On the Effectiveness and Optimization of Information Retrieval for Cross Media Content

Pierfrancesco Bellini, Daniele Cenni, Paolo Nesi

DISIT Lab, Department of Systems and Informatics
Faculty of Engineering, University of Florence, Florence, Italy
{pbellini, cenni, nesi}@dsi.unifi.it, http://www.disit.dsi.unifi.it

Abstract

In recent years, the growth of Social Network communities has posed new challenges for content providers and distributors. Digital contents and their rich multilingual metadata sets need improved solutions for an efficient content management. There is an increasing need of services for scaling digital services, searching and indexing. Despite the huge amount of contents, users want to easily find relevant unstructured documents in large repositories, on the basis of multilingual queries, with a limited waiting time. At the same time, digital archives have to be fully accessible even if a major restructuring is in progress, or without a significant downtime. Evaluating the effectivness of retrieval systems plays a fundamental role in the process of system assessment and optimization. This paper presents an indexing and searching solution for cross media content, developed for a Social Network in the domain of Performing Arts. The research aims to cope with the complexity of a heterogeneous indexing semantic model, with tuning techniques for discrimination of relevant metadata terms. Effectiveness and optimization analysis of the retrieval solution are presented with relevant metrics. The research is conducted in the context of the ECLAP project (`http://www.eclap.eu`).

***Keywords**: cross media content; indexing; searching; search engines; information retrieval; cultural heritage; stochastic optimization, test collections; social networks*

1. Introduction

During the last years, the dramatic growth of digital items on the Internet has opened new scenarios, addressing several issues for content providers and distributors. Large digital repositories, with heterogeneous formats and metadata are facing several issues, especially in the context of content retrieval and management. Cross media resources that are populating the Web need efficient indexing and searching solutions, tools for metadata management, mapping and extraction, content translation and certification services. Users want to get fast and relevant results in response to their queries, robust to typos or inflexions. Moreover, searching services are requested to offer faceting and sorting capabilities, for an easy and intuitive search refinement. Rich text documents require smart processing tools, for realtime full-text parsing, extraction and indexing, that are capable of supporting different types of content descriptors. Also, large content repositories are expected to be fully accessible, without significant downtime, in case of service maintenance or major updates of their indexing structure (e.g., redefinition of index schema, index corruption etc.).

Social Networks, while shortening the distances between users, are making possible a rapid integration of multilingual resources, with different metadata sets and encodings, difficult to deal with. Topics such as query or metadata translation are becoming even more relevant in the context of Information Retrieval (IR), with their pros and cons. Also, the evaluation of retrieval effectiveness, plays a determinant role when assessing a system, hence it is crucial to perform a detailed IR analysis, especially in huge multilingual archives. Ranking a retrieval system involves human assessors, and may contribute to find weakness and issues, that prevent a satisfactory and compelling search experience. A comparative evaluation of IR systems, usually follows the Cranfield model; in this context, the effectiveness of a retrieval strategy is calculated as a function of documents rankings, and each metric is obtained by averaging over the queries.

Tipically, the effectiveness evaluation starts by collecting information needs from a set of topics; following these needs, a set of queries is derived, and then a list of relevance judgments that map queries to corresponding relevant documents. Since people often disagree about a document relevance, collecting relevance judgments is a difficult task; in many cases, with an acceptable approximation, relevance is assumed to be a binary variable, even if it is defined in a range of values [10]. To overcome these limitations, some approaches start the retrieval evaluation without relevance judgments, making use of pseudo relevance judgments [9, 1, 11]. Ranking strategies are often performed by comparing rank correlation coefficients (e.g., Spearman [2], Kendall Tau) with TREC official rankings. The effectiveness of IR can be assessed by computing relevant estimation metrics such as precision, recall, mean average precision, R-precision, F-measure and normalized discounted cumulative gain (NDCG). IR test collections and evaluations series are used for a comparative study of retrieval effectiveness; some examples include TREC, GOV2, NTCIR and CLEF.

This paper describes a Social Network infrastructure, developed in the scope of the ECLAP project, and presents the effectiveness evaluation and optimization of an indexing and searching solution, in the field of Performing Arts. The research was conducted to overcome several other issues, in the context of cross media content indexing, for the ECLAP social service portal. The proposed solution is robust with respect to typos, runtime exceptions and index schema updates; above Information Retrieval metrics are calculated, on the basis of relevant Performing Arts topics. The solution deals with different metadata sets and content types of the ECLAP information model, thus enhancing the user experience with full text multilingual search, advanced metadata and fuzzy search facilities, faceted query refinement, content browsing and sorting solutions. The ECLAP portal includes a huge number of contents such as: MPEG-21, web pages, forums, comments, blog posts, images, rich text documents, doc, pdf, collections, play lists.

The paper is structured as follows: Section 2 depicts an overview of ECLAP; Section 3 introduces Searching and Indexing Tools implemented in the portal; Section 4 discusses IR evaluation and effectiveness, and describes a test strategy, developed for a fine tuning of index fields; Section 5 reports conclusions and future work.

2. ECLAP Overview

The ECLAP project aims to create an online digital archive in the field of the European Performing Arts; the archive will be indexed and searchable through the Europeana portal, using the so called EDM data model. ECLAP main goals include: make available on Europeana a large amount of Performing Arts digital cross media contents (e.g., performances, lessons, master classes, video lessons, audio, documents, images etc.); bring together European Performing Arts institutions, in order to provide their metadata contents for Europeana, and thus resulting in a Best Practice Network of European Performing Arts institutions.

ECLAP provides solutions and services for: Performing Arts institutions, final users (teachers, students, actors, researchers etc.). ECLAP is developing technologies and tools, to provide continuous access to digital contents, and to increase the number of online collected materials. ECLAP acts as a support tool for: content aggregators, working groups on Best Practice reports and articles, intellectual property and business models, digital libraries and archives. ECLAP services and facilities include: user group, discussion forums, mailing lists, integration with other Social Networks, suggestions and recommendations to users. Content distribution is available toward several channels: PC/Mac, iPad and Mobiles. ECLAP includes smart back office solutions, for automated ingestion and refactoring of metadata and content; multilingual indexing and querying, content and metadata enrichment, IPR modeling and assignment tools, content aggregation and annotations, e-learning support.

3. Searching and Indexing Tools

The ECLAP content model deals with different types of digital contents and metadata; at the core of the content model there is a metadata mapping schema, used for content indexing of resources in the same index instance. Resource's metadata share the same set of indexing fields, with a separate set for advanced search purposes. The indexing schema has a flexible and updatable hierarchy, that describes the whole set of heterogeneous contents. The metadata schema is divided in 4 categories (see Table 2): *Dublin Core* (e.g., title, creator, subject, description), *Dublin Core Terms* (e.g., alternative, conformsTo, created, extent), *Technical* (e.g., type of content, ProviderID, ProviderName, ProviderContentID), *Performing Arts* (e.g., FirstPerformance Place, PerformingArtsGroup, Cast, Professional), *ECLAP Distribution and Thematic Groups*, and *Taxonomical content related terms*. Multilingual support and management is available at metadata and resource content levels (e.g., multilingual web pages, taxonomy terms that may include a metadata language either mandatory, optional or not necessary). Performing Arts metadata are mapped to Dublin Core metadata before submission to Europeana. Multilingual index includes ECLAP partners languages (Catalan, Greek, English, Spanish, French, Hungarian, Italian, Dutch, Portuguese, Slovenian, Polish, German, Danish). Multilingual metadata are automatically translated from any source language, and then mapped into their index fields. This approach avoids query translations issues, while metadata editing tools

Table 1: ECLAP Indexing Model

Media Types	**DC (ML)**	**Technical**	**Performing Arts**	**Full Text**	**Tax, Group (ML)**	**Comments, Tags (ML)**	**Votes**
# of Index Fields*	468	10	23	13	26	13	1
Cross Media: html, MPEG-21, animations, etc.	Y_n	Y	Y	Y	Y_n	Y_m	Y_n
Info text: blog, web pages, events, forum, comments	T	N	N	N	N	Y_m	N
Document: pdf, doc, ePub	Y_n	Y	Y	Y	Y_n	Y_m	Y
Audio, video, image	Y_n	Y	Y	N	Y_n	Y_m	Y_n
Aggregations: play lists, collections, courses, etc.	Y_n	Y	Y	Y/N	Y_n	Y_m	Y_n

* = (# of Fields per Metadata type) $*$ (# of Languages)
ML: Multilingual; DC: Dublin Core; Tax: Taxonomy

are at user disposal in the ECLAP portal, for metadata assessment, correction and certification.

Notation used in Table 1, Y_n: yes with n possible languages (i.e., n metadata sets); Y: only one metadata set; Y/N: metadata set not complete; T: only title of the metadata set, Y_m: m different comments can be provided, each of them in specific language. Comments may be annidated, thus producing a forum discussion hierarchically organized.

ECLAP Index Model meets the metadata requirements of any digital content, while the indexing service follows a metadata ingestion schema. Twenty different partners are providing their digital contents, each of them with their custom metadata, partially fulfilling the standard DC schema. A single multilanguage index has been developed for faster access, easy management and optimization.

Index schema includes a set of catchall fields, automatically populated at indexing time for some relevant metadata; it provides a more compact way to build boolean queries to be sent to the searching service. In this model, the catchall field body includes full text of each content element. Relevant metadata are only indexed and not stored in the index, so to keep the index smaller, for a fast access and management; technical metadata are not analyzed and hence stored verbatim (e.g., video resolution, quality, format); descriptive metadata are fully processed (e.g., using token analyzers, lowercasing etc.). Metadata and resource parsed text (i.e., doc, docx, ppt, pptx, xls, xlsx, pdf, html, txt) are extracted with content identifier processing techniques (e.g., file extension, content type and encoding, MIME detection). Multilingual taxonomies are stored in a MySQL database with full references to each resource; each content related taxonomy path is indexed with other content metadata. Multilingual metadata can be enriched and edited on the portal, and then validated and certified before publishing on Europeana. Corrections of multiple metadata set at the same time is also available.

The index structure may be generated automatically from scratch, for example in case of corruption or major schema updates. The production service keeps running while a separate instance of the indexing service is processing the index structure, with accessory tasks (e.g., taxonomy/group extraction related to each digital resource). This strategy avoids significant service downtime, and synchronizes new contents with the newly build index, after indexing completion. A detailed logging and notifying system keeps trace and emails every thrown exception to the administrator.

The searching service aims to help users to easily find and sort digital contents in the ECLAP portal, and to refine queries and results. Main features include:

- Relevance is calculated by taking into account different weights for each indexed metadata field, and it is correlated with term frequency in documents.
- Queries are analyzed, tokenized and lowercased, to build a query expression that is a combination of weighted boolean clauses; in the case of advanced search, the user defines an arbitrary number of clauses.
- Search results are paginated, and presented with relevant metadata by descending relevance.
- Automatic suggesting is available from the portal settings, and provides realtime query suggestions, relevant to the typed query.

A fine tuning of term boosting, giving more relevance to certain fields with respect to others, is a major requirement for the system, in order to achieve an optimal IR performance.

4. Effectiveness and Optimization

ECLAP Metadata Schema, summarized in Table 2, consists of 541 metadata fields, divided in 8 categories; some important multilingual metadata (i.e., text, title, body, description, contributor, subject, taxonomy, and Performing Arts metadata) are mapped into a set of 8 catchall fields, for searching purposes. The scoring

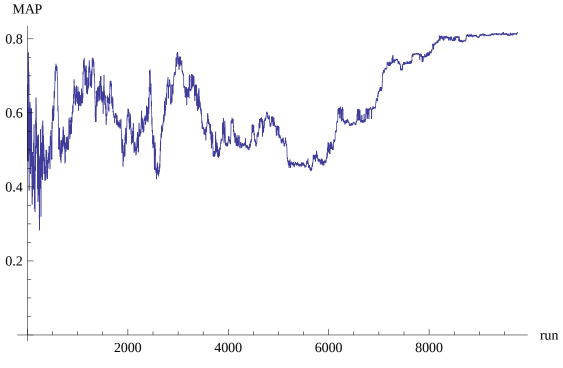


Figure 1: MAP vs test runs

system implements a Lucene combination of Boolean Model and Vector Space Model, with boosting of terms applied at query time. Documents matching a clause get their score multiplied by a weight factor. A boolean clause b, in the weighted search model, can be defined as

$$
\begin{aligned}
b := (\mathit{title}\colon q)^{w_1} \vee (\mathit{body}\colon q)^{w_2} \vee (\mathit{description}\colon q)^{w_3} \\
\vee (\mathit{subject}\colon q)^{w_4} \vee (\mathit{taxonomy}\colon q)^{w_5} \\
\vee (\mathit{contributor}\colon q)^{w_6} \vee (\mathit{text}\colon q)^{w_7}
\end{aligned}
$$

where w_1, w_2, ..., w_7 are the boosting weights of the query fields; title (DC resource name), body (parsed html resource content); description (DC account of the resource content; e.g., abstract, table of contents, reference), subject (DC topic of the resource content; e.g., keywords, key phrases, classification codes), taxonomy (content associated hierarchy term), contributor (contributions to the resource content; e.g., persons, organizations, services), text (full text parsed from resource; e.g. doc, pdf etc.); q is the query; DC: Dublin Core.

The effectiveness of the retrieval system was evaluated with the aim of the *trec_eval* tool. For this purpose, a set of 50 topics was initially collected, in the field of Performing Arts. The set of relevant topics was built starting from a list of popular queries, obtained with a query log analysis. The chosen number of topics is above a threshold, generally suitable for obtaining reliable results, as pointed out by Zobel [12], and by Sanderson and Zobel [8].

For each topic was formulated a query, and then a set of relevance judgments. Each judgment was collected by using a pooling strategy, which helps retrieving relevant items, by choosing a limited subset of the whole set. The method is reliable with a pool depth of 100; limiting the pool depth to 10 may change precision results, but doesn't affect relative performances of IR systems [12]. Moreover, a precise analysis of IR performance is possible, even with a relatively short list of relevance judgments [3].

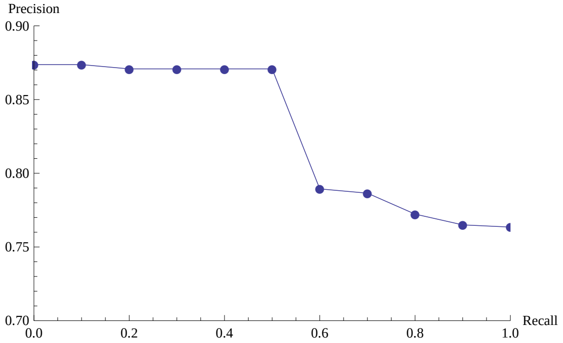


Figure 2: Precision-Recall graph

Table 2: ECLAP Metadata Schema

Metadata Type	# fields	Multilingual	Index fields	# fields/item
Performing Arts	23	N	23	n
Dublin Core	15	Y	182	n
Dublin Core Terms	22	Y	286	n
Technical	10	N	10	10
Full Text	1	Y	13	1
Thematic Groups	1	Y	13	20
Taxonomy Terms	1	Y	13	231
Pages Comments	1	N	1	n
Total	74	–	541	–

Table 3: Estimated IR Metrics for the optimal run

Metric	Value
# of queries	50
# of doc retrieved for topic	4312
# of relevant doc for topic	85
# of relevant doc retrieved for topic	84
MAP	0.8223
Geometric MAP	0.7216
Precision after retrieving R docs	0.7658
Main binary preference measure	0.9886
Reciprocal Rank of the 1^{st} relevant retrieved doc	0.8728

Full text searches on the ECLAP portal are performed through 7 relevant index fields (i.e., title, body, subject, description, text, taxonomy and contributor). In order to find optimal estimations of each index field weight, a minimization test was designed and implemented. Due to the high number of variables, the test implemented a simulation annealing strategy [6], which is a stochastic minimization method, introduced after the work of Metropolis et al. [7]. In order to minimize the energy function $f(\vec{w})$, one choice is to randomly select and increase/decrease only one variable for each iteration, or all the variables. The first option is likely to require a longer simulation time; also, considering the relatively limited excursion range for the random variable, this second approach was preferred.

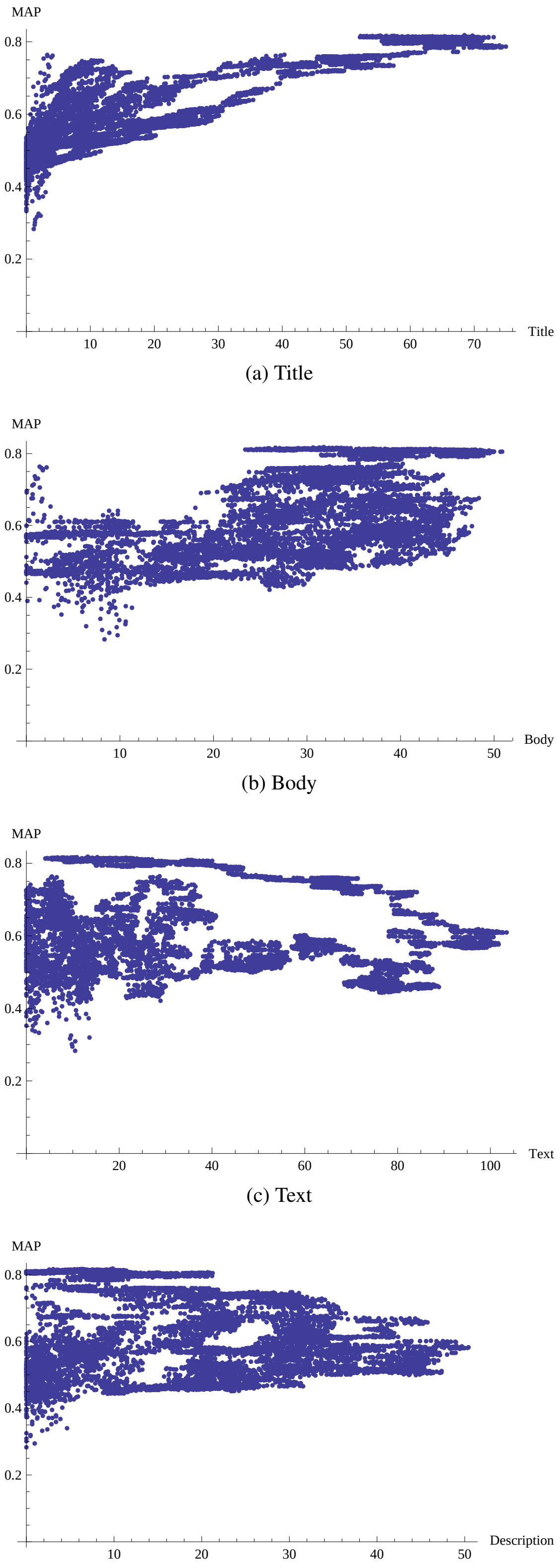
Different annealing schedules, initial state conditions and allowed transitions per temperature were tested. Simulations took place by defining the system state as a vector of field weights $\vec{w_i} = \{w_1, w_2, \ldots, w_7\}$. A run of 50 queries was performed for each state condition, to get corresponding search results with relevant metrics. For each run, Mean Average Precision (MAP) was computed and $(1-MAP)$ was assumed as the energy for the current state. Since MAP is defined as the arithmetic mean of average precision for the information needs, it can be thought as an approximation of the area under the Precision-Recall curve. Following the *Metropolis Criterion*, the probability p_t of a state transition is defined by

$$p_t = \begin{cases} 1 & \text{if } E_{i+1} < E_i \\ r < e^{-\Delta E/T} & \text{otherwise} \end{cases}$$

where E_{i+1} and E_i are respectively the energy states of w_{i+1} and w_i, T is the *synthetic temperature*, $\Delta E = E_{i+1} - E_i$ is the *cost function*, r is a random number in the interval [0,1]. The *annealing schedule* was defined as $T(i+1) = \alpha T(i)$, with $\alpha = 0.8$.

200 random transitions were proven for each temperature iteration. A smoother annealing schedule is more likely to exhibit convergence, but generally requires a bigger simulation time; alternative popular choices include logarithmic schedules such as $T(i) = c/log(1+i)$ [5, 4]. Stopping conditions were defined by counting the number of successful transitions occurred during each iteration. The best simulation schema, showing convergence and system equilibrium is reported in Fig. 1. Some semantic relevant index fields were found to give a limited contribution to the relevance scoring system (i.e., subject, taxonomy and contributor); reducing the number of boolean clauses to be processed by the retrieval system, would result in a higher search speed.

Scatter plots of field weights vs MAP, obtained during the test, exhibited a relevant dispersion across a considerable range of high energy values (see for example Fig.



(a) Title

(b) Body

(c) Text

(d) Description

Figure 3: Scatter plots of Title, Body, Text, Description weights

3). The observed behavior reasonably suggests a high sensitivity to initial conditions and random seeds. The minimization process resulted in an energy minimum at $w_1 = 68.4739$, $w_2 = 31.7873$, $w_3 = 0.2459$, $w_4 = 9.8633$, $w_5 = 13.2306$, $w_6 = 2.1720$, $w_7 = 3.9720$, with $MAP = 0.8223$ (see Precision-Recall graph in Fig. 2 and IR metrics in Table 3). The relatively high estimation of MAP is a direct consequence of a highly polarized assessment pool, toward Performing Arts topics.

5. Conclusions and Future Work

In this paper, an integrated searching and indexing solution for the ECLAP portal has been presented, with a IR evaluation analysis and assessment. The index scales efficiently with thousands of contents and accesses; the ECLAP solution aims to enhance the user experience, by speeding up and simplifying the information retrieval process. Further analysis, simulation and tuning of index weight fields are being conducted, with different optimization approaches. A user behavior study is in progress, in order to understand user preferences and satisfaction.

6. Acknowledgments

The authors want to thank all the partners involved in ECLAP, and the European Commission for funding the project. ECLAP has been funded in the Theme CIP-ICT-PSP.2009.2.2, Grant Agreement No. 250481.

References

[1] Javed A. Aslam and Robert Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 361–362, New York, NY, USA, 2003. ACM.

[2] Jamie Callan, Margaret Connell, and Aiqun Du. Automatic discovery of language models for text databases. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, SIGMOD '99, pages 479–490, New York, NY, USA, 1999. ACM.

[3] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 268–275, New York, NY, USA, 2006. ACM.

[4] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721 –741, nov. 1984.

[5] Bruce Hajek. Cooling schedules for optimal annealing. *Math. Oper. Res.*, 13(2):311–329, May 1988.

[6] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[7] Nicholas Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Medical Physics*, 21(6):1087–1092, 1953.

[8] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 162–169, New York, NY, USA, 2005. ACM.

[9] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 66–73, New York, NY, USA, 2001. ACM.

[10] Amanda Spink and Howard Greisdorf. Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the American Society for Information Science and Technology*, 52(2):161–173, 2001.

[11] Shengli Wu and Fabio Crestani. Methods for ranking information retrieval systems without relevance judgments. In *Proceedings of the 2003 ACM symposium on Applied computing*, SAC '03, pages 811–816, New York, NY, USA, 2003. ACM.

[12] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 307–314, New York, NY, USA, 1998. ACM.