# A Study of Self-Organizing Map in Interactive Relevance Feedback

Daqing He
School of Information Sciences
University of Pittsburgh
135 N. Bellefield Avenue, Pittsburgh, PA 15260
email: daqing@mail.sis.pitt.edu

## Abstract

*With the vast amount of potential relevant documents on the Web, a key question for a retrieval system is how to achieve a high accuracy retrieval under current Web setting. The work reported in this paper is our latest effort in examining the effectiveness of applying Self Organizing Maps (SOM) in clustering base interactive relevance feedback. Our experiment results demonstrate that SOM can be used as a clustering tool to generate terms for interactive relevance feedback, which produced a significant improvement over the baseline when measured by R-Precision. Our results also show that a preferred strategy for combining blind relevance feedback with interactive relevance feedback is to perform the former first then the latter.*

**Keywords**: *Self-Organizing Map, Interactive Relevance Feedback, Blind Relevance Feedback, High Accuracy Retrieval from Documents, TREC*

## 1 Introduction

With the advance of Web technology, searching for information is increasingly common in people's life. Batch mode information retrieval (IR), which essentially studies retrieval algorithms, receives a great deal of attention. Studies show that modern techniques based on "free text" indexing and ranked retrieval have proven to be scalable and robust into Web domain [4]. However, Web imposes some interesting requirements to the retrieval systems. With the vast amount of potential irrelevant documents on the Web, most time, the key for the success of a retrieval system is whether it can put truly relevant documents at the top of the search results. In other word, can the retrieval system achieve high precision in its search results?

We believe that the key of such success lies on the human users. This is because it is the users who pose the questions, interpret what they read, and determine when their needs have been met. However, it is widely accepted that interactive IR experiments are difficult to design, expensive to conduct, limited in its small scales, and hard to perform cross-site comparisons. Our past experience with interactive IR, especially experiments we conducted for interactive track of Cross Language Evaluation Forum (iCLEF) [7, 6], provides us some first-hand experience to these limitations.

We see the High Accuracy Retrieval from Documents (HARD), a track in Text REtrieval Conference (TREC), as a good framework for trying to address the problem of studying interaction between human users and retrieval systems, but at the same time keeping the TREC tradition of examining retrieval algorithms and achieving cross-site comparison [1]. For better representation of the actual retrieval process, HARD allows interactions between the users and the retrieval system. However, to avoid the difficulty of managing full interactions, HARD only allows one iteration of interaction, which is in the form of letting users to answer a set of clarification questions generated by retrieval systems. The most common types of clarification questions asked in the HARD track have been eliciting relevance feedback information on documents, clusters of documents, or terms inside documents/document clusters [1]. The system then uses the answers as the extra information to generate the final search results, which is a rank list of documents. The hope is that the search effectiveness would have improved via the interaction.

There are two types of relevance feedbacks. When a relevance feedback involves a human user in the relevance judgments, it is called interactive relevance feedback (IRF). The second form of relevance feedback does not involve human users. It uses the top N documents from the search result, which are assumed to be relevant, for relevance feedback. This is called blind relevance feedback (BRF).

Both IRF and BRF are effective techniques for improving retrieval results. The quality of IRF depends on the quality of the user!/s judgments, which are related to the knowledge of the user on the topic, the information presented in the judgment, the way the information is presented, and the amount of judgments. The quality of BRF depends on the quality of the top N documents, which could contain too much noise when there is no many truly relevant documents in the top N set.

One strategy developed in IRF for reducing the number of judgments required is to ask users to judge document clusters instead of individual documents. The assumption is that the users are interested in certain topics, which can be represented by clusters of documents, and any document in the user identified clusters would be relevant to the topic. The usage of clustering techniques in relevance feedback has been explored previously for both interactive relevance feedback [13, 5] and blind relevance feedback [3]. All these studies indicate that clustering approach needs further investigation.

In this paper, we will talk about one study we conducted within the HARD framework, where we examined the effectiveness of using Self Organizing Maps (SOM) as a clustering tool for grouping search results for IRF. We also studied the combination of IRF with BRF to further improve the precision of the search results.

In the reminder of this paper, we will first introduce our method of applying Self-Organizing Map (SOM) in interactive relevance feedback, which include a detail presentation of how such SOM was generated. Then we move on to talk about the experiment procedure, and the discussions of the experiment results. Finally, we will conclude with some discussion of further studies.

## 2 Applying SOMs in IRF

Self Organizing Map (SOM) is a technique invented by Professor Teuvo Kohonen. It reduces the dimensions of data through the use of self-organizing neural networks [8]. The way that SOM reduces high data dimensions is by producing a map of usually two dimensions, which plots data items by their similarities of each other.

The content of documents, no matter looking at them via term-based or concept-based representation, can be viewed as a function of all or most important terms/concepts in them, which clearly is in high dimensional space. This is the basis for some document representation models, say vector space model, to represent a document into a high dimensional vector [15]. SOM takes the same representation for documents, and when it is built for a collection of documents, it will reduce the high dimensional representation of documents into a two dimensional map. Through this process, documents with similar terms/concepts, thus similar content, would be grouped into either the same cells or adjacent cells in the SOM. This is the motivation for SOM to be applied in information retrieval for visualization and clustering purposes [10, 11].

In our study, instead of being used as a visualization tool, SOM was applied for grouping documents into clusters based on their content similarity, and generating the cluster representatives in the form of most representative terms so that the representatives can be used in the relevance judgment process.

To construct a SOM for each search topic, we first conduct a search based on the query generated from the topic statement. In TREC, a topic statement contains a title, which is a few representative keywords, a description, which is a natural language sentence describing the important aspect of the search topic, and a narrative, which is a detail discussion of the major relevant criteria. Figure 1 shows an example of HARD topic in TREC format. Top 400 documents were selected from the retrieval result, based on which, we extracted a 1000 word vocabulary for generating the vector representation of each of the 400 documents. The weights of the terms were calculated using BM25 formula [14].

Equipped with the document vectors, we used SOM_PAK version 3.1 to generate a SOM for each topic. SOM_PAK was developed by the SOM Programming Team of the Helsinki University of Technology Laboratory of Computer and Information Science [9]. We set the size of a SOM as 20 X 20 cells. The topology of the map was set to be rectangular, and the neighborhood function was Gaussian.

At the initiation stage, a randomly generated vector that shares the same 1000 word vocabulary was assigned to each cell. In this phase, the initial neighborhood radius was set at 10 and decreased to one. The initial learning rate was 0.05 and it also decreased to zero. Then at the fine tune stage, the initial neighborhood radius was set at 3 and the initial learning rate was 0.02. All 400 documents were fed sequentially into the map. The similarity between the vector of the document and that of each cell was calculated to identify the most similar cell to add the document. The based on the neighborhood radius and learning rate, the effect of adding that document into the cell is propagated to the surrounding cells. This process is repeated until the preset iteration number is reached. At that time, we have a SOM for the topic. Figure 2 shows a SOM generated for HARD Topic 325. Here for presentation purpose, the SOM is displayed as a 20 X 20 map, in which the color of the cell indicates the similarity between the cell and its surrounding cells. The lighter the color is, the more similar the cells are. The set of terms displayed in the figure are the representative terms for a cell.

Although SOM can group similar documents together, different to traditional clustering tools, it does not have a real sense of clusters of documents. A SOM only has cells, in which contains similar documents. We know that cells that are adjacent to each other are more similar than cells that are apart, but it could not tell whether two adjacent cells should belong to the same cluster or not. For the task

```
Topic Number: 325
<Title> Cult Lifestyles
<Description> Describe a cult by name and identify the cult members' activities
      in their everyday life.
<Narrative> A relevant document would include the name of the cult and offer
information about the members' lifestyles.  It may include how they dress or
what they do to attain the ultimate goal of the organization.  A relevant
document may tell what they eat or how they contribute to the cult. Just the
mention of the existence of a cult by name with no other clarifying information
would not be relevant.
```

**Figure 1. An example of HARD topics**

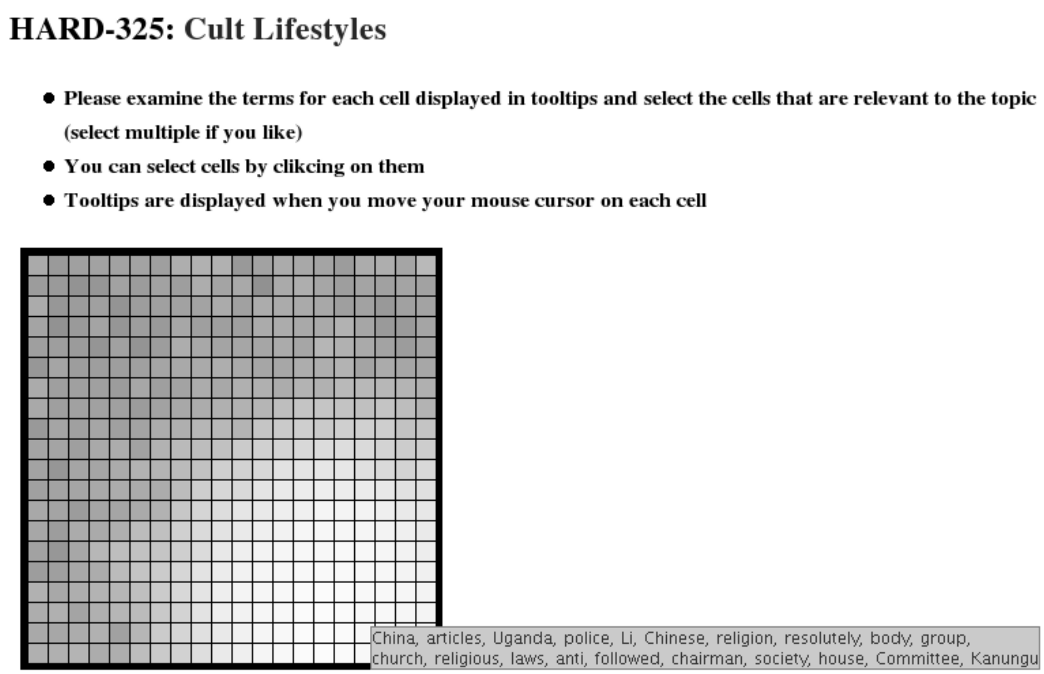


**Figure 2. The SOM generated for HARD topic 325 (The SOM is presented in graphical mode)**

of clustering-based interactive relevance feedback, we need to construct the clusters of documents based on the cells in the SOM. We achieved this via grouping cells using single link clustering method. The similarity between cells was calculated based on cosine, and the threshold was set at 0.85. Based on the grouping results, we then pick up the six largest clusters of cells for generating terms for IRF. The rationale for choosing large clusters was because we assume that the cells containing some relevant documents would probably be similar to each other, whereas cells contain irrelevant documents would contain rather random content. Therefore, the bigger the size of a cluster, the higher chance that it contains relevant documents. Top 20 weighed terms in each selected cluster of cells were extracted as the content of the clarification questions for IRF. Figure 3 shows an example of the IRF questions for a search topic.

## 3 Experiment

### 3.1 Retrieval system and Measures

We used Indri 2.0 as our retrieval system. Indri is a state-of-art retrieval system developed by University of Massachusetts Amherst and Carnegie Mellon University (http://www.lemurproject.org/indri/). It has similar rich query context as Inquery system that was also developed by University of Massachusetts Amherst.

We used R-Precision (R-Prec) as the measure. It is the precision at top R document where R stands for the number of relevant documents for a topic. This measure puts more

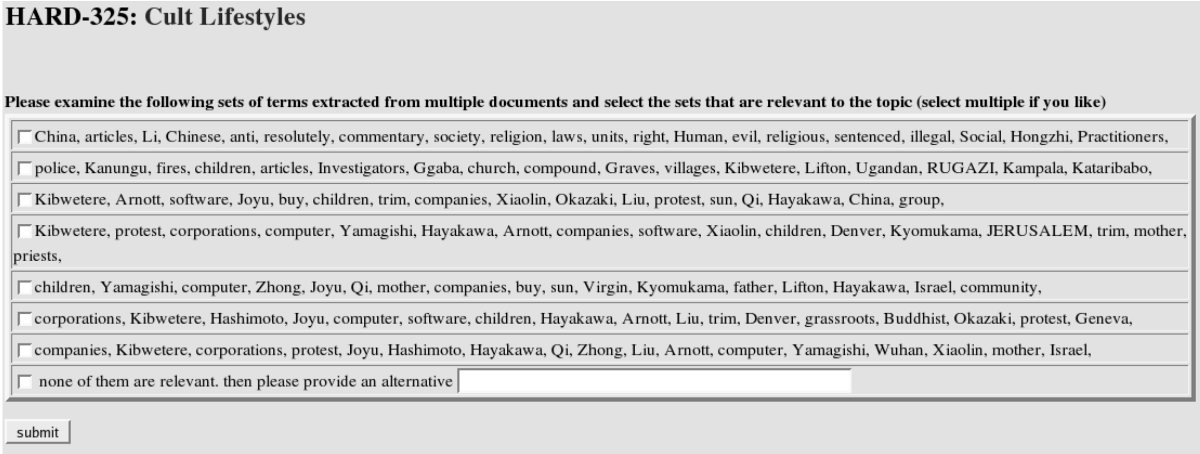


**Figure 3. A set of IRF questions for a HARD topic**

emphasis on the precision, but also avoids the limitation of a randomly selected N in the measure Precision at N.

### 3.2 The Initial training study

We conducted the initial training study on 50 HARD03 topics on AQUAINT collection, which contains 320380 documents. The purpose of this study is to establish the optimal parameter settings for the retrieval runs. In both the training and later evaluation experiments, the search queries were extracted from the title and the description part of the HARD topics (see Figure 1 for an example) for retrieving documents. This was due to the fact that we always obtained the best retrieval effectiveness using these two parts.

We used the BRF mechanism implemented in Indir system for our study of BRF. As shown in Table 1, BRF with the right parameters had achieved greatly improvement on R-Prec over the runs without (i.e., HDTRAN-PLAIN). Therefore, we used the BRF setting of HDTRAN-BRF2 for generating the initial search run HDEVAL, and used HDEVAL for generating the SOM for the questions for IRF.

### 3.3 The interactive relevance feedback in HARD

Once the questions for IRF were generated, they were sent to the human assessors to provide feedbacks. The assessors were recruited by NIST, and they had gone through extensive training on judging the relevance of documents. To make sure the relevance judgments are in a manageable scale, each assessor was asked to spend no more than three minutes per set of questions related to one topic. Assessors had the topic statements with them during the feedback time just in case they want to consult it. Because they were providing feedbacks to all the participant sites to the HARD track, the questions from each site are arranged in round robin style to minimize the learning and fatigue effect. After the feedback session, the feedback results were sent back to us.

### 3.4 Incorporating interactive feedback results

For incorporating IRF results, we used term based query expansion. To differentiate the original query from the expanded terms, we allocated 0.75 relative weight to the former, and 0.25 to the latter to reflect that believe that we trust more on the original query rather than expanded terms.

## 4 Results and Discussion

The final evaluation results are presented in Table 2. Overall, IRF base on our SOM generated positive results. The corresponding run HDEVAL-CF2NOB achieved 14% relative increase in R-Prec over the baseline HDEVAL, which was the run before IRF . The increase is statistical significant with the p value as 0.01 in the Wilcoxon Matched-Pairs Signed-Ranks Test. As stated, the baseline HDEVAL actually has blind relevance feedback. Therefore, the fact that HDEVAL-CF2NOB achieved significant improvement over HDEVAL demonstrates that BRF and IRF can be combined in producing further improvement.

| Runs | R-Prec |
|---|---|
| HDEVAL | 0.2903 |
| HDEVAL-CF2NOB | **0.3296** |
| HDEVAL-CF2B225 | **0.326** |

**Table 2. The HARD evaluation results**

| Runs | selected docs | selected terms | weight | R-Prec |
|---|---|---|---|---|
| HDTRAN-PLAIN | | | | 0.3818 |
| HDTRAN-BRF1 | 20 | 20 | 0.3 | 0.3919 |
| **HDTRAN-BRF2** | **20** | **20** | **0.5** | **0.4238** |
| HDTRAN-BRF3 | 20 | 20 | 0.8 | 0.4190 |
| HDTRAN-BRF4 | 15 | 20 | 0.5 | 0.4200 |
| HDTRAN-BRF5 | 25 | 20 | 0.5 | 0.4235 |
| HDTRAN-BRF6 | 20 | 15 | 0.5 | 0.4191 |

**Table 1. The training on 50 HARD topics. The parameters of the run in bold font was selected as the parameters for the evaluation baseline HDEVAL for generating clarification forms.**

One interesting question to ask here is whether another iteration of BRF still makes sense after the query expansion using terms from IRF. This was the purpose of the run HDEVAL-CF2B225. Again taking the advantage of the BRF mechanism implemented in Indri search engine, we set the BRF parameters as extracting 20 terms from top 20 documents and use 0.5 as the relative weight allocation between the query and new BRF terms (that is the two have the same weight). It generated 1% relative decrease measured by R-Prec. The decrease is not statistical significant. Therefore, there seems to be not much gain to perform BRF after IRF. It seems that a better sequence for combining BRF with IRF is applying BRF first, then IRF.

## 5 Related work

Applying Self-Organizing Map in information retrieval is not a novel idea. Because it reduce high dimensional information into two dimension map, its usage for visualizing the similarity among documents to support a user in browsing through a document collection has been studied extensively in the literature [12, 11, 2]. However, there has not been a study of SOM in clustering based interactive relevance feedback.

## 6 Conclusions

In this paper, we talk about the studies of applying Self-Organizing Map in Interactive Relevance Feedback, and the effect of combining blind relevance feedback with interactive relevance feedback. We conducted the studies in the framework of High Accuracy Retrieval from Documents, a track of TREC. Our experiment results demonstrate that SOM can be used as a clustering tool to generate terms for user's interactive relevance feedback. It produced significant improvement over the baseline when measured by R-Prec. Our studies also show that a preferred sequence of combining IRF with BRF is probably to perform BRF first then IRF.

Our further work includes studying SOM as a visualization tool in interactive relevance feedback, and exploring the data fusion approach of combining BRF and IRF for improving retrieval results.

## Acknowledgements

Thank Jae-wook Ahn for his help in generating SOMs for HARD topics. This work is supported by School of Information Sciences Dean's Research Fund.

## References

[1] James Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. In *The Twelfth Text Retrieval Conference*, 2003. to appear.

[2] Peter Brusilovsky, Rosta Farzan, and Jae-wook Ahn. Comprehensive personalized information access in an educational digital library. In *Proceedings for Joint Conference on Digital Libraries*, pages 9–18, 2005.

[3] Chris Buckley, Mandar Mitra, Janet A. Walz, and Claire Cardie. Using clustering and superconcepts within SMART: TREC 6. *Information Processing and Management*, 36(1):109–131, 2000.

[4] N. Craswell and D. Hawking. Overview of the trec 2004 web track. In *Proceedings of TREC 2004*, 2004.

[5] Daqing He and Dina Demner-Fushman. HARD Experiment at Maryland: From Need Negotiation to Automated HARD Process. In *Proceedings of Text REtrival Conference (TREC) 2003*, 2003.

[6] Daqing He, Jiangqiang Wang, Jun Luo, and Douglas W. Oard Oard. Summarization design for interactive

cross-language question answering. In *Proceedings of Cross-Language Evaluation Forum*, 2004.

[7] Daqing He, Jianqiang Wang, Douglas W. Oard, and Michael Nossal. Comparing user-assisted and Automatic Query Translation. In *Proceedings of CLEF'02*, 2002.

[8] Teuvo Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2000.

[9] Teuvo Kohonen, Jussi Hynninen, Jari Kangas, and Jorma Laaksonen. SOM_PAK: The Self-Organizing Map Program Package. Technical Report Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[10] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In E. Simoudis, J. Han, and U Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 238–243, Menlo Park, California, 1996. AAAI Press.

[11] Xia Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48:40–54, 1997.

[12] Xia Lin, Dagobert Soergel, and Gary Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 262–269, 1991. information retreival-visualization-SOM.

[13] Anick P. and Vaithyanathan S. Exploiting clustering and phrases for context-based information retrieval. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 314 – 323, Philadelphia, PA USA, 1997.

[14] S. Robertson, S. Walker, and K. Sparck Jones. Okapi at trec-3. In *Proceedings of TREC-3*, 1995.

[15] G Salton. *Automatic Text Processing*. Addison-Wesley, Reading, MA, 1989.