frontiers in Digital Humanities

ORIGINAL RESEARCH
published: 22 August 2016
doi: 10.3389/fdigh.2016.00007

Active Congruency-Based Reranking

Itai Ben Shalom[1], Noga Levy[1], Lior Wolf[1], Nachum Dershowitz[1], Adiel Ben Shalom[2], Roni Shweka[2], Yaacov Choueka[2], Tamir Hazan[3] and Yaniv Bar[2]*

[1] The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel, [2] Genazim Digital, The Friedberg Genizah Project, Jerusalem, Israel, [3] Faculty of Industrial Engineering & Management, Technion, Haifa, Israel

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. To this, we add an active relevance-based re-ranking process in order to leverage true matches, which have very low similarity to the query. The utility of the tool is demonstrated within the context of a visual search of documents from the Cairo Genizah. It is also demonstrated in a completely different domain or retrieving, given an input image of a painting, other related paintings.

Keywords: digital paleography and book history, active clustering, graphical Bayesian models, Genizah, image search

OPEN ACCESS

Edited by:
Arianna Ciula, University of Roehampton, UK

Reviewed by:
Nicholas R. Howe, Smith College, USA
Alicia Fornes, Universitat Autònoma de Barcelona, Spain

****Correspondence:***
Lior Wolf
wolf@cs.tau.ac.il

Specialty section:
This article was submitted to Digital Paleography and Book History, a section of the journal Frontiers in Digital Humanities

Received: *24 December 2015*
Accepted: *27 July 2016*
Published: *22 August 2016*

Citation:
Ben Shalom I, Levy N, Wolf L, Dershowitz N, Ben Shalom A, Shweka R, Choueka Y, Hazan T and Bar Y (2016) Active Congruency-Based Reranking. Front. Digit. Humanit. 3:7. doi: 10.3389/fdigh.2016.00007

1. INTRODUCTION

Scholarly search in digital archives is inherently different from casual goggling. When a query is performed as part of an ongoing research, the scholar is interested in collecting all of the relevant results to her study. Just obtaining the one result that best matches the intent of the query would not suffice. Rather, the scholar is interested in gathering all of the results that have some relevancy.

In this work, we focus on large-scale digital collections, where each query can potentially retrieve thousands of results. While many results might be superfluous, and arise due to the inaccuracies of the search algorithm, many others would require the scholar's attention. To address this and to support the development of practical search tools for researchers, we develop algorithms that consider the set of all retrieved documents and identify coherent groups that include the query. The exploration time needed to examine query results is, therefore, reduced, since the group elements reinforce each other. The researcher's attention is, thus, used more economically and spent on documents that match the query in a more meaningful way.

We consider two large-scale digital collections: one containing images of manuscript fragments and the other containing images of paintings. The manuscript collection contains digital versions of the Cairo Genizah manuscripts. It is a result of an ongoing digitization effort by the Friedberg Genizah Project (Glickman, 2011) and comprises 157,514 fragments.

The Cairo Genizah is a unique 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 manuscripts are written mainly in Hebrew, Aramaic, and Judeo-Arabic (in Hebrew characters). The collection has had an enormous impact on 20th century scholarship in a multitude of fields, including the Bible, rabbinics, liturgy, history, and philology. A major challenge in extracting knowledge from the Genizah is that pages and fragments from the

same original manuscript may now reside in different, remote collections. The identification of such connections is a challenging problem and a great deal of time and effort has been invested by human scholars on manually rejoining fragments of the same original book or pamphlet.

A visual similarity measure was previously developed (Wolf et al., 2011a), 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, by 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.

In addition to the Genizah collection, we also consider a dataset of digitized art images. This large collection was downloaded from the visual art encyclopedia www.wikipaintings.org, which thoroughly covers Western and modern art. The metadata of each image was also obtained. It includes the artist's name and nationality, a classification to art movement, a description of the content, etc. In our experiments, we automatically evaluate the relevancy of the retrieval results by employing this metadata: we regard as relevant images that were painted by the same artist and categorized as the same art movement and content.

2. RELATED WORK

This work deals with content-based image retrieval (CBIR). The vast majority of CBIR methods consider, for each database image, the direct similarity with the query. These methods ignore the correlations within the set of retrieved images. In our work, we exploit these correlations in order to better infer the relevancy of each retrieved result to the query at hand. In other words, the relevancy of a candidate item is estimated by considering both its similarity to the query and its similarity to other promising candidates.

Qin et al. (2011) exploit the correlations among dataset images by enhancing the similarity between reciprocal nearest neighbors. Daoud et al. (2008), working in the domain of personalized web queries, employ graphical models for information retrieval. The search results are enhanced by exploiting associations among items and by regarding personal preferences. Zhang et al. (2012) fuse query-specific ranking orders based on various similarity measures. In their work, the various rankings are represented as weighted graphs, which are integrated to a single fused graph. Either the pageRank method or a method for finding the maximal weight density are then employed on that graph.

We choose to employ probabilistic graphical models in our method since we wish to build upon local correlations among small subsets of the images. Such local correlations are effectively represented as the edges of 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 (Shental et al., 2003) assumes that the number of classes is known in advance. Pedestrian grouping identification with graphical models (Pellegrini et al., 2010) 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 (Wolf et al., 2011b).

Finding the maximum *a posteriori* (MAP) assignment in a graphical model involves searching in an 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 (Wainwright et al., 2005; Hazan and Shashua, 2010; Schwing et al., 2011). In our work, we employ the distributed computation method of (Schwing et al., 2011), to which we contribute a heuristic method for selecting the parameter of the inference algorithm.

The variables inferred from the graphical model can be used directly as scores of similarity to the query, or indirectly as intermediate similarities that are further processed. In the latter case, we suggest the spectral clustering co-occurrence stability method described in Section 4. This method repeatedly employs the spectral clustering method of (Ng et al., 2002) in order to identify a stable re-ranking. The stability of the clustering method can be analyzed by testing the influence of small perturbations in the data (Mavroeidis and Marchiori, 2013) and is arguably an appropriate measure of clustering-method quality (Ben-David et al., 2006). We propose a more local perspective: by considering only the cluster that contains the query at hand, we rank highly the documents which tend to co-occur with it. In the context of personalized web search, the idea of finding a good ranking based on co-occurrence frequencies was suggested by Lee and Borodin (2009). 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. In that previous work, "cluster" is used to describe a set of documents tagged by the same label.

3. IMPROVING THE SIMILARITY

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 regard to the query of interest.

3.1. Model Variables

We define a task-specific probabilistic graphical model. In this model, a set of binary variables l_{ij} denote linking between pairs of items. This includes the potential linking between the query and one of the retrieved candidates or between pairs of candidates. The set of admissible links is not arbitrary since groups need to be consistent, i.e., all items within a group need to be pairwise linked. This is modeled as a set of transitivity constraints between the linking variables. The prior probabilities of the linking variables l_{ij} are derived from the pairwise handwriting-based image similarity of i and j (Wolf et al., 2011a), and are expressed by the pairwise models $\gamma_{ij}(l_{ij})$. A $\gamma_{ij}(1)$ close to one indicates that the pair of images is visually similar. It is close to zero otherwise, and vice versa for dissimilar images. The visual similarity is further augmented in the following manner: we assign the potential of the top t ranked candidates to 1, since empirically these candidates are often linked to true matches or are a match themselves. Throughout our experiments, t is set to 10.

3.2. Transitivity Constraints

For a given query q, we examine the following triplets of pairs (l_{qi}, l_{qj}, l_{ij}). Assignments in which a triplet contains a single zero value violates transitivity, since in such assignments one image is similar to the two other images while these images are not similar to each other.

Transitivity potentials χ are constructed for every pair of images (i, j) as follows: the 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.

Since we examine long lists of retrieved results, the set of all transitivity constraints is too large to be supported by current inference engines. We, therefore, apply subsampling (see Section 3.3), and focus only on the triplets suspected as violating transitivity. The potentials χ then pushes toward assigning (1, 1, 1) of the corresponding variables.

The analog potentials used in previous work (Pellegrini et al., 2010; Wolf et al., 2011b) can either increase or decrease the linking between 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. As a result, transitivity can be either encouraged (obtaining a linking pattern of (1,1,1)), or the violation can be resolved by ignoring high similarities (obtaining one of the patterns (0,0,1), (0,1,0), (1,0,0), or (0,0,0)). It was observed in our experiments that our transitivity potential, which is more conservative, 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 Wolf et al. (2011b), 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

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

Unlike Wolf et al. (2011b), 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

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

We set β to 2 in all of the experiments, since transitivity is likely to exist when two images j and k resemble a third image with high probability. We, therefore, wish to identify these cases even for intermediate values of $\gamma_{jk}(0)$. The violation energy function (equation (2)) is evaluated for all triplets, and the N triplets with the maximal values are elected. In our experiments $N = 2000$, which leads to a sizeable decrease in the computational complexity of the inference process.

3.4. Optimization Problem

The graphical-model formulation we employ below follows Schwing et al. (2011), 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. A term capturing the variational entropy $H(b)$ is defined as

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

Inference is performed by a Convex Belief Propagation procedure directly optimizing the following problem:

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

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.

3.5. Calibration of the Trade-off Parameters

The objective function presented in equation (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 (Hazan et al., 2012). 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, depending on the subsampling process, a neighboring transitivity factor per each candidate $j \neq i$. On the other hand, for a binary variable that denotes the link between two candidates l_{ij}, there can be only one possible transitivity factor: $\alpha = (l_{qi}, l_{qj}, l_{ij})$.

Consider the entropy $H(b_\alpha)$: since each transitivity factor has a matching entropy expression, the belief of nodes l_{qi} with many neighboring factors is strongly tied to the entropy expression. Recall that the entropy is maximized when all states are equally probable. Therefore, the entropy pushes the distribution toward a 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. We set $c_\alpha = \eta / |\overline{N_\alpha}|$ in order to limit the effect of the entropy expressions on the beliefs of the binary variables. In all of our experiments, the parameter η is fixed at a value of 0.1.

4. SPECTRAL CLUSTERING CO-OCCURRENCE STABILITY

To reinforce congruent groups of similar images that are also similar to the query, 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 Ng et al. (2002).

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. RELEVANCE FEEDBACK

When the querier is provided with the query results, she inspects them one by one in order to identify the relevant items. As she goes through the list, marking the results she would like to collect, the system can make use of this information and improve the ranking of the results that were not inspected yet.

We have experimented with several alternative ways to alter the graphical model, thereby modifying the ranking of the remaining retrieval items. One way would be to strengthen the potentials of the links between the query and the results identified as relevant. In addition, one can weaken the potentials of results that are marked irrelevant. A third possibility would be to construct additional queries based on the known relevant results and integrate the graphical models (early integration) or the obtained ranked lists (late integration).

While each of these options might improve results, each has multiple parameters to consider. As a result, a proposed method can work well on one dataset and perform poorly on another. In order to provide a method that is robust and almost parameter-free, we employ in this work a simple elimination technique. In our system, the querier is provided with the top K results to inspect. She selects the relevant ones among them. All irrelevant results are removed from the graph and the computation is repeated. The top K results that were not already displayed are then presented and the process repeats.

This simple technique is enough to utilize the relevance feedback effectively. Results that are highly ranked have the greatest effect on the topology of the space of database items since they have high similarities to the query and among themselves. Removing results that are irrelevant causes significant changes to the underlying topology. As mentioned, this method has the advantage of not requiring extensive parameter tuning. In our experience, the results are largely independent of the parameter K as long it is small enough. The parameter can simply be set by taking into account user-interface and run-time considerations. In our experiments, we set the value to $K = 10$.

6. DATASETS

6.1. Synthetic Dataset

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

6.2. 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, handwriting-based image properties are calculated for all images, as described in Wolf et al. (2011a). 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 keyword 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 Wolf et al. (2011a).

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 (Rosenthal, 2010).

Halakhic Books: A second dataset contains a few dozen joins of halakhic books from the eighth and ninth centuries (Shweka, 2008), found manually by carefully inspecting all related Genizah fragments.

6.3. 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 matching 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 (Deac et al., 2006; Zujovic et al., 2009). We used the matlab implementation of steerable pyramid feature extraction described in Sheikh and Bovik (2006), available at live.ece.utexas.edu/research/quality. The pairwise similarity is measured by the euclidean distance between descriptors.

7. EXPERIMENTS

We evaluate the unique contribution of employing each of the underlying components of our system: (i) the graphical model; (ii) the SCCS technique; and (iii) relevance feedback. This is done by comparing the final retrieval accuracy of our method to the intermediate results. That is, the retrieval given by the "vanilla" image-based similarity scores, the retrieval of the beliefs learned by the graphical model, and the retrieval of the ranked list given to the user before he provides any feedback to the system.

The effect of user interaction is simulated in our experiments. Since we know the ground truth, we are able to provide, at each relevance feedback step, the true relevancy of the provided retrievals. In order to simulate a realistic scenario, feedback is provided to groups of ten results at once (i.e., $K = 10$); then, the list of retrievals is reranked and another group of ten results is evaluated. Of course, the results provided during this interactive process count as part of the retrieval results and are scored.

We also compare our method to baseline methods from the literature:

- The method of Wolf et al. (2011b), using the code shared by Wolf et al. 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. This method is too slow, however, to allow for repeated interactive application.
- Both ranking by Graph-pageRank and Graph-density suggested in Zhang et al. (2012), using the authors' 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 parameter n is set such that all other examples are used in consequent stages. The percentage of members of the class containing the query that are ranked among the top 50 retrieval results is reported as the success score in **Table 1**. All stages of our system seem to contribute significantly in the querying process.

TABLE 1 | Comparison of retrieval methods on the tested benchmarks.

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 (Zhang et al., 2012)	45.71	52.06	25.27	17.15
Graph-pageRank (Zhang et al., 2012) + relevance feedback	47.21	55.76	31.87	31.31
Graph-density (Zhang et al., 2012)	45.71	51.89	23.08	17.24
Graph-density (Zhang et al., 2012) + relevance feedback	45.19	54.63	29.67	30.14
Belief-based similarity (Pellegrini et al., 2010)	39.05	46.79	26.37	15.10
Belief-based similarity (Pellegrini et al., 2010) + 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**

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 Pellegrini et al. (2010) or Ben-Shalom et al. (2014). The relevance feedback is not evaluated with the method in Pellegrini et al. (2010), since this method is computationally demanding. Two graph-based methods (Zhang et al., 2012) are also evaluated. The results demonstrate the contribution of each of the components: belief-based similarity betterment, SCCS, and relevance feedback. Bold is an indicator for the best result.

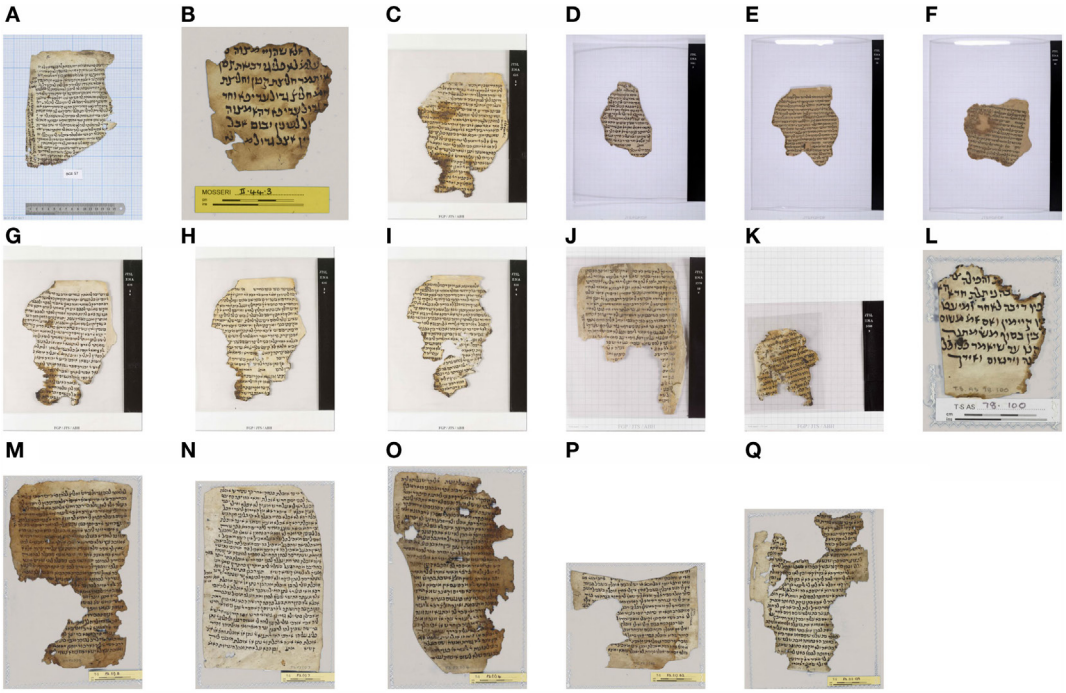

FIGURE 1 | **(B–Q) Samples of the Geneva join from which query (A) was taken**. Fragment **(B)** was discovered by our retrieval method. The graphical model ranks all joins except **(O–Q)** among the top 50 candidates, with a significant enhancement over the raw similarity score – over 90 positions – for fragments **(C–E)**. Fragments **(O–Q)** are upgraded to the top 50 candidates in the SCCS step.

In the two Genizah experiments, shelfmarks from either one of 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 contributions of both the similarity betterment and the SCCS step are evident, and there is a large performance gap compared to previous methods. The contribution of the interactive step is limited, perhaps due to a diminishing-returns effect.

Our multi-step retrieval method, without the interactive steps, discovered an unknown join by querying a fragment from the Geneva benchmark, shown in **Figure 1A**. The new join, shown in **Figure 1B**, 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 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 below 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 **Figures 1C–Q**.

A known join successfully retrieved by a query from the halakhic book dataset is shown in **Figure 2**. 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 (Margoliouth, 1905, p. 52) 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, even before any interactive step, 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 (as opposed to the top 50 retrievals in the Genizah dataset). The reason that we look further down the list is that the computed similarities for this dataset are very noisy; looking at the top 50 results would mean very low recall rates for all methods.

Figure 3 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 below 100. In **Figure 4**, 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.

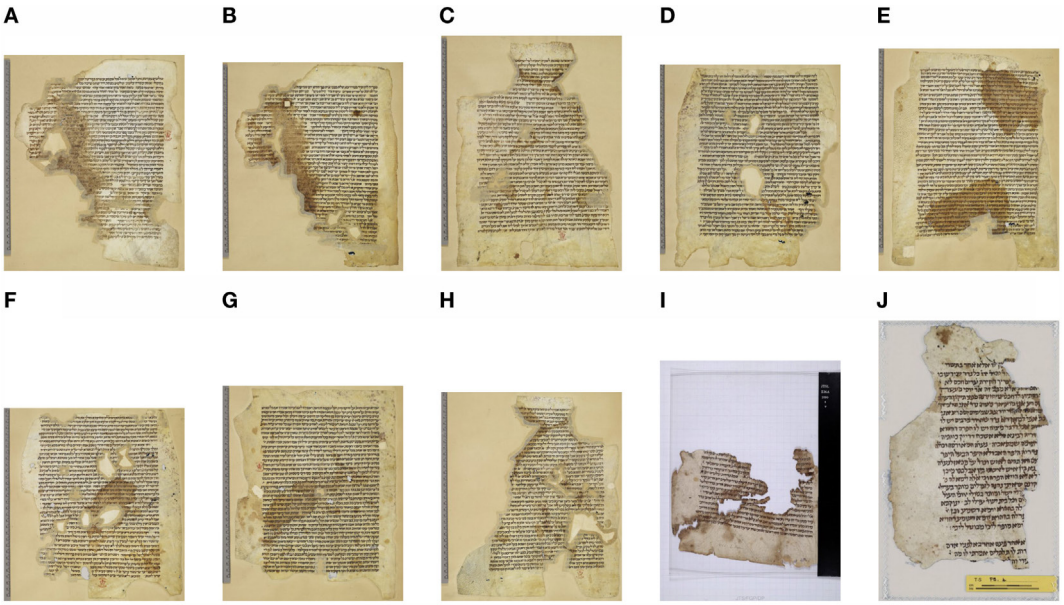

FIGURE 2 | **Samples from one join of the halakhic book dataset**. **(A)** is the query fragment. Fragment **(B)** is the most similar and is retrieved with **(C,D)** by the raw similarity scores. Fragments **(E–G)** are retrieved exclusively by our method, while **(H–J)** are retrieved by none of the compared methods.

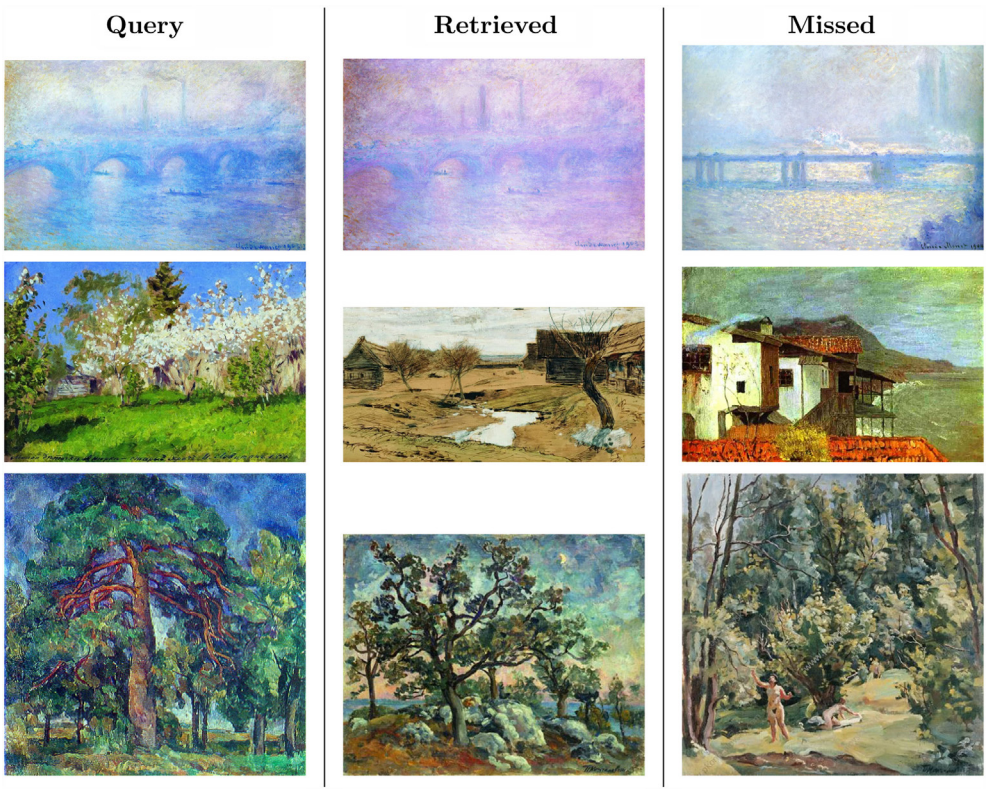

FIGURE 3 | **Images from the Art dataset**. Each row shows a query, one similar image retrieved and one missed by our method. Top – Monet, cityscape, impressionism; middle – Levitan, landscape, realism; bottom – Konchalovsky, landscape, post-impressionism.

8. DISCUSSION

Our method combines two different approaches – similarity betterment by graphical models and Spectral Clustering Co-occurrence Stability based on spectral 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 Pellegrini et al. (2010) and Wolf et al. (2011b), 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 the lack of transitivity violation triplets in the Art data. In addition, the features used in the art dataset capture only a fraction of the visual information in a painting.

The Graph-pageRank and Graph-density methods of Zhang et al. (2012), which were designed primarily to combine multiple similarities together, are not competitive in the context of our experiments. However, they 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.

Relevance feedback can provide a significant amount of actionable information. While in casual web search, it may not be realistic to expect the user to mark relevancy, collecting results is an inherent part of scholarly research of the type considered here. In this work, we employ such information to alter the graph structure of our problem. It might be beneficial to employ such information even at an earlier stage and alter the underlying pairwise similarities.

One limitation of our technique is the usage of multiple parameters. Due to computational reasons, it is infeasible to optimize these parameters in an exhaustive way. It is likely that results could improve further by discovering better parameter values. Note, however, that the parameters are intuitive and interpretable: the transitivity potentials $\chi(l_{qi}, l_{qj}, l_{ij})$ reinforces faint connections when appropriate, by assigning a high probability to the clique (1,1,1); we offer (Section 3.5) a unique way to calibrate the trade-off parameters c_{ij} and c_{α}; the meaning of the trade-off parameter β is given in Section 3.3; the parameters of SCCS and the relevance feedback are mostly discrete and have clear meanings.

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

FIGURE 4 | Ranking enhancement due to transitivity association. **(A)** is the query. **(B,C)** are the paintings with enhanced re-ranking. Painted by Sam Francis, abstract genre, abstract expressionism.

TABLE 2 | A summary of our contributions to the graphical model compared with Pellegrini et al. (2010) and Wolf et al. (2011b).

	Pellegrini et al. (2010)	Wolf et al. (2011b)	Our model
Inference	Dual decomposition (Komodakis et al., 2007)	Dual decomposition (Komodakis et al., 2007)	Message passing (Schwing et al., 2011), adjusted c_{α} assignment
Subsampling transitivity factors	Not required	Uses the energy function in equation (1)	Uses the energy function in equation (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

These modifications were made in response to the needs of the specific problem of query-based retrieval. Implementation details are described in Section 3.

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 and 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.

9. 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. Finally, a robust relevance feedback mechanism is employed.

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.

AUTHOR NOTES

A preliminary and partial version of this work has been published in Ben-Shalom et al. (2014) and is built upon herein under the standard terms and conditions of IEEE.

AUTHOR CONTRIBUTIONS

All authors contributed jointly to the work conducted as part of this project.

FUNDING

This research was supported in part by Grant #I-145-101.3-2013 from the German-Israeli Foundation for Scientific Research and Development, Grant 1330/14 from the Israel Science Foundation, and a grant from the Deutsch-Israelische Projektkooperation. ND's research benefited from a fellowship at the Paris Institute for Advanced Studies (France), with the financial support of the French state, managed by the French National Research Agency's "Investissements d'avenir" program (ANR-11-LABX-0027-01 Labex RFIEA+).

REFERENCES

Ben-David, S., Von Luxburg, U., and Pál, D. (2006). A sober look at clustering stability. In *Learning Theory: 19th Annual Conference on Learning Theory (COLT 2006)*, Edited by G. Lugosi and H.-U. Simon (Berlin: Springer), 5–19.

Ben-Shalom, I., Levy, N., Wolf, L., Dershowitz, N., Ben-Shalom, A., Shweka, R., et al. (2014). Congruency-based reranking. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)* (Washington, DC: IEEE Computer Society), 2107–2114. doi:10.1109/CVPR.2014.270

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

Deac, A.I., van der Lubbe, J., and Backer, E. (2006). Feature selection for paintings classification by optimal tree pruning. In *Proceedings of Multimedia Content Representation, Classification and Security: International Workshop, MRCS 2006, Istanbul, Turkey, Sep 11–13, 2006*, Edited by G. Bilge, A. K. Jain, A. M. Tekalp and B. Sankur (Berlin: Springer Berlin Heidelberg), 354–361. doi:10.1007/11848035_47

Glickman, M. (2011). *Sacred Treasure–the Cairo Genizah: The Amazing Discoveries of Forgotten Jewish History in an Egyptian Synagogue Attic*. Nashville, TN: Jewish Lights.

Hazan, T., Peng, J., and Shashua, A. (2012). Tightening fractional covering upper bounds on the partition function for high-order region graphs. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, Aug 14–18, 2012*, Edited by N. de Freitas and K. P. Murphy (Catalina Island, CA: AUAI Press), 356–366.

Hazan, T., and Shashua, A. (2010). Norm-product belief propagation: primal-dual message-passing for approximate inference. *IEEE Trans. Inf. Theory* 56: 6294–316. doi:10.1109/TIT.2010.2079014

Komodakis, N., Paragios, N., and Tziritas, G. (2007). MRF optimization via dual decomposition: message-passing revisited. In *International Conference on Computer Vision (ICCV)*, IEEE.

Lee, H.C., and Borodin, A. (2009). Criteria for cluster-based personalized search. *Internet Math.* 6: 399–435. doi:10.1080/15427951.2009.10390647

Margoliouth, G. (1905). *Catalogue of the Hebrew and Samaritan Manuscripts in the British Museum, II* (London), 52.

Mavroeidis, D., and Marchiori, E. (2013). Feature selection for k-means clustering stability: theoretical analysis and an algorithm. *Data Min. Knowl. Discov.* 28: 918. doi:10.1007/s10618-013-0320-3

Ng, A.Y., Jordan, M.I., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 2.

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

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

Rosenthal, D. (2010). *The Cairo Genizah collection in Geneva: Catalogue and studies*. Jerusalem: Magnes Press.

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

Sheikh, H.R., and Bovik, A.C. (2006). Image information and visual quality. *IEEE Trans. Image Process* 15: 430–44. doi:10.1109/TIP.2005.859378

Shental, N., Zomet, A., Hertz, T., and Weiss, Y. (2003). Pairwise clustering and graphical models. In *Advances in Neural Information Processing Systems*, 16.

Shweka, R. (2008). *Studies in Halakhot Gedolot – Text and Recension*. Jerusalem: Hebrew University.

Wainwright, M.J., Jaakkola, T.S., and Willsky, A.S. (2005). A new class of upper bounds on the log partition function. *IEEE Trans. Inform. Theory* 51: 2313–35. doi:10.1109/TIT.2005.850091

Wolf, L., Littman, R., Mayer, N., German, T., Dershowitz, N., Shweka, R., et al. (2011a). Identifying join candidates in the Cairo Genizah. *Int. J. Comput. Vision* 94: 118–35. doi:10.1007/s11263-010-0389-8

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

Zhang, S., Yang, M., Cour, T., Yu, K., and Metaxas, D.N. (2012). Query specific fusion for image retrieval. In *European Conference on Computer Vision (ECCV)*.

Zujovic, J., Gandy, L., Friedman, S., Pardo, B., and Pappas, T.N. (2009). Classifying paintings by artistic genre: an analysis of features & classifiers. In *Multimedia Signal Processing Workshop*, IEEE.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.