Towards "Cranfield" Test Collections for Personal Data Search Evaluation

Liadh Kelly
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
lkelly@computing.dcu.ie

Gareth J.F. Jones
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

Desktop archives are distinct from sources for which shared "Cranfield" information retrieval test collections[1]have been created to date. Differences associated with desktop collections include: they are personal to the archive owner, the owner has personal memories about the items contained within them, and only the collection owner can rate the relevance of items retrieved in response to their query. In this paper we discuss these unique attributes of desktop collections and search, and the resulting challenges associated with creating test collections for desktop search. We also outline a proposed strategy for creating test collections for this space.

Keywords

Test collection creation, pseudo desktop collections.

1. INTRODUCTION

Research progress in development of retrieval techniques for the personal search space is hindered by the lack of common shared test collections. To conduct experiments researchers have largely needed to create their own test collections consisting of individuals data, queries and result sets. There are two problems with this approach: 1) the effort required to create these collections; and 2) the difficultly in gaining large volumes of subjects for such experiments. In other spaces (e.g., web search) standardized collections exisit, hence eliminating these problems. The difficulty for standardization in the personal search space, is the personal nature of collections and individuals resulting unwillingness to share these collections. We foresee two possible avenues for standarization in this space: 1) through a blackbox technique where participating institutes submit their retrieval algorithms for evaluation on the personal collections of other participating institutes using an agreed task formation approach, etc; or 2) through development of pseudo desktop collections, queries and result sets. In this paper we focus on development of pseudo collections for standardized IR evaluation in this domain.

[1]Referred to as 'test collections' for remainder of paper.

ECIR Workshop on Evaluating Personal Search, April 18, 2011, Ireland.

2. TOWARDS TEST COLLECTIONS

To our knowledge the only existing work on pseudo desktop test collection creation is [2]. In this work the authors proposed amassing and created 3 pseudo desktop collections by extracting emails of 3 individuals prominant in the W3C collection and locating web pages, word documents, pdf files and powerpoint presentations related to these people by a web search query consisting of the persons name, organization and area of speciality (provided by TREC expert search track). They randomly chose known items from these collections and used a modification to the approach proposed by [1], for simulated query generation for web page re-finding, to generate simulted queries across multi-field personal items. This approach presents a promising new direction towards larger scale test collections creation for the desktop space and means to examine the utility of desktop retrieval approaches without the need for real users and their collections. However, these collections do not represent the diversity of real users collections, and hence may not provide a reliable way to evaluate the performance of retrieval algorithms intended for personal desktop collections. The created collections contain a limited number of item types and the same volume of each provided item type across the three collections (with the exception of emails). Given the personal nature of desktop collections, we can expect individuals to have different types of collections, with varying volumes and types of content, covering varying volumes of topics. Further the generated pseudo collections do not take account of the items individuals will actually want to retrieve from their collections. In addition, it is not known to what extent the query formulation approach used reflects what collection owners will actually recall about required items and hence the query terms they will use. Indeed the query generation approach of Azzopardi et al [1] which forms the core part of this multi-field query formation approach is acknowledged by its authors to require further analysis and refinement to exhibit more of the characteristics observed by individuals in web page re-finding.

To highlight the differences that can be present across real users collections, consider the personal collections of 3 subjects gathered through logging on their laptop and PCs over a period of 20 months[2], shown in Table 1. These individuals fit a common user profile of being computer science post graduates at the same university. However, as can be seen, even for these similar subjects large differences exist in the volumes of different item types in the collections and in the number of re-accesses of the individuals. Extended

[2]See http://www.cdvp.dcu.ie/iCLIPS for further details.

over the entire populous, with widely varying interests and requirements, we suspect much different variations would be noted not only in the volumes of information types contained within individuals collections, but also in the diversity and types of topics covered.

Personal collection owners will also have personal experiences and memories associated with the items in their archive, which will guide, depending on their information needs at given moments in time, the items they wish to retrieve from the archive and the query terms they will use in this retrieval process. Individuals will search their personal collections with different personal intentions, memories of required item and personal query generation styles.

We believe in creating pseudo desktop collections that the make up of real users collections need to be replicated as closely as possible to determine how sucessful retrieval approaches will be on real users collections containing varying types and volumes of data, with requirements for different types of item retrieval using different styles of query formation. In the next section we describe a means to gain an understanding of the make up of these collections.

3. STATISTICAL ANALYSIS

To successfully build pseudo desktop collections which represent 'real' users test collections a detailed understanding of the make up of real users desktop collections, items they retrieve from these collections and query formation styles is required. Part of such an analysis could take the form of observations, user studies, diary studies, etc, as are carried out in the PIM community. However, a detailed statistical analysis of the make up of the collections and querying behaviour of a large cross section of the populous is also required in order to move to a situation where real users collections can be replicated in a pseudo way.

To understand the make up of individuals desktop collections, statistics need to be built up on the volume of different information types in these collections, the volume of topics covered, the amount of similarity between items, etc. This analysis could potentially be conducted through a drive within the research community, with either clear guidelines on the statistics to gather or crawlers to automatically generate statistics from participants PCs provided.

We propose that required statistics for target result items would include: extension type of target item, distinctiveness of target item in collection as a whole, recency of last access to target item, etc. And that required statistics for user queries would include: query length, frequency of query terms in target item, frequency of query term in collection as a whole, etc. Similar to gaining statistics on the content of individuals desktops, a stand alone search application or a tool which plugs into individuals current search application (e.g., Google Desktop) could be provided to the research community to log statistics on the nature of queries performed and items retrieved on subjects computers.

Using gathered statistics we propose generating pseudo collections which mimic the characteristics of 'real' collections, described in the next section.

4. TEST COLLECTION CREATION

Using the statistics gathered for each individuals desktop contents, query format and items retrieved, we believe the techniques developed in [1] and [2] provide a strong foundation from which to build pseudo test collections which mimic the characteristics of 'real' test collections.

Type	Subject 1 Total	Subject 2 Total	Subject 3 Total
code files	590 (17)	183 (9)	2,220 (10)
excel	455 (8)	66 (4)	141 (9)
email	3,760 (1)	2,509 (2)	10,243 (2)
pdf	182 (5)	381 (3)	69 (4)
presentations	92 (10)	147 (3)	95 (19)
web	3,895 (3)	15,642 (2)	44,457 (3)
word	311 (7)	310 (6)	373 (13)
text files	381 (6)	81 (2)	308 (6)
other	7 (23)	32 (11)	40 (2)
TOTAL:	9,673 (4)	19,351 (2)	57,946 (3)

Table 1: Total number of distinct items. Average number of accesses to items provided in brackets.

We propose mimicking desktop content by using the statistics gathered on the make up of individuals desktop content to lay user profiles on top of an extension to the pseudo desktop collection creation approach proposed in [2]. In extending this approach, other information which could be mined in creating these collections includes the details provided by people on their homepage, e.g., many people provide lists of personal and work interests and details on co-workers (either explicitly or through inferred means, e.g., co-authorship of papers in the case of academics) on their homepages. We also envisage possibilities to extend the content gathering approach to include other item types and items generated from web content using exisiting summarization, extraction and rephrasing approaches, for example.

Having created pseudo desktop collections we propose extracting target result sets from each user's collection using the available statistics on what the 'real' user retrieves from their collection. To form the queries for the target items, a query generation process which uses the statistics on the 'real' users query formation for the given target item is required. We envisage the query generation approach proposed by [1] and refined to facilitate multi-field retrieval by [2], coupled with the information gained by our proposed statistical analysis would form a good starting point for development of a query generation process for this space.

5. CONCLUDING REMARKS

To drive the test collection creation approach presented in this paper and to both implement and evaluate its component part's against real collections will require formation of a consortium and possibly creation of TREC-like tracks.

The aim of this paper is to present a potential approach to move towards TREC-like collections for research in desktop search and to highlight the requirements of such collections. Thus generating debate and stimulating further progression on possible directions at the workshop.

6. REFERENCES

[1] L. Azzopardi, M. de Rijke, and L. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR'07*, July 2007.

[2] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *CIKM '09*, 2009.