Topic modelling of clickthrough data in image search

Morrison, Donn Alexander; Tsikrika, Theodora; Hollink, Vera; de Vries, Arjen P.; Bruno, Eric; Marchand-Maillet, Stéphane

How to cite

MORRISON, Donn Alexander et al. Topic modelling of clickthrough data in image search. In: Multimedia Tools and Applications, 2013, vol. 66, n° 3, p. 493–515. doi: 10.1007/s11042-012-1038-8

This publication URL: https://archive-ouverte.unige.ch//unige:114847

Publication DOI: 10.1007/s11042-012-1038-8

Multimed Tools Appl (2013) 66:493–515
DOI 10.1007/s11042-012-1038-8

Topic modelling of clickthrough data in image search

Donn Morrison · Theodora Tsikrika · Vera Hollink · Arjen P. de Vries · Éric Bruno · Stéphane Marchand-Maillet

Published online: 16 March 2012

Abstract In this paper we explore the benefits of latent variable modelling of clickthrough data in the domain of image retrieval. Clicks in image search logs are regarded as implicit relevance judgements that express both user intent and important relations between selected documents. We posit that clickthrough data contains hidden topics and can be used to infer a lower dimensional latent space that can be subsequently employed to improve various aspects of the retrieval system. We use a subset of a clickthrough corpus from the image search portal of a news agency to evaluate several popular latent variable models in terms of their ability to model topics underlying queries. We demonstrate that latent variable modelling reveals underlying structure in clickthrough data and our results show that computing document similarities in the latent space improves retrieval effectiveness compared to computing similarities in the original query space. These results are compared with baselines using visual and textual features. We show performance substantially

D. Morrison (✉) · É. Bruno · S. Marchand-Maillet
Computer Vision and Multimedia Laboratory, University of Geneva, Geneva, Switzerland
e-mail: donn.morrison@gmail.com

É. Bruno
e-mail: eric.bruno@ymail.com

S. Marchand-Maillet
e-mail: stephane.marchand-maillet@unige.ch

T. Tsikrika · V. Hollink · A. P. de Vries
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

T. Tsikrika
e-mail: theodora.tsikrika@acm.org

V. Hollink
e-mail: V.Hollink@cwi.nl

A. P. de Vries
e-mail: arjen@acm.org

better than the visual baseline, which indicates that content-based image retrieval systems that do not exploit query logs could improve recall and precision by taking this historical data into account.

Keywords Image retrieval · Latent variable modelling · User interaction · Clickthrough data

1 Introduction

Users exploring search results make internal judgements on the relevance of the ranked documents. Based on their assessment of a document summary (document title, text snippet or image thumbnail), users may select or *click* on it to further assess whether the document satisfies their information need. These search interactions can be logged without any additional cognitive burden to users performing the queries. Such search logs have proven to be a useful resource for gaining an understanding of how users interact with search engines [17] and for improving retrieval performance in a variety of IR tasks [6, 19, 26, 33] by considering these clicks as weak indicators of relevance [19].

Image retrieval is one such domain where clicks have been used as a way to improve the overall search experience [6, 28, 33]. It has been demonstrated that these implicit relevance judgements are accurate to the extent that they can represent high quality groundtruth [6] or training labels for supervised learning [33]. In fact, research suggests that the quality of clickthrough data in image search in terms of relevance is significantly better than in text search [6, 30].

It is in this context that we position our work. We posit that clickthrough data contains hidden topics and can be used to infer a lower dimensional latent space that can be subsequently employed to improve various aspects of the retrieval system. These topics can be likened to search trends realised by users interacting with the retrieval system. Following this formulation, we expose ourselves to the large body of research available on latent variable and topic modelling in areas such as text retrieval [5, 7, 16], collaborative filtering [5, 22], and computer vision [24]. Latent variable models are attractive in these and other domains due to the fact that they are unsupervised, require little parameter tuning, and act as a way of reducing dimensionality in data [5, 7, 16]. However, they have not yet been fully explored on image clickthrough data.

By adapting latent variable models to this problem domain, we aim to uncover the latent structure in image search logs and use it for the task of image retrieval. Specifically, we investigate the following points:

1. how rankings derived from document similarities computed in the latent topic space perform compared to baseline rankings;
2. whether graded implicit relevance judgements, when available, are beneficial;
3. and whether a text-based pre-processing step of merging queries is beneficial.

To this end, we compare several latent variable modelling approaches: the singular value decomposition (SVD), non-negative matrix factorisation (NMF), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA). Outside of the text modelling domain, these methods are rarely explicitly compared.

We begin the paper by detailing related work in long-term learning based on (explicit) relevance feedback and clickthrough data in image search. Next, we present the setting for this study by introducing definitions in the problem domain, how the models are adapted for image click data, and the methodology followed. Our experimental setup is then introduced and a description of the datasets, the groundtruth, and evaluation method is given. The experimental results and findings are then discussed.

2 Related work

For some time, research [12, 13, 27] has shown that content-based image retrieval systems supporting relevance feedback can benefit from what is known as *long-term* or *inter-query learning*,[1] where relevance feedback interactions are logged and then periodically mined for latent structure to aid subsequent queries. These studies deal with search logs of *explicit* relevance feedback, where the interaction data stem from users purposefully rating examples (usually on a positive/negative dichotomous scale $\{-1, +1\}$) in order to iteratively improve the search results. *Implicit* relevance feedback [21], on the other hand, stems from users interacting with documents (e.g. clicking on a document to view it) without explicitly expressing their satisfaction/dissatisfaction with a particular search result.

Müller et al. [27] approached the problem using the principle of market basket analysis and found that *tf-idf* weights for visual features could be updated using relevance judgements. Latent semantic analysis (LSA) [7], also explored in our study as the singular value decomposition, has been previously applied on search logs of explicit relevance judgements by Heisterkamp [13] and by He et al. [12] as a means of building a "semantic space" in which to index images in a collection. As in our work, a document-query matrix representation of the relevance feedback is used with the images as rows and the queries as columns. However, our study differs in that we focus on implicit relevance judgements in clickthrough data.

Difficulty in procuring large amounts of real-world relevance judgements led many of these early studies to make use of artificially generated search logs [12, 13]. The rise of web-based search engines in the last decade has seen the scale of search logs increase dramatically as millions of users submit queries daily. With this rise, the nature of search interaction evolved from explicit to *implicit* relevance judgements due in part to the fact that traditional relevance feedback saw little use in web IR settings [18]. Important to this transition has been a study by Joachims et al. [20] where differences between implicit and explicit relevance judgements were shown to be less than originally thought. This paved the way for research to consider implicit relevance judgements as training data for various machine learning-based improvements to IR [19, 26, 33].

The advantages of using clickthrough data are numerous: it can be easily collected by content owners, its quality allows it to be used as training data in various tasks [19, 26, 33], and its collection introduces no additional cognitive burden on the part of users performing the queries due to the implicit nature of the interaction [20].

[1] Conversely, *short-term* or *intra-query* learning refers to learning based on relevance feedback judgements for the current query only.

However, there are some important disadvantages that should also be discussed. It is highly sparse because, for a given query, a user will only ever see a fraction of the relevant images, and only actually click on a fraction of that. The degree of sparsity though may vary depending on the search domain. For example, multimedia archives for news agencies may yield larger groups of visually similar images that have been clicked because their users are typically trying to find images with the best angle or framing. This is contrasted with web image search, where visually similar (but not identical) images are not typically found in large quantities. This sparsity implies that no benefits can be seen for documents that have never been clicked (often referred to as the "cold start" problem).

There are a number of ways in which sparsity is addressed. In the study by Craswell and Szummer [6], where Markov random walks were applied to a bipartite image-query click graph, images not clicked in a specified query (but clicked for at least *one* query in the same connected component) could be associated by performing a backward walk through the graph. Furthermore, in long-term learning, images with no judgements can be given pseudo-relevance feedback judgements by classifying them based on visual or text features [13]. Our approach to partially alleviating sparsity is to merge identical queries based on the query text (see Section 3.3), although this does not have an effect for images that have never been clicked.

Noise resulting from spurious clicks on documents irrelevant to the query is another problem related to clickthough data. However, recent research [30] indicates that this problem is much less pronounced in image search. This is likely due to the fact that image thumbnails are much better document summaries compared to text snippets and document titles commonly seen in text retrieval [6].

Regarding the use of latent variable models, apart from the early studies that employed LSA on long-term explicit relevance feedback [12, 13], very few works have investigated their application on clickthrough data. Lin et al. [25] introduced a PLSA-based search personalisation model trained on implicit relevance judgements from clickthrough data in text-based web search. In addition to focusing on clickthrough data from image search, our study differs in that we model the clickthrough data as a whole, rather than on a per-user basis. The distinction is that we look for trends common to all users in a collaborative sense, whereas personalised search learns the search patterns of individual users.

In summary, although the use of latent variable models in situations where explicit feedback is available has been investigated [12, 13], to our knowledge, the effectiveness of these models based on implicit feedback in image search has not yet been studied.

3 Methodology

In this section we define the problem area, discuss the nature of clickthrough data in image search, and introduce the models used in our experiments.

3.1 Definitions

We define an M-element document collection as the set $\mathcal{D} = \{d_1, ..., d_M\}$ (we shall, for the sake of generality, refer to an image as the document d_i) and a collection of

N queries $\mathcal{Q} = \{q_1, ..., q_N\}$ over the elements of $\mathcal{D}$. Each query vector q_j represents the expression of a user's information need through relevance judgements over the documents $\mathcal{D}$. We represent these relevance judgements as a matrix of *document-query* co-occurrences $\mathbf{R} \in \mathbb{R}^{M \times N}$, i.e., each element $r_{ij} \in \mathbf{R}$ denotes implied relevance of document d_i to query q_j.

Clicks from users can be seen as weak indicators of relevance [6]. Under this assumption, the elements r_{ij} take binary values $\{0, 1\}$, where 1 indicates that document d_i is weakly relevant to query q_j, and 0 otherwise. We can extend this representation of implicit relevance to include other forms of user interaction recorded by the retrieval system by weighting actions according to a Likert rating scale.[2] For example, a *download* action, where a user purchases and saves a retrieved document locally, implies relevance of a higher degree than a click (which is seen as more exploratory), and so may be represented as the value 2, yielding ordinal values $r_{ij} \in \{0, 1, 2\}$. The query logs used in this study (formally introduced in Section 4.1) record photograph downloads, and as such we are afforded this additional relevance indication. The use of this graded scale of implicit feedback is investigated in Section 5.1 to determine whether it leads to better performance. To our knowledge such an investigation has not been performed for implicit feedback.

3.2 Noise and sparsity in click data

Noise may be present in clickthrough data if different users have the same information need but submit different queries or click on different results (subjectivity). Polysemy (e.g., "bank" as in river, and "bank" as in the monetary institution) is also a contributing factor. A user may also find a document attractive enough to click on despite its irrelevance to the query. Noise has been shown to be less pronounced in image search because thumbnails provide a concise summary of image content and are therefore more easily assessed during search [6, 30]. We posit that image click noise is even less pronounced in multimedia archives frequented by professional users (such as that considered in this study) compared to image retrieval on the web. Nevertheless, we are concerned with the effects of noise because it may obscure the structure of less frequent query concepts contained in the search logs (i.e. those comprising the tail of the power law).

Image click data, like other forms of clickthrough data, is very sparse because for a given query a user will only ever see a small fraction of the images in the collection. Furthermore, it has been shown that the problem of position bias, commonly seen in web IR [19] where clicks are more frequent on documents ranked higher, is also present in image click data [28]. Sparsity is problematic for learning latent structure because the number of observations is insufficient to make reliable inferences. In order to alleviate its effect we employ the following strategy to merge queries and evaluate its effect in the experiments in Section 5.

[2]The Likert scale is the rating scale used to record user-item preferences in collaborative filtering [14].

3.3 Merging identical queries

Our approach to alleviating sparsity in clickthrough data is as follows. As each search recorded in the query log is initiated by a text-based query, denoted $q_{i\,\text{text}}$, we use this to our advantage during the pre-processing stage and merge queries based on this attribute. If the text of different queries matches exactly, we make the assumption that the information need of the respective users is identical.

Algorithm 1 Merging identical queries based on query terms

```
Initialise hyper-query array H
for q_i ∈ Q(i = 1..N) do
  if ∃q_i text ∈ H then
    j ← index of q_i text ∈ H
    H_j ← H_j + q_i
  else
    Append q_i to H
  end if
end for
```

When two identical queries are merged, a new *hyper-query* vector is created by a vector sum of the original query vectors (see Algorithm 1). We are thus able to view the hyper-queries as containing co-occurrence counts. A document-query pair with a higher count indicates that the document was clicked or downloaded more than a document-query pair with a lower count. Another effect of merging queries is that we reduce the dimensionality of the data, thereby reducing computation requirements. The creation of hyper-queries has some implications, however. We may, for example, lose different meanings of homonyms and risk introducing noise into the data by merging unrelated queries. Merging queries is often implicitly considered in related work [1, 6], but to date its effect has not been empirically measured. We provide a comparative analysis of merging queries in Section 5.

3.4 Latent variable models

Latent variable models are unsupervised learning approaches traditionally used in the text retrieval domain where an $M \times N$ term-document co-occurrence matrix is factored into two component matrices generally following the form [31]:

$$X \approx \Phi \cdot \Theta, \tag{1}$$

where $X \in \mathbb{R}^{M \times N}$ is the matrix of observed co-occurrences, $\Phi \in \mathbb{R}^{M \times K}$ are the word distributions over topics and $\Theta \in \mathbb{R}^{K \times N}$ are the topic distributions over documents. The dot product $\Phi \cdot \Theta$ is an approximation of the original matrix X because the number of topics (K) is normally chosen to be less than N or M [16, 31], resulting in a dimension reduction. In text IR, this dimension reduction allows documents to be characterised as mixtures of topics.

Because we are modelling topics in image clickthrough data, we represent the co-occurrence matrix X as the relevance matrix $\mathbf{R} \in \mathbb{R}^{M \times N}$, where M denotes the number of documents and N denotes the number of queries. In our formulation, the matrix Φ corresponds to the images projected into the latent space and the matrix

Θ corresponds to the queries projected into the same space. Therefore, in this work we replace the traditional term-document matrix with our document-query relevance matrix.

Our aim in this study is to uncover latent structure in query logs. As the following sections will describe, we will decompose the document-query co-occurrence matrix into the factors Φ and Θ. We shall then focus on improving document retrieval by using the latent space Φ as an ad-hoc retrieval system where we posit that document similarities can be calculated more accurately.

3.4.1 Singular value decomposition

The singular value decomposition (SVD) is a well-known linear algebraic matrix approximation technique that has been shown to be useful in many fields including text retrieval as LSA [7], image retrieval based on long-term learning of relevance feedback [13], and collaborative filtering [22].

The calculation of the SVD involves an eigen-decomposition of the square matrices $\mathbf{R}\mathbf{R}^T$ and $\mathbf{R}^T\mathbf{R}$, yielding:

$$\mathbf{R} = U\Sigma V^T, \tag{2}$$

where U are the left singular vectors, Σ are the singular values, and V are the right singular vectors. The decomposition is such that U and V are orthonormal, and Σ is a diagonal scaling matrix with values in decreasing order. By retaining only the first K singular values in Σ, we arrive at a rank-K approximation of the original co-occurrence matrix:

$$\mathbf{R} \approx U_K \Sigma_K V_K^T, \tag{3}$$

where $U_K \in \mathbb{R}^{M\times K}$, $\Sigma_K \in \mathbb{R}^{K\times K}$, and $V_K \in \mathbb{R}^{N\times K}$. In this work, we focus on the U matrix, which projects the documents into the space spanned by the orthonormal basis vectors.

3.4.2 Non-negative matrix factorisation

Non-negative matrix factorisation (NMF) [24] offers us a straightforward approach to the problem of discovering latent topics or "parts" from observed data. NMF, given a non-negative matrix $\mathbf{R} \in \mathbb{R}^{M\times N}$, finds non-negative, non-unique factors giving:

$$\mathbf{R} \approx WH, \tag{4}$$

where $W \in \mathbb{R}^{M\times K}$ and $H \in \mathbb{R}^{K\times N}$ are such that the Frobenius norm $||\mathbf{R} - WH||^2$ is minimised. Following [9], we can define the diagonal scaling matrices $A_{kk} = \sum_{i=1}^{M} W_{ik}$ and $B_{kk} = \sum_{j=1}^{N} H_{kj}$ whose inverses are multiplied by W and H, respectively, to give us the normalised document-topic and topic-query factors WA^{-1} and $B^{-1}H$ (see Table 1).

3.4.3 Probabilistic latent semantic analysis

Probabilistic latent semantic analysis (PLSA) was introduced by Hofmann [16] as a probabilistic reformulation of LSA where term occurrences in documents are seen to

Table 1 Analogous components of the models used in this study

Model	$\Phi \in \mathbb{R}^{M \times K}$	$\Theta \in \mathbb{R}^{K \times N}$
SVD	U_K	V_K^T
NMF	WA^{-1}	$B^{-1}H$
PLSA	$P(d\|z)$	$P(z\|q)$
LDA	ϕ	θ

Φ are the document-topic components and Θ are the topic-query components

result from a generative process. In this study, we represent the generation of implicit relevance judgements as follows:

1. Select a query with probability $P(q_j)$,
2. Select a latent topic z with probability $P(z|q_j)$,
3. Generate a relevant document (i.e. a "click") with probability $P(d_i|z)$.

The elements of the document-query matrix $\mathbf{R}$ correspond to the observations needed to estimate the joint probability $P(q_j, d_i)$:

$$P(q_j, d_i) = P(q_j)P(d_i|q_j), \tag{5}$$

From here we expand the conditional probability:

$$P(d_i|q_j) = \sum_{z \in \mathcal{Z}} P(d_i|z)P(z|q_j). \tag{6}$$

We are particularly interested in the above conditional probability $P(d_i|q_j)$, because we wish to predict a set of relevant documents (i.e. those that would be rated relevant by a user) given a query. Using Bayes rule to transform $P(q_j)P(z|q_j)$ into $P(z)P(q_j|z)$, we can rewrite the joint probability as:

$$P(q_j, d_i) = \sum_{z \in \mathcal{Z}} P(z)P(q_j|z)P(d_i|z). \tag{7}$$

The log-likelihood function, $\mathcal{L}$, is commonly used to determine the fit of the latent variables to the observed data:

$$\mathcal{L} = \sum_{q_j \in \mathcal{Q}} \sum_{d_i \in \mathcal{D}} n(q_j, d_i) log P(q_j, d_i), \tag{8}$$

where $n(q_j, d_i) \in \{0, 1, 2\}$ is synonymous to the term frequency in the standard PLSA model, but here is used to denote whether or not the document is relevant to the query (i.e. a "click" or "download").

Expectation-maximisation (EM) [8] is used to maximise the log-likelihood function. The method consists of two steps: an E-step that calculates the posterior probabilities for the latent topics, and an M-step that updates the predictions based on the posteriors calculated in the E-step.

3.4.4 Latent Dirichlet allocation

Focusing on the shortcomings of the PLSA model, namely that it lacks generative assumptions about the mixture weights $P(d|z)$, Blei et al. [5] developed an extension where a conjugate Dirichlet prior α is placed on $P(d|z)$. Later, Griffiths and Steyvers [11] extended the model by introducing a prior β on $P(z|q)$ (our notation). These conjugate priors are seen as smoothing parameters. Although the main benefit of LDA is that it generalises better than PLSA on unseen documents, the goal of which is not the case in this study, we include it for comparison.

The generative model for a click in a given query q_j under LDA is:

1. Choose $\theta \sim \text{Dir}(\alpha)$,
2. Choose $\phi \sim \text{Dir}(\beta)$,
3. For each document d_i $(1 \leq i \leq M)$:
   (a) Choose a topic $z \sim \text{Multinomial}(\theta)$,
   (b) Choose a relevant document d_i from $P(d_i|z, \phi)$ (i.e. generate a "click" on document d_i).

We thus have the joint probability of a relevant document d_i occurring in a given query (corresponding to $\mathbf{R}$):

$$P(d_i, z, \theta, \phi|\alpha, \beta) = P(d_i|\theta, \phi)P(\theta|\beta)P(\phi|\alpha), \tag{9}$$

where

$$P(d_i|\theta, \phi) = \sum_{z \in \mathcal{Z}} P(d_i|z, \phi)P(z|\theta). \tag{10}$$

Gibbs sampling, used for inference in this study, is a popular method of estimating the posterior distribution over z because for large collections, computing ϕ and θ is intractable [11, 31].

3.5 Unifying the latent variable models

As we saw in (1), the matrix decomposition or factorisation of the latent variable model has a general form. For clarity, we unify the models in Table 1.

The SVD has no probabilistic interpretation due to the constraint that the left and right singular vectors need only be orthonormal; negative valued vector elements are possible following the decomposition. However, it is still possible to view the left and right singular vectors as related to the topic distributions of the general probabilistic topic model [16, 31].

For the purposes of this study, we focus our investigation on the document-topic component Φ because we want to relate documents with each other by measuring distances in this space. However, of equal interest could be the topic-query component Θ, which would allow us to calculate query similarities for query recommendation.

3.6 Model selection

The dimensionality of the latent space (also known as the number of topics for the generative models) is the main parameter in latent variable modelling. We denote

this to be K, where $K << N$ is typical. In choosing an appropriate K, one must keep in mind that choosing too small a value will yield topics that are very broad and contain possible sub-topics. In choosing a K that is too large, approaching N, the topics become less coherent and begin to model the noise. For LDA, the Dirichlet priors must also be chosen. A grid search yielded optimal priors of $\alpha = 0.1$ and $\beta = 0.01$.

3.7 Document similarities

Calculating document similarities in the query space, i.e. where each document d_i is represented as an N-element vector of relevance judgements, is straightforward and is analogous to many collaborative filtering approaches that do not implicitly or explicitly reduce the dimensionality before item recommendations are made. However, it has been shown that as the dimensionality of data increases (e.g. as more queries are performed, increasing N), similarities calculated within this space tend to concentrate, that is, the similarities tend to become equal [3, 32], resulting in lower quality document rankings.

Latent variable models yield a dimension reduction of the original query space to what is called the topic or *latent* space. Using the general notation of (1), we can calculate the similarity between two documents d_i and d_j in the latent space using their respective components Φ_i and Φ_j ($i, j = 1..M, i \neq j$). A number of similarity metrics can be used (for generative models Kullback-Leibler divergence is often used), but for this study, we use the cosine metric because it can be used for with all of the models in this study [31]:

$$D(\Phi_i, \Phi_j) = \frac{\Phi_i \Phi_j}{|\Phi_i||\Phi_j|}. \tag{11}$$

4 Experimental setup

This section details the image clickthrough dataset, groundtruth and baseline methods used in the experiments.

4.1 Clickthrough data

The clickthrough data used in this study is a subset of the search interactions logged by the image search portal of the Belga News Agency.[3] It comprises 1,588,037 images clicked and/or downloaded 5,697,287 times in the context of 824,813 queries, yielding a sparsity of 99.99%. The query text was lightly cleaned in order to facilitate the text-based merging of queries by conversion to lower-case, removal of some stopwords (e.g. "and" and "or"), URLs, punctuation, special characters, and names of major photo agencies. This full view of the click corpus is referred to as **Belga**.

A set of 25 concepts was manually annotated by Belga staff to serve as a groundtruth in the experiments (see Table 2). These concepts are overlapping, in that

[3] http://www.belga.be

Table 2 The 25 concepts used in this study

ID	Name	# Images	ID	Name	# Images
1	Airplane_flying	61	14	Highway	222
2	Airport	391	15	Logo	904
3	Anderlecht	846	16	Meadow	46
4	Athlete	1,714	17	Rally_motorsport	531
5	Basketball	709	18	Red_devils	807
6	Building	521	19	Sky	532
7	Club_brugge	645	20	Soccer	4,103
8	Crowd	538	21	Stadium	190
9	Farms	158	22	Team	151
10	Fashion_model	837	23	Tennis	1,107
11	Fire	334	24	Volleyball	547
12	Flood	468	25	War	477
13	Formula_one	991			

one image can be annotated with more than one concept, but they are not complete, meaning that not all images have been annotated for all concepts; for each concept, only images clicked for queries textually similar to that concept have been annotated. The concepts are also unbalanced, with those relating to sports comprising a majority of the images. Because only a fraction of the images in the entire clickthrough corpus are positively annotated with at least one of the 25 concepts, the scale of the available clickthrough data was substantially reduced. In total we made use of 8,776 images having been clicked and/or downloaded 21,974 times in 10,348 queries, yielding a sparsity of 99.97%. Per image, there is a mean of 0.65 downloads and 1.85 clicks. Per query, there is a mean of 0.55 downloads and 1.57 clicks. In this study, this view of the corpus is referred to as **Belga25**.

The text-based merging of queries, discussed in Section 3.3, yields another view of the corpus. The objective here is to determine whether merging queries is useful compared to the unmerged view. The merge reduced the number of individual queries to 4,835 and the number of clicks and downloads to 17,770 (now seen simply as non-zero co-occurrence counts), and yielding a sparsity of 99.95%. The number of images remained the same at 8,776. This particular view of the corpus is referred to as **Belga25merged**.

4.2 Evaluation

The fit of probabilistic topic models to observed data is often measured by what is known as *perplexity* [5, 11, 16]. However, this study is concerned with an examination of the underlying latent space, not with a generalisation of unseen samples. In addition, we want to use the same evaluation metric for all models, not only the probabilistic ones. Furthermore, the main goal of this study is to determine to what effect topics discovered in search logs can improve image retrieval. To this end, given a document (image) as query, the effectiveness of document rankings produced in the latent topic space is compared against appropriate baselines using mean average precision (MAP). The documents to be used as the queries are sampled from the collection and average precision is calculated against the groundtruth concepts by considering that two documents are relevant to each other if they share at least one

groundtruth concept. For the experiments, the number of documents sampled S was set to 1/10th of the total number of documents (S = 878). We take the mean of the average precision over this set of queries.

Four baselines are used for comparison against the ranking results derived from the latent variable modelling. The first, named **random**, is based on a ranking of random distances. The second consists of document similarities calculated in the original query space based on the *tf-idf* weights of the document-query matrix **R**; this baseline is named **orig**. The third baseline, named **visual**, comprises visual image features based on Integrated Weibull distributions [10] extracted from overlapping image regions. This yields a 120-dimensional feature vector F_v. Finally, the fourth baseline consists of text features and is named **textual**. Each image in the collection has associated meta-data in the form of a textual description of its context and visual content, including the names of places, events, and people. The rankings are performed using the PF/Tijah retrieval system [15]; each image's textual description is used as a query and the remaining images in the collection are ranked using the normalised log-likelihood ratio (NLLR) retrieval model [23].

4.3 Implementation

The SVD implementation used is the economical version available in Matlab (version 64-bit 2010), `svds`. For NMF, we also used the default Matlab implementation, `nnmf` with alternating least squares. The PLSA algorithm was written by Peter Gehler.[4] Convergence for PLSA was satisfied after 500 EM iterations or when the log-likelihood change was less than `1e-4`. For LDA, following a preliminary experiment, we set the Dirichlet priors to $\alpha = 0.1$ and $\beta = 0.01$ and the number of Gibbs sampling iterations to 1000. The implementation of LDA used was from the Topic Modelling Toolbox for Matlab by Steyvers and Griffiths.[5]

4.4 Algorithmic complexity

The complexity of a latent variable model is dependent on the number of documents M, number of observed co-occurrences $R = |\mathbf{R}|$ and the number of topics K (recall that there are N queries and in our case $M > N$). Specifically, SVD using the Lanczos approximation is roughly $O(MN(R/N)K) = O(MRK)$ where R/N is the average number of non-zero entries (relevance judgements) per query. NMF using the alternating least squares update method has complexity $O(MNK)$ [2]. For PLSA, each E-M iteration has complexity $O(RK)$ [16] and each iteration of the Gibbs sampler for LDA has complexity $O(RK)$ [29].

In decomposing the full (1,588,037 images × 824,813 queries) and evaluation (8,776 × 10,348) click corpora, we noted the computation times listed in Table 3 on an eight core 2.8 GHz Intel Xeon system running GNU/Linux with 32 GB RAM. These times are satisfactory considering the scale of the query log and show that

[4] http://www.kyb.tuebingen.mpg.de/bs/people/pgehler/code/index.html

[5] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

Table 3 Computation times for full and evaluation click corpus decompositions ($K = 100$ topics)

Method	**Belga** (s)	**Belga25** (s)
SVD	6,459	44
NMF	N/A	1,961
PLSA	9,400	65
LDA	12,773	46

decomposing such large matrices is feasible in both time and space. Additionally, as we are working in the news photography domain, image collections will have a smaller scale compared to image retrieval on the web.

5 Results and discussion

The main goal of our experiments is to investigate the image retrieval effectiveness of latent variable models on clickthrough data. In this context, we also examine two aspects that are currently unexplored: the effect of graded implicit relevance judgements and the reduction in sparsity resulting from a text-based merge of queries.

5.1 Weighted relevance judgements

To assess whether a higher weighting of *download* actions is beneficial, we performed the following experiment. We choose one model, SVD, because it is determined to be one of the better performing in the experiments in Section 5.2. The number of topics K is varied while decompositions are performed on the unmerged view **Belga25**. The decompositions are performed for the following scenarios:

- clicks only, no download actions present (*nodl*);
- downloads only, no click actions present (*nocl*);
- clicks and downloads equally weighted ($r_{ij} \in \{0, 1\}$);
- clicks equal to 1, downloads equal to 2 ($r_{ij} \in \{0, 1, 2\}$).

For each decomposition in each scenario, MAP is calculated using the 25 groundtruth concepts. Table 4 shows the results. In all cases, the inclusion of the *download* action ($dl = cl$ and $cl = 1, dl = 2$) is beneficial over not including it (*nodl*). With regard to weighting the action, the original space **orig** shows little difference. For SVD, apart from $K = 5$, we see an improvement by giving the *downloads* more weight than the *click* actions. Therefore, for a broad number of topics ($K > 5$), these results suggest that it is beneficial to include different relevance actions according to a graded (Likert) scale, where actions with higher relevance are weighted accordingly.

5.2 Document ranking

The experiments for the main contribution of our work examine the effect of clickthrough-based latent variable modelling on ranking documents. Since we have already shown evidence that weighting *download* actions improves MAP for a wide range of K, we only consider the case of weighted relevance judgements. For each

Table 4 Experiment on the efficacy of including and weighting *download* actions heavier than *click* actions (nodl: clicks only; nocl: downloads only; cl = dl = 1: clicks and downloads = 1; cl = 1, dl = 2: clicks = 1; downloads = 2)

Model	K	nodl	nocl	dl=cl	cl=1,dl=2
Orig	n/a	0.1920	0.2005	0.2165	0.2161
SVD	5	0.2885	0.1881	0.3115	0.2975
	10	0.2591	0.1875	0.2866	0.2985
	15	0.2579	0.1878	0.2964	0.2989
	20	0.2457	0.1875	0.2575	0.2717
	25	0.2457	0.1880	0.2608	0.2654
	30	0.2417	0.1882	0.2548	0.2622
	35	0.2371	0.1876	0.2502	0.2641
	40	0.2270	0.1874	0.2518	0.2608
	45	0.2400	0.1876	0.2469	0.2549
	50	0.2255	0.1889	0.2601	0.2646

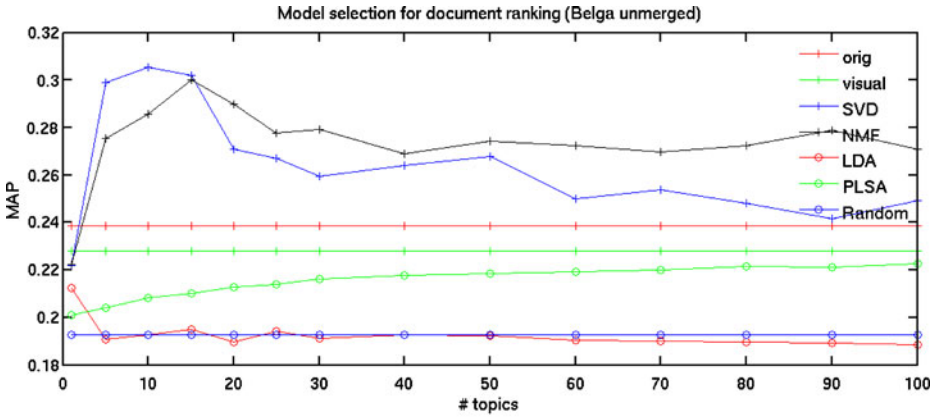
Empirical values represent mean MAP scores over five runs

model, we vary the number of topics K while computing MAP (see Section 4.2) to determine an optimal fitness. Figure 1a and b show the effects on MAP as model dimensionality increases for the **Belga25** and **Belga25merged** views, respectively. As the models are unsupervised, we ran each five times on the full dataset, and each point in Fig. 1 represent the mean over all runs. Significance testing was performed using ANOVA and p-values are included where appropriate.

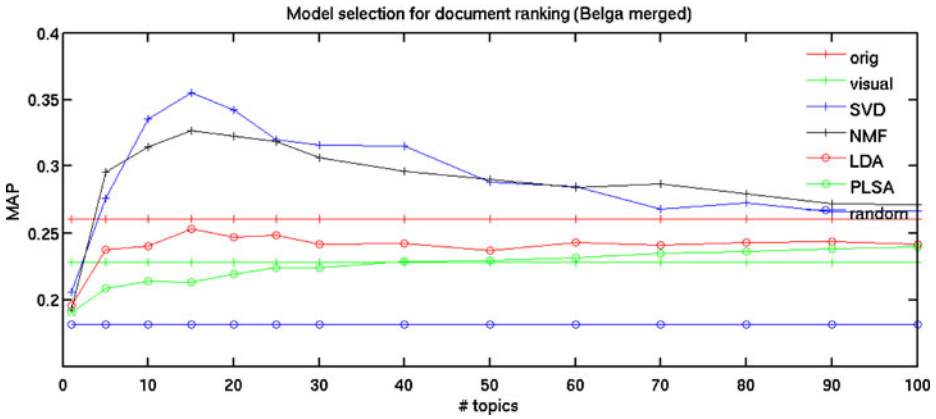
In comparing the document rankings of the latent variable models to the four baseline methods, we see that all perform better than the **random** baseline, although the behaviour of LDA is close to **random** in the unmerged view. With respect to **orig**, NMF and SVD perform better in both views, while LDA and PLSA perform worse in the unmerged view and converge towards it but never pass it in the merged view. With respect to the **visual** baseline, all models perform better in the merged view; in the unmerged view, NMF and SVD still perform better, while LDA and PLSA perform worse. Lastly, the **textual** baseline (MAP of 0.61) outperforms all models and comparative baselines. This is easily explained due to the high quality of the text meta-data associated with these particular images obtained from a professional archive.

SVD and NMF clearly show the benefit of computing document similarities in the latent space for both views of the Belga data. These findings support our initial assumption that for clickthrough data, document similarities computed in the latent space are more accurate than those in the original query space, which suggests that the dimension reduction performed by SVD and NMF helps by filtering noise. Furthermore, these models, as well as the **orig** baseline, yield better performance than the visual baseline in both views, indicating that better retrieval could be achieved by using only the clickthrough data.

Next, we compare the unmerged and merged views **Belga25** and **Belga25-merged**. This allows us to show the effects of our approach to reducing sparsity as well as representing relevance co-occurrences as counts (the reader is reminded that the unmerged view comprises categorical data {0, 1, 2} and the merged view comprises counts). The **orig** baseline is improved following the merge of queries. The **textual** and **visual** baselines are not affected by the query merge, and thus remain constant in



(a) Belga25



(b) Belga25merged

Fig. 1 Mean average precision (MAP) for all models as a function of varying the number of topics for (**a**) unmerged queries and (**b**) merged queries. The textual baseline, not included for readability reasons, has constant MAP at 0.61. Each point represents the mean over five runs

both views. There is a clear improvement in MAP for all models for the merged view over the unmerged view. For LDA, the differences are marked: in the unmerged view, performance is worse than **orig**, significantly worse than **visual**, and only marginally better than **random** ($p < 0.01$). In the merged view, LDA climbs steeply, while PLSA remains distanced from **orig** in the merged view. PLSA has previously shown poor performance in the factorisation of sparse 0–1 data [4] where this has been attributed to the fact that the zero entries do not contribute to the log-likelihood. This may explain why PLSA shows slightly better performance on the less sparse merged view. SVD and NMF, on the other hand, appear more robust to sparsity with the differences between the two views being comparatively less. These findings show that merging queries into hyper-queries yields an improvement in the inference of topics in clickthrough data. The cause for the increase in performance may be attributed to the fact that frequency counts are more discriminative in a less

sparse representation (10,348 queries to 4,385 hyper-queries and 99.97% to 99.95% sparsity). However, more experiments are required to determine the exact cause.

Optimal model fitness is usually represented by a maximum or minimum in the evaluation metric [5, 11, 12]. In the case of MAP, we should expect to see a peak for the optimal value of K, indicating that relevant documents are ranked higher. Furthermore, previous work has suggested that the optimum number of topics will be close to the cardinality of the concepts underlying the data [4, 11, 12]. In Fig. 1a, SVD exhibits a steep climb to a peak at $K = 10$, and then steadily decreases to converge with **orig**. In the merged view (Fig. 1b), the peak for SVD occurs at $K = 15$, a value closer to the number of concepts.

The reason why the models find a slightly different optimal number of topics (10–15 versus 25) underlying the clickthrough data is likely due to an artefact introduced during the creation of the annotated groundtruth. Although the annotators have close knowledge of the image database and retrieval system, they cannot be expected to know how the users will search, nor what (and how many) search topics were finally evident in the query logs. Therefore, when creating the list of groundtruth concepts, the cardinality may not match what exists in the data. Evidence of this can be seen by inspecting the list of concepts. For example, the concepts "athlete", "club_brugge", "anderlecht", "soccer" could be topically similar to the users, yet different enough to the annotators to warrant separate concepts. This indicates that several of the groundtruth concepts may in fact be better represented as one concept.

NMF peaks at $K = 10$ for the unmerged view and $K = 15$ for the merged view. These values are consistent with those of SVD, which performs significantly better ($p < 0.01$) and we can draw the same conclusions: that these models are optimal for a dimensionality close to the number of groundtruth concepts. One notable characteristic of NMF in the unmerged view is the relative invariance to the choice of the number of topics after $K = 15$.

For LDA, no peak is apparent in the unmerged view suggesting that LDA is deficient in problems with high sparsity or that it deals better with the count data contained in the hyper-queries. For the merged view, however, MAP is maximum at $K = 15$, more consistent with SVD and NMF, which lends support to this hypothesis. PLSA has no visible peak in either view, and it is difficult to draw any conclusion except to say that there is no clear optimal model due to invariance to the choice of the number of topics.

To summarise our findings on model selection in the unmerged view, SVD and NMF favour fewer topics. For LDA, we cannot draw any conclusions due to the poor performance. Likewise for PLSA, although the performance increases as K increases, there is no visible preferred model choice. In the merged view, all models, with the exception of PLSA, favour a number of topics close to the cardinality of the groundtruth, suggesting that after merging the queries, a realistic model choice is apparent. In the merged view, with the exception of PLSA, we see a trend of a peak MAP for between 15 and 20 topics, generally agreeing with the size of the concept annotations, a finding that is supported in the literature [4, 5, 11, 12].

Breaking up the MAP plot into the groundtruth concepts, it is possible to see the differences between the original space and the performance of the latent variable models on a per concept basis. Figure 2 shows the performance per concept in the original and latent spaces measured using average precision (AP). It is evident

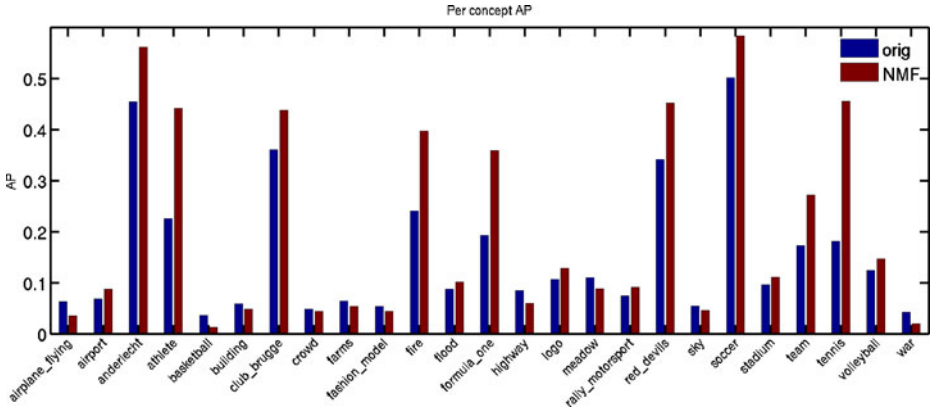


Fig. 2 AP performance in the original and latent spaces (NMF with $K = 15$) per concept for Belga25merged. The popular sports-related topics ("anderlecht", "club_brugge", "soccer") dominate the global MAP scores seen in Fig. 1. Less popular concepts such as "war" show poorer performance in the latent space due to the noise reduction properties of NMF

that sports-related topics ("anderlecht", "club_brugge", "soccer") dominate the global MAP scores seen in Fig. 1. Some concepts, for example "war", show poorer performance in the latent space, due to the noise reduction properties of NMF.

Figure 3 shows 15 example click topics derived using the SVD with $K = 50$ on the full Belga click corpus (1,588,037 images $\times$ 824,813 queries). Visual coherency across topics (columns) is clearly visible. This demonstrates that the models are able to useful reveal latent structure in the clickthrough data without any visual processing.

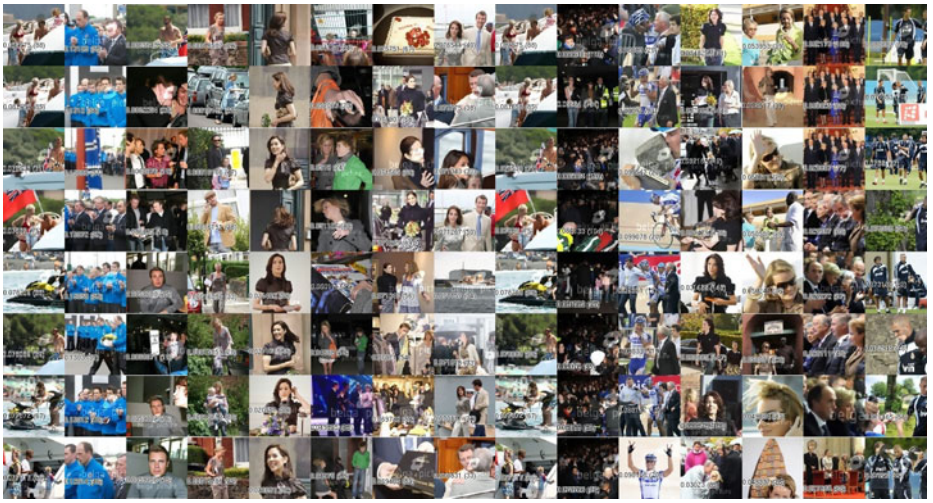


Fig. 3 Fifteen example click topics derived using SVD ($K = 50$) from the full Belga click corpus (1,588,037 images $\times$ 824,813 queries). Each *column* represents a topic with elements ranked *top* to *bottom*

In summary, despite widespread application and favourable performance of the generative models PLSA and LDA in other domains (e.g. text document modelling and collaborative filtering), the experiments conducted in this study do not support their application for sparse click logs. This also implies that a generative view of clickthrough data is not appropriate and that latent variable models without probability constraints such as NMF are more suitable. Our recommendations are, at the very least, that dimension reduction be carried out on sparse click logs before further processing, even if to simply perform a text-based merge of the recorded queries.

6 Conclusions

This study investigated the efficacy of latent variable modelling of clickthrough data in the image search domain. We compared several popular latent variable models, namely the singular value decomposition (SVD), non-negative matrix factorisation (NMF), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA), in terms of ability to model topics underlying implicit relevance clicks in image queries. We demonstrated that rankings were improved by computing similarities in the latent space. This has an important impact in query log mining where many studies calculate similarities in the original, high-dimensional spaces [1]. Most of the models not only outperformed the original query space, but also a baseline ranking using visual features. Although the textual features yielded the highest rankings due to the high quality of the text meta-data, we point out that they may be further improved through fusion with the rankings generated by the latent variable models.

In addition, we investigated the benefit of merging equivalent queries, showing that the resulting dimension reduction in the query space alleviates sparsity to an extent and in doing so improves rankings in both the query and latent spaces. This finding is important because merging queries based on query text is often implied in other studies [1, 6] but is never empirically justified. We also demonstrated that when additional relevance actions are afforded by the retrieval interface (e.g. a *download* action), it is useful to weight them according to a Likert scale, as is frequently seen in collaborative filtering. This is significant to the IR community because it shows that other implicit user actions already recorded in query logs, in addition to clicks, can be used for further performance gains.

This study can be extended in several directions. An enlargement of the groundtruth concepts would allow a larger subset of the clickthrough corpus to be evaluated. In this work, we performed an exact text-based merge of queries. However, dimensionality and sparsity could be further reduced by using a more sophisticated merging technique. Despite the fact that our experiments concentrate on improving document rankings, we should note that the models studied have a much wider range of application on clickthrough data, for example in query expansion and query suggestion. Finally, given that probabilistic topic models (LDA and PLSA) have been very effective in other domains, particularly language modelling [5], the poor performances in this study indicate that a generative formulation, while certainly intuitive, is not appropriate for clickthrough data. However, further analysis of other clickthrough datasets would be necessary to draw this conclusion.

Acknowledgements This research was funded by the Swiss National Science Foundataion (SNF) through IM^2 (Interactive Multimedia Information Management) and by EU-FP7-ICT.1.5 NoE PetaMedia. The authors would also like to thank the Belga News Agency for the use of the query logs.

References

1. Baeza-Yates R, Tiberi A (2007) Extracting semantic relations from query logs. In: Proceedings of ACM KDD'07. ACM, New York, NY, USA, pp 76–85. doi:10.1145/1281192.1281204
2. Berry MW, Browne M (2005) Email surveillance using non-negative matrix factorization. Comput Math Organ Theory 11(3):249–264. doi:10.1007/s10588-005-5380-5. URL: http://www.springerlink.com/content/p474382p18457228/
3. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is "nearest neighbor" meaningful? In: Int. conf. on database theory, pp 217–235
4. Bingham E, Kaban A, Fortelius M (2009) The aspect Bernoulli model: multiple causes of presences and absences. Pattern Anal Appl 12(1):55–78. doi:10.1007/s10044-007-0096-4
5. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
6. Craswell N, Szummer M (2007) Random walks on the click graph. In: Proceedings of ACM SIGIR'07, pp 239–246
7. Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 4:391–407
8. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Stat Methodol 39(1):1–38. doi:10.2307/2984875
9. Gaussier E, Goutte C (2005) Relation between PLSA and NMF and implications. In: Proceedings of ACM SIGIR'05, pp 601–602. doi:10.1145/1076034.1076148
10. van Gemert J, Geusebroek JM, Veenman C, Snoek C, Smeulders A (2006) Robust scene categorisation by learning image statistics in context. In: Proceedings of SLAM'06, p 105
11. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci U S A 101(Suppl 1): 5228–5235. doi:10.1073/pnas.0307752101
12. He X, King O, Ma WY, Li M, Zhang HJ (2003) Learning a semantic space from user's relevance feedback for image retrieval. IEEE Trans Circuits Syst Video Technol 13(1):39–48. doi:10.1109/TCSVT.2002.808087
13. Heisterkamp D (2002) Building a latent-semantic index of an image database from patterns of relevance feedback. In: Proceedings of the 16th international conference on pattern recognition, pp 134–137. citeseer.ist.psu.edu/heisterkamp02building.html
14. Herlocker JL, Konstan JA, Borchers A, Riedl J (1999) An algorithmic framework for performing collaborative filtering. In: Proceedings of ACM SIGIR'99, pp 230–237. doi:10.1145/312624.312682
15. Hiemstra D, Rode H, van Os R, Flokstra J (2006) PFTijah: text search in an XML database system. In: Proceedings of OSIR'06, pp 12–17. http://doc.utwente.nl/66798/
16. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of uncertainty in artificial intelligence. citeseer.ist.psu.edu/hofmann99probabilistic.html
17. Jansen BJ (2009) Understanding user-web interactions via web analytics. In: Synthesis lectures on information concepts, retrieval, and services, Morgan & Claypool
18. Jansen BJ, Spink A, Saracevic T (1999) The use of relevance feedback on the web: implications for web IR system design. In: Proceedings of WebNet'99, pp 500–555
19. Joachims T (2003) Evaluating retrieval performance using clickthrough data. In: Franke J, Nakhaeizadeh G, Renz I (eds) Text mining, pp 79–96
20. Joachims T, Granka L, Pang B, Hembrooke H, Gay G (2005) Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of ACM SIGIR'05, pp 154–161
21. Kelly D, Teevan J (2003) Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37(2):18–28
22. Koren Y (2009) Collaborative filtering with temporal dynamics. In: Proceedings of ACM SIGKDD'09, pp 447–456. doi:10.1145/1557019.1557072
23. Kraaij W (2004) Variations on language modeling for information retrieval. PhD thesis, Centre for Telematics and Information Technology, University of Twente
24. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791. doi:10.1038/44565

25. Lin C, Xue GR, Zeng HJ, Yu Y (2005) Using probabilistic latent semantic analysis for personalized web search. In: Proceedings of the 7th asia-pacific web conference. LNCS, vol 3399, pp 707–717. http://dblp.uni-trier.de/db/conf/apweb/apweb2005.html#LinXZY05
26. Macdonald C, Ounis I (2009) Usefulness of quality click-through data for training. In: Proceedings of the workshop on web search click data, pp 75–79. doi:10.1145/1507509.1507521
27. Müller H, Pun T, Squire D (2004) Learning from user behavior in image retrieval: application of market basket analysis. Int J Comput Vis 56(1–2):65–77
28. Poblete B, Bustos B, Mendoza M, Barrios JM (2010) Visual-semantic graphs: using queries to reduce the semantic gap in web image retrieval. In: Proceedings of CIKM'10, 26–30 October, Toronto, Canada. ACM Press, New York, NY
29. Porteous I, Newman D, Ihler A, Asuncion A, Smyth P, Welling M (2008) Fast collapsed gibbs sampling for latent dirichlet allocation. In: Proceedings of ACM SIGKDD'08. ACM, New York, NY, USA, pp 569–577. doi:10.1145/1401890.1401960
30. Smith G, Ashman H (2009) Evaluating implicit judgements from image search interactions. In: Proceedings of WebSci'09: society on-line. http://journal.webscience.org/148/
31. Steyvers M, Griffiths T (2005) Probabilistic topic models. Latent Semantic analysis: a road to meaning. Laurence Erlbaum
32. Szekely E, Bruno E, Marchand-Maillet S (2010) High-dimensional multimodal distribution embedding. In: IEEE ICDM 2010 workshop on visual analytics and knowledge discovery (VAKD'10), Sydney, Australia
33. Tsikrika T, Diou C, de Vries AP, Delopoulos A (2009) Image annotation using clickthrough data. In: Proceedings of CIVR'09

Donn Morrison studied Computer Science at the University of Victoria, BC, Canada. In 2005 he completed his masters degree at Massey University, Palmerston North, New Zealand, and in 2011 completed his PhD at the University of Geneva, Switzerland. He is currently a postdoctoral researcher at the Digital Enterprise Research Institute in Galway, Ireland where he focuses on modelling and analysis of user behaviour on social networks. His other research interests include latent variable modelling, user interaction in information retrieval and agent-based modelling and simulation.

Theodora Tsikrika received her Degree in Computer Science from the University of Crete, Greece and her MSc in Computer Science and PhD in Information Retrieval from Queen Mary, University of London, UK. During 2007–2010, she was a researcher at CWI, Amsterdam, The Netherlands and in 2011, she joined the University of Applied Sciences Western Switzerland as a researcher. Her research interests focus on (multimedia) information retrieval and are directed towards investigating the combination of statistical, social, and semantic evidence for modelling, analyzing and evaluating (multimedia) search processes and interactions. Since 2007, she has been involved in the coordination of international evaluation benchmarks (INEX, ImageCLEF).

Vera Hollink is a post-doctoral researcher at CWI, specialized in user interfaces and data analysis. Her research aims at enhancing the efficiency and efficacy of information search. It is strongly interdisciplinary, combining a human-computer interaction perspective with techniques from information retrieval, machine learning, and, more recently, semantic web. She received her Ph.D. at the University of Amsterdam, where she worked on statistical techniques for analyzing web browsing behavior and automatically improving link structures. Her current research focuses on the use of semantic information for search log analysis and interactive information retrieval. She has participated in nationally funded projects as well as the EU project VITALAS and is organizer of the international USEWOD workshop in 2011 and 2012 (held at WWW).

Arjen P. de Vries is a tenured researcher at CWI, and a full professor (0.2 fte) in the area of multimedia data management at the Technical University of Delft. De Vries received his PhD in Computer Science from the University of Twente in 1999, on the integration of (multimedia) information retrieval and database systems. He is especially interested in the design of database systems that support search in multimedia digital libraries. He has worked on a variety of research topics, including (multimedia) information retrieval, database architecture, query processing, retrieval system evaluation, and ambient intelligence. He has supervised several best student papers at ACM conferences. He has been general co-chair of the ACM SIGIR 2007 conference in Amsterdam and programme co-chair of CIKM 2011 and ECIR 2012. De Vries is a member of the TREC PC, and a steering committee member of INEX (the Initiative for the Evaluation of XML Retrieval).

Eric Bruno received the MS degree from the Engineers School of Physics, Strasbourg, France, in 1995, and the PhD degree in signal processing from Joseph Fourier University, Grenoble, France, in 2001. From 2002 to 2010, he has been with the Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland, as a research associate. His research interests focus on machine learning and multimodal fusion. Since 2010, Eric is working at Firmenich SA as a Senior Scientist in Data Mining and Knowledge Discovery.

Stéphane Marchand-Maillet has founded and is heading the Viper group (http://viper.unige.ch) in the Department of Computer Science at University of Geneva. His research is directed towards multimedia information retrieval with emphasis on Multimedia Content Abstraction, ie attaching semantic information to multimedia documents at cheapest cost. In particular, he is interested in all aspects related to multimedia information mining and retrieval and smooth acquisition of knowledge by enhancing user or group interaction. He has recently been appointed as Chair of the Technical Committee 12 of the International Association for Pattern Recognition (IAPR-TC12, "Multimedia and Visual Information Systems", http://www.iapr-tc12.org). He was the general co-chair of the ACM International Conference on Image and Video Retrieval (ACM-CIVR 2009, http://www.civr2009.org). He was also the general co-chair of the International Conference of the ACM-SIG on Information Retrieval in 2010 (ACM-SIGIR 2010, http://www.sigir2010.org).