

# Domain-Specific Information Retrieval Using Recommenders

Wei Li

Centre for Next Generation Localization  
School of Computing, Dublin City University  
Dublin 9, Ireland

wli@computing.dcu.ie

## ABSTRACT

The ever increasing volume of information available in our daily lives is creating ever greater challenges for people to find personally useful information. One approach used to address this problem is Personalized Information Retrieval (PIR). PIR systems collect a user's personal information from both implicit and explicit sources to build a user profile with the objective of giving retrieval results which better meet individual user's information needs than a standard Information Retrieval (IR) system. However, in many situations there may be no opportunity to learn about the specific interests of a user and build a personal model when this user is querying on a new topic, e.g. users visit a museum or exhibition which are unrelated to their normal interests. Under this condition, the experiences and behaviours of other previous users, who have made similar queries, could be used to build a model of user behavior in this domain. My PhD proposes to focus on the development of a new and innovative method of domain-specific IR model. My work seeks to combine recommender algorithms trained using previous search behaviours from different searchers with a standard ranked IR method to form a domain-specific IR model to improve the search effectiveness for a user entering a query without personal prior search history on this topic. The challenges for my work are: 1) how to provide users better results; 2) how to evaluate the results according to a user's interests in this specific domain; 3) in this particular domain, what data collection can be used to do experiment.

Since there are no suitable data collections available to enable us to explore our proposal, simulation plays a key role in this research. We simulate user behaviour in the physical environment, in order to make the collected data more realistic. We have conducted an initial experiment at study using the following steps:

- INEX 2009 Wikipedia documents collection [2] was used as our data collection, and 20 topics were chosen from INEX 2009 topic dataset;
- One or two words were randomly deleted from the original topic, 10 variations were made for each source topic;
- Use extended SMART retrieval system [3] to retrieve ranking for each variation topic, compare top 20 documents on this ranking against qrel file to identify true relevance documents on top 20 for each variation;
- Build a perfect scenario, generate rating value to each document: assigned 1 for true relevant documents obtained from last step and 0 for other documents;
- The true relevant documents and other documents are used together to trained visiting path information for each

variant, and 20 variants data for each original topic are integrated to train its corresponding recommender;

- For each new query, a browsing set can be obtained from extended SMART system, this set can be seen as a vector, every recommender also can be seen as a vector, compare the similarity between these two vectors, choose the most similar recommender for this query;
- Exploit the previous users information in this recommender to weighted slope-one collaborative algorithm [1] to predict ratings for every documents;
- For each document  $j$ , its final weight  $FW_j$  is calculated by linear combined its rating in initial ranking list from extended SMART system  $RW_j$  and rating in the prediction list from recommender algorithm  $RW_j$ .

$$FW_j = \alpha \cdot PW_j + (1 - \alpha) \cdot RW_j \quad [1]$$

The results ranking of each original topic retrieved by extended SMART system are used as our baseline. The MAP our experiment is increase 58% from baseline. The results show that the utilizing recommender algorithm in IR process can make the standard IR more efficient. My current work is focusing on developing a more realistic simulation model of the training environment, exploring the use of relevance feedback, automatically identifying topical domain and clustering topics for training recommenders, refining the results by learning from relevance feedback. Further work is identifying more suitable evaluation metrics, and exploiting other recommender algorithms to make this approach more effectively.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information Search and Retrieval, Information Filtering

**General Terms:** Recommender Algorithm, Simulation, Experimentation

**Keywords:** Domain-Specific IR, Recommender

## 1. REFERENCES

- [1] Lemire, D. and Maclachlan, A. *Slope One Predictors for Online Rating-Based Collaborative Filtering*. In Proceedings of SIAM Data Mining (SDM), 2005
- [2] Geva, S., Kamps, J., Lethonen, M., Schenkel, R., Thom, J.A., and Trotman, A. *Overview of the INEX 2009 Ad Hoc Track*, 2009
- [3] Debasis Ganguly, *Implementing Language Modeling in SMART*, Indian Statistical Institute, Calcutta, India, July 2008