Relevance feedback in probabilistic multimedia retrieval

Lioudmila Boldareva[†], Djoerd Hiemstra[†], Willem Jonker[†,‡]
† Database Group, University of Twente, The Netherlands,
{*L.Boldareva, hiemstra*}*cs.utwente.nl*
‡ Philips Research Eindhoven, The Netherlands, *willem.jonker@philips.com*

ABSTRACT

In this paper we propose a new method for data organisation in a (multimedia) collection. We use probabilistic approaches to indexing and interactive retrieval which enable to fill the semantic gap. Semi-automated experiments with TREC data showed that our approach is efficient and effective.

1. INTRODUCTION

In content-based information retrieval, there is a gap between the semantics of the document as percepted by a human and its actual representation at the lowest level in the storage. Careful selection of the features that represent the documents in the storage allows capture the semantics, especially in limited domains where the range of possible values is known.

A significant improvement of the performance of content-based retrieval systems can be achieved by using *relevance feedback*, a technique that allows the user to rate the (intermediary) search results [3, 12, 14, 2, 5]. The ranking and retrieval of objects from the collection is based on the feedback received from the user. In the domain of image retrieval, where the *semantic gap* is especially large, relevance feedback is often the only means to help the user with his/her information need.

In the present paper we approach the problem of content-based image indexing and retrieval from another perspective. Instead of multidimensional feature space we propose using a network of precomputed probabilistic similarity values. We call this network *multimedia dictionary*. It encodes higher-level context and makes use of multiple relevance feedback obtained from users. A probabilistic method is used at indexing time to construct the initial meta-data. The advantage of the proposed method is that in the end the system does not rely on physical-level features, but on the *common sense* of the many users. At the same time the data is organised in such a way that it can be extended to a larger collection quite easily.

We study various approaches to retrieval with relevance feedback in the light of our data organisation.

The rest of the paper is organised as follows: in Section 2 we briefly describe our approach to multimedia indexing and retrieval. The experiments and their results are described in Section 3. Finally, 4 contains discussion and further research directions.

2. BAYESIAN RETRIEVAL FRAMEWORK

2.1 Data organization

Consider a collection $\Im$ of objects i among which there is an object that the user is looking for — the target T. In the search session the user retrieves a set of candidate objects on the screen and feeds back to the system his/her opinion about their relevance to the target. Each object might look like the target the user has in mind, and then it is selected by the user as *relevant*. It is selected as *non-relevant*, if it doesn't resemble the target. For a relevant candidate object i we denote the event as $(\delta_i = 1)$, and for a non-relevant ones as $(\delta_i = 0)$. The feedback obtained from the user allows the system to infer the possible target and compose a new set of n elements, *the display set*, to show next. There may be several such iterations during one search session.

To make use of notions "relevant" and "non-relevant" for objects, it is necessary to organize the collection by introducing relations between the objects. We introduce a "measure of closeness" of an object i to an object x as a probability and denote it as $P(\delta_i|T = x)$.

DEF. 1. $P(\delta_i|T = x)$ is the probability of an object i being selected by the user on the condition that another object x is the target of the search.

The user's judgement about the relevance of objects is a necessary component of our system. It is reasonable to assume that $P(\delta_x|T = x) \equiv 1$, i.e. the user always identifies the target as relevant. We also put a constraint that the target exists in the collection and is unique: $P(T = j|T = x) \equiv 0$, $j \neq x$, $\sum_{i \in \Im} P(T = i) = 1$.

The data as we organise it is a bi-directional weighted graph with objects $i \in \Im$ as nodes and arcs $P(\delta_j|T = x)$ going from x to j. This graph has properties of a monolingual dictionary (or the WordNet system, "a lexical database for the English Language" [4]), where each article contains (several) entries with synonyms for the word, and often the most important antonyms. In the search process the meaning of the unknown word (the user's target) can be identified by looking at the article entries in the dictionary (entries that are relevant for the user). Two words are considered synonyms not when they have similar spelling, or when they appear in the same article, but when their dictionary articles contain many common entries.

In this way, each element in the collection can be *described* by a number of other elements linked to it, which, in turn, are linked to other elements. These associations that come from users judgements and refer to the hidden semantics of objects, serve as meta-data for the collection. The collection describes itself with the help of meaningful relations

observed in earlier retrieval sessions. Such representation of the multimedia dictionary (or: MD) refers to an object as a whole, and lower-level features do not explicitly play an important role.

The nodes in the MD need not be images only. Other types of media, such as video, audio or speech transcripts can be plugged into it as separate nodes. Note however, that integrating other types of media is not trivial. Our data organisation relies on multiple user interactions. Dynamic media such as video or audio may not stand many feedback loops, since assessing a video clip or a music fragment requires from the user more efforts and time compared to still images. Nevertheless, such nodes may be potential targets or, conversely, the starting points in a search session. Textual nodes are of particular interest for a retrieval system, since querying in the form of text is very convenient for the user.

2.2 Retrieval during the search session

We assume that (1) the user is consistent in his/her judgements; (2) does not forget what the target is; (3) the target object is unique and exists in the collection. The uniqueness of the target is observed with queries like "find me an image of a Golden Retriever puppy". Queries like "find me all pictures of Britney Spears" are not handled by the model directly. However, there is a way to retrieve ranked lists of objects most relevant in the context of current search session, which may be considered targets. In our framework we use the following definition of the target:

DEF. 2. *Target* is the object, after retrieval of which the user terminates the search successfully.

The goal of a retrieval system is to help the user find the target object (and possibly all similar objects) after few iterations, with a reasonably small amount of time spent on each round.

Probabilistic methods in information retrieval were developed for text collections [9, 11, 6]. The ideas were adapted to image retrieval [12, 3, 5]. In the content-based retrieval framework the user's information need is unknown and should be guessed. In general, retrieval with the use of relevance feedback can be formulated as follows:

> In the current data structure, having observed the user judgements in the search process, what is the object that the user wants to find?

We use the Bayes' rule to answer the question above and reformulate it as predicting the user's action of selecting/ deselecting relevant objects, given the target that he/she has in mind:

$$P(T = x, U|\delta^1 \dots \delta^n) = \frac{P(\delta^1 \dots \delta^n|T = x)P(T = x|U)P(U)}{P(\delta^1 \dots \delta^n)} \quad (1)$$

where U denotes the current user. Since we assume that the state of the (unknown) user variable does not change during one search session, and U affects $\delta^{(\cdot)}$ through T, we may omit the user notation in further formulae, to keep the notation short. The upper index in $\delta^1 \dots \delta^n$ denotes n displayed objects that either received positive ($\delta^s = 1$) or negative feedback ($\delta^s = 0$); $P(T = x)$ is the probability that the object x is the target, and $P(\delta^s|T = x)$ is the probability of a s-th object on the screen to be selected by the user given that x is his/her target.

Note that equation (1) is regarded as recursive, i.e. the posterior probability of being the target determined at step no. s as $P(T = x|\delta^1, \dots, \delta^n)$ serves as the prior $P(T = x)$ at the next iteration $s+1$. In each round the observed user response gives new distribution of $P(T = x)$. In the beginning, before any information from the user is received, each object has a certain prior probability to be the target[1]. The possible output of incorporated primary textual query or previous search sessions results may be used to define the prior value of $P(T = x)$ more accurately.

To determine the first term in the numerator of equation (1) we assume (for the time being) that given the target, the user picks each of n candidates independently of other objects on the screen. This assumption is similar to (conditional) term independence assumption used in text retrieval. Then equation (1) becomes

$$P(T = x|\delta^1, \dots, \delta^n) = \frac{\prod_{s=1}^{n} P(\delta^s|T = x)P(T = x)}{P(\delta^1, \dots, \delta^n)}. \quad (2)$$

Using the input from the user to change the distribution of $P(T = x)$ we want to achieve the state when all elements but the target have zero probability to be the user's information need. We *learn* this distribution from the relevance judgements provided by the user.

After T is initiated, i.e. the target of the search is identified, some conditional probabilities stored in the MD can be updated. The information obtained from a given retrieval session is saved to be used for *long-term learning*. The purpose of the multimedia dictionary update is to increase the probability to be selected by the user $P(\delta^s|T = x)$ with respect to the (found) target, for all objects that the user had indeed selected. At the same time the links to the objects that have been marked by the user as non-relevant, may be punished.

To initiate the contents of MD, the system uses lower-level features. In principle, the multimedia dictionary may be initiated quite arbitrarily and thus solely depend on learning from the relevance feedback, but we believe that pictorial features do contain valuable information. We assume that for a large number of objects in the collection the following holds:

$$P(\delta^s|T = x) = P(\delta^s), \quad (3)$$

i.e. the user cannot put it definite whether the object on the screen is relevant to his/her information need or not. These *uncertain* links are not stored, since they contain almost no information about the relevance of the objects to each other.

In the future we would like to receive some evidence about the user model, which may affect the update strategy, and the prior distribution. However, a simple assumption about the user who wants to find the target and responses consistently, can serve as a *generic* user.

3. EXPERIMENTAL EVALUATION

3.1 Design

We model the search process such that the user does not have to provide any example to his/her information need (which seems reasonable, for retrieving an example could

[1]Often equal prior probabilities are assigned to all elements of the collection. The importance of selecting the "good" priors is studied in, e.g. [7].

be a problem itself). The search process imitates random browsing through the collection, until the target is identified. We try to model the information need by learning from the user's actions and express it in terms of $P(T = x)$.

There are several parameters that affect the search. In the light of learning the multimedia dictionary contents from successful search sessions, we consider the following questions important:

- The form in which the user can give his/her judgements about the relevance. Should it be only positive feedback option (i.e. *unary*), or positive/ negative/ neutral (*ternary*), when the user has to mark all relevant and non-relevant objects explicitly? What about assumed negative feedback, when everything not marked by the user as relevant is assumed non-relevant (*binary* feedback)?
- The optimal contents of the multimedia dictionary with respect to search quality and performance. Should we store only elements that are *very relevant* to each other, or also *very non-relevant*, too? What is the performance compared to the `MD` containing all connections, including those that satisfy equation (3)?

We performed several simulations testing the questions stated above. In this paper we do not focus on testing the display update strategy. The objects for the new display set were selected randomly from the collection. The display set consisted of 12 objects and we made sure that the same object does not appear for the feedback twice.

Selecting new objects for the feedback is an important part of interactive information retrieval, and needs further elaboration. The present experimental setup allows us investigate the ability of the system to learn from the user interaction, given our data organisation and different feedback schemas. Smart display update strategy may further improve the search quality.

3.2 Setup

The test collection was the data for the video search task of TREC-2002 [1]. The multimedia dictionary was initialised using Gaussian mixture models described in detail in [13]. As two measures of similarity, approximated Kullback-Leibler divergence and Bhattacharyya distance [8] were used. Both methods are suited to measure difference between two distributions. We used the approximations under the assumptions articulated in [10], and further discussed in [13]. The distances were transformed into conditional probabilities in accordance with our model. The "uncertain" values were removed from the multimedia dictionary which left only *tails* containing the "very relevant" and "very non-relevant" neighbours of objects in the collection. The size of the trimmed `MD` with both relevant and non-relevant tails counted up to 8.5% of the original full multimedia dictionary.

The search task consisted of 25 different queries, or *topics* provided by TREC. As the basis for the simulated feedback, we used the relevance judgements that were gathered and summarised at TREC evaluation event. In the experiments that we performed the shots marked "relevant"/"non-relevant" received positive and negative feedback, respectively. Other shots not listed as relevant/non-relevant received either neutral ("don't care") or negative feedback depending on the schema used.

The advantage of such automated system is that the relevance judgements are the same for different setups, and the same target may be retrieved many times by different versions of the system. This stability of the feedback is not easily achieved when using humans in the experiments. Since the judgments are collected from different people, we considered them as feedback obtained from a generic user. The relevance judgements were not specifically developed for our test system.

The feedback data that we had was quite sparse for a collection of 14,500 images (on average 0.4% of shots are judged as relevant). To reduce the scarcity we selected the fraction contained in relevance judgements and added about same number of random key frames from the collection. That yielded 2875 images with about 5% of positive feedback.

We treated the data as an *image* collection, whereas the judgements were made for the *video shots*. The key frames were selected based on their middle position of common shot boundaries, which does not always determine the *key* frame of a shot. As a result, some key frames were not only visually, but also semantically different from the shot's true contents. These "inconsistencies" carried a role of a reference to some hidden semantics that the user has in mind. This semantics does not agree with the current contents of the multimedia dictionary. Note that the feedback that is supplied in the relevance judgements is not complete, i.e. some (non-)relevant key frames that have not been in the result set are missing.

To evaluate the performance, we looked at *recall* in the first hundred elements ranked according to their probability to be the target $P(T = x)$. As the base line for the performance measurement we chose the level corresponding to the prior distribution of $P(T = x)$, when no judgements have been input to the system. The prior distribution in these experiments is determined by the number of closest/remote neighbours in the multimedia dictionary. Naturally, if we initiate the multimedia dictionary based on lower-level features, this baseline is at least as good as a $k-$nearest neighbour search based on those features.

We averaged the value of recall over all topics and visualised the results in the graphs Fig. 1. It is necessary to note that, strictly speaking, the probabilistic formulation that we used allows only one target in the collection. However, in accordance with the definition (2), any of the elements marked in the TREC data as "relevant" may be the user's target. Thus, the more relevant elements are found in the first hundred, the more chances that the user is satisfied with the search quality, the more quality gets the user.

3.3 Analysis

The results of the experiments are presented in Fig. 1. As it is seen in the figures, some methods but not all, are capable of learning from the user's feedback, even if the browsing through the database is performed randomly. First few iterations make no or little distinction between the methods, but as the models learn, the difference appears.

The thick grey line (marked "`[Y] full`") corresponds to the full `MD`. Although it took the longest to conclude the experiments, it is clearly not the winner.

In the winning part of the graph, the multimedia dictionary with only very relevant neighbours (marked "`[Y] 1 tail`") performed as good as its sibling with both very relevant and very non-relevant neighbours (marked with "`[Y]`

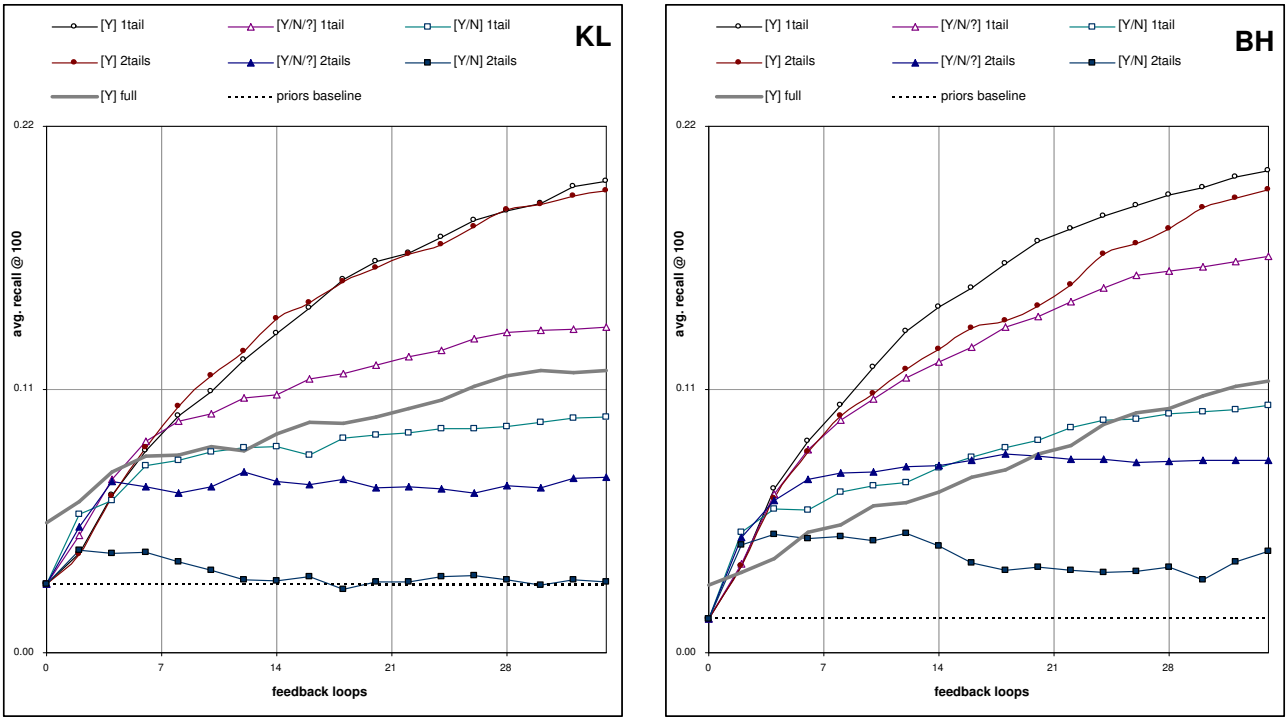

Figure 1: Recall for Kullback-Leibler divergence (a) and Bhattacharyya distance (b) based multimedia dictionary

`2 tails`").

To compare between the initialisation methods, the difference is not important for the trimmed `MD`. For a given initialisation method, Kullback-Leibler divergence or Bhattacharyya distance, if we look only at "`1 tail`" configuration, the recall curve becomes flatter with the increase of negative feedback fed to the system (maximum negative feedback is received in "`[Y/N]`" schema, and minimum in "`[Y]`" schema). If both very relevant and very non-relevant tails are considered, (marked "`2 tails`"), then the same effect is observed, but the corresponding recall is twice as low (on average 66% for Kullback-Leibler-based initialisation and 55% for the Bhattacharyya-based one). Schemas that did not use any negative feedback (marked with "`[Y]`") did not have this effect.

In general, when the display update is arbitrary, the more negative feedback is fed to the system, the worse is the average recall. The combination of binary feedback and both tails (marked with "`[Y/N] 2 tails`") did not show any improvement over iterations, and recall stayed at the lowest level.

Such dependencies mean that the true[2] relations of closest neighbours (the one indicated by relevant in TREC judgements) is in agreement with the multimedia dictionary. But the remote neighbours from the multimedia dictionary are contradicting the negative relevance judgements of the users, and these two pieces of information stand each other in the way. If the generic user at TREC is consistent and objective, then the multimedia dictionary is merely overtrained where it concerns the non-relevant elements according to pictorial features.

[2]i.e. the one that corresponds to the generic human perception

4. CONCLUSIONS

The results presented in this paper, suggest that the proposed method for content-based multimedia retrieval can be successfully used. The constructed multimedia dictionary that contains only about 8.5% of possible connections, performs much better than the full graph of `MD`, but the execution time for the latter is unacceptable. By trimming the `MD` we improved not only the efficiency of the method, but also the quality of the search. This suggests that data that was left out from the multimedia dictionary contained mostly noise.

Even with random browsing, a significant increase in recall can be achieved by using relevance feedback. Excessive negative feedback worsens the retrieval quality. Besides, the notion of non-relevant neighbours from the multimedia dictionary contradicts non-relevant judgements obtained from the generic TREC user. In the worst case there is no improvements over the iterations.

The initialisation of the `MD` can be done based on lower-level features e.g. Gaussian mixture models. Particular method of initialisation has secondary effect on the retrieval quality.

Further improvements in the quality will be training the initial multimedia dictionary to bring it in accordance with the "common sense" of the users. Applying a smarter display update schema enables better exploitation of the information that we can get from the user.

Acknowledgements

We thank Thijs Westerveld (CWI) for the Gaussian Mixture Models data that we obtained from him. We also thank Vojkan Mihajlovic (UT) and Arjen de Vries (CWI) for useful comments about the paper.

5. REFERENCES

[1] A.F.Smeaton and P.Over. The TREC-2002 video track report. In *The Eleventh Text Retrieval Conference TREC-2002*, pages 171–181, 2002.

[2] S. Aksoy and R. M. Haralick. Probabilistic vs. geometric similarity measure for image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition*, 06 2000.

[3] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Tran. On Image Processing*, 9(1):20–37, 2000.

[4] G.A.Miller. WordNet: an on-line lexical database. *Int. Journal of lexicography*, 4(3):235–312, 1990.

[5] D. Geman and R. Moquet. A stochastic feedback model for image retrieval. Technical report, Ecole Polytechnique, 91128 Palaiseau Cedex, France, 1999.

[6] D. Hiemstra. *Using language models for information retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[7] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34. ACM Press, 2002.

[8] C. Lee and D. Hong. Feature extraction using the bhattacharyya distance. In *IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 3, pages 2147–50, 1997.

[9] M. Maron and J. Kuhns. On relevance, probabilistic indexing, and information retrieval. *Journal of the ACM 7*, pages 216–244, 1960.

[10] N.M.Vasconcelos. *Bayesian Models for Visual Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[11] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[12] N. M. Vasconcelos and A. B. Lippman. Bayesian representations and learning mechanisms for content-based imagemark set retrieval. In M. M. Yeung, B.-L. Yeo, and C. A. Bouman, editors, *Proc. of SPIE: Storage and Retrieval for Media Databases 2000*, volume 3972, pages 43–54, 2000.

[13] T.Westerveld A.de Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *EURASIP special issue on Unstructured Information Mmanagement from Multimedia Data*, (to appear), 2003.

[14] P. Wu and B. S. Manjunath. Adaptive nearest neighbor search for relevance feedback in large image databases. In *Proc. of ACM International Multimedia Conference*, Ottawa, Canada, October 2001.