Harvesting Visual Concepts for Image Search with Complex Queries

Liqiang Nie[†], Shuicheng Yan[†], Meng Wang[§], Richang Hong[§] and Tat-Seng Chua[†]
[†] National University of Singapore
[§] Hefei University of Technology
{nieliqiang, eric.mengwang, hongrc.hfut}@gmail.com
{eleyans, chuats}@nus.edu.sg

ABSTRACT

The use of image reranking to boost retrieval performance has been found to be successful for simple queries. It is, however, less effective for complex queries due to the widened semantic gap. This paper presents a scheme to enhance web image reranking for complex queries by fully exploring the information from simple visual concepts. Given a complex query, our scheme first detects the noun-phrase based visual concepts and crawls their top ranked images from popular image search engines. Next, it constructs a heterogeneous probabilistic network to model the relatedness between the complex query and each of its crawled images. The network seamlessly integrates three layers of relationships, i.e., the semantic-level, cross-modality level as well as visual-level. These mutually reinforced layers are established among the complex query and its involved visual concepts, by harnessing the contents of images and their associated textual cues. Based on the derived relevance scores, a new ranking list is generated. Extensive evaluations on a real-world dataset demonstrate that our model is able to characterize the complex queries well and achieve promising performance as compared to the state-of-the-art methods. Based on the proposed scheme, we introduce two applications: photo-based question answering and textual news visualization. Comprehensive experiments well validate the proposed scheme.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.3.5 [**Information Systems**]: Information Storage and Retrieval

Keywords

Complex Image Query, Photo-based QA, News Visualization

1. INTRODUCTION

With the exponential growth of image contents available online, searching for images is becoming an indispensable activity in people's daily lives. Meanwhile, web image retrieval has attracted world-wide research attentions and achieved great success for simple textual queries [25, 13, 20, 31]. As the web surfers get increasingly savvy and specific with their search behaviors, the queries tend to be more complex and sophisticated. This phenomenon is consistent with the report from Hitwise [1] in late *2009*: the average query length is getting longer from *2007*. In addition, long queries are becoming more and more popular in various media search applications, such as multimedia question answering [24], text illustration [17, 11], and known item search [8].

A complex image query is defined as a natural language query comprising several inter-related concepts. One example is the query "a baby with an apple lying on the bed". Here there are three concepts: baby, apple and bed. These concepts are linked by internal relationships: baby with an apple, both baby and apple are on the bed. It is obvious that complex queries can express specific information needs more precisely than the shorter ones. However, current commercial web search engines do not, in general, perform well with verbose queries, especially for image retrieval[1]. This is due to the following reasons. First, compared to simple queries, long ones frequently consist of more concepts, which further widen the semantic gap between the textual queries and the visual contents. Second, a complex query usually depicts the intrinsic semantic relationships among its constituent visual concepts. These kind images have loose coupled relationships with the surrounding textual descriptions, causing poor text-based search performance. Third, while there are abundant positive samples and query logs for simple queries, the positive samples are rare for complex queries. This makes learning based model less effective. Therefore, it is not surprising that the returned images are often incorrectly ranked for complex queries.

Visual reranking techniques can drastically improve the traditional text-based image search results. The existing approaches generally fall into two categories. One is pseudo relevance feedback (PRF) based [23, 34, 20]. They treat a significant fraction of the top images as pseudo-positive examples and collect some bottom images as pseudo-negative examples. They then either learn a classifier or cluster the images to perform reranking. But, for complex queries, relevant samples are usually rare or not ranked at the top of the

MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$15.00.

[1] A study in [28] shows that a failed image query tends to be longer than the average successful query, which indicates longer queries' higher specificity of contents and also reveals the limitations of current web image search engines for complex queries.

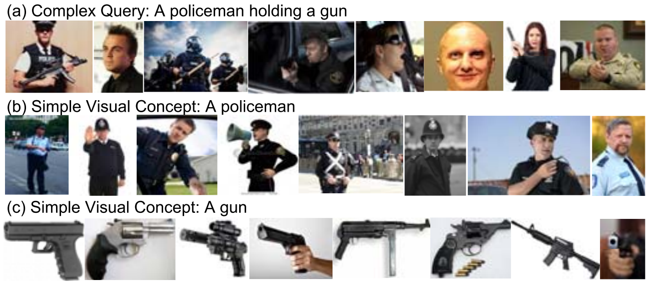
(a) Complex Query: A policeman holding a gun
(b) Simple Visual Concept: A policeman
(c) Simple Visual Concept: A gun

Figure 1: Image retrieval results comparison. The search results of a complex query are less visually consistent than those retrieved by its constituent visual concepts.

result list. This severely limits the ability to select pseudo positive and negative training samples. The other category is graph based [32, 12, 30] that propagates the initial ranking information over the whole graph until convergence. However, for complex query, many irrelevant images are frequently distributed in high ranked positions initially. These irrelevant images can hardly be pushed down by the graph-based methods, since they often have low similarities with other irrelevant images in the lower ranked positions [22]. Consequently new approaches towards image reranking for complex queries are highly desired.

To tackle this problem, we hypothesize that the search results of a complex query are less visually consistent and coherent than those retrieved by each of its constituent visual concept; and the latter characterize the former's partial features in terms of both semantics and visual exemplars. An example illustrating the assumption is intuitively demonstrated in Figure 1. Based on this assumption, we explore the information cues from visual concepts to enhance Web image reranking for complex queries. Specifically, we propose a scheme, which contains two main components as shown in Figure 2. The first component identifies the involved visual concepts by leveraging lexical and corpus-dependent knowledge, and collects the top relevant datapoints from popular image search engines. The second component constructs a heterogeneous probabilistic network to model the relevance between the complex query and each of its retrieved images. This network comprises three subnetworks, each representing a layer of relationship, including: (a) the underlying relationship among image pairs, (b) the cross-modality relationship between the image and the visual concept[2], and (c) the high-level semantic relationship between visual concept and the complex query[3]. The three layers are strongly connected by a probabilistic model. The layers mutually reinforce each other to facilitate the estimation of relevance scores for new reanking list generation. Most importantly, the whole process is unsupervised and can be extended to handle large-scale data.

Based on the proposed scheme, we introduce two potential application scenarios of web image reranking for complex queries: photo-based question answering (PQA) and textual news visualization (TNV) [17]. PQA is a sub-branch of multimedia question answering [24], aiming to answer questions with precise image information, which provides answer seekers with better multimedia experience. TNV is to complement the textual news with context associated images, which may better draw the readers' attention or help them grasp the textual information quickly. By conducting experiments on the real-world datasets, we demonstrate that our proposed scheme yields significant gains in reranking performance for complex queries, and achieves fairly satisfactory results for these two applications.

The main contributions of this research are:

1. To the best of our knowledge, this is the first that targets web images reranking for complex queries from the probabilistic perspective. This work unravels the unreliable initial ranking list problem of the existing image reranking approaches for complex queries.
2. It proposes a heuristic approach to detect noun-phrase based visual concepts from complex query, instead of just treating individual terms as possible concepts.
3. It proposes a heterogeneous probabilistic network to automatically estimate the relevance score of each image, which jointly couples three layer relationships, spanning from semantic level to visual level. This is different from the conventional complex query modelling approaches [35, 16, 36, 2] that either require human interactions or consider the query terms independently and neglect the connections among them.

The remainder is organized as follows. Sections 2 and 3 respectively review the related work and briefly introduce the reranking scheme. Sections 4 and 5 introduce visual concept detection and the proposed heterogeneous probabilistic network, respectively. Experimental results and analysis are presented in Section 6, followed by the applications in Section 7. Finally, Section 8 contains our remarks.

2. RELATED WORK

2.1 Complex Queries in Text Search

Several recent research efforts have been conducted for improving long query performance in text-based information retrieval. These efforts can be broadly categorized into automatic query term re-weighting [6, 5, 15, 7] and query reduction [13, 14, 4] approaches.

It has been found that assigning appropriate weights to query concepts has significant positive effects on retrieval performance [6]. Bendersky and Croft [5] developed and evaluated a technique that assigns weights to the identified key concepts in the verbose query, and observed improved retrieval effectiveness. Lease et al. [15] presented a regression framework to estimate term weights based on knowledge from past queries. A novel method beyond unsupervised estimation of concept importance was proposed in [7], which weights the query concept using a parameterized combination of diverse importance features.

Pruning the complex query to retain only the important terms is also recognized as one crucial dimension to improve search performance. Kumaran and Allan [13, 14] proposed an interactive query induction approach, by presenting the users with the top 10 ranked sub-queries along with corresponding top ranking snippets. The tabbed interface allows the user to click on each sub-query to view the associated snippet, and select the most promising one as their new query. A more practical approach was proposed in [4], utilizing efficient query quality prediction techniques to evaluate

[2] The underlying visual associations among visual concepts are also integrated.

[3] The semantic associations among visual concepts are also considered in this layer.

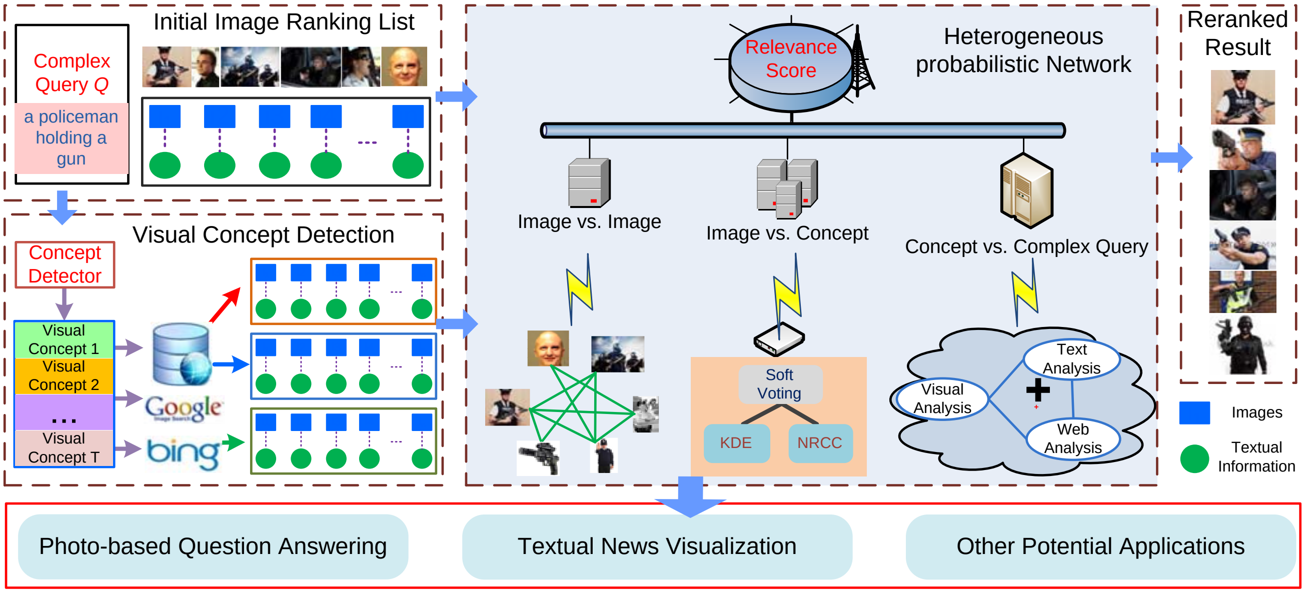

Figure 2: Illustration of the proposed web image reranking scheme for complex queries. It contains two components, i.e., visual concept detection and relevance estimation. This scheme facilitates many applications, including photo-based question answering, textual news visualization and others.

the reduced versions of the original query that were obtained by dropping one single term at a time. It can be incorporated into existing web search engines' architectures without requiring modifications to the underlying search algorithms.

Though great success has been achieved for complex query processing in text search domain, these techniques cannot be directly applied to the general media domain due to the different modalities between the query and search results.

2.2 Complex Queries in Media Search

Some research efforts have been conducted on modelling complex queries in media search. For example, Aly et al. [2] proposed fusion strategies to model combined semantic concepts by simply aggregating the search results from their constituent primitive concepts. However, such approach fails to characterize complex queries as it overlooks the mutual relationships among different aspects of complex queries. Image search by concept map was proposed in [33]. It presents a novel interface to enable users to indicate the spatial distribution among semantic concepts. However, the input model is not consistent with the current popular search engines and the concept-relationship is not limited to spacial arrangement. Yuan et al. [35] explored how to utilize the plentiful but partially related samples, as well as the users' feedbacks, to learn complex queries in interactive concept-based video search. This work gracefully compensates the insufficient relevant samples. Further, Yuan [36] moved one step beyond primitive concepts and proposed a higher-level semantic descriptor named "concept bundle" to enhance video search of complex queries. But these two works are supervised. Recently, harvesting social images for bi-concept search was proposed in [16] to retrieve images in which two concepts are co-occurring. However, it is unable to handle multiple concepts.

Overall, literature regarding complex queries in media search is still relatively sparse, and the existing approaches either view the query terms independently or require intensive human interactions. Differing from the existing works, our approach models the complex queries automatically, and jointly considers the relationships between concepts and the complex queries from high-level to low-level.

3. WEB IMAGE RERANKING SCHEME

As aforementioned, a complex query Q comprises several visual and abstract concepts as well as their intrinsic relations. As shown in the left part of Figure 2, we first perform visual concepts selection, since they have strong description in images. Supposing T visual concepts $\mathcal{C} = \{q_1, q_2, \dots, q_T\}$ are detected. The T visual concepts are then regarded as simple queries to a commercial search engine and retrieve a collection of images $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_L, y_L)\}$. Here the image $\mathbf{x}_i$ ($\mathbf{x}_i \in R^d$) is crawled using simple visual concept y_i ($y_i \in \mathcal{C}$). Complex query Q has an ordered image list $\mathcal{X} = \{(\mathbf{x}_{L+1}, \mathbf{x}_{L+2}, \dots, \mathbf{x}_{L+N}\}$. Our target is to explore the visual concepts and their partial relations to enhance the image relevance estimation with respect to the given complex query, i.e., $Score(Q, \mathbf{x}_u)$, $u = L+1, \dots, L+N$. Based on these relevance scores, a new refined ranking list will be generated.

To estimate the relevance score, we propose a heterogeneous probabilistic network as displayed in the middle part of Figure 2, which is inspired by the KL-divergence measure [3]. It is composed of several dissimilar sub-networks, which provide probabilistic estimations from different angles. But the constituents are of a conglomerate mass, strongly connected by a probabilistic model. It is formally formulated as,

$$Score(Q, \mathbf{x}_u) = -\sum_{q_c \in Q} P(q_c|Q) \times \log P(q_c|\mathbf{x}_u) \qquad (1)$$

where $P(q_c|Q)$ measures the importance of a visual concept q_c given the complex query Q, i.e., the high level semantic relatedness between a visual concept and the complex query. The second term in Eq.(1) can be further decomposed as,

$$P(q_c|\mathbf{x}_u) = \sum_{i=1}^{L} P(q_c|\mathbf{x}_i) \times P(\mathbf{x}_i|\mathbf{x}_u) \tag{2}$$

where $P(q_c|\mathbf{x}_i)$ involves two different modalities, specifically, the high level concept and the low level visual content; while $P(\mathbf{x}_i|\mathbf{x}_u)$ measures the underlying visual relatedness of image pairs.

The above formulation intuitively reflects that our proposed heterogeneous probabilistic network comprises three sub-networks, representing three different relationship layers: semantic level, cross-modality level and visual level.

4. VISUAL CONCEPT DETECTION

In this paper, a visual concept is defined as a noun phrase depicting a concrete entity with a visual form. Beyond visual concepts, complex queries tend to contain several redundant chunks. These redundant chunks have grammatical meaning for communication between humans to help understand the key concepts [26], but are hard to model visually. One example is the query, “find images describing the moment the astronaut getting out of the cabin”. In this query, only “the astronaut” and “the cabin” have high correspondence with the visual contents, while the use of other chunks may bring unpredictable noise to the image reranking method. Therefore, to differentiate the visual content related chunks from unrelated ones, we propose a heuristic framework for visual concept detection as illustrated in Figure 3. A central resource in this framework is an automatically constructed visual vocabulary. Now given a complex query, we extract its constituent visual concepts as follows:

1. We segment a given complex query Q into several chunks using the openNLP[4] tool.
2. For each chunk, we match it against our constructed visual vocabulary. If any of its terms matches a term in our visual vocabulary, the chuck is classified as a visual concept. This detected visual concept is used as a simple query to retrieve the top ranked images and their surrounding texts for reranking purpose.
3. We construct a flexible vocabulary containing visual related words, by leveraging the lexical and corpus-dependent knowledge. Specifically, we collect all the noun terms from our dataset utilizing the Part-Of-Speech Tagger[5], and remove stop words from the noun set. For each selected noun word, we traverse along its hypernyms path in the WordNet, until one of the five predefined high-level categories is reached. They are “color”, “thing”, “artifact”, “organism”, and “natural phenomenon”. These 5 categories cover almost all the key concepts in our dataset. The noun words that match to these 5 categories are recognized as visual related. This approach is analogous to [18].

Compared to the conventional single-word based visual concept definition [18, 31], the noun-phrase based definition is able to incorporate a lot of adjunct terms, such as “a red apple”, which carries additional color cue for “apple”.

5. HETEROGENEOUS NETWORK

In this section, we will discuss in greater detail each component of our proposed heterogeneous probabilistic network, namely, semantic relatedness estimation, visual relatedness estimation and cross-modality relatedness estimation.

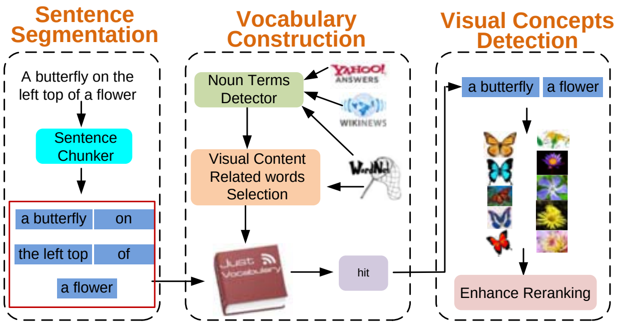

Figure 3: An illustration of visual concepts detection from a given complex query.

5.1 Semantic Relatedness Estimation

Different concepts play different roles in the given complex query, and concept weighting [5, 15, 7] has been studied for decades to quantify their importances. However, these conventional methods are developed for long query in text search domain; few of them take the visual information into consideration. Instead, our approach estimates the semantic relatedness in image search by linearly integrating multi-faceted cues, i.e., visual analysis, external resource analysis as well as surrounding text analysis.

First, from the perspective of underlying visual analysis, we respectively denote $\mathcal{X}_c$ and $\mathcal{X}$ to be the set of images retrieved by the visual concept q_c and complex query Q. Their relatedness can be defined as,

$$V(q_c, Q) = \frac{1}{|\mathcal{X}_c| \times |\mathcal{X}|} \sum_{\mathbf{x}_i \in X_c, \mathbf{x}_j \in X} K(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

$K(\cdot,\cdot)$ is the Gaussian similarity function, defined as,

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma^2}) \tag{4}$$

where the radius parameter, σ, is simply set as the median of the Euclidean distances of all related image pairs.

Second, actually the visual concepts detected from the same complex query are usually not independent. For example, for the complex image query “a lady driving a red car on the road”, the semantical relationship between “a red car” and “the road” is relatively high. Inspired by Google distance [10], we estimate the inter-concepts relatedness based on the frequency of their co-occurrence by exploring the Flickr image resource as the largest publicly available multimedia corpus,

$$NGD(q_c, q_j) = \frac{\max(\log f(q_c), \log f(q_j)) - \log f(q_c, q_j)}{\log M - \min(\log f(q_c), \log f(q_j))} \tag{5}$$

where M is the total number of images retrieved from Flickr, roughly estimated as 5 billion. $f(q_c)$ and $f(q_j)$ are respectively the numbers of hits for search concepts q_c and q_j, and $f(q_c, q_j)$ is the number of web images on which both q_c and q_j co-occur. Note that we define $NGD(q_c, q_j) = 0$, if $q_c = q_j$. Then the relatedness between q_c and the given

[4] http://incubator.apache.org/opennlp/
[5] http://nlp.stanford.edu/software/tagger.shtml

complex query Q is:

$$G(q_c, Q) = \frac{1}{T} \sum_{q_j \in \mathcal{C}} NGD(q_c, q_j) \tag{6}$$

where T is the number of visual concepts detected from Q. This estimation can be viewed as exploring the external web image knowledge to weight the visual concepts.

Third, we estimate the semantic relatedness by using the surrounding text-matching score. For each complex query Q, we first merge all surrounding textual information of its retrieved images, such as tag, title, description, etc, into a single document. The same operation is then conducted for all the T detected visual concepts, resulting in T documents. We then parse the $T+1$ documents using the OpenNLP tool. All nouns and adjectives are selected as salient words, since they are observed to be more descriptive and informative than verbs or adverbs. Based on these salient words, the tf-idf scores [36] are computed to represent the semantic relatedness between a visual concept q_c and the given complex query Q, denoted as $T(q_c, Q)$.

Finally, we linearly combine these three measures as,

$$P(q_c|Q) = \alpha_1 V(q_c, Q) + \alpha_2 G(q_c, Q) + \alpha_3 T(q_c, Q) \tag{7}$$

where α_i is the fusing weight with sum being 1. They are selected based on a training set comprising 20 complex queries, which are randomly sampled from our constructed complex query collection. We tune the weights to the values that optimize the average NDCG@50 with grid search.

5.2 Visual Relatedness Estimation

To explore the visual relationship between images, we perform Markov random walk over a K nearest neighbour graph to propagate the relatedness among images. The vertices of the graph are the $L+N$ images and the undirected edges are weighted with pair-wise similarity. We use $\mathbf{W}$ to denote the similarity matrix and W_{ij}, its (i,j)-th element, indicates the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Typically, it is estimated as

$$W_{ij} = \begin{cases} K(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mathbf{x}_j \in N_K(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_K(\mathbf{x}_j) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $N_K(\mathbf{x}_i)$ denotes the index set for the K nearest neighbours of image x_i computed by Euclidean distance. Noting that W_{ii} is set as 1, so that self-loop is included.

Denoting $\mathbf{A}$ as the one step transition matrix. Its element A_{iu} indicates the probability of the transition from node i to node u and is computed directly from the related weights,

$$A_{iu} = \frac{W_{iu}}{\sum_j W_{ij}} \tag{9}$$

The L_1 normalization of each row turns $\mathbf{A}$ into a stochastic transition matrix. Then the probability of a random walk, which initially starts from node i, and stops at node u after t steps, can be denoted as

$$P(\mathbf{x}_u(t)|\mathbf{x}_i(0)) = [\mathbf{A}^t]_{iu} \tag{10}$$

Based on Eq.(10), we can simply evaluate the probability that the walker starts from $\mathbf{x}_i$ at time 0 given that it ends at $\mathbf{x}_u$ at time t, with the reasonable assumption that the starting node is uniformly chosen,

$$\begin{aligned} P(\mathbf{x}_i(0)|\mathbf{x}_u(t)) =& \frac{P(\mathbf{x}_u(t)|\mathbf{x}_i(0)) \times P(\mathbf{x}_i(0))}{P(\mathbf{x}_u(t))} \\ =& \frac{P(\mathbf{x}_u(t)|\mathbf{x}_i(0))}{\sum_j P(\mathbf{x}_u(t)|\mathbf{x}_j(0))} \\ =& \frac{[\mathbf{A}^t]_{iu}}{\sum_j [\mathbf{A}^t]_{ju}} \end{aligned} \tag{11}$$

Since visual concepts and complex query are associated with the starting and ending images, respectively, Eq.(2) can be rewritten as

$$P(q_c|\mathbf{x}_u) = \sum_{i=1}^{L} P(q_c|\mathbf{x}_i) \times P(\mathbf{x}_i(0)|\mathbf{x}_u(t)) \tag{12}$$

5.3 Cross-Modality Relatedness Estimation

As mentioned above, $P(q_c|\mathbf{x}_i)$ in Eq.(2) measures the relatedness between two different modalities, the high-level concept and the low-level visual information. We now present two techniques to link these two modalities: kernel density estimation approach (KDE) [19] and normalizing relatedness cross concepts (NRCC) [29].

5.3.1 KDE Approach

For each image $\mathbf{x}_u$ retrieved by Q, $P(q_c)$ is identical and $P(\mathbf{x}_i)$ is assumed to be uniform. Therefore Eq.(12) can be restated as,

$$P(q_c|\mathbf{x}_u) \propto \sum_{i=1}^{L} P(\mathbf{x}_i|q_c) \times P(\mathbf{x}_i(0)|\mathbf{x}_u(t)) \tag{13}$$

where $P(\mathbf{x}_i|q_c)$ is the probability density function, representing the relevance of an image to the given visual concept. KDE approach is utilized to perform the estimation.

We use $\mathcal{X}_c$ to denote the set of images retrieved by the visual concept q_c, the KDE approach measures $P(\mathbf{x}_i|q_c)$ as

$$P(\mathbf{x}_i|q_c) = \frac{1}{|\mathcal{X}_c|} \sum_{\mathbf{x}_j \in \mathcal{X}_c} K(\mathbf{x}_i, \mathbf{x}_j) \tag{14}$$

The above equation can be intuitively interpreted as follows: q_c and each of its retrieved images in $\mathcal{X}_c$ can respectively be viewed as a family and family members. Then the closeness of an unknown image to this family is estimated by averaging the soft voting from all family members.

5.3.2 NRCC Approach

The drawback of the KDE approach is that it does not take the underlying associations among visual concepts belonging to the same complex query into consideration. To compensate for this limitation, we formally define $P(q_c|\mathbf{x}_i)$ as,

$$P(q_c|\mathbf{x}_i) = \frac{1}{Z_i} \sum_{\mathbf{x}_j \in \mathcal{X}_c} P(\mathbf{x}_j(0)|\mathbf{x}_i(t)) \tag{15}$$

where Z_u is a normalizing factor, and formulated as,

$$Z_i = \sum_{q_c \in \mathcal{C}} \sum_{\mathbf{x}_j \in \mathcal{X}_c} P(\mathbf{x}_j(0)|\mathbf{x}_i(t)) \tag{16}$$

As its formulation implies, this approach is named as normalizing relatedness cross concepts (NRCC), which has been preliminarily studied in [29].

Compared to the KDE approach, by regarding $\mathcal{C}$ as a community with several families, the relatedness between the given image $\mathbf{x}_i$ and a family q_c, is determined not only by the family members $\mathbf{x}_j$ in q_c, but also other community families in $\mathcal{C}$.

5.4 Discussions

To further study the impact of the number of transitions t on visual relatedness, we first define the stationary probability vector $\boldsymbol{\pi}$ of the stochastic transition matrix $\mathbf{A}$ that does not change under the power of $\mathbf{A}$. Mathematically, it is expressed as,

$$\boldsymbol{\pi}\mathbf{A} = \boldsymbol{\pi} \tag{17}$$

The Perron-Frobenius theorem [27] ensures every stochastic matrix has such vectors; and for a matrix with strictly positive entries, this vector is unique. It can be computed by observing that for any i,

$$\lim_{t\to\infty}[\mathbf{A}^t]_{iu} = \pi_u \tag{18}$$

where π_u is the u-th element of the row vector $\boldsymbol{\pi}$. It implies that when $t \to \infty$, the probability of being in a state u is independent of the initial state i. Namely, all the starting points become indistinguishable. In the other limiting case, when $t = 1$, we utilize only the neighbourhood graph, which will be totally influenced by K.

The local neighbourhood size K should be large enough to guarantee a singly connected graph. Meanwhile, K should be sufficiently small to avoid introducing more edges between the relevant and irrelevant samples, which may degrade the reranking performance drastically. However, too small a K will miss the "correct" edges between the relevant samples, resulting in the weakening of the key consistency.

The computational complexity of our approach mainly comes from two parts: feature extraction and transition matrix iteration. The former is the most computationally expensive step, but can be handled off-line. The cost of the latter scales as $O(d(L+N)^2 + t(L+N)^2)$, where d is the 1428-dimension features, and t is the number of transitions in dozens level. Since we only use the top results, $L+N$ is usually in the order of thousands. Thus the computational cost is very low. In our experiments, the process can be completed in less than 1 second if we do not take the feature extraction part into account (3.4GHz and 8G memory).

It is worthy emphasizing that our proposed scheme can also be applied for other image repositories, even for those images without surrounding texts by simply ignoring the last term in Eq.(7).

6. EXPERIMENTS

6.1 Experimental Settings

We collected a large real-world dataset from WikiAnswers, which contains $1,944,492$ unique QA pairs and covers a wide range of topics, including entertainment, life, education, etc. Based on this dataset, we constructed a visual word vocabulary, from which 100 most frequent visual words are selected. We then issued these terms into Google Image and selected 50 suggested complex queries according to our definition. Some representative samples are listed in Table 1. For each complex query and its embedded visual concepts, the top 500 images are crawled from Google Image.

Table 1: The representative complex queries generated based on our corpus and Google Image suggestion. Here we do not illustrate all the queries due to limited space.

ID	Complex Query
1	President Obama and troops
2	Women swimming in pool
3	Soldiers holding American flag on the mountain
4	Baby with an apple lying in the bed
5	A lady driving a red car on the street
6	A cowboy riding a horse at sundown
7	Lions attacking zebras on the grassland
8	A man walking his dog in the park
9	A lady wears sunglasses on the sea beach
10	Comparison between white iphone and black iphone

To obtain the relevance ground truth of each image, we conduct a manual labelling procedure. Five human annotators were involved in the process. Each image was labelled to be very relevant (score 2), relevant (score 1) or irrelevant (score 0) with respect to the given query. We performed a voting to establish the final relevance level of each image. For the cases that there were two classes having the same number of ballots, a discussion was carried out among the labelers to decide the final ground truths.

To represent the content of each image, we extracted the following features:

1. We used the difference of Gaussians to detect keypoints in each image and extracted their SIFT descriptors. By building a visual codebook of size 1000 based on K-means, we obtained a 1000-dimensional bag-of-visual-words histogram for each image.
2. We further extracted 428-dimensional global visual features, including 225-dimensional block-wise color moments based on 5-by-5 fixed partition of the image; 128-dimensional wavelet texture; and 75-D edge direction histogram.

When it comes to reranking performance evaluation, we adopted NDCG@n as our metric,

$$NDCG@n = \frac{rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}}{IDCG} \tag{19}$$

where rel_i is the relevance score of the i-th image in the ranked list, $IDCG$ is the normalizing factor that makes NDCG@n being 1 for a perfect ranking.

6.2 On Visual Concept Detection

Following our heuristic rules stated in Section 3, we first selected all the noun terms from WikiAnswers dataset the by Stanford Log-linear Part-Of-Speech Tagger. We filtered out the stop words from the noun set. We then went through the WordNet 3.0[6] hypernym hierarch within 10 steps, from bottom to top, to identify each selected word's hypernyms, until one of the five predefined high-level categories are matched: "color", "thing", "artifact", "organism", and "natural phenomenon". As shown in [18], these 5 categories cover a substantial part of many frequently used concepts in computer vision and multimedia domains. In this way, we constructed a visual word vocabulary containing $12,812$ noun entries. Table 2 illustrates their distribution statistics over the 5 categories in our vocabulary.

[6] http://wordnet.princeton.edu/

Table 2: The distribution of visual words over five predefined high-level categories.

Visual Category	Visual Words #	Percentage
Color	159	1.24%
Thing	919	7.17%
Artifact	4219	32.93%
Organism	7214	56.31%
Natural phenomenon	301	2.35%

Table 3: The confusion matrix of visual concept detection results. The prediction accuracy is 89.27%.

Class Prediction	Visual Concepts	Non Visual Concepts
Visual Concepts	102	5
Non Visual Concepts	17	81

From the selected 50 complex queries, 205 chucks are detected by OpenNLP, among which 119 chucks are manually voted as visual concepts by 5 volunteers. As mentioned previously, for each term in a given chuck, we search it in our constructed visual vocabulary, and this chuck will be categorized as a visual concept if at least one term is matched. Table 3 illustrates the confusion matrix obtained by our proposed visual concept detection. We can see that our approach achieves fairly good performance, i.e., 89.27%. The misclassification results mainly come from some chunks that are product names not archived in the WordNet, such as "iphone", "ipad", etc, and also from some verb chunks that have visual content descriptive attribute, such as "wear", etc. In our further work, we will broaden our visual word dictionary by incorporating product name list to boost our classification performance.

6.3 On Query Performance Analysis

We first conducted experiment to evaluate the retrieval effectiveness of the current dominant image search engines for simple and complex queries, respectively.

The selected 50 queries and their 119 involved visual concepts are regarded as complex queries and simple queries, respectively. Figure 4 displays the average search performance comparison. From the figure we can see that the search results of simple queries remarkably outperform those based on complex queries. And along with the increase of NDCG-depth n, the average performance of complex queries drops at a faster rate. This observation partially verifies our hypothesis that: compared to complex queries, the search results of simple queries are more visually consistent and coherent. Also, it reveals the fact that the current search engines, especially the image search engine, do not perform well for complex queries, even though great success has been achieved for simple queries. So effective image reranking for complex queries is highly desirable.

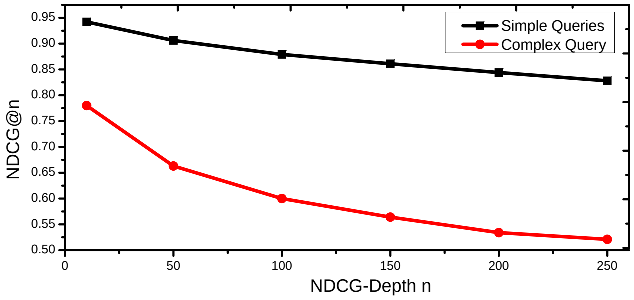

Figure 4: Retrieval performance comparison between complex queries and their belonging primitive visual concepts.

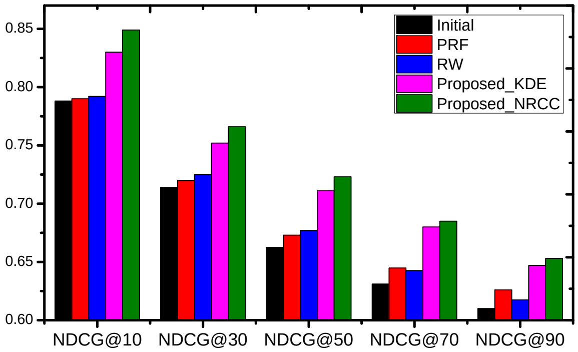

Figure 5: Performance comparison of different reranking approaches in terms of NDCGs.

6.4 On Reranking Performance Comparison

To demonstrate the effectiveness of our proposed approach, we comparatively evaluate the following unsupervised reranking methods:

- **RW**: Random walk reranking [12] is a typical graph-based reranking method jointly exploiting both initial ranking result and visual similarity between images. The stationary probability of random walk is used to compute the final relevance scores. (Baseline 1)
- **PRF**: Pseudo-Relevance Feedback [34]. A support vector machine (SVM) classifier is trained to perform the reranking based on the assumption that the top-ranked images for each query are more relevant than the low-ranked results in general. (Baseline 2)
- **Proposed_KDE**: Our proposed probabilistic reranking approach with cross-modalities relatedness estimation by KDE method.
- **Proposed_NRCC**: Our proposed probabilistic reranking approach with cross-modalities relatedness estimation by NRCC method.

For each method mentioned above, the involved parameters are carefully tuned, and the parameters with the best performances are used to report the final comparison results.

The experimental results are illustrated in Figure 5. It can be observed that our proposed approaches are consistently and substantially better than the current publicly disclosed state-of-the-art web image reranking algorithms across all evaluated NDCGs. From this figure, we can also observe that the improvements over the initial ranking result from RW and PRF are much slighter, especially for NDCGs with smaller n. The main reason is that they both have problems of unreliable initial ranking list, which frequently exists in complex query search. In contrast, our proposed scheme for complex queries is more robust, since it tends not to be affected too much by the initial ordering of images.

Further, it is observed that the proposed_NRCC stably outperforms the proposed_KDE approach. This is due to the fact that NRCC takes the relationship between visual concepts in the same complex query into consideration, while

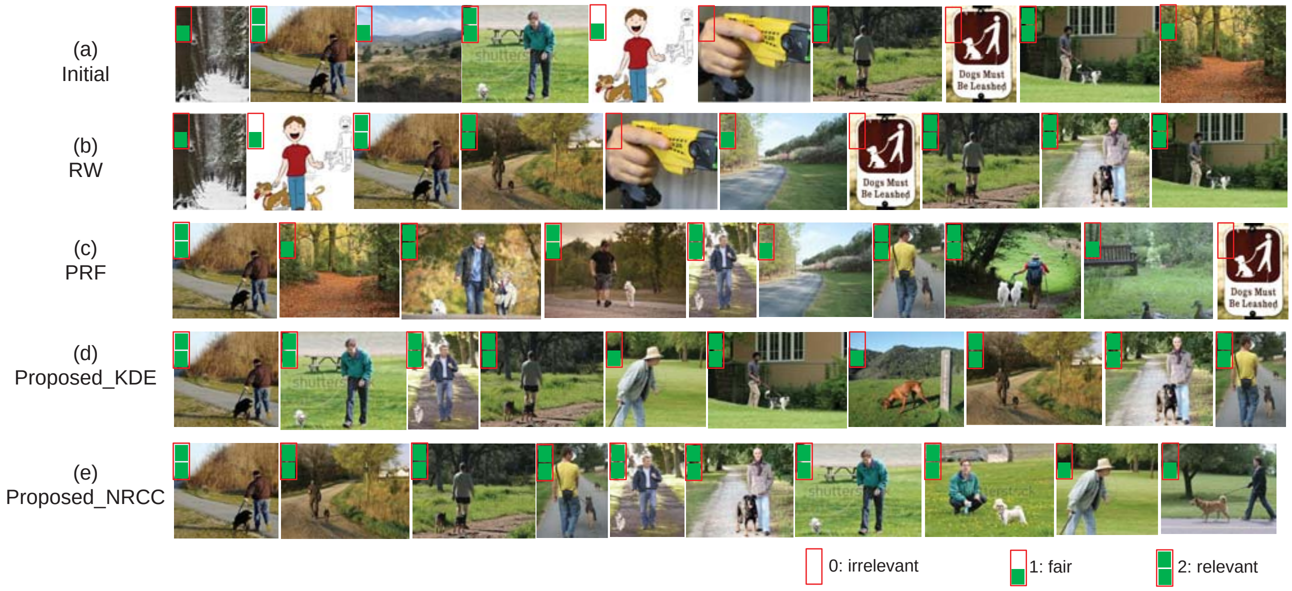

Figure 6: Illustrative results for complex query "soldiers holding American flag on the mountain" based on different reranking approaches. Our proposed approaches obtain the most satisfying results.

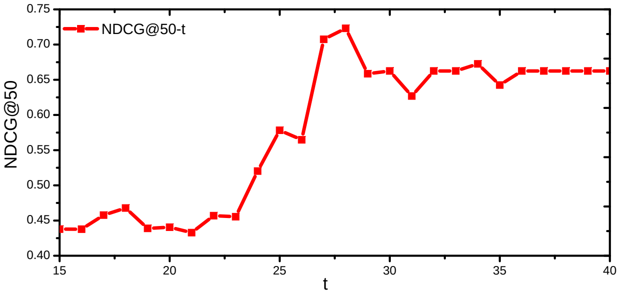

Figure 7: The performance with different t when K is fixed as 304.

KDE assumes that these visual concepts are independent, when estimating the relatedness between a visual concept and a given image.

Figure 6 illustrates the top 10 images before and after reranking for different reranking approaches for the complex query "soldiers holding American flag on the mountain". Obviously, our proposed approaches obtain the most satisfying results.

However, after examining the performance of each query, it is observed that our scheme fails to handle some queries that have explicit spatial or action relationship constraints between visual concepts. Examples include "a butterfly on the left top of the flower" and "a man listening to a mobile phone". We will integrate this kind information into our scheme for general complex queries in our future work.

6.5 On the Sensitivity of Parameters

As discussed above, both the number of transitions t and local neighbour size K are important parameters in our method. In this section, we further conduct experiments to investigate the effect of these parameters based on the proposed_NRCC. We first perform grid search with step size 1, to seek the t and K with optimal reranking performance. 28 and 304 are located for t and K, respectively.

The NDCG@50-t curve is presented in Figure 7 with K fixed as 304. As illustrated, the performance increases with t growing and arrives at a peak at a certain t, then the performance sharply decreases, and finally becomes relatively constant. This result is consistent with our previous analysis that when t tends towards infinite, all the starting points become indistinguishable.

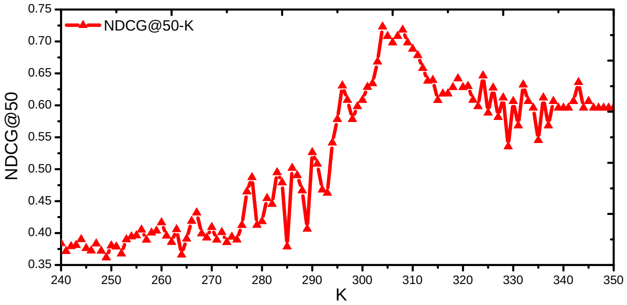

Figure 8: The performance with different K when t is fixed as 28.

Similarly, Figure 8 shows the NDCG@50-K curve with t fixed as 28, where the performance varies according to different K. With the gradual increase of K, more relevant samples are connected to each other, and "incorrect" edges between the relevant samples and irrelevant samples are potentially introduced. From Figure 8, it can be observed that NDCG@50 obtains the peak performance at $K = 304$, which is a trade-off value.

7. APPLICATIONS

In this section, we introduce two potential application scenarios of image reranking for complex queries: photo-based question answering and textual news visualization.

7.1 Photo-based Question Answering

7.1.1 Application Scenario

Community question answering (cQA) services have gained great popularity over the past decades [24, 21, 9], which encourage askers to post their specific questions on any topic and obtain answers provided by other participants. It also facilitates general users to seek information from the large repository of well-answered questions. However, existing cQA forums, such as Yahoo!Answers, Answerbag, MetaFilter, usually support only textual answers, which are not in-

Table 4: The distribution of visual concepts embedded in the generated queries for photo-based QA.

One Visual Concepts	Two Visual Concepts	More Than Two Visual Concepts
46.15%	38.85%	15.0%

tuitive for many questions, such as the question "what is the difference between alligators and crocodiles". Even when the answer is described by several very long sentences in Yahoo!Answers, it is still hard for users to grasp the appearance differences. Here it reflects the fact that a picture is worth a thousand words. However, noting that not all the QA pairs prefer image answers. Textual answer is sufficient when it comes to the quantity-type questions, such as "what is the population in China". Also video answers will be much more lively and interesting for procedure-oriented questions, such as "how to assemble a computer". Actually this is the so-called multimedia question answering [9], a rising topic in media search domain.

In this paper, we only focus on the QA pairs which may be better explained with images. However, as stated in [24], the queries generated from the textual QA pairs are usually very verbose and complex, not supported well by the current commercial image search engines. Based on our proposed approach, we develop a photo-based QA system, which automatically complements the original textual answers with relevant web images.

7.1.2 Experiments

To demonstrate the effectiveness of the PQA system, we conducted the experiment on 1000 non-conversational QA pairs, selected from Yahoo!Answers dataset [21], which contains 4,483,032 QA pairs. For each QA pair, five volunteers were invited to vote whether it can provide users with better experience by adding images instead of using purely texture descriptions. Around 260 QA pairs were selected. We then directly employed the method in [24] to generate a most informative query from each QA pair. Our statistics are shown in Table 4, which show that more than 53% of queries contain two or more visual concepts.

Accordingly, a query-aware reranking approach is proposed to select the top 10 relevant images. To be specific, if the query is simple, i.e., containing only one visual concept, then the RW [12] will be used directly. On the other hand, if the query is complex, we employ the proposed_NRCC. We compare our proposed approach with the following methods.

- **Naive Search**: Simply perform image search with each query on Google Image without reranking.
- **Naive Fusion**: Simply perform image search with each visual concept in the generated complex query, and then fuse the results.

Figure 9 shows the comparison of these three methods. It can be observed that our query-aware reranking approach outperforms the other two methods remarkably.

7.2 Textual News Visualization

7.2.1 Application Scenario

"Every picture tells a story" suggests to us the essence of visual communication via pictures. This phrase is also consistent with our common sense, i.e., pictures in textual news always facilitate and expedite our understanding, especially for elderly and juvenile. Meanwhile, searching the image database in order to provide several meaningful and illustrative pictures to their textual news is a routine task for news writers.

Table 5: The distribution of the number of pictures involved in news documents.

Without Any Picture	One Pictures	Two Pictures	> Two Pictures
46.15%	38.85%	15.0%	38.85%

However, the pictures contained in news documents are usually very few as shown in Table 5. which shows that more than 46% news documents do not contain any pictures. The statistical result is based on the experimental dataset. To assist news readers and news writers, we propose a scheme to automatically seek relevant web images that best contextualize the content of news.

7.2.2 Experiments

We directly used the news dataset in [17], crawled from ABCNews.com, BBS.co.uk, CNN.com and GoogleNews; it contains up to 48,429 unique documents after duplicate removal. To save manual labelling efforts, we randomly select 100 news documents from the whole data set for evaluation. It is observed that most of the news articles are fairly long, and it is not an easy task to extract descriptive queries. So we simply regard the expert generated titles of the news documents as complex queries due to their obvious summarizing attribute.

Further, it is observed that more than 43% of titles contain at least one person-related visual concept. So we propose to employ query dependent image representations for reranking. Specifically, let $\mathcal{X}_c$ and $\mathcal{X}$ the set of images retrieved by the visual concept q_c and complex query Q, respectively; and q_c is predicted as person related query by the method in [11]. Then for each image in $\mathcal{X}_c$ and $\mathcal{X}$, we performed face detection. We extracted the 256-dimensional Local Binary Pattern (LBP) features [28] from the largest face region for any $\mathbf{x}_i$ in $\mathcal{X}_c$; and the same features are extracted for all the detected faces for any $\mathbf{x}_u$ in $\mathcal{X}$. The similarity between $\mathbf{x}_i$ and $\mathbf{x}_u$ is then computed as,

$$W_{iu} = \max_{x \in \mathcal{O}_u} K(\mathbf{x}_i, \mathbf{x}) \tag{20}$$

where $\mathcal{O}_u$ is the set of LBP features extracted from the faces in image $\mathbf{x}_u$. Other image pair similarity is the same as previously introduced. We call this the query-aware presentation method.

To demonstrate the effectiveness of our proposed query-aware image presentation method, we compare it with the query independent unified image presentation method as described earlier, i.e., all the images are presented by the combination of bag-of-visual-words and global features. The result is presented in Figure 10, which shows that our query-aware image presentation is better than query-independent image presentation approach, even though both of them are based on our same reranking principles. The inial ranking performance reflects lower search performance. This is because the news titles generally contain some redundant terms, which overwhelm the key concepts and potentially confuse the search engines.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a scheme to rerank web images for complex queries, which is robust to the unreliable

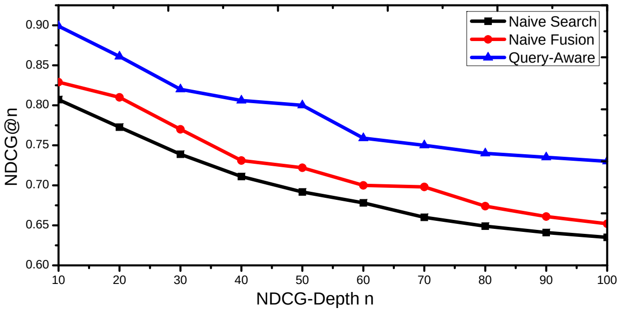

Figure 9: The average performance comparison of Photo-based QA System.

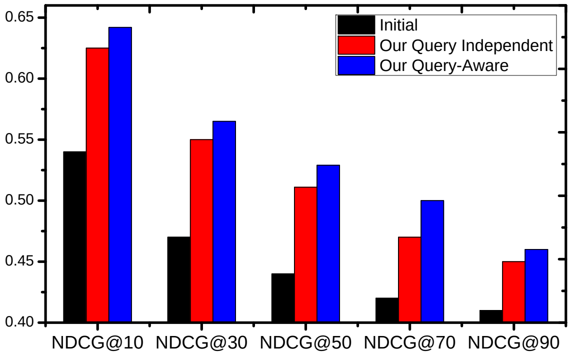

Figure 10: Performance comparison among different methods for textual news visualization.

initial ranking list. For a given complex query, it first detects the noun-phrase based visual concepts and collects their relevant images simultaneously. It then constructs a heterogeneous probabilistic network to estimate the image relevance score, which consists of three mutual reinforced sub-networks. These sub-networks represent different relationship layers, spanning from semantic level to visual level, which are established among the complex query and its detected visual concepts, by harnessing the content of images and their associated textual information. Based on these relevance scores, a new ranking list is generated. The experimental results showed that our scheme is significantly better than the other existing state-of-the-art approaches. We also introduced two application scenarios, which can benefit from our scheme, namely photo-based question answering and textual news visualization.

A limitation of current work is that it ignores the relationships explicitly described by the complex query, which have no uniform patterns and are notoriously hard to model. We will integrate this kind information into our scheme for general complex queries in our future work.

9. ACKNOWLEDGEMENTS

This work was supported by NUS-Tsinghua Extreme Search project under the grant number: R-252-300-001-490.

10. REFERENCES

[1] Hitwise. see `http://weblogs.hitwise.com/alan-long/2009/11/searches_getting_longer.html`.

[2] R. Aly, D. Hiemstra, and R. Ordelman. Building detectors to support searches on combined semantic concepts. In *MIR Workshop*, 2007.

[3] J. Bai, D. Song, P. Bruza, J.-Y. Nie, and G. Cao. Query expansion using term relationships in language models for information retrieval. In *CIKM*, 2005.

[4] N. Balasubramanian, G. Kumaran, and V. Carvalho. Exploring reductions for long web queries. In *SIGIR*, 2010.

[5] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *SIGIR*, 2008.

[6] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In *WSDM*, 2010.

[7] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *SIGIR*, 2011.

[8] X. Chen, J. Yuan, L. Nie, Z.-j. Zha, S. Yan, and T.-s. Chua. Known-item search by nus. In *NIST TRECVID*, 2010.

[9] T.-S. Chua, R. Hong, G. Li, and J. Tang. From text question-answering to multimedia qa on web-scale media resources. In *LS-MMRM*, 2009.

[10] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *TKDE*, 2007.

[11] D. Delgado, J. Magalhaes, and N. Correia. Assisted news reading with automated illustration. In *MM*, 2010.

[12] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *MM*, 2007.

[13] G. Kumaran and J. Allan. A case for shorter queries, and helping users create them. In *NAACL-HLT*, 2007.

[14] G. Kumaran and J. Allan. Effective and efficient user interaction for long queries. In *SIGIR*, 2008.

[15] M. Lease, J. Allan, and W. B. Croft. Regression rank: Learning to meet the opportunity of descriptive queries. In *ECIR*, 2009.

[16] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Harvesting social images for bi-concept search. *TMM*, 2012.

[17] Z. Li, M. Wang, J. Liu, C. Xu, and H. Lu. News contextualization with geographic and visual information. In *MM*, 2011.

[18] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *MM*, 2010.

[19] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *WWW*, 2009.

[20] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *ICME*, 2008.

[21] M. C. Mihai Surdeanu and H. Zaragoza. Learning to rank answers on large online qa collections. In *ACL*, 2008.

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

[23] A. P. Natsev, M. R. Naphade, and J. TešiĆ. Learning the semantics of multimedia queries and concepts from a small number of examples. In *MM*, 2005.

[24] L. Nie, M. Wang, Z.-j. Zha, G. Li, and T.-S. Chua. Multimedia answering: enriching text qa with media information. In *SIGIR*, 2011.

[25] L. N. Nie, M. Wang, Z. Zha, and T.-S. Chua. Oracle in image search: A content-based approach to performance prediction. *TOIS*, 2012.

[26] J. H. Park and W. B. Croft. Query term ranking based on dependency parsing of verbose queries. In *SIGIR*, 2010.

[27] S. U. Pillai, T. Suel, and S. Cha. The Perron-Frobenius theorem: some of its applications. *Signal Processing Magazine, IEEE*, 2005.

[28] H.-T. Pu. An analysis of failed queries for web image retrieval. *Journal of Information Science*, 2008.

[29] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. *NIPS*, 2002.

[30] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *MM*, 2008.

[31] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 2008.

[32] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards a relevant and diverse search of social images. *TMM*, 2010.

[33] H. Xu, J. Wang, X.-S. Hua, and S. Li. Image search by concept map. In *SIGIR*, 2010.

[34] R. Yan, E. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *ICVR*, 2003.

[35] J. Yuan, Z.-J. Zha, Y.-T. Zheng, W. Meng, X. Zhou, and T.-S. Chua. Utilizing related samples to enhance interactive concept-based video search. *TMM*, 2011.

[36] J. Yuan, Z.-J. Zha, Y.-T. Zheng, M. Wang, X. Zhou, and T.-S. Chua. Learning concept bundles for video search with complex queries. In *MM*, 2011.