University of Amsterdam at the TREC 2013 Contextual Suggestion Track: Learning User Preferences from Wikitravel Categories

Marijn Koolen[1] **Hugo Huurdeman**[2] **Jaap Kamps**[2,3]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[3] ISLA, Informatics Institute, University of Amsterdam

Abstract: This paper describes our participation in the TREC 2013 Contextual Suggestion Track. The goal of the track is to evaluate systems that provide suggestions for activities to users in a specific location, taking into account their personal preferences. As a source for travel suggestions we use Wikitravel, which is a community-based travel guide for destinations all over the world. From pages dedicated to cities in the US we extract suggestions for sightseeing, shopping, eating and drinking. Descriptions from positive examples in the user profiles are used as queries to rank all suggestions in the US. Our user-dependent approach merges the per-query rankings of the positive examples of a single user. We automatically classified the rated examples according to the Wikitravel categories—Buy, Do, Drink, Eat and See—and derived a user-specific prior probability per category. With these we re-rank Wikitravel suggestions. The ranked suggestions are then filtered based on the location of the user.

1 Introduction

Wikitravel[1] is a collaboratively created site for travel and tourist information, with lists of things to see and do in places all over the world. Locations are neatly structured in countries, states, regions, districts, cities and suburbs and have a dedicated page, and the places to visit within each location are presented in lists and tables in each page. This information provides travellers with easy access to a list of options for sightseeing, shopping, eating, drinking and sleeping. If you find yourself in a particular city, it is easy to browse this list. For larger cities, the number of options can be very large and is often spread over multiple pages, making it hard to find options that you like. For smaller places the list can be very short and not contain anything of interest in the immediate area, but pages on nearby places may have better options.

[1] URL: http://wikitravel.org/

Our aim for the TREC 2013 Contextual Suggestion Track is to use Wikitravel as a source for suggestions based on the user's current location, which are ranked by distance and how well they match the user's known preferences.

We use the descriptions of the suggestions as document representations and the descriptions of preferred items in the profiles as queries to retrieve and rank suggestions.

The rest of this paper is organised as follows. We first describe our experimental setup in Section 2. We discuss our results in Section 4 and provide a more detailed analysis in Section 5. We summarize our findings in Section 6.

2 Experimental Setup

2.1 Data collection

Wikitravel is an open platform where anyone can add, edit and delete travel information about places in the world. There are many pages, each dedicated to a specific city or town, with sections describing how to get there and things to see and do. Most pages are structured according to some general rules, to get a consistent travel guide, with clearly separated sections for transportation, sightseeing, shopping and accommodation. Activities, attractions, restaurants and bars are usually presented in lists or tables, with the name of the shop, museum, park, restaurant or hotel, a short description and often a hyperlink to the homepage of a dedicated site. These are provided by a community of travellers and locals and can be used as a source for contextual suggestions.

We crawled all Wikitravel pages of locations within the US, starting with the page on the United States of America as the seed list. We extracted site-internal links from all the *States*, *Regions*, *Cities*, *Districts* and *Burroughs* sections. The pages within the *Districts* and *Burroughs* categories describe neighbourhoods in large cities. While extracting links from each of these sections, a mapping is stored that identifies how the source is connected to the target page. For instance, in the *Regions* section of the page for the U.S. state Oregon we extract links for the regions *Cascade Mountains*,

Table 1: Number of suggestions and examples in each Wikitravel category

Category	# suggestions	%	# examples	%
Buy	2496	11	10	20
Do	5841	27	12	24
Drink	2476	11	7	14
Eat	6333	29	9	18
See	4726	22	12	24
Total	21,872			

Central Oregon, *Columbia Gorge* and four other regions. With each link we store a mapping indicating that that region is a region in the state *Oregon*. This hierarchical mapping can be used as an indication of distance between the location of the user and other locations. When there are not enough suggestions in the city where the user is located, we can add suggestions from cities in the same region. From the *City*, *District* and *Burrough* pages we extracted suggestions from the sections *Do*, *See*, *Buy*, *Eat*, and *Drink*. Each suggestion is identified by either a paragraph, list item or table row in html markup. We only considered items that have an hyperlink to an external web page as suggestions and used the surrounding text in the list item or table row element as description. We extracted a total of 21,872 suggestions from 1735 cities and towns. For some locations there is only a single suggestion, the median (mean) number of suggestions is 4 (13). The place with the highest number of suggestions is Chicago (816 suggestions).

The number of suggestions from each category is shown in column 2 of Table 1. The Buy and Drink categories are the smallest, with 2496 and 2476 suggestions respectively. The Eat category is the biggest, with 6333 suggestions (29%).

3 Category Priors

Some users may prefer *Do* suggestions over *Drink* suggestions, or *Eat* over *Buy* suggestions. From the ratings of the examples, the system could derive a predicted category, but for the 50 examples provided by the Track organisers, the Wikitravel categories are unknown (although some of them may be on the Wikitravel page for Philadelphia). Therefore, we use the descriptions from the 50 examples and the 21,872 Wikitravel suggestions to assign the 50 examples to the 5 Wikitravel categories. To assign the examples to the categories, we crawled all 50 example websites, downloading the homepage from each example, and following site-internal links up to one level deep. Subsequently, we extracted and concatenated the plain text from all crawled pages for each separate site, and tokenized it, followed by basic stop word filtering. Using a tf-idf measure, we extracted the top 30 keywords for each example website, that could serve as queries. The crawled and concatenated text of each of the 5 Wikitravel categories served as document representations, which we indexed using Indri. Issuing the generated queries based on the top 30 keywords per site resulted in a ranked list of the 5 candidate categories for each given example website. With the ratings per user, we can then compute predicted ratings per category. Each profile contains the ratings of the 50 examples by a single user, on a 5 level scale: strongly uninterested (0), uninterested (1), neutral (2) interested (3) and strongly interested (4). Of the 50 examples, 10 are assigned to the *Buy* category (column 4 in Table 1), 12 to *Do*, 7 to *Drink*, 9 to *Eat* and 12 to *See*. The distribution is somewhat different over the 50 examples than over the Wikitravel suggestions. The number of ratings per category is small and may be too low to predict a rating useful for re-ranking. We investigate this by comparing a ranking based on document retrieval scores alone with a ranking based on both document scores and category-based predicted ratings.

Table 2: Statistics on the user ratings across profiles, examples and Wikitravel categories

Aggregate	#	min	max	medn	mean	stdev
Profiles	562	0.38	4.00	2.38	2.38	0.51
Examples	50	1.41	3.49	2.37	2.38	0.44
Category						
Buy	4496	0	4	3	2.37	1.28
Do	7306	0	4	2	2.14	1.31
Drink	3372	0	4	2	2.11	1.35
Eat	3934	0	4	3	2.68	1.19
See	8992	0	4	3	2.54	1.20

We provide statistics on the user ratings (based on the ratings after seeing the full document) in Table 2. In the top part of the table, the we see the rating distribution over profiles, where the ratings of the 50 examples are averaged per profile. For the statistics on the examples we first average over the 562 profiles. The preferences of users are highly varied. Some users are at best neutral towards examples. Some users are positive about a few things but negative about most other things, with a median score of 0. Others are mostly positive, with median ratings of 4. One user gave all examples a rating of 4. The ratings over the examples are distributed more evenly, with the lowest rated example having an average rating of 1.41 and the highest 3.49. In the bottom half of Table 2 we show rating statistics per Wikitravel category, based on the estimated category per example. The Do and Drink categories are the least liked while the Eat category is the highest rated. Per profile the category ratings vary strongly. Some strongly prefer the See category while others prefer the Buy or Drink categories. These preferences can be captured by the user's personal rating distribution over categories.

The average rating $\bar{r}_u$ of examples D_E by a user u is given

as:

$$\bar{r}_u = \frac{1}{|D_E|} \sum_{d \in D_E} r_u(d) \quad (1)$$

The average rating of example websites D_C in Wikitravel category C by user u is:

$$\bar{r}_u(C) = \frac{1}{|D_C|} \sum_{d \in D_C} r_u(d) \quad (2)$$

We use these average ratings as category-based predicted ratings to rerank retrieved suggestions.

3.1 Indexing and Retrieval

Each suggestion is a document with the description as representation, which we indexed with Indri. We used Krovetz stemming and removed common stopwords. The topic set consists of 50 examples, 562 user profiles and 50 contexts. The examples are suggestions in Philadelphia and consist of a short description and a URL to a dedicated website. Each user profile contains judgements from a single user on all 50 examples, with an initial judgement based on the description of the example suggestion and a final judgement after visiting the website. The contexts contain a location (city and state in the US). In the user profiles, the description of each positive example (where the user rated the example positive (score 3 or 4) based on seeing the actual website) was used as a query, resulting in the set Q_u^+.[2] We ranked suggestions per query (default language model with Dirichlet smoothing, $\mu = 2500$) and scores are merged over all queries per profile using CombSUM. The score of each retrieved suggestion is the sum of all its scores for all queries q for user u. Formally, score $S(d)$ for suggestion d is computed as:

$$S(d) = \sum_{i=1}^{|Q_u^+|} P(d|q_i) \quad (3)$$

The language model score $P(d|q_i)$ is computed as:

$$P(d|q_i) = P(d) \cdot P(q_i|d) \quad (4)$$

where $P(d)$ is a document prior probability, which is $P(d) = 1$ in the baseline system and $P(d) = \bar{r}_u(C)/r_{max}$, with $r_{max} = 4$ when we use the category-based predicated rating. This produces a location-independent ranking of suggestions, which can be updated each time the user adds new information to her profile. When the user wants suggestions based on where she is, the ranking is filtered on distance to her location. All suggestions within the city where the user is located are ranked first, then suggestions within the same region, then within the same state, then the rest of the suggestions. The top 50 suggestions are returned to the user. For large cities this often means all suggestions are within the same city. For smaller locations, with only a small number of suggestions, this often means the suggestions further down the list require some travelling. In Section 5 we analyse the difference between suggestions for small and large cities.

[2]Profile IDs 146 and 420 have no positively rated examples. For these profiles, we use all neutrally rated examples as queries.

3.2 Official Runs

For this year's Contextual Suggestions Track, systems have to provide 50 suggestions for each pair of user and context. There are $562 \cdot 50 = 28,100$ user/context pairs. We submitted one run:

UAmsTF30WU : this is a baseline run without category priors. Suggestions are ranked per profile/context pair based on the positive examples in the profile and filtered on the context location, with additional suggestions from other cities in the same region or state if there are fewer than 50 suggestions in the context location itself.

In addition, we prepared another run, which we unfortunately could not submit in time:

UAmsTF30WUC : This run is the same as the one above, but with the category prior probability assigned to each suggestion.

These runs allow us to investigate the value of the category priors. Is a category-based document prior effective for ranking? Or are category preferences already captured by using only the descriptions of positively rated examples?

4 Results

We submitted only the baseline run so there are no official results for the run with category priors. However, to find out if the category prior has potential value for improving the ranking, we take the top 5 results of the baseline run and rerank them based on the ranking of the *UAmsTF30WUC* run and designate this run UAmsTF30WU$_{cat}$. Note that this results in a different ranking from the *UAmsTF30WUC* run, as that run may have different suggestions in the top 5 than the baseline. However, if the category prior is an effective relevance indicator, we expect it to improve performance scores.

Suggestions are judged on 3 aspects: the description (Desc) of the suggestion, the document (Doc) and the geographical location (G). Suggestions are considered relevant if the G score is at least 1 (marginally geographically appropriate) and Desc and Doc scores are at least 3 (*interesting*) for P@5 and MRR. For TBG, the Desc has to be at least 2 (neutral) and Doc at least 3. All dimensions are judged on the same 5 level scale as the examples.

The evaluation results are shown in Table 3. The Track Median is the mean of the per topic Median scores. Our baseline *UAmsTF30WU* scores well above the Track Median

Table 3: Evaluation results for the TREC 2013 Contextual Suggestion Track. The run marked * is not an official submission

Run	P@5	MRR	TBG
Track Median	0.2368	0.3415	0.8593
UAmsTF30WU	0.3121	0.4803	1.1905
UAmsTF30WU$_{cat}$	0.3237	0.5036	1.2413

Table 4: Distribution of judgements based on descriptions and documents

# results	description	document (%)
does not load	0 (0)	184 (17)
strongly uninterested	68 (6)	74 (7)
uninterested	161 (15)	134 (13)
neutral	251 (24)	149 (14)
interested	474 (44)	365 (34)
strongly interested	113 (11)	161(15)

on all three performance measures. The reranking of the top 5 results based on the category prior, UAmsTF30WU$_{cat}$, further improves performance. The category prior seems to be a useful signal for ranking. We analyse our runs in more detail in the next section.

5 Analysis

In this section we take a closer look at differences between users, the per topic performance of our two methods and the impact of user-dependent result merging on the final ranking.

There are a few notable differences between the description and document judgements (see Table 4). First, 17% of the documents fail to load, which does not happen with the descriptions. Because of this there are fewer *neutral* and *interested* document-level judgements compared to description-level judgements. Second, the number of negatively rated suggestions remains relatively stable. However, the number of *strongly interested* suggestions increases at the document judgement level. In general, going from the description-based ratings to the document-based ratings there are some small shifts from the moderate ratings to the more extreme ratings.

In Table 5 we see the relation between the description-based judgements and the document-based judgements. Row 1 shows how the suggestions initially given a negative rating (229 in total) and how these are rated after seeing the full document. Most suggestions are still rated negatively (143 or 62%), but 18% (28 and 14 out of 229) are rated higher. Suggestions initially rated neutral (row 2) tend to get a non-neutral after seeing the full document: 43 are rated negatively (17%) and 72 positively (29%). Of the suggestions rated positively upon seeing the descriptions, the majority are also rated positively based on the document (440 or 75%). Of the rest, most change due to pages failing to load (90 or 15%), while for 57 (10%) the pages turn out to be less than interesting. In total, 86 judgements shift from *negative* or *neutral* to *positive* while only 57 shift form positive to *neutral* or *negative*. Again, we see that ratings become less neutral upon seeing the full document, but when the page loads, the majority of the top 5 results are rated positively.

Table 5: Change in judgement from description to document

		Change to			
Description	total	not load	negative	neutral	positive
negative	229	44	143	28	14
neutral	251	50	43	86	72
positive	587	90	22	35	440
total	1067	184	208	149	526

Table 6: Distribution of categories over suggestions in the top 5 retrieved results and in the index

Category	Retrieved	Indexed
Buy	0.09	0.11
Do	0.28	0.27
Drink	0.05	0.11
Eat	0.18	0.29
See	0.39	0.22

5.1 Categories

Finally, we look at the categories of the top 5 retrieved results. Recall that the Wikitravel suggestions all have explicit categories, whereas for the examples we had to estimate a category.

In Table 6 we see the distribution of Wikitravel categories over the top 5 retrieved suggestions and over all suggestions in the index. The See category is overrepresented in the top 5, whereas the Eat and Drink categories are underrepresented. The Buy and Do categories are similarly distributed in the top 5 and the index. The high number of suggestions from the See category may be due to the relatively high ratings of the examples in the See category. However, the examples from the Eat category were rated even higher but fail to push Eat suggestions to the top of the ranking.

6 Conclusions

In this paper, we detailed our official runs for the TREC 2013 Contextual Suggestion Track. We extracted a larger number of suggestions from Wikitravel pages on cities and towns in the US and created two systems that generate geographically independent rankings. Per geographic context the ranked suggestions are filtered on location. This year we experimented with the Wikitravel suggestion categories for buying, doing, drinking, eating and seeing. By estimating the Wikitravel category for the provided examples, we created personalised category prior probabilities. For the baseline system, suggestions are ranked per user profile based on their positively rated examples and filtered on the geographic context. We compare this against a system that incorporates the personalised category prior. Unfortunately, due to time constraints, only the baseline run was submitted so we cannot properly measure the impact of the category prior on performance. However, by reranking the top 5 results of the baseline according to how the system with category prior would rank them and using the same relevance judgements we found that the category prior improves both early precision and Time-Based Gain. Part of making good suggestions is knowing what type of activities a user likes.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO projects # 640.005.001) and by the European Communitys Seventh Framework Program (FP7 2007/2013, Grant Agreement 270404).