IPAL at CLEF 2008: Mixed-Modality based Image Search, Novelty based Re-ranking and Extended Matching

Sheng Gao, Jean-Pierre Chevallet and Joo-Hwee Lim
IPAL, Institute for Infocomm Research, A*Star, Singapore
{gaosheng,viscjp,joohwee}@i2r.a-star.edu.sg

Abstract

This paper introduces the IPAL participation at CLEF 2008 on the new TEL collection and on the ad-hoc photographic retrieval ImageClef. Following the changes in evaluation criterion this year in ImageClef, i.e. promoting diversity in the top ranked images, we have integrated the novelty measure in our similarity based system developed in ImageCLEF 2007. The novelty score is calculated between an image in the ranked list and the images ranked higher than it. The system is still an automatic and mixed-modality based image search, which is similar to the previous years. 10 runs are submitted this year in ImageClef. In the overall ranking, our group stands at the 3rd place in 25 participants. 4 runs are submitted for the TEL collection. In this working note, we will share our experience in participating these two tasks.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Language model, information retrieval, multimodality fusion, content-based image retrieval, text based image search, clustering

1 Introduction

This year IPAL group continues to participate in the task of ad-hoc photographic retrieval. The evaluation image database still uses the previous year's, which contains 20,000 images attached with some text description (e.g. title, short description, narrative description, location, date, etc) about the image. The obvious change of this year is the evaluation criterion. Rather than simply focusing on mean average precision, i.e. *MAP*, over all queries, this year is to promote the diversity at the top ranking images. It means that *a good image search engine ensures that duplicate or near-duplicate documents retrieved in response to a query are hidden from the user and ideally the top results*[1] of a ranked list will contain diverse items representing different sub-topics within the

[1] http://www.imageclef.org/2008/photo

results. 39 queries which are parts of 60 queries used in the previous year are re-defined for this year. An additional tag, i.e. cluster tag, is added in the query. Here is a query example (query 2):

```
<top>
<num> Number: 2 </num>
<title> church with more than two towers </title>
<cluster> city </cluster>
<narr> Relevant images will show a church, cathedral or a mosque with three
or more towers. Churches with only one or two towers are not relevant.
Buildings that are not churches, cathedrals or mosques are not relevant even
if they have more than two towers. </narr>
<image> images/16/16432.jpg </image>
<image> images/37/37395.jpg </image>
<image> images/40/40498.jpg </image>
</top>
```

Evaluation is based on two measures: precision at 20 (p20) and instance recall at rank 20 (cr20), which calculates the percentage of different clusters represented in the top 20.

Although it is possible to learn a ranking function to maximize the p20 or cr20 metric of the retrieved system, a few annotated samples must be provided. This is suitable for an interactive search but we prefer to set up a fully automatic retrieval system. The baseline system is similar to our system in 2007. Both are content-based image retrieval systems (*CBIR*) with multiple visual features. The text based image retrieval system (*TBIR*) is using a language model approach [7] and their combination using cross-media pseudo-relevance feedback method [6, 5]. To improve the diversity, we introduce a *novelty score* for each image in the ranked list and combine it with it similarity score to generate the ranking score for ranking images. Novelty is calculated from the pair-wise distance between the images. To incorporate the hidden clustering information, we apply unsupervised clustering algorithm, affinity propagation based clustering, to get the cluster size, the representative image in each cluster and the cluster identity of each image. However, we do not use the cluster tag provided in the query in our current systems. For the textual part only, we experiment this year an *extended matching* that consists of a fusion of the matching inner product with probability links computed on Wikipedia source text.

In the next section, we introduce the details of our systems and the submitted runs.

2 Systems Details

We build various CBIR and TBIR systems using different indexing methods and similarity functions. Totally we submit 10 runs for ImageCLEF and 4 for TEL. In the following, the details are given.

2.1 CBIR System

To enrich the visual content representation, 4 types of low-level visual feature are extracted from the local regions or global images. They are detailed in the following:

COR: Auto color correlogram with the depth 2 in the HSV space. It is extracted from the whole image and is represented by one 324-dimensional vector.

HSV: Histogram in HSV and gray-level space with 162-dimension for HSV plus 4-dimension for gray-level. An image is represented by a 166-dimensional vector.

GABOR: texture feature using Gabor filter in the uniformly segmented 5x5 grids at the 2-scale and 12-orientations. Thus, the mean and variance are calculated at each grid to generate 48-dimensional feature vector. We concatenate the vectors from 25 grids into one 1,200-dimensional vector.

EDGE: 18-dimensional edge orientation histogram is calculated from an image.

Then we apply SVD to remove correlation among the feature components. Assuming the full rank is N, we empirically select the top $N \times 0.8$ eigenvectors and index the image in the eigenspace. The cosine function is used to calculate similarity score. Thus, we have 4 CBIR systems based on each of visual features in the above.

2.2 TBIR System

For the LM-based TBIR used in ImageClef only, we first build a lexicon dictionary (7,866 words) from all text documents (including title, narrative, location) and then train a unigram language model only based on the attached text document for each image in the database. This is done using the CMU-Cambridge Statistical Language Modeling Toolkit [2] with the Witten-Bell smoothing. Thus, each image is indexed by the word probabilities in the lexicon. Given a query topic and any image in the database, the probability of query words generated by the corresponding image LM can be calculated. The image documents in the database are ranked by the probability from the highest to the lowest.

2.3 Pseudo-Relevance Feedback

Learned from the ImageClef 2006 [6, 5], the cross-modality pseudo-relevance feedback (PRF) can improve the system performance, i.e. the TBIR can be boosted by the top-N (here 10 documents are selected) documents from the CBIR as the feedback and vice versa. This year PRF is also adopted.

2.4 Novelty based Re-ranking

In our system, the novelty scores are used for re-ranking the images in the top-1000 generated by the traditional similarity based ranking. Two methods are used to calculate the novelty score for the image in the top-1000. Both derive a novel measure through calculating the pair-wise distance between the image and all images ranked higher than it. The pair-wise distance can be calculated from the low-level image feature (in the first approach) or from both the low-level feature and the cluster identity assigned by unsupervised clustering (in the second approach). Let $I(i)$ be the *i-th* image in the top-1000 list as well as its corresponding feature vector, and $r(i)$ be its ranking position in the list. Given $I(i)$, we denote by $R(i)$ the set of images whose ranking position is higher than $r(i)$. Thus the novelty score $novelty(i)$ of the image $I(i)$, is defined as,

$$novelty(i) = max_{j \in R(i)} f(I(i), I(j)) \tag{1}$$

where $f(x, y)$ is a distance function. Higher value the function has, more novelty $I(i)$ is. In the LM-based indexing, KL-distance is used and the cosine distance is used for visual features.

Besides the low-level feature, we also incorporate the cluster identity of images in the novelty score. We apply affinity propagation based clustering in the top-1000 list [4]. Unlike the k-means clustering which generally needs to input the cluster number and the cluster is described by the samples mean in the cluster, this method [4] can automatically find the cluster number from the input pair-wise similarity matrix and the cluster is represented by the representative sample in the cluster rather the mean. In our case, the representative image can be selected for describing the cluster and has the higher novelty than other images in the same cluster. Then, a new novelty is derived by fusing the cluster based novelty with the low-level feature based.

When the novelty values are calculated, they are combined with the similarity scores to reranking the top-1000 images.

2.5 Maximum Similarity Extended Matching

Both document collections we have worked on this year (ImageCLEF and TEL) are very small: only few sentences. For this reason we had the idea to use another source of information to enhance the matching between short query and short documents.

So this year, we continue to use Wikipedia information, but in a different way: we have modified the matching function to directly incorporate weighted links between words. So, neither document nor query are extended statically, but it is the matching function that dynamically (at querying time) choose the best matching between one word of the query and one word of the document.

We compute probability links between terms by computing the probability of a word w_1 to appear in a Wiki document, knowing that a word w_2 is in this document : $P(w_1|w_2)$. This forms a probability graph. This computation is done by counting all concurrencies of all terms in all wikipedia documents. The raw couple frequency is called "support" in text mining. By imposing a *minimum support*, we force the computed probability to be computed using a minimum of occurrences. This should enhance their quality: the biggest support, then the more significant the relation should be. But it also reduces the number of links in the graph and may miss some interesting ones.

In practice we have filtered Wikipedia using words from documents and queries test collection in order to fasten the concurrency computation, otherwise the number of possible concurrency couples is too large and the time to compute them is too long (days). For the TEL collection we limit the computation to the most frequent 300 million of possible relations, in order to feet into 6Mb of main memory of our server. Using only main memory speed up the computation. These probabilistic links are then used directly into the matching function in this way:

$$\text{for a given } t,\ t^* = \underset{t' \in D}{argmax}(p(t'|t)) \tag{2}$$

$$RSV_{max(Q \triangleright D)}(D, Q) = \sum_{t \in Q} k \times p(t^*|t) \times w_d(t^*) \times w_q(t) \tag{3}$$

When computing the matching between document D and query Q (eq. (3)), for each term t of the query, we select the term t^* of the document D (eq. (2)) that maximize the probabilistic link computed in wikipedia: $p(t^*|t)$. Of course, it is done only when the term t does not appear in the document D. If t is in document D, then obviously $p(t|t) = 1$. If not, we replace the missing t from D, by t^*, and we use its weighting. This method, that we call "maximum similarity extended matching", is an extension of the classical inner product and enables us to retrieve document with a very small term intersection, even with no term intersection at all. This matching technic comes from a work of Crestani [3]. In this way, we can expand the matching process with links but still use classical document weighting.

2.5.1 Extended Matching in ImageCLEF

The IPAL CLEF Photo test collection is mainly composed of image annotations. So our extended matching technics should be well adapted to this collection. The proposed runs use the Divergence From Randomness (DFR) document weighting [1]. The unique constant of this weighting is kept at 1.0. Standard stemming and stop-word removal is applied. For query we only use the "title" field.

We have used our extended maximum matching using Wikipedia documents as a source for the term probability link. We use the WIKI file of January 2008 (enwiki-20080103-pages-articles.xml), which has about 14Gb of text.

All runs (pt = probabilistic term extension) use DFR weighting with constant k at 1.0, the extended matching uses also an arbitrary constant k set to 0.01 to combine the term weight with term link. Finally, we have tested different "support". It is in fact the minimum of couple frequency in Wikipedia that is necessary to keep the link in the similarity graph. We have tested 10, 100, and 1000, with correspond to the 3 proposed runs (`IPALpt1`, `IPALpt2`, `IPALpt3`).

2.5.2 Extended Matching in AD-HOC Task TEL

This collection is also composed of small documents. We apply the same technic described before. For all run (except `IPAL04`), we have used the following document field: `oai_dc:dc`, `dc:title`, `dc:subject`, `dcterms:alternative`, `dc:description`. For queries, only the "title" field is used. We have performed a standard stemming and stop-word removal. All runs use the Deviation From Randomness weighting (DFR), with constant left to 1.0. The minimum support (the minimum number of couple concurrencies) if fixed to 10 for all runs. The computation of the Wiki term dependency graph is based on the filtered version of wiki. We first remove all the stop-words and then do a standard stemming, the same applied to the document and queries. Finally, we filter wikipedia, keeping only stemmed terms that effectively appears in the collection and in the query. Unfortunately, this filtering is not enough to reduce significantly the size of Wiki, and hence to enable a full computation of the concurrencies. So because of these technical reasons (to long computation, not enough main memory when running the process), in these run, *only a very small part* of the wiki is effectively used the the runs about the first 2400 Wiki documents. This may be the explanation of the very little influence of this technic to the results (see below). Since these experiments, we managed to compute the full Wiki concurrency but not in time to be used in the run. We will evaluate later the impact of the size of the Wiki used.

2.6 Description of Submitted Runs

A total of 14 runs has been submited: 10 runs are submitted for ImageClef, including the similarity based CBIR run, TBIR run, cross-modality run, and the novelty based runs and 4 runs for the TEL collection. To combine the ranking scores from different runs, the linear fusion method is utilized. The coefficients of each system are equally set. Now we will describe the condition of each type of run.

IPAL01V_4RUNS_EQWEIGHT: a visual run by equally combining 4 CBIR system;

IPAL02T_LM: a LM-based text run;

IPAL03TfV_LM_FB: a mixed-modality run using cross-media pseudo-relevance feedback from `IPAL01V_4RUNS_EQWEIGHT` (Top-5 documents in `IPAL01V_4RUNS_EQWEIGHT` are used to boost `IPAL02T_LM`);

IPAL04T_LM_Tnov: a text run with the novelty score to re-rank `IPAL02T_LM`;

IPAL05TfV_LM_FB_Tnov: a mixed-modality run with novelty scores calculated from the text and visual features (Baseline is `IPAL03TfV_LM_FB`);

IPAL06T_LM_Tnov_Cluster: a text run with novelty score calculated from text feature and cluster identity (Baseline is `IPAL04T_LM_Tnov`);

IPAL07TfV_LM_FB_Tnov_Cluster: a mixed-modality run with novelty scores calculated from the text and visual features and cluster identity (Baseline is `IPAL05TfV_LM_FB_Tnov`);

IPALpt1: the text based Deviation From Randomness with maximum similarity extended matching using a support of 10;

IPALpt2: The same as previous but with a support of 100;

IPALpt3: The same as `IPALpt1` but with a support of 1000;

The next four runs concerns the Had hoc task for the TEL document collection.

IPAL01: a simple reference run with DFR and no extended matching;

IPAL02: the maximum similarity extended matching using wikipedia (small part see above). The k mixing constant is set to 0.001;

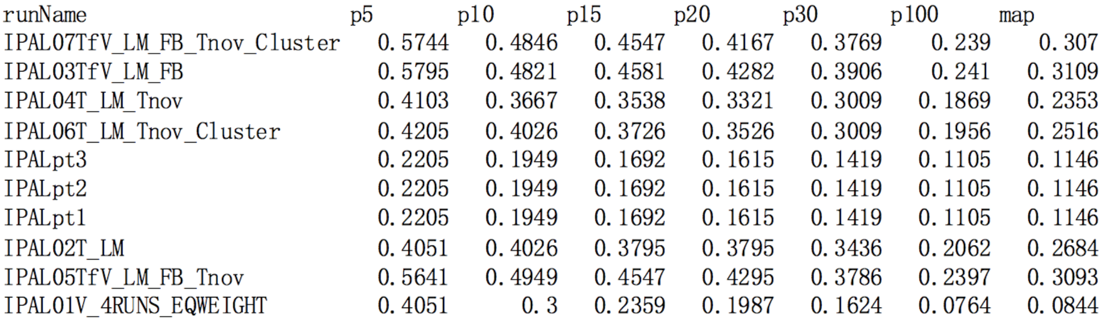
IPAL03: the same at previous, but with a stronger influence of the Wiki extension because the k constant is set to 0.01;

IPAL04: Because we think that the field dc:description is not a good source of information, and can produce noise, we rerun the previous run without this field;

3 Results

The official evaluation results of 10 runs for ImageClef are reported in table 1 for precision at top-20 and Table 2 for instance recall at top-20. In terms of AP@20, the best run is `IPAL05TfV_LM_FB_Tnov`. Its AP is 0.4295. Similarly, the run is also best in terms of CR@20 with 0.4235. However, comparing with the corresponding baseline, `IPAL03TfV_LM_FB`, which has 0.4282 and 0.4217 respectively, the improvement is not obvious. When comparing `IPAL02T_LM` with `IPAL04T_LM_Tnov`, the performance is degraded due to the introduction of novelty score. From these results, we found that little benefit is obtained from the novelty score compared with the traditional similarity search. Similarly, the cluster identity of images discovered from unsupervised clustering has little help to improve the instance recall. It may be helpful when the cluster tag in the query is used. We will evaluate it in future. The use of DFR measure in all `IPALpt` is too low compared with language model, and the use of Wikipedia does not help. The respective MAP for the four runs on TEL collection are: 0.2624, 0.2623, 0.2618 and 0.2579. It shows a degradation of performances using Wikipedia probabilistic links. We thinks that the quality of these extracted link may not be hight enough to show any improvement.

Finally, we summarize the top-10 runs in all submitted runs from 25 participants in Table 3, of which 3 runs are from IPAL.

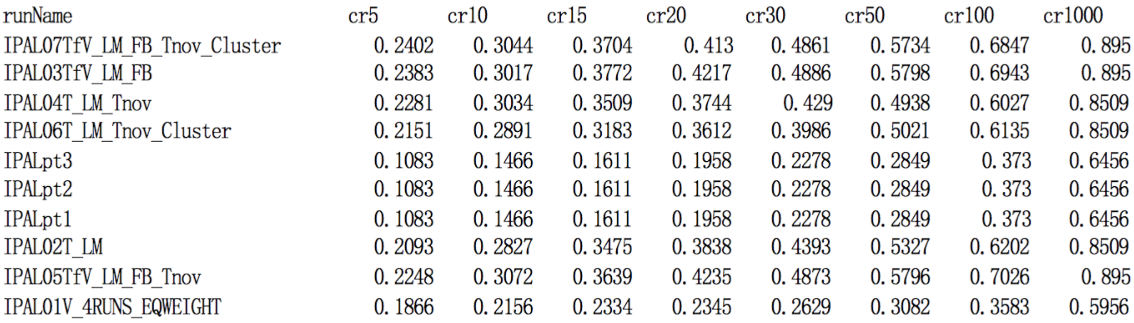


runName	p5	p10	p15	p20	p30	p100	map
IPAL07TfV_LM_FB_Tnov_Cluster	0.5744	0.4846	0.4547	0.4167	0.3769	0.239	0.307
IPAL03TfV_LM_FB	0.5795	0.4821	0.4581	0.4282	0.3906	0.241	0.3109
IPAL04T_LM_Tnov	0.4103	0.3667	0.3538	0.3321	0.3009	0.1869	0.2353
IPAL06T_LM_Tnov_Cluster	0.4205	0.4026	0.3726	0.3526	0.3009	0.1956	0.2516
IPALpt3	0.2205	0.1949	0.1692	0.1615	0.1419	0.1105	0.1146
IPALpt2	0.2205	0.1949	0.1692	0.1615	0.1419	0.1105	0.1146
IPALpt1	0.2205	0.1949	0.1692	0.1615	0.1419	0.1105	0.1146
IPAL02T_LM	0.4051	0.4026	0.3795	0.3795	0.3436	0.2062	0.2684
IPAL05TfV_LM_FB_Tnov	0.5641	0.4949	0.4547	0.4295	0.3786	0.2397	0.3093
IPAL01V_4RUNS_EQWEIGHT	0.4051	0.3	0.2359	0.1987	0.1624	0.0764	0.0844

Table 1: Official evaluation results for 10 submitted runs (Precision at top-N, p@)

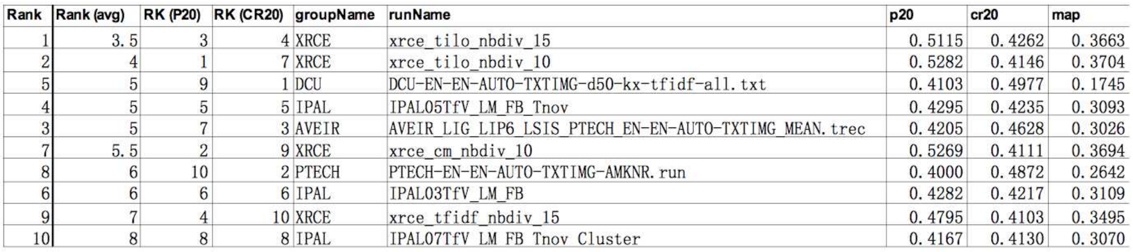
4 Conclusion

In this paper we introduced our ad-hoc photographic retrieval system submitted to ImageClef 2008 and experiments using Wikipedia. None improvement is shown using probabilistic links from Wikipedia. On the image collection, we calculate the novelty score from pair-wise distances among the top-1000 ranked images and then integrate them with the similarity score in order to improve the diverse at the top ranked images. However, the improvement is not significant when comparing with traditional similarity based system and the cluster identity of images cannot give us benefit as we expected. This year, cluster tag in the query is not used. We would like to get some positive effect of unsupervised clustering on the performance when combining with the cluster tag.



runName	cr5	cr10	cr15	cr20	cr30	cr50	cr100	cr1000
IPAL07TfV_LM_FB_Tnov_Cluster	0.2402	0.3044	0.3704	0.413	0.4861	0.5734	0.6847	0.895
IPAL03TfV_LM_FB	0.2383	0.3017	0.3772	0.4217	0.4886	0.5798	0.6943	0.895
IPAL04T_LM_Tnov	0.2281	0.3034	0.3509	0.3744	0.429	0.4938	0.6027	0.8509
IPAL06T_LM_Tnov_Cluster	0.2151	0.2891	0.3183	0.3612	0.3986	0.5021	0.6135	0.8509
IPALpt3	0.1083	0.1466	0.1611	0.1958	0.2278	0.2849	0.373	0.6456
IPALpt2	0.1083	0.1466	0.1611	0.1958	0.2278	0.2849	0.373	0.6456
IPALpt1	0.1083	0.1466	0.1611	0.1958	0.2278	0.2849	0.373	0.6456
IPAL02T_LM	0.2093	0.2827	0.3475	0.3838	0.4393	0.5327	0.6202	0.8509
IPAL05TfV_LM_FB_Tnov	0.2248	0.3072	0.3639	0.4235	0.4873	0.5796	0.7026	0.895
IPAL01V_4RUNS_EQWEIGHT	0.1866	0.2156	0.2334	0.2345	0.2629	0.3082	0.3583	0.5956

Table 2: Official evaluation results for 10 submitted runs (Instance recall at top-N, cr@)



Rank	Rank (avg)	RK (P20)	RK (CR20)	groupName	runName	p20	cr20	map
1	3.5	3	4	XRCE	xrce_tilo_nbdiv_15	0.5115	0.4262	0.3663
2	4	1	7	XRCE	xrce_tilo_nbdiv_10	0.5282	0.4146	0.3704
5	5	9	1	DCU	DCU-EN-EN-AUTO-TXTIMG-d50-kx-tfidf-all.txt	0.4103	0.4977	0.1745
4	5	5	5	IPAL	IPAL05TfV_LM_FB_Tnov	0.4295	0.4235	0.3093
3	5	7	3	AVEIR	AVEIR_LIG_LIP6_LSIS_PTECH_EN-EN-AUTO-TXTIMG_MEAN.trec	0.4205	0.4628	0.3026
7	5.5	2	9	XRCE	xrce_cm_nbdiv_10	0.5269	0.4111	0.3694
8	6	10	2	PTECH	PTECH-EN-EN-AUTO-TXTIMG-AMKNR.run	0.4000	0.4872	0.2642
6	6	6	6	IPAL	IPAL03TfV_LM_FB	0.4282	0.4217	0.3109
9	7	4	10	XRCE	xrce_tfidf_nbdiv_15	0.4795	0.4103	0.3495
10	8	8	8	IPAL	IPAL07TfV_LM_FB_Tnov_Cluster	0.4167	0.4130	0.3070

Table 3: Top-10 runs in all submitted runs from 25 participants

References

[1] Gianni Amati and Cornelis Joost van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transaction on Information Systems*, 20(4):357–389, October 2002.

[2] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Eurospeech97*, pages 2707–2710, 1997.

[3] Fabio Crestani. Exploiting the similarity of non-matching terms at retrievaltime. *Journal of Information Retrieval*, 2(1):27–47, 2000.

[4] Brendan J. J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, February 2007.

[5] Sheng Gao, Jean-Pierre Chevallet, Thi Hoang Diem Le, Trong Ton Pham, and Joo Hwee Lim. Ipal at imageclef 2007 mixing features, models and knowledge. In *Working Notes for the CLEF 2007 Cross Language Evaluation Forum, Budapest, Hungary*, 19–21 September 2007.

[6] Nicolas Maillot, Jean-Pierre Chevallet, Vlad Valea, and Joo Hwee Lim. Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In *Working Notes for the CLEF 2006 Workshop, 20-22 September , Alicante, Spain*, 2006.

[7] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.