Overview of the INEX 2010 Data Centric Track

Andrew Trotman

Department of Computer Science
University of Otago
Dunedin
New Zealand

Qiuyue Wang

School of Information
Renmin University of China
Beijing
China

Abstract. The INEX 2010 Data Centric Track is discussed. A dump of IMDb was used as the document collection, 28 topics were submitted, 36 runs were submitted by 8 institutes, and 26 topics were assessed. Most runs (all except 2) did not use the structure present in the topics; and consequently no improvement is yet seen by search engines that do so.

1 Introduction

2010 sees the introduction of the Data Centric Track at INEX. The results of INEX up-to and including 2009 showed that whole document retrieval was effective. This result was, perhaps, a consequence of the IEEE and Wikipedia collections used in the past. It is reasonable to assume that the Wikipedia will include a whole document result to almost any ad hoc query.

In the Data Centric Track we ask: Can whole document retrieval approaches outperform focused retrieval on highly structured document collection?

To answer this question a new highly structured collection was developed and made available for download from the INEX website. That collection was a snapshot of the IMDb taken early in 2010. Highly structured queries were solicited from participants. Together with the assessments, these form the new INEX Data Centric Collection.

Most of the runs submitted to the track did not use the structure present in the topics. This is not surprising in a new track because participants are inclined to use their existing systems on a new collection before making modifications to it. Consequently the track has not yet seen improvements in precision from structure. It is hoped that in future years participating groups will prefer to conduct experiments using the structure present in the topics. The track has generated a topic set that can be used for training.

2 The Task

In its first year, the track focused on ad hoc retrieval from XML data. An XML document is typically modeled as a rooted, node-labeled tree. An answer to a keyword query was defined as a set of *closely related* nodes that are *collectively relevant* to the query. So each result could be specified as a collection of nodes from one or more XML documents that are related and collectively cover the relevant information[1]. The task was to return a ranked list of results estimated relevant to the user's information need. The content of the collections of nodes was not permitted to overlap. This is similar to the focused task in the ad hoc track, but using a data-centric XML collection and allowing the construction of a result (i.e. a collection of nodes) from different parts of a single document or even multiple documents.

3 INEX Data Centric Track Collection

3.1 Document Collection

The track used the IMDb data collection newly built from www.IMDb.com. It was converted from the plain text files (April 10, 2010) published on the IMDb web site. The plain text files were first loaded into a relational database by the Java Movie Database system[2]. Then the relational data are published as XML documents according to the DTDs. There are two kinds of objects in the IMDb data 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. Please refer to Appendix A and B for the movie DTD and person DTD respectively.

Information about one movie or person is published in one XML file, thus each generated XML file represents a single object, i.e. a movie or person. In total, 4,418,102 XML files were generated, including 1,594,513 movies, 1,872,492 actors[3], 129,137 directors who did not act in any movies, 178,117 producers who did not direct or act in any movies, and 643,843 other people involved in movies who did not produce or direct nor act in any movies.

3.2 Topics

Each participating group was asked to create a set of candidate topics, representative of a range of real user needs. Both Content Only (CO) and Content And Structure (CAS) variants of the information need were requested. In total 30 topics were submitted by 4 institutes (IRIT / SIG, Renmin University of China, Universidade

[1] However, as it is unclear how to evaluate this the standard INEX metrics were eventually used
[2] http://www.jmdb.de/
[3] 21 of the actor files were empty and removed.

Federal do Amazonas, and Universitat Pompeu Fabra). From these a total of 28 topics were selected based on uniqueness, preciseness, and being correctly formed. An example topic (2010001) is given in Fig**.** 1:

```
<topic id="2010001" ct_no="3">
<title>Yimou Zhang 2010 2009</title>
<castitle>//movie[about(.//director, "Yimou Zhang")
        and (about(.//releasedate, 2010) or about(.//releasedate, 2009))]</castitle>
<description>I want to know the latest movies directed by Yimou Zhang.</description>
<narrative>
        I am interested in all movies directed by Yimou Zhang,
        and I want to learn the latest movies he directed.
</narrative>
</topic>
```

Fig. 1. INEX 2010 Data Centric Track Topic 2010001

4 Submission Format

The required submission format was a variant of the familiar TREC format used by INEX, the so called TREC++ format. The following information was collected about each run:

- The participant ID of the submitting institute,
- Whether the query was constructed automatically or manually from the topic,
- Topic fields used (from: Title, CASTitle, Description, and Narrative),

A run was permitted to contain a maximum of 1000 results for each topic. A result consisted of one or more nodes from a single or multiple XML documents. A node is uniquely identified by its element path in the XML document tree. The standard TREC format is extended with one additional field for specifying each result node: <qid> Q0 <file> <rank> <rsv> <run_id> <column_7>

Here:

- the first column is the topic number.
- the second column is the query number within that topic (unused and should always be Q0).
- the third column is the file name (without .xml) from which a result node is retrieved.
- the fourth column is the rank of the result. Note that a result may consist of one or more related nodes, so there can be multiple rows with the same rank if these nodes belong to the same result.
- the fifth column shows the score that generated the ranking. This score must be in descending (non-increasing) order and is important to include so that assessment tools can handle tied scores (for a given run) in a uniform fashion (the evaluation routines rank documents from these scores, not from ranks). If you want the precise ranking that you submit to be evaluated, the scores should reflect that ranking.

- the sixth column is called the "run tag" and should be a unique identifier from within a participating group. It should also include a brief detail of the method used. The run tags contained 12 or fewer letters and numbers, with no punctuation.
- the seventh column gives the element path of the result node. Element paths are given in XPath syntax. To be more precise, only fully specified paths are allowed, as described by the following grammar:

Path ::= '/' ElementNode Path | '/' ElementNode | '/' AttributeNode
ElementNode ::= ElementName Index
AttributeNode ::= '@' AttributeName
Index ::= '[' integer ']'

For Example the path */article[1]/bdy[1]/sec[1]/p[1]* identifies the element which can be found if we start at the document root, select the first *article* element, then within that, select the first *body* element, within which we select the first *section* element, and finally within that element we select the first *p* element.

An example submission is:

1 Q0 9996 1 0.9999 I09UniXRun1 /article[1]/bdy[1]/sec[1]
1 Q0 9996 1 0.9999 I09UniXRun1 /article[1]/bdy[1]/sec[2]/p[1]
1 Q0 9888 1 0.9999 I09UniXRun1 /article[1]/bdy[1]/sec[3]
1 Q0 9997 2 0.9998 I09UniXRun1 /article[1]/bdy[1]/sec[2]
1 Q0 9989 3 0.9997 I09UniXRun1 /article[1]/bdy[1]/sec[3]/p[1]

Here there are three results. The first result contains the first section and first paragraph of the second section from 9996.xml, and the third section from 9888.xml. The second result only consists of the second section in 9997.xml, and the third result consists of the first paragraph of the third section from 9989.xml.

5 Submitted Runs

Participants were permitted to submit up to 10 runs. Each run was permitted to contain a maximum of 1000 results per topic, ordered by decreasing value of relevance. Runs were permitted to use any fields of the topics, but only runs using either the <title>, or <castitle>, or a combination of them were regarded as truly automatic. The results of one run was contained in one submission file and so up to 10 files per group could be submitted.

In total 36 runs were submitted by 8 institutes. Those institutes were: Benemérita Universidad Autónoma de Puebla, Indian Statistical Institute, Kasetsart University, Peking University, Renmin University of China, Universidade Federal do Amazonas, Universitat Pompeu Fabra, and the University of Otago. Only 29 runs were assessed since other runs were submitted after the deadline. Of note is that more runs were

submitted than topics, and more institutes submitted runs than that submitted topics. This suggests an increase in interest in the track throughout the year.

6 Assessment and Evaluation

Shlomo Geva ported the tool to work with the IMDb collection and in doing so identified some problems with the document collection. The collection was, consequently, cleaned for use with the assessment tool. The new collection will most likely be used in 2011, if the track continues.

Assessment was done by those groups that submitted runs. In total 26 of the 28 topics were assessed. Topics 2010003 and 20100013 were not assessed, all others were. The evaluation results presented herein were computed using just the 26 assessed topics with the other 2 topics dropped from the runs.

Jaap Kamps used the (unmodified) INEX and TREC evaluation tools on the runs. The TREC MAP metric was used to measure the performance of the runs at whole document retrieval. The INEX thorough retrieval MAiP metric and the INEX Relevant-in-Context MAgP T2I(300) metrics were used to measure Focused Retrieval[4]. Although the run submission permitted the use of aggregated retrieval, it has not yet become clear how to measure aggregation and so the track organisers chose to fall-back to more traditional measures for 2010. Descriptions of the INEX and TREC measures are not given herein as they are well known and described elsewhere (see the ad hoc track overview paper pre-proceedings).

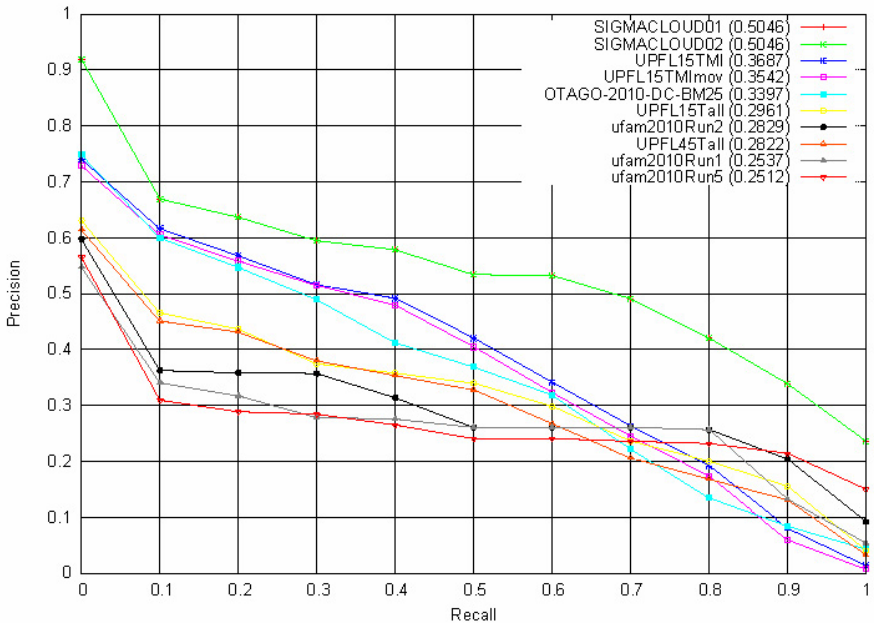

Fig. 2. Best runs measured with MAP

[4] See the ad hoc track overview paper (in this volume) for details on the metrics

7 Results

The performance of the runs using the whole document based MAP metric are presented in Fig. 2. The best run, SIGMACLOUD01 was submitted by Peking University and performed substantially better than the next best run at all recall points. We note that this run used the description and narrative of the topic whereas the other runs did not (formally it is not an INEX automatic run and must be considered a manual run). The runs from Benemérita Universidad Autónoma de Puebla used the castitle and all other runs used the title. That is, despite being data centric, most runs did not use structure in ranking.

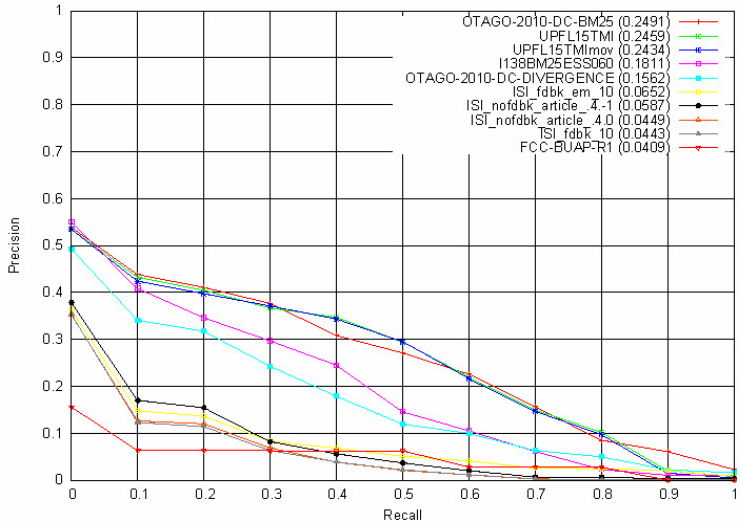

Fig. 3. Best runs measured with MAgP T2I(300)

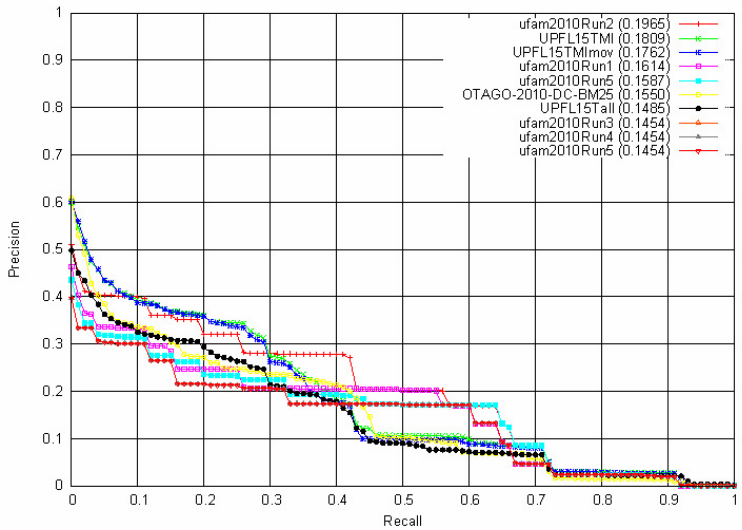

Fig. 4. Best runs measured with MAiP

From visual inspection, there is little difference between the next three runs. The Otago run (that placed 3rd amongst the automatic runs) is a whole document run generated from the title of the topic by using the BM25 ranking function trained on the INEX 2009 document collection – it is equivalent to the ad hoc reference run. It can be considered a baseline for performance.

When measured using the MAgP T2I(300) metric (see Fig. 3 the Otago reference-like run performs best, however there is a cluster of 3 runs performing at about (from visual inspection) the same level. When measured using MAiP (Fig. 4) the reference-like run shows high early precision but quickly decreases. Of course, whole document retrieval is not a good strategy for thorough retrieval because precisely 1 element is returned per document. Those runs that exhibited overlap were not evaluated using the MAgP metric.

8 Conclusions

The track has successfully produced a highly structured document collection including structured documents (IMDb), structured queries, and assessments. The participants of the track submitted runs and those runs were evaluated. Because most runs did not use structured queries no claim can be made about the advantage of doing so. This is expected to change in future years. The track was overly ambitions in allowing result aggregation. No method of measuring the performance of aggregated retrieval was developed for the track in 2010 and is left for future years.

9 Acknowledgements

Thanks are given to the participants who submitted the topics, the run, and performed the assessment process. Special thanks go to Shlomo Geva for porting the assessment tools, and to Jaap Kamps for performing the evaluation. Finally, some of the contents of this paper was taken from the INEX web site which was authored by many people – we thank each of those for their contribution to the text in this paper. Qiuyue Wang is supported by the 863 High Tech. Project of China under Grant No. 2009AA01Z149.

Appendix A: Movie DTD

```
<!ELEMENT movie (title, url, overview?, cast?, additional_details?, fun_stuff?)>
<!ATTLIST movie xmlns:xlink CDATA #FIXED "http://www.w3.org/1999/xlink">
<!ELEMENT title (#PCDATA)>
<!ELEMENT url (#PCDATA)>

<!ELEMENT overview (rating?, directors?, writers?, releasedates?, genres?, tagline?,
plot?, keywords?) >
<!ELEMENT rating (#PCDATA)>
<!ELEMENT directors (director+)>
<!ELEMENT director (#PCDATA)>
<!ELEMENT writers (writer+)>
<!ELEMENT writer (#PCDATA)>
<!ELEMENT releasedates (releasedate+)>
<!ELEMENT releasedate (#PCDATA)>
<!ELEMENT genres (genre+)>
<!ELEMENT genre (#PCDATA)>
<!ELEMENT tagline (#PCDATA)>
<!ELEMENT plot (#PCDATA)>
<!ELEMENT keywords (keyword+)>
<!ELEMENT keyword (#PCDATA)>

<!ELEMENT cast (actors?, composers?, editors?, cinematographers?, producers?,
production_designers?, costume_designers?, miscellaneous?)>
<!ELEMENT actors (actor+)>
<!ELEMENT actor (name, character?)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT character (#PCDATA)>
<!ELEMENT composers (composer+)>
<!ELEMENT composer (#PCDATA)>
<!ELEMENT editors (editor+)>
<!ELEMENT editor (#PCDATA)>
<!ELEMENT cinematographers (cinematographer+)>
<!ELEMENT cinematographer (#PCDATA)>
<!ELEMENT producers (producer+)>
<!ELEMENT producer (#PCDATA)>
<!ELEMENT production_designers (production_designer+)>
<!ELEMENT production_designer (#PCDATA)>
<!ELEMENT costume_designers (costume_designer+)>
<!ELEMENT costume_designer (#PCDATA)>
<!ELEMENT miscellaneous (person+)>
<!ELEMENT person (#PCDATA)>

<!ELEMENT additional_details
(aliases?,mpaa?,runtime?,countries?,languages?,colors?,certifications?,locations?,com
panies?,distributors?)>
```

```
<!ELEMENT aliases (alias+)>
<!ELEMENT alias (#PCDATA)>
<!ELEMENT mpaa (#PCDATA)>
<!ELEMENT runtime (#PCDATA)>
<!ELEMENT countries (country+)>
<!ELEMENT country (#PCDATA)>
<!ELEMENT languages (language+)>
<!ELEMENT language (#PCDATA)>
<!ELEMENT colors (color+)>
<!ELEMENT color (#PCDATA)>
<!ELEMENT certifications (certification+)>
<!ELEMENT certification (#PCDATA)>
<!ELEMENT locations (location+)>
<!ELEMENT location (#PCDATA)>
<!ELEMENT companies (company+)>
<!ELEMENT company (#PCDATA)>
<!ELEMENT distributors (distributor+)>
<!ELEMENT distributor (#PCDATA)>

<!ELEMENT fun_stuff (trivias?,goofs?,quotes?,movielinks?)>
<!ELEMENT trivias (trivia+)>
<!ELEMENT trivia (#PCDATA)>
<!ELEMENT goofs (goof+)>
<!ELEMENT goof (#PCDATA)>
<!ELEMENT quotes (quote+)>
<!ELEMENT quote (#PCDATA)>
<!ELEMENT movielinks (movielink+)>
<!ELEMENT movielink (#PCDATA?, link, #PCDATA?)>
<!ELEMENT link (#PCDATA)>
<!ATTLIST link xlink:type CDATA #IMPLIED>
<!ATTLIST link xlink:href CDATA #IMPLIED>
```

Appendix B: Person DTD

```
<!ELEMENT person (name, overview?,filmography?, additional_details?)>
<!ELEMENT name (#PCDATA)>

<!ELEMENT overview (birth_name?, birth_date?, death_date?, height?, spouse*, trademark*, biographies?, nicknames?, trivias?, personal_quotes?, where_are_they_now?, alternate_names?, salaries?) >
<!ELEMENT birth_name (#PCDATA)>
<!ELEMENT birth_date (#PCDATA)>
<!ELEMENT death_date (#PCDATA)>
<!ELEMENT height (#PCDATA)>
<!ELEMENT spouse (#PCDATA)>
<!ELEMENT trademark (#PCDATA)>
<!ELEMENT biographies (biography+)>
<!ELEMENT biography (#PCDATA, by)>
<!ELEMENT by (#PCDATA)>
<!ELEMENT nicknames (name+)>
<!ELEMENT trivias (trivia+)>
<!ELEMENT trivia (#PCDATA)>
<!ELEMENT personal_quotes (quote+)>
<!ELEMENT quote (#PCDATA)>
<!ELEMENT where_are_they_now (where+)>
<!ELEMENT where (#PCDATA)>
<!ELEMENT alternate_names (name+)>
<!ELEMENT salaries (salary+)>
<!ELEMENT salary (#PCDATA)>

<!ELEMENT filmography (act?, direct?, write?, compose?, edit?, produce?, production_design?, cinematograph?, costume_design?, miscellaneous?)>
<!ELEMENT act (movie+)>
<!ELEMENT direct (movie+)>
<!ELEMENT write (movie+)>
<!ELEMENT compose (movie+)>
<!ELEMENT edit (movie+)>
<!ELEMENT produce (movie+)>
<!ELEMENT production_design (movie+)>
<!ELEMENT cinematograph (movie+)>
<!ELEMENT costume_design (movie+)>
<!ELEMENT miscellaneous (movie+)>
<!ELEMENT movie (title, year, character?)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT character (#PCDATA)>
```

<!ELEMENT additional_details (otherworks?, public_listings?)>
<!ELEMENT otherworks (otherwork+)>
<!ELEMENT otherwork (#PCDATA)>
<!ELEMENT public_listings (interviews?, articles?, biography_prints?, biographical_movies?, portrayed_ins?, magazine_cover_photos?, pictorials?)>
<!ELEMENT interviews (interview+)>
<!ELEMENT interview (#PCDATA)>
<!ELEMENT articles (article+)>
<!ELEMENT article (#PCDATA)>
<!ELEMENT biography_prints (print+)>
<!ELEMENT print (#PCDATA)>
<!ELEMENT biographical_movies (biographical_movie+)>
<!ELEMENT biographical_movie (#PCDATA)>
<!ELEMENT portrayed_ins (portrayed_in+)>
<!ELEMENT portrayed_in (#PCDATA)>
<!ELEMENT magazine_cover_photos (magazine+)>
<!ELEMENT magazine (#PCDATA)>
<!ELEMENT pictorials (pictorial+)>
<!ELEMENT pictorial (#PCDATA)>