ICTNET at Context Suggestion Track TREC 2013

Bingyang Liu[1,2], Yanqin Zhong[1,2], Yue Liu[1], Dayong Wu[1],Xueqi Cheng[2]

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

2. University of Chinese Academy of Sciences, Beijing, 100190

{Liubingyang , zhongyanqin}@software.ict.ac.cn

{liuyue,wudayong, cxq}@ict.ac.cn

Abstract

This year we did not use any external structured resources like 'Yelp'. An "Information Retrieval" scheme is implemented. We built index of the ClueWeb12-B-13 using Lucene and use manually and automatically constructed queries to fetch pages from the data subset. Finally we ranked the pages based on user preferences.

1 Data Preparation

As we have datasets to choose this year, we try to extract the suggestion pages from the datasets themselves not external structured data. It is a challenge to computation and storage using the full Clueweb 2012 dataset, so we choose the ClueWeb12-B-13instead. Lucene is used to build index.

2 Query Generation

Queries are the basics of our scheme. We want to build good queries for the search engine to return us with good suggestions.

Each query consists of base part and keyword part.

The base part is geographical informationmanually chosen providing the longitude and latitude. In the base part, we give the name of the state, city, block and street.

The keyword part is automatically generated by user preference. We extract all the nouns in their preferred pages and ranking these words using tf, idf and PMI. Only the top 25 words with their weight are used for each profile. The difference between RUN1 and RUN2 is: RUN2 also uses the disliked pages and calculates minus weight for each word.

3 Search and Combine

Now we get a query base part and 25 keyword parts with weights for each profile. We use 25 queries for each profile to fetch 25 batches of pages. Each of the pages is returned with a score from Lucene. We need to combine/re-rank the 25 batches of pages.

A simple model is set:

$$\text{New Score(page)} = \sum \text{Influence(query)} * \text{Weight(query)} * \text{Score(page)}$$

Which means multiply the weight of a query and the weight of a page returned by this query and plus them all together. The influence of each query should be learned but we use a constant here in this year. After calculating the new score of each page in each batch, we re-order all the pages and use the top 50 as our final result.

4 Result Analysis

One of the shortages of this simple model is obvious: pages that contain sets of interesting places can accumulate very high score, such as restaurant list pages. We need to add some penalty to this model to filter out these very unspecific pages.

The P@5 results confirmed this shortage: we get very low P@5 score for nearly all the profiles.

```
Run     profile     context     score  metric
RUN1    290         71          0.0000 P_5
RUN1    372         66          0.0000 P_5
RUN1    535         66          0.0000 P_5
RUN1    471         95          0.2000 P_5
RUN1    496         61          0.0000 P_5
...
```

We manually checked the pages after the results were published and found that the pages with lists werethe major problems. We should provide pages that are specific to one interesting place instead of pages consist of lists of places. Data clean process (identify which pages are lists and which pages are specific ones) should be carried out before we build the search index.

5 Acknowledgements

We would like to thank all organizers and assessors of TREC and NIST. This work is sponsored by NSF of China Grants No. 61202058 , and by 242 Program of China Grants No. 2013G130, and by the National Key Technology R&D Program (2012BAH39B04).

6 References

[1] https://sites.google.com/site/treccontext/

[2] Dean-Hall, Adriel, et al. "Overview of the trec 2012 contextual suggestion track." Proceedings of TREC.Vol. 12. 2012.

[3] Sappelli, Maya, Suzan Verberne, and Wessel Kraaij. "TNO and RUN at the TREC 2012 Contextual Suggestion Track: Recommending personalized touristic sights using Google Places." 21st Text REtrieval Conference Notebook Proceedings (TREC 2012). 2013.