Searching Cultural Heritage Data: Does Structure Help Expert Searchers?

Marijn Koolen[1] Jaap Kamps[1,2]
[1]Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2]ISLA, Faculty of Science, University of Amsterdam

ABSTRACT

On-line search requests of cultural heritage (CH) material are often very short and mainly focused on names and dates,while the data provides much more detail and is highly structured, based on classification systems and ontologies. Apparently, typical users make no use of the available information and structure. Expert users such as museum curators have extensive knowledge of the objects in the collection and the classification systems used to describe them, and have complex information needs. In this paper we investigate the impact of exploiting the metadata structure on retrieval effectiveness of complex queries. Our findings are that 1) expert queries require little smoothing as all terms are important for identifying the right objects, 2) the field structure of CH descriptions can help improve early precision, 3) combining free-text retrieval and structured Boolean retrieval leads to significant improvements on both approaches alone. Finally, from analysing the questions send to a museum, we find that non-experts have more complex information needs than what search logs show us, suggesting they can benefit from systems that exploit structure as well.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Retrieval Models

General Terms: Experimentation, Measurement, Performance

Keywords: Structured queries, Metadata, Cultural heritage

1. INTRODUCTION

The field of Information Retrieval has been showing, for over 50 years, that best-match text retrieval is at least as effective as exact-match Boolean search [3]. In this light it is remarkable that Boolean exact match systems are still pervasive in the professional search systems used in the cultural heritage sector. The only possible explanation is that the Boolean systems are in fact effective for the typical search requests of the professionals. The apparent standoff between the proven superiority of best match text search and the pervasiveness of exact match Boolean search in professional applications, prompts an in-depth study of expert searchers in the wild.

We look at the impact of the data, the searchers and their search requests, and simulate the effectiveness of a broad range of systems. We focus on the domain of cultural heritage (CH), specifically on a major fine arts museum, having a typical collection of structured object descriptions. How does this data differ from standard text collections? Is the structure helpful for promoting retrieval effectiveness? The expert searchers are a group of curators that are clearly experts on the their domain of expertise, as well as intimately familiar with the collection and way of description. differ? Do they issue more complex search request? Can these naturally be cast as fielded or structured queries? Based on the search requests and known relevant objects, we can simulate a wide range of systems varying from plain text search to Boolean exact match, and various combinations. What is the relative effectiveness of exact match and best match search? Do field restrictions help retrieval? Can we fruitfully combine exact and best match querying?

Search log analysis of museum web sites often show the same patterns. Users type very short queries consisting mainly of artist names [12–14]. Occasionally they search on other aspects like title, date, object type, and only rarely type more complex queries. This is in stark contrast with the wealth of detailed information that is described by cataloguers and indexers in CH institutions.

These object descriptions are highly structured, with information about the title, creator, size and material of the object. This metadata often contains terms from controlled vocabularies to create precise descriptions that are, at least in theory, consistent in terminology, complete and free of errors, so that a search on a particular object type from a particular period results in a list of all and only descriptions of the right type of objects from the right period.

Expert users like museum curators have extensive knowledge of the object collections and the terminology and format of the descriptions, and might have complex information needs. They might want to exploit their knowledge of the collection when searching through the object descriptions more efficiently and effectively.

Databases are good for exploiting semantic structure and for retrieving sets of exactly matching documents. However, descriptions are often heterogeneous, created using different standards, even within a single institution. Information is often incomplete and might contain inconsistent or incorrect keywords and spelling errors. This might have a significant impact of the effectiveness of database retrieval and might call for a less strict retrieval approach. Modern information retrieval systems have been developed to work well with any type of document representation and deal with inconsistency and incompleteness by using a best-matching approach. Most retrieval systems ignore any available structure and treat the document text as a bag of words. For CH descriptions consisting of semantically labelled keywords, the available structure would seem an important feature for experts to use. This prompts the question:

- Can we exploit the structure of cultural heritage descriptions to improve retrieval effectiveness of expert queries?

The effectiveness of structural information for information re-

RIAO'10, 2010, Paris, France.

trieval is an old question. The Cranfield experiments were designed to evaluate different indexing languages, which were based on different classification systems [11].The experiments were conducted on collections on scientific literature. One of the findings was that searching on combinations of words was more effective than using formally defined classifications. Nowadays it is common practice to simply index all the words in a document. Exploiting the document structure has been studied extensively [1, 2, 9], however, most document collections studied in IR have very little structure.

More specifically, we want to know:

- Do experts (museum curators) have more complex information needs than non-experts?
- Can expert searchers benefit from structured queries?

In the following sections we discuss information requests in cultural heritage (§ 2), and experiments with structured queries (§ 3). We discuss the utility of structured queries for non-expert users (§ 4), and conclude with a discussion of our findings (§ 5).

2. INFORMATION REQUESTS

In this section, we look at the search requests of both experts and non-experts.

2.1 Search Log Queries

We studied the search log data of the Gemeentemuseum from 9 January 2004 to 15 October 2007. The web site allows users to search through a collection of 1,100 of the highlights of the museum object collection. The web site provides images and some metadata for these objects. The search logs contain over 7,500 queries with corresponding click-through data. Of the 10 most frequent queries, 9 are artist names and one is the name of a painting. All 10 queries consist of one words. Why are these queries short and atomic? Is it because users of the system have only very simple, atomic information needs? Is it because they lack knowledge about the system and the data to be able to formulate a more complex query? Or is it because the system does not support complex queries that target multiple aspects of object descriptions?

The log only shows the clicks on images from the results list, including the query that generated the results list. The logs show when a user clicks the search button, but not the actual query. The only information we can get therefore, is the combination of a requested object and the corresponding query. If no image is clicked, or no result is returned in response to a query, we have no information. We note that the logs only contain the "successful" queries, where at least some results were returned and the user clicked on at least one of those results. But from these queries and clicks, we can still derive some information. First of all, user queries can target creators and titles. However, the top 10 queries show mainly creator names, which is in line with earlier findings in search log analysis of art institutions [12–14]. Second, we cannot see whether the observed queries are typed by users with simple information needs, users with limited knowledge or users with more complex information needs who simplified an initial complex query that returned no results. In any of those cases, users still typed those queries and requested an image in response to that query.

2.2 Expert Searchers

We want to look at users who have extensive knowledge of the collection, to find out if they have more complex information needs and whether the available structure is of use to them. To this end, we obtained information requests from curators. They are closely involved in creating the object descriptions, therefore know the terminology, the field structure and the objects in the collection. These descriptions are edited to ensure high quality records with precise terminology and are developed for Boolean search. The search terms suggested by the curators will come from the same vocabularies as the terms in the object descriptions. These keywords form queries that can be seen as Boolean queries. A good example of the "Boolean" way of thinking about retrieval in CH is nicely described in Oddy and Barker [10].

Table 1: Distribution of topics over number of aspects.

	Experts		Non-experts	
aspects per query	**# queries**	**%**	**# queries**	**%**
1	14	31.8	2	7.7
2	16	36.4	9	34.6
3	13	29.5	10	38.5
4	1	2.3	3	11.5

Table 2: Distribution of topics over different aspects.

	Experts		Non-experts	
aspects	**# queries**	**%**	**# queries**	**%**
acquisition method	5	11.4	4	5.6
creator	16	36.4	21	29.2
location	9	20.5	2	2.8
material	5	11.4	4	5.6
period	8	18.2	0	0.0
physical description	4	9.1	6	8.3
size	0	0.0	1	1.4
style	5	11.4	0	0.0
title	0	0	11	15.3
type	24	54.5	12	16.7

We asked nine curators of the *Gemeentemuseum* to write down information requests that are based on or resemble their actual information needs. This resulted in a list of 44 topics. We obtained relevance judgements for 24 of these topics, from seven curators, through a combination of asking them to write down any relevant objects they could think of and compiling lists through manual searches and asking the curators to judge these lists of objects. There are 928 relevant objects for the 24 assessed topics.

We looked at the complexity of all 44 curators topics to see if experts use more complex queries. First, we look at the number of aspects appearing in the requests in Table 1, in columns 2 and 3 (the last two columns will be discussed in § 4). Clearly, the curators requests are more complex than the queries found in the search logs. Still over 30% of the queries contain only one aspect, but the majority of queries have two or more aspects.

Second, we look at the categories appearing in the requests in Table 2, columns 2 and 3. As with the search log queries, many expert requests contain *creator* names, but there are many more aspects that are requested frequently. In fact, the majority of the curator requests contain the object type.

We obtained a full copy of the object database of the *Gemeentemuseum*, which contains 116,493 museum object descriptions. These descriptions are structured with fields like *creator*, *title* and *date*. The 24 *judged* topics have 68 query terms (2.83 terms per query). On average each query term appears in 11.37 distinct fields. The meaning of the term is often determined by the field it appears in. This shows the potential for using field restrictions.

Curators often write and update the descriptions themselves, using agreed upon terminology—reducing the vocabulary mismatch. The experts broadly know which terms were used and should be used to describe characteristics and aspects of objects, and therefore, which terms to use to search for objects. Because they are used to storing and retrieving information in a highly structured way, they are able to generate very precise queries. Their queries should reflect the underlying information need well.

As a result, we may expect that for expert search requests, each query term is important and should occur in each relevant description (consistency and completeness). Descriptions where all but one query terms occur are less relevant than descriptions where all query terms occur. This is related to co-ordination level ranking [4, 5]. In the language modelling framework, this can be achieved by using little smoothing. The contribution of the background model for missing query terms will be very small in comparison to the contribution of the document model for query terms that do occur in the document. The impact of term weights is strongly reduced.

3. EXPERIMENTS AND RESULTS

We used Indri [6] for indexing and retrieving. To use the available structure, we indexed the field information as well, allowing us to retrieve objects that have query terms in specific fields. We experimented with different sets of queries for the 24 expert topics.

Text: a set of unstructured bag of words queries as is typical in many standard text retrieval systems.

Fields: a set of structured queries based on the Indri query language. A query term gets a field restriction if the curators specified one. For instance, *hobo.(object_name)* means the query term *hobo* must occur in the *object_name* field.

Fields+Doc: To see the impact of the document context, these queries balance fields restrictions with an unrestricted equivalent.

That is, we use *#wsum(1.0 hobo.(object_name) 1.0 hobo)* so that documents that have *hobo* in another field still receive some score. We experimented with different weights but found that using unit weights gives the best results.

Fields, Boolean: the same set of structured queries as above, but with a Boolean AND operator, thereby treating the queries in a database-like fashion. All terms must occur for a document to be retrieved. Indri still weights terms and ranks accordingly.

Note that the structured queries were mediated by the experimenter. Structured query languages give users more expressive power [15], but many users have problems formulating good structured queries [7, 8]. We assume here that expert searchers can cope with this.

3.1 Smoothing for Expert Queries

The amount of smoothing determines the importance of whether a term occurs in a document. With little smoothing we can approximate co-ordination level ranking: all query terms are considered important, so document are heavily penalised for missing query terms. We have ranged the smoothing parameter λ between 0 and 1 and found a low value of $\lambda = 0.1$ to be the most effective for MAP, although the differences with higher values is very small. If the smoothing parameter has little impact, this means that the background model has little impact. It seems that even with heavy smoothing we have a co-ordination level ranking. This could be a consequence of the characteristics of CH descriptions. Each term is selected with care and often appears only once in each description. As a result, descriptions matching all terms are automatically ranked above description matching only a fraction of the query terms. With no smoothing ($\lambda = 0$) the results are much worse, showing it is very effective to have at least some smoothing.

3.2 Unstructured and Structured Queries

We compare the effectiveness of the unstructured queries against the two sets of structured queries with field restrictions. One set of structured queries, *Fields*, only matches terms in specified fields, while the other set, *Fields+doc*, matches terms against both the specified fields and the entire document. The first three lines in Table 3 shows the results. With fields we see a small decrease in MRR but a larger improvement in P@10. The semantic information of the fields reduces the number of spurious matches. The slightly lower recall shows some descriptions can only be found by ignoring the field structure. If we match query terms both in the specified field and in the full description, we get improvements at all rank cut-offs, and thereby a larger improvement in MAP. However, it results in a slightly lower P@10 compared to the Fields only run. We test improvements upon the *Text* baseline for significance using the one-tailed bootstrap test with 100,000 resamples. None of the improvements are statistically significant.

Table 3: Comparison of structured and unstructured queries. Significance levels are 0.05 ($^{\circ}$), 0.01 ($^{◉}$) and 0.001 ($^{\bullet}$)

Queries	MAP	MRR	P@10	Recall
Text λ_c*=0.1*	0.4525	0.6691	0.4000	**0.8722**
Fields, λ_c*=0.1*	0.4600	0.6542	0.4542	0.8418
Fields+doc,λ_c*=*λ_d*=0.1*	**0.4771**	**0.6850**	**0.4375**	**0.8722**
Fields, Boolean	0.4482	0.6108	0.4708	0.5534
Fields, Boolean, no rank	0.4612	0.6648	0.4896	0.5534
Fields, Boolean, no rank + Text	0.5636$^{\bullet}$	0.7582$^{\bullet}$	0.5458$^{\bullet}$	**0.8722**

The semantics of the field structure can help precision if expressed properly in the query. On the other hand, taking the larger document context into account when matching terms is more effective for overall precision. This shows that although the fields structure is informative, the document context is also important for ranking.

3.3 Boolean Queries

We compare Structured and Unstructured queries against structured Boolean AND queries. To emulate database retrieval, which considers any document exactly matching the query as relevant and equally relevant, we use the Boolean AND queries and repeatedly randomise scores and evaluate. This is done because we use *trec_eval* to evaluate the runs, which orders documents with equal score by document ID. To remove any possible bias introduced by this ordering on document ID, we randomise the scores and evaluate and repeat this 1,000 times to obtain average scores. This run is denoted *Fields, Boolean, no rank*.

What if we combine the document ranking of the *Text* run with the set-based retrieval of the structured Boolean AND run? If we sum the scores, the set of documents returned by the Boolean run will be ranked before the set of documents returned only by the *Text* run, but within each of these two sets, the documents will be ranked by their *Text* score. What can the unstructured query score add to Boolean retrieval apart from larger recall?

The results of the Boolean runs are shown in Table 3. The overall results indicate that the Boolean-style retrieval is reasonably effective for these expert queries. The Boolean run with ranking returns far fewer relevant documents, but has a much higher early precision. Also, with a recall of 0.5534, it shows that many relevant documents do not exactly match expert queries. We can substantially improve results by linearly combining Boolean set-based retrieval with unstructured text retrieval. The combined document score $S(d)$ for a document d is computed as: $S_{Bool.+Text}(d) = 0.5 \cdot S_{Bool.}(d) + 0.5 \cdot S_{Text}(d)$. The combined run is the only run that shows significant improvements. The fact that the set-based *Bool., no rank* outperforms the rank-based *Bool.* run shows that term frequency and document frequency are not as important for edited CH descriptions containing precise keywords picked by CH experts as for typical natural language texts.

To summarise, smoothing has very little impact on unstructured queries, but little smoothing is more effective than heavy smoothing. Expert searchers know the right terminology and which terms to choose to identify a set of objects. Many terms in CH descriptions consist mainly of labelled keywords which occur only once per description, so term frequency might not be an important factor in ranking. All query terms are important, making document frequency possibly less important. On the one hand, the effectiveness of exactly matching the query gives an indication that subtle term weighting for natural language texts is less needed for keywords based CH descriptions. On the other hand, the results in Table 3 show that set-based retrieval can be greatly improved by combining it with a standard text-retrieval method.

4. DISCUSSION

The search logs showed that non-experts use simple queries, but this is partly due to the system not supporting more complex queries. Perhaps non-experts could also benefit from a more complex query language. We obtained a large number of e-mail questions sent to the museum by outsiders to study the complexity of "non-expert" information requests. Of course, not all questions can be answered by searching through the object database, but from a sample of 75 questions sent to the *Gemeentemuseum*, we found that 26 questions could. Analysis suggests these non-expert have more complex information needs than the simple requests from the search logs. Columns 4 and 5 of Table 1 show the number of aspects per query for the non-expert queries, using the same analytic approach for the e-mail questions as for the expert requests in § 2.2. Similarly, columns 4 and 5 in Table 2 show the distribution over different aspects. Although many questions contain the name of an artist–similar to earlier studies–there are several other aspects that frequently appear in questions. Apart from that, the average complexity of the questions shows that "non-experts" and experts have similarly complex information needs. The average complexity of the e-mail questions is 2.88 aspects and are a further indication that users might use information from more categories in the descriptions if they had access to this information.

5. CONCLUSIONS

In this paper we investigated the potential of using the structure of cultural heritage descriptions to improve retrieval effectiveness for expert searchers. Although CH descriptions are very precise and detailed and highly structured, on-line search logs of cultural heritage institutions show users typically issue simple information requests. We obtained a collection of search requests from expert searchers of CH data and found that they have more complex information needs. We have created a test-collection of museum object descriptions, expert search topics and corresponding relevance judgement, which we used to investigate the importance of the semantics of the available field structure. Both the object descriptions and the search request of are very high quality, with keywords carefully selected using precise terminology.

Because each query term is important and precise, little smoothing is required to compensate for documents missing query terms. Still, smoothing is essential to obtain high recall, showing that incompleteness and inconsistency in descriptions is a significant problem for retrieval. Structured queries that exploit the field semantics lead to more precise search results. However, the improvements are not significant and for overall precision they are small. This indicates the query terms they choose are indeed very precise and unambiguous. Strict Boolean queries are very effective for obtaining precise retrieval results, but lead to low recall because of errors and omissions in the descriptions. Set-based retrieval is roughly as effective as standard text retrieval and suggests that for CH descriptions, standard term weighting based on term frequency and document frequency are less important than for natural language text documents. The effectiveness of the database-like Boolean queries shows the importance of the field structure of the descriptions while the effectiveness of the unrestricted free-text queries shows the importance of using the larger document context and best-matching. The ranking can be significantly improved by combining the set-based retrieval of the Boolean AND queries and the relevance ranking of the unstructured bag of words queries, showing that structure and context are complementary evidence required to obtain high quality retrieval results.

Acknowledgments

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 640.001.501.

REFERENCES

[1] F. J. Burkowski. Retrieval activities in a database consisting of heterogeneous collections of structured text. In *SIGIR '92*, pages 112–125, New York, NY, USA, 1992. ACM. ISBN 0-89791-523-2.

[2] C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. An Algebra for Structured Text Search and a Framework for its Implementation. *The Computer Journal*, 38(1):43–56, 1995.

[3] C. W. Cleverdon. Report on the first stage of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1960.

[4] W. S. Cooper. Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science*, 34(1):31–39, 1983.

[5] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *ECDL '98*, pages 569–584, London, UK, 1998. Springer-Verlag. ISBN 3-540-65101-2.

[6] Indri. Language modeling meets inference networks, 2009. http://www.lemurproject.org/indri/.

[7] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Structured queries in XML retrieval. In A. Chowdhury, N. Fuhr, M. Ronthaler, and H.-J. Schek, editors, *CIKM'05*, pages 2–11. ACM Press, New York NY, USA, 2005.

[8] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Understanding content-and-structure. In A. Trotman, M. Lalmas, and N. Fuhr, editors, *INEX 2005*, pages 14–21. University of Otago, Dunedin New Zealand, 2005.

[9] V. Mihajlovic, H. E. Blok, D. Hiemstra, and P. M. G. Apers. Score region algebra: building a transparent XML-IR database. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken, editors, *CIKM*, pages 12–19. ACM, 2005.

[10] W. A. Oddy and H. Barker. A feature card information-retrieval system for the general museum laboratory. *Studies in Conservation*, 16 (3):89–94, August 1971.

[11] S. Robertson. On the history of evaluation in IR. *J. Information Science*, 34(4):439–456, 2008.

[12] J. Sledge. Points of view. In *Multimedia Computing and Museums, Selected papers from ICHIM95*, pages 335–346, 1995.

[13] J. Sledge and M. Case. Looking for MR Rococo: Getty Art History Information Program Point of View Workshop. *Archives and Museum Informatics*, 9(1):124–129, 1995.

[14] J. Trant. Understanding Searches of a Contemporary Art Museum Catalogue: A Preliminary Study, 2006.

[15] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *INEX*, volume 3493, pages 16–40. Springer, 2004.