A Framework for Test Topic Generation

Jiangping Chen[1], Min Namgoong[1], Gaohui Cao[2]
[1]Department of Library and Information Sciences, University of North Texas
[2]School of Information Management, Central China Normal University

Abstract
This study proposes a test topic generation framework through an analysis of existing literature. The framework contains three components, including a list of questions for eliciting users' information needs, a mechanism for topic generators to interact with the document collection, and a list of criteria to assess the quality of generated test topics. An application of this framework for generating test topics for a collection of library metadata records is presented.

Keywords: Information retrieval; test collection; test topics; topic generation
doi: 10.9776/16451

Acknowledgements: This study is supported by the Institute of Museum and Library Services (IMLS) under the National Leadership Grant LG-06-13-0187-13. The authors would like to thank Shravan Kumar Anugam for developing part of TGS and all participants.
Contact: Jiangping.Chen@unt.edu

1 Introduction

Information Retrieval (IR) systems are often evaluated using test collections. A test collection usually contains a document collection, a set of test topics, and relevance judgments associated with those topics. Test collections have played a crucial role in advancing IR research and practice. The most influential test collections were those developed by the organizers of the three major IR forums: TREC, CLEF, and NTCIR. Many of these collections are about news stories and web pages. However, existing test collections are not always sufficient to satisfy the different needs of IR research and evaluation. In particular, very few test collections have been developed containing library metadata records.

As one of the important components of a test collection, test topics are based on "user need" statements from which queries sent to an IR system can be derived (Harman, 1993). A number of methods for generating test topics have been investigated, as described below. However, guidance is needed for the IR community to choose effective and efficient topic generation approaches. Furthermore, the evaluation of generated test topics is an area that has not be systematically explored.

This study aims to investigate a framework that can guide test topic generation practices. Through a test topic generation exercise guided by the proposed framework, our hope is to provide guidelines for IR researchers and practitioners on test topic generation.

2 Related Literature

Test topics express users' information needs (Harman, 1993). The literature shows that there are various ways to generate test topics using human participants. TREC has used retired intelligence analysts to create test topics and develop relevance judgments for different IR tasks or tracks. NTCIR and CLEF have occasionally used third-party companies for test topic generation (Mitamura *et al*., 2010).

Kelly and Fu (2007) explored a technique that allowed users to express their information needs more fully. Three clarification questions were asked so that users could clarify their background knowledge about a topic/query, their reasons for inquiry, and additional keywords.

Lykke and others (2010) created a test collection involving physics-related library records, papers, and other objects. They used a questionnaire consisting of five questions to collect topics, including current user information needs, the user's background on the topic, the user's current knowledge state, expected answers, and possible search terms.

Oard *et al*. (2004) built a test collection for the retrieval of spontaneous conversational speech. They chose 70 representative requests from more than 250 collected from scholars, educators, and documentary filmmakers, and formulated them into TREC-type test topics, consisting of a title, a short description, and a narrative description.

Mitamura *et al*. (2010) used SEPIA (Standard Evaluation Package for Information Access) topic creation tools for topic generation. The topics were actually questions involving multiple tasks, including

IR and question answering. SEPIA consists of an interface for topic development, nugget (a question/topic answer) extraction, and nugget voting via a pyramid method. The interface includes a topic creation form and the Lemur Project's Indri Search Engine used by the topic developers to search for documents relevant to each topic.

The automatic method can be applied to generate test topics as well. Graf and Azzopardi (2008) proposed a methodology with eight steps for the construction of a patent test collection for prior art searches. Most of these steps can be performed automatically.

3 Test Topic Generation Framework

Our analysis of the limited literature on test topic generation found that the following principles are representative:

- Topics should reflect real user needs. It is desirable to recruit real users of the document collection as topic generators or developers. Also, techniques should be applied to allow generators to elicit their information needs;
- Most topics should have relevant documents in the document collection. Generators should be able to interact with the document collection; and
- The characteristics of the document collection should be considered. For example, a document collection of patents may have different document structures, styles, and content than one consisting of library metadata records.

Guided by the above principles, we propose a test topic generation framework that contains the following components:

- A set of questions to elicit the information needs of users,
- A mechanism that allows generators to interact with the document collection, and
- A list of criteria to assess the quality of topics individually and as a whole.

3.1 Questions to Elicit Information Needs

Based on the characteristics of the document collection, the researcher can develop a list of questions for the purposes of understanding the topic generators' information needs. These questions should help the generators to provide as much information as possible regarding what she/he wants to know about the topic. Questions presented by Kelly and Fu (2007) and Lykke and others (2010) are ones for possible consideration and use.

3.2 A Mechanism to Interact with the Document Collection

Because the test topics are to be generated for a specific document collection, it is important to make sure that the topics do have relevant documents from that collection. Therefore, a mechanism should be in place so that generators can check the topics against the document collection and develop their topics appropriately. In many cases, the IR system serves as the mechanism for the interaction.

3.3 Criteria for Quality Test Topics

A few studies have touched on how test topics should be selected once they are generated (Harman, 1993; Kando *et al*., 1999; Eguchi, Kuriyama, & Kando, 2002; Mandl & Womser-Hacker, 2004; Grubinger, Leung, and Clough, 2005; Mitamura *et al*., 2010). As a result, some IR evaluations at TREC, NTCIR, and CLEF contained failed topics that have to be excluded from the test collection for various reasons. In this framework, a quality test topic should meet the following criteria:

- Unambiguity. A test topic should be presented in clear natural language;
- No Duplication. A topic does not duplicate or overlap other topics; and
- Cultural Appropriateness. A topic is suitable for translation into other languages and not culturally unacceptable.

Furthermore, the final set of chosen topics should meet the following criteria:

- Diversity. Test topics should cover different subjects within the scope of the targeted document collection; each topic should be different from the other topics;

- Relevancy. Test topics should be associated with the document collection to varying degrees. Some should retrieve more relevant documents than others; and
- Complexity. Test topics need to be at different levels of difficulty so that the topic set can be effectively used to test IR systems.

Among the above six criteria, complexity is the most difficult to estimate. Whether a topic is difficult or not can be detected, in many cases, only after the test collection has been used and none of the participating IR systems do a good job on that topic. The other criteria are comparatively easy to assess by analyzing the individual topics.

4 Generating Test Topics for a Collection of Metadata Records

We have applied the above framework to develop a set of test topics for a document collection consisting of 1 million library catalog records. These records contain up to six metadata elements: title, creator, subject, description, publisher, and coverage, for print or digital objects, such as books, CDs, and DVDs. We used the following six questions to obtain test topics from a group of participants:

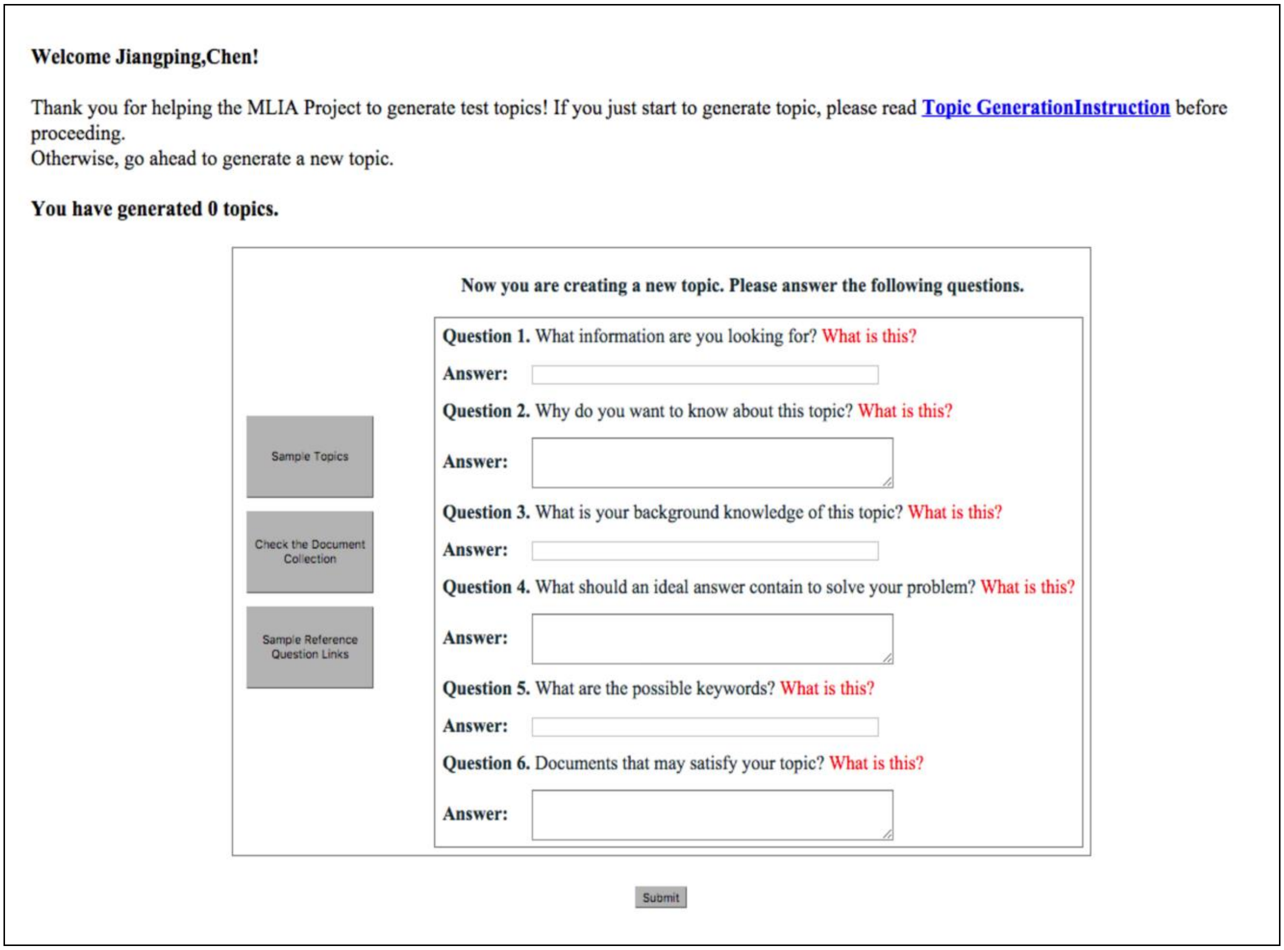
1) What information are you seeking?
2) Why do you want to know about this topic?
3) What is your background knowledge of this topic?
4) What should an ideal answer contain to solve your information need?
5) What are possible keywords?
6) What are 3-6 possible metadata records that may satisfy your topic?

To facilitate topic generation, we developed a system called TGS (Topic Generator System: http://txcdk-v10.unt.edu/TGS/). TGS is a public, web-based database system developed with open-source technologies. It contains the following functions:

- User Management. Interested participants can register on TGS by completing an online registration form providing topic-generation related information, such as educational background, major, how they use libraries, and contact information. They are also asked to provide a username and password for their login to use TGS, if approved. Later, the participants can modify their profile and password;
- Test Topic Generation. Once a participant logs in, TGS presents its main page with the six questions listed above, as well as related information and links. Figure 1 is a screen shot of the test topic generation page. It contains a welcome message at the top that includes a link to an instruction page for the participants, which describes the purpose of the study and the steps necessary to generate a topic. Below the welcome message, the main window presents the questions and a textbox for each question on the right, as well as three big square buttons on the left. The first button “Sample Topics” is a link to a web page providing three sample topics following the six question format; the second button “Check the Document Collection” links to a web portal that allows the participant to search the document collection to verify their topics and to answer the sixth question in topic generation; and the third button “Sample Reference Question Links” leads to a web page to further assist participants by providing three hyperlinks to Internet resources that list reference questions.
- Topic Verification and Editing. Once a topic is generated and submitted, TGS will again present the topic and ask the participant to review/edit.

We recruited eight generators, all of them college graduate students. Using TGS, participants generated 47 topics online in two weeks. These topics cover areas such as health, pet care, furniture repair, music, and history. Using the criteria in 3.3, the authors reviewed all topics and revised some of them for clarity. We were able to remove duplicate topics, as well as ones that didn’t retrieve any relevant documents.

Because TGS is a web-based system, it enables test topic generation via crowd sourcing – collecting test topics from web users. TGS has the potential to be used by other researchers and practitioners for generating test topics.



Welcome Jiangping,Chen!

Thank you for helping the MLIA Project to generate test topics! If you just start to generate topic, please read Topic GenerationInstruction before proceeding.
Otherwise, go ahead to generate a new topic.

You have generated 0 topics.

Now you are creating a new topic. Please answer the following questions.

Question 1. What information are you looking for? What is this?

Answer:

Question 2. Why do you want to know about this topic? What is this?

Answer:

Question 3. What is your background knowledge of this topic? What is this?

Answer:

Question 4. What should an ideal answer contain to solve your problem? What is this?

Answer:

Question 5. What are the possible keywords? What is this?

Answer:

Question 6. Documents that may satisfy your topic? What is this?

Answer:

Sample Topics
Check the Document Collection
Sample Reference Question Links
Submit

Figure 1. TGS Main Page

5 Future Work

This study is part of the test collection development, still in progress. We will continue to investigate other possible criteria for quality test topics. We will also explore more extensively how to score complexity of test topics, as well as how to develop automatic approaches to assess a test topic set for its targeted document collection. Also, we will review and revise the functions and content of TGS system. The TGS system and the test collection will be available for public use once the project is completed.

6 References

Chen, J., Knudson, R., & Namgoong, M. (2014). An investigation of effective and efficient multilingual information access to digital collections. In *Proceedings of the 77th Annual ASIS&T Conference*, Seattle, WA, Nov. 2-5.

Eguchi, K., Kuriyama, K., & Kando, N. (2002). Sensitivity of IR evaluation to topic difficulty. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Esmaili, K. S., Abolhassani, H., Neshati, M., Behrangi, E., Rostami, A., & Nasiri, M. (2007). Mahak: A test collection for evaluation for Farsi information retrieval systems. In *Proceedings of Computer Systems and Application*, 639-644.

Graf, E., & Azzopardi, L. (2008). A methodology for building a patent test collection for prior art search. In *The Second International Workshop on Evaluating Information Access (EVIA)*, 60-71. Retrieved from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/11-EVIA2008-GrafE.pdf.

Grubinger, M., Leung, C., & Clough, P. (2005). Towards a topic complexity measure for cross-language image retrieval. In *Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop*.

Harman, D. (1993). Overview of the first TREC conference. In *ACM SIGIR'93*.

Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., & Hidaka, S. (1999, September). Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 11-44.

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. *Information Processing and Management*, 43(1), 30-46.

Lykke, M., Larsen, B., Lund, H., & Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In C. Gurrin *et al*. (Eds.), Lecture Notes in Computer Science: Vol. 5993. *Advances in Information Retrieval*, 627-630. Milton Keynes, UK: Springer-Verlag.

Mandl, T., & Womser-Hacker, C. (2004). Linguistic and statistical analysis of the CLEF topics. In *Advances in Cross-Language Information Retrieval*, Vol. 2785, 505-501.

Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., *et al*. (2010). Overview of the NTCIR-8 ACLIA tasks: Advanced cross-lingual information access. In *Proceedings of NTCIR-8 Workshop Meeting*, June 15-18, Tokyo, Japan.

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., *et al*. (2004). Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 41-48. New York, NY: ACM Press. Retrieved from http://www.ece.umd.edu/~oard/pdf/sigir04.pdf.

Paramita, M. l., Sanderson, M., & Clough, P. (2009). Developing a test collection to support diversity analysis. In *SIGIR Workshop on Redundancy, Diversity, and Interdependent Document Relevance*. Retrieved from http://ir.shef.ac.uk/cloughie/papers/idr09.pdf.