Contextual Suggestion from Wikitravel: Exploiting Community-based Suggestions

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

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

Abstract: This paper describes our participation in the TREC 2012 Contextual Suggestion Track. The goal of the track is to evaluate systems that provide suggestions for activities to users in a specific location, at a specific time, 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 baseline approach merges the per-query rankings of all positive examples of all users. Our user-dependent approach merges the per-query rankings of the positive examples of a single user. The rankings suggestions are then filtered based on the location of the user. We ignore the temporal aspects of the context. The user-dependent rankings are more effective for contextual suggestion than user-independent rankings. The two systems show similar perform on the geographical dimension, but the user-dependent system provides more interesting suggestions. Our results show that information on user preferences is valuable for providing appropriate suggestions.

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. Our aim for the TREC 2012 Contextual Suggestion Track [1] 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 3 and provide a more detailed analysis in Section 4. We discuss some aspects of the relevance judgements in Section 5 and summarise 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 Wiki Travel 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 de-

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

scribe 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*, *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 20,200 suggestions from 1587 cities and towns. For some locations there is only a single suggestion, the median (mean) number of suggestions 4 (13). The place with the highest number of suggestions is Chicago (816 suggestions).

2.2 Retrieval

Each suggestion is a document with the description as representation, which we indexed with Indri [2]. We used Krovetz stemming and removed common stopwords. In the user profiles, the description of each positive example (where the user rated the suggestion positive both when reading the description and seeing the actual page) was used as a query, resulting in the set Q_u^+. 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 it's score for all queries q for user u. Formally, score $S(d)$ for suggestion d is computed as:

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

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 4 we analyse the difference between suggestions for small and large cities.

2.3 Official Runs

The topic set consists of 34 user profiles, 49 examples and 50 contexts. The examples are suggestions in Toronto and consist of a short description and a URL to a dedicated website. Each user profile contains judgements from a single user on all 49 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), time of day (morning, afternoon or evening) and season of the year (spring, summer, autumn, winter).

For this year's Contextual Suggestions Track, systems have to provide 50 suggestions for each pair of user and context. There are $34 \cdot 50 = 1700$ user/context pairs. We submitted two runs:

UAmsCS12wtSUMb : this is a baseline run that is user-independent. The ranking is the based on the positive examples of all user profiles.

UAmsCS12wtSUM : This is a location-dependent run, where suggestions in the user's location are ranked first, then suggestions in the same region, then suggestions in the same state, etc.

These runs allow us to investigate the value of individual user profiles. Is the profile of an individual user better for ranking suggestions than a general profile?

3 Results

Suggestions are judged on 4 aspects: the description (D) of the suggestion, the website (W), the geographical location (G) and the temporal aspect (T). Evaluation is focused on the W, G and T dimensions. All dimensions are judged on a 3 level scale: not appropriate/interesting (0), marginally appropriate/interesting (1) and appropriate/interesting (2). The official evaluation reports on three measures: W, GT and WGT. To be relevant for the GT measure, a suggestion has to score both G=2 and T=2. For WGT it has to score 2 for W, G and T.

Evaluation results for our two official runs are shown in Table 1. Columns 2, 3 and 4 are averaged over all profiles per context. Columns 5, 6 and 7 are averaged over all contexts per profile. The Median is based on the median per topic score of all submitted runs, and is calculated as the average over all median per topic scores.

The user-independent baseline UAmsCS12wtSUMb generally scores below the Median except for the website dimension per context. The user-dependent run UAmsCS12wtSUM scores above the Median on the website dimension, but well below the Median for the geo-temporal dimension. Our system ignores the temporal aspects of the context, so the higher Median scores suggests other participating systems did incorporate temporal aspects.

Table 1: Evaluation results for the official submissions, averaged per context (columns 2–4) and averaged per profile (columns 5–7). Best scores are in bold

	P@5					
	Context			Profile		
Run	WGT	W	GT	WGT	W	GT
Median	**0.1456**	0.3193	**0.5214**	0.0943	0.3400	**0.4729**
UAmsCS12wtSUM	0.1429	**0.3743**	0.2438	**0.1211**	**0.3772**	0.2253
UAmsCS12wtSUMb	0.0743	0.3286	0.2360	0.0632	0.3035	0.1971
	MRR					
	Context			Profile		
UAmsCS12wtSUM	**0.1781**	**0.5321**	**0.3174**	**0.1629**	**0.5471**	**0.3089**
UAmsCS12wtSUMb	0.1109	**0.5321**	0.2813	0.0962	0.5015	0.2401

When we compare our two runs to each other, it is clear that using the preferences of a single user improves performance on the website dimension. Interestingly, it also improves performance on the geo-temporal dimension, even though both runs use the same distance-based filtering method and both ignore the temporal aspects of the context. In the next section we analyse performance for the individual relevance dimensions and individual topics to try and find an explanation for this phenomenon.

For the context averages, the GT and WGT scores are higher than for the profile averages. For each profile, between 16 and 19 contexts have been judged. For each context, the number of profiles judged varies between 2 and 34. From this we expect the profile means to be more stable and show less variance. The mean per context scores range from 0.0 and 0.9 while the mean per profile scores range from 0.1556 to 0.3125. It is clear that our system fails for some contexts, regardless of which user it provides suggestions for. It seems the higher G and WGT scores for context averages are caused by high scoring outliers.

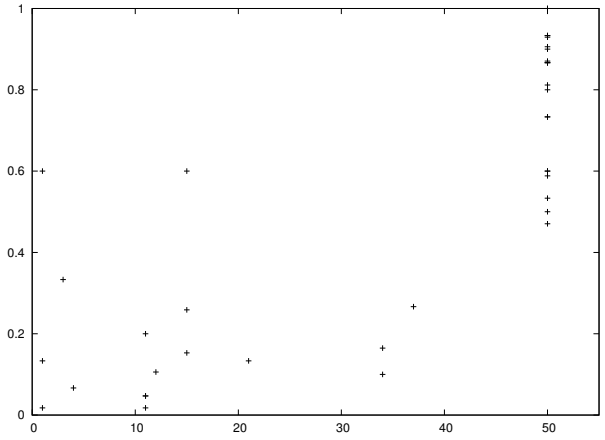

Figure 1: Relation between the number of intra-city suggestions and geographical precision (P@5(G).

4 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.

4.1 Suggestions per City

Is there a relation between the number of suggestions available in the context city and the number of suggestions that are geographically relevant? If suggestions from outside the context cities are geographically irrelevant, we should focus on finding other sources for suggestions in those cities where few are provided on Wikitravel.

Figure 1 shows the relation between the number of suggestions in the context city and the fraction of geographically relevant suggestions in the top 5. The x axis represents the number of suggestions in the top 50 that are located in the context city. Data points below 50 represent contexts where there are fewer than 50 suggestions on Wikitravel for that city. For instance, the data points at $x = 11$ represent cities where only 11 suggestions were extracted from Wikitravel. The remaining 39 suggestions in the top 50 come from other cities in the region or state. The y axis represents the precision on the geographical dimension, P@5(G). The scores below 0.4 all represent contexts where fewer than 50 suggestions are available. Almost all scores above 0.4 represent contexts where all top 50 suggestions are in the context city.

There is a clear relation between the number of suggestions available in a city and the P@5(G) score. The back-off strategy to add suggestions from other cities in the same region seems ineffective. It might be more effective to search the web for other suggestions in the context city.

Table 2: Strict pairwise agreement of different users on the provided examples

	# Pairs	Min	Max	Median	Mean	Std.dev
Initial	561	0.18	0.76	0.41	0.41	0.09
Final	561	0.20	0.69	0.41	0.42	0.08
Initial	9	0.00	0.80	0.40	0.44	0.25
Final	9	0.40	0.75	0.50	0.54	0.13

4.2 Comparing Users

To what extent do users judge the examples differently? We compare the overlap in examples and within those sets the overlap of the judgements. All 34 users provided initial judgements based on descriptions and final judgements based on websites for 49 examples. This allows us to compare the pairwise agreement for each pair of users. In Table 2 we show statistics on the strict pairwise agreement for the examples (rows 2 and 3), in which we consider user agreement only if two users choose the exact same level of interest (0 for not interested, 1 for marginally interested, or 2 for interested). There is little difference between agreement levels for initial and final judgements. The mean (median) agreement for initial judgements is 0.41 (0.41) and 0.41 (0.42) for final judgements. The maximum agreement is 0.76 for initial judgements and 0.69 for final judgements. Clearly, users have different preferences, indicating there is value in adjusting suggestions to personal preferences.

To what extent do users judge the suggestions of our systems differently? The user-independent ranking (UAmsCS12wtSUMb) gives the same suggestions in the same order for all users, which allows us to compute user agreement on the judged suggestions. These are shown in rows 4 and 5 of Table 2. The agreement on the initial judgements are similar to those for the examples, with a lower minimum and a larger variance. This is probably due to the low number of data points per pair of users. Only the top 5 suggestions are judged by both users, compared to the 49 data points for the examples. The agreement on the final judgements has a slightly higher mean and median than for the examples, but this may again be due to the low number of pairs and the low number of data points per pair. However, in general the agreement for judged suggestions is similar to that of the examples. This indicates judgement behaviour is similar across suggestions and examples.

4.3 Comparing Methods

Recall that both our submissions use the descriptions of positively judged examples as queries to rank suggestions. The UAmsCS12wtSUMb run ignores individual users and merges the rankings from all these queries, while the UAmsCS12wtSUM run merges the rankings for the positive examples of a single user. How different are these rankings? And are the performance differences stable and similar across all user profiles and contexts, or do the two systems perform differently on only a small number of profiles and contexts? We look at the overlap in the rankings of the two systems and the per topic differences in P@5 for the different relevance dimensions.

Table 3: Number of topics for which user preference improves performance

Change	G	T	GT	W	WGT
↑	124	267	155	19	14
=	339	201	357	15	29
↓	142	137	93	10	1

The overlap between the two runs starts at 26% at rank 1—that is, on average 0.26 of the results of the top 1 results overlap. At rank 5 the overlap is still 26%, but then steadily grows to 38% overlap at rank 50. The top of the rankings are substantially different. The ranking is very sensitive to which queries are used in producing the location-independent ranking. That the overlap increases further down the ranking is due to the decreasing number of different suggestions to choose from when creating the location-dependent ranking.

In Table 3 we show the number of topics for which the user-dependent method improves performance (row 2, ↑), decreases performance (row 4, ↓) and achieves the same performance (row 3, =) as the user-independent baseline, for five relevance dimensions: geographical (column 2, G), temporal (column 3, T), geo-temporal (column 4, GT), website (column 5, W) and geo-temporal and website (column 6, WGT). Here we see that the user-independent baseline more often outperforms the user-dependent system on geographical relevance than vice versa. This is perhaps due to the larger number of queries used per context. The baseline UAmsCS12wtSUMb run uses all positive examples from all users as queries, and may match with more suggestions in the context city than the user-dependent UAmsCS12wtSUM run. However, in the majority of cases, the two systems have the same score. This is not surprising, given that they use the same geographical filtering technique.

Although both systems ignore temporal aspects of the contexts and the suggestions, there is a large difference in per-

formance on the temporal relevance dimension. The user-dependent system scores higher than the baseline on 267 profile–context pairs, but lower on only 137 pairs. Although this can explain the difference in P@5(GT) in Table 1, it is not clear why this performance difference on temporal relevance occurs.

Figure 2 shows the per topic differences for P@5(G) (top), P@5(T) (centre) and P@5(W) (bottom). This further demonstrates that both methods perform similarly on the geographical dimension, but the user-dependent system performs better on the temporal and website dimensions.

5 Discussion

Our analysis has brought up questions about the relevance dimensions. We are not able to explain how it is possible that when both systems ignore the temporal aspects, one still clearly outperforms the other on temporal relevance. More details on the relevance judgements criteria and procedures could be insightful. How is temporal relevance judged? Is the only necessary condition that a suggested attraction or activity is available at the specified time of day? That is, do judges only consider whether the time of day falls within the opening hours of a suggested attraction? If this is the case, perhaps the user-independent baseline offers suggestions that have more restricted opening hours than the user-dependent system. The user-independent baseline probably focuses on suggestions that most users like, which may be mostly day time activities or only night time activities.

For geographical relevance, we would also like to know more about the relevance criteria. How should geographic relevance be treated? A suggestion for a bar to have a coffee that is 50 miles from the user's location is perhaps less relevant than a museum with famous paintings that the user likes. One is less likely to travel far for a coffee than for some famous site or a special concert. The uniqueness of what the suggestion offers plays a role in what the user considers an acceptable distance. The mood of the user may play a role as well (*I don't feel like travelling far*), but is not part of the provided context. But we argue that geographical relevance depends on how appropriate the travelling distance is for the provided suggestion.

6 Conclusions

In this paper, we detailed our official runs for the TREC 2012 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. One system also ignores individual user preferences, while the other tries to take those preferences into account when ranking suggestions. Geographical filtering is done per context city, providing fast query-time result selection.

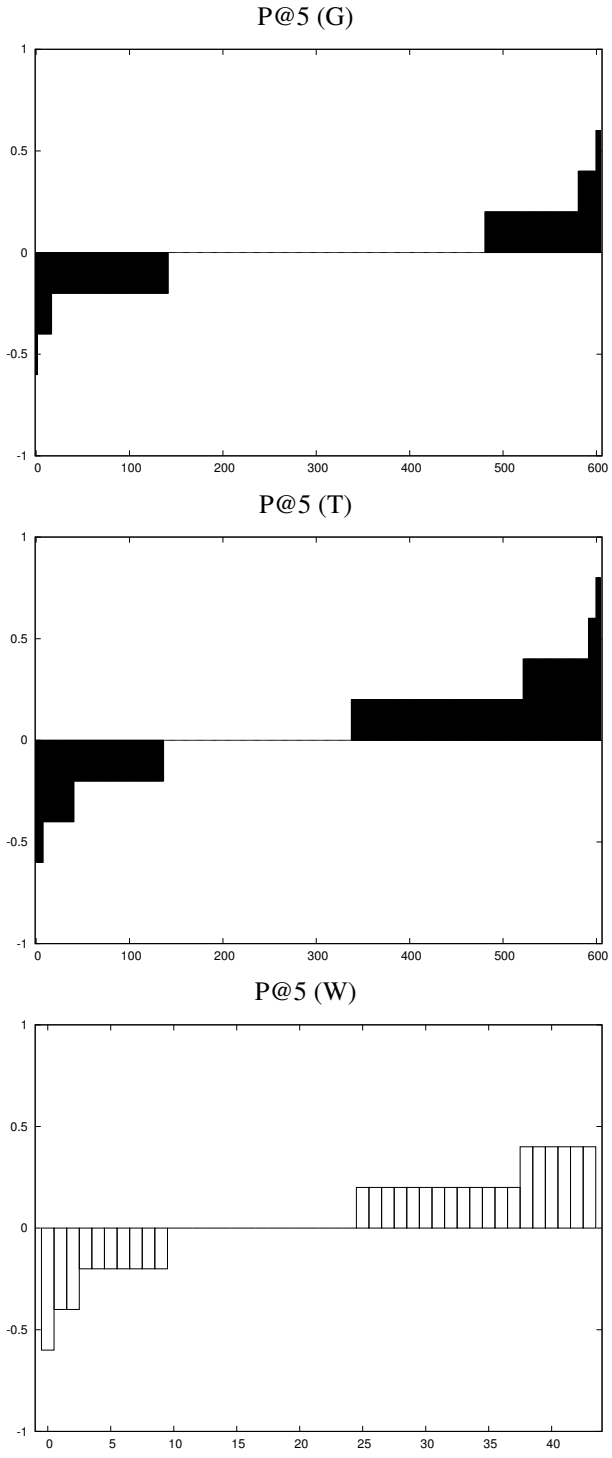

Figure 2: Improvement per topic of user-dependent run (UAmsCS12wtSUM) over user-independent run (UAmsCS12wtSUMb) for MRR(W) (top) and P@5(W) (bottom).

Our results show that ranking suggestions based on user preferences outperforms those of the user-independent baseline. Analysis shows that the two systems have little overlap in the top suggestions, showing that tailoring the results to specific user preferences has a large impact on the ranking. Since both systems use the same geographical filtering technique, they show similar performance on the geographical dimension.

Both systems ignore the temporal aspect of the context (time of day and season), but for some unknown reason, the user-dependent ranking performs better on the temporal dimension than the user-independent baseline.

For future work, we want to incorporate temporal aspects into retrieval model by taking the opening hours of suggestions and the time of day of the context into account. We also would like to expand the number of suggestions, especially for cities with a small number of available suggestions. Sources for these suggestions could be local travel sites or the results from specific queries to general purpose search engines.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO projects # 612.066.513, 639.072.601, and 640.005.001) and by the European Communitys Seventh Framework Program (FP7 2007/2013, Grant Agreement 270404).

References

[1] A. Dean-Hall, C. Clarke, J. Kamps, P. Thomas, and E. Voorhees. Overview of the TREC 2012 Contextual Suggestion Track, 2012.

[2] Indri. Language modeling meets inference networks, 2012. http://www.lemurproject.org/indri/.