Building Pseudo-Desktop Collections

Jinyoung Kim and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
{jykim,croft}@cs.umass.edu

ABSTRACT

Research on the desktop search has been constrained by the lack of reusable test collections. This led to a high entry barrier for new researchers and difficulty in the comparative evaluation of existing methods. To address this point, we introduce a method for creating reusable pseudo-desktop collections by gathering documents and generating queries that have similar characteristics to actual collections. Our method involves a new query generation method and a technique for evaluating the similarity of generated queries with user-generated queries.

Categories and Subject Descriptors

H.4 [**Database Management**]; D.2.8 [**Information Storage and Retrieval**]: [Information Search and Retrieval]

Keywords

Desktop Search, Test Collection Generation

1. INTRODUCTION

Although desktop search plays an important role in personal information management, past research has been limited by the lack of availability of shareable test collections. For instance, desktop search prototypes such as Stuff I've Seen [2] and Connections [4] employ evaluation methods based on real users' desktop collections and queries. Based on actual use cases, this type of evaluation is certainly valuable. Yet this approach requires a fully functional desktop search engine and the lack of reusability makes it difficult, if not impossible, to repeat experiments and make comparisons to alternative search techniques.

In this paper, we suggest a methodology for automatically building reusable pseudo-desktop collections, consisting of document gathering and query generation. The resulting collections have many of the characteristics of typical desktop collections and, importantly, are free from the privacy concerns that are common with personal data.

While we cannot claim that a generated test collection is an ideal substitute for a real desktop environment with actual user queries, we tried to make the collection generation procedure as realistic as possible, and verify the validity of the resulting test collection for retrieval experiments by comparison to actual instances of desktops and user queries.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

2. GENERATING A PSEUDO-DESKTOP

2.1 Collecting Documents

As a first step, we need a collection of documents that has the characteristics of a typical desktop. The criteria that we used for the documents in a desktop were that the documents should be related to a particular person, there should be of a variety of document types, the different document types should have metadata or fields. The privacy of the target individual was another concern.

Given these conditions, our choice of a document collection method was to focus on people mentioned in the email collection from the TREC Enterprise track (crawl of the W3C website) and fetch a variety of publicly-available documents on the web related to those people. More details will be provided in Section 3.1.

2.2 Generating Known-Item Queries

Azzopardi et al. [1] suggested a set of methods for generating a known-item query in a multilingual web collection by algorithmically selecting a set of terms from a target document, based on a observation that an user may formulate query by taking whatever terms she can remember from the document.

However, since we assume that a user's querying behavior would be somewhat different in desktop search, we adapted their generation method by incorporating the selection of fields in the generation process, which results in the following algorithm:

1. Initialize an empty query $q = ()$ and select the query length s with probability $P_{length}(s)$
2. Select document d_i to be the known-item with probability $P_{doc}(d_i)$
3. Repeat s times:
   - 3-1. Select the field $f_j \in d_i$ with probability $P_{field}(f_j)$
   - 3-2. Select the term t_k from field language model of f_j $P_{term}(t_k|f_j)$ and add t_k to the query q
4. Record d_k and q to define a known-item/query pair

The only step added here is step 3.1, where we choose the field from which the query term is selected. We call this modification field-based generation method to contrast with document-based generation method suggested in previous work [1]. For P_{term}, we use random selection, TF-based selection, IDF-based selection and TF*IDF-based selection, as suggested in Azzopardi et al. [1].

Table 1: Number and average length of documents for each pseudo-desktop collection.

Type	Jack		Tom		Kate	
email	6067	(555)	6930	(558)	1669	(935)
html	953	(3554)	950	(3098)	957	(3995)
pdf	1025	(8024)	1008	(8699)	1004	(10278)
doc	938	(6394)	984	(7374)	940	(7828)
ppt	905	(1808)	911	(1801)	729	(1859)

2.3 Evaluating Equivalence to Manual Queries

Azzopardi et al. [1] introduced the notion of predictive and replicative validity to show that generated queries are equivalent to hand-built queries. Predictive validity means whether the data (e.g., query terms) produced by the model is similar to real queries, while replicative validity indicates the similarity in terms of the output (e.g., retrieval scores).

2.3.1 Verifying Predictive Validity

In verifying predictive validity, we need to evaluate how close the generated queries are to hand-built queries. While previous work [1] introduced only the idea of predictive validity, we suggest using the generation probability $P_{term}(Q)$ of the manual query Q with the term distribution P_{term} from the given query generation method, as follows:

$$P_{term}(Q) = \prod_{q_i \in Q} P_{term}(q_i) \quad (1)$$

For document-based query generation method [1], we can just use the simple maximum-likelihood estimates for each word. For the field-based query generation method, since every field has different P_{term}, we need to take the linear interpolation of P_{term} for all fields.

2.3.2 Verifying Replicative Validity

Azzopardi et al. [1] measured replicative validity by the two-sided Kolmogorov-Smirnov test (KS-test) using the score samples of real and generated queries as input. Since KS-test determines whether two samples are from the same distribution, we can conclude that two distributions are equivalent if resulting p-value is greater than a certain threshold.

3. EXPERIMENTS

3.1 Building a Pseudo-desktop Collection

As described in Section 2, we built each pseudo-desktop collection so that it may contain typical file types in desktop like *email*, webpage (*html*) and office document (*pdf*, *doc* and *ppt*) related to specific individuals. Table 1 lists the statistics from the resulting pseudo-desktop collections corresponding to three pseudo-users – "Jack", "Tom" and "Kate".

To get the emails related to a person, we filtered the W3C mailing list collection where the name occurrence of each person was tagged, which enabled us to identify several individuals whose activities in W3C were prominent. For other document types, using the Yahoo! search API with the combination of name, organization and speciality (provided by TREC expert search track) of each pseudo-user as query words, we collected up to 1,000 documents for each individual and document type, which roughly matches the statistics of previously used desktop collections [3].

3.2 Generated Queries

Table 2: P-values of Kolmogorov-Smirnov test for different query generation methods.

Extent	P_{term}	DLM	PRM-S	PRM-D
Document	Uniform	0.068	**0.417**	**0.129**
	TF	0.058	**0.619**	**0.244**
	IDF	0.000	**0.116**	0.003
	TF*IDF	0.000	**0.266**	0.007
Field	Uniform	**0.621**	**0.299**	**0.406**
	TF	**0.456**	**0.207**	**0.605**
	IDF	**0.110**	0.027	0.061
	TF*IDF	**0.227**	0.030	0.066

We generated queries using methods described in Section 2.2 and verified its predictive and replicative validity using three pseudo-desktops each with 50 queries written by three people for random sample of documents. For predictive validity, the field-based generation method showed higher generation probability (-13.7 in log scale) than the document-based generation method (-13.9 in log scale). We also verified the replicative validity using three retrieval models – document query likelihood (DLM), PRM-S [3] and the interpolation of DLM and PRM-S (PRM-D). The result in Table 2 confirms the replicative validity of field-based generation methods, especially when query-terms were selected randomly or based on term frequency. All document-based generation methods show replicative validity only for some of the retrieval models.

4. CONCLUSION

In this paper, we described a method for generating a reusable test collection for desktop search experiments and showed that pseudo-desktop collections generated with the field-based method are valid based on various criteria. For future work, we can refine the generation procedures using more sophisticated query generation models or scale the collection by adding more file types and metadata fields. We are also working on verifying the result in pseudo-desktops with the actual desktops.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-0707801, and NSF grant #IIS-0711348. Any opinions, findings and conclusions or recommendations expressed are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07*, pages 455–462, New York, NY, USA, 2007. ACM.

[2] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03*, pages 72–79, New York, NY, USA, 2003. ACM.

[3] J. Kim and W. B. Croft. *Retreival Experiments using Pseudo-Desktop Collections*. CIIR Technical Report. 2009.

[4] S. Shah, C. A. N. Soules, G. R. Ganger, and B. D. Noble. Using provenance to aid in personal file search. In *ATC'07: 2007 USENIX*, pages 1–14, Berkeley, CA, USA, 2007. USENIX Association.