A Cross-Platform Collection for Contextual Suggestion

Mohammad Aliannejadi, Ida Mele, and Fabio Crestani
Faculty of Informatics, Università della Svizzera italiana (USI)
Lugano, Switzerland
{mohammad.alian.nejadi,ida.mele,fabio.crestani}@usi.ch

ABSTRACT

Suggesting personalized venues helps users to find interesting places on location-based social networks (LBSNs). Although there are many LBSNs online, none of them is known to have thorough information about all venues. The Contextual Suggestion track at TREC aimed at providing a collection consisting of places as well as user context to enable researchers to examine and compare different approaches, under the same evaluation setting. However, the officially released collection of the track did not meet many participants' needs related to venue content, online reviews, and user context. That is why almost all successful systems chose to crawl information from different LBSNs. For example, one of the best proposed systems in the TREC 2016 Contextual Suggestion track crawled data from multiple LBSNs and enriched it with venue-context appropriateness ratings, collected using a crowdsourcing platform. Such collection enabled the system to better predict a venue's appropriateness to a given user's context. In this paper, we release both collections that were used by the system above. We believe that these datasets give other researchers the opportunity to compare their approaches with the top systems in the track. Also, it provides the opportunity to explore different methods to predicting contextually appropriate venues.

CCS CONCEPTS

•**Information systems** →**Test collections;** *Recommender systems;*

KEYWORDS

Collection; Venue Suggestion; Context-Awareness

1 INTRODUCTION

Location-based social networks (LBSNs), such as Yelp, TripAdvisor, and Foursquare, allow users to share check-in data using their mobile devices. Such platforms also collect valuable information about users' mobility records such as check-in data and users' feedback (e.g., ratings, tags, and reviews). One important related task consists in suggesting personalized venues to a user who is exploring a new venue or visiting a new city [3]. The Contextual Suggestion track in the Text REtrieval Conference (TREC) aimed at providing a standard evaluation setup for researchers in which they can compare their proposed approaches [5]. For each user, the participants had to produce a ranked list of venues to recommend to a user visiting a new city. They were given the users' history of preferences expressed in one or two previously visited cities (30-60 venues per user) and had to consider both preferences and context of the users for making their suggestions.

In the last two years of the track, the organizers attempted to create reusable datasets. Consequently, the track consisted of two phases. *Phase 1*[1]: participants were free to suggest any venue in the target cities as long as they existed in a reference venue collection (see Table 1). *Phase 2*[2]: participants had to rerank a given list of venues in the target city for each user. Therefore, the coordinators were able to release the ground truth for Phase 2, that is an essential step toward making a collection reusable.

In spite of such attempts, there are still some drawbacks with the current collection that can limit its reusability. First, even though the organizers released a crawl of the collection in 2016, it is unstructured and does not introduce a homogeneous set of data (see Table 1). Hence most top-ranked systems ignored it and crawled their collections. Second, the collection has a set of contextual data such as type and duration of the trip. However, the contextual data was neglected by most of the participants. In particular, many of them just ignored the contextual information or used it with handcrafted rule-based methods. It could be due to the current structure of the collection which does not give the researchers many options concerning context-aware recommendation.

In an attempt to address these limitations, we present a collection that was crawled very carefully to be homogeneous, cross platform, and context aware. More specifically, we release the collection that we used for our participation at the TREC Contextual Suggestion track performing best in both phases of the track. The collection was crawled from two major LBSNs: Foursquare[3] and Yelp[4]. We searched for the venues present in the TREC dataset on the LBSNs to find their corresponding profiles and verified the retrieved data very carefully to prevent adding any noise to the dataset. It is worth noting that more than half of the submitted systems to TREC 2016 had crawled data from either Yelp or Foursquare or both. More specifically, we observed that for 12 tasks, namely, the last 2 years with 2 phases each and taking into account the top 3 systems, 11/12 (=92%) of the systems had crawled data from one or both sources and 7/12 (=58%) of the systems crawled data from more than one LBSN. Hence, it is clear that there is the need for a unified, comprehensive dataset of this information which is available publicly. As for the contextual information, we created a secondary dataset

SIGIR '17, August 7-11,2017, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080752

[1] *Phase 1* is the 2016 equivalent of *Live Experiment* in 2015. For simplicity we refer to both of them as *Phase 1*.

[2] *Phase 2* is the 2016 equivalent of *Batch Experiment* in 2015. For simplicity we refer to both of them as *Phase 2*.

[3] http://www.foursquare.com

[4] http://www.yelp.com

Table 1: Sample rows from the collection of venues used in TREC Contextual Suggestion 2015 and 2016.

ID	Context	URL	Title
TRECCS-00001768-394	394	http://shop.hobbylobby.com	Hobby Lobby
TRECCS-00001776-394	394	http://dominos.com	Dominos Pizza
TRECCS-00001777-394	394	http://www.arbys.com	Arbys
TRECCS-00001927-182	182	http://www.milwaukeepublicmarket.org	Milwaukee Public Market
TRECCS-00001934-182	182	http://www.botanasrestaurant.com	Botanas Restaurant

of contextual information that enables researchers to investigate different approaches to predict the contextual appropriateness of venues and study the influence of context on user behavior.

The released collection[5] comprises:

- more than 300K crawled venues from Foursquare providing enough information for venue recommendation research;
- more than 15K crawled venues from Foursquare and Yelp providing information for cross-platform recommendation;
- a human-annotated contextual appropriateness dataset containing human-tailored features for almost 2K context example features.

The release of this new collection will provide researchers with a unique opportunity to develop context-aware venue recommender systems under the same setting and data as the one of the best-submitted systems in the TREC 2016. This will enable them to compare their work with state-of-the-art approaches and explore the brand new venue-context appropriateness dataset.

The remainder of the paper is organized as follows: Section 2 briefly describes the TREC Contextual Suggestion track. Section 3 describes the method we used for the dataset construction. In Section 4 we provide more details on the dataset. Finally, Section 5 concludes the paper.

2 TREC CONTEXTUAL SUGGESTION TRACK

In 2012 TREC introduced the task of Contextual Suggestion[6] which provided a common evaluation framework for participants who wanted to deal with the challenging problem of improving contextual suggestions. More in details, given a set of example venues as user's preferences (profile) and some contextual information, such as geographical, temporal, and personal contexts, the task consisted in returning a ranked list of candidate venues fitting the user's profile and context.

The crawled dataset of this paper covers the places that were part of the last two years of the track, that is 2015 and 2016. For 2015 the dataset includes the venues for Phase 2, while for 2016 the dataset covers the venues for both Phase 1 and Phase 2. The additional contextual appropriateness dataset can also be used for 2015 and 2016 collections. In these tasks, there is a number of users, and for each user there is a list of 30 to 60 venues that a particular user has previously rated. Additionally, there is a set of contextual information for each user. Given such information, the task is to return a ranked list of candidate suggestions of venues based on their relevance to the user's needs and context. The collection has also a ground-truth made of relevance assessments which indicate whether a candidate suggestion is relevant to the user or not (i.e., whether the user likes the candidate suggestion or not).

For each venue the profile includes a rating in the following range **4:** very interested, **3:** interested, **2:** neutral, **1:** uninterested, **0:** very uninterested, and **-1:** not rated. Venues may also have tags that indicate why the user liked the particular venue. Contextual information is represented by: location, time, type of trip (*Business*, *Holiday*, or *Other*), duration of trip (*Night out*, *Day trip*, *Weekend trip*, or *Longer*), group of people involved (*Alone*, *Friends*, *Family*, or *Other*), and season (*Winter*, *Summer*, *Autumn*, or *Spring*). Moreover, user's age and gender are optionally included. This information can be exploited to have a better understanding of user's context in order to recommend appropriate venues. For example, you know that a user liked or disliked some venues in "New York City", then she goes to "Boston" for a *weekend* trip, she is *alone*, and it is *winter*. Given such information, the recommendation system should be able to rank the candidate suggestions so that the top-ranked places are the most appropriate venues for the user to visit in "Boston".

3 METHODOLOGY

We chose Foursquare and Yelp not only because they are two of the most popular LBSNs where many users leave their check-in data, but also because the type of information provided by Yelp is a perfect complement for the type of information on Foursquare. Moreover, as we will show in the statistics of the dataset, there is a considerable overlap of venues that have a profile on both LBSNs. However, there are places in the TREC dataset that appear only on one of the two crawled LBSNs, hence crawling data from both of them allows making the data gathering more complete.

Deveaud et al. [4] showed that venue-centric features which were extracted from Foursquare play a key role in venue recommendation. On the other hand, Chen et al. [3] argued that user reviews on venues provide a wealth of information that can be leveraged to address the data-sparsity and the cold-start problems for venue recommendation. Also, Yang et al. [6] showed that the accuracy of a recommender system can be significantly improved by extracting opinions from user reviews in Yelp. Also, almost all of the best performing systems in the TREC Contextual Suggestion track 2015 and 2016 [5] crawled data from these LBSNs. In particular, our previous work which were among the best runs in both 2015 [2] and 2016 [1] benefited from a comprehensive crawled dataset from Yelp and Foursquare. We also showed that a system can benefit from multiple LBSNs, and systems using reviews from Yelp and venue tags from Foursquare had the best performance.

[5] Available at http://inf.usi.ch/phd/aliannejadi/data.html

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

Table 2: Sample rows from the collection on venues contextual appropriateness. The numbers in parentheses show the degree of agreement between different assessors and ranges from −1 (full agreement on *no*) to +1 (full agreement on *yes*). Hence, 0 means that there is no agreement between the assessors and so the task is subjective.

Features				Output
Category	Trip Type	Trip Duration	Group Type	Appropriateness
Candy Store	Business (−0.60)	Day Trip (+0.60)	Friends (+1.00)	No (−1.00)
Library	Holiday (−1.00)	Weekend Trip (−0.58)	Family (+0.62)	No (−1.00)
Pharmacy	Holiday (−0.60)	Night out (−0.60)	Family (−0.18)	No (−1.00)
BBQ Joint	Holiday (+1.00)	Night out (−0.18)	Friends (+1.00)	Yes (+1.00)
Sandwich Place	Business (−0.60)	Longer trip (+0.20)	Alone (+0.60)	Yes (+0.68)
Lounge	Holiday (+0.58)	Weekend trip (+1.00)	Friends (+1.00)	Yes (+1.00)

3.1 Data Crawling

For our collection, we crawled data from two LBSNs: Foursquare and Yelp. Foursquare provides an easy-to-access API[7] which makes crawling quite easy. We used Foursquare's API to crawl a large number of venues and scraped a very smaller fraction of venues for additional information on the website. Yelp's API[8], on the other hand, has more restrictions and therefore we were able to crawl much fewer venues on Yelp. For TREC 2016 Phase 1, we only crawled data using the Foursquare's API since there were virtually 630K venues to crawl in a very limited time and thus the only option was using the Foursquare's API. For Phase 2 of TREC 2015 and 2016, there was more time and much fewer venues to crawl so that we could crawl data from both LBSNs.

The data was crawled in two time periods: July - August 2015 and July - August 2016. To find the corresponding profiles of venues on LBSNs, we used two search engines: 1) Foursquare venue search engine and 2) Google Custom Search. For TREC 2015 there were 8,794 venues from which we crawled 6,427 places from Yelp and 5,639 from Foursquare, with a considerable overlap between data crawled from Yelp and Foursquare. For TREC 2016 there were 18,808 venues from which we crawled 13,868 places from Yelp and 13,417 from Foursquare, again emerging a big overlap between the two sources. As all the venues are in the US cities, we expected that most of the users who reviewed the venues were English speakers.

Query Structure. For each venue in the collection, we created a query to search on the LBSNs. The query consisted of a venue's name and its location. We cleaned the venues' names in the TREC collection since many contained unrelated terms such as the host service (e.g., Facebook, Wikipedia). Finally, the query we used to search for venues was in the form:

$$\text{query} = \text{venue's name} + \text{venue's city} + \text{venue's state}\ .$$

Search Result Validation. Since we could not trust the results of search and in order to minimize the noise, we validated the returned results from the search engine following these steps:

(1) We first validated the city and state of the returned venue.
(2) We then measured the similarity between the name of the venue and the name of the returned place using Levenshtein distance.
(3) If the similarity between the two names (calculated in Step 2) was more than a threshold (70%), we considered that result as a match. If not, we continued steps 1-3 for other returned results up to the 5th result.

Note that the high similarity threshold (70%) was set to prevent adding possible noise to the collection.

3.2 Crowdsourcing

We used the CrowdFlower[9] crowdsourcing platform to collect judgments of contextual appropriateness of venues and create the additional contextual-appropriateness dataset. We asked a number of crowdworkers to judge if a venue category is appropriate for a trip description. For instance, if a trip was described in the collection as trip type: *business*, trip duration: *one day*, and group type *family*, for a venue with category *Pizza Place* then we asked crowdworkers to judge if the venue was appropriate to the trip. In particular, we asked them: "Is a *Pizza Place* appropriate for a *business* trip?", "Is it appropriate to go to a *Pizza Place* on a *one-day* trip?", "Is it appropriate to go to a *Pizza Place* with *family*?" While assessing such tasks could seem trivial and objective, in fact it is subjective in many cases (e.g., going to a *pharmacy* with *family*). Therefore, we asked at least 5 crowdworkers to provide their judgment to each row. If we found no agreement among the assessors, we considered the task as subjective. We considered the answer "appropriate" as a +1 score and "inappropriate" as a −1 score. Thus, the assessment agreement is the average of assessment scores. We asked workers to judge the context/category pairs for almost all possible pairs regardless of their existence in the TREC collections. This makes this dataset general enough to be used for other purposes.

We made sure to explain the task clearly to the workers and asked them to assess such appropriateness regardless of their personal preferences over categories. Also, we performed a training step and allowed only top-quality workers to do the task.

4 COLLECTION

The released collection contains more than 330K venues from Foursquare for TREC 2016 Phase 1 and 15,765 venues from both Foursquare and Yelp for TREC 2016 Phase 2. As we can see in Table 4 there were many broken or unrelated links in the the TREC collection (300K out of 600K), however, there were much fewer unrelated links

[7] https://developer.foursquare.com/docs/
[8] https://www.yelp.com/developers/documentation/v2/overview
[9] http://www.crowdflower.com

Table 3: Statistics on the crawled collection

	Phase 1	Phase 2	
	TREC'16	TREC'15	TREC'16
# venues in TREC collection	633,009	8,794	18,808
# venues crawled: Yelp	-	6,427	13,868
# venues crawled: Foursquare	336,080	5,639	13,417
# Yelp and Foursquare overlap	-	4,844	11,520
avg. reviews per venue	-	117.34	66.82
avg. categories per venue	1.35	1.63	1.57
avg. tags per venue	-	8.73	7.89
avg. user tags per user	3.61	1.46	3.61
# distinct user tags	150	186	150

Table 4: Statistics on the crowdsourced contextual appropriateness collection

Number of categories	179
Number of category-context pairs	1969
Number of assessments	11,487
Average assessments per pair	5.83
Average assessment agreement	85%
Number of full travel annotations	760

for Phase 2 (3K out of 18K). For each venue we release all available information: venue name, address, category, tags, ratings, reviews, check-in count, menu, opening hours, parking availability, etc.

The contextual-appropriateness collection consists of 1,969 pairs of trip descriptors and venue categories as features. In order to enable researchers to train their models using the contextual appropriateness of venues, we created another collection providing ground truth assessments for the contextual appropriateness of the venue categories. It completes the contextual information (i.e., trip type, group type, trip duration) for 10% of the whole TREC collection. This collection contains 760 rows including the features we already created using crowdsourcing and the context-appropriateness labels for venues. The 10% of labeled data allows to model the venues' contextual appropriateness given the users' context and to make prediction for the remaining 90% of the data, as we did in [1].

In Table 2 we report some rows of the collection, and Table 4 lists some statistics of the crawled collection. Figure 1 shows the histogram of venue-appropriateness features assessed by the workers. We divided the assessments in three groups based on appropriateness scores: [−1.00, −0.40): not appropriate (objective), [−0.40, +0.40]: no agreement (subjective), (+0.40, +1.00]: appropriate (objective). To categorize the tasks as subjective and objective, we assumed that those tasks for which there was a high agreement between the assessors could be considered objective since everybody agreed on their (in)appropriateness. While we assumed that those tasks with relatively lower agreement between the assessors could be considered subjective.

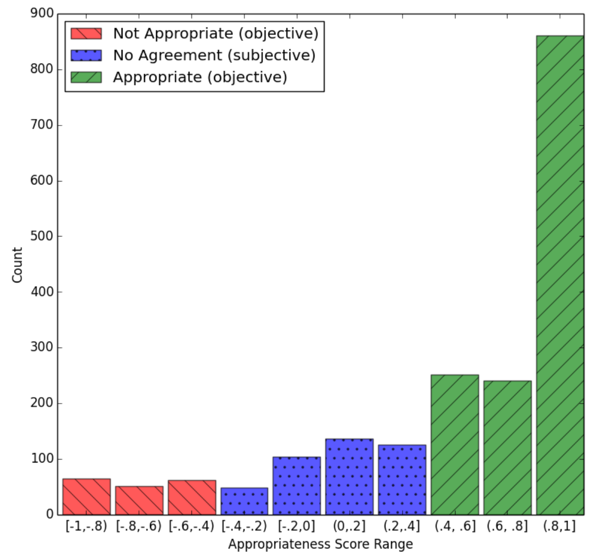

Figure 1: Histogram of venue-context appropriateness score ranges. We partition the histogram into 3 parts based on the scores range. Scores below −0.4 represent *inappropriateness* and score higher than +0.4 represent *appropriateness.* Scores between −0.4 and +0.4 do not provide much information and show no agreement among assessors (subjective task).

5 CONCLUSIONS

In this paper we present the dataset we used for our participation to the TREC Contextual Suggestion tracks. We crawled information of venues used in the TREC dataset from two popular LBSNs. Also we collected, using crowdsourcing, ratings on the venue-context appropriateness which allows to make predictions on the appropriateness of a recommended venue to a given user's context. Such collection can be helpful for other researchers interested in comparing their venue-recommendation techniques against state-of-the-art approaches. It could also foster further research on contextual suggestions of venues.

ACKNOWLEDGMENTS

This research was partially funded by the RelMobIR project of the Swiss National Science Foundation (SNSF).

REFERENCES

[1] Mohammad Aliannejadi and Fabio Crestani. 2017. Venue Appropriateness Prediction for Personalized Context-Aware Venue Suggestion. In *SIGIR 2017*. ACM.
[2] Mohammad Aliannejadi, Ida Mele, and Fabio Crestani. 2016. User Model Enrichment for Venue Recommendation. In *AIRS 2016*. Springer, 212–223.
[3] Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender systems based on user reviews: the state of the art. *UMUAI* 25, 2 (2015), 99–154.
[4] Romain Deveaud, M-Dyaa Albakour, Craig Macdonald, and Iadh Ounis. 2015. Experiments with a Venue-Centric Model for Personalised and Time-Aware Venue Suggestion. In *CIKM 2015*. ACM, 53–62.
[5] Seyyed Hadi Hashemi, Charles L. A. Clarke, Jaap Kamps, Julia Kiseleva, and Ellen M. Voorhees. 2016. Overview of the TREC 2016 Contextual Suggestion Track. In *TREC 2016*. NIST.
[6] Peilin Yang, Hongning Wang, Hui Fang, and Deng Cai. 2015. Opinions matter: a general approach to user profile modeling for contextual suggestion. *Information Retrieval Journal* 18, 6 (2015), 586–610.