Modelling long-term relevance feedback

Donn Morrison
Computer Vision and Multimedia Laboratory
University of Geneva
Geneva, Switzerland
donn.morrison@unige.ch

Stéphane Marchand-Maillet
Computer Vision and Multimedia Laboratory
University of Geneva
Geneva, Switzerland
stephane.marchand-maillet@unige.ch

Eric Bruno
Computer Vision and Multimedia Laboratory
University of Geneva
Geneva, Switzerland
eric.bruno@unige.ch

ABSTRACT

We propose a general relevance model, called the User Relevance Model, that formalises the decisions taken by a user during a query with respect to relevance judgements. Starting from a keyword-based query, the user is allowed to refine the document search using relevance feedback iterations where some subset of the result set is marked as relevant, and another subset is marked as non-relevant. The model postulates that observed relevance judgements stem from the existence or lack thereof of underlying topics or concepts common to both documents and the query. Furthermore, it explains the underlying concepts through the estimation parameters using a latent-variable model, non-negative matrix factorisation. Experiments are carried out on artificial relevance feedback judgements generated using the model.

Keywords

User modelling, long-term learning, inter-query learning, relevance feedback

1. INTRODUCTION

Content-based image retrieval has a fundamental limitation: the semantic gap. Humans are able to infer high-level semantics from images, while computers can only describe images in low-level detail: shape, colour, texture, etc. In order to improve image retrieval and indexing beyond what has been accomplished in the traditional content-based domain, this gap must be narrowed.

Exploiting user interaction, whether explicit or implicit, is increasingly seen as a promising approach to this problem given that users have the ability to quickly summarise and compare documents. In the context of a typical retrieval scenario, this user interaction may be motivated by the need to broaden or narrow a query. Relevance feedback (RF) is one way this has been accomplished: the user begins and iteratively refines a query by specifying relevant and non-relevant examples from the results. These interactions, comprising the relevant and non-relevant documents, can be aggregated over a long period of time (and from many users) and used to improve subsequent searches. This technique is often referred to as *long-term* or *inter-query*[1] learning because these judgements are used long after the pertinent query has terminated.

Long-term learning differs from search personalisation in that there are no distinct sub-profiles (i.e. user profiles) within the data; queries from users are combined and treated as a whole. In the context of social networks, the aggregated interactions from many users can contribute to the efficacy of content retrieval and improve and propagate content metadata and tags.

Several studies have looked at using long-term learning to annotate and index images, propagate meta-data, and improve the search experience [13, 7, 2]. However, as yet no research has proposed a model that formally describes the characteristics of long-term relevance feedback data and under what conditions it is generated. We propose a new model to address this, called the User Relevance Model. The proposed model postulates that observed relevance judgements stem from the existence or lack thereof of underlying topics or concepts common to both documents and the query. By formalising these assumptions, we are able to justify the extraction and analysis of the underlying concepts using a latent-variable model, non-negative matrix factorisation.

We start by briefly introducing related studies in the area which have attempted to learn long-term relevance feedback data. We then present the notion of hidden topics or concepts that underly document collections. Next, we formalise this notion with the User Relevance Model and demonstrate how the underlying concepts manifest themselves in relevance feedback data provided by users. Finally, we show the role latent-variable models play in extracting and analysing concepts observed in long-term RF judgements.

2. RELATED WORK

Relevance feedback, traditionally used in a query-by-example (QBE) paradigm, is a method of allowing the user to iteratively refine a query by marking relevant and non-relevant examples. The most pervasive approach was introduced by Rocchio in 1971 for the SMART retrieval system [16]. A query is modified based on relevant and non-relevant exam-

[1]Conversely, *intra-query* learning refers to the learning of relevance feedback judgements for the current query only.

ples such that a revised query

$$q' = aq + b\frac{1}{|R|}\sum_{d_j \in R} d_j - c\frac{1}{|R|}\sum_{d_j \in R} d_j \quad (1)$$

is formulated, where d is the document in relevance space R, q is the original query, weighted by a, and b, c weight the positive and negative examples. Many variations have appeared in the literature since Rocchio but it was only until recently that research began focusing on using long-term learning on cumulative relevance feedback instances.

Müller *et al.* studied the use of query logs to improve image search based on market basket analysis [13]. Latent semantic analysis (LSA), an inspiration for our study, has also been effective. Heisterkamp [8] used LSA on artificial RF logs and addressed the problem of data sparsity with pseudo-relevance feedback based on image features. LSA was also applied in a study by He *et al.* where low-level features were used in conjunction with simulated long-term RF data to improve retrieval [7]. Automatic image annotation was a goal for Cord *et al.* in a study that clustered labelled feature vectors around concept centres [2]. Simulated relevance feedback judgements were also used in place of real user interaction.

From the reviewed literature, it is evident that there is a general trend of using artificially generated relevance feedback data. From a machine learning perspective, artificial data has benefits over the use of real-world data, namely parameter control and the ability to benefit from a ground truth. Carefully generated artificial data following a flexible parameterised model can be extremely useful in developing and tuning learning algorithms. We propose such a model, attempting to fill a void in the literature, that formalises the manner and conditions under which users make relevance judgements.

3. USER RELEVANCE MODEL

3.1 Notion of a concept space

The information retrieval community has long modelled so-called *topics* for text and multimedia retrieval. Just as words give an idea of the topics in text documents, visual features give an idea of the concepts in images and video. Concepts are a natural way of organising documents and are a key notion behind dimensionality reduction. They are difficult to quantify yet are visibly evident when we examine documents. They can also be subjective; what one person considers a prevalent concept in a document, another may consider as non-existent. For this reason, in the following subsection, while we formalise the notion of a concept space, we also intentionally leave it abstract.

3.2 User relevance model

We first define documents at the collection level. To remain general, we shall refer to images as documents; the model can apply to any multimedia data type because we assume user interaction is independent of the nature of the collection.

Definition. Consider a collection of M documents $\mathcal{D} = \{d_1, ..., d_M\}$. We define a document $d_i \in \mathcal{D}$ as the set of concept vectors $c_k \in \mathcal{C}$, where $\mathcal{C} = \{c_1, ..., c_K\}$ defines an underlying concept space. Considering the traditional vector space model notation, we can represent d_i as a linear combination of these basis vectors:

$$d_i = \sum_{k=1}^{K} \beta_{ik} c_k. \quad (2)$$

where $\beta_{ik} \in \{0, 1\}$ and denotes whether the concept c_k exists or not in document d_i [15].

Definition. Given a query-by-example retrieval system that affords relevance feedback, we can assume that, at any given stage, users will have invoked a set of N queries $\mathcal{Q} = \{q_1, ..., q_N\}$ over the set of documents $\mathcal{D}$. Using the vector space notation, a query is represented as a linear combination of the basis vectors:

$$q_j = \sum_{k=1}^{K} \gamma_{jk} c_k. \quad (3)$$

where $\gamma_{jk} \in \{0, 1\}$ and denotes whether the concept c_k exists or not in the query q_j.

The scalar product of d_i and q_j can be used to determine whether a user will mark the document relevant to the query via relevance feedback during the query process:

$$\begin{aligned} \langle d_i, q_j \rangle &= \textstyle\sum_{k=1}^{K} \beta_{ik} c_k \cdot \sum_{l=1}^{K} \gamma_{jl} c_l \\ &= \textstyle\sum_{k,l=1}^{K} \beta_{ik} \gamma_{jl} \langle c_k, c_l \rangle. \end{aligned} \quad (4)$$

A document-query relevance matrix $\mathcal{R} \in \mathbb{R}^{M \times N}$ can then be defined, where each element

$$\mathcal{R}(i,j) = \begin{cases} +1 & \text{if } \langle d_i, q_j \rangle > 0 \\ -1 & \text{otherwise,} \end{cases} \quad (5)$$

for $i \in \mathbb{R}^M, j \in \mathbb{R}^N$, where $+1$ indicates that the document d_i is relevant to query q_j, and -1 indicates that it is not relevant. Thus, a document is marked relevant to the query if the document and query share at least one concept in common.

For the scope of this paper we impose two limitations on Eq. (5). First, we assume that each document d_i and query q_j contain at most one concept. Second, we shall generalise the model to accommodate only positive relevance judgements[2]. Following these limitations, we can reformulate Eq. (5) as:

$$\mathcal{R}(i,j) = \langle d_i, q_j \rangle \quad (6)$$

and thus a document is marked relevant to the query if and only if the document and query share exactly one concept in common.

It is important to note at this stage that the relevance matrix $\mathcal{R}$ is complete, i.e. for every query, every document is rated as being either relevant or non-relevant. However, because relevance judgements by users are subjective (subject to noise and distortion), and because no user will likely specify relevance for every document in the collection, we must account for these cases in our model.

Case 1: Subjectivity and user error. Two users having the same query in mind will not always agree to rate documents

[2]This enables an additional flexibility in modelling the weak relevance judgements observed in click-through data [3].

similarly. Some of these relevance ratings may even conflict (e.g. two users with the same information need mark a document d_r as both relevant and non-relevant to their respective queries). In the same vein is user error, arising due to misinterpretation of documents in the result set leading to erroneous relevance judgements. We treat subjectivity and user error by allowing for corruption in the User Relevance Model. By introducing a threshold parameter ξ_n, we can control the amount of corruption introduced by the model. We augment the relevance matrix $\mathcal{R}$ such that for each element:

$$\mathcal{R}'(i,j) = \begin{cases} \mathcal{R}(i,j) & \text{if} \quad \mathcal{U}(0,1) \geq \xi_n \\ flip(\mathcal{R}(i,j)) & \text{otherwise.} \end{cases} \tag{7}$$

For example, by setting $\xi_n = 0.02$, we introduce 2% uniform noise into the data realised as flipping the relevance judgement.

Case 2: Sparsity in the data. Sparsity in the data can be the result of two causes. First, the efficacy of the retrieval system and the heterogeneity between document classes may result in searches ending after only a few relevance feedback iterations, meaning that the user has found what he or she was looking for quickly. Second, for even modestly sized collections, a user cannot be expected to specify relevance for every document in relation to the current query. Depending on the number of results displayed and the number of relevance feedback iterations, a user may evaluate anywhere from a handful to hundreds of documents before the query finishes. Sampling again from a uniform distribution, we simulate sparsity by setting a second threshold ξ_s and augmenting the elements of $\mathcal{R}'$ accordingly:

$$\mathcal{R}''(i,j) = \begin{cases} \mathcal{R}'(i,j) & \text{if} \quad \mathcal{U}(0,1) \geq \xi_s \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The effect is that elements are uniformly randomly deleted from the relevance matrix. Previous studies, including those on collaborative filtering, whose sparsity is generally lower because ratings are cumulative, typically show sparsity levels between 95 to 99% ($0.95 \leq \xi_s \leq 0.99$) [17, 5, 18, 14].

The proposed User Relevance Model formalises a missing aspect of previous studies in long-term learning. Assumptions made in previous studies to generate and model artificial data ignore the concept-basis relationship between a document and a query, noise introduced due to user error, and oversimplify the user decision in judgement-making. This new formalisation permits the understanding of how long-term RF data is generated by users based on perceived concepts in the documents and queries.

3.3 Treatment using latent-variable models

Latent-variable modelling is a branch of unsupervised statistical learning that, through a decomposition of a matrix of co-occurrences, attempts to explain patterns in the observed data through one more latent or hidden variables. Widely used in the social sciences for analysis of personality traits, political affiliation, and the like [1], latent-variable models, such as latent semantic analysis (LSA), have also proved to be a powerful tool in text retrieval because of their ability to resolve problems of polysemy and synonymy [4].

In the problem of long-term learning, where each entry of the matrix represents document relevance to a query (rather than term occurrence in a document), the hidden variables uncovered by a latent-variable model explain the observations by representing the underlying concepts that users have decided make a given document relevant to a query. It is therefore a natural method to apply.

The idea is to examine the RF process to discover the underlying concepts present in the documents and queries. Essentially, we want to discover the matrices of weights β_{ik} and γ_{jk} from Eqs. (2) and (2) which project the documents and queries into the underlying K-dimensional concept space. In other words, we want to discover to what extent each concept $c_k \in \mathcal{C}$ exists in $d_i \in \mathcal{D}$ and $q_j \in \mathcal{Q}$.

Generally, decompositions made by latent-variable models are not unique and therefore the interpretation of the latent variables can be problematic [1]. However, the latent space present in the component matrices can be interpreted in light of the values of the rows and columns in the co-occurrence matrix. For example, consider two images d_1 and d_2 depicting horses. Through a decomposition both documents are seen to have a high component of concept c_1. In the absence of further information, we could say that concept c_1 may represent something to do with horses.

Non-negative matrix factorisation (NMF) offers a straightforward approach to the problem of discovering latent concepts from observed data. NMF, given a non-negative matrix $\mathcal{R}'' \in \mathbb{R}^{M \times N}$, finds non-negative, non-unique factors giving:

$$\mathcal{R}'' \approx WH, \tag{9}$$

where $W \in \mathbb{R}^{M \times K}$ and $H \in \mathbb{R}^{K \times N}$ and such that $W \cdot H$ minimises the Frobenius norm $||\mathcal{R}'' - WH||^2$ [10]. In our case, the resulting component matrix W yields a projection of the documents into the space defined by the latent basis vectors.

Another popular method is latent semantic analysis (LSA) which uses the singular value decomposition (SVD) to decompose the co-occurrence matrix into three components:

$$\mathcal{R}'' = U\Sigma V^T, \tag{10}$$

where U are the left singular vectors, Σ are the singular values, and V are the right singular vectors [4]. 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 have an rank-K approximation of the original co-occurrence matrix:

$$\mathcal{R}''_K \approx U_K \Sigma_K V_K^T, \tag{11}$$

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 choosing the number of latent variables appropriately in NMF and the SVD, we can conveniently extract the underlying concepts in the User Relevance Model. We can identify the concept weight matrices β_{ik} and γ_{jk} as having the same dimensionality as W, U_K and H, V_K^T, respectively. In practice, we will never know the exact nature of the underlying concepts in β_{ik} and γ_{jk}, and so K must be chosen such that it is large enough to capture diversity in the data yet small enough that some interpretable clustering is observed.

Because the singular value decomposition does not impose non-negativity assumptions on the co-occurrence matrix or the resulting component matrices, interpretation of the latent variables in U_K and V_K^T is not straightforward [9]. NMF is preferable in this sense, because the latent variables lend

themselves to a modelling of non-negative concept weights, which we note leads to a probabilistic formulation [6].

4. EXPERIMENTS

The experiments are conducted on a small subset of the Corel image collection. The subset comprises 1,000 images uniformly spanning 10 categories (100 images per category). Although small, this dataset allows us to quickly and easily visualise performance. Document categories are contiguous in the matrix $\mathcal{R}''$ and therefore similarly in all figures to make interpretation of the results easier.

All sessions of relevance feedback are generated according to the User Relevance Model described in Section 3.2. In other words, given the ground truth image categories, we generate the full relevance matrix $\mathcal{R}$ and subsequently account for sparsity and noise, yielding $\mathcal{R}''$. We empirically fix all parameters as follows: sparsity is set to 98% ($\xi_s = 0.98$), noise is set to 10% ($\xi_n = 0.1$), the number of underlying concepts matches that in the data ($K = 10$), and the number of relevance feedback sessions is set to 5,000. All parameters are varied in the experiments, with the number of latent variables varied to discover the optimum value of K in the data.

We evaluate NMF in comparison to latent semantic analysis using the singular value decomposition, which we have also applied in previous studies [11, 12].

Performance is measured using mean average precision (MAP). MAP emphasises retrieving relevant documents first and provides a quantifiable measure of the clustering of the documents into latent concept classes.

In the experiments presented, we concentrate on modelling the underlying document concepts β_{ik}. However, the same principles can be applied to discovering the underlying query concepts γ_{jk}. Applications of this include profiling of user queries and query suggestion via query similarities.

5. DISCUSSION

The subplots of Figure 1 show the effects of varying various parameters in the User Relevance Model. We know from previous work that the MAP should be at its maximum when the value of K is equal to the actual number of concepts underlying the data [11]. In Figure 1 (a), MAP is highest when $K \approx 10$. Figure 1 (b) demonstrates that the MAP increases as we collect more relevance feedback judgements (sessions). Despite a modest level of noise, it is evident that the MAP approaches 1 as the number of relevance feedback sessions M increases. A significant improvement in mean average precision is observed with as little as 6,000 RF sessions. Figure 1 (c) shows the effects on the MAP as we increase sparsity by augmenting the threshold parameter ξ_s in the model. As sparsity approaches nearly 100%, the latent-variable models can no longer estimate the concept weights β_{ik}. Figure 1 (d) shows the effects of uniform noise introduced by the parameter ξ_n. We see that as we pass 10% noise, both latent-variable models begin to fail. It should be noted however that this is in the presence of high sparsity which is held fixed at 98%.

Figures 2 and 3 demonstrate the success of NMF's modelling of the concept weights β_{ik} from the User Relevance Model. Each figure shows the highest ranked images for the particular concept. In the case of Figure 3, five concepts, lions, wolves, elephants, bears and foxes, have been overlapped in the modelled concepts. This can be attributed to the noise parameter ξ_n. If we set the noise threshold to zero, the overlap disappears and the images cluster. Images with the highest concept weights β_{ik} are shown. The uncovered concept weights β_{ik} can be subsequently normalised and thresholded to obtain equivalent binary estimates to the underlying document concepts described in Eq. (2). A global view of the clustering accuracy is given in Table 1 where we show a matrix measuring cluster confusion.

Table 1: Confusion matrix with columns indicating the underlying concepts and rows indicating cluster membership size. The columns sum to the number of documents per category and the rows sum to the number of documents in each cluster.

3	2	2	3	85	0	8	3	2	2
3	1	5	1	1	1	14	2	80	1
4	4	4	13	4	2	5	4	3	2
3	6	6	2	3	84	10	4	4	5
2	2	68	1	0	3	11	6	3	3
78	2	6	5	1	2	16	2	3	4
1	8	3	3	2	5	9	5	1	80
3	1	0	63	1	2	6	2	1	2
1	70	1	6	2	1	17	4	2	0
2	4	5	3	1	0	4	68	1	1

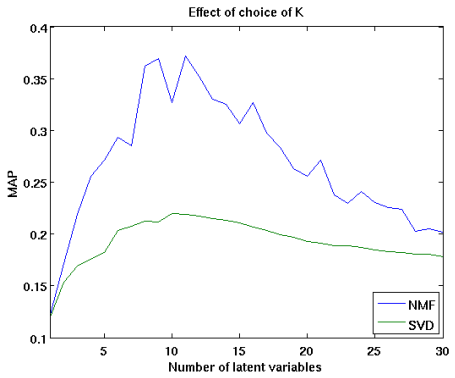
Figure 4 shows the reconstructed concept matrix W containing weights β_{ik} for each document. Columns corresponds to the bar plots in Figures 2 and 3. Due to the inherent non-uniqueness of latent-variable models, the columns of W and the original document-concept matrix will not correspond. What is important to note is that the documents are grouped into similar clusters.

6. CONCLUSION

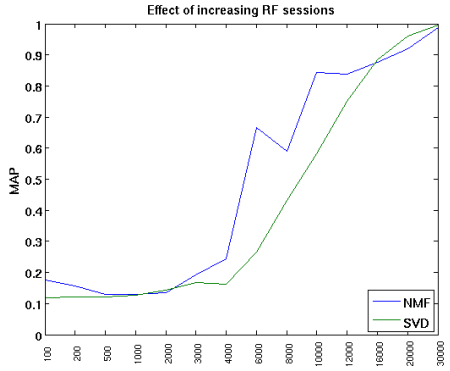
This paper introduced the User Relevance Model, based on notation and principles from the traditional vector space model in information retrieval, which formalises the modelling of relevance feedback judgements for documents given a query based on the principle of concept intersection via the scalar product. The model permits an understanding of how long-term relevance feedback data is generated and describes an underlying concept space that can be readily linked to components of latent-variable models. We demonstrated how one such model, non-negative matrix factorisation, can be used to recover the underlying concepts present in the documents.

We are currently working on a probabilistic version of the User Relevance Model which will make the generative aspect more natural. Due to the non-negativity constraints, non-negative matrix factorisation is readily adaptable to a probabilistic interpretation, permitting a more intuitive analysis of the concept weights. In realising a probabilistic version, we intend to make the modelling of documents and queries more powerful, allowing them to contain overlapping concepts. We also intend to develop a more complicated formalisation of the query, supporting a more realistic boolean query.

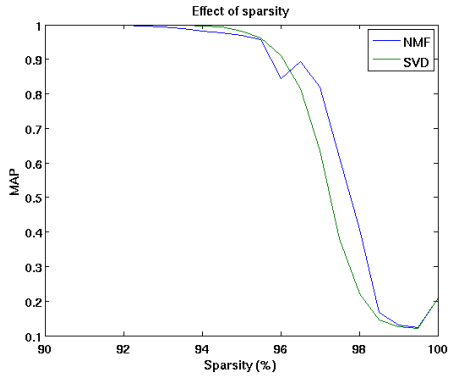
7. ACKNOWLEDGEMENTS



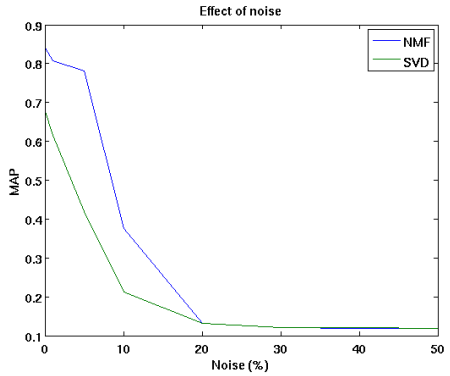
(a) Choice of number of latent variables (K)



(b) Number of RF sessions



(c) Effects of sparsity ξ_s



(d) Effects of noise ξ_n

Figure 1: Various parameters of the User Relevance Model are varied to show the MAP under different conditions.

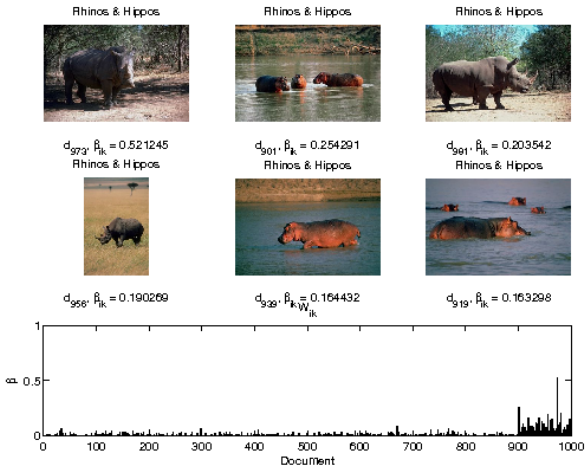


Figure 2: Example underlying concept (bottom) with corresponding images depicting images of rhinoceri and hippopotami. Beneath each document is the corresponding concept weight β_{ik}.

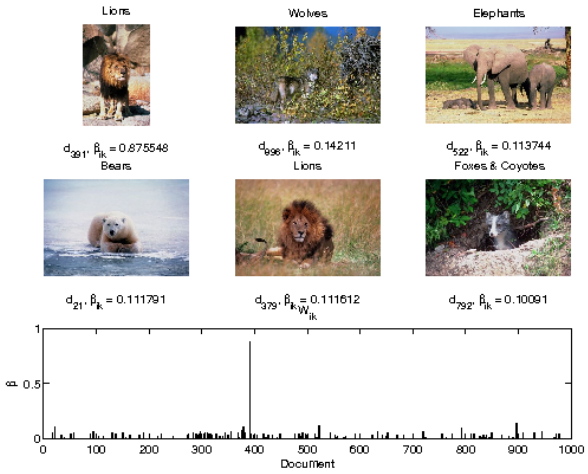


Figure 3: Example underlying concept (bottom) with corresponding images where the dominant concept appears to be lions. Beneath each document is the corresponding concept weight β_{ik}. Concepts are overlapping due to artificial noise introduced in the User Relevance Model.

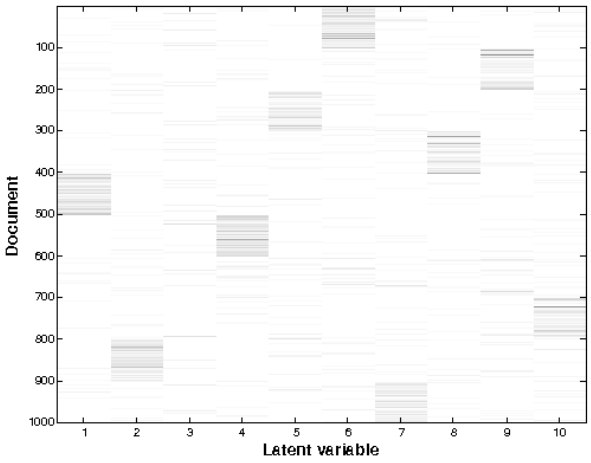


Figure 4: Recovered document-concept relationships in W. Documents are clustered into the latent concepts. The concept in column 4 shows overlap between images. This is attributed to noise introduced by the parameter ξ_n in the User Relevance Model.

This research was funded in part by the Swiss National Science Foundataion (SNF) through IM^2 (Interactive Multimedia Information Management) and by EU-FP7-ICT.1.5 NoE Petamedia.

8. REFERENCES

[1] D. J. Bartholomew and M. Knott. *Latent variable models and factors analysis*. Oxford University Press, Inc., New York, NY, USA, 1999.

[2] M. Cord and P. H. Gosselin. Image retrieval using long-term semantic learning. In *IEEE International Conference on Image Processing*, 2006.

[3] N. Craswell and M. Szummers. Random walks on the click graph. In *In Proceedings of SIGIR 2007*, 2007.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 4:391–407, 1990.

[5] M. Deshpande and G. Karypis. Item-based top-N recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, 2004.

[6] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 601–602, New York, NY, USA, 2005. ACM.

[7] X. He, O. King, W.-Y. Ma, M. Li, and H.-J. Zhang. Learning a semantic space from user's relevance feedback for image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(1):39–48, 2003.

[8] D. Heisterkamp. Building a latent-semantic index of an image database from patterns of relevance feedback. 2002.

[9] A. Kabán and M. A. Girolami. Fast extraction of semantic features from a latent semantic indexed text corpus. *Neural Process. Lett.*, 15(1):31–43, 2002.

[10] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[11] D. Morrison, S. Marchand-Maillet, and E. Bruno. Semantic clustering of images using patterns of relevance feedback. In *Proceedings of the 6th International Workshop on Content-based Multimedia Indexing*, London, UK, June 18-20 2008.

[12] D. Morrison, S. Marchand-Maillet, and E. Bruno. Capturing the semantics of user interaction: A review and case study. In R. Chbeir, A.-E. Hassanien, A. Abraham, and Y. Badr, editors, *Studies in Computational Intelligence: Emergent Web Intelligence*. Springer-Verlag, 2009.

[13] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Long-term learning from user behavior in content-based image retrieval. Technical report, Université de Genève, 2000.

[14] Netflix. The Netflix Prize. Web site: `http://www.netflixprize.com/`, 2006.

[15] V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–287, January 1999.

[16] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System*, pages 456–484. Prentice Hall, New Jersey, 1971.

[17] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender systems–a case study, 2000.

[18] J. Wang, A. P. de Vries, and M. J. Reinders. A user-item relevance model for log-based collaborative filtering. In *Proc. of European Conference on Information Retrieval (ECIR 2006), London, UK*, 2006.