Overview of the INEX 2011 Data-Centric Track

Qiuyue Wang[1,2], Georgina Ramírez[3], Maarten Marx[4], Martin Theobald[5], Jaap Kamps[6]

[1] School of Information, Renmin University of China, P. R. China
[2] Key Lab of Data Engineering and Knowledge Engineering, MOE, P. R. China
qiuyuew@ruc.edu.cn
[3] Universitat Pompeu Fabra, Spain
georgina.ramirez@upf.edu
[4] University of Amsterdam, the Netherlands
maartenmarx@uva.nl
[5] Max-Planck-Institut für Informatik, Germany
martin.theobald@mpi-inf.mpg.de
[6] University of Amsterdam, the Netherlands
kamps@uva.nl

Abstract. This paper presents an overview of the INEX 2011 Data-Centric Track. Having the *ad hoc search task* running its second year, we introduced a new task, *faceted search task*, which goal is to provide the infrastructure to investigate and evaluate different techniques and strategies of recommending facet-values to aid the user to navigate through a large set of query results and quickly identify the results of interest. The same IMDB collection as last year was used for both tasks. A total of 9 active participants contributed a total of 60 topics for both tasks and submitted 35 ad hoc search runs and 13 faceted search runs. A total of 38 ad hoc search topics were assessed, which include 18 subtopics for 13 faceted search topics. We discuss the setup for both tasks and the results obtained by their participants.

1 Introduction

As the de facto standard for data exchange on the web, XML is widely used in all kinds of applications. XML data used in different applications can be categorized into two broad classes: one is document-centric XML, where the structure is simple and long text fields predominate, e.g. electronic articles, books and so on, and the other is data-centric XML, where the structure is very rich and carries important information about objects and their relationships, e.g. e-Commerce data or data published from databases. The INEX 2011 Data Centric Track is investigating retrieval techniques and related issues over a strongly structured collection of XML documents, the IMDB data collection. With richly structured XML data, we may ask how well such structural information could be utilized to improve the effectiveness of search systems.

The INEX 2011 Data-Centric Track features two tasks: the *ad hoc search task* and the *faceted search task*. The *ad hoc search task* consists of informational requests to be answered by the entities contained in the IMDB collection (movies, actors, directors, etc.); the *faceted search task* asks for a restricted list of facet-values that

will optimally guide the searcher towards relevant information in a ranked list of results, which is especially useful when searchers' information needs are vague or complex.

There were 49 institutes or groups interested in participating in the track, from which 8 (Kasetsart University, Benemérita Universidad Autónoma de Puebla, University of Amsterdam, IRIT, University of Konstanz, Chemnitz University of Technology, Max-Planck Institute for Informatics, Universitat Pompeu Fabra) submitted 45 valid ad hoc search topics and 15 faceted search topics. A total of 9 participants (Kasetsart University, Benemérita Universidad Autónoma de Puebla, University of Amsterdam, IRIT, University of Konstanz, Chemnitz University of Technology, Max-Planck Institute for Informatics, Universitat Pompeu Fabra, Renmin University of China, Peking University) submitted 35 ad hoc search runs and 13 faceted search runs. 38 ad hoc topics were assessed, which included 18 subtopics for 13 faceted search topics.

2 Data Collection

The track uses the cleaned IMDB data collection used in INEX 2010 Data-Centric Track [1]. It was generated from the plain text files published on the IMDB web site on April 10, 2010. There are two kinds of objects in the collection, movies and persons involved in movies, e.g. actors/actresses, directors, producers and so on. Each object is richly structured. For example, each movie has title, rating, directors, actors, plot, keywords, genres, release dates, trivia, etc.; and each person has name, birth date, biography, filmography, etc. Each XML document contains information about one object, i.e. a movie or person, with structures conforming to the movie.dtd or person.dtd [1]. In total, the IMDB data collection contains 4,418,081 XML documents, including 1,594,513 movies, 1,872,471 actors, 129,137 directors who did not act in any movie, 178,117 producers who did not direct nor act in any movie, and 643,843 other people involved in movies who did not produce nor direct nor act in any movie.

3 Ad-Hoc Search Task

The task is to return a ranked list of results, i.e. objects, or equivalently documents in the IMDB collection, estimated relevant to the user's information need.

3.1 Topics

Each participating group was asked to create a set of candidate topics, representative of a range of real user needs. Each group had to submit a total of 3 topics, one for each of the categories below:

- **Known-item**: Topics that ask for a particular object (movie or person). Example: "I am searching for the version of the movie 'Titanic' in which the two major characters are called Jack and Rose respectively". For these topics the relevant

answer is a single (or a few) document(s). We will ask participants to submit the file name(s) of the relevant document(s).

- **List**: Topics that ask for a list of objects (movies or persons). For example: "Find movies about drugs that are based on a true story", "Find movies about the era of ancient Rome".
- **Informational**: Topics that ask for information about any topic/movie/person contained in the collection. For example: "Find information about the making of The Lord of the Rings and similar movies", "I want to know more about Ingmar Bergman and the movies she played in".

All the data fields in the IMDB collection can be categorized into three types: categorical (e.g. genre, keyword, director), numerical (e.g. rating, release_date, year), and free-text (e.g. title, plot, trivia, quote). All submitted topics had to involve, at least, one free-text field. The list of all the fields along with their types is given in Appendix 1. We asked participants to submit challenging topics, i.e. topics that could not be easily solved by a current search engine or DB system. Both Content Only (CO) and Content And Structure (CAS) variants of the information need were requested. TopX provided by Martin Theobald was used to facilitate topic development.

After cleaning some duplicates and incorrectly-formed topics, there were a total of 25 valid topics (11 list, 7 known-item, 7 informational). An example of topic is shown in Fig. 1.

*<topic id="**2011105**" guid="**20**">*
*<task>**AdHoc**</task>*
*<type>**Known-Item**</type>*
*<title>**king kong jack black**</title>*
*<castitle>**//movie[about(.//title, king kong) and about(.//actor, jack black)]**</castitle>*
*<description>**I am searching for the version of the movie "King Kong" with the actor Jack Black.**</description>*
*<narrative>**Cause i've heard that this is the best King Kong movie, I am searching for the version of the movie "King Kong", with the actor Jack Black.**</narrative>*
</topic>

Fig. 1. INEX 2011 Data Centric Track Ad Hoc Search Topic 2011105

3.2 Submission Format

Each participant could submit up to 3 runs. Each run could contain a maximum of 1000 results per topic, ordered by decreasing value of relevance. The results of one run had to be contained in one submission file (i.e. up to 3 files could be submitted in total). For relevance assessment and evaluation of the results we required submission files to be in the familiar TREC format:

<qid> Q0 <file> <rank> <rsv> <run_id>

Here:

- The first column is the topic number.
- The second column is the query number within that topic. This is currently unused and should always be Q0.

- The third column is the file name (without .xml) from which a result is retrieved.
- The fourth column is the rank of the result.
- The fifth column shows the score (integer or floating point) that generated the ranking. This score MUST be in descending (non-increasing) order and is important to include so that we can handle tied scores (for a given run) in a uniform fashion (the evaluation routines rank documents from these scores, not from ranks).
- The sixth column is called the "run tag" and should be a unique identifier that identifies the group and the method that produced the run. The run tags must contain 12 or fewer letters and numbers, with NO punctuation, to facilitate labeling graphs with the tags.

An example submission is:

```
2011001 Q0 9996 1 0.9999 2011UniXRun1
2011001 Q0 9997 2 0.9998 2011UniXRun1
2011001 Q0 person_9989 3 0.9997 2011UniXRun1
```

Here are three results for topic “2011001”. The first result is the movie from the file 9996.xml. The second result is the movie from the file 9997.xml, and the third result is the person from the file person_9989.xml.

4 Faceted Search Task

Given a vague or broad query, the search system may return a large number of results. Faceted search is a way to help users navigate through the large set of query results to quickly identify the results of interest. It presents the user a list of facet-values to choose from along with the ranked list of results. By choosing from the suggested facet-values, the user can refine the query and thus narrow down the list of candidate results. Then, the system may present a new list of facet-values for the user to further refine the query. The interactive process continues until the user finds the items of interest. The key issue in faceted search is to recommend appropriate facet-values for the user to refine the query and thus quickly identify what he/she really wants in the large set of results. The task aims to investigate and evaluate different techniques and strategies of recommending facet-values to the user at each step in a search session.

4.1 Topics

Each participating group was asked to create a set of candidate topics representative of real user needs. Each topic consists of a general topic as well as a subtopic that refines the general topic by specifying a particular interest of it. The general topic had to result in more than 1000 results, while the subtopics had to be restrictive enough to be satisfied by 10 to 50 results.

Each group had to submit 4 topics: two from the set of general topics given by the organizers, and two proposed by the participants themselves. The given set of general

topics was: {"trained animals", "dogme", "food", "asian cinema", "art house", "silent movies", "second world war", "animation", "nouvelle vague", "wuxia"}.

After removing incorrectly-formed topics, we got a total of 15 general topics along with their 20 subtopics (2 subtopics for "Food", 3 subtopics for "Cannes" and 3 subtopics for "Vietnam"). An example of topic is shown in Fig. 2. The general topic is specified in the <general> field of the <topic> element, while the other fields of <topic>, e.g. <title> and <castitle>, are used to specify the subtopic, which is the searcher's real intention when submitting this general topic to the search system. The participants running the faceted search task could only view the 15 general topics, while the corresponding 20 subtopics were added to the set of topics for the ad hoc search task. The relevance results for these subtopics were used as the relevance results for their corresponding general topics. Thus, altogether we got 45 topics for the ad hoc search task and 15 topics for the faceted search task.

```
<topic id="2011202" guid="28">
  <task>Faceted</task>
  <general>animation</general>
  <title>animation fairy-tale</title>
  <castitle>//movie[about(.//genre, animation) and about(.//plot, fairy-tale)]</castitle>
  <description>I am searching for all animation movies based on a fairy-tale.</description>
  <narrative>I like fairy-tales and their animations remakes.</narrative>
</topic>
```

Fig. 2. INEX 2011 Data Centric Track Faceted Search Topic 2011202

4.2 Submission Format

Each participant had to submit up to 3 runs. A run consists of two files: one is the result file containing a ranked list of maximum 2000 results per topic in the ad hoc search task format, and the other is the recommended facet-value file, which can be a static facet-value file or a dynamic faceted search module.

(1) Facet-Value File. It contains a hierarchy of recommended facet-values for each topic, in which each node is a facet-value and all of its children constitute the newly recommended facet-value list as the searcher selects this facet-value to refine the query. The maximum number of children for each node is restricted to be 20. The submission format is in an XML format conforming to the following DTD.

```
<!ELEMENT run    (topic+)>
<!ATTLIST run    rid    ID     #REQUIRED>
<!ELEMENT topic  (fv+)>
<!ATTLIST topic  tid    ID     #REQUIRED>
<!ELEMENT fv     (fv*)>
<!ATTLIST fv     f      CDATA  #REQUIRED
                 v      CDATA  #REQUIRED>
```

Here:

- The root element is <run>, which has an ID type attribute, *rid*, representing the unique identifier of the run. It must be identical with that in the result file of the same run.

- The <run> contains one or more <topic>s. The ID type attribute, *tid*, in each <topic> gives the topic number.
- Each <topic> has a hierarchy of <fv>s. Each <fv> shows a facet-value pair, with *f* attribute being the facet and *v* attribute being the value. The facet is expressed as an XPath expression. The set of all the possible facets represented as XPath expressions in the IMDB data collection can be found in Appendix 1. We allow only categorical or numerical fields to be possible facets. Free-text fields are not considered. Each facet-value pair represents a facet-value condition to refine the query. For example, <fv f="/movie/overview/directors/director" v="Yimou Zhang"> represents the condition /movie/overview/directors/director="Yimou Zhang".
- The <fv>s can be nested to form a hierarchy of facet-values.

An example submission is:

```
<run rid="2011UniXRun1">
  <topic tid="2011001">
    <fv f="/movie/overview/directors/director" v="Yimou Zhang">
      <fv f="/movie/cast/actors/actor/name" v="Li Gong">
        <fv f="/movie/overview/releasedates/releasedate" v="2002"/>
        <fv f="/movie/overview/releasedates/releasedate" v="2003"/>
      </fv>
      <fv f="/movie/cast/actors/actor/name" v="Ziyi Zhang">
        <fv f="/movie/overview/releasedates/releasedate" v="2005"/>
      </fv>
    </fv>
    …
  </topic>
  <topic tid="2011002">
      ...
  </topic>
…
</run>
```

Here for the topic "2011001", the faceted search system first recommends the facet-value condition /movie/overview/directors/director="Yimou Zhang" among other facet-value conditions, which are on the same level of the hierarchy. If the user selects this condition to refine the query, the system will recommend a new list of facet-value conditions, which are /movie/cast/actors/actor/name="Li Gong" and /movie/cast/actors/actor/name="Ziyi Zhang", for the user to choose from to further refine the query. If the user then selects /movie/cast/actors/actor/name="Li Gong", the system will recommend /movie/overview/releasedates/releasedate="2002" and /movie/overview/releasedates/releasedate ="2003". Note that the facet-value conditions that are selected to refine the query form a path in the tree, e.g. /movie/overview/directors/director="Yimou Zhang" → /movie/cast/actors/actor/name = "Li Gong" → /movie/overview/releasedates/releasedate ="2003". It is required that no facet-value condition occurs twice on any path.

(2) Faceted Search Module. Instead of submitting a static hierarchy of facet-values, participants are given the freedom to dynamically generate lists of recommended facet-values and even change the ranking order of the candidate result list at each step in the search session. This is achieved by submitting a self-implemented dynamically linkable module, called Faceted Search Module (FSM). It implements the *FacetedSearchInterface* defined as the following:

```
public interface FacetedSearchInterface {
   public String[] openQuery(String topicID, String[] resultList);
   public String[] selectFV(String facet, String value, String[] selectedFV);
   public String[] refineQuery(String facet, String value, String[] selectedFV);
   public String[] expandFacet(String facet, String[] selectedFV);
   public void closeQuery(String topicID);
}

public class FacetedSearch implements FacetedSearchInterface {
    // to be implemented by the participant
}
```

The User Simulation System (USS) used in evaluation will interact with the FSM to simulate a faceted search session. The USS starts to evaluate a run by instantiating a FacetedSearch object. For each topic to be evaluated, the USS first invokes openQuery() method to initialize the object with the topic id and initial result list for this topic. The result list is actually the list of retrieved file names (without .xml) in the third column of the result file. The method would return a list of recommended facet-values for the initial result list. A facet-value is encoded into a String in the format "<facet>::<value>", for example, "/movie/overview/directors/director::Yimou Zhang".

After opening a query, the USS then simulates a user's behavior in a faceted search system based on some user model as described in Section 5. When the simulated user selects a facet-value to refine the query, the selectFV() method would be called to return a new list of recommended facet-values; and the refineQuery() method would be called to return a list of candidate results in the initial result list that satisfy all the selected facet-value conditions. The inputs to both methods are the currently selected facet and value, as well as a list of previously selected facet-values. A facet-vaue pair is encoded into a String in the format shown above.

If the user could not find a relevant facet-value to refine the query in the recommended list, he/she could probably expand the facet-value list by choosing a facet among all possible facets, examine all its possible values and then select one to refine the query. In such a case, the USS invokes the expandFacet() method with the name of the facet to be expanded as well as a list of previously selected facet-values as input and the list of all possible values of this facet as output. Observe that in the specification of FacetedSearchInterface, we do not restrict facet-value comparisons to be of equality, but can be of any other possible semantics since the interpretation of facet-value conditions is capsulated into the implementation of FacetedSearchInterface. Thus, given the same facet, different systems may give different sets of all possible values depending on if they will cluster and how they will cluster some values.

When the search session of a query ends, the closeQuery() method is invoked. The FacetedSearch object will be used as a persistent object over the entire evaluation of a run. That is, different topics in the same run will be evaluated using the same FacetedSearch object. But different runs may have different implementations of the FacetedSearch class.

5 Assessments and Evaluations

In total 35 ad hoc search runs and 13 faceted search runs were submitted by 9 active participants. Assessment was done using the same assessment tool as that used in INEX 2010 Data-Centric Track provided by Shlomo Geva. 38 ad hoc topics among 45 ones were assessed by those groups that submitted runs. Among the assessed topics, there are 9 list type topics, 6 known-item type topics, 5 informational type topics, and 18 subtopics for 13 faceted search topics.

Table 1 shows the mapping between the subtopics in ad hoc search task and the general topics in faceted search task. The relevance results of subtopics are treated as the intended results for their corresponding general topics. Note that some general topics, e.g. 2011205, 2011207 and 2011210, have more than one intention/subtopic. For these general topics, we take the subtopics that have the least number of relevance results. For example, compared with topic 2011120 and 2011142, topic 2011141 has the least number of relevance results, whose relevance results are then chosen as the relevance results for topic 2011205. The chosen subtopics are underlined in Table 1. Since the subtopics 2011121 and 2011139 were not assessed, we have no relevance results for topics 2011206 and 2011215 in the faceted search task.

Table 1. Mapping between the faceted search topics and subtopics in ad hoc task.

General Topics	Subtopics
2011201	2011111
2011202	2011114
2011203	2011118
2011204	2011119
2011205	2011120,2011141,2011142
2011206	2011121
2011207	2011112,2011140
2011208	2011129
2011209	2011130
2011210	2011135,2011144,2011145
2011211	2011143
2011212	2011136
2011213	2011137
2011214	2011138
2011215	2011139

The TREC MAP metric, as well as P@5, P@10, P@20 and so on, was used to measure the performance of all ad hoc runs at whole document retrieval.

For the faceted search task, since it is the first year, we used the following two types of evaluation approaches and metrics to gain better understanding to the problem.

- **NDCG of facet-values:** The relevance of the hierarchy of recommended facet-values is evaluated based on the relevance of the data covered by these facet-values, measured by NDCG. The details of this evaluation methodology are given in [2].

- **Interaction cost:** The effectiveness of a faceted search system is evaluated by measuring the interaction cost or the amount of efforts spent by a user in meeting his/her information needs. To avoid expensive user study and make the evaluation repeatable, we applied user simulation methodology like that used in [3, 4] to measure the costs.

We can use two metrics to measure the user's interaction cost. One is the number of results, facets or facet-values that the user examined before he/she encounters the first relevant result, which is similar to the Reciprocal Rank metric in traditional IR. Here we assume that the effort spent on examining each facet or facet-value is the same as that spent on examining each result. The other is the number of actions that the user performs in the search session. We only consider the click actions.

As in [3, 4], we assume that the user will end the search session when he/she encounters the first relevant result, and the user can recognize the relevant results from the list of results, and can distinguish the relevant facets or facet-values that match at least one relevant result from the list of facets or facet-values.

The user begins by examining the first page of the result list for the current query. It is assumed that each page displays at most 10 results. If the user finds relevant results on the first page, the user selects the first one and ends the session. If no relevant result is found, the user then examines the list of recommended facet-values. If there are relevant facet-values, the user then clicks on the first relevant facet-value in the list to refine the query, and the system returns the new lists of results and facet-values for the refined query. If none of the recommended facet-values is relevant, the user chooses the first relevant facet in the list of all possible facets to expand and select the first relevant value in this facet's value list to refine the query. If the user does not find any relevant facet to expand, the user begins to scan through the result list and stops at the first relevant result encountered. Fig. 3 shows the flowchart of the user interaction model and cost model used in the evaluation. Notation used in Fig. 3 is given in Table 2.

Table 2. Notation used in Fig. 3.

Symbol	Meaning
q	The current query
R_q	The result list of query q
FV_q	The list of recommended facet-values for query q
F_q	The list of all possible facets for query q
loc(x,y)	A function returns the position of item x in the list y
cost	The number of results, facet-values or facets examined by the user
actionCount	The number of click actions performed by the user

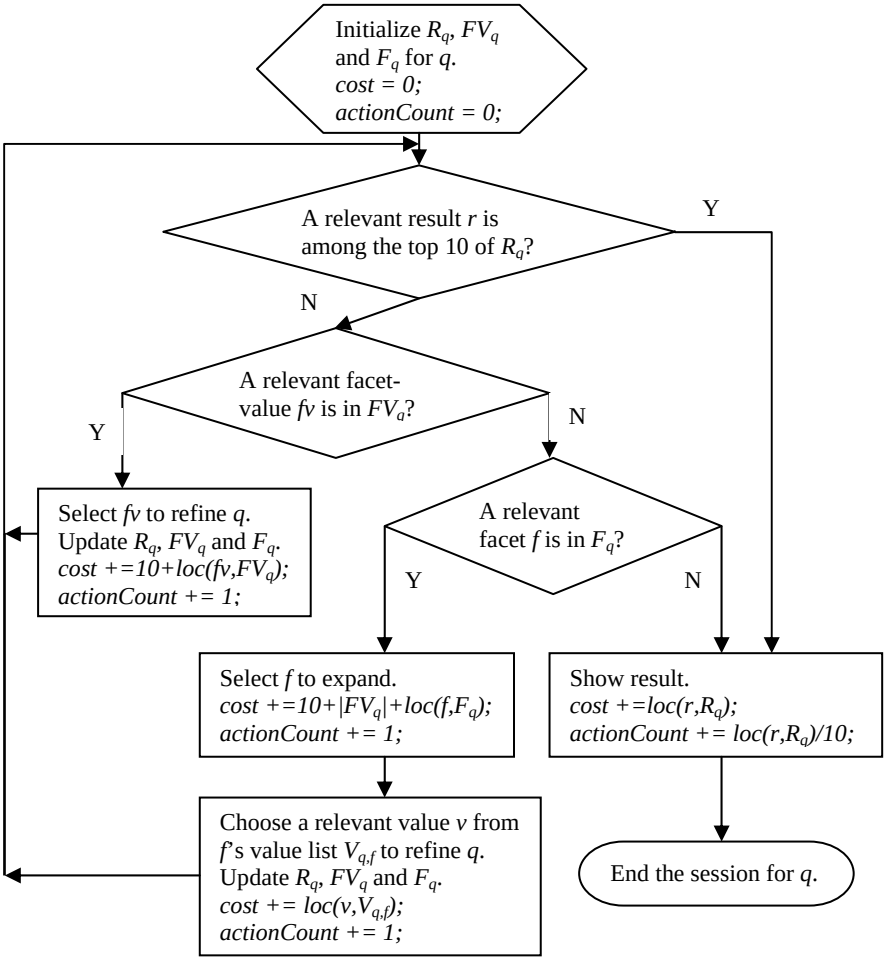

Fig. 3. Flowchart of the Simulated User Interaction Model with Faceted Search System

6 Results

6.1 Ad Hoc Search Results

As mentioned above, a total of 35 runs from 9 different institutes were submitted to the ad hoc search task. This section presents the evaluation results for these runs. Results were computed over the 38 topics assessed by the participants using the TREC evaluation tool. The topic set is a mixture of informational, known-item, list, and faceted (sub)topics. We use MAP as the main measure since it averages reasonably well over such a mix of topic types.

Table 3 shows an overview of the 10 best performing runs for this track. Over all topics, the best scoring run is from the University of Amsterdam with a MAP of 0.3969. Second best scoring team is Renmin University of China (0.3829). Third best scoring team is Kasetsart University (0.3479) with the highest score on mean reciprocal rank (1/rank). Fourth best team is Peking University (0.3113) and the highest precision at 10. Fifth best team is Universitat Pompeu Fabra, with a MAP of 0.2696 but the highest scores for precision at 20 and 30.

Table 3. Best performing runs (only showing one run per group) based on MAP over all ad hoc topics.

Run	map	1/rank	P@10	P@20	P@30
p4-UAms2011adhoc	**0.3969**	0.6991	0.4263	0.3921	0.3579
p2-ruc11AS2	0.3829	0.6441	0.4132	0.3842	0.3684
p16-kas16-MEXIR-2-EXT-NSW	0.3479	**0.6999**	0.4316	0.3645	0.3298
p77-PKUSIGMA01CLOUD	0.3113	0.5801	**0.4421**	0.4066	0.3851
p18-UPFbaseCO2i015	0.2696	0.5723	0.4342	**0.4171**	**0.3825**
p30-2011CUTxRun2	0.2099	0.6104	0.3684	0.3211	0.2965
p48-MPII-TOPX-2.0-co	0.1964	0.5698	0.3684	0.3395	0.3289
p47-FCC-BUAP-R1	0.1479	0.5120	0.3474	0.2763	0.2412
p12-IRIT_focus_mergeddtd_04	0.0801	0.2317	0.2026	0.1724	0.1702

Interpolated precision against recall is plotted in Fig. 4, showing quite solid performance for the better scoring runs.

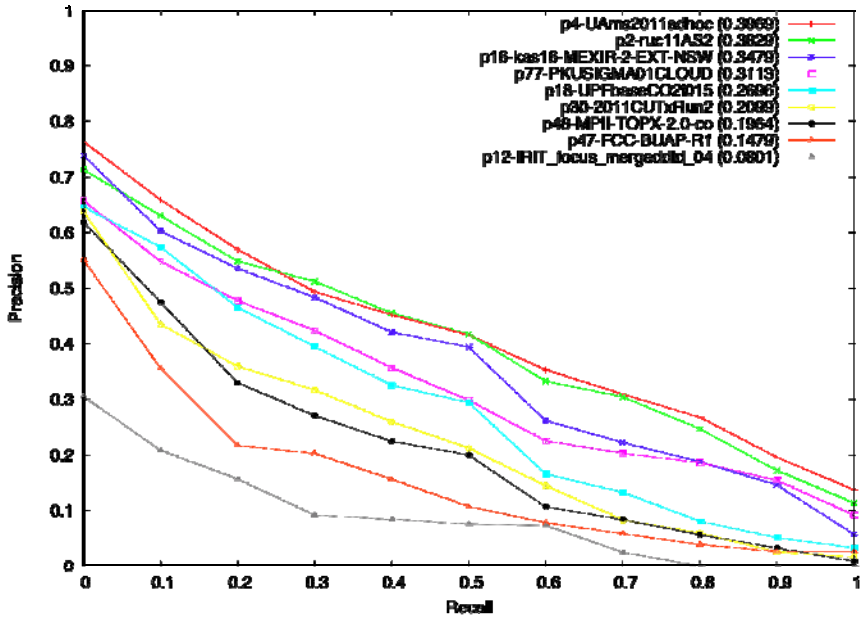

Fig. 4. Best run by each participating institute measured with MAP

Breakdown over Topic Types

In this section, we will analyze the effectiveness of the runs for each of the four topic types. Let us first analyze the topics and resulting judgments in more details. Table 4 lists the topics per topic type, and Table 5 lists statistics about the number of relevant entities.

Table 4. Breakdown over Topic Types

Topic Type	Topics created	Topics Judged	Topics with relevance
Informational	7	5	5
Known-Item	7	6	6
List	11	9	8
Faceted subtopics	20	18	18
All	45	38	37

Table 5. Relevance per Topic Type

Topic Type	Topics	Min	Max	Median	Mean	Std.	Total
Informational	5	6	327	40	125.8	150.4	629
Known-Item	6	1	416	2	71.3	168.9	428
List	8	5	299	32	98.6	118.1	789
Faceted subtopics	18	23	452	148	168.3	123.8	3,029
All	37	1	452	72	168.3	134.0	4,875

What we see in Table 4 is that we have 5 (informational) to 18 (faceted sub-) topics judged for each type. Given the small number of topics per type, one should be careful with drawing final conclusions based on the analysis, since the particular choice of topics may have had a considerable influence on the outcome.

While all topics have been judged “as is” without special instructions for each of the topic types, the statistics of the relevance judgments in Table 5 is confirming the differences between these topic types. The known-item topics have a median of 2 relevant documents, the list topics have a median of 32 relevant documents, and the informational topics have a median of 40. The faceted (sub)topics, which were based on a general seed topic, have even a median of 148 relevant documents. For all topic types the distribution over topics is skewed, and notable exceptions exist, e.g. a known-item topic with 416 relevant documents.

Table 6 shows the results over only the informational topics. We see that Kasetsart (0.3564), Chemnitz (0.3449), and BUAP (0.3219) now have the best scores, and that there are less differences in scores amongst the top 5 or 6 teams. Over all 34 submissions the system rank correlation (Kendall’s tau) with the ranking over all topics is moderate with 0.512.

Table 6. Best performing runs (only showing one run per group) based on MAP over the 5 informational ad hoc topics.

run	map	1/rank	P@10	P@20	P@30
p16-kas16-MEXIR-2-EXT-NSW	**0.3564**	0.8000	0.5000	0.4200	0.3600
p30-2011CUTxRun2	0.3449	0.7067	0.5000	**0.4700**	0.4333
p47-FCC-BUAP-R1	0.3219	**1.0000**	**0.5600**	0.4300	0.4133
p2-ruc11AMS	0.3189	0.6500	0.4200	0.4500	**0.4600**
p4-UAms2011adhoc	0.3079	0.6750	0.3800	0.3100	0.2600
p18-UPFbaseCO2i015	0.2576	0.6346	0.4600	0.4400	0.3800
p77-PKUSIGMA02CLOUD	0.2118	0.5015	0.4400	0.4200	0.3133
p48-MPII-TOPX-2.0-co	0.0900	0.3890	0.2600	0.1800	0.2000
p12-IRIT_focus_mergeddtd_04	0.0366	0.3022	0.2200	0.1100	0.0733

Table 7 shows the results over only the known-item topics, now evaluated by the mean reciprocal rank (1/rank). We observe that Amsterdam (0.9167), Renmin (also 0.9167), and MPI (0.7222). Hence the best teams over all topics score also well over the known-item topics. This is no surprise since the known-item topics tend to lead to relatively higher scores, and hence have a relatively large impact. Over all 34 submissions the system rank correlation based on MAP is 0.572.

Table 7. Best performing runs (only showing one run per group) based on 1/rank over the 6 known-item ad hoc topics.

run	map	1/rank	P@10	P@20	P@30
p4-UAms2011adhoc	**0.8112**	**0.9167**	**0.3167**	**0.2417**	**0.2167**
p2-ruc11AS2	0.7264	**0.9167**	**0.3167**	**0.2417**	**0.2167**
p48-MPII-TOPX-2.0-co	0.2916	0.7222	0.2333	0.1833	0.1778
p18-UPFbaseCO2i015	0.3752	0.7104	0.2500	0.2083	0.1944
p16-kas16-MEXIR-2-EXT-NSW	0.4745	0.6667	0.0833	0.0417	0.0278
p77-PKUSIGMA01CLOUD	0.5492	0.6389	**0.3167**	**0.2417**	**0.2167**
p30-2011CUTxRun2	0.3100	0.5730	0.2667	0.1750	0.1667
p47-FCC-BUAP-R1	0.2500	0.3333	0.0333	0.0167	0.0111
p12-IRIT_large_nodtd_06	0.0221	0.0487	0.0167	0.0333	0.0222

Table 8 shows the results over the list topics, now again evaluated by MAP. We see the best scores for Kasetsart (0.4251), Amsterdam (0.3454), and Peking University (0.3332). The run from Kasetsart outperforms all other runs on all measures for the list topics. Over all 34 submissions the system rank correlation is 0.672.

Table 8. Best performing runs (only showing one run per group) based on MAP over the 8 list ad hoc topics.

run	map	1/rank	P@10	P@20	P@30
p16-kas16-MEXIR-2-EXT-NSW	**0.4251**	**0.7778**	**0.4778**	**0.3833**	**0.3741**
p4-UAms2011adhoc	0.3454	0.6674	0.4222	0.3500	0.3222
p77-PKUSIGMA02CLOUD	0.3332	0.5432	0.3889	0.3667	0.3481
p2-ruc11AS2	0.3264	0.6488	0.4111	0.3333	0.2963
p48-MPII-TOPX-2.0-co	0.2578	0.4926	0.3000	0.3333	0.3259
p18-UPFbaseCO2i015	0.2242	0.5756	0.3556	0.3278	0.2741

p12-IRIT_focus_mergeddtd_04	0.1532	0.2542	0.2333	0.2111	0.2148
p30-2011CUTxRun3	0.0847	0.5027	0.1889	0.1611	0.1667
p47-FCC-BUAP-R1	0.0798	0.3902	0.2889	0.2500	0.2259

Table 9 shows the results over the faceted search subtopics (each topic covering only a single aspect). We see the best performance in the runs from Renmin (0.3258), Amsterdam (0.3093), and Peking University (0.3026), with Peking University having clearly the best precision scores. Given that 18 of the 37 topics are in this category, the ranking corresponds reasonably to the ranking over all topics. Over all 34 submissions the system rank correlation is high with 0.818.

Table 9. Best performing runs (only showing one run per group) based on MAP over the 18 facted ad hoc topics.

run	map	1/rank	P@10	P@20	P@30
p2-ruc11AS2	**0.3258**	0.5585	0.4722	0.4778	0.4722
p4-UAms2011adhoc	0.3093	0.6492	0.4778	0.4861	0.4500
p77-PKUSIGMA02CLOUD	0.3026	**0.7400**	**0.5722**	**0.5361**	**0.5315**
p16-kas16-MEXIR-2-EXT-NSW	0.2647	0.6443	0.5056	0.4472	0.4000
p18-UPFbaseCO2i015	0.2605	0.5072	0.5278	0.5250	0.5000
p30-2011CUTxRun2	0.2130	0.6941	0.4611	0.4083	0.3741
p48-MPII-TOPX-2.0-co	0.1635	0.6078	0.4778	0.4389	0.4167
p47-FCC-BUAP-R1	0.0995	0.4969	0.4222	0.3333	0.2778
p12-IRIT_focus_mergeddtd_04	0.0810	0.2754	0.2500	0.2278	0.2296

6.2 Faceted Search Results

In the faceted search task, 5 groups, University of Amsterdam (Jaap), Max-Plank Institute, University of Amsterdam (Maarten), Universitat Pompeu Fabra, and Renmin University of China, submitted 12 valid runs. All runs are in the format of static hierarchy of facet-values except that one run from Renmin is in the format of a self-implemented faceted search module. So we only present the evaluation results for the 11 static runs. Most of the runs are based on the reference result file provided by Anne Schuth, who generated the reference result file using XPath and Lucene. Two runs from Amsterdam (Jaap) are based on a result file generated by Indri and one run from Max-Plank Institute is based on the result file generated by TopX.

13 out of 15 general topics have relevance results. Table 10 shows, for each topic, the number of relevant results, and the rank of the first relevant result in the three result lists generated by Indri, Lucene and TopX respectively, which is in fact the cost that users sequentially scan through the list of results to find the first relevant answer without using the faceted-search facility. We call it raw cost, which is actually equal to 1/RR. "-" means that the result file contains no relevant result for this topic. It can be observed that the Indri result file contains relevant results for all topics and ranks them quite high. The TopX result file ranks the first relevant results for 7 topics highest among the three result files, but it fails in containing relevant results for 3 topics. The Lucene reference result file, however, is the worst one.

Table 10. Raw costs (1/RR) of faceted search topics on 3 different result files.

Topic ID	Number of relevant results	Raw cost of Indri result file	Raw cost of Lucene result file	Raw cost of TopX result file
2011201	48	**45**	-	97
2011202	327	**11**	19	85
2011203	138	**114**	451	-
2011204	342	**306**	989	-
2011205	141	9	316	**1**
2011207	23	69	850	**44**
2011208	285	2	11	**1**
2011209	76	**1**	2	**1**
2011210	23	217	-	**49**
2011211	72	61	45	**40**
2011212	156	1110	-	**344**
2011213	35	828	-	-
2011214	176	**4**	44	16

We use two metrics to evaluate the effectiveness of recommended facet-values by each run. One is the interaction cost based on a simple user simulation model, and the other is the NDCG of facet-values [2].

As described in Section 5, the interaction cost is defined as the number of results, facets or facet-values that the user examined before he/she encounters the first relevant result. This cost can be compared with the raw cost, which is the number of results sequentially examined in the result list without using faceted search facility, to see if the faceted search facility is effective or not. We name their difference as the *Gain* of faceted search. To compare systems across multiple topics, we define the *Normalized Gain (NG)* and *Average Normalized Gain (ANG)* as the following. Note that *NG* is a number between 0 and 1.

$$NG = \max(0, (rawCost - Cost) / rawCost) \quad (1)$$

$$ANG = \frac{1}{|Q|} \sum_{i=1}^{|Q|} NG_i \quad (2)$$

Table 11 shows the evaluation results for all the 11 runs in terms of *NG* and *ANG*. Two runs from Amsterdam (Jaap), p4-UAms2011indri-c-cnt and p4-UAms2011indri-cNO-scr2, are based on the Indri result file, and p48-MPII-TOPX-2.0-facet-entropy (TopX) from Max-Plank is based on the TopX result file. All the other 8 runs are based on the Lucene result file. Because the Indri result file is superior to the TopX and Lucene result files, the two runs based on it perform also better than other runs, and the best one is p4-UAms2011indri-cNO-scr2 (0.35). Among all the 8 runs based on the Lucene result file, p2-2011Simple1Run1 (0.33) from Renmin performs best in terms of ANG. It is followed by p4-UAms2011Lucene-cNO-lth (0.24) from Amsterdam (Jaap), p18-2011UPFfixGDAh2 (0.21) from Universitat Pompeu Fabra and p4-2011IlpsNumdoc (0.20) from Amsterdam (Maarten).

The NDCG scores calculated using the method described in [2] for all 11 static runs are listed in Table 12. For *p* we chose 10 (we thus consider the top 10 documents per facet-value) and we also limited the number of facet-values to be evaluated to 10. Note that we did not evaluate the runs using NRDCG.

Table 11. Evaluation results of all static runs in terms of NGs and ANG.

run / NG	p4-UAms2011indri-c-cnt	p4-UAms2011indri-cNO-scr2	p4-UAms2011lucene-cNO-lth	p48-MPII-TOPX-2.0-facet-entropy (TopX)	p48-MPII-TOPX-2.0-facet-entropy (Lucene)	p4-2011IlpsFtScore	p4-2011IlpsNumdoc	p18-2011UPFfixG7DAnh	p18-2011UPFfixGDAh	p18-2011UPFfixGDAh2	p2-2011Simple1Run1
201	0.64	0.60	-	0	-	-	-	-	-	-	-
202	0	0	0.21	0	0	0	0.21	0.11	0.11	0.11	0.21
203	0	0	0	-	0	0.83	0.82	0	0	0	0.86
204	0.63	0.75	0.94	-	0	0	0.90	0	0	0.98	0.91
205	0	0	0.81	0	0.75	0.79	0.72	0.95	0.95	0.95	0.81
207	0	0.77	0	0	0	0	0	0	0	0	0.94
208	0	0	0	0	0	0	0	0	0	0	0
209	0	0	0	0	0	0	0	0	0	0	0
210	0.75	0.74	-	0	-	-	-	-	-	-	-
211	0.18	0	0.53	0	0	0	0	0.71	0.71	0.71	0.60
212	0.89	0.88	-	0	-	-	-	-	-	-	-
213	0.76	0.76	-	-	-	-	-	-	-	-	-
214	0	0	0.64	-	0	0	0	0	0.09	0	0
ANG	0.30	**0.35**	0.24	0	0.06	0.12	0.20	0.14	0.14	0.21	**0.33**

Table 12. Evaluation results for the 11 statics runs in terms of NDCG. Results are per topic and the mean over all topics. Highest scores per topic are highlighted.

run / NDCG	p4-UAms2011indri-c-cnt	p4-UAms2011indri-cNO-scr2	p4-UAms2011lucene-cNO-lth	p48-MPII-TOPX-2.0-facet-entropy (TopX)	p48-MPII-TOPX-2.0-facet-entropy (Lucene)	p4-2011IlpsFtScore	p4-2011IlpsNumdoc	p18-2011UPFfixG7DAnh	p18-2011UPFfixGDAh	p18-2011UPFfixGDAh2	p2-2011Simple1Run1
201	**0.03**	**0.03**	0	0	0	0	0	0	0	0	0
202	0	0	0	0	0	0	0	0	0	0	0
203	0	0	0	0	0	0	0	0	0	0	0
204	0	0	0	0	0	0	0	0	0	0	0
205	0	**0.43**	0.21	0	0	0	0.13	0	0	0	0.07
207	0	**0.16**	0	0	0	0	0	0	0	0	0
208	0	**0.45**	0	0	0	0	0	0	0	0	0
209	0	0	0	0	0	0	**0.24**	0	0	0	0
210	0	0	0	0	0	0	0	0	0	0	0
211	0	0	0	0	0	0	0	0	0	0	0
212	0	0	0	0	0	0	0	0	0	0	0
213	0	0	0	0	0	0	0	0	0	0	0
214	**0.18**	**0.18**	0.09	0	0	0	0	0	0	0	0
mean	0.02	**0.10**	0.02	0	0	0	0.03	0	0	0	0.01

Note that the NDCG calculation used the union of relevance judgments in case there were multiple subtopics for a topic. Statistics for the relevance judgments used for the NDCG evaluation are listed in Table 13.

Table 13. Relevance judgments for faceted search topics.

Topic Type	Topics	Min	Max	Median	Mean	Std.	Total
Faceted	13	35	774	156	233	229.3	3029

7 Conclusions and Future Work

We presented an overview of the INEX 2011 Data-Centric Track. This track has successfully run its second year and has introduced a new task, the *faceted search* task. The IMDB collection has now a good set of assessed topics that can be further used for research on richly structured data. Our plan for next year is to extend this collection with related ones such as DBpedia and Wikipedia in order to reproduce a more realistic scenario for the newly introduced faceted search task.

Acknowledgements

Thanks are given to the participants who submitted the topics, runs, and performed the assessment process. Special thanks go to Shlomo Geva for porting the assessment tools, to Anne Schuth, Yu Sun and Yantao Gan for evaluating the faceted search runs, and to Ralf Schenkel for administering the web site.

References

1. A. Trotman, Q. Wang, Overview of the INEX 2010 Data Centric Track, INEX 2010.
2. A. Schuth, M.J. Marx, Evaluation Methods for Rankings of Facetvalues for Faceted Search, Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation 2011.
3. J. Koren, Y. Zhang, X. Liu, Personalized Interactive Faceted Search, WWW 2008.
4. A. Kashyap, V. Hristidis, M. Petropoulos, FACeTOR: Cost-Driven Exploration of Faceted Query Results, CIKM 2010.

Appendix 1: All the Fields or Facets in IMDB Collection

Field Type	Field (or Facet) expressed in XPath
free-text	/movie/title
numerical	/movie/overview/rating
categorical	/movie/overview/directors/director
categorical	/movie/overview/writers/writer
numerical	/movie/overview/releasedates/releasedate
categorical	/movie/overview/genres/genre
free-text	/movie/overview/tagline

free-text	/movie/overview/plot
categorical	/movie/overview/keywords/keyword
categorical	/movie/cast/actors/actor/name
categorical	/movie/cast/actors/actor/character
categorical	/movie/cast/composers/composer
categorical	/movie/cast/editors/editor
categorical	/movie/cast/cinematographers/cinematographer
categorical	/movie/cast/producers/producer
categorical	/movie/cast/production_designers/production_designer
categorical	/movie/cast/costume_designers/costume_designer
categorical	/movie/cast/miscellaneous/person
free-text	/movie/additional_details/aliases/alias
categorical	/movie/additional_details/mpaa
numerical	/movie/additional_details/runtime
categorical	/movie/additional_details/countries/country
categorical	/movie/additional_details/languages/language
categorical	/movie/additional_details/colors/color
categorical	/movie/additional_details/certifications/certification
categorical	/movie/additional_details/locations/location
categorical	/movie/additional_details/companies/company
categorical	/movie/additional_details/distributors/distributor
free-text	/movie/fun_stuff/trivias/trivia
free-text	/movie/fun_stuff/goofs/goof
free-text	/movie/fun_stuff/quotes/quote
categorical	/person/name
categorical	/person/overview/birth_name
numerical	/person/overview/birth_date
numerical	/person/overview/death_date
numerical	/person/overview/height
categorical	/person/overview/spouse
free-text	/person/overview/trademark
free-text	/person/overview/biographies/biography
categorical	/person/overview/nicknames/name
free-text	/person/overview/trivias/trivia
free-text	/person/overview/personal_quotes/quote
free-text	/person/overview/where_are_they_now/where
categorical	/person/overview/alternate_names/name
numerical	/person/overview/salaries/salary
free-text	/person/filmography/act/movie/title
numerical	/person/filmography/act/movie/year
categorical	/person/filmography/act/movie/character
free-text	/person/filmography/direct/movie/title
numerical	/person/filmography/direct/movie/year
categorical	/person/filmography/direct/movie/character
free-text	/person/filmography/write/movie/title
numerical	/person/filmography/write/movie/year
categorical	/person/filmography/write/movie/character
free-text	/person/filmography/compose/movie/title
numerical	/person/filmography/compose/movie/year
categorical	/person/filmography/compose/movie/character
free-text	/person/filmography/edit/movie/title
numerical	/person/filmography/edit/movie/year
categorical	/person/filmography/edit/movie/character
free-text	/person/filmography/produce/movie/title
numerical	/person/filmography/produce/movie/year
categorical	/person/filmography/produce/movie/character
free-text	/person/filmography/production_design/movie/title
numerical	/person/filmography/production_design/movie/year
categorical	/person/filmography/production_design/movie/character

free-text	/person/filmography/cinematograph/movie/title
numerical	/person/filmography/cinematograph/movie/year
categorical	/person/filmography/cinematograph/movie/character
free-text	/person/filmography/costume_design/movie/title
numerical	/person/filmography/costume_design/movie/year
categorical	/person/filmography/costume_design/movie/character
free-text	/person/filmography/miscellaneous/movie/title
numerical	/person/filmography/miscellaneous/movie/year
categorical	/person/filmography/miscellaneous/movie/character
free-text	/person/additional_details/otherworks/otherwork
free-text	/person/additional_details/public_listings/interviews/interview
free-text	/person/additional_details/public_listings/articles/article
free-text	/person/additional_details/public_listings/biography_prints/print
free-text	/person/additional_details/public_listings/biographical_movies/biographical_movie
free-text	/person/additional_details/public_listings/portrayed_ins/portrayed_in
free-text	/person/additional_details/public_listings/magazine_cover_photos/magazine
free-text	/person/additional_details/public_listings/pictorials/pictorial