Yosef Gaon to the book of Leviticus, chapter 19. Sa'adiah Gaon is considered the founder of Judeo-Arabic literature and is known for his works on Hebrew linguistics, Halakha, and Jewish philosophy. The chosen shelfmark has eight known joins in the Genazim database, all from the Jewish Theological Seminar Library.

As a first step, the similarity between the chosen shelfmark and approximately 100,000 shelfmarks in the Genizah was computed using four similarity measures that are then combined together as described in the Methods section.

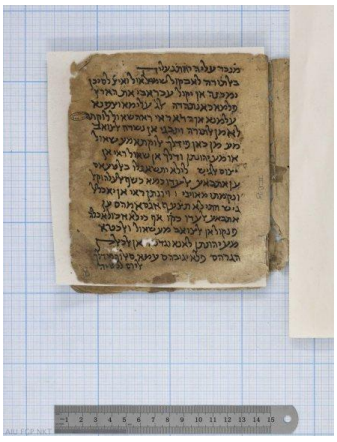
To narrow down the candidate shelfmarks we chose only candidates that have a positive similarity score with the shelfmark of interest. Additionally, duplicated shelfmarks were recognized and only the most similar fragment in each selfmark remained.

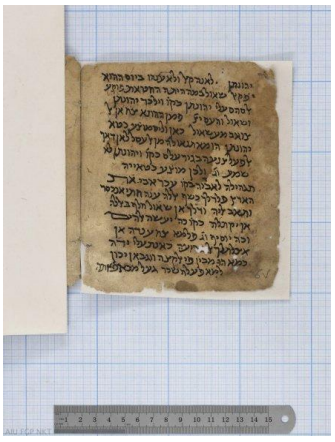
After narrowing down, the number of shelfmarks is tremendously decreased to 390. However, going over 390 fragments is a complicated, time-consuming task for a human researcher, and this number might be much higher for other queries. Our goal is to gain further insight on the relations among the candidate shelfmarks and hand the researcher only the most promising ones.

The shelfmarks of interest were used to build the graphical model. This model contains 390 vertices, each vertex represents a shelfmark. The edges in the graph are weighted by the similarity between the shelfmarks they connect.

To find interesting connections in the graph, a statistical clustering algorithm was applied (Methods). We then considered the returned beliefs found (these are referred to as variables l_{ij} in Methods), and extracted the maximal cliques using the Bron-Kerbosch algorithm [Brom73].

The largest clique found contains eleven fragments. The fragments belonging to this clique are presented in Figure 1.



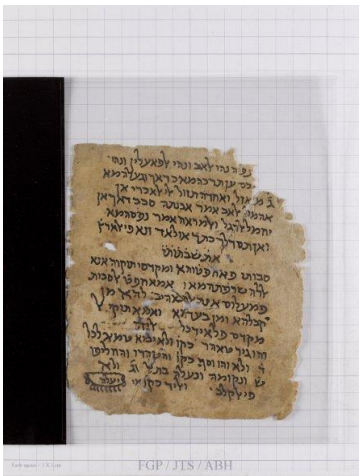


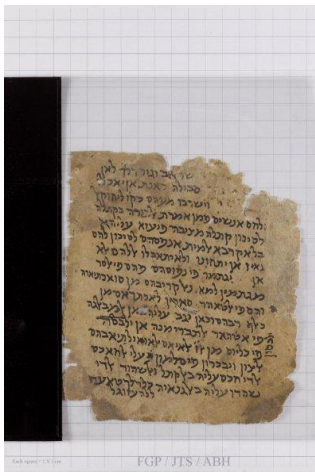
1. AIU: II.B.79

2. JTS: ENA 1696.10

3. JTS: ENA 1696.9

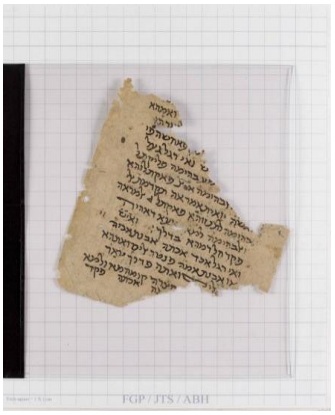
4. JTS: ENA 2815.12

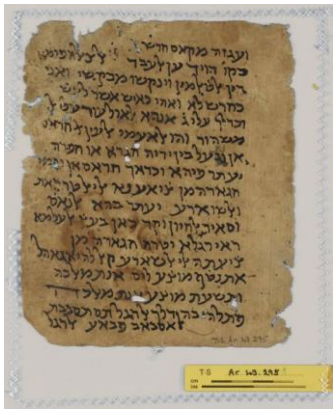


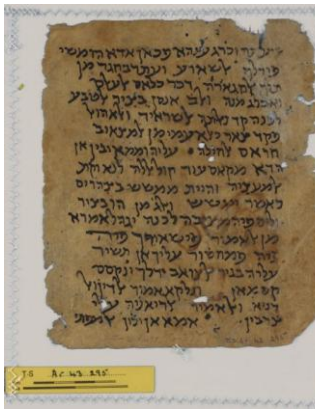


5. JTS: ENA 2931.30

6. JTS: ENA 2931.31





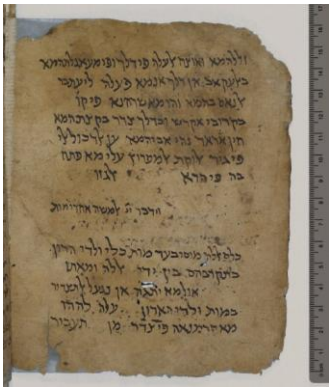


7. JTS: ENA 3202.6

8. T-S Ar.43.295

9. T-S Ar.51.198

10. T-S NS 164.11



11. OR 10821.10

Figure 1: The shelfmarks contained in the largest clique. The recto and verso of every document are presented. The first shelfmark is the investigated document. shelfmarks 2 to 7 are known joins.

The shelfmarks chosen by the graphical model as the most promising turned out to be of proven connection with the inspected shelfmark, AIU:II.B.79. Out of the ten shelfmarks, six are known joins (2 to 7 in Table 1). Shelfmarks T-S Ar.43.295 (8 in Table 1) and T-S Ar.51.198 (9 in Table 1) have catalogue identifications indicating that they also belong to the exegesis by Saadiah Gaon to Leviticus, chapter 19, and it is likely that these shelfmarks are unknown joins to AIU:II.B.79.

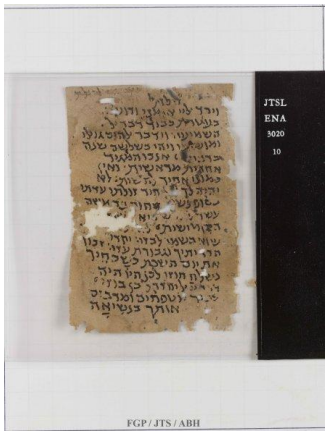
The other shelfmarks are also very relevant. shelfmark 10 in the table, T-S NS 164.11, is identified in the catalogue as a responsum referring to Leviticus 19:16. The last shelfmark in the table, OR 10821.10, is relevant as well, being part of an arabic tifsir (exegesis) to Leviticus 17 estimated to the 14th century.

The other large cliques found do not contain the shelfmark investigated, hence might not be as valuable to the researcher interested in a specific shelfmark, but some of them are of their own value. For example, the second largest clique contains ten shelfmarks all catalogued as liturgical poems, indicating that there are potential joins, previously unknown, among them.

JTS: ENA 3026.18

JTS: T-S NS 151.7

JTS: ENA 2829.4



JTS: ENA 3020.10

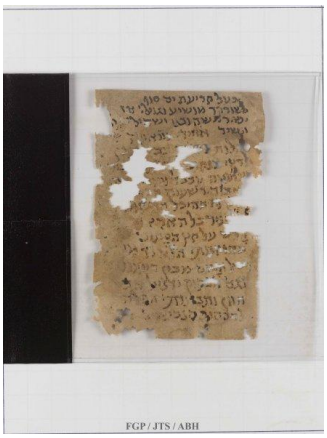
JTS: ENA 2829.3

Strasbourg: 4077/15

JTS: ENA 2829.1

JTS: ENA 2829.2

T-S NS 151.6



JTS: ENA 3020.11

Figure 2: All shelfmarks in the second largest clique. The recto and verso of every document are presented. All the shelfmarks in this clique are catalogued as liturgical poems.

Methods

In the pre-processing step, handwriting-based image properties are calculated for the documents in all shelfmarks, as described in [Wolf11a]. Each image is segmented into fragments that are binarized and aligned horizontally by rows. We then detect keypoints in an image, and calculate their descriptors. All descriptors from the same document are combined into a single vector.

Given a query, we apply a pairwise similarity measure between the vector corresponding to the queried shelfmark and the vectors of all fragments in the collection. The similarity scores are calculated as described in [Wolf11a] by employing both simple and learned similarity scores and, in addition, combine several scores together by a technique called stacking .

Ranking the shelfmarks by their similarity to the fragment of interest provides an effective way to scan the dataset for relevant shelfmarks; However, this method considers only direct correlation with the query. The shelfmarks collected also correlate with each other to a varying degree, and from these local correlations one would like to better infer their relevancy to the queried shelfmark . That is, a candidate

shelfmark relevancy is estimated not merely by its similarity to the queried shelfmark, but also by its similarity to other promising candidates.

Statistical Clustering

We apply statistical clustering to discover structures that contain groups of shelfmarks. Conventional clustering algorithms including centroid based algorithms, spectral clustering techniques, various hierarchical clustering schemes, and mixture models, often fail to produce meaningful results for large and heterogeneous datasets, where the number of clusters is unknown in advance and the size of each cluster can be as small as one (a singleton, which is a very common case) or as large as a hundred items.

Graphical models are suited to our data since we wish to find local correlations among small subsets of the shelfmarks. These local correlations can be effectively represented as edges in a graphical model. In addition, graphical models enable reasoning about hidden long-range correlations, through connectivity and influences. The connectivity is typically expressed as the maximum a-posteriori (MAP) problem with potential function ϕ on single and groups of data points (shown here for pairs):

$$argmax_{x_1,\ldots x_n} \sum_i \phi_i(x_i) + \sum_{i,j} \phi_{i,j}(x_i, x_j)$$

.

As our task is to search for the best assignment out of the exponentially many assignments, this formalization manages to discover correlations between its variables even when they are not explicitly correlated through a pairwise potential function.

In practice, message-passing algorithms provide an efficient framework for these tasks. These algorithms act locally by sending messages along the graph edges, and yet achieve globally consistent result that directly corresponds to long-range correlations.

While the use of graphical models for explicit clustering problems is relatively limited, the act locally infer globally capability made graphical models an effective tool for various clustering-like problems, such as image segmentation (clustering of similar and nearby pixels), object detection (clustering of nearby pixels that have some pre-learned object-like properties), pose estimation of human bodies from images (clustering of 3D placements of neighboring joints), and depth estimation in stereo images (clustering of nearby patches that have similar disparities to facets).

There is a computational gap between these types of specific clustering-like applications and the conventional clustering problem, in which we are interested, in that these applications produce graphs of a limited degree. In the general clustering problem every two data points are potentially connected, and enforcing consistent grouping of the data points into non-overlapping subsets requires even more involved connections.

Earlier contributions in the field of graphical model based clustering [Shental03] assume a known number of classes. A recent contribution in the field of pedestrian grouping [Pellegrini10] enforce transitivity constraints by considering all triplets of pedestrians. Similar transitivity constraints were used at a much larger scale by our group to cluster Genizah documents in a semi-supervised manner [Wolf11c].

Finding the maximum a-posteriori (MAP) assignment in a graphical model involves searching in exponentially large space. The MAP problem can be described by a linear program, where the variables of the program are zero-one probability distributions which agree on their marginal probabilities. Since this linear program has integer constraints, it has high complexity. In the last decade a considerable

effort was made to construct a scalable solver for large-scale linear programs. One of the first approaches was based on spanning trees over the graphical models and is known as the tree re-weighted belief propagation [Wainwright05]. Hazan and Shashua [Hazan10] continued this line of work, presenting the convex belief propagation algorithms for inference. This approach emerges from methods which decompose the large-scale MAP inference to many small-scale MAP inference problems, with interdependent messages sent along the edges of the graphical model.

Our Graphical Model

We employ a tailor-made graphical model clustering solution, in which binary variables l_{ij} denote linking between every two data points, and the consistent grouping constraint is modeled as multiple transitivity constraints between the linking variables $l_{ij} \wedge l_{jk} \rightarrow l_{ik}$.

The prior probabilities of the l_{ij} variables are derived from the pairwise handwriting-based image similarity of i and j, and is expressed by the pairwise models $\gamma_{ij}(l_{ij})$. For two documents that are visually similar, $\gamma_{ij}(l_{ij})$ is close to one for l_{ij} of a value of 1, and close to zero otherwise. The situation is the opposite for documents that do not look similar.

Transitivity is enforced by adding models $\chi(l_{ij},l_{ik},l_{jk})$ for every triplet of documents to capture the constraints $l_{ij} \wedge l_{ik} \rightarrow l_{jk}$, $l_{ij} \wedge l_{jk} \rightarrow l_{ik}$, and $l_{ik} \wedge l_{jk} \rightarrow l_{ij}$.

Such models are created for every lexicographically ordered pair (i,j) < (i,k) < (j,k).

The integrated log-probability of the grouping variables is calculated as:

$$logP(\{l_{ij}\}) = \sum_{ij} \gamma_{ij}(l_{ij}) + \sum_{ijk} \chi(l_{ij},l_{ik},l_{jk}) - log\ Z$$

,

with Z being the normalization factor which ensures that the probabilities are feasible.
Here we use the method of [Hazan10] since it is scalable enough to support our needs: in our model for every pair of shelfmarks there exists a linking variable, and for every triplet of linking variables there is a transitivity constraint. Therefore, for n documents we have $O(n^2)$ linking variables that we need to solve under $O(n^6)$ constraints. Convex belief propagation provides a tractable solution for such systems. This is in contrast to the Dual Decomposition method previously used in [Wolf11c] which requires the use of a simpler graph where the constraints are sub-sampled.

Postprocessing

The transitivity constraints ensure that the links are arranged in a consistent manner, i.e., in cliques. However, due to the probabilistic nature of the problem and the approximations performed during optimization, the obtained solution requires further processing in order to make it discrete and entirely coherent.

We perform this step using the Bron-Kerbosch maximal-clique algorithm [Brom73]. Bron-Kerbosch finds maximal cliques by going over all cliques in the graph and extending them using recursive calls. The algorithm keeps track of vertices that preserve complete connectivity of the clique at hand. A clique is extended by adding a vertex from these vertices, and keeping only neighbours of the newly added vertex as potential members of the clique. When there are no more vertices to add, a maximal clique is found.

Discussion

Digital Paleography holds the promise of scalability: it allows not only the processing of sizable collections of documents, but also, and even more practically at the moment, the comparison of all small subsets of such collections, thereby finding links that were previously unknown.

As the Genizah visual search engine is making its online debut as one of the first digital paleography tools that are fully accessible to the non-technical research community, an effort is made to closely match the typical work patterns that are employed by the scholars when using more traditional tools. Some researchers consider the search engine to be an “extended google” and feel comfortable scanning the list of results obtained by the it looking for the documents that are of interest to them. Other researchers expect the system to provide more structured results and are not satisfied with linear scanning of lists.

In this work we explore the use of advanced clustering tools in order to provide structure to the list of retrievals returned by the visual search. We demonstrate that such a treatment can produce meaningful results at the “front page” of the results, provide new insights, and effortlessly locate unknown joins.

By using scalable methods, the underlying computational tasks are solved in minutes and can be further sped-up by using multiple machines and by caching of previous computations. It is our plan to run our inference engine for each feasible Genizah query (based on a single document) in advance and to avoid any latency when presenting the results to the researchers.

We hope that the incorporation of such mid-level capabilities to the emerging tools that are aimed at the non-technical research community will help make them even more useful and popular, provide additional value to more researchers, and help make the impact of digital paleography more significant in the very near future.

References

[Brom73] **C. Bron and J. Kerbosch.** *Algorithm 457: finding all cliques of an undirected graph. Commun. ACM* 16, 1973.

[Glickman10] **M. Glickman**. *Sacred Treasure*, the Cairo Genizah. 2010.

[Hazan10] **T. Hazan and A. Shashua.** *Norm-Prodcut Belief Propagtion: Primal-Dual Message-Passing for LP-Relaxation and Approximate-Inference*. IEEE Trans. on Information Theory, Dec. 2010.

[Pellegrini10] **S. Pellegrini, A. Ess, and L. V. Gool**. *Improving data association by joint modeling of pedestrian trajectories and groupings.* In ECCV, 2010

[Shental03] **Noam Shental, A. Zomet, Tomer Hertz, and Yair Weiss.** *Pairwise Clustering and Graphical Models*. Neural Information Processing Systems (NIPS), 2003.

[Wainwright05] **M. J. Wainwright, T. Jaakkola and A. S. Willsky.** *A new class of upper bounds on the log partition function*. IEEE Trans. on Information Theory, vol. 51, page 2313--2335, July 2005.

[Wolf11a] **L. Wolf, R. Littman, N. Mayer, T. German, N. Dershowitz, R. Shweka, and Y. Choueka** *Identifying Join Candidates in the Cairo Genizah*. International Journal of Computer Vision (IJCV) , 2011.

[Wolf11b] **L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka**. *Automatic paleographic exploration of Genizah manuscripts.* Codicology and Palaeography in the Digital Age II, 2011.

[Wolf11c] **L. Wolf, L. Litwak, N. Dershowitz, R. Shweka, and Y. Choueka** *Active Clustering of Document Fragments using Information Derived from Both Images and Catalogs. IEEE International Conference on Computer Vision (ICCV)*, 2011.