The Importance of Document Ranking and User-Generated Content for Faceted Search and Book Suggestions

Frans Adriaans[1,2], Jaap Kamps[1,3], and Marijn Koolen[1]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] Department of Psychology, University of Pennsylvania
[3] ISLA, Faculty of Science, University of Amsterdam

Abstract. In this paper we describe our participation in INEX 2011 in the Books and Social Search Track and the Data Centric Track. For the Books and Social Search Track we focus on the impact of different document representations of book metadata for book search, using either professional metadata, user-generated content or both. We evaluate the retrieval results against ground truths derived from the recommendations in the LibraryThing discussion groups and from relevance judgements obtained from Amazon Mechanical Turk. Our findings show that standard retrieval models perform better on user-generated metadata than on professional metadata. For the Data Centric Track we focus on the selection of a restricted set of facets and facet values that would optimally guide the user toward relevant information in the Internet Movie Database (IMDb). We explore different methods for effective result summarisation by means of weighted aggregation. These weighted aggregations are used to achieve maximal coverage of search results, while at the same time penalising overlap between sets of documents that are summarised by different facet values. We found that weighted result aggregation combined with redundancy avoidance results in a compact summary of available relevant information.

1 Introduction

Our aim for the Books and Social Search Track was to look at the relative value of user tags and reviews and traditional book metadata for ranking book search results. The Social Search for Best Books task is newly introduced this year and uses a large catalogue of book descriptions from Amazon and LibraryThing. The descriptions are a mix of traditional metadata provided by professional cataloguers and indexers and user-generated content in the form of ratings, reviews and tags.

Because both the task and collection are new, we keep our approach simple and mainly focus on a comparison of different document representations. We made separate indexes for representations containing a) only title information, b) all the professional metadata, c) the user-generated metadata, d) the metadata from Amazon, e) the data from LibraryThing and f) all metadata. With these indexes

S. Geva, J. Kamps, and R. Schenkel (Eds.): INEX 2011, LNCS 7424, pp. 30–44, 2012.

we compare standard language model retrieval systems and evaluate them using the relevance judgements from the LibraryThing discussion forums and from Amazon Mechanical Turk. We break down the results to look at performance on different topic types and genres to find out which metadata is effective for particular categories of topics.

For the Data Centric Track we focus on the selection of a restricted set of facets and facet values that would optimally guide the user toward relevant information. We aim to improve faceted search by addressing two issues: a) weighted result aggregation, and b) redundancy avoidance.

The traditional approach to faceted search is to summarise search results by providing counts of the number of documents that are associated with different facet values [6, 14]. Those facet values that have the highest number of counts are returned to the user. We extend this approach by exploring the aggregation of results using weighted document counts. The underlying intuition is that facet values with the *most* documents are not necessarily the *most relevant* values [2]. That is, buying a dvd by the director who directed the most movies does not necessarily meet the search demands of a user. It may be more suitable to return directors who made a large number of important (and/or popular) movies. More sophisticated result aggregations, acknowledging the importance of an entity, may thus provide better hints for further faceted navigation than simple document counts. We therefore explore different methods for effective result summarisation by means of weighted aggregation.

Another problem in faceted search concerns the avoidance of overlapping facets [8]. That is, facets whose values describe highly similar set of documents should be avoided. We therefore aim at penalising overlap between sets of documents that are summarised by different facet values. We expect that weighted result aggregation combined with redundancy avoidance results in a compact summary of the available relevant information.

We describe our experiments and results for the Books and Social Search Track in Section 2 and for the Data Centric Track in Section 3. In Section 4, we discuss our findings and draw conclusions.

2 Book Track

In the INEX 2011 Books and Social Search Track we participated in the Social Search for Best Books task. Our aim was to investigate the relative importance of professional and user-generated metadata. The document collection consists of 2.8 million book description, with each description combining information from Amazon and LibraryThing. The Amazon data has both traditional book metadata such as title information, subject headings and classification numbers, and user-generated metadata as well as user ratings and reviews. The data from LibraryThing consists mainly of user tags.

Professional cataloguers and indexers aim to keep metadata mostly objective. Although subject analysis to determine headings and classification codes is somewhat subjective, the process follows a formal procedure and makes use of

controlled vocabularies. Readers looking for interesting or fun books to read may not only want objective metadata to determine what book to read or buy next, but also opinionated information such as reviews and ratings. Moreover, subject headings and classification codes might give a very limited view of what a book is about. LibraryThing users tag books with whatever keywords they want, including personal tags like *unread* or *living room bookcase*, but also highly specific, descriptive tags such *WWII pacific theatre* or *natives in Oklahoma*.

We want to investigate to what extent professional and user-generated metadata provide effective indexing terms for book retrieval. The Cranfield tests [4] showed that using natural language terms from documents for indexing was at least as effective for retrieval as using controlled vocabularies. However, controlled vocabularies still hold the potential to improve completeness and accuracy of search results by providing consistent and rigorous index terms and ways to deal with synonymy and homonymy [7, 13]. [5] found that "if subject headings were to be removed from or no longer included in catalog records, users performing keyword searches would miss more than one third of the hits they currently retrieve." Authors, indexers and searchers all have different vocabularies [3] which, when all used in a single search process, may very well lead to the possibility of term mismatches. Bates [1, p.7] states that users of library catalogues prefer to use keyword search, which often does not match the appropriate subject headings.

One of the interesting aspects of user-generated metadata in this respect is that it has a smaller gap with the vocabulary of searchers [9]. User tags may (partially) compensate for missing subject headings. Yi and Chan [16] explored the possibility of mapping user tags from folksonomies to Library of Congress subject headings (LCSH), and found that with word matching, they could link two-thirds of all tags to LC subject headings. [10] looked at the retrieval effectiveness of tags taking into account the tag frequency. They found that the tags with the highest frequency are the most effective.

2.1 Experimental Setup

We used Indri [12] for indexing, removed stopwords and stemmed terms using the Krovetz stemmer. We made 5 separate indexes:

Full: the whole description is indexed.
Amazon: only the elements derived from the Amazon data are indexed.
LT: only the elements derived from the LibraryThing data are indexed.
Title: only the title information fields (title, author, publisher, publication date, dimensions, weight, number of pages) are indexed.
Professional: only the traditional metadata fields from Amazon are indexed, including the title information (see Title index) and classification and subject heading information.
Social: only the user-generated content such as reviews, tags and ratings are indexed.

Table 1. Evaluation results for the Social Search for Best Books task runs using the LT suggestion Qrels. Runs marked with * are official submissions.

Run	nDCG@10	P@10	MRR	MAP
xml_amazon.fb.10.50	0.2665	0.1730	0.4171	0.1901
*xml_amazon	0.2411	0.1536	0.3939	0.1722
*xml_full.fb.10.50	0.2853	0.1858	0.4453	0.2051
*xml_full	0.2523	0.1649	0.4062	0.1825
xml_lt.fb.10.50	0.1837	0.1237	0.2940	0.1391
*xml_lt	0.1592	0.1052	0.2695	0.1199
xml_prof	0.0720	0.0502	0.1301	0.0567
*xml_social.fb.10.50	**0.3101**	**0.2071**	**0.4811**	**0.2283**
*xml_social	0.2913	0.1910	0.4661	0.2115
xml_title	0.0617	0.0403	0.1146	0.0563

The topics are taken from the LibraryThing discussion groups and contain a *title* field which contains the title of a topic thread, a *group* field which contains the discussion group name and a *narrative* field which contains the first message from the topic thread. In our experiments we only used the *title* fields as queries and default settings for Indri (Dirichlet smoothing with $\mu = 2500$). We submitted the following six runs:

xml_amazon: a standard LM run on the Amazon index.
xml_full: a standard LM run on the Full index.
xml_full.fb.10.50: a run on the Full index with pseudo relevance feedback using 50 terms from the top 10 results.
xml_lt: a standard LM run on the LT index.
xml_social: a standard LM run on the Social index.
xml_social.fb.10.50: a run on the Social index with pseudo relevance feedback using 50 terms from the top 10 results.

Additionally we created the following runs:

xml_amazon.fb.10.50: a standard LM run on the Amazon index.
xml_lt.fb.10.50: a standard LM run on the LT index.
xml_prof: a standard LM run on the Professional index.
xml_title: a standard LM run on the Title index.

2.2 Results

The Social Search for Best Books task has two sets of relevance judgements. One based on the lists of books that were suggested on the LT discussion groups, and one based on document pools of the top 10 results of all official runs, judged by Mechanical Turk workers. For the latter set of judgements, a subset of 24 topics was selected from the larger set of 211 topics from the LT forums.

Table 2. Evaluation results for the Social Search for Best Books task runs using the AMT Qrels. Runs marked with * are official submissions.

Run	nDCG@10	P@10	MRR	MAP
xml_amazon.fb.10.50	0.5954	0.5583	0.7868	0.3600
*xml_amazon	**0.6055**	**0.5792**	0.7940	0.3500
*xml_full.fb.10.50	0.5929	0.5500	**0.8075**	**0.3898**
*xml_full	0.6011	0.5708	0.7798	0.3818
xml_lt.fb.10.50	0.4281	0.3792	0.7157	0.2368
*xml_lt	0.3949	0.3583	0.6495	0.2199
xml_prof	0.1625	0.1375	0.3668	0.0923
*xml_social.fb.10.50	0.5425	0.5042	0.7210	0.3261
*xml_social	0.5464	0.5167	0.7031	0.3486
xml_title	0.2003	0.1875	0.3902	0.1070

We first look at the results based on the Qrels derived from the LT discussion groups in Table 1. The runs on the Social index outperform the others on all measures. The indexes with no user-generated content–Professional and Title—lead to low scores. The user-provided content seems to add more useful information to the title fields than the professional metadata. The LT index also leads to better performance than the Professional index, suggesting tags can indeed compensate and improve upon controlled subject access. The indexes that have reviews—Amazon, Full and Social—outperform the LT index which has user tags but no reviews. Reviews seem to be effective document representations. Feedback is effective on the four indexes Amazon, Full, LT and Social.

Next we look at the results based on the Mechanical Turk judgements in Table 2. Here we see a different pattern. With the top 10 results judged on relevance, all scores are higher than with the LT judgements. This is probably due in part to the larger number of judged documents, but perhaps also to the difference in the tasks. The Mechanical Turk workers were asked to judged the topical relevance of books—is the book on the same topic as the request from the LT forum—whereas the LT forum members were asked by the requester to recommend books from a possibly long list of topically relevant books. Another interesting observation is that feedback is not effective for the AMT evaluation on the Full, Amazon and Social indexes, whereas it was effective for the LT evaluation. The main difference between the Full, Amazon and Social indexes on the one hand and the LT index on the other hand is that the LT index has no reviews. This might suggest the AMT workers paid more attention to the tags than to the reviews when making their judgements. A rationale for this could be that tags provide a faster way to judge a book than reviews, which is in the interest of workers who wish to minimise the time spent on a HIT.

Perhaps another reason is that the two evaluations use different topic sets. To investigate the impact of the topic set, we filtered the LT judgements on the 24 topics selected for AMT, such that the LT and AMT judgements are more directly comparable. The results are shown in Table 3. The pattern is similar

Table 3. Evaluation results for the Social Search for Best Books task runs using the LT recommendation Qrels for the 24 topics selected for the AMT experiment. Runs marked with * are official submissions.

Run	nDCG@10	P@10	MRR	MAP
xml_amazon.fb.10.50	0.2103	0.1625	0.3791	0.1445
xml_amazon	0.1941	0.1583	0.3583	0.1310
xml_full.fb.10.50	0.2155	0.1708	0.3962	0.1471
xml_full	0.1998	0.1625	0.3550	0.1258
xml_lt.fb.10.50	0.1190	0.0833	0.3119	0.0783
xml_lt	0.1149	0.0708	0.3046	0.0694
xml_prof	0.0649	0.0500	0.1408	0.0373
xml_social.fb.10.50	**0.3112**	**0.2333**	**0.5396**	**0.1998**
xml_social	0.2875	0.2083	0.5010	0.1824
xml_title	0.0264	0.0167	0.0632	0.0321

to that of the LT judgements over the 211 topics, indicating that the impact of the topic set is small. The runs on the Social index outperform the others, with the Amazon and Full runs scoring better than the LT runs, which in turn perform better than the Official and Title runs. Feedback is again effective for all reported measures. In other words, the observed difference between the LT and AMT evaluations is not caused by difference in topics but probably caused by the difference in the tasks.

2.3 Analysis

The topics of the SB Track are labelled with topic type and genre. There are 8 different type labels: *subject* (134 topics), *author* (32), *genre* (17), *series* (10), *known-item* (7), *edition* (7), *work* (3) and *language* (2). The genre labels can be grouped into fiction, with genre label *Literature* (89 topics) and non-fiction, with genre labels such as *history* (60 topics), *biography* (24), *military* (16), *religion* (16), *technology* (14) and *science* (11).

We break down the evaluation results over topic types and take a closer look at the *subject*, *author* and *genre* types. The other types have either very small numbers of topics (*work* and *language*), or are hard to evaluate with the current relevance judgements. For instance, the *edition* topics ask for a recommended edition of a particular work. In the relevance judgements the multiple editions of a work are all mapped to a single work ID in LibraryThing. Some books have many more editions than others, which would create in imbalance in the relevance judgements for most topics.

The evaluation results are shown in Table 4. For most runs there is no big difference in performance between *fiction* and *non-fiction* topics, with slightly better performance on the *fiction* topics. For the two runs on the Social index the difference is bigger. Perhaps this is due to a larger amount of social metadata for fiction books. The standard run on the LT index (xml_lt) performs better on

Table 4. Evaluation results using the LT recommendation Qrels across different topic genres and types. Runs marked with * are official submissions.

	nDCG@10				
Run	**Fiction**	**Non-fiction**	**Subject**	**Author**	**Genre**
xml_amazon.fb.10.50	0.2739	0.2608	0.2203	0.4193	0.0888
*xml_amazon	0.2444	0.2386	0.1988	0.3630	0.0679
*xml_full.fb.10.50	0.2978	0.2765	0.2374	0.4215	0.1163
*xml_full	0.2565	0.2491	0.2093	0.3700	0.0795
xml_lt.fb.10.50	0.1901	0.1888	0.1597	0.2439	0.0850
*xml_lt	0.1535	0.1708	0.1411	0.2093	0.0762
xml_prof	0.0858	0.0597	0.0426	0.1634	0.0225
*xml_social.fb.10.50	**0.3469**	**0.2896**	**0.2644**	**0.4645**	0.1466
*xml_social	0.3157	0.2783	0.2575	0.4006	**0.1556**
xml_title	0.0552	0.0631	0.0375	0.1009	0.0000

the non-fiction topics, suggesting the tags for non-fiction are more useful than for fiction books. Among the topic types we see the same pattern across all measures and all runs. The *author* topic are easier than the *subject* topics, which are again easier than the *genre* topics. We think this is a direct reflection of the clarity and specificity of the information needs and queries. For author related topics, the name of the author is a very clear and specific retrieval cue. Subject are somewhat broader and less clearly defined, making it harder to retrieve exactly the right set of books. For genre-related topics it is even more difficult. Genres are broad and even less clearly defined. For many genres there are literally (tens of) thousands of books and library catalogues rarely go so far in classifying and indexing specific genres. This is also reflected by the very low scores of the Official and Title index runs for *genre* topics.

3 Data Centric Track

For the Data Centric Track we participated in the Ad Hoc Task and the Faceted Search Task. Our particular focus was on the Faceted Search Task where we aim to discover for each query a restricted set of facets and facet values that best describe relevant information in the results list. Our general approach is to use weighted result aggregations to achieve maximal coverage of relevant documents in IMDb. At the same time we aim to penalise overlap between sets of documents that are summarised by different facet values. We expect that this results in a compact summary of the available relevant information. Below we describe our setup and results.

3.1 Experimental Setup

We use Indri [12] with Krovetz stemming and default smoothing (Dirichlet with $\mu = 2500$) for indexing. All XML leaf elements in the IMDb collection are indexed

as fields. Documents were retrieved using title fields only. The maximum number of retrieved documents was set to 1000 (Ad Hoc Task) and 2000 (Faceted Search Task). We submitted one run for the Ad Hoc Search Task and three runs for the Faceted Search Task.

Ad Hoc Task: One run was generated using the settings described above: *UAms2011adhoc*.

Faceted Search Task: Two Ad Hoc result files were used as a basis for facet selection: the *2011-dc-lucene.trec* file provided by the INEX organisation, and an Ad Hoc run that was created using Indri. The maximum number of results for this run was set to 2000. We submitted three Faceted Search runs: *UAms2011indri-c-cnt*, *UAms2011indri-cNO-scr2*, *UAms2011lucene-cNO-lth*. In each run, a hierarchy of recommended facet values is constructed for each topic. A path through the hierarchy represents an accumulated set of conditions on the retrieved documents. The search results become more refined at every step, and the refinement ultimately narrows down to a set of potentially interesting documents.

3.2 Facet Selection

The set of candidate facets consists of all numerical and categorical fields in the IMDb collection. The goal is to select useful facets (and values) from the set of candidate facets.

Result Aggregation. We explored two different methods of weighted result aggregation. The first method aggregates document lengths rather than number of documents. Since popular movies in IMDb have larger entries (which we measure by file size), we assume that document lengths push facet values associated with popular movies to the top of the ranked set of facet values. The second method aggregates documents from the Ad Hoc run by summing retrieval scores. The idea is that higher-ranked documents display facet values that are most likely to be of interest to the user. Note that document length is a static ('global') measure of document importance, whereas retrieval scores are dynamic ('local'), resulting in different degrees of importance for different topics. We compare both methods to traditional non-weighted aggregation of search results using document counts. The result aggregations form the basis of facet selection.

Coverage. For facet selection we use the intuition that facets which provide compact summaries of the available data allow fast navigation through the collection. This intuition was implemented as *facet coverage*: the number of documents that are summarised by a facet's top n values. Two types of coverage were implemented. The first version, *coverage*, sums up the (weighted) document counts that are associated with the facet's top n values. A potential pitfall of this approach is that this method favours redundancy. That is, the sets of documents that are associated with different facet values may have a high degree of overlap. For example, the keywords 'murder' and 'homicide' may point to almost identical sets of documents. We assume a user wants compact overviews

Table 5. Selected facets and values for the query 'Vietnam' (topic 2011205). Facets are ranked by coverage based on document counts.

Rank	Coverage	Facet	Top-5 values
1	945	genre	Drama (306) Documentary (207) War (199) Action (157) Comedy (76)
2	850	keyword	vietnam (286) vietnam-war (220) independent-film (162) vietnam-veteran (110) 1960s (72)
3	477	language	English (400) Vietnamese (42) French (16) Spanish (10) German (9)
4	437	country	USA (345) UK (30) Canada (27) France (19) Vietnam (16)
5	397	color	Color (291) Color - (Technicolor) (45) Black and White (40) Color - (Eastmancolor) (11) Color - (Metrocolor) (10)

of different, non-overlapping sets of documents that may be of interest to the searcher. Therefore, we implemented a second version: *coverageNO* ('coverage, no overlap') counts the number of unique documents that are summarised by the facet's top n values. As a consequence, redundancy in facet values is penalised.

Coverage-based facet selection is applied recursively. Starting with the complete set of Ad Hoc results (corresponding to the root node of the facet hierarchy), the facet with the highest coverage is chosen. The set of results is then narrowed down to the set of documents that are covered by this facet. In this new set, a second facet is chosen with the highest coverage. This selection process continues until a specified number of facets has been selected. We apply facet selection to movie facets and person facets independently, since these facets describe different types of documents (i.e., you cannot drill-down into person files after you have narrowed down the results using a movie facet). An example of a ranked set of movie facets for the query 'Vietnam' is given in Table 5.

3.3 Path Construction

The facet hierarchy is based on the selected set of facets and corresponding top n ranked values. Each path starts with a value from the first facet, followed by a value from the second facet, etc. The paths are ordered by rankings of the values within a facet. Not all logically possible paths are considered relevant. As a formal criterium, we assume that only paths leading to between 10 and 20 documents are useful recommendations for the user. Paths that lead to fewer documents are deemed too specific. Paths to a larger number of documents are deemed too general, and the system will attempt to branch into a deeper, more specific level. We generate trees for 'movies' and 'persons' independently and join them in the order of the largest number of paths. (For most queries there were more movie paths than person paths.) As an example, we display a partial tree corresponding to the query 'Vietnam', using the facets from Table 5:

```
<topic tid="2011205">
<fv f="/movie/overview/genres/genre" v="Drama">
  <fv f="/movie/overview/keywords/keyword" v="vietnam">
    <fv f="/movie/additional_details/languages/language" v="Vietnamese">
      <fv f="/movie/additional_details/countries/country" v="USA">
        <fv f="/movie/additional_details/colors/color" v="Color"/>
          </fv>
        </fv>
      </fv>
  <fv f="/movie/overview/keywords/keyword" v="vietnam-war">
    <fv f="/movie/additional_details/languages/language" v="English">
      <fv f="/movie/additional_details/countries/country" v="USA">
        <fv f="/movie/additional_details/colors/color" v="Color - (Technicolor)"/>
          </fv>
        </fv>
    <fv f="/movie/additional_details/languages/language" v="Vietnamese">
      <fv f="/movie/additional_details/countries/country" v="USA"/>
        </fv>
      </fv>
  <fv f="/movie/overview/keywords/keyword" v="vietnam-veteran">
    <fv f="/movie/additional_details/languages/language" v="English">
      <fv f="/movie/additional_details/countries/country" v="USA">
        <fv f="/movie/additional_details/colors/color" v="Color - (Technicolor)"/>
          </fv>
        </fv>
      </fv>
    </fv>
<fv f="/movie/overview/genres/genre" v="Documentary">
  <fv f="/movie/overview/keywords/keyword" v="vietnam">
    <fv f="/movie/additional_details/languages/language" v="English">
      <fv f="/movie/additional_details/countries/country" v="USA">
        <fv f="/movie/additional_details/colors/color" v="Black and White"/>
...
```

Table 6. Experimental parameters. The values of the first three parameters were combined to generate a total of 2x4x2 = 16 different runs. The other parameters (4-7) were kept constant.

Parameter	Values
1. Ad hoc input	Indri, Lucene
2. Document weights	count (cnt), length (lth), score (scr), score^2 (scr2)
3. Selection method	coverage (c), coverageNO (cNO)
4. Number of facets	5
5. Number of values	5
6. Min. number of path results	10
7. Max. number of path results	20

3.4 The Faceted Seach Runs

We generated a total of 16 runs by varying the parameters listed in Table 6. From this set, three runs were selected for submission to the INEX workshop:

UAms2011indri-c-cnt: This is our baseline run which implements the standard approach of selecting those facet values that summarize the largest number of documents.

UAms2011indri-cNO-scr2: This run uses weighted result aggregation (using retrieval scores, in contrast to unranked aggregation in the baseline run). This run also penalises overlap between document sets that correspond to different facet values.

UAms2011lucene-cNO-lth: The third run uses the Lucene reference results file that was provided by INEX. The run uses weighted result aggregation based on document lengths (file sizes, as opposed to retrieval scores).

3.5 Results and Discussion

Our run for the Ad Hoc Task was the best scoring run out of a total of 35 submitted runs by 9 different institutes, with a MAP of 0.3969 [15]. The success of our Ad Hoc run indicates that indexing the complete XML structure of IMDb is not necessary for effective document retrieval. It appears, at least for the Ad Hoc case, that it suffices to index leaf elements.

The runs for the Faceted Search Task were evaluated with respect to two different metrics. The first measure assesses the effectiveness of a faceted system by calculating the interaction cost. This is defined as the number of results, facets, or facet values that the user examines before encountering the first relevant result. The measure is referred to as the Normalised Gain (NG), and the Average Normalised Gain (see [15] for more details). The second measure is the Normalised Discounted Cumulated Gain (nDCG), which assesses the relevance of a hierarchy of facet values based on the relevance of the results that are associated with the values [11].

Table 7. Evaluation results for Faceted Search runs in terms of NGs and ANG

topic	UAms2011indri-c-cnt	UAms2011indri-cNO-scr2	UAms2011lucene-cNO-lth
201	**0.64**	0.60	-
202	0	0	**0.21**
203	0	0	0
204	0.63	0.75	**0.94**
205	0	0	**0.81**
207	0	**0.77**	0
208	0	0	0
209	0	0	0
210	**0.75**	0.74	-
211	0.18	0	**0.53**
212	**0.89**	0.88	-
213	**0.76**	**0.76**	-
214	0	0	**0.64**
ANG	0.30	**0.35**	0.24

Note. c = coverage, cNO = coverage with no overlap, cnt = count, lth = length, scr = retrieval score, scr2 = retrieval score^2. Best scores are in bold.

Table 7 shows the NG and ANG scores of the three runs that we submitted for the Faceted Search Task. While the results vary substantially between topics, our UAms2011indri-cNO-scr2 run (which uses retrieval scores as document weights, and penalises facet values with overlapping sets of documents) has a higher overall score than our baseline run (UAms2011indri-c-cnt). This confirms our expectation that faceted search can be improved by exploiting information from the Ad Hoc results list, and by penalising redundancy. The two Indri-based runs outperform the Lucene-based run (which uses document lengths as weights). The superior performance could thus be due to two factors: the underlying Ad Hoc run, or the aggregation method. Our run UAms2011indri-cNO-scr2 had the highest ANG score out of the 12 runs that had been submitted to the workshop by 5 different groups [15]. Most other groups used the Lucene reference result file, so, again, it is possible that the superior performance of our run is due to a better underlying results file, rather than to effective facet selection. We therefore examine a larger set of runs, allowing us to analyse the results in a more systematic way.

Table 8 shows the nDCG scores for all of our runs (including the three runs that we had submitted).[1] The nDCG scores confirm that the UAms2011indri-cNO-scr2 run was our best one, and the run has a higher mean nDCG than any of the runs that had been submitted by other participating groups (as reported in

[1] Out of all 16 different runs described in Table 6 only 12 produced positive results on the nDCG metric.

Table 8. Evaluation results for the Faceted Search runs in terms of nDCG

	Indri							Lucene				
topic	c cnt	c lth	c scr	c scr2	cNO cnt	cNO lth	cNO scr2	c cnt	c scr2	cNO lth	cNO scr	cNO scr2
201	**0.035**	**0.035**	**0.035**	**0.035**	**0.035**	**0.035**	**0.035**	0	0	0	0	0
202	0	0	0	0	0	0	0	0	0	0	0	0
203	0	0	0	0	0	0	0	0	0	0	0	0
204	0	0	0	0	0	0	0	0	0	0	0	0
205	0	0	0	0	**0.429**	0.198	**0.429**	0	0	0.215	0.209	0.066
207	0	0	0	0	0	0	**0.162**	0	0	0	0	0
208	0	0	0	0	**0.455**	0.452	**0.455**	0	0	0	0	0
209	0	0	0	0	0	0	0	0	0.360	0	0.360	0.360
210	0	0	0	0	0	0	0	0	0	0	0	0
211	0	0	0	0	0	0	0	0	0	0	0	0
212	0	0	0	0	0	0	0	0	0	0	0	0
213	0	0	0	0	0	0	0	0	0	0	0	0
214	**0.185**	0	0.160	0	**0.185**	**0.185**	**0.185**	0.091	0	0.091	0.091	0.022
mean	0.017	0.003	0.015	0.003	0.085	0.067	**0.097**	0.007	0.028	0.024	0.051	0.034

Note. c = coverage, cNO = coverage with no overlap, cnt = count, lth = length, scr = retrieval score, scr2 = retrieval score^2. Best scores are in bold.

[15]).[2] The nDCG results show that the Indri results file indeed provided a better basis for the selection of facet values than the Lucene reference file. However, if we compare different runs that are based on the same Lucene results file, we find that retrieval scores (scr and scr2) improve performance as compared to the Lucene run that we submitted (which was based on document length). Although we have to be careful with interpreting these results where scores for most topics are zero, the success of our run seems to be due to both the results file and the aggregation method. Moreover, the results indicate that our method for penalising overlapping facet values was effective: the *coverage, no overlap* runs had higher nDCG means than their *coverage* counterparts. Result aggregation using retrieval scores proved to be especially useful in combination with the overlap penalty.

In sum, the results confirm our expectation that weighted result aggregation combined with redundancy avoidance results in a compact summary of available relevant information. The findings show the importance of good Ad Hoc results as a basis for faceted search (the well-known 'garbage in, garbage out' principle), and the importance of penalising redundancy in different facet values. Finally, while the effect of different result aggregations varies, it seems that retrieval scores are useful for the detection of relevant facet values.

[2] The mean nDCG score of our run is still quite low, and the fact that many topics yielded NDGC = 0 suggests that either the topic set, the collection and/or the metric may have been inappropriate for the evaluation of faceted search systems.

4 Conclusion

In this paper we discussed our participation in the INEX 2011 Books and Social Search Track and the Data Centric Track.

In the Books and Social Search Track we participated in the Social Search for Best Books task and focused on comparing different document representations based on professional metadata and user-generated metadata. Our main finding is that standard language models perform better on representations of user-generated metadata than on representations of professional metadata.

In our result analysis we differentiated between topics requesting fiction and non-fiction books and between subject-related topics, author-related topics and genre-related topics. Although the patterns are similar across topic types and genres, we found that social metadata is more effective for fiction topics than for non-fiction topics, and that regardless of document representation, all systems perform better on author-related topics than on subject related topics and worst on genre-related topics. We expect this is related to the specificity and clarity of these topic types. Author-related topics are highly specific and target a clearly defined set of books. Subject-related topics are broader and less clearly defined, but can still be specific. Genre-related topics are very broad—many genres have tens of thousands of books—and are also more vague information needs that are closer to exploratory search.

In future work we will look closer at the relative value of various types of metadata and directly compare individual types of metadata such as reviews, tags and subject headings. We will also look at the different search scenarios underlying the relevance judgements and topic categories, such as subject search, recommendation and exploratory search.

In the Data Centric Track we participated in the Ad Hoc and Faceted Search Task. Our main finding is that faceted search can be improved through aggregation of search results that are weighted by their Ad Hoc retrieval score, expressing the local importance of different documents in the results list. In addition, we found that avoiding redundancy leads to a more compact representation of the results list. Although the results are based on a small number of topics, weighted result aggregation and redundancy avoidance together seem to provide an effective means of creating a compact summary of available relevant information.

Acknowledgments. Frans Adriaans was supported by the Netherlands Organization for Scientific Research (NWO) grants # 612.066.513 and 446.010.-027. Jaap Kamps was supported by NWO under grants # 612.066.513, 639.-072.601, and 640.005.001. Marijn Koolen was supported by NWO under grant # 639.072.601.

References

[1] Bates, M.J.: Task Force Recommendation 2.3 Research and Design Review: Improving user access to library catalog and portal information. In: LoC Bicentennial Conf. on Bibliographic Control for the New Millennium (2003)
[2] Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R.: Beyond basic faceted search. In: WSDM 2008 (2008)
[3] Buckland, M.: Vocabulary as a Central Concept in Library and Information Science. In: Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of CoLIS3 (1999)
[4] Cleverdon, C.W.: The Cranfield tests on index language devices. Aslib 19, 173–192 (1967)
[5] Gross, T., Taylor, A.G.: What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. College & Research Libraries 66(3) (2005)
[6] Hearst, M.A., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.-P.: Finding the flow in web site search. Communications of the ACM 45, 42–49 (2002)
[7] Lancaster, F.W.: Vocabulary control for information retrieval, 2nd edn. Information Resources Press, Arlington (1986)
[8] Li, C., Yan, N., Roy, S.B., Lisham, L., Das, G.: Facetedpedia: Dynamic generation of query-dependent faceted interfaces for wikipedia. In: Proceedings of WWW 2010 (2010)
[9] Mathes, A.: Folksonomies - cooperative classification and communication through shared metadata (December 2004), `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`
[10] Peters, I., Schumann, L., Terliesner, J., Stock, W.G.: Retrieval Effectiveness of Tagging Systems. In: Grove, A. (ed.) Proceedings of the 74th ASIS&T Annual Meeting, vol. 48 (2011)
[11] Schuth, A., Marx, M.: Evaluation Methods for Rankings of Facetvalues for Faceted Search. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 131–136. Springer, Heidelberg (2011)
[12] Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis (2005)
[13] Svenonius, E.: Unanswered questions in the design of controlled vocabularies. JASIS 37(5), 331–340 (1986)
[14] Tunkelang, D.: Faceted Search. Morgan and Claypool Publishers (2009)
[15] Wang, Q., Ramírez, G., Marx, M., Theobald, M., Kamps, J.: Overview of the INEX 2011 Data Centric Track. In: Geva, S., Kamps, J., Schenkel, R. (eds.) INEX 2011. LNCS, vol. 7424, pp. 118–137. Springer, Heidelberg (2012)
[16] Yi, K., Chan, L.M.: Linking folksonomy to Library of Congress subject headings: an exploratory study. Journal of Documentation 65(6), 872–900 (2009)