PREDICTING THE EFFECTIVENESS OF QUERIES FOR VISUAL SEARCH

Bing Li[†‡] *Ling-Yu Duan*[‡⋆] *Yiming Chen* [‡] [‡]*Rongrong Ji*[‡] *Wen Gao*[‡]

[†]The Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[‡]The Institute of Digital Media, School of EE & CS, Peking University, Beijing, 100871, China
{libing, lingyu, ymchen, rrji, wgao}@pku.edu.cn

ABSTRACT

Poor retrieval performance significantly degenerates users' experience of visual search, especially in mobile search. Ideally, users would like to be alerted when bad queries are present, which helps eliminate latency as well as waste of bandwidth, especially in 3G wireless environment. In this paper, we propose a visual query performance prediction (v-QPP) approach to predict the retrieval effectiveness. We employ latent dirichlet allocation (LDA)to derive latent topics from image database. From the collection statistics, we model the query's specificity based on topics. High specificity helps a retrieval system to derive user's search intent exactly. Moreover, as low discriminative content is difficult to search in terms of distinguishing relevant images from irrelevant one, we propose a topics based inverse concept frequency (t-ICF) model to deal with specific queries but difficult to discriminate in the reference database. Comparison experiments over MPEG CDVS benchmarking datasets have shown our method significantly outperforms existing approaches in document retrieval.

***Index Terms*—** query performance prediction, visual search, mobile, topic model

1. INTRODUCTION

With the proliferation of camera embedded mobile devices, visual search have received a wide range of attentions from both academia and industry. Huge efforts have worked on improving a particular retrieval system. However, state-of-the-art systems still suffer from poor or unstable performance in visual search. Even a very promising system (i.e., high mean average precision (mAP)) would probably return poor retrieval for a particular query.

Undoubtedly, any worse retrieval may deteriorate user experience. Users definitely concern their current queries' performance rather than mAP. User information needs would be better respected by alerting the unsuitability of the query and asking for refinement(e.g. Fig.1). Especially in mobile search, delivering a query image incurs high latency and costly bandwidth, so that we propose to perform v-QPP to predict the visual retrieval effectiveness.

Query performance prediction (QPP) is important, as evidenced by extensive research activity [3]. Existing QPP methods are divided into two classes: pre-retrieval[4][9][5] and post-retrieval[10]. Post-retrieval is usually superior to pre-retrieval, while pre-retrieval is efficient as none of retrieval process is involved. Targeting mobile search, we focus on pre-retrieval QPP. To the best of our knowledge, our v-QPP is the first to tackle QPP in pure visual search. [6][12][11] attempted QPP methods for image retrieval, but their methods work on pure text (e.g. tags) rather than visual cue.

⋆ Corresponding Author

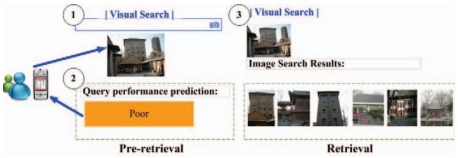


Fig. 1: A v-QPP scenario of mobile users. (1) Users snap a visual query to search; (2) Mobile APP predicts query performance before the actual retrieval; (3) If v-QPP is high, mobile delivers the query to a retrieval system; otherwise, user are alerted of poor performance or query refinement (e.g. adjusting photograph manners).

Predicting the visual query performance is a challenging problem. Firstly, existing QPP methods in document retrieval do not suffice for visual search, due to the unavailability of semantic meaningful visual words. Bag-of-Words (BoW) has been successfully applied to object recognition and scene categorization by considering an image as an analogy to a document consisting of visual words. Unlike textual words, visual words (often quantized from a large collection of local descriptors) are much less semantic meaningful or discriminative. Existing pre-retrieval methods[4][9][5] rely on the statistics of query terms to analyze the specificity expression power of each query term, so that they cannot work well in visual search. Secondly, a visual query is often more sensitive to distractors (e.g., irrelevant foreground or background objects), which would confuse a retrieval system to infer the actual search intent of users. Clearly, it is impractical to request users to take perfect "close-up" of an object or scene. These issues do not exist in document retrieval, as query terms basically reflect the user's search intent. In this paper, we investigate a wide range of queries with high or low performance over diverse datasets. Two practical factors are proposed to model the visual quality performance predictor. Beyond visual words, we introduce LDA to come up with latent semantics for measuring the effects of these two factors on retrieval performance.

Our main contributions are twofold. First, we figure out two practical factors in predicting the visual query performance. Second, we propose a topic model based v-QPP approach to fix the unsuitability of applying most existing QPP methods to visual words.

2. THE V-QPP MODEL

The v-QPP model investigates two essential factors:

(1) Query based information needs specificity (q-INS). q-INS predicts a query to perform better with increased specificity. For example, in Fig.3a, the "Seilbahn" house dominates the image to produce a specific information need, whereas two objects (i.e. tree and buildings) in Fig.3b blurs the information need. As shown in Fig.3, the LDA derived topic distribution may qualitatively evaluate q-INS.

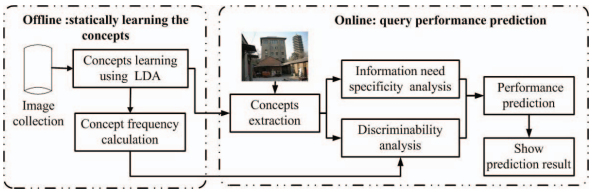


Fig. 2: The framework of proposed query performance prediction.

(2) Collection based discriminability of search content (c-DSC). Even with a high q-INS, a query may perform poor if relevant visual contents are visually similar to irrelevant contents in terms of concepts. A concept corresponds to an object or a part of an object, while different objects could share similar visual words. c-DSC attempts to recover the concurrent statistics of visual words to infer the discriminability of search content over the collection. As we focus on pre-retrieval QPP, c-DSC cares more about whether syntactic or semantic concepts would probably be distinguished in general in terms of visual words, rather than the semantic gap between each individual concept and low-level features.

2.1. Query based information needs specificity (q-INS)

Each visual query may contain multiple objects or different facades of an object. Given a visual query Q, a mix of concepts in query Q are denoted as $C_Q = \{c_k\}$, c_k is the k-th concept. $p(c_k|Q)$ is the proportion of concept c_k . Instead of modeling the concepts from low-level features (i.e. training a concept classifier), we attempt to estimate the potential concepts' distribution in a visual query, measuring q-INS according to the proportions of concepts as shown in Fig.3.

We measure q-INS by the distribution statistics of concepts. Information need is usually determined by selecting those concepts with higher proportions. It is natural to imagine that for a very flat distribution, any retrieval system would be confused as the query cannot express user need exactly. Given a query Q, q-INS is thus defined as follows:

$$q-INS = \log_2 \sqrt{\frac{1}{|C_Q|}\sum_{c_k \in C_Q}(p(c_k|Q) - \frac{1}{|C_Q|}\sum_{c_k \in |C_Q|} p(c_k|Q))^2} \quad |C_Q| > 1 \tag{1}$$

where $|C_Q|$ is the total number of concepts of the query Q. When K equals 1, we set a high $q-INS$ score (say 10,000), as one concept clearly indicates a very specific information need.

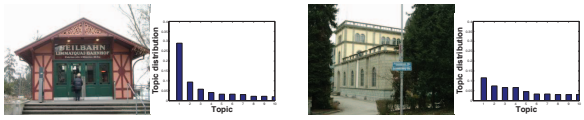
2.2. Collection based discriminability of search content (c-DSC)

c-DSC is measured by the discriminability of query concepts over a collection. We propose an inverse concept frequency model(ICF) [1] to measure the discriminability of a concept. Let cf_k denote the total number of images containing concept c_k in reference database. A higher value cf_k indicates that concept c_k occurs in many documents and c_k is less discriminative, and vice versa.

A query's c-DSC is thus defined as follows:

$$c-DSC(Q) = \sum_{c_k \in C_Q} \log_2 \frac{N}{r_k^2 \cdot cf_k + \mu} \tag{2}$$

[1]Note that ICF is different from inverse document frequency (IDF), as IDF works on each single word while ICF works on concepts (semantic or syntactic, each involving multiple words or low-level features.)



(a) (b)

Fig. 3: Two queries of different q-INS and their LDA based topic distributions, ranked by their proportions.

where N denotes the total image number in database, r_k is the rank (in decreasing order) of concept c_k by its proportion. $\frac{1}{r_k^2}$ is the weight of concept frequency cf_k. μ is used to avoid zero denominator from cf_k. We empirically set $\mu = 0.0000001$.

We have two points to explain c-DSC. First, although a concept does not closely relate to information need, those concepts' concurrent statistics may contribute to retrieval. For example, in landmark search, a little statue close to a landmark may facilitate search, so v-QPP need to consider concurrent concepts. Generally speaking, we should analyze all the concepts for better c-DSC prediction. Second, the concept with higher proportion in a query contributes more to v-QPP prediction. So large weights are assigned to the concept with higher proportion in c-DSC.

2.3. Combining q-INS and c-DSC

As two factors jointly affect v-QPP, we currently apply linear weighting to combine two scores. Given a query Q, the v-QPP score is defined as follows:

$$v-QPP(Q) = \lambda \cdot q-INS(Q) + (1-\lambda) \cdot c-DSC(Q) \tag{3}$$

where λ is defined in the interval [0,1], empirically set λ =0.5. Ideally, the optimal weight depends on the correlation between AP and q-INS/c-DSC, which vary with different datasets.

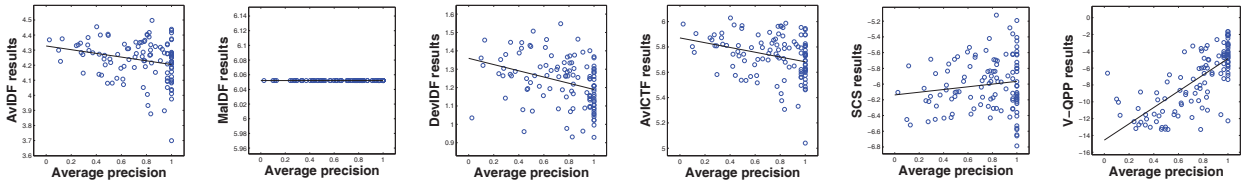
3. LDA BASED V-QPP IMPLEMENTATION

We employ Latent Dirichlet Allocation (LDA)[1] to implement our v-QPP model. In textual analysis, LDA based topic models are widely used to discover the semantic structure based on co-occurrence statistics. Recently, LDA has been successfully applied in scene classification (e.g.[2]) or object discovering (e.g.[7]). The recovered topics are assumed to correspond to an object (e.g. airplane, landmark). To implement v-QPP model, we employ LDA to discover the concepts.

In essence, LDA model learns a generative process over a collection. Given an image consisting of visual words, LDA models the BoW histogram as a mix of multiple topics. Let $W_I = \{w_i\}_{i=1}^{N_I}$ denote the visual words of image W_I, N_I the number of words in image I, z_k is the topics. The LDA probabilistic generative process is given as:

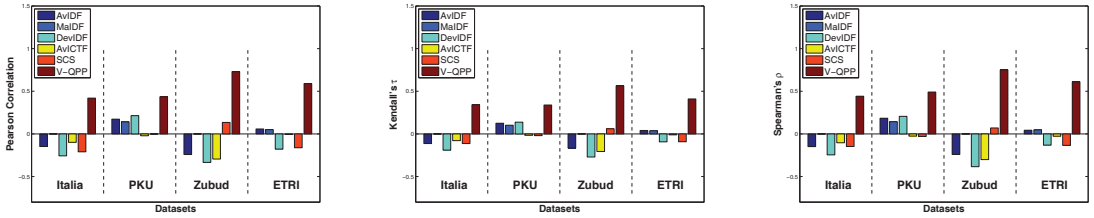
$$p(W_I|\alpha,\beta) = \int p(\theta|\alpha)(\Pi_{i=1}^{N_I}(\sum_{k=1}^{K} p(z_k|\theta)p(w_i|z_k,\beta)))d\theta \tag{4}$$

where α, β are the parameters of dirichlet distribution for generating topic distribution and word distribution, θ is the multinomial distribution of topics. Over the reference database, we sample images to train LDA model. With topic number K, the words distribution over each topic $\{p(w_i|z)\}$ as well as topic distributions over an image $\{p(z_k|I)\}$ can be learnt.



(a) (b) (c) (d) (e) (f)

Fig. 4: Scatter plots of AP values versus QPP results on Zubud dataset by different methods: (a) AvIDF, (b)MaiDF, (c) DevIDF, (d) AvICTF, (e) SCS, (f) our v-QPP. Each point corresponds to a query.



(a) (b) (c)

Fig. 5: Comparison of six QPP Methods on four datasets: (a)Pearson Correlation γ, (b)Kendall's τ, (c)Spearman's ρ

With the learnt topics, we implement q-INS and c-DSC based on learnt topics as shown in Fig.2 (the topics based c-DSC also named as t-ICF.) First, we calculate the concept frequency in the collection. Given an reference image I, we determine the concepts based on the topic distribution. The concept number of an image is assumed to be no more than M.

By the topic proportions, the top M topics $z_{T_1}, \ldots, z_{T_M}$ are selected as the concepts of image I, denoted as C_I=$\{c_{T_1}, \ldots, c_{T_M}\}$. z_{T_j} is the top j topic. The topic frequency $cf(I, c_k)$ of image I relying on the top M topic is defined as:

$$cf(I, c_k) = \begin{cases} 1 & p(c_k|I) > \delta \quad c_k \in C_I \\ 0 & otherwise \end{cases} \tag{5}$$

We set K, δ, M empirically, say $K = 20, \delta = 0.1, M = 6$.

Thus, the concept frequency is calculated as follows:

$$cf_k = \sum_{i=1}^{N} cf(I_i, c_k) \tag{6}$$

where N is the total number of images.

For query Q, the concept distribution of query $\{p(z_k|Q)\}$ is generated by the word distributions over topics $\{p(w_i|z)\}$. Determining the query's concepts is the same as the reference image and $\{p(c_k|Q)\}$ is obtained. Substituting cf_k and the concept distribution of query $\{p(c_k|Q)\}$ in equation3, we finally obtain the v-QPP score.

Table 1: Four datasets for evaluating QPP methods

Dataset	Queries	References	Avg refer. per query	mAP
Zubud	115	425	5	0.7495
ETRI	83	2057	24.78	0.4939
PKU	567	5007	8.83	0.2671
Telecom Italia	1620	360	2	0.8240

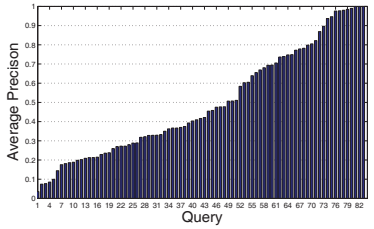


Fig. 6: The TF-IDF based AP distribution of all queries in ETRI. TF-IDF retrieval suffers from a wide range of performances, indicating the necessity of effective v-QPP.

4. EXPERIMENTS

4.1. Dataset

We conduct v-QPP evaluation over MPEG CDVS benchmarking datasets for mobile visual search[8]. Four datasets are selected including ZuBud, ETRI, PKU, Telecom Italia. Regarding dataset selection, we have two considerations: (1) Evaluating QPP on a wide range of queries (both difficult and easy ones)with various mAP performance; (2) Evaluating QPP over multiple datasets to investigate the generality. Tab.1 lists the dataset attributes, where the retrieval performance (in mAP) of TF-IDF is rather diverse among datasets. Fig.6 shows the varying performance of different queries over ETRI.

4.2. Baselines

To evaluate v-QPP, the ground truth retrieval is by TF-IDF. Each query's performance is measured by average precision (AP). As few work is on visual QPP, we compare our v-QPP with five representative pre-retrieval QPP methods widely used in document retrieval: (1) Averaged Inverse Document Frequency (AvIDF)[4]. (2) Maximum Inverse Document Frequency (MaxIDF) [9]. (3) Standard Deviation of IDF (DevIDF) [5]. (4) Averaged Inverse Collection Term Frequency (AvICTF) [5]. (5) Simplified Clarity Score(SCS) [5].

Table 2: Comparison results of different pre-retrieval QPP methods on ETRI and zubud datasets using Pearson correlation γ, Kendall' s τ, and Spearman's ρ . The best results are highlighted by bold. The QDP methods with $p > 0.05$ are marked in italic, indicating this QDP method is not statistically significant.$\times$ means constant QPP results do not yield correlation values. (Due to space limit, the detailed results over two other datasets are not listed here.)

Method	ETRI						Zubud					
	γ	p	τ	p	ρ	p	γ	p	τ	p	ρ	p
AvIDF	0.06	*0.60*	0.04	*0.59*	0.04	*0.70*	-0.24	9.1e-03	-0.17	1.1e-02	-0.24	9.6e-03
MaIDF	0.05	*0.65*	0.04	*0.68*	0.05	*0.67*	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$
DevIDF	-0.18	*0.11*	-0.09	*0.21*	-0.13	*0.24*	-0.33	2.7e-04	-0.27	3.3e-05	-0.38	2.3e-05
AvICTF	-0.01	*0.96*	-0.01	*0.89*	-0.03	*0.81*	-0.29	1.4e-03	-0.21	1.6e-03	-0.30	1.1e-03
SCS	-0.16	*0.14*	-0.09	*0.22*	-0.13	*0.23*	0.13	*0.15*	0.06	*0.36*	0.07	*0.46*
V-QPP	**0.59**	6.2e-09	**0.41**	3.5e-08	**0.61**	6.4e-10	**0.73**	2.1e-20	**0.56**	5.2e-18	**0.75**	4.2e-22

4.3. Evaluation

We evaluate QPP effectiveness by measuring the correlation between the predicted QPP score and the actual retrieval performance (AP) over test queries. We consider three correlation coefficients, Pearson correlation γ (linear correlation),Kendall's τ and Spearman's ρ (rank correlation). The correlation value is defined in the interval [-1, 1]. The higher absolute correlation value, the stronger correlation two variable lists have. All comparing baselines assume the positive correlations (high score suggests better retrieval performance),so high positive correlation value indicates better QPP. In addition, p-value is to test the statistic significance of the correlation. p>0.05 indicate the correlation is probably by chance; otherwise, it is statically significant.

Referring to Fig.5 and Tab.2, textual QPP baselines are ineffective in visual search. For example, over ETRI dataset, the correlation value of AvIDF is weakly positive, while MaIDF,SCS, AvICTF and DevIDF are even negative, which is against the positive correlation rationale of baselines. Morever, the correlation is not statistically significant for p>0.05. The failure of these textual QPP may be attibuted to visual words' lack of semantics. Unlike test words, visual words are much less discriminative, and visual words' statistics (e.g., IDF, TF) are poorly related to the specificity of a query. Therefore, we argue that directly applying textual QPP methods to visual search is inappropriate. Compared with baselines, our v-QPP is shown to be much more effective in predicting the visual query quality. Our v-QPP has yielded stronger positive correlation values as shown in Fig.4 and Tab.2. In particular, p value is less than 0.05, which indicates v-QPP is statistically significant.

Strong positive correlation values have demonstrated that our proposed v-QPP factors (q-INS, c-DSC) are effective in predicting the effectiveness of visual query. As a sort of formula to measure a query's discriminability, both AvIDF and MaIDF scores weakly correlate with AP. But the topics based concepts are more effective to implement v-QPP factors than visual words. The consistently better performance over four dataset has shown our method is generic and stable. Note that the combination of QPP methods(e.g.[5]) is not included in QPP baselines, as any single QPP method of not statistically significance would make the combination not statistically significant.

5. CONCLUSIONS

We have proposed a novel v-QPP method to predict the visual query performance. This pre-retrieval v-QPP is expected to significantly improve user experiences in visual search, which could benefit the saving of bandwidth cost as well as power consumption in mobile environments. Experiments across several datasets have demonstrated our v-QPP's effectiveness. As the ranked lists of results retrieved in response to a query are not considered, our v-QPP may fail in a more complex retrieval system. How to immigrate this model to mobile platforms is included in our future work.

6. ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China under contract no. 2009CB320902, in part by grants from the Chinese National Nature Science Foundation under contract no. 60902057, and in part by the CADAL Project Program.

7. REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *PAMI*, 2008.

[3] David Carmel and Elad Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Morgan and Claypool Publishers, 2010.

[4] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *SIGIR*, 2002.

[5] Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *In Proc. Symposium on String Processing and Information Retrieval*, 2004.

[6] Enamul Hoque, Grant Strong, Orland Hoeber, and Minglun Gong. Conceptual query expansion and visual search results exploration for web image retrieval. In *AWIC*, 2011.

[7] J. Philbin, J.and Sivic and A. Zisserman. Geometric latent dirichlet allocation on a matching graph for large-scale image datasets. *IJCV*, 2010.

[8] Yuri Reznik. Evaluation framework for compact descriptors for visual search. In *Requirements Subgroup. MPEG N12202*, 2011.7.

[9] Falk Scholer, Hugh E. Williams, and Andrew Turpin. Query association surrogates for web search: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 2004.

[10] Anna Shtok, Oren Kurland, and David Carmel. Predicting query performance by query-drift estimation. In *ICTIR*, 2009.

[11] Aixin Sun and Sourav S. Bhowmick. Quantifying tag representativeness of visual content of social images. In *MM*, 2010.

[12] Y.and Xing, X.and Zhang and M Han. Query difficulty prediction for contextual image retrieval, 2010.