PRIS at TREC2012 Contextual Suggestion Track

Lin Qiu, JunRui Peng, QianQian Wang, Yue Liu, ZhiHua Zhou
Weiran Xu, Guang Chen, Jun Guo
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications
Beijing, P.R. China, 100876
buptly@yahoo.com.cn

Abstract

The system to Contextual Suggestion Track at TREC2012 includes information crawling and preprocessing, context filtering, user modeling, similarity computing and ranking, description generating. Some third party tool kits are used, such as URLPARSE. TF-IDF (term frequency–inverse document frequency) and cosine similarity is also used for building user models and computed similarities between users and candidate items.

1. Introduction

The Contextual Suggestion Track investigates search techniques for complex information needs that are highly dependent on context and user interests. This year is the first year of the Contextual Suggestion Track, which aims at addressing a recommend task whereby not only user's bias but also contextual information is take account of. There're thirty-four users and fifty groups of contextual information located on 36 cities. And fifty suggestions should be returned for each combination of a user and a context. That means 85,000 suggestions should be generated. The difficulty is how to get enough candidate items because of limitation of time. However, it is not the unique difficulty. User profiles and context profiles provided too much information needing to processing and modeling. And finally, different description must be generated according to different user. As a result, a rounded system is needed.

2. System Structure

The system is a rounded system which includes four parts of data acquisition, data processing, data analysis and result generating.

2.1 Information Crawling

Candidate items were crawling from open web because the new ClueWeb12 corpus was not ready on time.

A spider framework is designed to crawl with information about candidate items, such as attractions or restaurants etc., from some specific website (for example, www.tripadvisor.com). The framework mainly made use of these third-party libraries including PYCURL, URLPARSE, and URLLIB.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
PRIS at TREC2012 Contextual Suggestion Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Beijing University of Posts and Telecommunications,School of Information and Communication Engineering,Beijing, P.R. China, 100876,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT
The system to Contextual Suggestion Track at TREC2012 includes information crawling and preprocessing, context filtering, user modeling, similarity computing and ranking, description generating. Some third party tool kits are used, such as URLPARSE. TF-IDF (term frequency?inverse document frequency) and cosine similarity is also used for building user models and computed similarities between users and candidate items.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

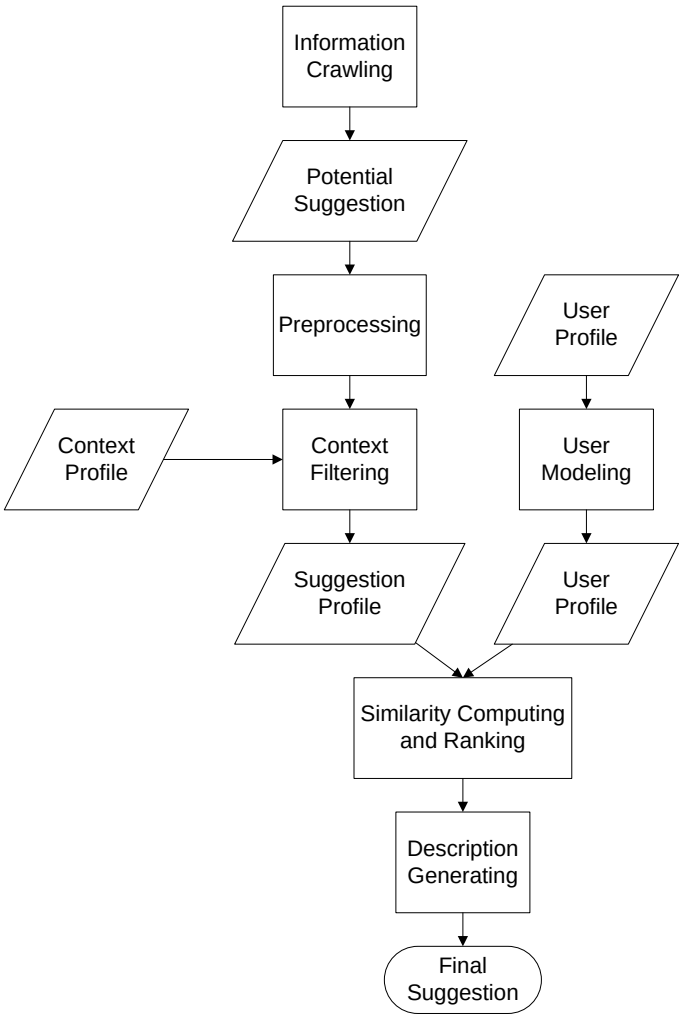

Figure 1. The structure of contextual suggesting system

Not all items listed on the website should be downloaded because of the limitation of contextual information. Contextual profile provides 50 groups of contextual information located on 36 cities. The requirement “the user has up to five hours available to follow a suggestion and has access to appropriate transportation (e.g., a car)” makes it impossible to choose some attractions far away. As a result, attractions should be chosen firstly in or nearby some city. Meanwhile, homepages provided by the website also must be validation. And there’re some small towns where candidate items were not enough, so manual method for searching for more suggestions from nearby towns is necessary.

2.2 Preprocessing

A candidate item is downloaded means web pages related to the suggestion are downloaded. Useful information, including name, homepage, rate and comment, should be separated from web pages by regular expression. Comments represent a candidate items. Furthermore, the homepages of some items contains many pictures instead of text information. To make use of information hiding in the draws, they were labeled manually.

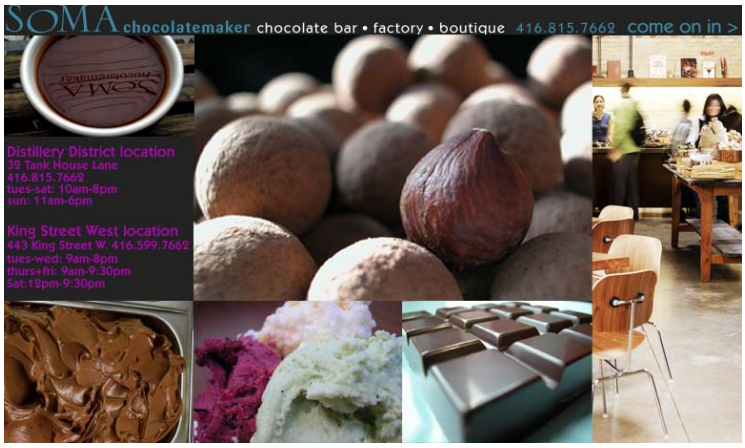

Figure 1. A picture from home page of an item: its labels are "coffee bean", "chocolate", and "ice-cream".

2.3 Context Filtering

Context profile provided four kinds of information——location, season, time, and weather. The location information has been used for choosing appropriate items when download candidate items. The other information would be used for making candidate items more suitable. Opening time of every item was checked to make it fitting the time provided by context. And suitable weight is given up to season and weather.

2.4 User Modeling

A user is assumed to be represented by attractions which the user's attitude is 1, and the attractions consist of documents on its homepage. Then a user can be supposed to be represented by some documents. After removing stop words, TF-IDF value of words for documents representing each user was calculated. TF-IDF value of words for description of each user was calculated in the same way. The value from the documents is named TF-IDF-doc, and the value from the description TF-IDF-des. Thus a user model can be structured with the words whose TF-IDF value was bigger than others. The first ten words will be chosen. Exactly, a scalar which composed of words represents a user model.

Those TF-IDF-des values of items which is judged "1" by user initially is also combined together and called TF-IDF-init-user.

More details about TF-IDF value can be found in references.

2.5 Similarity Computing and Ranking

Similarly, TF-IDF value of each candidate item waiting to be recommended was calculated. A scalar also composed of words represents a candidate item.

Cosine similarity was used to represent similarity between a user and an item. The similarity and the weight generated because of season and weather was combined together to produce final weight to every item, and then final ranking produced. The method used is as follows: Because items to be recommended are organized according to cites, similarities can be computed between

user and items of one city without the consideration of the geography information. To represent the influence of the time and season for one user to choose the suitable one, the weight was a plus to similarities.

2.6 Description Generating

The last question is how to generate the description.

Because the items are crawled from open web, the documents of candidate items are composed of assessment from real users. They are short, not rigorous, even syntactically wrong, and users usually express their feeling in their assessment.

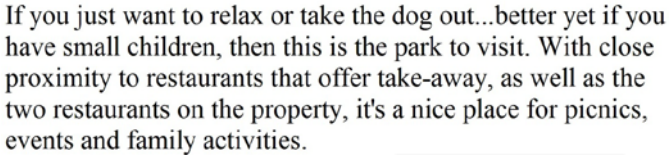

Figure 3. An assessment from a real person

As a result, descriptions are generated as follows: Texts about an item is split into small sentences with the punctuation and line breaks at first. Then sentences which have the words with greater TF-IDF value in the user model are chosen. To be noticed, sentences having these words like “I”, “we” or “you” are dropped because these sentences may be not suitable for description because they are probably the feeling of customers or owners. Because description fields may contain up to 512 characters from tag to tag, sentences are added into the description fields until the length of the description is larger than 400. So that only when the length of the last sentence is more than 112, the length of description is larger than the demand. For a small amount of status, some descriptions are cut off.

3. References

[1] http://en.wikipedia.org/wiki/Tf%E2%80%93idf

[2] http://en.wikipedia.org/wiki/Cosine_similarity