Building Test Collection from Old IR Literature

Anirban Chakraborty, Kripabandhu Ghosh and Swapan Kumar Parui
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
203, B.T. Road, Kolkata - 700108, India
{chakraborty.abhi89, kripa.ghosh, swapan.parui}@gmail.com

ABSTRACT

Standard test collections form the very basis of Information Retrieval research and evaluation. Important datasets have been created to promote empirical research and experimentation. In this paper, we describe our endeavour in creating a test collection from old, archived writings of IR stalwarts. The documents are created in text format from the scanned and OCRed version. The test collection consists of a set of documents in TREC format along with a set of expert queries and their relevance assessments. This dataset, though small in size, would be of paramount interest for researchers and students of IR since it contains valuable discourses on the discipline from its very inception. Also, to the best of our knowledge, no standard IR dataset has been built so far comprising old research articles. Furthermore, this is a dataset without the original error-free digital text version. So, the resulting collection would expect researchers to run retrieval experiments on the erroneous collection without the scope of error modeling. This would invite new research ideas.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Test Collection, Old Literature, OCR Errors

1. INTRODUCTION

Test collections are indispensable in an empirical science like Information Retrieval. Unlike any data retrieval task where the content of a correct response to a query is easily specified, IR tasks treat correctness of a response as a matter of opinion. A returned document is considered *relevant* with respect to a query if the user of the search system would wish to see that document, and *non relevant* otherwise [13]. The notion of *relevance* is determined by a group of experts, known an *assessors*, who would attach a degree of relevance to a document-query pair. The history of IR evaluation has been shaped by a research paradigm broadly known as "Cranfield tradition". The *Cranfield tradition* refers to a set of guidelines laid down at Cranfield Institute of Technology in the 1960's [3]. According to the tradition, a test collection should be composed of three components :

- A set of text descriptions of information needs, referred to as *topics*, *requests*, or *queries*.
- A static collection of text documents.
- A set of manually produced assessments, known as *relevance judgements*, specifying (binary or graded) relevance of each document in the collection corresponding to each query.

Several small collections were built following the Cranfield tradition norms. However, the real major initiative came from TREC in 1992 by the U. S. National Institute of Standards and Technology (NIST) [16]. TREC has produced several useful standard datasets for various tasks which have evolved over time. TREC inspired other evaluation forums like CLEF [14], NTCIR [9], INEX [11] and FIRE [12].

In this work, we have made an effort to form a test collection from the scanned articles archived by ACM SIGIR[1]. It contains articles on IR by Cyril W. Cleverdon, Gerard Salton, Joseph John Rocchio, K. Sparck Jones and other experts. The archive, known as ACM SIGIR Museum, allows free download of these articles. However, it will be an important contribution if a Cranfield-style dataset can be constructed from this repository. That will allow IR researchers to deploy their algorithms on the resulting dataset.

The rest of the paper is organised as follows :

We discuss the related works in Section 2. We then describe building the test collection in Section 3. Finally, we conclude in Section 4.

2. RELATED WORKS

The Cranfield collection[2] was the pioneering test collection in allowing precise quantitative measures of information retrieval effectiveness. It was collected in the United Kingdom starting in the late 1950s. It contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgements of all (query, document)

FIRE '14, December 05-07, 2014, Bangalore, India

DOI: http://dx.doi.org/10.1145/2824864.2824869

[1] http://sigir.org/resources/museum/

[2] http://ir.dcs.gla.ac.uk/resources/test_collections/cran/

pairs. TREC produced several test collections for IR research and evaluation. The U.S. National Institute of Standards and Technology (NIST) has run a large IR test bed evaluation series since 1992. Within this framework, there have been many tracks over a range of different test collections, but the best known test collections are the ones used for the TREC Ad Hoc track during the first 8 TREC evaluations between 1992 and 1999. In total, these test collections comprise 6 CDs containing 1.89 million documents (mainly, but not exclusively, newswire articles) and relevance judgements for 450 information needs which are called topics and specified in detailed text passages. Individual test collections are defined over different subsets of this data. The early TRECs each consisted of 50 information needs, evaluated over different but overlapping sets of documents. In more recent years, NIST has done evaluations on larger document collections, including the 25 million page GOV2 web page collection. NII Test Collections for IR Systems or NTCIR project has built various test collections of similar sizes to the TREC collections, focusing on East Asian languages and cross-language information retrieval, where queries are made in one language over a document collection containing documents in one or more other languages[3]. Cross Language Evaluation Forum (CLEF) has concentrated on European languages and cross-language information retrieval[4]. FIRE has focused mainly on building test collections in Indian languages like Bengali, Hindi, Marathi, Punjabi, Tamil, and Telugu[5].

When it comes to datasets created by scanning and OCRing documents, two important such datasets were created by TREC for the Confusion Track and the Legal Track. The TREC Confusion track was part of TREC-4 (1995) [5] and TREC-5 (1996) [10]. In this track, thousands of documents were printed, scanned and then OCRed to form a noisy collection. However, electronic text for the same documents was available for comparison. In TREC legal track [7] [15] [4] [2] the IIT CDIP 1.0 collection was prepared by OCRing 7 million scanned English business documents for Ad Hoc, Relevance Feedback and Batch tasks from 2006 to 2009. This collection was about 57 GB in size and was extremely noisy. Moreover, it lacked the error-free electronic version which made error-modeling difficult. In addition to TREC, some work has been done on IR from OCRed collections. Among the earliest works, Taghva et al. [8] applied probabilistic IR on OCRed text. Singhal et al. [1] showed that linear document length normalization models were better suited to collections containing OCR errors than the quadratic (cosine normalization) models.

Test collections based on previously published research papers have been used in TREC[6] and CLEF[7]. Donna et al. [6] embarked on a project of scanning old IR reports such as the Cranfield reports, ISR reports along with Karen Sparck Jones' Information Retrieval Experiment book. The initial archive at SIGIR Museum[8] was built with SIGIR funding. More scanned documents have been added to this archive in the course of time. We have attempted to convert this archive to standard TREC test collection format. We have created queries and relevance assessments suitable for IR evaluation.

[3]http://research.nii.ac.jp/ntcir/data/data-en.html
[4]http://www.clef-campaign.org/
[5]http://www.isical.ac.in/~fire/
[6]http://www.trec-cds.org/2014.html
[7]http://bioasq.org/
[8]http://sigir.org/resources/museum/

3. BUILDING OF THE TEST COLLECTION

3.1 Document Collection

3.1.1 Scanned Collection

The source collection is archived as ACM SIGIR Museum at http://sigir.org/resources/museum/. It contains important notes like:

- "Report on the first stage of an investigation onto the comparative efficiency of indexing systems" and "Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems" by Cyril W. Cleverdon
- "Factors determining the performance of indexing systems; Volume 1: Design" by Cyril W. Cleverdon, Jack Mills, E. Michael Keen
- "Information Storage and Retrieval: Scientific ISR Reports" by Gerard Salton
- "Report on a Design Study for the 'IDEAL' Information Retrieval Test Collection" by K. Sparck Jones, R.G. Bates
- "Information Retrieval Experiment" by Karen Sparck Jones
- "New Models in Probabilistic Information Retrieval" by C.J. van Rijsbergen, S.E. Robertson, M.F. Porter
- "A front-end for IR experiments" by S.E. Robertson, J.D. Bovey

It contains 237 scanned PDF documents. The scanning was done at 600 dpi grayscale [6]. Each document can span from a single page to 114 pages. The whole collection comprises 8790 pages.

3.1.2 Image Collection

Each PDF file is converted to JPEG format using Linux *convert* command. The density was chosen as 500, quality was kept at 100 and the sharpness was 0x1.0. One .jpg image file was produced for one page of PDF file. So, each multi-page PDF file produced many .jpg files. For example, Salton.pdf file contains 3 pages. Then, *convert* command will produce three .jpg files, say, Salton-1.jpg, Salton-2.jpg and Salton-3.jpg; Salton-1.jpg corresponds to the first page of Salton.pdf, Salton-2.jpg corresponds to the second page of Salton.pdf and Salton-3.jpg corresponds to the last page of Salton.pdf.

3.1.3 Text Collection

Each .jpg file produced is now OCRed using Tesseract[9], an open source optical character recognition engine. Here, the files Salton-1.jpg, Salton-2.jpg and Salton-3.jpg are OCRed and stored as Salton-1.txt, Salton-2.txt and Salton-3.txt respectively. Finally, Salton-1.txt, Salton-2.txt and Salton-3.txt are merged using *cat* command in Linux platform into a single file, say Salton.txt

[9]http://code.google.com/p/tesseract-ocr/

Information	Value
No. of documents	237
No. of pages	8790
No. of unique terms	37,828
Total number of terms	5,61,215
Number of test queries created	24

Table 1: Collection Statistics

Correct	Misrecognised
formulation	eoemelaeion
block diagram	blockhdiagram
facility	facilityj
processes	processesyff
system	sxstem
formalizer	fobmalizer
syntactic	nyntactic
relations	AfnâĂŸtâĂŸpqrelations
desirable	desirïñĆÃl'ble
specify which	spbectifyd
retrieved	eeÃl'sieoeo
perturbations	pertubations
request	reguest

Table 2: OCR errors : terms

3.1.4 Formatting the Text Collection

Finally, the resulting text collection is converted to standard TREC document format. A sample TREC format document looks as follows:

```
<DOC>
<DOCNO>acm_sigir_1</DOCNO>
<TEXT>
7 CHAPTER 5 âĂŤ

SEARCH REQUEST EOEMELAEION
.3

1. Introduction . _ ' \

The dialogue initiated by user-generated inputs to a
......

</TEXT>
</DOC>
```

The XML tags <DOC>$\cdots$</DOC> encloses a whole document. Each document can be uniquely identified by the *document number* enclosed by <DOCNO>$\cdots$</DOCNO>. The OCRed text is placed inside the tags <TEXT>$\cdots$ </TEXT>.

3.1.5 OCR Errors

OCRing the documents may result in recognition errors. The absence of ground truth clean text makes automatic error detection infeasible. So, we have performed a manual test on the text collection for error analysis. We have randomly selected 15 pieces of text from the final text collection and made a word-wise comparison with the corresponding scanned source version. We have found that the OCR produces about 95% accurate results at word-level.

Table 1 shows the statistics of the text collection. The table also shows the number of test queries created by us. We will discuss about query creation in Section 3.2. Note that, here "term" refers to the smallest unit of text produced delimited by white spaces after preprocessing steps like stopword removal, case-folding and stemming. We have used the Indri[10] toolkit for indexing the text collection. The Indri standard stopwords list was used. Porter stemmer[11] implemented in Indri was used for stemming. Despite high OCR accuracy, some erroneous terms were introduced to the collection. For example, the sample TREC format document snapshot in section 3.1.4 has the word "FORMULATION" misrecognised as "EOEMELAEION". Table 2 shows some terms with their misrecognised forms. Misrecognitions of different types have been found in the corpus. We see that some characters have been wrongly recognised, e.g., in the term "formulation" the characters 'f', 'r', 'u' and 't' have been recognised as 'e' and consequently, the whole word has been recognised as "eoemelaeion". In some cases, two valid terms have been joined by a character to form an invalid term. For example, "block diagram" has become "blockhdiagram". The addition of stray characters has corrupted some valid words. For example, "facility" and "processes" have become "facilityj" and "processesyff" respectively. Moreover, some characters have been omitted. For example, "perturbations" has been misrecognised as "pertubations", where the character 'r' has been missed by the OCR. Misrecognitions of important terms may lead to poor retrieval performance. This highlights the importance of error-handling during retrieval exercises on the text collection.

We are looking to create a text-only collection devoid of any pictorial or tabular representations. However, the source scanned collection has diagrams, graphs, mathematical equations, formulae, tables etc. which are not recoverable from the image format through OCRing. Such illustrations form an integral part of the research articles. This valuable information is lost while converting the scanned images into text format. Also, these images contribute invalid terms produced from the figures as well as from the misrecognised text portions of the same. Therefore, some thought should be put on how the lost pictorial and tabular information should be recovered and reconstructed so as to reinstate the integrity of the source collection. This is a vital aspect without which the text collection will be far from being complete and sound.

3.2 Query Creation

This is a domain-specific collection. Any information-seeking activity on this collection would need basic knowledge about IR. Hence, the queries on this collection would be questions on IR concepts and experimentation. So, for the purpose of query development, we formed a group of people who had some experience in IR. However, this group was divided into two broad categories based on their level of knowledge, viz., *students* and *experts*. The category *students* comprised Masters students doing their dissertation on IR, young researchers and scholars pursuing a Ph.D. degree in IR. The group of *experts* consisted of Post Doctoral fellows and teachers of IR. The idea behind this categorization was to study the information seeking behaviour of people with different levels of knowledge and expertise on the subject.

3.2.1 Topic Selection

[10] www.http://sourceforge.net/projects/lemur/
[11] www.tartarus.org/martin

We selected a set of documents from the whole collection to find topics from them. We made a thorough search of the collection to ensure a good coverage of the research issues discussed in the lecture notes over the collection. Then we selected a representative subset of documents and assigned them to the group of *students* and *experts*. We requested them to find topics from the subset that represented important IR questions. The members from the group were asked to do the job on their own. They were instructed not to consult with the other members of the group. They came up with interesting topics on IR. However, we found a marked difference between the *students* and *experts* in the basic approach. The *students* usually found general topics like "Information Retrieval definition", "Boolean retrieval", "Rocchio Relevance Feedback", etc. On the other hand, the *experts* selected more specific topics like "Effect of document length on retrieval performance", "Query-specific clustering of pseudo-relevant documents", "Word associations in improving retrieval", etc.

3.2.2 *Query Formation*

After the topic selection phase, we talked with each member of the group separately. In order to get the actual queries, we had sessions with a search engine. From the topics, we made several rephrasing exercises to reach a query which is neither too broad in scope nor too narrow. We consulted the corresponding person(s) who selected the topic about the correctness of the queries thus formed. Finally, a total of 24 queries were selected. Table 3 shows some selected queries with the "title" and the "description" parts. The topics suggested by both the *students* and *experts*, were unanimously chosen. "Relevance Assessments and Retrieval System Evaluation", "Pseudo relevance feedback effectiveness" are some of such queries.

3.3 Relevance Assessment

The relevance assessment phase is an integral part of query formation. This is because, each time a new form of a query is created from the topic, the aptness of the query was judged from the top retrieved documents returned by the search engine. This process also helped the *students* and *experts* to have a clearer idea of their notion of relevance associated with each topic selected by them. Too many relevant documents at the top of the ranked list indicated over-broadness or ambiguity in the query. On the other hand, too specific queries tended to have a very few relevant documents at the top of the ranked list. Finally, after a set of queries was agreed upon, we had a formal pooling process. The same search engine was used for generating different runs by changing the retrieval models and parameters. For each query, a pool was formed from the top 100 documents per ranked list. For a given query, relevance assessment was done by at least two persons in the group. We made sure that at least one of the assessors for a query is the creator of the query in order to maintain integrity in the notion of relevance. However, the team for a given query also had one neutral annotator, i.e., who was not a creator of the query. This was done to get an unbiased view about the relevance of the query. In case of assessor disagreements, we had a discussion with all the assessors together for the query.

As mentioned earlier, a total of 24 queries were selected. Relevance assessments were created for all these queries.

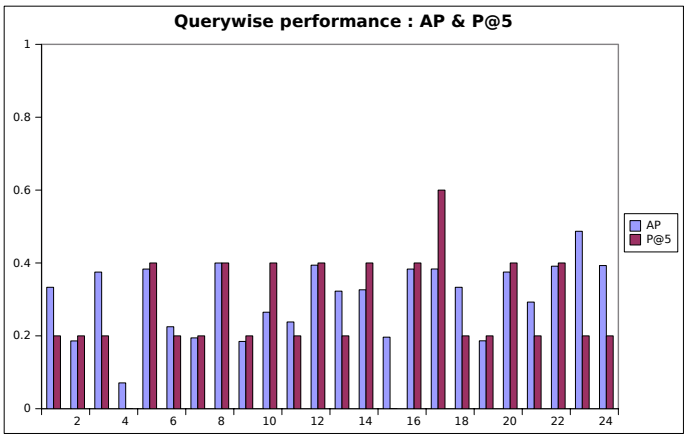

Figure 1: Querywise Performance

Figure 1 shows querywise Average Precision (AP) and Precision at 5 (P@5).

4. CONCLUSION

Creation of an IR dataset from old literature is a novel effort. The ACM SIGIR Museum is a stepping stone towards building a test collection comprising research articles in different domains and perhaps, different languages. In this paper, we have made a humble effort of forming a document collection consisting of old IR notes. We have also presented a small and yet carefully created set of expert test queries and relevance assessments. In future, we plan to create a bigger collection by incorporating old lecture notes of other disciplines like Mathematics, Physics, etc. from the original hard-copy version. Another issue with such collections is the presence of errors introduced by the OCRing of the scanned source files. The absence of error-free text version makes error-modeling difficult. This would encourage new retrieval challenges from the erroneous corpus.

5. REFERENCES

[1] G. S. A. Singhal and C. Buckley. Length normalization in degraded text collections. pages 149–162. In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, 1996.

[2] J. B. B. Hedin, S. Tomlinson and D. Oard. Overview of the trec 2009 legal track. The Eighteenth Text Retrieval Conference, 2009.

[3] C. Cleverdon. Readings in information retrieval. chapter The Cranfield Tests on Index Language Devices, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[4] S. T. D. Oard, B. Hedin and J. Baron. Overview of the trec 2008 legal track. The Seventeenth Text Retrieval Conference, 2008.

[5] D. Harman. Overview of the fourth text retrieval conference. pages 1–24. The Fourth Text Retrieval Conference, 1995.

[6] D. Harman and D. Hiemstra. Saving and accessing the old ir literature. *ACM SIGIR Forum*, 42(2):16–21, 2008.

Query no.	Title	Description
2	Pseudo relevance feedback effectiveness	What is the relation of the effectiveness of pseudo relevance feedback (PRF) with the number of relevant documents retrieved at top ranks after the initial retrieval?
4	Relevance Assessments and Retrieval System Evaluation	The procedure of relevance assessment with the use of pooling and query-specific document annotation.
9	Deriving word-word associations	How are word-word associations derived automatically using a collection of documents or other external resources?
10	Effect of document length on recall	What effect does the indexed length of a document have on recall performance? Does indexing only the abstract, which is a summary of the key points of a document, help improve recall?
24	Cluster hypothesis in Information Retrieval	Explain the cluster hypothesis of IR and how is it useful for pseudo-relevance feedback.

Table 3: Selected queries

[7] D. L. J. Baron and D. Oard. The trec-2006 legal track. The Fifteenth Text Retrieval Conference, 2006.

[8] J. B. K. Taghva and A. Condit. Results of applying probabilistic ir to ocr text. pages 202–211. In The Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994.

[9] N. Kando, T. Mitamura, and T. Sakai. Introduction to the ntcir-6 special issue. 7(2):4:1–4:3, Apr. 2008.

[10] P. Kantor and E. Voorhees. Report on the trec-5 confusion track. pages 65–74. The Fifth Text Retrieval Conference, 1996.

[11] G. Kazai, M. Lalmas, N. Fuhr, and N. GÃűvert. A report on the first year of the initiative for the evaluation of xml retrieval (inex'02). *Journal of the American Society for Information Science and Technology*, 55(6):551–556, 2004.

[12] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Mitra, A. Sen, and S. Pal. Text collections for fire. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 699–700, New York, NY, USA, 2008. ACM.

[13] D. W. Oard, J. R. Baron, B. Hedin, D. D. Lewis, and S. Tomlinson. Evaluation of information retrieval for e-discovery. *Artif. Intell. Law*, 18(4):347–386, 2010.

[14] C. Peters and M. Braschler. European research letter: Cross-language system evaluation: The clef campaigns. *J. Am. Soc. Inf. Sci. Technol.*, 52(12):1067–1072, Oct. 2001.

[15] J. B. S. Tomlinson, D. Oard and P. Thompson. Overview of the trec 2007 legal track. The Sixteenth Text Retrieval Conference, 2007.

[16] E. Voorhees and H. DK. The text retrieval conference. pages 3–19. MIT Press, Cambridge, TREC: experiment and evaluation in information retrieval, 2005.