# A Nearest Neighbor Approach to Contextual Suggestion


Sandeep Avula
School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
asandeep@live.unc.edu

John O'Connor
School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
oconnorv@live.unc.edu

Jaime Arguello
School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
jarguello@unc.edu



## ABSTRACT

The School of Information and Library Science at the University of North Carolina at Chapel Hill (UNCSILS) submitted two runs to the Contextual Suggestion Track. Given a geographical context, both our runs (UNCSILS_BASE and UNCSILS_PARAM) scored venues from the same candidate set gathered using the Yelp API. Our baseline run (UNCSILS_BASE) followed a nearest neighbor approach. For a given profile/context pair, the candidate venues were scored using the weighted average rating associated with the venues in the profile. The weighting was implemented based on the cosine similarity between the candidate venue and the profile venue using TF.IDF term weighting. The goal of this approach was to score each candidate venue based on the rating associated with the most similar venues in the profile. Our experimental run (UNCSILS_PARAM) boosted the contribution from the profile venue with the greatest similarity with the candidate venue and rating. The experimental run (UNCSILS_PARAM) outperformed the baseline run (UNCSILS_BASE) by a small, but statistically significant margin.


## 1. INTRODUCTION

Contextual suggestion is the task of recommending venues (places to visit) for a particular person in a geographical context. The recommendations can be informed by ratings provided by the same person in a different geographical context, which is referred to as the individual's *profile*. For TREC 2013, each profile consisted of ratings for 50 sample venues from a seed city. The seed city this year was Philadelphia, PA. Each person rated the same 50 venues from Philadelphia on a scale of 0-4, where 0 indicated no interest in that venue and 4 indicated great interest. Because each person's profile is different, venue recommendations for each individual in a new city should be relatively unique. Track participants were required to submit a ranking of venues for each profile/context pair. There were a total of 500 profiles and 50 contexts. The contexts were US cities of various sizes—the largest metro area being Washington, DC and the smallest being Lewiston, ID. Participants could submit rankings of up to 50 recommendations for each profile/context pair. Ultimately, 223 profile/context pairs were judged.

## 2. CORPUS

One necessary step in contextual suggestion is to identify a set of candidate venues for a given geographical context. This year, participants could either make use of the full ClueWeb12 collection[1], a subset of ClueWeb12 created specifically for the Contextual Suggestion Track[2], or the open web. We identified candidate venues from the open web using the Yelp API[3], Yelp website[4], and individual venue web pages.

Each context was defined by a latitude/longitude coordinate. We found that using a single coordinate yielded too few results from the Yelp API. In order to augment the single coordinate, we used the GeoLite City database created by Maxmind.[5] We expanded the given coordinate using coordinates from the GeoLite City database associated with the same metropolitan area (metroCode) and the same area code (areaCode). This allowed us to expand the region from which we pulled venues using the Yelp API. Duplicate venues were discarded.

The Yelp API only returned the URL for the Yelp listing of a particular venue. We needed the venue's own URL in order to submit to TREC. Taking the Corcoran Gallery of Art as an example, we needed http://www.corcoran.org instead of http://www.yelp.com/biz/the-corcoran-gallery-of-art-washington. In order to retrieve the venue's URL, we crawled the Yelp webpage for each venue and scraped the venue's URL from the HTML. If the venue did not have its own URL, we discarded it from potential use.

Each venue was also required to have a description associated with it for submission to the Track. For every remaining candidate venue, we crawled its website for the HTML description and used that as our description. If a website did not have a description, we simply repeated the name of the venue as the description.

## 3. ALGORITHMS

Our end goal is to output a ranking of venues for an input user $u$ and context $c$. In general, our approach is to score candidate venues based on the ratings associated to the most similar profile venues (in a text-similarity sense). This resembles a *weighted* $k$-nearest neighbor approach using the entire user profile as the nearest neighbors (i.e., $k = 50$).

Let $\mathcal{V}_p$ denote the set of 50 venues used to construct each user profile and $\mathcal{V}_c$ denote the set of candidate venues pulled from context $c$. Our baseline approach (UNCSILS_BASE)

[1] lemurproject.org/clueweb12/
[2] lemurproject.org/clueweb12/related-data.php
[3] www.yelp.com/developers/documentation
[4] www.yelp.com/
[5] dev.maxmind.com/geoip/legacy/geolite/

scores candidate venue $v_c$ for user $u$ according to,

$$\mathcal{S}_{base}(v_c, u) = \frac{1}{|\mathcal{V}_p|} \sum_{v_p \in \mathcal{V}_p} \phi_{sim}(v_c, v_p) \times rating(u, v_p), \quad (1)$$

where $\phi_{sim}(v_c, v_p)$ refers to the normalized cosine similarity between candidate venue $v_c$ and profile venue $v_p$ and $rating(u, v_p)$ refers to the rating (0-4) assigned by user $u$ to profile venue $v_p \in \mathcal{V}_p$. Cosine similarities were computed using tf.idf term vectors and normalized to zero minimum and unit maximum using the minimum and maximum cosine similarities between all candidate-venue/profile-venue pairs.

In addition to our baseline run, we also submitted an experimental run (UNCSILS_PARAM) which boosted the contribution from the profile venue with the greatest combination of similarity to the candidate venue and rating. This experimental run scores candidate venue $v_c$ for user $u$ according to,

$$\mathcal{S}_{param}(v_c, u) = \lambda \mathcal{S}_{max}(v_c, u) + (1 - \lambda) \mathcal{S}_{base}(v_c, u), \quad (2)$$

where,

$$\mathcal{S}_{max}(v_c, u) = \text{MAX}_{v_p \in \mathcal{V}_p} \Big( \phi_{sim}(v_c, v_p) \times rating(u, v_p) \Big). \quad (3)$$

Equation 2 corresponds to the linear combination of our baseline approach and the score given by the profile venue with the greatest combination of similarity and rating.

## 4. PARAMETER TUNING

Parameter $\lambda$ was tuned by splitting the 50 profile venues ($\mathcal{V}_p$) into two sets of 30 (for training) and 20 (for testing). For each user, the set of 30 venues was used as a truncated user profile and the held-out set of 20 was used as the candidate venues to be ranked. Parameter $\lambda$ was tuned by maximizing average $\mathcal{P}$@5 across all users. Venues with a rating of 3-4 were considered "relevant" and those with a rating of 2 or lower were considered "non-relevant". Parameter $\lambda$ was tuned across values ranging from 0.0 to 1.0 in increments of 0.1. The optimal value was found to be 0.4, which places slightly more weight on the score given by our baseline approach. We admit that this is a slightly different formulation of the contextual suggestion task. In reality, we would expect the profile and candidate venues to come from *different* geographical contexts. However, we believe this was a reasonable way of tuning $\lambda$.

## 5. RESULTS

Results for both our runs (UNCSILS_BASE are UNCSILS_PARAM) are presented in Table 1 in terms of precision at rank 5 ($\mathcal{P}$@5), Mean Reciprocal Rank (MRR) and Time-based Gain (TBG). Time-based Gain was originally proposed by Smucker & Clark [2] and motivated as an appropriate evaluation metric for contextual suggestion in Dean-Hall *et al.* [1]. For all metrics, we present average performance across all 223 profile/context pairs that were judged. UNCSILS_PARAM improved upon UNCSILS_BASE in terms of all three metrics, although the improvement was statistically significant only in terms of TBG.

**Table 1: Evaluation results. A ▲ denotes a statistical significant improvement over UNCSILS_BASE at the $p < .05$ level.**

| | $\mathcal{P}$@5 | MRR | TBG |
|---|---|---|---|
| UNCSILS_BASE | 0.257 | 0.414 | 1.137 |
| UNCSILS_PARAM | 0.278 | 0.427 | 1.311▲ |

Figure 1 illustrates our performance for each profile/context pair compared to the median and best performance achieved for that profile/context pair. In terms of $\mathcal{P}$@5, MRR, and TBG, UNCSILS_BASE performed equal to or greater than the median 78.0%, 68.6%, and 66.4% of the time, and equal to the best 9.9%, 32.7%, and 5.4% of the time. Similarly, in terms of $\mathcal{P}$@5, MRR, and TBG, UNCSILS_PARAM performed equal to or greater than the median 78.0%, 69.1%, and 69.1% of the time, and equal to the best 11.7%, 34.1%, and 9.9% of the time. Consistent with the results in Table 1, UNCSILS_PARAM improved over UNCSILS_BASE mostly in terms of TBG.

## 6. DISCUSSION

The main obstacle in implementing both of our runs was to gather a large-enough set of candidate venues for every given geographical context. Several contexts yielded only about 20 venues from the Yelp API. These mostly corresponded to rural communities. Yelp relies on users to add and update venue information. This is a great model for large urban areas with many attractions. However, for rural areas, where most of the population already knows about all the local venues, Yelp is less useful and therefore less up to date. In order to have a large enough corpus for each venue, we were required to expand the radius from which we pulled venues. This may have resulted in a loss of contextual relevance.

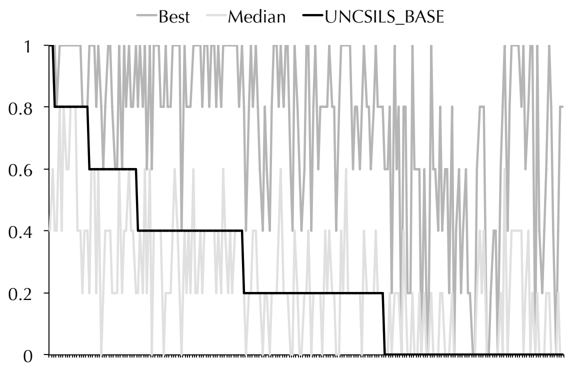
## 7. CONCLUSION

This was the first year that the School of Information and Library Science at University of North Carolina at Chapel Hill participated in the Contextual Suggestion Track. Our group (UNCSILS) submitted two runs. Both runs scored candidate venues gathered from the open web using the Yelp API. In cases where the Yelp API returned fewer than 50 results, we expanded the geographical context using nearby coordinates provided by the GeoLite City database. Our baseline run (UNCSILS_BASE) scored candidate venues using a weighted nearest-neighbor approach. Candidate venues were scored based on the ratings given to the most similar profile venues. Our experimental run (UNCSILS_PARAM) combined our baseline approach with the score given by the profile venue with the greatest similarity and rating. Our experimental run outperformed our baseline run across all metrics, but was only significantly better in terms of TBG.

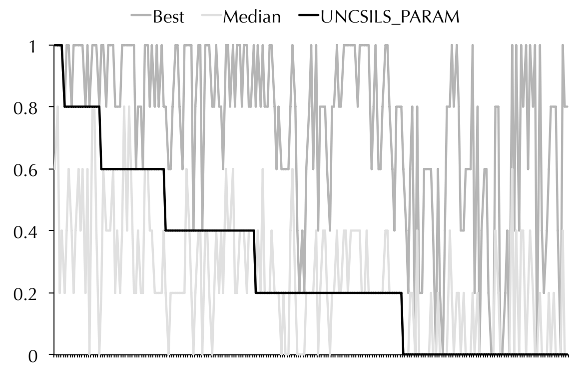
## 8. REFERENCES


[1] D.-H. Adriel, C. Clark, J. Kamps, and P. Thomas. Evaluating contextual suggestion. In *EVIA '13*, 2013.

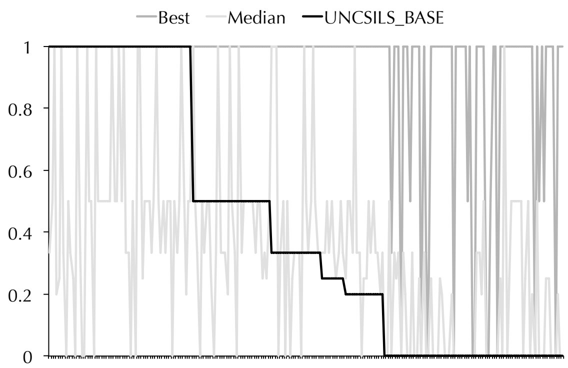
[2] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *SIGIR '12*, pages 95–104. ACM, 2012.



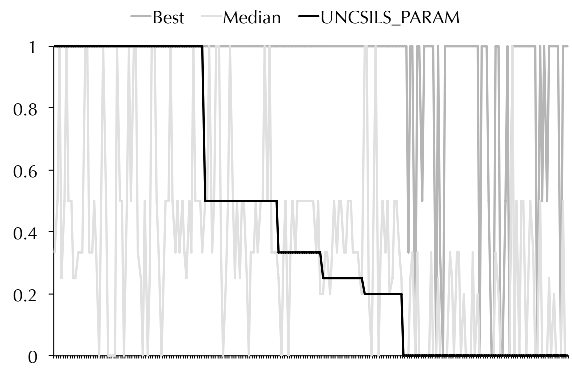

(a) P@5



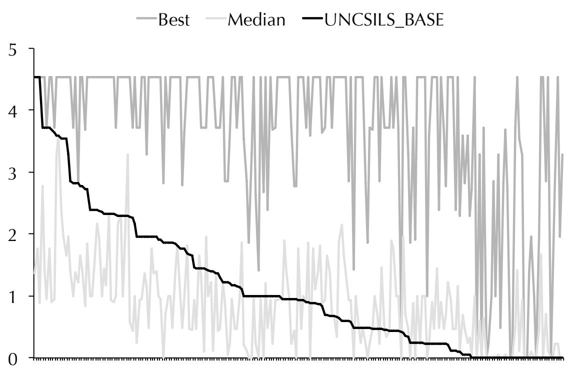

(b) P@5



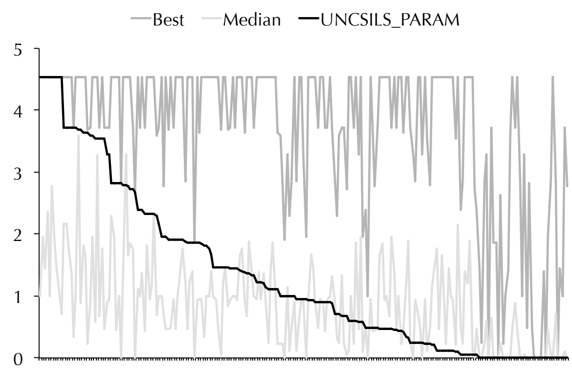

(c) MRR




(d) MRR




(e) TBG




(f) TBG

**Figure 1: Results for UNCSILS_BASE (left) and UNCSILS_PARAM (right) in terms of P@5 (top), MRR (middle), and TBG (bottom).**