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

Pierfrancesco Bellini, Daniele Cenni and Paolo Nesi
DISIT Lab, Department of Systems and Informatics, Faculty of Engineering, University of Florence, Florence, Italy

Keywords: Cross Media Content, Indexing, Searching, Search Engines, Information Retrieval, Cultural Heritage, Stochastic Optimization, Test Collections, Social Networks.

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. 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)

1 INTRODUCTION

The effectiveness evaluation of Information Retrieval (IR) plays a determinant role when assessing a system, following the Cranfield paradigm or other approaches (Krsten and Eibl, 2011); 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. 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; the 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, playlists. 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 the 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 Europeana Data Model (EDM). ECLAP main goals include: making available on Europeana a large amount of digital cross media contents (e.g., performances, lessons, master classes, video lessons, audio, documents, images etc.); bringing together the European Performing Arts institutions, in order to provide their metadata contents for Europeana, thus creating 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

Bellini P., Cenni D. and Nesi P..
On the Effectiveness and Optimization of Information Retrieval for Cross Media Content.
DOI: 10.5220/0004142703440347
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 344-347
ISBN: 978-989-8565-29-7

groups on Best Practice reports and articles, intellectual property and business models, digital libraries and archives. ECLAP services and facilities include: user groups, 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, Intellectual Property Rights 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 upgradeable 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*.

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 a specific language. Comments may be annidated, thus producing a hierarchically organized discussion forum. The 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. 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.

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

4 EFFECTIVENESS AND OPTIMIZATION

The 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 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

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

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.

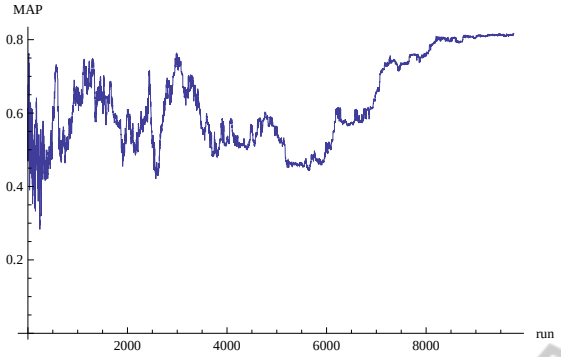


Figure 1: MAP vs test runs.

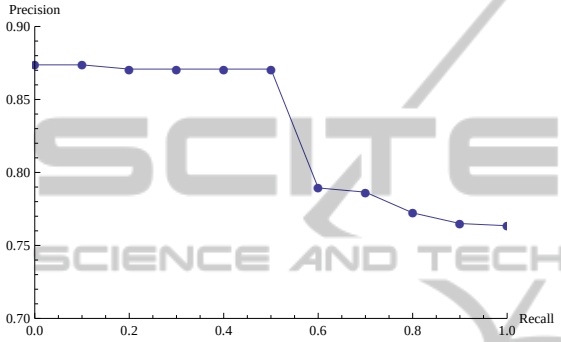


Figure 2: Precision-Recall graph.

For this purpose, a set of 50 topics (a common choice for TREC runs, see for example (Robertson, 2011)) 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 (Armstrong et al., 2009). For each topic a query was formulated, 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 20 (Craswell et al., 2011) or 10 may change precision results, but doesn't affect the relative performances of IR systems. Moreover, a precise analysis of IR performance is possible, even with a relatively short list of relevance judgments (Carterette et al., 2006).

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 the optimal estimations of each index field's weight, a minimization test was designed and implemented. Due to the high number of variables, the test implemented a simulation annealing strategy (Kirkpatrick et al., 1983). Different annealing schedules, initial state conditions and allowed transitions per temperature were tested. Simulations

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.

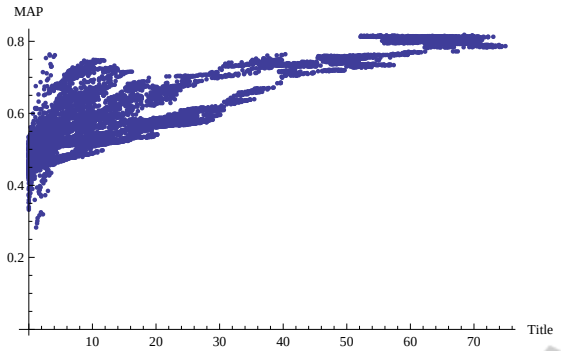
Metric	Value
# of queries	50
# of doc retrieved for topic	4312
# of relevant doc for topic	85
# of relevant doc retrieved for topic	84
MAP	0.8223
Geometric MAP	0.7216
Precision after retrieving R docs	0.7658
Main binary preference measure	0.9886
Reciprocal Rank of the 1^{st} relevant retrieved doc	0.8728

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 the 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

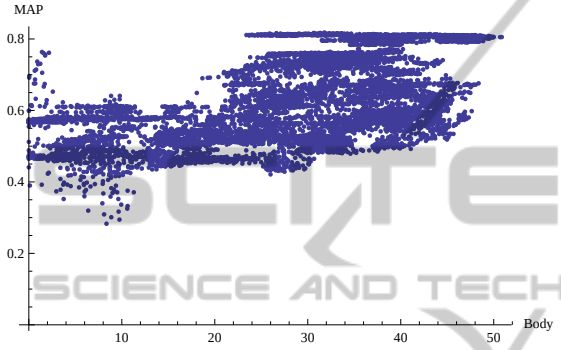
$$p_t = \begin{cases} 1 & \text{if } E_{i+1} < E_i \\ r < e^{-\Delta E/T} & \text{otherwise} \end{cases}$$

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. 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 Figure 1. Some semantic relevant index fields were found



(a) Title



(b) Body

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

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 Figure 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 Figure 2 and IR metrics in Table 3). Before the tests, the weight values used in the production server ($w_1 = 3.1$, $w_2 = 0.5$, $w_3 = 1.7$, $w_4 = 2.0$, $w_5 = 0.5$, $w_6 = 0.8$, $w_7 = 0.8$) produced a $MAP = 0.7552$, thus the optimization strategy yielded an increase in *MAP* of 8.885064%.

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 the fields' weight are being conducted, with different optimization approaches. A user behavior study is in progress, in order to understand both the user's preferences and satisfaction.

ACKNOWLEDGEMENTS

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

Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. (2009). Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 601–610, New York, NY, USA. ACM.

Carterette, B., Allan, J., and Sitaraman, R. (2006). 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. ACM.

Craswell, N., Fetterly, D., and Najork, M. (2011). The power of peers. In *Proceedings of the 33rd European conference on Advances in information retrieval*, pages 497–502. Springer-Verlag.

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

Krsten, J. and Eibl, M. (2011). A large-scale system evaluation on component-level. In *Advances in Information Retrieval*, volume 6611, pages 679–682. Springer Berlin / Heidelberg.

Robertson, S. (2011). On the contributions of topics to system evaluation. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 129–140. Springer Berlin / Heidelberg.