Topic Based Query Suggestions for Video Search

Kong-Wah Wan[1], Ah-Hwee Tan[2], Joo-Hwee Lim[1], and Liang-Tien Chia[2]

[1] Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632
{kongwah,joohwee}@i2r.a-star.edu.sg
[2] School of Computer Engineering, Nanyang Technological University, Singapore
{asahtan,asltchia}@ntu.edu.sg

Abstract. Query suggestion is an assistive technology mechanism commonly used in search engines to enable a user to formulate their search queries by predicting or completing the next few query words that the user is likely to type. In most implementations, the suggestions are mined from query log and use some simple measure of query similarity such as query frequency or lexicographical matching. In this paper, we propose an alternative method of presenting query suggestions by their thematic topics. Our method adopts a document-centric approach to mine topics in the corpus, and does not require the availability of a query log. The heart of our algorithm is a probabilistic topic model that assumes that topics are multinomial distributions of words, and jointly learns the co-occurrence of textual words and the visual information in the video stream. Empirical results show that this alternate way of organizing query suggestions can better elucidate the high level query intent, and more effectively help a user meet his information need.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Query Suggestion.

1 Introduction

Query formulation has long been a critical component in information retrieval, and most modern search engines now have mechanisms to help users refine their queries. One such mechanism is the auto-query completion that automatically suggests possible complete queries as a user type into the query box [1]. However, most existing query suggestion methods suffer from the following two problems:

First, the query suggestions in most systems are obtained by utilizing a large-scale query log [2,3]. These methods typically work by looking at the clickthrough patterns of similar queries. However, the applicability of such methods for customized search domains such as personal desktop search or intranet search is severely limited. This is because often times there is either no available query log, or the size of the user base and the number of past queries is too small for any meaningful query mining.

Second, query suggestions are mostly obtained using simple lexicographical matches and then ranked by their frequencies. The resulting suggestions often appear ad-hoc and disorganized. An example is shown on the left in figure 1. Our

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 288–299, 2012.

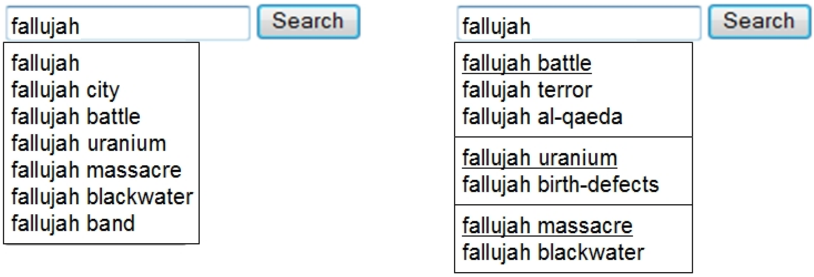

Fig. 1. Left: Existing query suggestions; Right: Presenting suggestions by thematic topics: three topics relevant to the query "fallujah" are shown, each titled and underscored by the highest probability word in the discovered topic

approach is to identify the *topics* among the documents that are relevant to the original query, and to group the query suggestions according to these topics. This new layout is shown on the right in figure 1. Note that query suggestions are now clustered into topics that are labeled by a keyword for easy navigation. The query suggestion "fallujah birth-defects" is also a new suggestion that gets presented as being part of the "fallujah uranium" topic. This allows a user to uncover possible new query interests that were hidden previously. The new presentation is clearly an improvement over the original.

Motivated by the above observations, in this paper, we propose a document-centric topic-based query suggestion method for the video search domain. We base our topic-elucidation methodology on the latent dirichlet allocation (LDA), which aims to uncover the hidden thematic structures in documents [4]. Using LDA to model query suggestions offers the following two advantages: (a). the relationships among the observed words, documents and latent topics is based on a theoretically robust probabilistic framework. (b). the learned topics are multinomial distributions of words, from which the high probability words are natural candidates for query word suggestion.

1.1 Challenges for the Video Domain

Despite provened capable of mining semantic topics in text collections, the use of topic model for the video domain poses some new challenges. First, naively feeding the speech transcripts (usually from Automatic Speech Recognition, or ASR) of video as textual input to the topic model will likely not yield good results. This is because ASR transcripts are noisy (ASR word error rates are generally above 20%), whereas most successful application of LDA are reported on clean text (e.g. newswire and publications). Apart from the few sporadic work in [5], the utility of LDA to noisy text source remains suspect.

The second challenge concerns the quality of LDA output: LDA often produces word distributions that are coarse, with no apparent meaning amongst high probability words. This can degrade the quality of detected topics. The common approach to deal with this problem is to introduce side-information into the modeling [6,4]. In this paper, we explore using the visual information in the shot keyframes to constrain topic development. There are two motivating intuitions: First, video footages are often repeated for similar or related news stories, and hence are highly correlated with the spoken (ASR) words. Second, different topics of a query may use different sets of words, and the same set of recurring visual shots become a bridge between these words, allowing them to be learned as distinct topics. To compute recurring visual shots, we use the Near Duplicate Image (NDI) detection method of Chum *et al* [7]. We derive a Gibbs Sampling-based inference and parameter estimation algorithm to jointly account for the NDI shots and the ASR words as distinct but correlated observations.

The third challenge relates to how to choose words from a learned topic to label suggestions. The difficulty arises because the semantic *theme* of a learned topic is only *collectively* conveyed by the high probability words, without any preferential order. Hence, a judicious choice of words from these high probability words is still needed to yield meaningful suggestions.

1.2 Our Contributions

In this paper, our contributions are:

(a) We propose a topic-based query suggestion method that can effectively help users refine and formulate their queries
(b) We develop a variant of the latent dirichlet allocation to constrain the development of topics during the inference process. The variant works by jointly modeling the text and visual information of a video stream.
(c) We develop a way to select high probability words in a topic to form suggestions.
(d) We perform extensive evaluation of our approach using real-life datasets and user studies.

In the remaining of this paper, we first discuss related works in Section 2. Section 3 provides details of our joint topic model. Experimental results are presented in Section 4, before we conclude the paper in Section 5.

2 Related Works

Query Completion: Machine-predicted text was first used mainly as an assistive mechanism for the physically handicap, but in recent years, has also found widespread benefit for mobile web search [8]. Today, query completion is featured in all major search engines. However, the focus of most work is on deriving a set of semantically related queries [9], rather than clustering them by topics. Jain and Mishne [10] proposed clustering query suggestions by simple phrasal keywords. There are only a few works on query suggestions without query log. Bhatia *et al*proposed a probabilistic model for generating query suggestions from the corpus, again using phrasal keywords. However, the limitations of phrasal keywords as a text summary have been highlighted in [12].

Table 1. Sample expansion words given the words on the left

Condolezza	state, bush, secretary, security, stanford
Jintao	china, brazil, sino, economic, aids
Basketball	points, conference, group, match, nba

Topic Modeling: Topic models are first derived as multinomial distributions of unimodal text data, and the joint modeling of multiple data types such as visual and text is not straightforward. The authors in [13,14] model annotated images using the visual features and text annotations, for automatic annotation and retrieval respectively. Two ways of combining the two modalities are explored: feature concatenation and hierarchical modeling. The former treats the two modalities equally, while the latter first models each individually and then fuse them at a later stage. Our work in this paper differs from the above in that we perform joint modeling of visual features and text in the video domain. Our modeling granularity is coarser: our visual features are not at the local patch-level but rather at the keyframe-shot-level. This most closely resembles Wu *et al*'s video representation with visual shot duplicates [15].

Because topic models are unsupervised methods, often best results are obtained by incorporating a priori knowledge about the desired output (e.g. must-link constrains in clustering). Adding observations from cross-media types as a way to constrain topic modeling is proposed by several authors. Jain *et al* guide topic formation of news photo caption by correlating the names with a face recognizer [6]. Blei and Jordan describe an image annotation model to learn the correspondence between an image region and a word in the caption [4].

3 Our Method

3.1 Overall Framework

In this subsection, we present a topic model to hypothesize a set of K topics $\{f_1, f_2, ..., f_k\}$ in a video collection D. Our overall approach is shown in figure 2. Two feature extraction tracks act on an input video simultaneously to compute text and visual features. These are then jointly combined using a generative model to compute topics. We discuss each of these key modules in the following.

Text Features: While important keywords are generally detected by ASR, the presence of the many misrecognized words can result in erroneous topic formation. Apart from stemming and discarding rare words, we use the document expansion approach of Wan to alleviate the problem [17]. The main idea is to introduce additional words to form an expanded text document vector. These words are selected based on their high mutual information in a parallel news corpus. Such a corpus is readily obtainable since news content is widely available

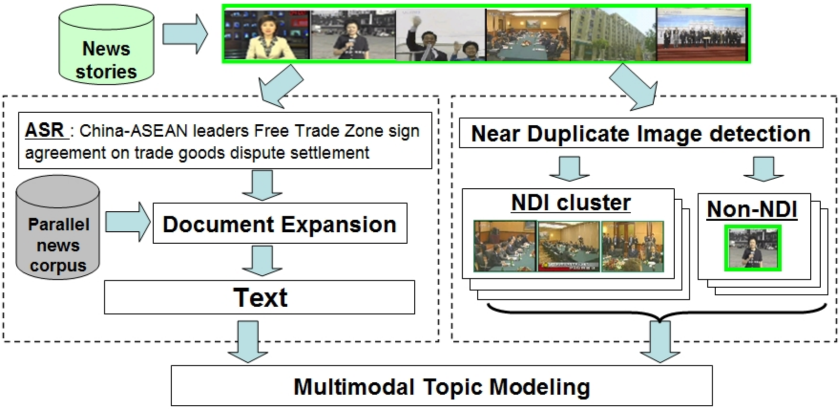
News stories
ASR : China-ASEAN leaders Free Trade Zone sign agreement on trade goods dispute settlement
Parallel news corpus
Document Expansion
Text
Near Duplicate Image detection
NDI cluster
Non-NDI
Multimodal Topic Modeling

Fig. 2. Multimodal topic modeling framework for hypothesizing topics in video

over the internet. For the TRECVID-2005 dataset used in this paper, we build the parallel corpus by issuing to Google Archives News our query description and restricting retrieval time-range to the period when the dataset was collected (Nov-Dec 2004). Table 1 shows some examples of expansion words.

Visual features: To capture a higher level of visual information, we generalize the common practice of modeling images as bag-of-features to represent video as *bag-of-keyframes*. Following the approach of Wu *et al* in [15], a keyframe is classified as whether it is a Near Duplicate Image (NDI) to other keyframe(s) or not. By assigning unique IDs to keyframes, they can be treated as visual words. All keyframes in a NDI-cluster are visually similar and are given the same ID. We can now generalize the TD-IDF weighting in text domain to these visual words. Figure 3 shows some examples of videos represented by the term frequencies (tf) and document frequencies (df) of visual words.

To compute near duplicates images, we use a color histogram combined with a spatial pyramid over the image to jointly encode global and local information. This approach is inspired by Chum *et al* [7], who applied the method to efficiently handle NDI detection amidst jitter and noise. Figure 4 shows the spatial pyramid configuration which is arranged so that an increasingly granular grid (i.e. from global to more localized) of color information is stored as we move up the level. The setup provides a highly compressed representation for each image that makes histogram comparison efficient. Given a query image, the NDIs are defined to be those within a specified Euclidean distance from the query. To compute a NDI cluster, we maintain a NDI list that initially only contains the query image. This list is then repeatedly populated with NDI of the new members in the list. The final NDI cluster is then given by the transitive closure of the NDI list.

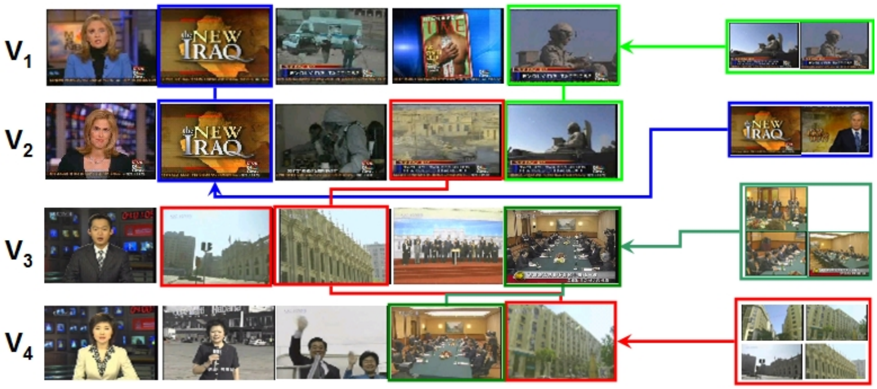

Fig. 3. TFIDF weighting of keyframe-based visual words. Within the 4 videos, four NDI clusters are shown and colored differently. For the red NDI cluster, it appears in V_2, V_3, V_4, hence its df=3, tf=2 for V_3, tf=1 for V_2, V_4. All non-NDI in a video have df=1 and tf=1. Best viewed in color.

3.2 Joint Modeling of Visual and Text

We now consider a corpus **D** of news video, each comprising of text **W** and visual words **V**. Each video d is modeled as a mixture of latent topics, to *simultaneously* account for **W** and **V** as distinct set of observations. Our model is motivated by Blei and Jordan's Corr-LDA model for text and images [4], and also Jain *et al*'s People-LDA model in [6]. We call our model cLDA-VT, denoting the use of both visual and text modality. Figure 5 shows the graphical representation of cLDA-VT. The generative process of cLDA-VT is as follow:

- Draw a multinomial ϕ over K topics: $\phi \sim \text{Dir}(\alpha)$
- For each topic $k = 1 \ldots K$,
  - draw multinomial $\theta_k \sim \text{Dir}(\eta_w)$ for text words
  - draw multinomial $\gamma_k \sim \text{Dir}(\eta_v)$ for visual words
- For each text word index n in d, $n = 1$ to N_d
  - Sample a topic z_n from ϕ: $z_n \sim \text{Multinomial}(\phi)$
  - Sample a text word w_n from θ_{z_n}
- For each visual word index m in d, $m = 1$ to M_d
  - Sample a topic y_m from ϕ: $y_m \sim \text{Multinomial}(\phi)$
  - Sample a visual word v_m from γ_{y_m}

where N_d and M_d is respectively the number of text words and visual words in video d, and η_w and η_v are Dirichlet priors for the text and visual words distribution respectively. The above cLDA-VT model results in the following joint distribution on text **W**, visual **V** and the latent topics:

$$P(\mathbf{W}, \mathbf{V}, \phi, \mathbf{z}, \mathbf{y}) = P(\phi|\alpha)\Big(\prod_{n=1}^{N_d} P(z_n|\phi)P(w_n|z_n, \theta)\Big) \Big(\prod_{m=1}^{M_d} P(y_m|\phi)P(v_m|y_m, \gamma)\Big) \quad (1)$$

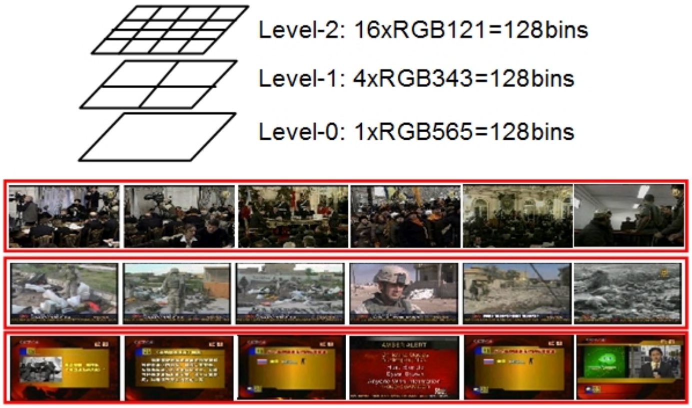
Level-2: 16xRGB121=128bins
Level-1: 4xRGB343=128bins
Level-0: 1xRGB565=128bins

Fig. 4. Top shows the spatial division of the image at each level of the RGB pyramid. At each level, the RGB suffix refers to the number of bits in the color channel. E.g. at level-1, we have 4 divisions each with 3 bits red, 4 bits green, 3 bits blue, totaling 128 bins. Each bin count uses 2 bytes. Therefore, each image is represented by 768 bytes. Bottom shows sample NDI given the leftmost image as query.

The main difference of our model from the Corr-LDA in Blei and Jordan [4] is that we use two multinomial distributions for the text and visual words. The sampling of visual words is essentially the same as the sampling text words. However, within a video, ϕ is a higher-level factor that is held fixed and it governs the ensemble of all text word and visual word observations. The topic-word multinomial θ and γ now learns the combined co-occurrence of important text words across video documents and also the complementary visual words.

Several approaches to learning the cLDA-VT parameters exist in the literature, such as Variational inference [4] and Gibbs Sampling [18]. We choose a simple extension of the latter by iterating over each text word, visual word and video document, each time resampling a single topic of the word (text or visual) based on the current topic assignment for the document and all other observed words (text and visual). A perplexity measure on a held-out set is used to determine learning convergence. On the 1028 TRECVID-2005 video documents comprising of 210K text words and 95K visual words, our implementation on a standard 3Ghz PC takes about 10 minutes to compute. We noticed that varying the Dirichlet priors η and α did not affect performance too much. We used the same value of 0.2 in the experiments below.

3.3 Selecting Topics and Terms for Query Suggestions

After the cLDA-VT learning has converged, the K latent topic distributions of both the text words and visual words are given by θ_k and γ_k, k=1..K. In particular, the probability of a text-word w in the kth latent topic is given by $P(w|\theta_k)$. We now make the assumption that the text multinomial θ_k represents the kth topic f_k in $\mathbf{D}$.

Topic Selection: From the K topics $\{f_1, f_2, ..., f_K\}$, we select a few as suggestion clusters. We can do this by a ranking approach, where the kth topic f_k is ranked by its relevance to the query as follow. Given a query Q, first note that $p(f_k|Q) \propto p(Q|f_k)p(f_k)$. By assuming $p(f_k)$ are uniformly distributed and query words occur independently, we can write

$$p(f_k|Q) \propto p(Q|f_k) = \prod_{q \in Q} p(q|f_k). \tag{2}$$

We select the top S topics with highest $p(f_k|Q)$ as suggestion clusters.

Terms Selection: Next, from each of the top S topics, we wish to select a few query suggestions as representative queries of the topic. For example, for the "fallujah battle" topic in figure 1, we have selected three suggestions: "battle", "terror" and "al-qaeda". While each of these suggestions is anchored on the "battle" topic of the query "fallujah", they also bear distinctive aspects within the topic. Our proposed way to achieve this is to select from amongst the high probability words candidate terms that are "compatible" with the current topic. By compatibility, we mean the following: suppose a candidate term t is also associated with a multinomial distribution of words $p(t|C)$, where C is a parallel context corpus. Then we can compare this multinomial distribution with the topic multinomial word distribution f_k using a suitable distribution measure such as the KL divergence:

$$\begin{aligned}\mathbb{KL}(t, f_k) &= -\sum_w p(w|f_k) \log \frac{p(w|f_k)}{p(w|t, C)} \\ &= \sum_w p(w|f_k) \log \frac{p(w, t|C)}{p(t|C)p(w|f_k)} \\ &\propto \sum_w p(w|f_k) \mathrm{PMI}(w, t|C) \end{aligned} \tag{3}$$

where PMI is the *pointwise mutual information* between the candidate term t and the terms in the topic model over the context corpus C. The PMI of two words is usually used as a measure of the semantic relationship between them. Intuitively, equation 3 assigns greater weights to a candidate term if it has a stronger semantic relationship to the important topic words. Hence, selected candidate terms are better representative of the entire topic f_k.

4 Experiments

To evaluate our methods, we used a subset of the TRECVID-2005 dataset [19]. It comprises of about 127 hours of Chinese and English news broadcast from 5 different sources (e.g. CCTV4, CNN). The dataset includes computed story boundaries from CMU Informedia. For queries, we used the annotated set of 33 queries from Wu *et al* [15]. Sample queries include "Bush visits Canada", "Mideast Peace", "Arafat health". For the full list of queries, refer to [15].

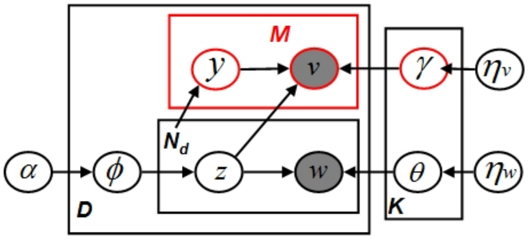

Fig. 5. Graphical representation of cLDA-VT. The red box encloses the additional observations from the visual modality, which constrain the topic formation in the text modality.

4.1 Perplexity Comparison

A commonly used quantitative evaluation of a topic model is how well it can predict the words in a test-set W_{test} after learning from a train-set W_{train}: that is, we are interested in which model provide a better predictive distribution $p(w \in W_{test}|w \in W_{train})$. We adopt the popular *perplexity* measure [4]:

$$\text{Perplexity}(\Psi) = \left(\prod_{d=1}^{D} \prod_{w \in W_{test}} p(w|\Psi, w \in W_{train}) \right)^{\frac{-1}{\sum_{d=1}^{D}(|doc_d|-|W_{train}|)}} , \qquad (4)$$

where Ψ denotes the model parameters of a learned LDA or cLDA-VT model. Mathematically, the perplexity of a word distribution is defined as the inverse of the per-word geometric average of the probability of the observations. Informally, it can be thought of as the effective number of equally likely words according to the model. Note that a lower number denote more predictive power. Figure 6 compares the predictive perplexity obtained by the regular LDA and cLDA-VT. In the former, only the ASR words are fed to the inference process. As can be seen from the graph, there is less uncertainty in the cLDA-VT than in LDA. This indicates that the topics learned in cLDA-VT are more robust and descriptive.

4.2 Qualitative Inspection

We show sample topic-based query suggestions in Table 2. In Table 3, we qualitatively show how the text-words multinomial θ_k has benefited from the inclusion of visual features during topic learning. On each of two queries "War on Fallujah" and "Mideast peace", we introspectively pick a learned cLDA-VT topic that contains high probability words that are meaningful to the queries. Observe that words from multimodal model are more intuitive and correlate better to the query topic.

4.3 User Study

We are also interested to see if users would find it beneficial to be provided with our query suggestions. To do this, we conducted a small-scale user study.

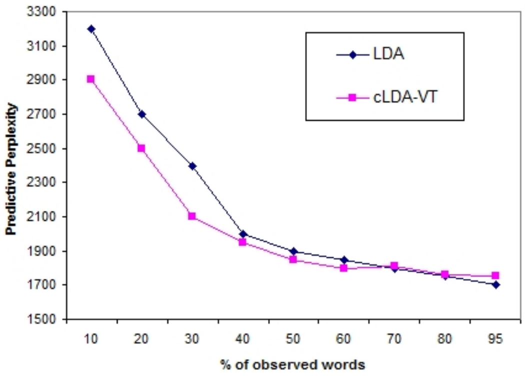

Fig. 6. Predictive Perplexity plot of LDA and cLDA-VT on a held-out set from the TRECVID-2005 video dataset. Lower is better.

Table 2. Sample topic-based query suggestions. The first line is the original query.

Fallujah	Middle-east	George Bush
fallujah battle	Middle-east Yasser Arafat	George Bush visits
fallujah terror	Middle-east Palestinian	George Bush canada
fallujah al-qaeda	Middle-east Israel	George Bush falungong
fallujah uranium	Middle-east peace	George Bush protests
fallujah birth-defects	Middle-east minister	George Bush pentagon
fallujah massacre	Middle-east election	George Bush rice
fallujah blackwater	Middle-east leader	George Bush cheney

Table 3. Top-10 probability words of 2 learned cLDA-VT θ_k topics, for "War on Fallujah" (top) and "Mideast peace" (bottom). The topics are picked introspectively. Meaningful words highlighted in bold.

LDA (ASR-text-only)	cLDA-VT (Visual+text)
iraq people time good meet **united states baghdad** chinese	**iraq iraqi** people **military** govern **arm kill force attack baghdad**
arafat know **yasser** leader thing **peace** minister people just **israel**	**palestinian arafat peace israel** president leader election **yasser east middle**

We asked three subjects to look at two versions of suggestions displayed for all 33 queries. The first version uses a simple lexicographical match and spelling corrector to find the top ten similar queries as query suggestions. They form the non-clustered suggestions as are currently obtained on most search engines. The second version uses the results from our automatic topic-based query suggestions. The subjects were then asked to put down on a scale of one to five their preference liking to both methods. We consider a method to be significantly better or worse liked for a user if the difference in his rating on the two methods is equal or more

Table 4. User preferences for query suggestions. See text for details.

Category	Percentage of time
Topic-based clustered suggestion significantly better	30
Topic-based clustered suggestion marginally better	22
No preference	28
Unclustered suggestions marginally better	18
Unclustered suggestions significantly better	2

than two, and marginally better or worse liked if the rating difference is one. The results are shown in Table 4. They indicate that our topic-based query suggestion method has potential value.

5 Conclusion and Future Works

Because the average length of a typical query is only two or three words long, most search engines are faced with a difficult task of discerning a user's search intent. Query auto-completion is an important mechanism to ameliorate the problem, and to facilitate a user in finding his information need. However, most existing query suggestion methods are based on simple lexicographical matching, and suggestions are placed in an ad-hoc manner. In this paper, we propose an alternative form of query suggestions that presents to users a list of possible complete queries grouped by their thematic topics. We argue for the benefits of such an arrangement, and demonstrate empirical results that show a high level of user acceptance and satisfaction with the idea.

As opposed to current query suggestion techniques, our method is document-centric and do not require a query log. Such a platform is particularly helpful when the user is not aware of how to phrase his query, or when the query words he chooses are not found in the documents. By mining topics on the documents directly, the resulting query suggestions are guaranteed to be found in the documents, making the suggestive framework useful not only for predicting what the user is likely to type, but also for uncovering new queries that may be of interest to the user.

Several aspects of the current implementation will be considered for future works. One relates to the run-time efficiency. Another pertains to the use of more sophisticated methods for query suggestion term selection.

Acknowledgments. The authors would like to express appreciation for the labeling help rendered by the three students from Temasek Polytechnic.

References

1. Feuer, A., Savev, S., Aslam, J.A.: Evaluation of phrasal query suggestions. In: Proc. ACM Conference on Information and Knowledge Management, pp. 841–848 (2007)

2. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query Recommendation Using Query Logs in Search Engines. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 588–596. Springer, Heidelberg (2004)
3. Boldi, P., Bonchi, F., Castillo, C., Donato, D., Vigna, S.: Query suggestions using query-flow graphs. In: Proc. Workshop on Web Search Click Data, pp. 56–63 (2009)
4. Blei, D., Jordan, M.: Modeling annotated data. In: Proc. ACM Conference on Research and Development in Information Retrieval, pp. 127–134 (2003)
5. Cao, J., Li, J., Zhang, Y., Tang, S.: LDA-based retrieval framework for semantic news video retrieval. In: Proc. International Conference on Semantic Computing, pp. 155–160 (2007)
6. Jain, V., Learned-Miller, E., McCallum, A.: People-LDA: Anchoring topics to people using face recognition. In: Proc. International Conference on Computer Vision (2007)
7. Chum, O., Philbin, J., Isard, M., Zisserman, A.: Scalable near identical image and shot detection. In: Proc. ACM International Conference on Image and Video Retrieval, pp. 549–556 (2007)
8. Church, K., Thiesson, B.: The wild thing! In: Proc. Association for Computational Linguistics (2005)
9. Zhang, Z., Nasraoui, O.: Mining search engine query logs for query recommendation. In: Proc. International World Wide Web Conference (2006)
10. Jain, A., Mishne, G.: Organizing query completions for web search. In: Proc ACM International Conference on Information and Knowledge Management, pp. 1169–1178 (2010)
11. Bhatia, S., Majumdar, D., Mitra, P.: Query Suggestions in the Absence of Query Logs. In: Proc. ACM Conference on Research and Development in Information Retrieval (2011)
12. Madnani, N., Dorr, B.: Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. Journal of Computational Linguistics 36(3), 341–387 (2010)
13. Monay, F., Gatica-perez, D.: Modeling semantic aspects for cross-media image retrieval. IEEE PAMI 29, 1802–1817 (2007)
14. Lienhart, R., Romberg, S., Horster, E.: Multilayer plsa for multimodal image retrieval. In: Proc. ACM International Conference on Image and Video Retrieval, pp. 1–18 (2009)
15. Wu, X., Hauptmann, A., Ngo, C.: Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In: Proc. ACM Multimedia, pp. 168–177 (2007)
16. Cutting, D., Karger, D., Pedersen, J., Tukey, J.: Scatter/Gather: A cluster-based approach to browsing large document collections. In: Proc. ACM Conference on Research and Development in Information Retrieval, pp. 318–329 (1992)
17. Wan, K.: Exploiting story-level context to improve video search. In: Proc. Internatonal Conference on Multimedia and Expo. (2008)
18. Griffiths, T., Steyvers, M.: Finding scientific topics. Proc. National Academy Science U.S.A 101 (Supp. 1), 5228–5235 (2004)
19. TRECVID (2005), `http://www-nlpir.nist.gov/projects/trecvid`