ICTNET at Session Track TREC 2011

Mingxhuan Wei[1,2] ,Yuanhai Xue[1,2], Chen Xu[1,2], Xiaoming Yu[1], Yue Liu[1], Xueqi Cheng[1]

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190

2. Graduate School of Chinese Academy of Sciences, Beijing, 100190

1.Introduction

Following several methods of some past publications, among these we refined an important idea that related queries could be derived from queries submitted within the same session. As the browsers not just provide the previous queries but more informations,so we could use these informations to build a expert database for a special user, base on the database ,we could analysis some fixed behaviors of the user, so then forecast what information he really wants, rerank the results from search engine and give them to the user finally. The whole process is the session trec works on:Providing the really information to different user, so we can make the search engine more efficient and smarter.

The rest of paper is structured as follows. In section 2 we discuss the ideal of session type. In section 3 we descripte the classification we used . In section 4 we explain the experiment and result we submitted ,Finally we make a brief conclusion and a plan for the future work.

2.Session Type

According our expreience,when using the web search engine to search our konwledge ,we may not always find the most suitable result at once, especially for some unkonwn problems ,we even cannot use the right query to describe it.At the time ,we may firstly describe it using some words to from a query as closely as we thought,then we scan the results provided by search engine, during this time ,according to the information provides by the webpages we scaned, we may know some new konwledge and then use it to organize a query to search engine, and then repeat the previous behavior until we find the satisfied answer.Based the previous work,we could know the current query and the previous ones could form some session types,and when modifing query is adjusting different types to get better result. In general,The session types have three as follows[2]:

1.**Generalisation.** **2. Specication.** **3. Drifting/Parallel Reformulation.**

The three types are all have some a ssociations in query content or meanings,but in real work we can find the adjacent two query contain different purpose totally,they belong to two different search respectively,so we add a type to **Separation.** In this case,two search have no relationship (or concern) with each other.

3. Classification

For each previous query,the offical data have provide the top ten or even more search results from Yahoo after manual handling. For each final result,we could provide up to 2000 items for each query,So how could we distinguish them not just by the search engine.In our experiment,we use the classification to do this.We consider the top result as different category, then extract and purify keywords from their sinnpet content,The may be several specific methods ,and we use tf-idf to achieve the goal,and then make a vsm model for each category.

For the other result, rpeat the former behavior to make a vsm model,the handle content may be the snippet from the search engine,or the anchor text,this depends on the information the search engine could provide .Then there also may different methods to cassification.We use two methods as follows:

1.Match the keywords each keyword has its own ranking and weight after tf-idf,so after match,we can calculate a result for each category,get the max result.So we can know which category each result belongs to.

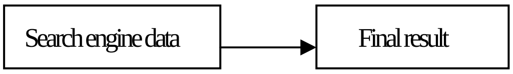
2.Calculate the similarity between result and the category,we use the cos-distance as the evalution parameter. Consider the two text as two vector ,then calculate the cos result of the vectors.Sort the cos-distances,get the max result.

The former method is only suitable for RL3 result,as for RL4,we should add the query's narration and describe into the analyse information.So for a item,it can get two classification results, mixed them to form a data,sort it and get the final category data.

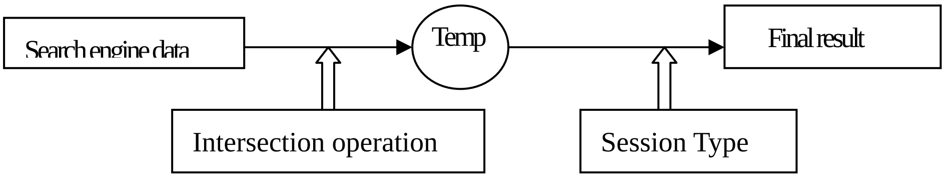
For RL4 xml file, we should consider the clicked items, when surfing on the website,we may clicked some pages but donot spend too much time to scan it as we thind it worths little.So we simulate the same progress,fit a timer threshold to distinguish the user's willing. For each clicked item,we simply consider it important or no important accordint the clicked gap time,when the gap time is greater than the threshold, we make it important and promote the category's score; on the contrary, make it no important and reduce the category's score.All above is our group's program to deal with the session task.

4.Experiment flow charts

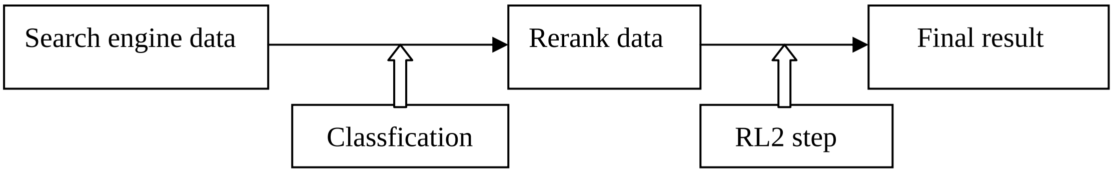
As we know ,the whole procedures should generate four different results in proper order for RL1 to RL4.So next we give a brief explaintion for the four results.RL1: Because this file only contains the final querys,so it neednot any optimazition,the result is provided by our own search engine. The flow chart shows as follow:



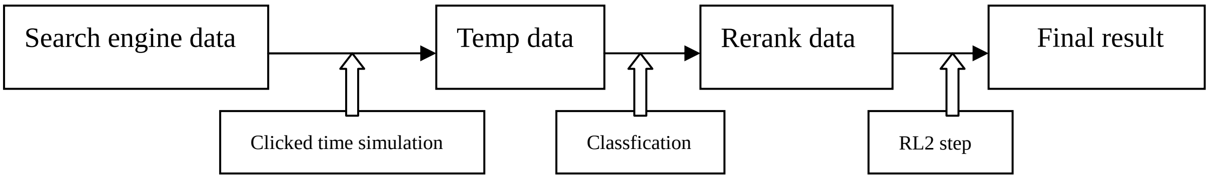
RL2: The file contains each query and previous query lists,so we could make a intersection operation to get the intersection ,and then according the session type to rerank the result.The flow chart shows as follow:



RL3: Comparing with the RL2 file,We use the classification to classify the data to rerank one time, then repeat the RL2 step. The flow chart shows as follow:



RL4:Comparing with the RL3 file,we should add the clicked time simulation ,then the classification parameter should add the query describe and narration. The flow chart shows as follow:



5.Conclusion

This year session trec provides a good platform to participants to prove their search ideals,meanwhile,the task is Closely linked the hot research current: intelligent search, so related researchers can communicate with each other by this platform.

Our group have make the promising results,First,it follows the important ideal of session type,second,it bring up the ideal of classfication,mixed the two ideals,we indeedly could get some better results, but it also have a bit drawbacks,as the classfication is two absolute,rely on the first search data given by search engine excessively.

So for the future,we should make a more clear program to optimize the result.

Acknlowledgement：

Our work was supported by 863 Program of China Grants 2010AA012502 ,2010AA012503 , NSF of China Grants No. 60903139 ,No.60873243.

References

1. TREC 2011 Session Track http://ir.cis.udel.edu/sessions/
2. Autoadapt at the Session track in TREC 2010
   Author: M-Dyaa Albakour, Udo Kruschwitz, Jinzhong Niu, Maria Fasli
   School of Computer Science and Electronic Engineering, University of Essex,
   Wivenhoe Park, Colchester, CO4 3SQ, UK
3. Session Track at TREC 2010
   Author: Evangelos Kanoulas Paul Clough Ben Carterette Mark Sanderson
4. The University of Amsterdam at TREC 2010
5. University of Lugano at TREC 2010
6. Webis at the TREC 2010 Sessions Track