Int J Multimed Info Retr (2014) 3:147–159
DOI 10.1007/s13735-014-0058-8

REGULAR PAPER

Optimization of information retrieval for cross media contents in a best practice network

Pierfrancesco Bellini · Daniele Cenni · Paolo Nesi

Received: 15 July 2013 / Revised: 19 March 2014 / Accepted: 10 April 2014 / Published online: 8 May 2014

Abstract Recent challenges in information retrieval are related to cross media information in social networks including rich media and web based content. In those cases, the cross media content includes classical file and their metadata plus web pages, events, blog, discussion forums, comments in multilingual. This heterogeneity creates large complex problems in cross media indexing and retrieval for services that integrate qualified documents and user generated content together. Problems are also related to scalability, robustness and resilience to errors. Moreover, users expect to have fast and efficient indexing and searching services, from social media in best practice network services. This paper presents a model and an indexing and searching solution for cross media contents, addressing the above issues, developed for the ECLAP Social Network, in the domain of Performing Arts. Effectiveness and optimization analysis of the retrieval solution are presented with relevant metrics. The research aimed to cope with the complexity of a heterogeneous indexing semantic model, using stochastic optimization techniques, with tuning and discrimination of relevant metadata terms. The research was conducted in the context of the ECLAP European Commission project and services (http://www.eclap.eu).

P. Bellini · D. Cenni · P. Nesi (✉)
Distributed Systems and Internet Technology Laboratory,
Department of Information Engineering, University of Florence,
via S. Marta, 3, Florence, Italy
e-mail: paolo.nesi@unifi.it

P. Bellini
e-mail: pierfrancesco.bellini@unifi.it

D. Cenni
e-mail: daniele.cenni@unifi.it

Keywords Cross media content · Indexing · Searching · Search engines · Information retrieval · Stochastic optimization · Social networks · Best practice network

1 Introduction

The rapid growth of digital resources on the Web has opened new challenges in developing efficient and robust information retrieval solutions. A wide variety of contents, with different formats and metadata types, constitutes a heterogeneous set of resources difficult to deal with. A relevant example is provided by cross-media resources, which often include a rich set of metadata and mixed media, addressing serious issues, for example when building a digital content index. Typically, there is a need of tools for metadata extraction, schemas and metadata mapping rules and tools, multilingual metadata and content translation and certification. Information retrieval (IR) systems are required to give coherent answers with respect to typos or inflexions and must be efficient enough while sorting huge result lists. Search refinement, sorting and/or faceting techniques are major topics, especially in large multimedia repositories. Document parsing algorithms have to be fast enough to index high volumes of rich text documents and to support different types of content descriptors. Indexes and repositories have to be fully accessible, without significant downtime, in case of failures or major updates of the index structure, in production services (e.g., redefinition of index schema, index corruption).

Multilingual documents require query or metadata translation for information retrieval. The first approach reduces the memory usage and each document is stored only once in the index [35], while the second produces larger indexes and avoids query translation issues. Indeed, the automatic query translation process could create word ambiguity, poly-

semy, inflection and homonymy issues [1], especially in the case of short queries [25]. Disambiguation techniques can be applied, for example using co-occurrences of pair terms [58], or a general statistical approach. Query expansion [6], for example pseudo-relevance feedback technique [4,7], thesauri such as WordNet [20] or structured translation [49] can be used to increase the efficiency of a retrieval system.

Other possible approaches for dealing with multilingual documents refer to Self-Organizing Maps (SOMs) [28] or make use of sentence clustering before the translation process [18]. An alternative query translation approach involves the use of parallel or comparable Corpora [40]. They consist in a collection of natural language texts, where each document is translated in various languages; aligned parallel corpora are annotated to match each sentence in the source document with their respective translations. Thus, documents are comparable when they use the same vocabulary and deal with the same topic [32].

Ranking algorithms consist in ordering the output results list from the most to the least likely item [8]. Generally, ranking is based on location and frequency; documents with higher term occurrences are ranked higher. A notable example is the PageRank algorithm [13], which determines a page's relevance with a link analysis. Relevance feedback algorithm is based on the concept that a new query follows a modified version of the old one, derived by increasing the weight of terms in relevant items, and decreasing the weight of terms in non-relevant items. In order to overcome the limitations of traditional keyword-based search engines, fuzzy approaches are exploited too. In this case synonyms or typos are evaluated in terms of similarity with the current indexed tokens, to provide more complete results.

Relevant examples of fuzzy techniques application include semantic search [27], ontologies [2], Cloud Computing [30], image text analysis [12], query expansion [51], clustering [34] and popular search platforms such as Apache Lucene. Multidimensional dynamic taxonomies models (i.e., faceted search [43,52]) are also very popular, especially in e-commerce sites, where the user needs a way to easily explore the contents, and each facet can be represented with a taxonomy [42]. Document type detection and parsing algorithms for metadata extraction are a valuable key factor for integrating rich text resources (e.g., semi-structured or unstructured documents) in digital indexes, with the aim of Natural Language Processing (NLP) techniques; example approaches include machine learning methods [24], table metadata extraction (e.g., from PDFs [31]), context thesauri in conjunction with document analysis [46], DOM-based content extraction [22]. Typically, extracted information from unstructured documents can be organized as entities (i.e., noun phrases) and relationships between them, adjectives, tables and lists [45].

The evaluation of effectiveness plays a determinant role when assessing a retrieval 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 paradigm or other approaches [26]; 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.

Typically, 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 the queries to their 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 [50].

To overcome these limitations, some approaches start the retrieval evaluation without relevance judgments, making use of pseudo-relevance judgments [3,48,55]. Ranking strategies are often performed by comparing rank correlation coefficients (e.g., Spearman [15], Kendall Tau) with TREC official rankings. The IR effectiveness is assessed by computing relevant metrics such as precision, recall, mean average precision, R-precision, F-measure and normalized discounted cumulative gain (NDCG) [32]. Test collections and evaluations series are often used for a comparative study of retrieval effectiveness (e.g., TREC, GOV2, NTCIR and CLEF).

In the context of IR optimization, stochastic approaches have been exploited to improve the IR effectiveness; for example genetic algorithms have been used for improving the effectiveness of IR systems [36,37,41], for query reformulation [38], for query selection [17] and improving [57]. Other techniques make use of Fuzzy algorithms [33,47], local context analysis [56], clustering [59], and ranking improvement [54].

In this paper, the problem of cross media indexing was addressed. In particular, the case in which several different kinds of digital resources presenting heterogeneous number and types of metadata is considered. To this end, an indexing and searching solution was developed addressing problems such as heterogeneity, sparse and missing metadata fields, different languages and typos, with the aim of IR effectiveness optimization. These problems are typical of cases in which different kinds of content are indexed together such as forums, groups, blogs, events, pages, archives, audios, blogs, braille music, collections, documents, ePub, excel, flash, html, images, pdf, playlists, slides, smil, tools and videos.

Therefore, the proposed cross media content indexing solution was designed and tuned for the ECLAP social portal, best practice networks, in the area of Performing Arts. The technical solution is capable to cope with runtime exceptions, index schema updates, different metadata sets and content types. The ECLAP information model for cross-media integrates sources coming from 35 different international institutions [9,10].

It enhances and facilitates the user experience with full-text multilingual search, for a large range of heterogeneous types of content, with advanced metadata and fuzzy search, faceted search, content browsing and sorting techniques. The defined indexing and searching solution for ECLAP portal enabled a set of features including a range of rich content such as MPEG-21 (ISO IEC TR 21000-1:2001), web pages, forums and blogs posts, comments, events, images, rich text documents, doc, pdf, collections, playlists, ePub, 3D and animations. Due to the computational complexity of the ingestion process, the indexing service was implemented as a distributed parallel architecture.

This paper is structured as follows: Sect. 2 depicts an overview of ECLAP; Sect. 3 introduces the metadata model used; Sect. 4 discusses the Information Retrieval facilities at disposal in the ECLAP portal; Sect. 5 reports details about optimization tests and strategies followed in order to enhance the Information Retrieval effectiveness of the searching solution; Sect. 6 reports results data about the assessment of the ECLAP services; Sect. 7 reports conclusions and a sketch of future work.

2 ECLAP overview

ECLAP aims to create an online archive and portal in the field of Performing Arts to provide services to users from international institutions (mainly students and researchers). The ECLAP content and information is also indexed and searchable through the Europeana portal in the so-called EDM data model [19]. ECLAP Performing Arts material is inherently cross media and includes performances, lessons, master classes, teaching material in the form of videos, audio, documents and images. ECLAP can be seen as a support and tool for content aggregators (e.g., for content enrichment and aggregation, metadata quality assessment, preparing content for Europeana and for content distribution); working groups on best practice reports and articles, about tools for education and training, intellectual property and business models, digital libraries and archiving [11].

ECLAP networking and social services facilities include user group, discussion forums, mailing lists, connection with Social Networks, suggestions and recommendations to users, as intelligence tools (e.g., potential colleagues, using metrics based on static and dynamic user aspects, similar contents). Content distribution is available toward several channels: PC/Mac, tablets and mobiles. ECLAP portal features a large set of item formats, accessible through a search service with faceting refinement and ordering.

In ECLAP, users are able to deal with the above-mentioned content kinds, such as forums, groups, blogs, events, pages, archives, audios, blogs, braille music, collections, documents, ePub, excel, flash, html, images, pdf, playlists, slides, smil, tools and videos. Depending on credentials and a set of grants, each user can upload, create, improve and/or edit digital resources and their corresponding metadata.

3 Metadata model

ECLAP provides access to cross-media content. ECLAP Content Providers and Working Groups have associated the above-mentioned content containing: operas, performances, music scores, posters, lyrics, cards, video, pictures, events and all items related to performing arts production and performances. Moreover, *events* are present and shown to the users in the calendar. They are used to provide information on forthcoming events like conferences. ECLAP provides a multilingual taxonomy of terms for the classification of contents (for a total of 231 terms) organized in six different areas: Subject (e.g., Teaching, Philosophy, Multiculture), Genre (e.g., Comedy, Comic, Drama), Historical period (e.g., Contemporary, Classical, XX Century), Movement and style (e.g., Experimental, Theatre of the absurd), Performing arts type (e.g., Dance, Ballet, Music, Rock, Theatre, Noh), Management and organization (e.g., Performance, Choreography). Moreover, the full taxonomical associations and thematic groups related to each cross-media resource are indexed with the content, for full-text search purposes. There are also present aggregated cross media contents such as collections and playlists that aggregate other cross media contents.

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 metadata schema, designed to build the IR infrastructure of the ECLAP portal, consists of seven sets of metadata and descriptors (see [9,10] for further details): *Performing Arts* specific metadata, *Dublin Core* (DC) and *Dublin Core Terms* generic multilingual metadata (e.g., title, description, subject), multilingual *Taxonomy* and *Group* association, multilingual *Comments* and *Tags* associated with content, *Technical* metadata extracted from the digital resource, *Votes* provided by users and *Full Text* of the resource (for documents and web pages).

Table 1 ECLAP indexing model

Media types	DC (ML)	Technical	Performing arts	Full text	Tax, group (ML)	Comments, tags (ML)	Votes
No. of index fields[a]	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

ML multilingual, *DC* Dublin core, *Tax* taxonomy
[a] (No. of fields per metadata type) × (No. of languages)

Performing arts metadata include the information about the performance date and place (i.e., venue, city and country) and the information about the *premiere* of the performance depicted in the digital resource, as well as the details on the people involved in the creation process with their specific role (e.g., actor, director, light designer and choreographer). In the basic 15 *Dublin Core* elements and the extended *Dublin Core Terms* are present the more generic information on the digital resource, that can be also provided in multiple languages, while the *Technical* metadata contains information extracted from the digital resource (e.g., resource type, duration, dimension and number of pages) and information on the upload (e.g., upload date, user making the upload, content provider and associated groups).

Since the content was collected from 20 different partners, metadata sets differently fulfilled the standard DC schema. The ECLAP Index Model meets the metadata requirements of any digital content, while the indexing service follows a metadata ingestion schema. A single multilingual index was 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, to achieve an optimal IR performance.

In the indexing model of 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 nested, thus producing a hierarchically organized discussion forum.

4 Searching facilities

The goal of the searching service is to allow the users to easily locate and sort each type of content and to refine their queries for a more detailed result filtering, through a fast search interface, robust with respect to mistyping. High granularity of data is at disposal (i.e., advanced metadata search), with a detailed search interface. Textual contents (e.g., web pages, forums, comments, groups and events) and media contents are fully searchable in the ECLAP portal, and queries may produce heterogeneous list of results (e.g., blog posts, groups, events, comments and PDFs). Querying for a term contained in a page, blog, forum or cross-media content, produces a match with the set of resources containing that search term, thus producing a list of formatted results. Queries related to taxonomy terms attached to a content provide a pertinent match too. Relevance scoring has to take into account different weights for each document's metadata field; a same term occurring in different document fields is expected to provide different scoring results (i.e., a higher field's weight means a higher relevance of that field).

In order to simplify the users' work, searching is provided as an easy to use full text and advanced search service. The full text frontal search is in the top center of the portal. Each query is automatically tokenized and lowercased, before assembling the query string (i.e., a combination of weighted OR boolean clauses, with escaping of special characters) and then sent to the indexing service.

Depending on the enabled languages in the portal, each active language field is included in the query string for full-text search. Advanced search is reachable from the top center portal menu and provides language, partner and metadata filtering. The user is allowed to compose an arbitrary number of boolean clauses in the advanced search page, thus allowing the building of a rich metadata query; for example, by restricting the search to some metadata fields that only match any or all of them (OR/ALL).

Fuzzy logic is transparently applied in full-text queries; hence even a query with typos can return coherent results. The query term is compared to similar terms in the index, for retrieving documents with a high degree of similarity (e.g., "documant" should match "document"), thus allowing an efficient search in case of mistyping. The string metric used (Levenshtein or edit distance [29]) allows measuring the similarity between two search terms by evaluating the minimum number of transformations needed to change one search term into another.

Table 2 ECLAP metadata schema

Metadata type	No. of fields	Multilingual	Index fields	No. of 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	13	*N*	13	*n*
Votes	1	*N*	1	1
Total	87	–	554	–

This fuzzy similarity weight is customizable by the administrator in the portal (a weight $w < 1$ means fuzzy logic, while a weight $w = 1$ means boolean logic). In the frontal search service, a deep search checkbox is also available, allowing the user to enable/disable such functionality. If enabled, the query string is prefixed and suffixed with a special wildcard, in a transparent way to the user, to allow searching of substrings in the index (e.g., query "test" matches "testing").

Boosting of terms is customizable on the portal, for tuning and enhancing the importance of certain metadata. On the basis of the performed experiments, the best appreciation was obtained by giving more relevance to some fields with respect to others (i.e., title, subject, description). The administrator is able to change the boosting of the main search fields, though boost values can be extended to the whole set of metadata. Each field of the ECLAP document structure is boosted with its predefined value at query time.

Faceted search is activated on the results of both simple and advanced search. In order to accomplish the faceting count, each faceted term is indexed un-tokenized in the ECLAP index. Facet parameters are appended to the query term, and facet counts are evaluated from the output result by a service module, before rendering. The user can select or remove any facet in any order to refine the search. Adding or removing a facet results in adding or removing a search filter and performs again the search query with or without it.

- Dublin core: resource category, format, type, classification, creator, content language;
- Technical: duration, video quality, device, publisher, source metadata language and upload time;
- Group, taxonomy: genre, historical period, performing arts, coded subject.

These facets can be subject to change. For instance, locations and dates, different for each historical period, can be added.

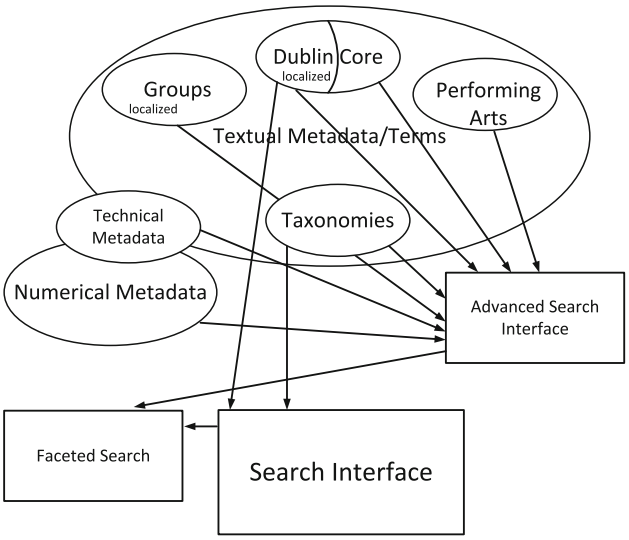

Fig. 1 Metadata mapping

Given the metadata schema of Table 2, the index structure was built mapping each metadata in a separate index field, with support for localizations (see Fig. 1. Some semantic metadata (e.g., textual metadata such as taxonomical terms and parsed text from rich text documents) were mapped in a separate field for each available localization; other textual metadata (i.e., Performing Arts Metadata) are provided in single language, numerical metadata (e.g., video resolution and content id) were mapped in non-localized fields, and technical metadata are typically numerical with a few of them as enumerate string. Fields of date type were mapped in special timestamp fields, to allow the use of range queries (i.e., search filtering in a range interval). Full-text searches are conducted using the most relevant metadata from those of textual type (see Sect. 5.1). Seventeen metadata were also mapped to special index fields to allow a faceting count for search refinements. Considering that facets are applied after issuing a query, to speed up the rendering of the results, only a subset of the available facets was enabled, that are resource format, thematic group, Dublin core language, partner, default metadata language, upload time and some taxonomical terms (i.e., genre, historical period, subject and type of performing art).

5 Information retrieval optimization

5.1 Weighted query model

The indexing schema consists of 554 metadata fields, belonging to eight large categories (see Table 2). With such a large metadata schema, that is also sparsely populated, a flat indexing would lead to poor IR effectiveness and thus unsatisfactory results beneath the users' expectations. Thus the opti-

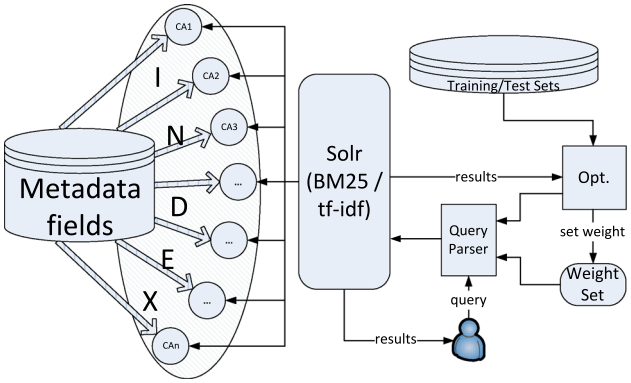

Fig. 2 Metadata identification and optimization

mization model is applied to maximize the precision and recall, obtaining back the weights and relevant fields.

On the other hand, considering the huge dimension of the metadata set, an exhaustive optimization approach addressing all the 554 fields and weights would be unachievable. A possible solution is to identify the most relevant fields or groups of them that may play the role of good descriptors for the cross-media content. The goal of this phase can be to reduce the number of fields or groups of them to a manageable number, for example by reducing the optimization process to less than ten dimensions.

Figure 2 illustrates the structure to identify the most relevant fields or groups, and to find the best weights that optimize the effectiveness of the retrieval system (for details, see Sect. 5.2). In Fig. 2, the Solr block represents the scoring algorithms at disposal (i.e., BM25 and tf-idf) used during the tests.

A platform to automatically issue full-text queries was built upon the searching facility, making use of a predefined number of queries and corresponding to training and test sets. The optimization block implements stochastic algorithms for dynamically tuning the weighted parameters of the query, to be sent to the query parser.

In order to find a reasonable number of field groups, a number of combinations was tested, by indexing together different combinations of metadata fields belonging to the same domain, thus obtaining different field group categories ($CA_1, \ldots, CA_n$). At the end of the process, the metadata groups leading to the most relevant results were selected, producing the results illustrated in Table 3. For example, field group category of titles includes all versions of titles and their corresponding translations.

Table 3 reports the top significant metadata groups according to our metric analysis on the ECLAP digital library. *Dim* is the maximum dimension of the field group (i.e., a metadata group consisting of L languages and M different instances would have dimension $N = L \times M$), *Type* is the type of metadata (i.e., text, date, or ID that is an identifier), *Instances* is the number of items with that metadata populated, *Chars* is the total chars used in the field group, *Distincts* is the number of different occurrences of a field, *Avg Chars* is the average number of chars per instance and *Avg Distincts* is the product *Avg Chars* × *Distincts*.

This last metric constitutes a measure of the informative metadata's content, representing the total amount of information for an index field, and was used to sort the field groups. Thus, several experiments were performed to identify a man-

Table 3 Top significant metadata

Field group	Dim	Type	Instances	Chars	Distincts	Avg chars	Avg distincts
Text	*N*	Text	52,288	2,674,345,614	52,288	51,146.45	2,677,145,614.00
Description	*N*	Text	219,448	42,620,235	88,174	194.22	17,124,770.34
Body	*N*	Text	891	8,251,017	856	9,260.40	7,926,902.98
Title	*N*	Text	182,956	8,348,823	101,364	45.63	4,625,538.90
Extraction date	1	Date	172,314	3,273,966	170,659	19.00	3,242,521.00
Subject	*N*	Text	126,186	4,481,193	42,935	35.51	1,524,733.50
Identifier	1	ID	125,911	1,477,569	104,594	11.74	1,227,413.43
Description table of contents	*N*	Text	23,163	1,236,725	22,965	53.39	1,226,153.33
Taxonomy	*N*	Text	123,925	1,048,797	79,275	8.46	670,916.94
Source	1	ID	54,738	2,445,586	14,324	44.68	639,968.10
Relation references	*M*	ID	4,612	578,659	2,436	125.47	305,640.36
Relation is referenced by	*M*	ID	3,556	1,383,479	530	389.05	206,199.06
Creator	1	ID	146,523	2,436,848	8,639	16.63	143,676.62
Names extraction date	1	Date	172,324	2,240,212	8,617	13.00	112,021.00
Publisher	1	ID	76,063	2,259,989	2,377	29.71	70,625.58
Date	1	Date	112,405	697,096	10,012	6.20	62,090.88
Contributor	*N*	Text	50,477	1,209,163	2,465	23.95	59,048.41
…	…	…	…	…	…	…	…

ageable number of field groups that can constitute a trade off between effectiveness and performance, also taking into account the population of the metadata set to be indexed. Therefore, subsets of the most informative field groups were used to perform rounds of optimization, to assess the IR effectiveness. This allowed to identify the most appropriate field groups, according to the user expectations. At the end of this iterative process the most significant field groups were selected.

As a result, some metadata fields presented a limited contribution (e.g., *abstract*, *coverage*, *accrualPolicy*, *relation* that occur sporadically). Numerical or date related metadata did not provide a semantic contribution and thus were not considered (e.g., *date*, *uploadTime*, *objectId*, *resourceType*). The metadata *Description table of contents* was not relevant in influencing the IR effectiveness, since it contains data already present in *title* and *description*.

The most semantic relevant seven metadata field groups (i.e., *text*, *title*, *body*, *description*, *contributor*, *subject*, *taxonomy*) were identified. They are all multilingual fields adopted as catchall, on which the search module can apply a field boosting to every metadata field group at query time (i.e., documents matching a query term have their score multiplied by a weight factor). A boolean weighted clause b can be defined as

$$\begin{aligned} b := {} & (title\colon q)^{w_1} \vee (body\colon q)^{w_2} \vee (description\colon q)^{w_3} \\ & \vee \ (subject\colon q)^{w_4} \vee (taxonomy\colon q)^{w_5} \\ & \vee \ (contributor\colon q)^{w_6} \vee (text\colon q)^{w_7}, \end{aligned} \tag{1}$$

where (i) $(w_1, w_2, \ldots, w_7)$ are the boosting weights related to each query field; (ii) *title* includes the set of DC titles in all their languages; (iii) *body* is the parsed content of a html resource; (iv) *description* includes the set of descriptions in all their languages (e.g., abstract, table of contents, reference); (v) *subject* includes the set of subjects in all their languages (e.g., keywords, key phrases, classification codes); (vi) *taxonomy* is a content-related taxonomical classification in all languages with hierarchies; (vii) *contributor* is a contribution to the content (e.g., persons, organizations, services) and (viii) *text* is a full-text content, parsed from the resource (e.g. doc, pdf); q is the search query.

The methodology described is general enough to be applied to a large range of cross-media contents and data sets, to discriminate the most appropriate field groups to use in an information retrieval system. Moreover, in the proposed solution all the other metadata fields are accessible and were indexed to be retrieved by advanced full-text and specific queries and interface. The advanced search interface provides AND/OR operators for combining single search fields, thus allowing to exclude from the most informative field groups those that are dates or IDs. This allows recovering the cross media contents even by using their ID (sub pattern of the ID) or dates.

5.2 Scoring formulas optimization

The platform to analyze the optimal estimations for each index field's weight in (1) included two stochastic minimization tests. Considering the relatively high number of variables, the tests implemented a simulated annealing strategy (testing various annealing schedules, initial state conditions and allowed transitions per temperature), or a genetic algorithm-based evolution process. The tests were conducted issuing subsequent queries to the retrieval system, using both the tf-idf scoring and the BM25 scoring formulas [5,39].

In this paper the test configurations are labeled with SA_1, SA_2, GA_1, GA_2, indicating the simulated annealing based tests with tf-idf and BM25 scoring, and the genetic algorithm-based tests with tf-idf and BM25 scoring, respectively. For the purpose, 200 topics were collected for the training phase and 50 topics for testing the results, with corresponding human-assessed relevance judgments. Fifty topics is a common choice used when evaluating IR systems, for example in TREC or other contexts [14,16,44,53,60].

In the following section the results regarding the abovementioned test configurations are presented by providing for each of them optimal weights and effectiveness related measures.

5.2.1 Simulated annealing

Simulations took place by defining the state of the system as a vector of field weights $w_i = (w_1, w_2, \ldots, w_7)$. A run of 200 queries was issued for each state condition, to retrieve the corresponding search results with the most relevant IR measures. For each run, the Mean Average Precision (*MAP*) was computed and $(1 - MAP)$ was set as the energy for the current state. *MAP* is defined as the arithmetic mean of average precision for the information needs so that it can be thought as an approximation of the area under the precision–recall curve. Considering the *Metropolis Criterion*, a state transition probability p_t is defined by

$$p_t = \begin{cases} 1, & \text{if } E_{i+1} < E_i \\ r < e^{-\Delta E/T}, & \text{otherwise,} \end{cases} \tag{2}$$

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* and r is a random number in the interval $(0, 1)$. The *annealing schedule* was defined as $T(i + 1) = \alpha T(i)$, with $\alpha = 0.8$, with $T = 0.001$ as stopping condition, since no other significant improvements can be observed. Thirty random transitions were tested for the temperature

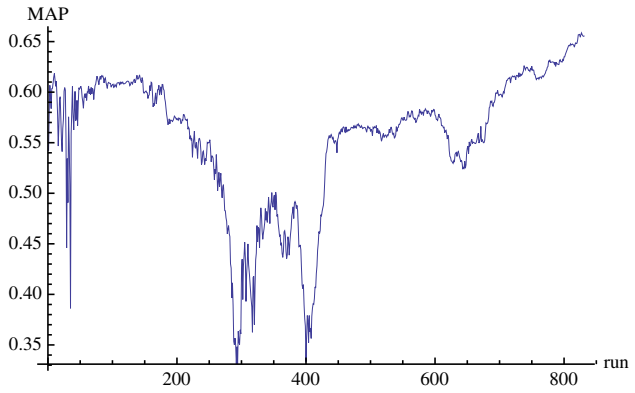

Fig. 3 MAP vs test runs (SA_1)

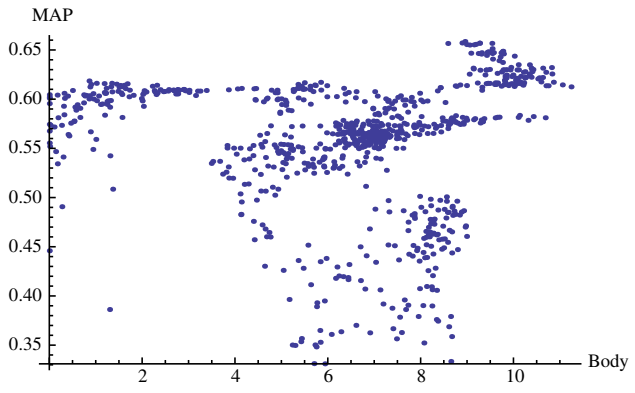

Fig. 4 Body vs MAP (SA_1)

of each iteration. A smoother annealing schedule is more likely to exhibit convergence, but generally requires a bigger simulation time. Stopping conditions were assumed by counting the number of successful transitions that occurred during each iteration. Other popular choices include logarithmic schedules such as $T(i) = c/log(1+i)$ [21,23].

Figure 3 reports the best simulation configuration, obtained with (SA_1), exhibiting convergence and system equilibrium. Some metadata fields were found to have a limited relevance weight, with respect to the relevance score (i.e., subject, taxonomy and contributor). Reducing the number of boolean clauses to be processed by the IR system is indeed an advantage that produces a higher search speed. Scatter plots of field weights vs *MAP*, collected during the test run, showed a relevant dispersion across a huge range of high energy values, both for tf-idf and BM25 scoring (see Figs. 4, 5, 6, 7). SA_1 tests gave a better result than SA_2.

For tf-idf scoring, the minimization strategy resulted in an energy minimum at $w_1 = 6.08$, $w_2 = 8.95$, $w_3 = 0.0$, $w_4 = 9.46$, $w_5 = 0.75$, $w_6 = 1.69$, $w_7 = 0.20$, with $MAP = 0.6586$. The optimal state was then validated with the test set obtaining $MAP = 0.6609$ (see the precision–recall curve for SA and GA in Fig. 8, *MAP* scatter plot for SA_1 in Fig. 9 and IR measures for all tests in Table 4). The scatter *MAP* of

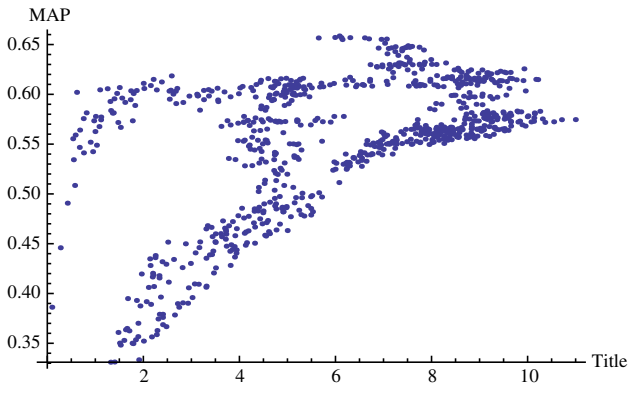

Fig. 5 Title vs MAP (SA_1)

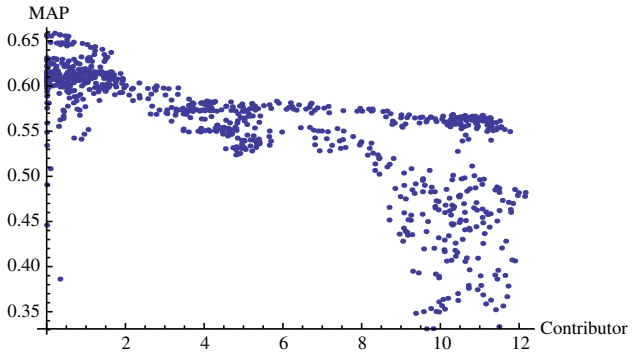

Fig. 6 Contributor vs MAP (SA_1)

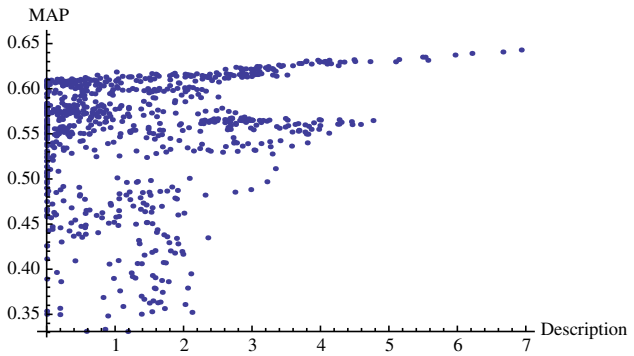

Fig. 7 Description vs MAP (SA_1)

Fig. 9 depicts the behavior of *MAP* at time t versus that at time $t+1$, thus giving the evidence of a convergent behavior. For BM25 scoring, the minimization strategy resulted in an energy minimum at $w_1 = 6.75$, $w_2 = 0.14$, $w_3 = 0.99$, $w_4 = 5.65$, $w_5 = 6.07$, $w_6 = 5.40$, $w_7 = 3.14$, with $MAP = 0.6047$. The optimal state was then validated with the test set obtaining $MAP = 0.6279$. The observed patterns thus suggest a relevant sensitivity to initial conditions and random seeds. The behavior of the precision–recall curve for both SA_1 and SA_2, during some test runs, is depicted in Fig. 10.

Before the optimization tests, the weight values used in the production server ($w_1 = 3.1$, $w_2 = 0.5$, $w_3 = 1.7$,

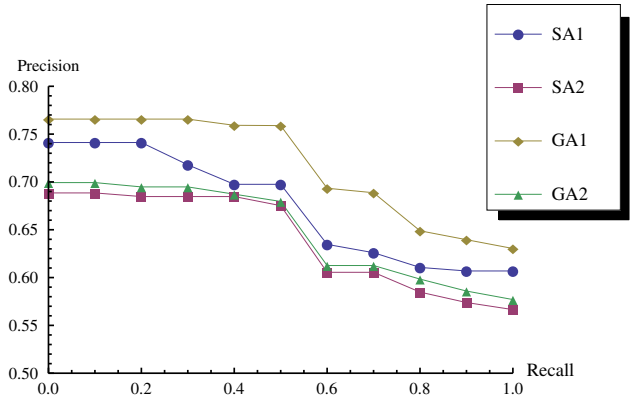

Fig. 8 Precision–recall (SA, GA)

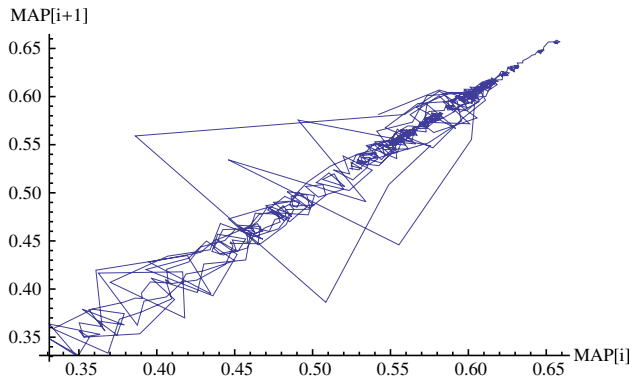

Fig. 9 MAP scatter (SA_1)

$w_4 = 2.0$, $w_5 = 0.5$, $w_6 = 0.8$, $w_7 = 0.8$), produced $MAP = 0.5957$. The optimization strategy yielded an increase in MAP of ~9.86 %.

5.2.2 Genetic algorithms

Another stochastic approach to IR optimization makes use of Genetic Algorithms. Each field weight, that constitutes the boolean query expression in (1), was defined as a gene of the sample chromosome. The test was built with a population of 100 chromosomes (the more the chromosomes, the larger the number of solutions, but with a longer computation time, due to the fact that the population will require more time to evolve for each round). The upper limit of maximum allowed evolutions was set to 10. The field's weight values of the fitness function f, were evaluated by computing their corresponding gene values for the current chromosome. For each vector of weights $\mathbf{w_i} = (w_1, w_2, \ldots, w_7)$ the indexing service was queried in order to find the corresponding Mean Average Precision.

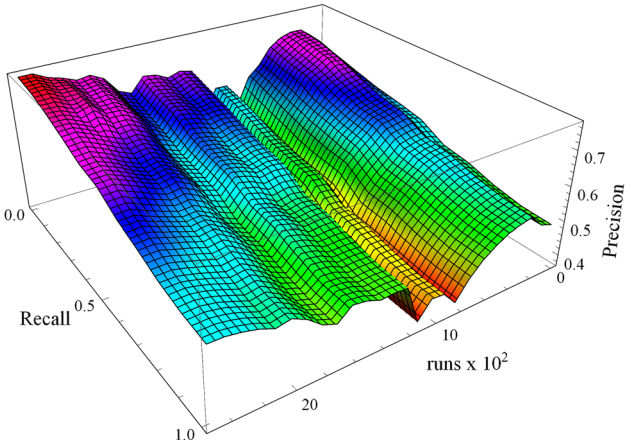

Fig. 10 Precision–recall vs test runs (annealing)

The fitness function f was then normalized to exaggerate the difference between the higher values, assuming $f_n = 2^{10f}$. Figure 11 shows the convergence of MAP across the test runs for GA_1. Table 4 shows the most relevant IR measures collected for this simulation strategy. Also in this case a considerable dispersion across a huge range of energy values was noticed, for every index field (see Figs. 12, 13, 14, 15). The MAP_{GA} value was consistent to what obtained with the annealing strategy. GA_1 tests gave a better result than GA_2.

Table 4 IR measures for the optimal run

Measure	SA_1	GA_1	SA_2	GA_2
No. of learning queries	200	200	200	200
No. of test queries	50	50	50	50
No. of cross media retrieved for topic	4,365	4,376	4,380	4,380
No. of relevant cross media for topic	85	85	85	85
No. of relevant cross media retrieved for topic	81	78	81	82
MAP	0.6609	0.7022	0.6279	0.6371
Geometric MAP	0.4185	0.3468	0.3431	0.4279
Precision after retrieving R cross media	0.5574	0.6273	0.5563	0.5605
Main binary preference measure	0.9545	0.9280	0.9583	0.9697
Reciprocal rank of the first relevant retrieved cross media	0.7265	0.7474	0.6734	0.6843

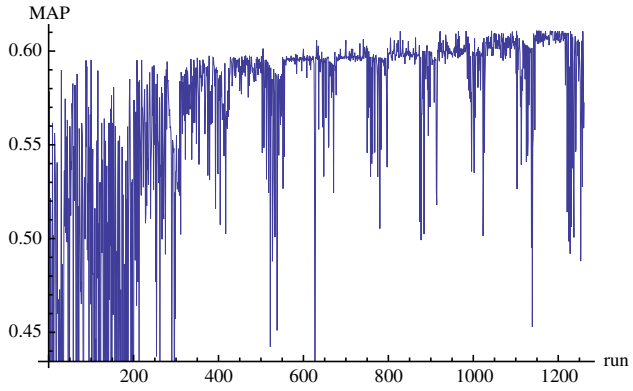

Fig. 11 MAP vs test runs (GA_1)

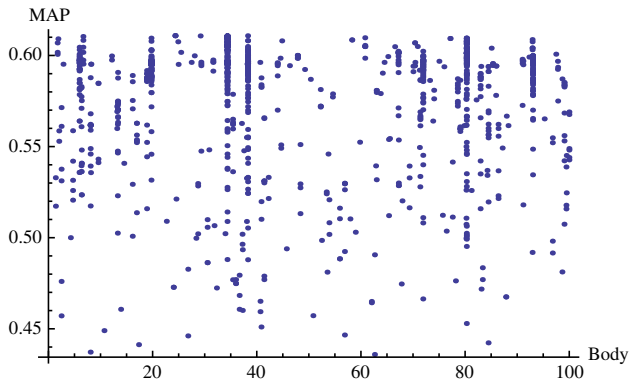

Fig. 12 Body vs MAP (GA_1)

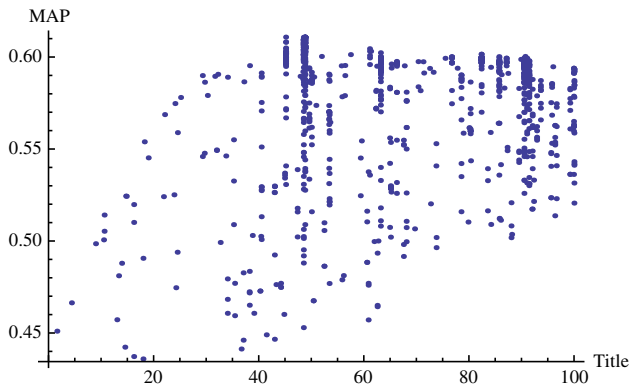

Fig. 13 Title vs MAP (GA_1)

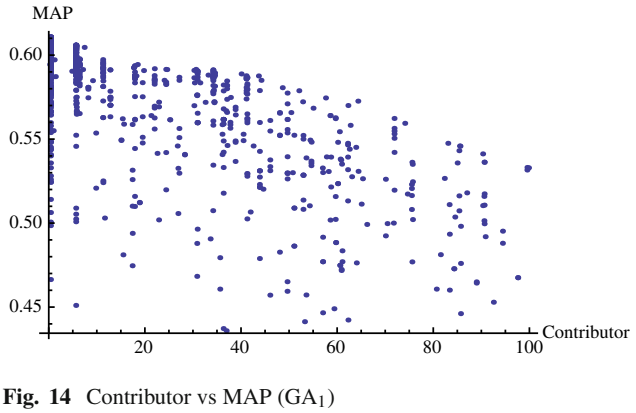

Fig. 14 Contributor vs MAP (GA_1)

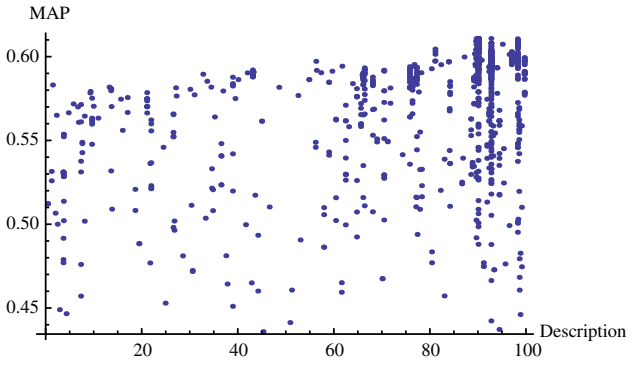

Fig. 15 Description vs MAP (GA_1)

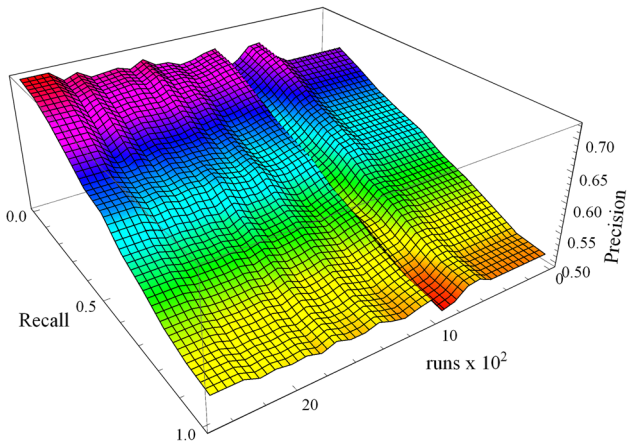

Fig. 16 Precision–recall vs test runs (GA)

For tf-idf scoring, the minimization strategy resulted in an energy minimum at $w_1 = 48.77$, $w_2 = 24.30$, $w_3 = 0.50$, $w_4 = 90.02$, $w_5 = 12.55$, $w_6 = 19.86$, $w_7 = 0.36$, with $MAP = 0.6109$. The optimal state was then validated with the test set obtaining $MAP = 0.7022$.

For BM25 scoring, the minimization strategy resulted in an energy minimum at $w_1 = 97.31$, $w_2 = 83.24$, $w_3 = 15.30$, $w_4 = 80.83$, $w_5 = 27.00$, $w_6 = 30.52$, $w_7 = 35.17$, with $MAP = 0.6067$. The optimal state was then validated with the test set obtaining $MAP = 0.6371$. The estimated MAP difference (ΔMAP) between the best two tests SA_1 and GA_1 was ~5.88 % ($\Delta MAP = MAP_{GA_1} - MAP_{SA_1} = 0.7022 - 0.6609 = 0.04$). Figures 8, 17, show, respectively, the MAP scatter plot for GA_1, and the precision–recall curve

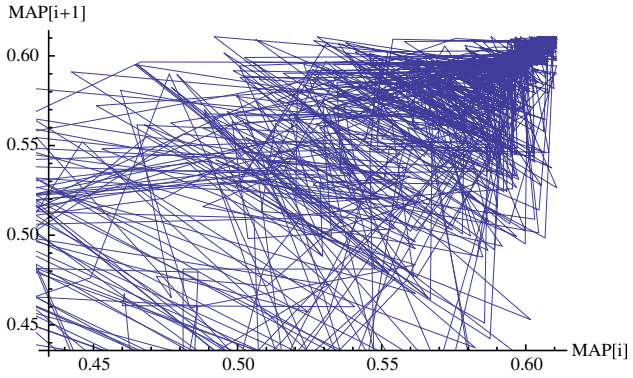

Fig. 17 MAP scatter (GA_1)

for SA and GA, obtained during their best performing simulation runs. The progress of the Precision–recall curve for GA_1, collected through some test runs, is depicted in Fig. 16.

6 Results assessment

As results, the four test environments considered (SA_1, SA_2, GA_1, GA_2) produced different precision/recall results as depicted in Fig. 8. It is evident that the best results were obtained with GA_1; details are reported in Table 4. It is worth noting that the weights reported in the previous sections were estimated in the learning phases by using a set of 200 topics, and they were applied in the weighted model for validation against 50 topics.

6.1 Search facility assessment

The analysis was performed in the period July 1, 2012–June 30, 2013. Some of the data were collected with the aim of Google Analytics, while others were directly collected with internal logs. In that period, a total number of 55,631 visits to the portal (of which 34,109 unique visitors) was registered. The portal collected a total of 598,820 views of pages/resources, and thus there were 10.76 views per visit. These data were associated with 5.01 min of mean time of web sessions. A total of 609,560 content accesses were registered (view, play and download, downloads are 1.76 % of the total). Table 5 depicts some data about searching activities performed by the ECLAP community (sorted by partnership), through queries and static menu lists available on the portal (the numbers are referred to the same period). The first column reports the number of performed full text queries, obtaining a high ratio of query per visit (62.94 %). This means that the 62.94 % of visitors performed at least one query (35,019 full text queries, 1,444 faceted queries, and 557 advanced queries). Most of the queries were issued by anonymous users. Registered users are those that are regularly registered on the portal, and do not belong to one of the institutions that have signed an agreement with ECLAP as partners or affiliated partners. In Table 5, the data related to other search results is reported to put in evidence the usage of faceted search, last posted, featured and the most popular content lists. The last line of the table reports the number of clicks performed after a search or a click on those content lists. Clicks on last posted contents and featured contents were performed through the portal menu, at the top of the home page.

7 Conclusions

This paper proposed a specific model for weighting metadata contributions, and a corresponding indexing and searching solution for cross-media contents, addressing the typical issues related to heterogeneity and sparsity of associated content descriptors. It was developed and validated for the ECLAP Social Network, in the domain of Performing Arts. Effectiveness and optimization analysis of the retrieval solution were presented with relevant metrics, obtaining good results using stochastic optimization techniques (genetic algorithm combined with the tf-idf scoring IR formula). The research aimed to cope with the complexity of a heterogeneous indexing semantic model, using stochastic optimization techniques, with tuning and discrimination of relevant metadata terms. The research was conducted in the context of the ECLAP European Commission project and services (http://www.eclap.eu).

Table 5 Queries and content lists

Users	No. of full text queries	No. of faceted queries	No. of last posted contents	No. of featured contents	No. of popular contents
Registered	4,747	167	34	56	55
Partners	6,665	325	30	91	31
Anonymous	23,607	952	1,469	533	706
Total	35,019	1,444	1,533	680	792
Clicks after query	17,756	589	1,150	7,448	3,407

This paper reported all the experience details that allow replicating the results in different contexts, in which similar problems can be faced.

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. Abusalah M, Tait J, Oakes MP (2005) Literature review of cross language information retrieval. In: WEC (2)'05, pp 175–177
2. Akhlaghian F, Arzanian B, Moradi P (2010) A personalized search engine using ontology-based fuzzy concept networks. In: Proceedings of the 2010 international conference on data storage and data engineering. IEEE Computer Society, Washington, DC, DSDE '10, pp 137–141. doi:10.1109/DSDE.2010.30
3. Aslam JA, Savell R (2003) 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. ACM, New York, SIGIR '03, pp 361–362. doi:10.1145/860435.860501
4. Attar R, Fraenkel AS (1977) Local feedback in full-text retrieval systems. J ACM 24(3):397–417
5. Baeza-Yates RA, Ribeiro-Neto B (1999) Modern information retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston
6. Ballesteros L, Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. SIGIR Forum 31(SI):84–91
7. Ballesteros L, Croft WB (1998) Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, SIGIR '98, pp 64–71
8. Belkin NJ, Croft WB (1987) Retrieval techniques. In: Williams ME (ed) Annual review of information science and technology, vol 22. Elsevier Science Inc., New York, pp 109–145. http://dl.acm.org/citation.cfm?id=42502.42506
9. Bellini P, Nesi P (2013) A linked open data service for performing arts. In: ECLAP, 2nd international conference on information technologies for performing arts, media access and entertainment, LNCS. Springer, Berlin
10. Bellini P, Cenni D, Nesi P (2012) On the effectiveness and optimization of information retrieval for cross media content. In: KDIR, 4th international conference on knowledge discovery and information retrieval. SciTePress, pp 344–347
11. Bellini P, Bruno I, Cenni D, Nesi P, Paolucci M, Serena M (2013) A new generation digital content service for cultural heritage institutions. In: Nesi P, Santucci R (eds) Information technologies for performing arts, media access, and entertainment, Lecture Notes in Computer Science, vol 7990. Springer, Berlin, pp 26–38. doi:10.1007/978-3-642-40050-6_3
12. Berkovich S, Inayatullah M (2004) A fuzzy find matching tool for image text analysis. In: Proceedings of international symposium on information theory, 2004. ISIT 2004, pp 101–105. doi:10.1109/AIPR.2004.2
13. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117
14. Buckley C, Voorhees EM (2000) Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '00, pp 33–40
15. Callan J, Connell M, Du A (1999) Automatic discovery of language models for text databases. In: Proceedings of the 1999 ACM SIGMOD international conference on management of data. ACM, New York, USA, SIGMOD '99, pp 479–490. doi:10.1145/304182.304224
16. Carterette B, Allan J, 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. ACM, New York, SIGIR '06, pp 268–275. doi:10.1145/1148170.1148219
17. Cecchini RL, Lorenzetti CM, Maguitman AG, Brignole NB (2008) Using genetic algorithms to evolve a population of topical queries. Inf Process Manage 44(6):1863–1878
18. Chen HH, Kuo JJ, Su TC (2003) Clustering and visualization in a multi-lingual multi-document summarization system. In: Proceedings of the 25th European conference on IR research. Springer, Berlin, ECIR'03, pp 266–280
19. Europeana portal. http://europeana.eu
20. Fellbaum C (ed) (1998) WordNet an electronic lexical database. The MIT Press, Cambridge
21. Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741. doi:10.1109/TPAMI.1984.4767596
22. Gupta S, Kaiser G, Neistadt D, Grimm P (2003) Dom-based content extraction of html documents. In: Proceedings of the 12th international conference on World Wide Web. ACM, New York, WWW '03, pp 207–214. doi:10.1145/775152.775182
23. Hajek B (1988) Cooling schedules for optimal annealing. Math Oper Res 13(2):311–329. doi:10.1287/moor.13.2.311
24. Han H, Giles CL, Manavoglu E, Zha H, Zhang Z, Fox EA (2003) Automatic document metadata extraction using support vector machines. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries. IEEE Computer Society, Washington, JCDL '03, pp 37–48. http://dl.acm.org/citation.cfm?id=827140.827146
25. Hull DA, Grefenstette G (1996) Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '96, pp 49–57
26. Kürsten J, Eibl M (2011) A large-scale system evaluation on component-level. Advances in information retrieval, vol 6611. Springer, Berlin, pp 679–682
27. Lai LF, Wu CC, Lin PY, Huang LT (2011) Developing a fuzzy search engine based on fuzzy ontology and semantic search. In: 2011 IEEE international Conference on fuzzy systems (FUZZ), pp 2684–2689. doi:10.1109/FUZZY.2011.6007378
28. Lee CH, Yang HC (2000) Towards multilingual information discovery through a som based text mining approach. In: PRICAI workshop on text and web mining, pp 80–87
29. Levenshtein V (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 10:707
30. Li J, Wang Q, Wang C, Cao N, Ren K, Lou W (2010) Fuzzy keyword search over encrypted data in cloud computing. In: INFOCOM, 2010 proceedings IEEE, pp 1–5. doi:10.1109/INFCOM.2010.5462196
31. Liu Y, Mitra P, Giles CL, Bai K (2006) Automatic extraction of table metadata from digital documents. In: Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries. ACM, New York, JCDL '06, pp 339–340. doi:10.1145/1141753.1141835

32. Manning CD, Raghavan P, Schtze H (2008) Introduction to information retrieval. Cambridge University Press, New York
33. Martin-Bautista M, Vila MA, Sanchez D, Larsen H (2000) Fuzzy genes: improving the effectiveness of information retrieval. In: Proceedings of the 2000 congress on evolutionary computation, vol 1, pp 471–478
34. Matsumoto T, Hung E (2010) Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation. In: 2010 IEEE international conference on fuzzy systems (FUZZ), pp 1–8. doi:10.1109/FUZZY.2010.5584771
35. McCarley JS (1999) Should we translate the documents or the queries in cross-language information retrieval? In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, Stroudsburg, ACL '99, pp 208–214
36. Pathak P, Gordon M, Fan W (2000a) Effective information retrieval using genetic algorithms based matching function adaptation. In: Proceedings of the 33rd Hawaii international conference on system science (HICSS)
37. Pathak P, Gordon M, Fan W (2000b) Effective information retrieval using genetic algorithms based matching functions adaptation. In: Proceedings of the 33rd annual Hawaii international conference on system sciences, 2000, vol 1, p 8. doi:10.1109/HICSS.2000.926653
38. Pérez-Agüera JR (2007) Using genetic algorithms for query reformulation. In: Proceedings of the 1st BCS IRSG conference on future directions in information access. British Computer Society, Swinton, UK, FDIA'07, pp 15–15
39. Pérez-Iglesias J, Pérez-Agüera JR, Fresno V, Feinstein YZ (2009) Integrating the probabilistic models BM25/BM25F into lucene. CoRR abs/0911.5046
40. Picchi E, Peters C (1998) Cross-language information retrieval: A system for comparable corpus querying. In: Grefenstette G (ed) Cross-language information retrieval, The Springer International Series on Information Retrieval, vol 2. Springer, US, pp 81–92. doi:10.1007/978-1-4615-5661-9_7
41. Radwan AAA, Latef BAA, Mgeid A, Ali A, Sadek OA (2006) Using genetic algorithm to improve information retrieval systems
42. Sacco G (2007) Research results in dynamic taxonomy and faceted search systems. In: 18th international workshop on database and expert systems applications, 2007. DEXA '07, pp 201–206. doi:10.1109/DEXA.2007.75
43. Sacco GM, Tzitzikas Y (2009) Dynamic taxonomies and faceted search: theory, practice, and experience, 1st edn. Springer, New York
44. Sanderson M, Zobel J (2005) 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. ACM, New York, SIGIR '05, pp 162–169. doi:10.1145/1076034.1076064
45. Sarawagi S (2008) Information extraction. Found Trends databases 1(3):261–377. doi:10.1561/1900000003, http://dx.doi.org/10.1561/1900000003
46. Shepherd M, Watters C, Young J (2004) Context thesaurus for the extraction of metadata from medical research papers. In: Proceedings of the 37th annual Hawaii international conference on system sciences (HICSS'04), Track 6, vol 6. IEEE Computer Society, Washington, DC, HICSS '04, pp 60138.2. http://dl.acm.org/citation.cfm?id=962754.963037
47. Snášel V, Abraham A, Owais S, Plato J, Krömer P, (2009) Optimizing information retrieval using evolutionary algorithms and fuzzy inference system. In: Abraham A, Hassanien AE, Carvalho A (eds) Foundations of computational intelligence, vol 4, Studies in computational intelligence, vol 204. Springer, Berlin, pp 299–324
48. Soboroff I, Nicholas C, Cahan P (2001) Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '01, pp 66–73. doi:10.1145/383952.383961
49. Sperer R, Oard DW (2000) Structured translation for cross-language information retrieval. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '00, pp 120–127
50. Spink A, Greisdorf H (2001) Regions and levels: measuring and mapping users' relevance judgments. J Am Soc Inf Sci Technol 52(2):161–173
51. Takagi T, Tajima M (2001) Query expansion using conceptual fuzzy sets for search engine. In: The 10th IEEE International conference on fuzzy systems, 2001, vol 3, pp 1303–1308. doi:10.1109/FUZZ.2001.1008898
52. Tunkelang D (2009) Faceted search. Synthesis lectures on information concepts, retrieval, and services. Morgan and Claypool Publishers
53. Voorhees EM, Buckley C (2002) The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '02, pp 316–323
54. Wang J, Zhu J (2010) On statistical analysis and optimization of information retrieval effectiveness metrics. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '10, pp 226–233
55. Wu S, Crestani F (2003) Methods for ranking information retrieval systems without relevance judgments. In: Proceedings of the 2003 ACM symposium on applied computing. ACM, New York, SAC '03, pp 811–816. doi:10.1145/952532.952693
56. Xu J, Croft WB (2000) Improving the effectiveness of information retrieval with local context analysis. ACM Trans Inf Syst 18(1):79–112
57. Yang JJ, Korfhage R, Rasmussen EM (1992) Query improvement in information retrieval using genetic algorithms—a report on the experiments of the trec project. In: TREC, pp 31–58
58. Yuan SA, Yu SN (2007) A new method for cross-language information retrieval by summing weights of graphs. In: Proceedings of the fourth international conference on fuzzy systems and knowledge discovery, vol 02. IEEE Computer Society, Washington, DC, FSKD '07, pp 326–330
59. Zhang J, Gao J, Zhou M, Wang J (2001) Improving the effectiveness of information retrieval with clustering and fusion. Comput Linguist Chin Lang Process 6(1):1–18
60. Zobel J (1998) 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. ACM, New York, SIGIR '98, pp 307–314