Active Congruency-Based Reranking☆

Itai Ben Shalom[a], Noga Levy[a,*], Lior Wolf[a,**], Nachum Dershowitz[a], Adiel Ben Shalom[b], Roni Shweka[b], Yaacov Choueka[b], Tamir Hazan[c], Yaniv Bar[a]

[a] *The Blavatnik School of Computer Science, Tel Aviv University, Israel*
[b] *Genazim Digital, The Friedberg Genizah Project*
[c] *The Department of Computer Science, The University of Haifa, Israel*

Abstract

We present a tool for re-ranking the results of a specific query by considering the matrix of pairwise similarities among the elements of the set of retrieved results and the query itself. The re-ranking thus makes use of the similarities between the various results and does not employ additional sources of information. The tool is based on graphical Bayesian models, which reinforce retrieved items strongly linked to other retrievals, and on repeated clustering to measure the stability of the obtained associations. The utility of the tool is demonstrated within the context of visual search of documents from the Cairo Genizah and for retrieval of paintings by the same artist and in the same style.

Keywords: query retrieval, ranking, computer vision, graphical models, active learning, clustering

1. Introduction

Searching digital collections as part of ongoing research is inherently different from everyday use of Internet search engines. A scholar 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

☆A preliminary and partial version of this work has been published in [1]

*This research was carried out in partial fulfillment of the requirements for the Ph.D. degree
**e-mail: wolf@cs.tau.ac.il, phone: 972-3-6406700, postal address: Tel Aviv University, P.O.B. 39040, Ramat Aviv, Tel Aviv 69978, Israel

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 methods that consider the various retrieved documents and identify coherent groups that include the query. Thus the exploration time needed to examine query results is reduced by having group elements reinforce each other, ultimately pointing the researcher's attention to results that are more likely to match the query in a meaningful way.

One collection that we consider in this work is the digital collection of the Cairo Genizah manuscripts, comprising 157,514 fragments collected and maintained by the Friedberg Genizah Project [2]. The Cairo Genizah is a large collection of discarded codices, scrolls, and documents, written predominantly 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). 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 cataloged. Unfortunately, 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 fragments of the same original book or pamphlet.

Previously, a visual similarity measure was developed [3], which 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 the digital Genizah database and the system returns the shelfmarks of Genizah fragments that are the most similar to the query. In a system recently put online (`www.jewishmanuscripts.org`), 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 build algorithmic tools to help her explore more efficiently, ranking the more relevant results higher. We base our work on the assumption that if several fragments are similar to the query fragment and 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.

The second collection we consider is a large dataset of digitized art images obtained from the visual art encyclopedia `www.wikipaintings.org` thoroughly covering Western and modern art. For each image, available metadata includes artist name and nationality, art movement, content, etc. We automatically evaluate the retrieved results by regarding images painted by the same artist and categorized as the same art movement and content as similar.

2. Related Work

Query-driven image retrieval contributions are mainly concerned with finding an informative visual representation. Ranking images by their similarity to the query of interest provides an effective way to scan the dataset for relevant images. However, this method considers only direct correlation with the query. The retrieved images collected also correlate with each other to varying degrees, and from these local correlations one would like to better infer their relevancy to the query at hand. That is, a candidate relevancy is estimated not merely by its similarity to the query, but also by its similarity to other promising candidates.

One possible way to achieve similarity betterment is by enhancing the similarity between reciprocal nearest neighbors as suggested in [4]. Another alternative is to employ graphical models for information retrieval, examined in the context of personalized web search in [5], where retrieval results are improved by exploiting associations among items and by regarding personal preferences. Fusion of query-specific ranking orders based on various similarity measures is suggested in [6]. Each ranking is represented as a weighted graph, and all graphs are integrated into a single graph. The final ranking is chosen either by

employing the pageRank method or by finding the maximal weight density in the integrated graph.

Graphical models are suited to our data since we wish to build upon local correlations among small subsets of the images. 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. This "act locally, infer globally" capability made graphical models an effective tool for various clustering-like problems, such as image segmentation, object detection, pose estimation of human bodies from images, and depth estimation in stereo images. Previous research on graphical-model based clustering [7] assumes that the number of classes is known in advance. Pedestrian grouping identification with graphical models [8] encourages transitivity by adding constraints for all triplets of pedestrians. Similar transitivity constraints were subsequently used at a much larger scale to cluster Genizah documents in a semi-supervised manner [9].

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 that 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 tree re-weighted belief propagation [10]. This line of work is continued by [11, 12], presenting the convex belief propagation algorithms for inference. This approach emerges from methods that decompose the large-scale MAP inference into many small-scale MAP inference problems, with interdependent messages sent along the edges of the graphical model. In our work, we employ the distributed method of [12], and contribute a heuristic method to select the parameter of the inference algorithm.

The variables inferred from the graphical model can either be used to form improved similarity scores to re-rank the images with regard to the query of

interest or as intermediate results that are further processed. We suggest the spectral clustering co-occurrence stability method described in Section 4, which employs the spectral analysis of [13] and finds a stable re-ranking of the images by repeated runs of k-means. Stability of spectral analysis often refers to the stability of the clusters derived from the spectral data (continuous clustering) and the stability of the clusters obtained by employing k-means afterwards (discrete clustering). The stability of the clustering method can be analyzed by testing the influence of small perturbations in the data [14] and is arguably an appropriate measure of clustering-method quality [15]. Our perspective is somewhat different and more local: we consider only the cluster that contains the query at hand and rank highly those documents that tend to co-occur with it. The idea of finding a good ranking based on co-occurrence frequencies also appears in [16] in the context of personalized web search. However, our method of finding co-occurrences is unsupervised, while that setting is supervised multi-label classification, with classes referring to geographical location, the content of the page, etc. The term "cluster" is used in that paper to describe a set of documents tagged by the same label.

3. Similarity Betterment

We deal with a very noisy similarity matrix, in the context of query-driven image retrieval. Ranking the images by their similarity to the query of interest provides a baseline way to scan the dataset for relevant images. However, this method considers only direct correlation with the query and ignores correlations among candidate images. We try to leverage these local correlations to better infer the results' relevancy to the query at hand. The applicability of a candidate image is estimated both by its similarity to the query and by its similarity to other leading candidates. We concentrate on the n top candidates according to the initial similarity matrix, and aim to improve their ranking with regards to the query of interest.

3.1. Model Variables

We employ a tailor-made graphical model solution, in which binary variables l_{ij} denote linking between the query of interest and one of the candidates or between pairs of candidates, and the consistent grouping constraint is modeled as multiple transitivity constraints between the linking variables. The prior probabilities of the l_{ij} variables are derived from the pairwise handwriting-based image similarity of i and j, and are expressed by the pairwise models $\gamma_{ij}(l_{ij})$. For two images that are visually similar, $\gamma_{ij}(1)$ is close to one, and close to zero otherwise, and vice versa for dissimilar images. Our model assigns the potential of the top t ranked candidates to 1 (in our experiments, t is set to 10), since empirically these candidates are often linked to true matches or are a match themselves.

3.2. Transitivity Constraints

We examine triplets of pairs, (l_{qi}, l_{qj}, l_{ij}), where q is the query of interest. Transitivity is violated in assignments that contain a single zero value, as one image is similar to both other images but they are not similar to each other.

Let q be the query index, then for each pair (i, j) of images, the transitivity potential $\chi(l_{qi}, l_{qj}, l_{ij})$ equals 0.9 if $(l_{qi}, l_{qj}, l_{ij}) = (1, 1, 1)$ and 0.1 otherwise.

After applying subsampling (see Sect. 3.3), the triplets remaining in the model are those suspected as transitivity violating groups. The chosen potential function pushes towards assigning (1, 1, 1) to their corresponding variables.

In comparison, the transitivity potential in [9, 8], can either increase or decrease the beliefs of the images suspected as violating transitivity. Their potential function assigns low values to the transitivity violation states $((0, 1, 1)$, $(1, 0, 1)$, $(1, 1, 0))$, and high values to all other states, whether they encourage transitivity (that is, $(1, 1, 1)$) or solve the violation by ignoring high similarities $((0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$, $(0, 0, 0))$. Empirically, our transitivity potential outperforms the transitivity potential previously proposed.

3.3. Subsampling Transitivity Potentials

Due to the large scale of our datasets, subsampling of the transitivity constraints is required. In [9], all triplets of images are considered, and the subsampling selects triplets with energy above a predefined threshold. The transitivity violation in their model is measured by the energy function

$$\gamma_{ij}(1)\gamma_{ik}(1)\gamma_{jk}(0) + \gamma_{ij}(1)\gamma_{ik}(0)\gamma_{jk}(1) + \gamma_{ij}(0)\gamma_{ik}(1)\gamma_{jk}(1) \ . \tag{1}$$

Unlike [9], we consider only triplets containing the query and use our own energy function, which has a trade-off parameter β, to weight the minimal potential with regard to the other potentials. Without loss of generality, for every triplet (i, j, k), let (j, k) be the pair with the minimal potential value, $\gamma_{jk}(1) = min(\gamma_{ij}(1), \gamma_{ik}(1), \gamma_{jk}(1))$, then our violation energy function is

$$E(l_{ij}, l_{ik}, l_{jk}) = \gamma_{ij}(1) + \gamma_{ik}(1) + \beta\ \gamma_{jk}(0). \tag{2}$$

We set β to 2 in all of the experiments, as transitivity is likely to exist when two images resemble a third image with high probability, and we want to find these cases even for intermediate $\gamma_{jk}(0)$ values. Subsampling is performed by calculating the energy scores and selecting the N maximal triplets. In our experiments $N = 2000$.

3.4. Optimization Problem

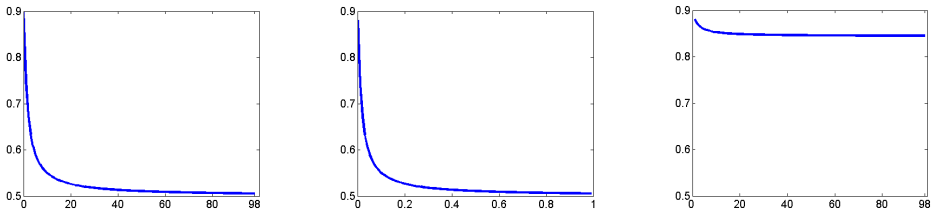
Our formalization of the model follows [12], and for consistency we denote by x_α the variables involved in each transitivity constraint (l_{ij}, l_{jk}, l_{ik}). Beliefs are denoted b_{ij} and b_α. The variational entropy $H(b)$ is approximated by

$$\widetilde{H}(b) = \sum_\alpha c_\alpha H(b_\alpha) + \sum_{ij} c_{ij} H(b_{ij}),$$

The objective function of our Convex Belief Propagation optimization problem is given by

$$\max \sum_{\alpha, x_\alpha} b_\alpha(x_\alpha) \ln \chi_\alpha(x_\alpha) + \sum_{ij, l_{ij}} b_{ij}(l_{ij}) \ln \gamma_{ij}(l_{ij}) + \varepsilon \widetilde{H}(b), \tag{3}$$

s.t. $\forall i, j, l_{ij}, \alpha \in N_{ij}, \sum_{x_\alpha \setminus l_{ij}} b_\alpha(x_\alpha) = b_{ij}(l_{ij})$, where γ_{ij} and χ_α are the potentials and N_{ij} stands for all nodes α for which $l_{ij} \in x_\alpha$. Parameter ε is set to 1 in our experiments.



(a) (b) (c)

Figure 1: The belief of the chosen pair (i, j) inferred by the graphical model (explained in the text) as a function of (a) the number of transitivity factors when $c_\alpha = 1$, (b) the value of c_α when the number of transitivity constraints is fixed to 98, and (c) the number of transitivity factors when c_α is calibrated as suggested.

3.5. Calibration of the Trade-off Parameters

The objective function presented in Eq. (3) contains the potential functions γ and χ, and the entropy approximation. The entropy approximation consists of an entropy expression for each factor l_{ij} and x_α, weighted by the trade-off parameters c_{ij} and c_α, respectively. Each l_{ij} variable has exactly one matching entropy expression, and the default assignment of $c_{ij} = 1$ for all $H(b_{ij})$ expressions is reasonable. However, the calibration of c_α is not straightforward and has to be done with care [17]. The complication stems from the difference in $|N_{ij}|$, the number of neighboring transitivity factors of the variables l_{ij}.

Let l_{qi} be a binary variable that denotes linking between the query and some candidate i, then l_{qi} may have (up to subsampling) a neighboring transitivity factor per each candidate $j \neq i$. For a binary variable l_{ij}, denoting the link between two candidates, on the other hand, the only possible factor is for $\alpha = (l_{qi}, l_{qj}, l_{ij})$.

Since each transitivity factor has a matching entropy expression, the belief of nodes l_{qi} with many neighboring factors is strongly influenced by the en-

tropy expressions $H(b_\alpha)$. As entropy is maximized when all states are equally probable, their distribution is pushed towards uniform distribution. Let $|\overline{N}_\alpha| = (|N_{ij}| + |N_{ik}| + |N_{jk}|)/3$ be the average number of neighbors of the binary variables in α. To limit the effect of the entropy expressions on the beliefs of the binary variables, c_α is set to $\eta/|\overline{N}_\alpha|$. The parameter η is manually set to 0.1 in our experiments.

We synthetically demonstrate the effect of imbalanced entropy maximization expressions on the model's beliefs. The synthetic data contains a random continuous similarity matrix $S \in [0,1]^{100\times100}$, $S_{ii} = 1$ for $i = 1, \ldots, 100$. One pair of nodes with matching variable l_{ij} is randomly chosen and its similarity is set to $\gamma_{ij}(1) = 0.9$. We examine the impact of adding a subset of the transitivity factors containing i, j and a third node k from the remaining 98 nodes, with a constant potential function $\chi(l_{ij}, l_{ik}, l_{jk}) = 0.5$ for every state of the variables and various c_α values.

Figure 1(a) shows the belief of l_{ij} as a function of the number of transitivity factors (0 to 98) when $c_\alpha = 1$. When none of the transitivity factors is added, the belief equals the prior, $b_{ij}(1) = 0.9$. As more factors are added, the belief approaches the uniform distribution value of 0.5. This behavior illustrates the problem of adding transitivity factors without balancing their matching entropy expressions.

Figure 1(b) depicts the inferred belief of l_{ij} as a function of the c_α when all 98 transitivity constraints are added. When the entropy expressions are not weighted ($c_\alpha = 1$), the belief of l_{ij} goes to uniform distribution. However, when c_α approaches the reciprocal value of the transitivity count, the belief is closer to the prior similarity, implying that this choice of the c_α value re-balances the objective function. Finally, Figure 1(c) depicts the belief of l_{ij} as a function of the number of transitivity factors, when c_α is calibrated as suggested here. As can be seen, the belief of l_{ij} remains close to the prior probability.

4. Spectral Clustering Co-occurrence Stability

To reinforce congruent groups of similar images that are similar to the query as well, we employ a second method on the similarity matrix derived from the graphical model. The similarity between i and j is the belief $b_{ij}(1)$ of the graphical model. Our method uses spectral clustering as described in [13].

The spectral clustering algorithm receives an affinity matrix $A \in [0,1]^{n \times n}$ that represents the pairwise similarities within a set of n elements. Let D be the diagonal matrix with elements $D_{ii} = \sum_{j=1}^{n} A_{ij}$. The normalized Laplacian of A is calculated as $L = D^{-1/2} A D^{-1/2}$. Let X be a matrix whose columns are the s eigenvectors corresponding to the s largest eigenvalues of L, with rows normalized to unit vectors. Each row in X can be regarded as an s-dimensional representation of the elements. These s-dimensional vectors are clustered by employing the k-means algorithm.

The Spectral Clustering Co-occurrence Stability (SCCS) algorithm works as follows: first, the spectral embedding of the data (X) is found. Then, k-means is applied repeatedly for either a fixed or varying number of clusters, with random initialization. In all of our experiments we employ k-means 200 times and set the number of clusters to 100. The computed relevancy score of an image (its similarity to the query) is the frequency in which it was clustered together with the query image. Given the high number of retrieved images (3000 in the Genizah datasets, 500 in the Art dataset), the noisy nature of the similarity matrix, and the large number of clusters, it is not surprising that the results of the clustering algorithm depict a large amount of variability between runs. This variability translates to rather continuous relevancy scores in all of our experiments.

5. Datasets

5.1. Synthetic Dataset

We simulate pairwise similarities between 1200 images by randomly generating a 1200×1200 matrix whose values are normally distributed around 0.3.

Figure 2: Samples of the Geneva join from which query (a) was taken. Fragment (b) was discovered by our retrieval method. The graphical model ranks all joins except (o), (p) and (q) among the top 50 candidates, with a significant enhancement over the raw similarity score – over 90 positions – for fragments (c), (d) and (e). Fragments (o), (p) and (q) are upgraded to the top 50 candidates in the SCCS step.

We create 40 imbalanced classes of images by sampling for each class a prior probability from the distribution [0.2 .. 1] and dividing by the sum of all 40 samples. Each image is randomly assigned to a class based on these priors. For each image, one to three pairwise similarities with class members are increased to values normally distributed around 0.9.

5.2. Genizah Datasets

The digital image collection of Cairo Genizah manuscripts contains 157,514 fragments. We experiment with two subsets of the Genizah, the Geneva benchmark, and a well-studied dataset containing halakhic books.

For each Genizah shelfmark, a pre-processing step is applied. First, hand-

writing based image properties are calculated for all images, as described in [3]. Each image is segmented into fragments that are binarized and aligned horizontally by rows. Keypoints are then detected in the image by identifying connected components, and local SIFT descriptors are calculated. All descriptors from the same image are combined into one vector using bag-of-features with a 500 keywords dictionary.

The similarity scores employ both simple and learned similarity scores and combine several scores together by the stacking technique. The similarity scores were taken from [3].

The Geneva Genizah Collection is a small collection of 150 Genizah fragments that were brought from Cairo to the Bibliothèque Publique et Universitaire of Geneve in 1896 and were stored there for over a century. Since their rediscovery in 2005, they have been studied intensively, and recently a full catalog of the collection was published [18].

Halakhic Books. A second dataset contains a few dozen joins of halakhic books from the eighth and ninth centuries [19], found manually by carefully inspecting all related Genizah fragments.

5.3. Art Dataset

We present a new dataset describing 81,449 unique digitized paintings, covering almost the entire Western and modern art. This dataset was collected from the visual art encyclopedia `www.wikipaintings.org`, a complete and well-structured online repository of fine art. We will make the dataset collected publicly available.

For each painting, there exists metadata specifying the artist name and nationality, art movement, year of creation, material, technique, painting dimensions and the gallery it is presented at. This collection contains over a thousand different artists, and is categorized to 27 art movements such as renaissance and impressionism, and to 45 genres such as abstract, graffiti and landscape.

We describe a painting by its gray level texture information, based on Steerable Filter Decomposition descriptors. These descriptors approximate a match-

ing set of Gabor filters with different frequencies and orientations. The descriptors are 28-dimensional, consisting of the mean and variance of a low pass filter, a high pass filter, and 12 sub-band filters from three scales and four orientation decompositions. The mean and variance roughly correspond to the sub-band energy, and characterize the strokes utilized by the artist [20, 21]. We used the matlab implementation of steerable pyramid feature extraction described in [22], available at `live.ece.utexas.edu/research/quality`. The pairwise similarity is measured by the euclidean distance between descriptors.

6. Experiments

We evaluate the unique contribution of employing the graphical model and the SCCS technique by comparing the final retrieval accuracy of our method to the intermediate results, that is to the retrieval given by the "vanilla" image-based similarity scores and to the retrieval of the beliefs learned by the graphical model.

We then compare our method to method [9]. Their method is also motivated by the Genizah dataset, and shares the aim of finding joins of images, as well as the use of a graphical model. We compare the retrieval of their learned beliefs, and also after applying our spectral clustering variant on these beliefs.

Finally, we compare to both ranking by Graph-pageRank and Graph-density suggested in [6], and use their publicly available code. These methods are designed for fusion of multiple ranking lists based on different similarity matrices. We only have one similarity matrix that defines a single ranking order, which can explain the low performance of these methods in our experiments.

We conduct the experiment on the synthetic data over 40 queries, one per class. The percentage of members of the class containing the query that are ranked among the top 50 retrieval results is reported in Table 1.

In the two Genizah experiments, shelfmarks from either one of the the Genizah subsets are used as the query source. The parameter n is set to 3000, and performance is evaluated by measuring the percentage of the known joins that

are retrieved among the 50 highest ranked results. The results are presented in Table 1. The contribution of both the similarity betterment and the SCCS step is evident, and there is a large performance gap compared to previous work.

Our retrieval method discovered an unknown join by querying a fragment from the Geneva benchmark, shown in Figure 2(a). The new join, shown in Figure 2(b), is cataloged as a page from the Babylonian Talmud tractate *Yevamot*, and was identified by a Talmud expert as a small part of another page already recorded as a join. The new join was ranked 468 by the raw similarity scores, a position typically overlooked by researchers, and was advanced by our method to the top 50. The Geneva catalog contains accurate and up-to-date information on the joins in the collection; hence a new join discovered by our method is surely unknown to the Genizah research community.

According to the Geneva catalog there are 27 known joins of the queried fragment, 22 out of them are in the Friedberg Genizah collection. The raw similarity scores of seven of these fragments were ranked bellow 3000 and discarded before the graphical model step, and one fragment for each of the remaining 15 joins was ranked among the 50 highest scores by our system. The retrieved fragments are presented in Figure 2(c)–(q).

A known join successfully retrieved by a query from the halakhic book dataset is shown in Figure 3. The fragments shown in (a)–(h) belong to the British Library collection. Query (a) and fragment (b) are erroneously identified in the library's catalog [23] as separate from the other fragments. The raw similarity scores retrieve (b), as well as fragments (c) and (d) from the group that does not contain the query. Our method was the only one to associate three additional members from the second group, shown in (e), (f) and (g). Three known joins, (h), (i) and (j), are ranked below 50 by all compared methods.

For the Art dataset, we randomly query 100 paintings, and use $n = 500$. For evaluation, images painted by the same artist and categorized as belonging to the same art movement and genre are regarded as similar. The accuracy presented in Table 1 is the ratio of similar paintings (by the above definition) out of all possible true matches (recall rate), among the top 100 retrieved images

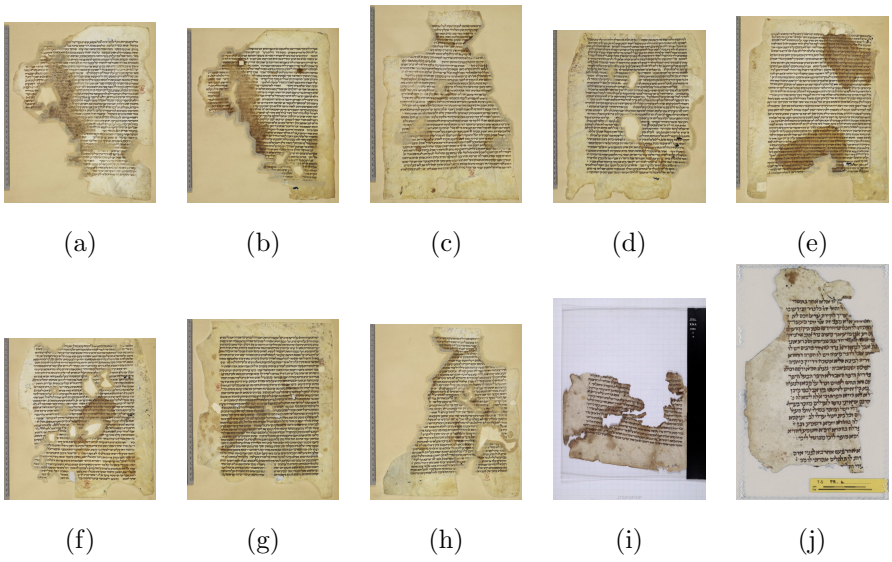


Figure 3: Samples from one join of the halakhic book dataset. (a) is the query fragment. Fragment (b) is the most similar and is retrieved with (c) and (d) by the raw similarity scores. Fragments (e), (f) and (g) are retrieved exclusively by our method, while (h), (i) and (j) are retrieved by none of the compared methods.

(as opposed to the top 50 retrievals in the Genizah dataset). We look further down the list since this similarity is very noisy.

Figure 4 shows three query images from the Art dataset. For each query, we show two images that are categorized by the same painter, movement and genre as the query, one of them is ranked by our method within the highest 100 candidates, and the other is ranked bellow 100.

In Figure 5, we show images whose ranking significantly increased due to the transitivity constraints in the graphical model. The query image is presented in (a), the ranking of image (b) climbed from 123 to 65, and the ranking image (c) increased from 114 to 83.

7. Discussion

Our method combines two different approaches – similarity betterment by graphical models and Spectral Clustering Co-occurrence Stability based on spec-

Table 1: Comparison of retrieval methods on the tested benchmarks. The results depict the recall rate within a fixed number of the top listed results. The methods are based on the raw similarity scores or on the scores after the belief-based method of [8] or [1]. The relevance feedback is not evaluated with the method in [8], as this method is computationally demanding. Two graph based method [6] are also evaluated. The results demonstrate the contribution of each of the components: belief-based similarity betterment, SCCS, and relevance feedback.

Method	**Geneva**	**Halakhic**	**Art**	**Synthetic**
Similarity	44.76%	51.07%	26.37%	17.32%
Similarity + SCCS	74.29%	42.50%	26.37%	53.16%
Similarity + SCCS + relevance feedback	75.36%	46.20%	30.77 %	67.32%
Graph-pageRank [6]	45.71%	52.06%	25.27%	17.15%
Graph-pageRank [6] + relevance feedback	47.21%	55.76%	31.87%	31.31%
Graph-density [6]	45.71%	51.89%	23.08%	17.24%
Graph-density [6] + relevance feedback	45.19%	54.63%	29.67%	30.14%
Belief-based similarity [8]	39.05%	46.79%	26.37%	15.10%
Belief-based similarity [8] + SCCS	63.81%	55.35%	28.57%	45.05%
Belief-based similarity (ours)	53.33%	52.55%	29.67%	38.31%
Belief-based similarity (ours) + SCCS	**77.14%**	59.47%	32.97%	56.66%
Belief-based similarity (ours) + SCCS + relevance feedback	**77.14%**	**63.17%**	**35.16%**	**70.82%**

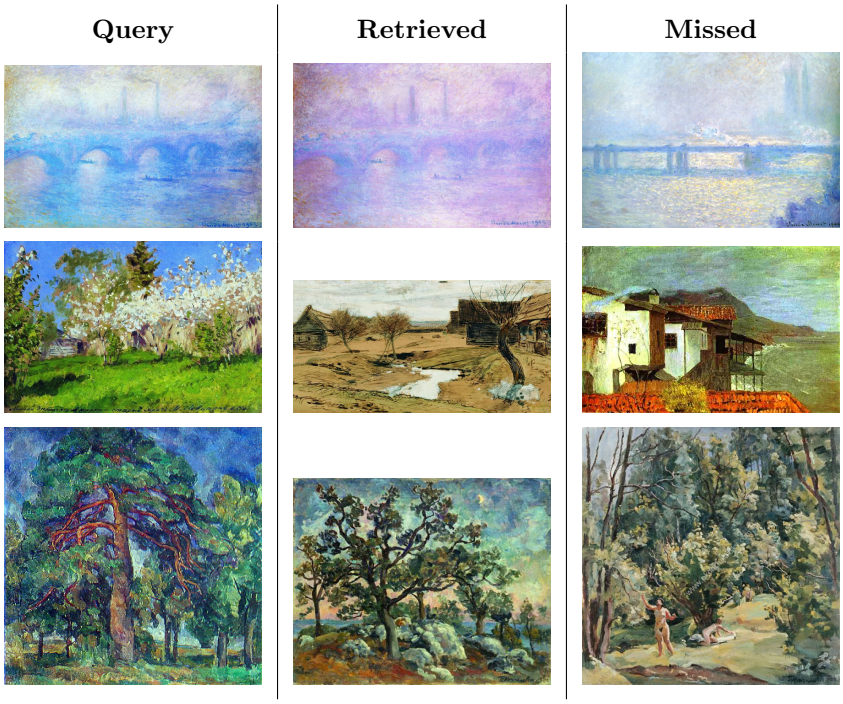


Figure 4: Images from the Art dataset. Each row shows a query, one similar image retrieved and one missed by our method. Top - Monet, cityscape, impressionism; middle - Levitan, landscape, realism; bottom - Konchalovsky, landscape, post-impressionism.

(a) (b) (c)

Figure 5: Ranking enhancement due to transitivity association. (a) is the query. (b) and (c) are the paintings with enhanced re-ranking. Painted by Sam Francis, abstract genre, abstract expressionism.

	Pellegrini et al. [8]	**Wolf et al. [9]**	**Our model**
Inference	Dual Decomposition [24]	Dual Decomposition [24]	message passing [12], adjusted c_α assignment
Subsampling transitivity factors	not required	uses the energy function in Eq. 1	uses the energy function in Eq. 2
Transitivity potential	penalize transitivity violations (1,1,0),(1,0,1),(0,1,1)	same as Pellegrini et al.	encourage linked triplets (1,1,1)
Prior probability	similarity values	similarity values	similarity values + increased prior of top candidates

Table 2: A summary of our contributions to the graphical model compared with [8, 9]. These modifications were made in response to the needs of the specific problem of query-based retrieval. Implementation details are described in Section 3.

tral analysis. The experiments demonstrate that the contributions of these two steps, which both tap into group congruency, albeit using very different approaches, partly overlap.

It is worth noting that the graphical model suggested in [8] and [9], which partially resembles our similarity betterment method, was not designed to be query specific, and therefore it does not consistently improve retrieval results. The main differences between the suggested model and the previous ones are summarized in Table 2.

The SCCS procedure, while highly effective on the Genizah dataset, was much less effective on the Art dataset. We hypothesize that this stems from lack of transitivity violation triplets in the Art data.

The Graph-pageRank and Graph-density methods of [6], which were designed primarily to combine multiple similarities together are not competitive in the context of our experiments, but did extremely well (in the original paper)

when combining local and holistic features. Note that the nature of the experiment is different, since previous work focused on the fusion of ranking from various sources, while we deal with a single, very noisy, similarity matrix.

Our method is partly motivated by join finding based on handwriting similarity in the Cairo Genizah. Digital Paleography holds the promise of scalability: it allows processing of sizable collections of images, 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 work patterns that are employed by the scholars when using more traditional tools (personal communication). Some researchers consider the search engine to be an "extended Google" and feel comfortable scanning the results obtained by it, looking for the images 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. This is a separate research direction not developed here.

8. Summary

In this work we explore the use of re-ranking tools in order to improve 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 help locate unknown joins. We suggest a graphical model adapted for query specific retrieval that focuses on factors containing the query variable, and strengthens specific transitivity connections through the potential function. The model parameters are set such that the entropy expressions in the objective function remain proportional and do not artificially force the beliefs to uniform probability. Our Spectral Clustering Co-occurrence Stability score measures the relevancy of an image to the query by the frequency of their co-occurrences within the same cluster as it emerges from multiple k-

means runs.

By using scalable and distributed inference methods, the underlying computational tasks are solved in minutes and can be further sped up by using multiple machines and by caching previous computations. For the exploration of the Genizah dataset, it is our plan to suggest the use of pre-computed results obtained by our re-ranking method to each feasible Genizah query.

References

[1] I. Ben-Shalom, N. Levy, L. Wolf, N. Dershowitz, A. Ben-Shalom, R. Shweka, Y. Choueka, T. Hazan, Y. Bar, Congruency-based reranking, in: Computer Vision and Pattern Recognition (CVPR), 2014.

[2] M. Glickman, Sacred Treasure–the Cairo Genizah: The Amazing Discoveries of Forgotten Jewish History in an Egyptian Synagogue Attic, Jewish Lights, 2011.

[3] L. Wolf, R. Littman, N. Mayer, T. German, N. Dershowitz, R. Shweka, Y. Choueka, Identifying join candidates in the Cairo Genizah, International Journal of Computer Vision 94 (1) (2011) 118–135.

[4] D. Qin, S. Gammeter, L. Bossard, T. Quack, L. Van Gool, Hello neighbor: accurate object retrieval with k-reciprocal nearest neighbors, in: Computer Vision and Pattern Recognition (CVPR), IEEE, 2011.

[5] M. Daoud, L. Tamine-Lechani, M. Boughanem, Using a concept-based user context for search personalization, in: International Conference of Data Mining and Knowledge Engineering (ICDMKE), London, UK, 2008.

[6] S. Zhang, M. Yang, T. Cour, K. Yu, D. N. Metaxas, Query specific fusion for image retrieval, in: European Conference on Computer Vision (ECCV), 2012.

[7] N. Shental, A. Zomet, T. Hertz, Y. Weiss, Pairwise clustering and graphical models, Advances in Neural Information Processing Systems 16.

[8] S. Pellegrini, A. Ess, L. Van Gool, Improving data association by joint modeling of pedestrian trajectories and groupings, in: European Conference on Computer Vision (ECCV), Springer, 2010.

[9] L. Wolf, L. Litwak, N. Dershowitz, R. Shweka, Y. Choueka, Active clustering of document fragments using information derived from both images and catalogs, in: International Conference on Computer Vision (ICCV), 2011.

[10] M. J. Wainwright, T. S. Jaakkola, A. S. Willsky, A new class of upper bounds on the log partition function, Information Theory, IEEE Transactions 51 (2005) 2313–2335.

[11] T. Hazan, A. Shashua, Norm-product belief propagation: Primal-dual message-passing for approximate inference, Information Theory, IEEE Transactions on 56.

[12] A. Schwing, T. Hazan, M. Pollefeys, R. Urtasun, Distributed message passing for large scale graphical models, in: Computer Vision and Pattern Recognition (CVPR), IEEE, 2011.

[13] A. Y. Ng, M. I. Jordan, Y. Weiss, et al., On spectral clustering: Analysis and an algorithm, Advances in Neural Information Processing Systems (NIPS) 2.

[14] D. Mavroeidis, E. Marchiori, Feature selection for k-means clustering stability: theoretical analysis and an algorithm, Data Mining and Knowledge Discovery.

[15] S. Ben-David, U. Von Luxburg, D. Pál, A sober look at clustering stability, in: Learning Theory, 2006.

[16] H. C. Lee, A. Borodin, Criteria for cluster-based personalized search, Internet Mathematics 6 (3).

[17] T. Hazan, J. Peng, A. Shashua, Tightening fractional covering upper bounds on the partition function for high-order region graphs, arXiv:1210.4881.

[18] D. Rosenthal, The Cairo Genizah collection in Geneva: Catalogue and studies, Magnes Press, Jerusalem.

[19] R. Shweka, Studies in Halakhot Gedolot – Text and Recension, Jerusalem: Hebrew University, 2008.

[20] A. I. Deac, J. van der Lubbe, E. Backer, Feature selection for paintings classification by optimal tree pruning, in: Multimedia Content Representation, Classification and Security, Springer, 2006.

[21] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, T. N. Pappas, Classifying paintings by artistic genre: An analysis of features & classifiers, in: Multimedia Signal Processing workshop, IEEE, 2009.

[22] H. R. Sheikh, A. C. Bovik, Image information and visual quality, Image Processing.

[23] G. Margoliouth, Catalogue of the Hebrew and Samaritan manuscripts in the British Museum.

[24] N. Komodakis, N. Paragios, G. Tziritas, MRF optimization via dual decomposition: Message-passing revisited, in: International Conference on Computer Vision (ICCV), IEEE, 2007.