Search-Based Relevance Association with Auxiliary Contextual Cues

paper ID: 6

ABSTRACT

In this work, we target at solving the Bing challenge provided by Microsoft. The task is to design an effective and efficient measurement of query terms in describing the images (image-query pairs) crawled from the web. We observe that the provided image-query pairs (e.g., text-based image retrieval results) are usually related to their surrounding text; however, the relationship between image content seems to be ignored. Hence, we attempt to integrate the visual information via traditional retrieval-based method and similarity propagation model for better ranking results. In addition, we found that plenty of query terms are related to people (e.g., celebrity) and user might have similar queries (click logs) in the search engine. Therefore, in this work, we propose a relevance association by investigating the effectiveness of different auxiliary contextual cues (i.e., face, click logs, visual similarity). Experimental results demonstrate the effectiveness of our methods. Finally, due to the consideration of efficiency, we primarily adopt visual similarities, i.e., Modified PageRank Model (MPM) as our final model to compete for the final prize.

1. INTRODUCTION

Aiming at the Bing Grand Challenge, in this work, we are dealing with the relevance of each image-query pair. That is, given a pair of tag and image (cf. Figure 2), we need to measure the relevance of the image according to the tag. And the database contains 23 millions of tag-image pairs with their corresponding click number log. Both the queries and database are real data from Microsoft. This challenge might be related to visual re-ranking [2][3][5][8] or image search result refinement [6]; hence, it is essential to bridge the semantic gap between visual and textual information [7]. In addition, the response time (latency) of the proposed method should be take into account for online image-query relevance measurement. Hence, in this work, we investigate the effectiveness of different auxiliary contextual cues which seldom be considered in the prior work.

To deal with the large scale real data, we adopt three major strategies. First, the search-based system consists of two parts, images search and tag similarity association. For each image-query pair, similar images are retrieved based on visual content (tag). Then the relevance is measured by the tag (visual) similarities (cf. Figure 1(b)). Further, we observe that a significant portion of the image-query pairs in Bing dataset comprise faces in images and names in their tags (as an example image-query pair in Figure 3). This phenomenon is obvious via the statistics of name detection in tags and face detection in images. If we only look at the 1,000 unique tags in development set, around 318 unique tags consist of at least one name (either given name or family name), totally 31.8% over the whole unique tags.[1] According to the statistics, we believe the correspondence between face in image content and name in the associated tag

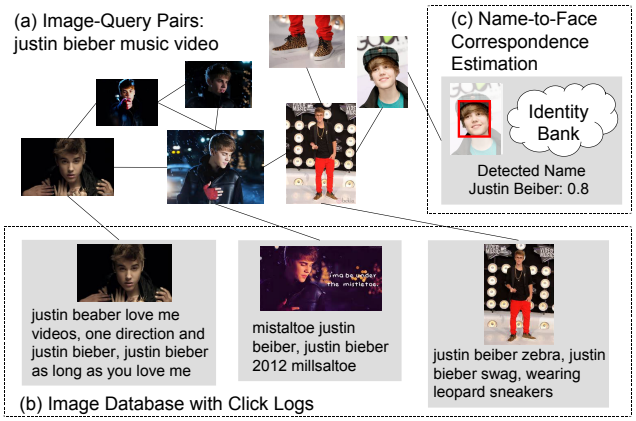

Figure 1: The proposed search-based relevance association with auxiliary contextual cues. We observe that the provided image-query pairs are usually related to the query tag; however, the relationship between image content seems to be ignored. Therefore, to tackle this challenge, we propose to measure the relevance via auxiliary contextual cues (i.e., face, click logs, visual similarity).

will benefit measuring image-query relevance. Therefore, as Figure 1(c) shows, once an image-query pair comprises at least a face and a name, the relevance of this query will be measured by the proposed name-to-face correspondence estimation.

Third, as Figure 2 shows, another observation is that the provided image-query pairs might be the text-based search results. That is, relevant (users' desired) images might be the majority corresponding to the tag. Further, the desired images usually have similar visual content in the image-query pairs (cf. Figure 4). As a result, we find the classical PageRank algorithm [11] may solve the problem if we have the full list of images corresponding to the tag. As Figure 1(a) shows, the graph is constructed for each query. And each image in the list is a node and the weights of the edges are set by the visual similarities. However, in this challenge, we have to reply the relevance score once receiving each image-query pair, which means we do not have full images corresponding to a tag. To solve this, we propose two modified PageRank methods which does not need full image set.

In this report, we will show the integrated results from each model and provide extensive experiments to support our methods and arguments under development set. The experimental results demonstrate the effectiveness of our proposed methods, which achieve 14.5% and 45.7% relative improvement in all queries and people-related queries in the development set, respectively.[2] On the other hand, in the final online contest, due to efficiency consideration, we

[1]For celebrity, there exists 8.98% image-query pairs, and 86.20% of them contain at least one face in image content.

[2]Note that the preliminary results were presented in [14]. We extend the original method to a modified version for tackling the online challenge.

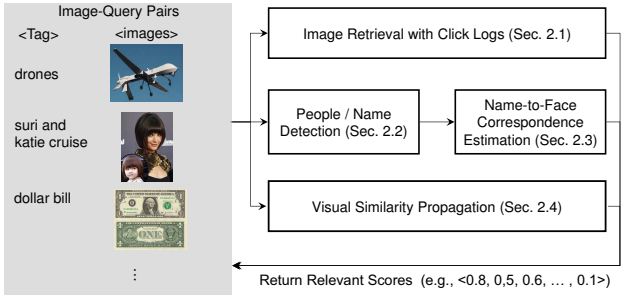

Figure 2: A system diagram of the proposed method. When an image-query pair comes in, we integrate three different approaches to measure the relevant scores for each image-query pair.

primarily adopt the modified PageRank methods (MPM), which have the best experimental results.

2. RELEVANCE ASSOCIATION WITH AUXILIARY CONTEXTUAL CUES

Our system flow is shown in Figure 2. Image-query pairs are first fed into three different scoring methods, which are content-based method (Sec. 2.1), human-based method (Sec. 2.2 and 2.3) and query-association method (Sec. 2.4). The outputted scores are fused together as the final relevant scores.

2.1 Image Retrieval with Click Logs (IRCL)

The most intuitive way to measure the relevance score of an image-query pair is to retrieve visually and textually similar images from database (cf. Figure 1(b)). Therefore, it is essential to effectively retrieve similar images by applying content- (IRCLV) or text-based image retrieval (IRCLT).

1) Content-based image retrieval (IRCLV). To efficiently retrieve similar images from large-scale database, we apply the state-of-the-art bag-of-words (BoW) model with 1M vocabulary [12]. We first extract the visual features for the query image, and retrieve similar database images under the inverted indexing structure. Based on the top-ranked visually similar images, we can calculate the tag similarity between the query tag and click logs. To avoid misspelling (e.g., "fridsy") and synonym issues, we apply the standard pre-precessing before calculating the similarity.

2) Query tag expansion from hash tags in Twitter (IRCLT). For text-based image retrieval, we further apply tag expansion for query and tag clustering for database to ensure the quality of the top-ranked images. For tags in database images, we apply affinity propagation [1] for better representation. For each query tag, we first retrieve similar tweets from Twitter and then extract top frequently appeared hash tags as the query expansion results. We found that the hash tags are succinct with less noise than other methods such as snippet from search engine, and created by the users so its more close the what users really expect. Based on the expansion results, we can efficiently retrieve textually similar images. Eventually, we combine the textual and visual similarity as the final relevant scores.

2.2 People/Name Detection

As mentioned in Sec. 1, a significant portion of image-query pairs comprise names; hence, it is essential to detect whether the image-query pair contains a name. We propose two name detection methods as follows:

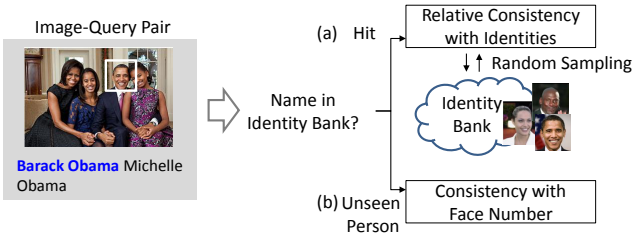

Figure 3: The framework of name-to-face correspondence estimation. Given an image-query pair, if the name in the tag is included in Identity Bank, the relevance will be measured by its corresponding identity model and its ranking in the results of randomly samlped models (as (a)). We also measure the consitency between the number of faces in image content and the number of names in its tag, which can deal with persons without training data (as (b)).

1) Name list method (NLM). We first collect a list of celebrity names from the web,[3] which is denoted as $\mathbb{L}$ later. For each multiple-word celebrity name in $\mathbb{L}$, the first and the last words of the name are added into the $FirstNameSet$ and $LastNameSet$ respectively. After constructing these information, for each new coming tag, if it contains a name in the name list $\mathbb{L}$, we said a name is detected. However, unseen names cannot be handled by using this method only. For that reason, an additional method is applied to enhance the recall. That is, a detection is also claimed if we can find two consecutive words, in which the first and the second words are in the $FirstNameSet$ and $LastNameSet$ respectively.

2) Social media method. After using tag snippets returned by the search engine as query expansion, an interesting observation has been found. If a query tag contains names, it has higher probability that social media website can be found in the title of top-10 related website, such as Linkedin, Twitter and Facebook.

2.3 Name-to-Face Correspondence Estimation

This section includes how we train visual-based face identity models - *Identity Bank* - to discover name-to-face correspondences and how we measure the consistency between the number of faces and the number of names for finding unseen person's name-to-face correspondence (Figure 3(b)).

2.3.1 *Relative Consistency with Identity Bank (IB)*

Firstly, we address how Identity Bank is constructed and used for name-to-face correspondence measurement. Given a set of image-query pairs, we leverage face detection [10] and name detection (cf. Sec. 2.2) to collect training face images and their name annotations. Totally, Identity Bank consists of 6,762 identity models trained by 35,092 face images from Bing training dataset. Further, we randomly sample more models from Identity Bank for testing and use the multiple test results to confirm the confidence of correspondence measurement. As an image-query pair example in Figure 3, we will choose the model of "Barack Obama" for measuring the name-to-face correspondence probability. Meanwhile, we randomly select another two models, e.g., "Angelina Jolie" and "Michael Jordan" and obtain their test results as background model. The correspondence probability of "Barack Obama" will be further adjusted by its rank within the three test results. Higher confidence scores will

[3]http://www.posh24.com/celebrities/

Figure 4: Images expected by the users usually have similar visual content but are not directly the same as the semantic of the tag. "Waldo Alabama" is often not referred to the town but the covered bridge famous in the town.

be derived if the target name is more salient than the background ones.

2.3.2 *Consistency with Face Number (FN)*

Further, it is critical to deal with unseen persons for a online system because it should adapt to incrementally updated data. From our statistics in Bing development set, around 29.96% image-query pairs are with names (persons) never shown in any training image and thus without corresponding models in Identity Bank. For these unseen persons, it is difficult to train models due to the lack of training images. We propose to measure the consistency by matching the number of faces in image content and the number of names detected in its tag (cf. Sec. 2.2). In this scenario, no training images are required such that it is applicable to the unseen persons out of Identity Bank. If a name is detected in a tag, the associated image should comprise at least one face. We also consider the word "and" after a name, for example, Justin Bieber and his mother. In this case, the number of faces should greater than one though only one name is detected. The image-query pairs following the two rules will be given higher relevance; otherwise, the relevance will be decreased.

2.4 Visual Similarity Propagation

We observe the given image-query pairs in this contest might come from text-based image retrieval (TBIR) results or from user's click through. Thus, we assume that the majority of retrieved images are relevant to the query.[4] Furthermore, relevance depends on what users really want, which usually focuses on a few famous features directly/indirectly regarding to the tag. Hence, the image-query pairs often contain similar visual content, frequently grouping into some clusters as shown in Figure 4. Although "Waldo Alabama" is a town in the U.S., what most users really like to see from tag is the famous covered bridge in Alabama. Thus, the relevant images often contain certain famous covered bridge in Alabama and form some clusters in the images list. As a result, if an image in the list is relevant to the tag, there should be some other images with strong visual similarities. In short, those having strong visual connections with others deserve more relevance scores.

As a result, it is reasonable to assign the relevance score to an image based on its connections (similarities) with others. The stronger the connections are, the higher the score should be. The concept is highly resembles PageRank based on the connection of hyperlinks. Thus, if we have the full list of images corresponding to a tag, the relevance scores ($\mathbf{s}$) could be formulated as $\mathbf{s} = (\alpha\mathbf{P} + (1-\alpha)\mathbf{v}\mathbb{1}^{\mathbf{T}})\mathbf{s}$, where

[4]Note that the assumption can be extended to real case by using TBIR to retrieval some potential candidates.

Table 1: The performance of content-based method (IRCL) and name-to-face correspondence measurement. The proposed method (IB+FN) can reach 45.71% relative improvement for the name of queries appeared in Identity Bank (I), The results confirm the benefits from the proposed name-to-face correspondence estimation for relevance measurement.

Full queries	Initial	IRCLV (+FN)		IRCLT	Ideal
DCG@25 (A)	0.469	0.484 (0.489)		0.495	0.684
Name queries	Initial	IB	FN	IB+FN	Ideal
DCG@25 (N)	0.481	0.496	0.508	0.516	0.702
DCG@25 (I)	0.350	0.496	0.500	0.510	0.672

$\mathbb{1}$ is a vector of one, and $\mathbf{v}$ is encoded with initial normalized prior ($\mathbb{1}$) or prior knowledge from previous sections. Further, the transition probability ($\mathbf{P}$) is equivalent to the normalized similarity between image-query pairs (i.e., $P(i,j) = sim(i,j)/\sum_i sim(i,j)$).

However, due to the limitation of latency, we have to reply the relevance score for each image-query pair in limited time latency to the remote query server. That is, it is impossible and impractical to wait for all the images corresponding to a tag, and then apply PageRank to assign the relevance scores. Thus, in this challenge, we inherit the spirit of PageRank and modify the formulations to fit our purpose. That is, we score each image-query pair based on all previous information. The main idea of RageRank is that the one with higher similarities to other images of the same tag should have higher relevance score. To approximate this property, we keep all image-query pairs and assign the score by averaging its similarities to other images of the same tag recorded. Thus, the relevance scores can be formulated as $\mathbf{s_i} = (\mathbf{1/N_i}) \sum_j \mathbf{sim}(i,j)$, where j indicates the images with the same tag recorded earlier, and N_i is the number of images with the same tag shown before $i-th$ query pair. We name method as modified PageRank model (MPM).

To further avoid the scores being influenced by noisy images, only the top-5 most similar images are considered (top-5 MPM). That is, if the number of the recorded images with the same tag is greater than 5, we just average the top-5 similarities of the recorded images. By doing this, better scores for the images can be achieved. The experimental results show that the above methods to approximate the PageRank algorithm can have competitive results with limited performance degradation in Sec. 3.4.

3. EXPERIMENTAL SETUP AND RESULTS

3.1 Experimental Setting

Dataset. We use the entire dataset provided by Bing challenge, which contains 1M images with 23M click logs. In addition, they also provide a development set that contains 1,000 query terms with around 80K images. To tackle this challenge, we extract different kinds of visual features—bag-of-words (BoW) [13], the vector of locally aggregated descriptors (VLAD) [4], grid color moment (GCM), and local binary pattern (LBP) [9] for facial images.

Evaluation. We evaluate the relevant scores of the same query terms by Discounted Cumulated Gain at 25 (DCG@25) as defined by the challenge. The averaged DCG@25 of the original ranking (Initial) and the ideal ranking is 0.469 and 0.684, respectively. Because some queries are less than 25 excellent images, the averaged DCG@25 will not be 1.

Table 2: The performance of query-association method on three different features. Top-5 MPM has 14.5% relative improvement compared to the original ranking (0.469) in the development dataset.

Method	Average Similarity			PageRank (PR)			PR with Prior	PR with Late Fusion	MPM	top-5 MPM
Feature	GCM	VLAD	BoW	GCM	VLAD	BoW	BoW	BoW	BoW	BoW
DCG@25 (A)	0.492	0.538	0.541	0.490	0.527	0.543	0.543	0.544	0.528	0.537

3.2 Performance of IRCL

We first evaluate the performance of the entire queries (A) by content-based method that considers both visual and tag similarities. As the first row of Table 1 shows, the proposed IRCLV (0.484) is better than the initial ranking (0.469). This is because users might have similar types of queries (e.g., poster, logo) in the search engine. Therefore, by considering both visual and tag similarities, we can obtain partial knowledge from the provided database. In addition, as mentioned in Sec. 2.3, the query might contain the name so we should also consider the number of faces in images. Hence, we can further improve the accuracy to 0.489 (IRCLV+FN). Then we expand query tags by Twitter. Each query tag is viewed as a hash tag of Twitter. We use the hash tag of Twitter to find related tags to the corresponding query tag. In our experiments, this approach (IRCLT) can improve the accuracy to 0.495. Nevertheless, the accuracy is still far from the ideal case (0.684) because most of queries are quite different from the database. Thus, the relationship between the query images will be considered in Sec. 3.4.

3.3 Name-to-Face Correspondence Estimation

To evaluate the effect from name-to-face correspondence estimation, we conduct the experiments on the test queries with at least one name detected (dataset (N)), totally 7,180 image-query pairs, which corresponds to around 9% of the whole test queries in Bing dataset. The baseline for comparison is assigning each query with the same relevance. As shown in Table 1 (N), relevance measurement by Identity Bank (IB) and by Face Number (FN) report 1.5% improvement and 2.7% improvement compared with the baseline (0.481), respectively. In addition, their late fusion (0.516) surpasses the baseline with 3.5% improvement. We also conduct the experiments on the test queries with names included in Identity Bank (I), that is, the identities with training data. As shown in Table 1 (I), these test queries are more challenging because the DCG@25 in Ideal case (0.672) is lower than that for dataset (N) (0.702) and the result of baseline decreases to 0.350. In these challenging queries, the result of proposed method (IB+FN), 0.510, substantially outperforms the result of baseline with 16% improvement, that is, 45.71% relative improvement. In summary, the proposed name-to-face correspondence measurement can deal with celebrities (with training images) as well as unseen persons (without training images) and obviously improves image-tag relevance measurement in the queries with names.

3.4 Visual Similarity Propagation

Table 2 shows the experiments on visual propagation methods (Sec. 2.4). First, we conduct experiment with full list of images corresponding to the same tag, and measure the relevance based on the fully average similarity (i.e., $s = \sum_j sim(i,j)$, where j is all other images with the same tag as i) and PageRank (PR) Model. The experiment manifests our observations in Sec. 2.4. That is, more similar images the one has, higher relevance score the one deserves. PageRank can further improve the DCG@25 to 0.543.

Then, the experiments (MPM, top-5 MPM) that individually reply the relevance score to the remote query server are also simulated. The DCG@25 will be 0.528 and 0.537 respectively. Though slightly degrading, the results are still nice compared to other methods. And the degradations are expectable because we cannot not see the whole images corresponding to the tag. Thus, some information is definitely missed. From Table 2, we can find that fusing the results of other sources only have little influence on the performance of PageRank model. Thus, considering both the effectiveness and efficiency, we only use modified PageRank models in the online system as a tradeoff. ***The performances of MPM and top-5 MPM in the final online contest system are 0.5356 and 0.5371 respectively.***

4. CONCLUSIONS

We propose a search-based relevance association framework to measure the relevance of image-query pairs. By utilizing different auxiliary contextual cues, we propose three methods – content-based, human-based, and query-association – to conquer the challenge. The preliminary experiments manifest the effectiveness of our proposed methods, which successfully uplift the image-query relevance measurement in Bing challenge. Further, the visual similarity propagation gives the best performance within our methods. Therefore, we adopt the visual similarity propagation as our primary method in the final contest.

5. REFERENCES

[1] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 2007.

[2] W. H. Hsu et al. Video search reranking via information bottleneck principle. ACM MM, 2006.

[3] W. H. Hsu et al. Video search reranking through random walk over document-level context graph. ACM MM, 2007.

[4] H. Jégou et al. Aggregating local image descriptors into compact codes. *IEEE TPAMI*, Sept. 2012.

[5] W. Liu et al. Noise resistant graph ranking for improved web image search. In *CVPR*, 2011.

[6] J. Lu et al. Image search results refinement via outlier detection using deep contexts. In *CVPR*, 2012.

[7] H. Ma et al. Bridging the semantic gap between image contents and tags. *IEEE TMM*, 12(5):462–473, 2010.

[8] N. Morioka and J. Wang. Robust visual reranking via sparsity and ranking constraints. In *ACM MM*, 2011.

[9] T. Ojala et al. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002.

[10] Omron. Okao vision. 2008. `http://www.omron.com/r_d/vision/01.html`.

[11] L. Page et al. The pagerank citation ranking: Bringing order to the web. Technical report, 1999.

[12] J. Philbin et al. Object retrieval with large vocabularies and fast spatial matching. In *CVPR07*.

[13] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[14] C.-C. Wu et al. Search-based relevance association with auxiliary contextual cues. In *ACM MM*, 2013.