Ranking Without Learning: Towards Historical Relevance-based Ranking of Social Images

Min Min Chew [‡] Sourav S Bhowmick[‡] Adam Jatowt[§]
[‡]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[§]Graduate School of Informatics, Kyoto University, Japan
assourav@ntu.edu.sg,adam@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Tag-based Social Image Retrieval (TagIR) aims to find relevant social images using keyword queries. State-of-the-art TagIR techniques typically rank query results based on relevance, temporal or popularity criteria. However, these criteria may not always be sufficient to match diverse search intents of users. In this paper, we present a novel ranking scheme that ranks query results (images) based on their *historical relevance*. Informally, an image is *historically relevant* if its visual content is *relevant* to the query and it depicts objects, scenes, or events that are *related* to human history. To this end, we propose a *learning-agnostic* technique that leverages Wikipedia to quantify *historical relevance* of images. We empirically demonstrate the effectiveness of our ranking scheme using *Flickr* dataset.

ACM Reference Format:
Min Min Chew [‡] Sourav S Bhowmick[‡] Adam Jatowt[§]. 2018. Ranking Without Learning: Towards Historical Relevance-based Ranking of Social Images. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210100

1 INTRODUCTION

Given a keyword query, a *Tag-based Social Image Retrieval* (TagIR) search engine returns a *ranked* list of images annotated with query words. The state-of-the-art approaches to rank images are primarily based on the *relevance score* [4, 9] (images annotated with the most *relevant* tags to the query are ranked high) or *result diversity* [10]. However, this may not always be sufficient criteria to match images with user search intents due to the increasing content diversity of image sharing platforms coupled with high diversity of user search needs calling for novel and specialized retrieval paradigms. For instance, recently millions of tagged historical images were posted in *Flickr* (www.bbc.com/news/technology-28976849). Yet, traditional relevance score-based techniques are ineffective in ranking images based on their *historical relevance*. Intuitively, an image is *historically relevant* if (a) its visual content is relevant to the query *and* (b) it depicts objects, scenes, or events *related* to human history (formally defined later).

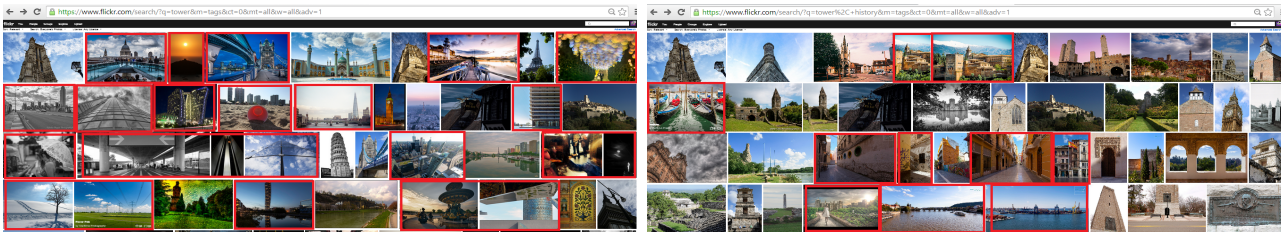
SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210100

For example, consider the following scenario. Jane, a student, wishes to find images of towers with historical importance. Intuitively, *Leaning Tower of Pisa* would be more relevant to her information need than *Petronas Tower* in Malaysia due to the former's rich affiliation with European history. She submits the query `"tower"` on *Flickr* (using "tags only" setting) which returns several images related to towers as depicted in Figure 1(a). As Jane is an impatient user, she does not cherish the idea of sifting through all images to find desired ones, neither of studying world history to find clues on towers that played key roles in human history. She is willing to browse only the top-30 results in the list. Although some of the top-ranked images indeed have strong historical significance (*e.g., Leaning Tower of Pisa*), many do not (highlighted in red boxes) as they are either not related to tower or are just tall modern buildings. Jane then expands the query by adding the keyword `"history"`, hoping to retrieve more historically-relevant images. Although now there are more images related to historical buildings (Figure 1(b)), not all are relevant as they are not towers (*e.g., Charles bridge*). More importantly, images related to *Leaning Tower of Pisa* are no more in the top-ranked result set as they may not be annotated with the `"history"` tag. Then, how can Jane rank her search results according to the historical relevance?

In this paper, we address the aforementioned problem by laying down the vision of ranking images of a search query based on their historical relevance. In particular, we propose a technique to associate an image with a *historical relevance score* which is then used to rank the images in the query results. Note that it is challenging to measure historical relevance of a social image. This is because the relationship between the image content and history is not necessarily captured explicitly by its tags, partly, due to the incompleteness and noisiness of tagging process.

Given the tremendous success deep learning-based frameworks have achieved in recent times in the area of object recognition, it may seem that such a framework can be leveraged to search and rank images based on their historical relevance. Indeed, in recent times such framework is used for retrieving domain-specific images such as fashion [6]. However, the success of a deep learning-based framework depends on the availability of huge volumes of training data. Unfortunately, it is hard to generate such data in our problem setting as it demands experts with deep knowledge and coverage of human history. Furthermore, it is impossible to determine historical relevance by simply looking at image pixels or tags. Hence, *in this paper we depart from this popular strategy and explore if it is possible to rank historical images effectively without utilizing a learning framework.* To this end, we leverage Wikipedia to quantify the *historical relevance score* due to its high coverage



(a) Query "tower" (b) Query "tower, history"

Figure 1: Query results in *Flickr*.

and capability to provide historical information related to many real-world entities.

It is worth noting that our proposed vision is not intended to replace existing relevance score-based ranking scheme in TagIR. Certainly, historical relevance-based ranking is not going to be appropriate for all queries. For instance, rarely a user would like to rank the result images of the query "orange" based on their historical relevance. Such ranking is more meaningful when the query keyword(s) has historical significance (*e.g.,* egypt, architecture, tower) and a user wishes to rank the results based on it (*e.g., Flickr* allows one to choose a ranking criteria ("relevant", "recent", "interesting")). Hence, the proposed ranking scheme is meant to *complement* existing schemes in TagIR.

We make the following contributions: (1) We introduce a novel research problem of ranking social images according to their historical relevance. (2) We propose an unsupervised, learning-agnostic technique that outputs images not only relevant but also historically significant. (3) We experimentally demonstrate its effectiveness.

2 RELATED WORK

TagIR research can be broadly categorized into three types, *indexing*, *scoring*, and *ranking*. *Indexing* an image in TagIR involves two complementary tasks: (i) determining the set of tags best describing the image [1, 13], and (ii) quantifying how accurately each of these tags describes the visual content of the image (also known as *tag relevance* [4]). Specifically, Li *et al.* in [4] proposed to learn tag relevance by visual nearest neighbor voting. Since then several works have been proposed to refine tag relevance computation [2, 5]. Neighbor-voting based tag relevance has also been used in re-ranking the tags of a tagged image [12]. In contrast, we explore techniques to measure tag relevance based not only on visual content they represent but also historical relevance of the tagged image. *Scoring* aims to compute a *relevance score* between a keyword query and a tagged image. Relevance score computation typically considers the matching score between the keyword and the image tags as well as other factors derived from search logs and user click-through data [3, 8]. In contrast to these efforts, in our study we compute a *historical relevance score* between a query and a tagged image that considers the historical relevance of the image. Although, the default image *ranking* is based on the relevance score [4], more sophisticated methods diversify search results [10]. Our work is orthogonal to them as we rank images based on its historical relevance.

3 HISTORICAL RELEVANCE-BASED RANKING

Given a social image collection $\mathcal{S}$ and the set of all tags $\mathcal{T}$ in $\mathcal{S}$, each image $s \in \mathcal{S}$ is user-annotated by a set of tags $T_s \subseteq \mathcal{T}$. The set of images annotated by tag t is denoted as $S(t)$. Given $T \subseteq \mathcal{T}$, we denote the set of images annotated by all tags in T as $S(T)$.

Given a search query Q composed of n query tags $t_1, \ldots, t_n$, let $R(Q, \mathcal{S})$ (or simply $R(Q)$ when the context is clear) denote the image search result list of Q on $\mathcal{S}$. For simplicity, we shall assume $t_i \in \mathcal{T}, \forall (i = 1, 2, \ldots, n)$. Each result image $s \in R(Q)$ is annotated by query tags $t_1, \ldots, t_n$ (*i.e.,* $Q \subseteq T_s$) and is associated with a *relevance score*, denoted as $rel(s, Q)$. $R(Q)$ is assumed to be listed by descending order of the relevance scores. By abusing the notation of lists, we denote the list of images in $R(Q)$ as the image set $S(Q)$. We use $R_m(Q)$ or $S_m(Q)$ interchangeably to denote the top-m images in $R(Q)$. Given a search query Q, in this paper, we present a technique to compute for each result image $s \in R(Q)$ a *historical relevance score*, denoted as $rel_h(s, Q)$, and we rank the images in descending order of $rel_h(s, Q)$.

Historical Tag Affinity. We determine the *historical relevance* of a social image in query results as follows. First, we precompute the *historical tag affinity* of all tags in $\mathcal{T}$ offline. Next, we leverage on this score to compute $rel_h(s, Q)$ at the query time. We elaborate on the first step here.

Observe that social images platforms do not classify a concept or tag according to their historical relevance. Furthermore, a tag (*e.g.,* egypt) may not necessarily co-occur with a tag "history" or any other tag denoting the concept of history. Hence, the underlying tag collection $\mathcal{T}$ in $\mathcal{S}$ and traditional tag co-occurrence techniques cannot be leveraged for computing the historical relevance of a tag $t \in \mathcal{T}$. In order to address this challenge, we resort to the external source, *Wikipedia*, to quantify the historical relevance of t. Specifically, the term "history" in Wikipedia matches the disambiguated topic of *History* (http://en.wikipedia.org/wiki/History_(disambiguation)). Hence, we shall use it to measure the *strength* of relationship between t and "history" to quantify the former's historical relevance. For example, when the "history" term is compared with the term "egypt", the common connective articles between them include articles on *"Ancient history"* and *"History of the World"*, which shows that "egypt" is related to history.

We use the notion of *historical tag affinity*, denoted as $a_h(t)$, to measure the strength of the relationship of a tag t with history. Specifically, we leverage the *Wikipedia Link Measure* (WLM), an efficient and accurate technique to measure the similarity between two Wikipedia articles using hyperlinks [7], towards our goal. Computing $a_h(t)$ using WLM consists of two steps namely, *disambiguation* and *article similarity* computation. *Disambiguation* corresponds to the mapping process from a term to the corresponding article. *Article similarity* between two articles x and y, denoted as $docSim(x, y)$, is computed using WLM as follows:

$$docSim(x, y) = 1 - \frac{\log(max(|X|, |Y|)) - \log(|X \cap Y|)}{\log(|W|) - \log(min(|X|, |Y|))} \quad (1)$$

where X and Y are the sets of all articles that link to x and y, respectively, and W is the entire Wikipedia. The intuition behind the above formula is that two related articles tend to be referred to by lots of common articles. Hence, $a_h(t)$ of a tag t and the term "history" is computed as the article similarity of their matching articles. That is, $a_h(t) = docSim(t, \text{"}history\text{"})$. For example, consider the tags `cairo` and `alabama`. The historical tag affinity of these tags are 0.43 and 0.28 as $docSim(cairo, history) = 0.43$ and $docSim(alabama, history) = 0.28$. Intuitively, `cairo` has stronger historical relevance than `alabama`. Note that if t does not exist in Wikipedia, then $a_h(t) = 0$.

Historical Relevance of Images. Observe that the historical tag affinity $a_h(t)$ of a tag t as computed above is not influenced by the relevance of the tag to an image. However, a tag t may annotate more than one image in $\mathcal{S}$ and its relevance in describing the visual content of an image may vary across all images it is associated with. Furthermore, an image may have tags with high historical tag affinity values but they may be irrelevant to the query. Such an image should not be returned as it does not answer the user's query. Hence, we cannot simply use $a_h(t)$ as a proxy to measure *historical relevance* of an image in a query result. Intuitively, given a query Q, the *historical relevance* of an image s in $R(Q)$ is influenced by two key factors, namely, the *relevance* of s to Q and the *historical affinity* of s. We use the notion of *historical relevance score* to measure the *historical relevance* of an image w.r.t a query. In our approach, an image is considered to have high *historical relevance score* with respect to a query Q if it is highly relevant to Q and it has a high *historical affinity* score. The *historical relevance score* of an image $s \in R(Q)$ is then represented as follows.

$$rel_h(s, Q) = rel(s, Q) \times \mathbb{A}_h(s) \quad (2)$$

In the above equation, $rel(s, Q)$ is the *relevance* of the image s to query Q in $R(Q)$ and $\mathbb{A}_h(s)$ denotes the *historical affinity* of an image s. In the literature, several techniques have been proposed to compute the relevance of an image to a query [3, 8, 9]. In this paper, we adopt the best performing configuration in [9]. We now describe our approach to compute $\mathbb{A}_h(s)$.

Intuitively, *historical affinity* of an image s, denoted as $\mathbb{A}_h(s)$, is a *weighted* aggregation of the historical tag affinity values of its *candidate tags*. Note that tags associated with s in $\mathcal{S}$ are not listed according to their *relatedness* in describing the visual content of s. Hence, for each image, we extract its top-k related tags ordered by *tag relatedness* (computed using the neighbor voting scheme [4]) where $k = max(\alpha, \lceil \rho * |T_s| \rceil)$. Both α and ρ are configurable in our approach ($\rho = 0.1$ and $\alpha = 5$ by default). That is, for each image, we want to consider a reasonably small set of tags that best visually describe the image in order to compute the historical affinity of the image. The main reason is that tags are noisy in nature and as a result many of them do not effectively describe the images. On the other hand, some images may not be well tagged and $|T_s|$ can be very small; α is introduced to avoid a very small k. The extracted tags are then considered as *candidate tags* for computing $\mathbb{A}_h(s)$.

Observe that if the candidate tags have high aggregated historical tag affinity values, then it is highly likely that the image has high historical relevance. Note that a tag may have high $a_h(t)$ but if it is lowly ranked in the candidate tag list then its influence is discounted as it may not be very relevant to the image content. Hence, the *historical affinity* of an image s is computed as follows.

$$\mathbb{A}_h(s) = \frac{1}{k} \sum_{i=1}^{k} w(t_i) \times a_h(t_i) \quad (3)$$

As discussed above, Equation 1 is used to compute $a_h(t_i)$. The *weight* $w(t_i)$ is used to promote historical affinity of top-ranked tags that describe the contents of the image well and to discount the impact of the historical tag affinities of lower ranked tags. We set $w(t_i)$ as follows. If $i = 1$ then $w_i = 1$. Otherwise, $w_i = \frac{\tau(t_i)}{\log_2 i}$ where $\tau(t_i)$ denotes the tag relatedness of t_i.

Implementation Overview. The proposed technique is implemented in Java 1.7 using Lucene 3.0.3 as the underlying index engine and MySQL as the database for storing the image collection. We precompute the historical tag affinity values of all tags in $\mathcal{S}$. Furthermore, the tag relatedness between an image and any of its annotated tags is computed offline using the neighbor voting scheme [4] and stored. Given a query Q containing one or more query keywords, we retrieve images that are annotated with Q to form the initial search results $S(Q)$. Any superior TagIR algorithm can be adopted to achieve this goal. In this paper, we adopt the best performing framework in [9][1]. Note that this framework returns the images in $S(Q)$ in descending order of their relevance scores $rel(s, Q)$. The images in $S(Q)$ are then indexed in an in-memory Lucene index for further processing. This index is used to rank the images according to their historical relevance scores (computed using Equation 2) and $S_m(Q)$ is returned (m is user-defined).

4 EXPERIMENTS

We present the performance of our historical relevance-based ranking scheme (denoted by HRR) and compare it with [2], a state-of-the-art relevance-based ranking strategy (denoted by RR). All experiments are conducted on an Intel Xeon X5570 machine with 12GB memory.

Dataset. Since off-the-shelf image search engines (*e.g., Flickr*) typically disallow full access to their data and some statistics (*e.g.,* relevance scores) which are required by HRR and RR, we cannot evaluate our approach directly on top of such search engines. Hence, we are confined to conduct experiments on smaller, constrained collection. In particular, we use the NUS-WIDE dataset containing 269,648 images from *Flickr* [1]. The underlying TagIR system used in our experiments follows the best performing configuration in [9] for multi-tag queries. We used the English Wikipedia dump released on 30 January, 2010 containing 3.2 million articles and more than 266 million hyperlinks.

Query Set and User Study. We invited 14 unpaid volunteers (undergraduate and graduate students in computer science, business, and history majors) to rate the results of 10 queries[2]. To avoid any bias on the evaluation, all the participants were selected such that they did not have any knowledge about the proposed technique[3]. Results generated by HRR and RR are presented to the subjects but without the names of the technique producing the

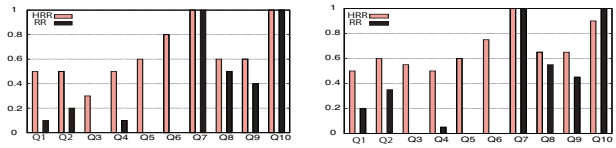
[1] Note that the image retrieval step is orthogonal to the problem addressed in this paper.

[2] Since there is no benchmark query set for this problem, queries were selected based on their association to historical objects, locations, scenes, and events.

[3] None of the volunteers are authors of this paper.

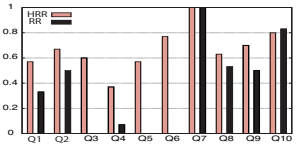
Table 1: Sample queries.

Id	Query	$S_m(Q_{HRR_i}) \cap S_m(Q_{RR_i})$ Top-10	Top-20	Top-30
Q_1	architecture	0	0	0
Q_2	buddha	3	8	11
Q_3	people	0	0	0
Q_4	general	1	3	7
Q_5	car	0	0	0
Q_6	apple	0	0	0
Q_7	pyramid	3	9	22
Q_8	architecture, history	1	2	4
Q_9	tower, history	5	10	20
Q_{10}	rome, architecture	2	7	14

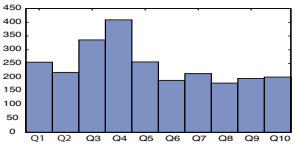


(a) P@10

(b) P@20



(c) P@30



(d) Execution time (msec.)

Figure 2: Performance results [Best viewed in color].

results. Each participant was given one query at a time in random order (all queries). Then, for each pair of query and its results, we asked the volunteers to label which result images are relevant to history in RR and HRR.

Performance Metric. We conducted experiments to evaluate the *effectiveness* and *efficiency* of the two schemes. The *effectiveness* is measured by *Precision@K* (*P@m*), which is the ratio of the relevant images among the top-*m* retrieved images for a query. Since a keyword query may match a large number of images and a user is unlikely to sift through all of them, we believe that *P@m* better reflects a user's perception about a TagIR system in our setting. To the best of our knowledge, this is the first work to rank images in TagIR based on historical relevance. Hence, there is a lack of benchmark dataset for the evaluation. The labels annotated by the volunteers are used as ground-truth labels.

Experimental Results. First, we investigate the number of common images in top-*m* results of a query Q_i in HRR and RR schemes (*i.e.,* $S_m(Q_{HRR_i}) \cap S_m(Q_{RR_i})$). It is reported in the last column of Table 1. Observe that it is rather small for many queries. *This shows that the top-ranked images produced by HRR scheme are different from those produced by traditional RR.*

Second, we study the effectiveness of HRR compared to RR. Figures 2(a)-(c) plot the average *P@m* ($m \in \{10, 20, 30\}$) of HRR and RR for each query (average of the 14 volunteers). Note that RR produces no historically relevant images for queries Q_5 and Q_6 ($P@m = 0$). Among the 10 sampled queries, almost all underwent increase in *P@m* under HRR scheme. For instance, in Q_2 HRR returns the statues of buddha, old buddhist temples and shrines among the top-ranked results, whereas RR includes the images of buddhist monks and children in the result set. In Q_3 all the top-ranked images returned by RR are faces of people. In contrast, HRR returns several images depicting people carrying banners in historic protest marches. Similarly, in Q_4, HRR returns several images of army generals during world wars (*e.g.,* Hitler), whereas RR failed to return such images. Instead, the latter retrieves images related to Nazi flag, museums, etc (note that although they have high historical relevance, they are not relevant to Q_4). Interestingly, HRR retrieves many images of old or vintage cars as top-ranked results for Q_5 whereas all images returned by RR are racing cars. In Q_6, buildings in NYC and Amsterdam along with image of old macintosh computer are returned by HRR. In contrast, RR retrieved images of fruits or *Apple* products such as computers, iphones, and ipads. Observe that although HRR performs better than RR for Q_8 and Q_9, the difference is not significant as these queries retrieve images annotated with the tag `"history"`. Hence, many of them have historical significance. However, as remarked in Section 1, such queries fail to retrieve historically-relevant images that are *not* tagged with this keyword. Lastly, observe that Q_7 and Q_{10} perform well for both HRR and RR. This is because a large number of images tagged with these keywords are strongly related to history. Specifically, the tag `pyramid` is primarily used for images related to Egyptian, Mesoamerican, and Louvre pyramids. Similarly, there are many images depicting historic architectures of Rome.

Finally, we analyze the execution time. Figure 2(d) reports the average execution times of HRR for $m = 30$. Observe that it is cognitively negligible (150-400 msec).

5 CONCLUSIONS & FUTURE WORK

We have proposed a novel vision to rank query results in a TagIR framework based on their historical relevance. In contrast to traditional relevance-based ranking techniques, our approach ranks images not only based on the relevance of its visual content to the query but also its *relatedness to human history*. Specifically, we introduced an unsupervised, learning-agnostic, Wikipedia-based approach to quantify the historical relevance scores of images and rank them accordingly. Our empirical study demonstrates the effectiveness of the proposed ranking scheme. It is not difficult to observe that the proposed Wikipedia-driven solution can be easily extended to other topics where traditional relevance score-based ranking is ineffective. As future work, we intend to leverage text mining-based techniques to further improve the effectiveness.

REFERENCES

[1] T-S. Chua, et al. NUS-WIDE: A Real-world Web Image Database from National University of Singapore. In *ACM CIVR*, 2009.
[2] Y. Gao, M. Wang, et al. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. In *IEEE Trans. on Image Processing*, 22(1), 2013.
[3] V. Jain, M. Varma. Learning to Re-rank: Query-dependent Image Re-ranking Using Click Data. In *WWW*, 2011.
[4] X. Li, et al. Learning Social Tag Relevance by Neighbor Voting. *IEEE Trans. Multimedia*, 11(7), 2009.
[5] X. Li, et al. Unsupervised Multi-feature Tag Relevance Learning for Social Image Retrieval. In *CIVR*, 2010.
[6] Z. Liu, et al. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *CVPR*, 2016.
[7] D. Milne, I. A. Witten. An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *Proc. AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008.
[8] G. Smith, et al. Evaluating Implicit Judgments from Image Search Clickthrough Data. *JASIST*, 63(12), 2012.
[9] A. Sun, S. S. Bhowmick, et al. Tag-based Social Image Retrieval: An Empirical Evaluation. *JASIST*, 62(12), 2011.
[10] R. van Leuken, et al. Visual Diversification of Image Search Results. *WWW*, 2009.
[11] L. Wu, Y. Wang, and J. Shepherd. Efficient Image and Tag Co-ranking: A Bregman Divergence Optimization Method. In *ACM MM*, 2013.
[12] J. Xiao, et al. Exploring Tag Relevance for Image Tag Re-ranking. In *SIGIR*, 2012.
[13] G. Zhu, et al. Image Tag Refinement Towards Low-rank, Content-tag Prior and Error Sparsity. In *ACM MM*, 2010.