Recommending personalized touristic sights using Google Places

Maya Sappelli, Wessel Kraaij
TNO and Radboud University Nijmegen, The Netherlands
m.sappelli@cs.ru.nl, w.kraaij@cs.ru.nl

Suzan Verberne
Radboud University Nijmegen, The Netherlands
s.verberbe@cs.ru.nl

ABSTRACT

The purpose of the Contextual Suggestion track, an evaluation task at the TREC 2012 conference, is to suggest personalized tourist activities to an individual, given a certain location and time. In our content-based approach, we collected initial recommendations using the location context as search query in Google Places. We first ranked the recommendations based on their textual similarity to the user profiles. In order to improve the ranking of popular sights, we combined the initial ranking with rankings based on Google Search, popularity and categories. Finally, we performed filtering based on the temporal context. Overall, our system performed well above average and median, and outperformed the baseline — Google Places only — run.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation, User profiles and alert services*

Keywords

Recommender Systems, Contextual Suggestion

1. INTRODUCTION

According to a report from the The Second Strategic Workshop on Information Retrieval in Lorne (submitted to SIGIR Forum, 2012), "Future information retrieval systems must anticipate to user needs and respond with information appropriate to the current context without the user having to enter an explicit query". At TREC 2012, a new track was organized: the contextual suggestion track [1], in order to evaluate such proactive systems. In this track the goal was to suggest personalized tourist activities to an individual, given a certain geo-temporal context.

As input to the task, each group participating in the track was provided with a set of 34 profiles, 49 examples of tourist activities and 50 geo-temporal contexts in XML format.

Each tourist activity example consisted of a title and a short description of the activity as well as an associated URL. The tourist activity examples were a collection of bars, museums and other tourist activities in the Toronto area.

Each profile corresponded to a single user and consisted of a list of rated URLs of the tourist activity examples in Toronto. The ratings were divided into an initial rating, based on the title and description of the URL and a final rating, which was given by the user after he/she viewed the website. These ratings could be used as training data to infer the particular preferences for this user.

For testing, systems needed to generate suggestions for 50 geo-temporal contexts. Each context consisted of spatial information (city-name, state-name, latitude, longitude) and categorical temporal information (day, time and season). The day could be a weekday or a weekend day and the time was either morning, afternoon or evening.

The task for the participating teams was to build a system that automatically provides a ranked list of 50 suggestions for each profile/context pair. Each suggestion should contain a title, description and associated URL. The description of the item could be personalized. The suggestions should be appropriate to the profile as well as the geo-temporal context. Time-wise, the user has five hours available for the suggestion, limiting acceptable locations of suggestions.

For evaluation, a selection of these suggestions were rated by the persons that provided the profiles, and the suggestions were assessed on their fit to the spatial and the temporal context by professional assessors as well.

Although there is quite some research in the area of mobile tourist guides, only a few works describe automatic recommendation of tourist places based on interests and context. Ardissono et al. [1] describe their Intrigue system which presents a user with tourist information in the Turin, Italy region. They define heterogeneous tourist groups (such as families with children) and recommendations are made while taking possibly conflicting preferences into account. Preferences are given by the users themselves and reflect geographic features, essential information such as opening hours, basic information such as price, specific characteristics such as the historical period of an attraction, and properties such as historical value. In a conflicting group the preferences of individuals are weighted and compared to the properties of an activity to determine its rank.

Schwinger et al [8] do not present a ready to use system,

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

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

but study the strengths and weaknesses of several mobile tourist guides. They note that current systems tend to use their own selection of content data. This gives the developer more control over the presented information, but it also means that rich tourist-content websites are not used. Some systems adapt to the user's interests, but they require the user to provide these interests or at least explicit feedback on the points of interest.

Buriano [2] shares his views on the importance of social context in tourist activities. He notes that people enjoy sightseeing in groups and that they involve their social networks by sharing pictures for example. He suggests that these social relations should be included in recommender systems for tourist activities.

These works suggest that it would be wise to exploit the expertise of specialized websites. Also automatic personalization is an interesting approach, with the note that the social context should play a role as well.

In the contextual suggestion track, however, the user profiles were anonymous. We did not have any demographic information of the user, or information about the user's social situation. This limited our options. Therefore we have taken a content-based recommendation approach. We selected potential tourist activities from Google Places using the context information and re-rank these potential places to match the user's preferences. In section 2 we describe our recommendation approach. The results were evaluated in several ways, which is described in section 3, after which we finish with a discussion in section 4.

2. METHOD

Our method comprises 5 steps: (1) Collecting a first set of potential recommendations, (2) building the user profiles, (3) ranking the recommendations for the user profile, (4) re-ranking the list of recommendations, (5) filtering the recommendations using the temporal context. A more detailed description of these steps can be found in [7];

(1) Collecting potential recommendations

The first step was to collect potential recommendations for tourist places. We used the Google Places API for that purpose. Longitude and latitude of the location were used together with the keyword "tourist attractions" to retrieve relevant places. Short descriptions of the search results were obtained by querying the Google Custom Search API with the URL of the search result from Google Places.

(2) Building the Profiles

We described a user with two term profiles, one consisting of terms from the tourist activity examples that the user had judged as positive and the other consisted of terms from the tourist activity examples that the user had judged as negative. Terms from the title and description from the examples where put in the positive term profile if the *initial* rating was positive, and in the negative if the *initial* rating was negative. Terms from the categories, reviews and events from Google Places where put in the positive profile when the *final* rating was positive or in the negative if the *final* rating was negative. Terms with a neutral association were ignored. We did not use the content of a website, because the websites contained either too much noise (e.g. advertisement data) or we could not extract the content easily (flash content). Overall, this collection of terms results in the user profile $U = \{R_p, R_n\}$ in which R_p is the term frequency vector representation of the "positive" profile and R_n of the "negative" term profiles.

(3) Ranking recommendations

To rank the potential recommendations based on the user models we used two different methods: a similarity based method and a language modeling method.

In the similarity method, each term in the term profiles was weighted using the tf-idf measure [6] to determine the importance of each term in the profile.

We represented the potential tourist sight by a tf-idf term vector as well, based on its title, description, reviews and events. The fit of this potential recommendation was determined by taking the cosine similarity between the potential suggestion and the positive and negative profiles. The suggestions are ranked on their similarity scores. We order each items descending on their $cos_{positive}$ score. However, when $cos_{negative} > cos_{positive}$ we place the item at the bottom of the list (i.e. after the item with the lowest $cos_positive$ score, but with $cos_{positive} > cos_{negative}$). Originally, we discarded the items with a better fit to the negative profile than to the positive profile, but we needed them to be able to meet the number of requested recommendations (50 recommendations per person/context combination).

The alternative method we used to rank the potential recommendations was using a language modeling approach. In this variant the Kullback-Leibler divergence was used to weigh each term. We used point-wise Kullback-Leibler divergence [5], as suggested by [3]. It functions as a measure of term importance that indicates how important the term is to distinguish the "positive" examples from all examples.

A potential recommendation is better when it has many terms that are important in the "positive" examples. For each potential recommendation we derived its score by taking the sum of the Kullback-Leibler scores for the terms describing the search result. The potential recommendations were ordered descendingly on their scores. This approach benefits suggestions with more textual data, since the likelihood that it contains terms that also occur in the profiles is larger.

(4) Re-ranking the list of recommendations

During the development phase, we had no evaluation material. Therefore, we had to evaluate our methods manually. We created our own personal profile and we looked at which order of suggested activities appealed more to us.

We noticed in the suggestions given by the two runs, that famous tourist attractions did not rank very well. This is likely to be an artefact of the example data. For example, the Statue of Liberty does not resemble any of the examples in the tourist activity examples in Toronto, so it is no surprise that it does not receive a high rank. However, we believe that these famous sites should rank well. Therefore we use elements from the Google Places API to increase the rank of these items, independently of the user profiles.

We take an approach in which we created 4 ordered ranked lists: (A) Our personalized ranking based on KL-divergence or tf-idf; (B) a ranking based on the prominence of a place given by the original order of Google Places; (C) a ranking based on ratings of people that visited the place as indication of the overall perceived quality of a place; and (D) a ranking based on the a priori category likelihood. This latter ranking is based on the idea that some people have preferences for certain categories of activities (such as museums) rather than preferences for individual items. We derived the ranking from the Google categories and the times that this

category appeared in positive and negative examples. This final rating was smoothed (using +1 smoothing) to account for categories that did not occur in the example set. Since these were quite a lot and we did not want this to influence the results too much we weighted this rank half as much.

The final rank is determined by the weighted average rank of the search result in these 4 ordered lists. The weights we used were $\{1, 1, 1, 0.5\}$

(5) Filtering based on temporal context

In the last phase, we filter out the search results that do not match the temporal part of the given context using manually defined rules. We use the opening hours as registered in Google Places as reference material for determining whether a result matches the temporal context or not. For example, when the temporal context is evening, we do not suggest search results that have opening hours until 5pm.

(6) Presentation of the results

The first impression of a search result is very important for its relevance assessment by the user. However, some Google snippets contained advertisements or unclear descriptions. Therefore, we decided to use positive reviews as descriptions for the suggested places. Even though they might not always be good descriptors for the suggestion we hope that the positiveness may make people more inclined to give a positive rating.

3. RESULTS

In this section we present the accuracy and precision@5 results that we obtained with the two runs we submitted: (1) run01TI ranking based on tf-idf with cosine similarity and (2) run02K ranking based on point-wise Kullback-Leibler divergence scores. There were only 44 out of 1750 profile/context pairs taken into account during evaluation (i.e. not all contexts, and not all profiles were evaluated) and only the top 5 suggestions were evaluated. All results in this section are based on these 220 (i.e. $44 * 5$) datapoints.

Table 1: Precision @5 results for both runs and the –Google Places only– baseline

	Website*GeoTemporal	Description	Website
run01TI	0.19	0.42	0.40
run02K	0.22	0.41	0.47
baseline	0.18	0.30	0.41
	Geotemporal	Geo	Temporal
run01TI	0.54	0.89	0.56
run02K	0.57	0.90	0.58
baseline	0.51	0.79	0.57

Table 1 shows the precision results for the different measures, as well as a baseline (baselineA) provided by TREC, which is based on the original order of Google Places. To calculate precision, only items that have scored a rating of 2 (i.e good fit, or interesting) on each dimension are considered relevant. The results show that the differences between the tf-idf measure and the Kullback-leibler divergence measure are very small. Both measures seem to perform better than the baseline. Interestingly the geographical fit of this baseline is lower, which is likely caused by a different query method. The results of our runs show a particularly high precision at rank 5 for the geographical fit. The precision in terms of the rating on description and website shows room for improvement. Also the precision on the combination of personal ratings (e.g. website) and geo-temporal fit is not very high. However, the neutral items are interpreted as bad suggestions, making this measure quite conservative.

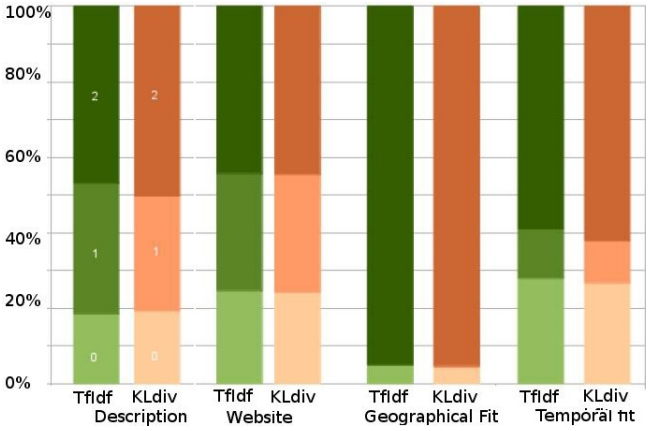

Figure 1: Distribution of positive (2), neutral (1) and negative (0) ratings

A more detailed look on the distribution of positive, neutral and negative ratings is given in figure 1. The two left-most columns of figure 1 show that approximately half of the suggestions are perceived as interesting (rating 2) when it comes to the opinion of the users. Many items (a third) are perceived as neutral (rating 1). This may mean that the user is not yet sure if he/she would want to follow up on the suggestion, in any case the user is not negative on the suggestion. Overall, around 80% (the sum of the 1 and 2 ratings) of the suggestions are perceived as positive when only the description is shown. When the website is shown the users are a little less positive.

The two right-most columns of figure 1 show a big difference between the accuracy of the suggestions in terms of the geographical fit to the context and the temporal fit to the context. The difference between the tf-idf measure and the Kullback-leibler divergence measure is again neglectable. 95% of the suggestions fit the geographical context.

The temporal context is matched in 62% of the suggestions. This leaves room for improvement. After inspection we noticed that theatres and night clubs tend to be suggested during the day as well. This is caused by the opening hours of the box office, which are usually in the afternoon and thus according to our algorithm a suitable suggestion for the afternoon context.

3.1 Impact of mixing rank-methods

The impact of each of the ranking methods on the final ranking was assessed using Kendall's τ [4].

Table 2: correlations between the ranking methods (A,B,C,D) and the final ranking (Kendall's τ)

	with Final Ranking
(A) Tf-idf	0.59
(A) KL-divergence	0.56
(C) Ratings from other people	0.36
(D) A-priori category likelihood	0.20
(B) Place Prominence	0.17

Table 2 shows the average rank correlations (Kendall's τ) with the final ranking for the various ranking methods from section 2. Overall we see that the rankings based on the user profiles (by either KL-divergence or tf-idf) are correlated the most with the final ranking. The prominece of a place (based

on the original Google Places order) has the least influence on the final ranking.

The tf-idf measure and the Kullback-Leibler measure show a correlation with each other of $\tau = 0.47$, showing that the methods are actually quite similar in the proposed order of suggestions, even though the actual ranks may vary. Also both methods are slightly correlated with rankings based on ratings from other people ($\tau = 0.17$ for KL-divergence and $\tau = 0.21$ for tf-idf).

4. DISCUSSION

We encountered a number of challenges in the implementation of our approach. First, it was difficult to obtain 50 suggestions for each context. This was mainly because of the limitations of the Google Places API. However, since only the top 5 suggestions were evaluated this did not have an effect on our results.

A second problem was the little variation between suggestions for one person and the other. This was a result of a high similarity between user profiles, which was caused by the limited example set. Each individual rated the same example places and they tended to be very positive about them as well. The rating may be positively biased, since the training examples were places from the area of residence of the users. It is possible that when rating places that you are familiar with, you have other preferences then when it comes to places that you have not visited before.

In general, it is still a point for debate how much the influence of personal characteristics should be when suggesting tourist sights. After all, people often go to the main points of interest when they visit a city anyway. It is important that these are part of the suggestions. But, when a person visits the place for a second time, personal characteristics might be more important, since the person has likely visited the main points of interests already. For some types of suggestions, e.g. places to eat, personal characteristics are likely to be more important than for other types of suggestions. This would be an interesting point for future research.

Most other teams used a similar method for collecting search results. Some groups included more specialized search engines such as Yelp. Many teams used a recommender system based approach in which search results were collected first, and ranked according to their match to term profiles, although a few teams took an approach in which a query was generated based on the user's preferences. Some teams used the terms from examples, others focused more on conceptualizing examples by recognizing categories from them.

There were two teams in the top 5 results using tf-idf weighting with cosine similarity to calculate the match between profile and search results, while our tf-idf run was at position 11. These two teams did not mix their results with other rankings like we did, used different descriptions and also had a slightly different approach in acquiring search results. Our runs both performed better than average and median and even had the best performance for a few of the contexts.

5. CONCLUSION

We think we have several strong points in our approach. Overall it is attractive that our approach is completely automated. Our suggested places matched the geographical contexts very well. This is because we used search results from Google Places, which allowed us to use precise location information in the search query. However, even though opening hours were provided by Google Places as well, it was more difficult to obtain a good fit on the temporal context, because these hours were sometimes erroneous but also because not everybody had the same interpretation of the categorical values of the temporal context.

Secondly, we think it is attractive to mix several ranking methods. This way we could find a balance between personalized suggestions and more generic famous places suggestion. Additionally, we could use the opinion of people that have visited the sight already. Our analysis of the rank correlations for the ranking methods show that the personalized ranking method (either by tf-idf or KL-divergence) had the most impact on the final ranking. Interestingly, both the tf-idf measure and KL-divergence measure rankings correlated slightly with rankings based on the ratings from other people. This means that a personal measure gives to some extent the same ranking order as a collective measure based on ratings by many people.

And finally, we think the use of reviews as a description for the search result is attractive, since it gives a personal touch to the suggestion even though the descriptions are not personalized. A positive review may influence people, making them more enthusiastic about the suggestion. Overall, people responded a little better to our descriptions than to the website (see table 1).

We could make some improvements by investigating the influence of the keyword that is used to collect potential places. Additionally, the weights of the 4 ranking methods could be optimized, once there is more data available.

More generally speaking, the TREC contextual suggestion track provides a platform to evaluate the "zero query term problem" in which the search engine can pro-actively suggest resources given a context. In the future this can be expanded with context types, other than geo-temporal context, such as social context, or content context.

6. ACKNOWLEDGEMENTS

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

7. REFERENCES

[1] L. Ardissono, A. Goy, G. Petrone, M. Segnan, and P. Torasso. Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence*, 17(8-9):687–714, 2003.

[2] L. Buriano. Exploiting social context information in context-aware mobile tourism guides. *Proc. of Mobile Guide 2006*, 2006.

[3] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, 2001.

[4] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[5] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[6] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[7] M. Sappelli, S. Verberne, and W. Kraaij. TNO and RUN at the TREC 2012 contextual suggestion track: Recommending personalized touristic sights using google places. In *21st Text REtrieval Conference Notebook Proceedings (TREC 2012)*, 2013.

[8] W. Schwinger, C. Grün, B. Pröll, W. Retschitzegger, and A. Schauerhuber. Context-awareness in mobile tourism guides–a comprehensive survey. *Rapport Technique. Johannes Kepler University Linz*, 2005.