Overview of the TREC 2013 Contextual Suggestion Track

Adriel Dean-Hall
University of Waterloo

Charles L. A. Clarke
University of Waterloo

Jaap Kamps
University of Amsterdam

Paul Thomas
CSIRO

Nicole Simone
University of Waterloo

Ellen Voorhees
NIST

1 Introduction

1.1 Summary for Previous Participants

For participants familiar with the 2012 Contextual Suggestion Track we have provided a list of the main changes to this year's track:

- Contexts no longer include a temporal component (day of week, time of day, and season), contexts consist of only a location.
- Users were recruited from a crowdsourcing service (Mechanical Turk) as well as from the University of Waterloo student body.
- Suggestion attractiveness judgements are given on a 5-point, rather than a 3-point, scale.
- Submissions based off of the ClueWeb12 corpus were allowed in addition to submissions based off of the open web.
- A modified Time-Biased Gain (TBG) metric was used in addition to P@5 and MRR. This metric is described in section 4.3.
- The option to submit solely based on context or solely based or user profiles was removed.
- The file format used for profiles, contexts, and suggestions was switched from XML to CSV and JSON.

If you are already familiar with this track you can skip to section 5 which provides an overview of the approaches used by participants and section 6 which contains the results.

1.2 Task Description

The contextual suggestion track investigates search techniques for complex information needs that are highly dependent on context and user interests. For example, imagine an information retrieval researcher with a November evening to spend in Gaithersburg, Maryland. A contextual suggestion system might recommend a beer at the Dogfish Head Alehouse, dinner at the Flaming Pit, or even a trip into Washington on the metro to see the National Mall. The primary goal of this track is to develop evaluation methodologies for such systems.

This track ran for the second time as part of TREC 2013 after a positive response at TREC 2012. This year participants were again given, as input, a set of user profiles and set of geographical contexts. The task was to take these profiles and contexts and to produce a list of up to 50 ranked suggestions for each profile-context pair. Participants could choose to gather suggestions from either the open web or the ClueWeb12 dataset.

Each profile corresponds to a single user and indicates that user's preference with respect to each sample suggestion. For example, if one sample suggestion is a beer at the Dogfish Head Alehouse, the profile might indicate a negative preference to that suggestion. Each suggestion includes a title, short description, and an

associated URL. Each context corresponds to a particular location at the granularity of a city. For example, a context might be Gaithersburg, Maryland.

As with last year each groups was allowed to submit up to two runs. A total of 19 groups submitting 34 runs participated in the track this year, an increase by 7 runs from last year. This included two baseline runs submitted by the track organizers which are described later in this report in section 3.3. 7 of these runs comprised suggestions from the ClueWeb12 dataset, the other 27 runs comprised suggestions from the open web.

2 Detailed Task Description

Profiles and contexts were distributed to participants as CSV and JSON files. For this track we generated 562 profiles and 50 contexts, below we describe how these were generated. An experimental run consists of a single suggestion (CSV) file generated automatically from the profile and context files.

2.1 Profiles

Profiles indicate a user's preferences to a list of 50 example suggestions within Philadelphia, PA. These profiles are built by conducted a survey advertised to both University of Waterloo students and Mechanical Turk users.

Profiles are split into two files: examples2013.csv and profiles2013.csv[1]. examples2013.csv contains a list of 50 suggestions which each consist of an id, a title, a description, and a url. Below are two suggestions from the file:

- **ID** 51
  Title Elfreths Alley Museum
  Description Elfreths Alley Museum is a reputable museum. A lovely little piece of history. Definitely a must while visiting Philadelphia... To walk down the oldest residential street in the country is just something I think everyone should do at least once if in the area! I really enjoyed it.
  URL `http://www.elfrethsalley.org`
- **ID** 65
  Title Red Mango
  Description Red Mango is committed to providing the healthiest and best tasting all-natural nonfat frozen yogurt and fresh fruit smoothies. No wonder Zagat ranked us #1, twice.
  URL `http://www.redmangousa.com`

The second file contains a list of ratings for each suggestion in examples2013.csv given by each user, below are a few example lines from profiles2013.csv:

```
...
534,51,1,2
534,52,4,4
534,53,2,1
...
534,64,1,4
534,65,4,4
534,66,2,3
...
```

Listing 1: An excerpt from profiles2013.csv.

[1]These two files are also distributed as JSON files.

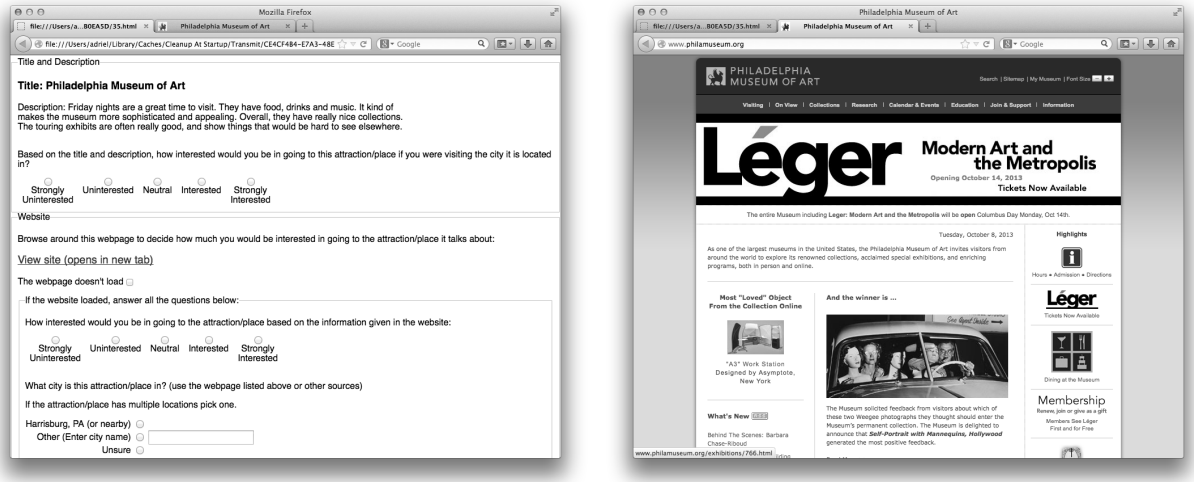


Figure 1: Screenshots of survey seen by users.

The first line means that user id 534 gave example suggestion number 51 a description rating of 1 and a website rating of 2.

2.1.1 Generating Example Suggestions

First we need to generate example suggestion which will rated by users in a survey. A suggestion consists of a title, short description, and a website URL. 50 example suggestions were taken from attractions submitted by participants in the TREC 2012 version of this track. These example suggestions were all from the Philadelphia, PA area (one of the contexts used in 2012). Example suggestions were chosen if they had a high quality description and a website that was available; example suggestions were also chosen so that there was diversity in the types of attractions in our set.

2.1.2 Gathering Attraction Preferences

Profiles distributed to participants indicated users' preference towards the example suggestions. In order to form the profile users, recruited from Mechanical Turk and the University of Waterloo, where asked to complete an online survey. In the survey sample suggestions were presented to users in a random order. Users were asked to give two 5-point ratings for each attraction, one for how interesting the attraction seemed to the user based on its description and one for how interesting the attraction seemed to the user based on its website. The survey interface, as presented to University of Waterloo users, can be seen in figure 1. Mechanical Turk users saw a similar interface but it also included some Mechanical Turk interface elements that were not controlled by us.

In total 62 Waterloo students and 500 Mechanical Turk users responded to the survey.

2.2 Contexts

Contexts describe which city a user is currently located in. There were 50 cities chosen randomly from the list of primary cities in metropolitan areas in the United States (which are not part of a larger metropolitan area) excluding Philadelphia, PA (the seed city). The list of metropolitan areas was taken from Wikipedia[2].

[2] `http://en.wikipedia.org/wiki/List_of_metropolitan_areas_of_the_United_States`

Contexts are distributed to participants in the file contexts2013.csv (as with the profile files, a JSON file with the contexts is also distributed).

```
...
71,Monroe,LA,32.81513,-92.20569
72,Tampa,FL,27.94752,-82.45843
...
78,Lewiston,ID,46.41655,-117.01766
79,Lima,OH,40.74255,-84.10523
...
```

Listing 2: An excerpt from contexts2013.csv.

Here the first line means that context number 71 represents Monroe (city), LA (state) with a latitude of 32.81513 and a longitude of -92.20569. For contexts the latitude and longitude are provided as a convenience and are synonymous with the city and are not meant to represent the exact position of the user. Contexts represent locations at the granularity of a city-level.

2.3 Collections

Participants were able to gather suggestions from either the open web, ClueWeb12[3], ClueWeb12 B13, or ClueWeb12 CS. ClueWeb12 and ClueWeb12 B13 are datasets prepared by Jamie Callan's research group at CMU. ClueWeb12 CS was prepared for the track by the track organizers.

The ClueWeb12 CS subcollection was created by issuing a variety of queries for each context location against a commercial search entire. Returned results that had URLs which matched documents in ClueWeb12 were grouped by context and included in the subcollection. URLs were normalized before they were matched, for example forward-slashes were removed from the end of URLs. In total the subcollection contains 30 144 documents.

2.4 Submitted Suggestions

Each submitted run consists of up to 50 ranked suggestions for each profile-context pair. Similarly to the example suggestions, profiles consist of a title, description, and URL that correspond to an attraction. The URL can be substituted with a ClueWeb12 DocID. Suggestions also contain a group id, run id, profile id, context id, and rank.

In order to generate suggestions participants were allowed to use whatever resources they wished to use, for example review websites such as Yelp. The goal was that each suggestion should be tailored to the profile and located within the context that was being targeted. Ideally, the description of the suggestion would be tailored to reflect the preferences of the user.

Here are two of the suggestions we received:

- **Group ID** udel_fang
  Run ID UDInfoCS1
  Profile ID 534
  Context ID 71
  Rank 1
  Title Waterfront Grill
  Description Waterfront Grill is a seafood restaurant. HERE ARE REVIEWS FROM OTHER PEOPLE:Great views of the water while dining on great real food. There is a perfect view of the bayou, and on a nice day, you can sit on the patio. As with any place, there are good days and bad days but overall, they possess a wonderful track record.

[3] http://lemurproject.org/clueweb12/

URL/Doc ID `http://www.yelp.com/biz/waterfront-grill-monroe`

- **Group ID** udel_fang
  Run ID UDInfoCS1
  Profile ID 534
  Context ID 78
  Rank 1
  Title M J Barleyhoppers
  Description M J Barleyhoppers is an american (traditional) restaurant. HERE ARE THE DESCRIPTIONS FROM ITS WEB SITE:Lewiston's red lion hotel is the home of m.j. barleyhopper's, the area's largest microbrewery. HERE ARE REVIEWS FROM OTHER PEOPLE:I've only been here a few times but i'll be back many more. This little brewpub has the best beers i know of that you can find in the valley. THIS PLACE IS SIMILAR TO OTHER PLACE(S) YOU LIKED, i.e. Fergie's Pub.
  URL/Doc ID `http://www.redlionlewiston.com/mj_barleyhoppers.htm`

3 Judging

Judging was split up into two tasks. Suggestions were judged with respect to their profile relevance by users and with respect to the contextual relevance by accessors at NIST as well as users.

3.1 Profile Relevance

In order to judge the relevance of suggestions with respect to profiles a second survey was conducted, which was similar to the first one. Some users were invited back to give ratings for the attraction descriptions and websites of the top 5 ranked suggestions for each run for their profile and one or two randomly chosen contexts.

The judgements given were one of:

-2 Could not load
0 Strongly uninterested
1 Uninterested
2 Neutral
3 Interested
4 Strongly interested

Some users were not invited back to give judgements on suggestions. While completing both surveys users were asked whether suggestions were geographically appropriate and the amount of time user took to make judgements was also recorded. In the initial survey used to generate profiles 5 suggestions **not** in Philadelphia, PA were included with the 50 suggestions in Philadelphia, PA. A score was generated for each user after the first survey that was based on how long users took to make judgements and how many geographical judgements users got correct. If users took too little time in making judgements or got too many geographical judgements incorrect they were not invited back.

Approximately the top 80% of users were invited back. Some users did not respond to our invitation to the second survey. In total 223 context-profile pairs were judged by users.

Judgements of relevance of suggestions with respect to profiles are distributed in desc-doc.qrels.

```
...
UDInfoCS1 534 71 http://www.yelp.com/biz/cotton-monroe 2 3 31 13
UDInfoCS1 534 71 http://www.yelp.com/biz/waterfront-grill-monroe 3 4 -1 108
```

```
UDInfoCS1 534 78 http://www.redlionlewiston.com/mj_barleyhoppers.htm 3 3 10 6
...
```

Listing 3: An excerpt from desc-doc.qrels.

Here the first line means that the user was neither interested nor uninterested (neutral, 2) in the attraction based on the description provided by run UDInfoCS1 for profile 534, context 71, and the website `http://www.yelp.com/biz/cotton-monroe` but the user was interested (3) in the attraction based on the content of the website. The last two numbers mean that the user took 31 sec. to rate the description and 13 sec. to rate the website. A -1 means that no timing data is available. This timing data is not used as part of the scoring calculations for runs.

3.2 Geographical Relevance

In order to judge the geographical relevance of suggestions users were asked, during the survey, whether the attraction was in the city it was submitted for or not. Additionally accessors at NIST where also asked to make the same judgement for attractions. The list of context-profile pairs judged by users and those judged by NIST were not the same list however there was an overlap of approximately nine thousand judgements. Of the documents judged for context by both NIST accessors and users there was an agreement on judgements of 77% if judgements of "marginally appropriate" and "appropriate" are considered the same.

-2 Could not load
0 Not geographically appropriate
1 Marginally geographically appropriate
2 Geographically appropriate

Note that only NIST accessors explicitly made judgements of 1, users made judgements of either -2, 0, or 2, however some of the user judgements are reported as 1 when users didn't agree with each other on whether an attraction was geographically appropriate. For purposes of calculating final metric scores if both NIST accessors and users disagree on whether a suggestion is contextually appropriate the value the NIST accessors gave is taken.

Judgements of geographical appropriateness are distributed in geo-nist.qrels and geo-user.qrels for NIST assessments and user assessments respectively.

```
...
71 http://www.yelp.com/biz/waterfront-grill-monroe 2
71 http://yelp.com/biz/la-perla-3-mexican-restaurant-and-grocery-store-johnson-city 0
...
78 http://www.redlionlewiston.com/meriwethers_american_grill.htm 2
78 http://www.redlionlewiston.com/mj_barleyhoppers.htm 2
...
```

Listing 4: An excerpt from geo-nist.qrels.

Here the first line means that for context 71 the website `http://www.yelp.com/biz/waterfront-grill-monroe` is geographically appropriate (2).

3.3 Baseline Runs

Two baseline runs were submitted. BaselineA takes the top 50 attractions returned by the Google Places API when provided with the city in the context. For the description, a Google Places provided description, review, or a blurb from the meta-description tag on the website is used. BaselineB uses the same strategy except that suggestions not in ClueWeb12 were filtered out and the remaining suggestion URLs were mapped to

ClueWeb12 document ids. Exact URL matches were not needed, for filtering suggestion URLs had forward-slashes removed from the end of them before being matched to ClueWeb12 documents. Personalization was not attempted for either baseline run.

4 Measures

Three measures are used to rank runs. Our main measure, Precision at Rank 5 (P@5), is supplemented by Mean Reciprocal Rank (MRR) and a modified version of Time-Biased Gain (TBG)[1].

4.1 P@5

An attraction is considered relevant for P@5 if it has a geographical relevance of 1 or 2 and if the user reported that both the description and document were found to be interesting (3) or strongly interesting (4). A P@5 score for a particular topic (a profile-context pair) is determined by how many of the top 5 ranked attractions are relevant, divided by 5.

4.2 MRR

For MRR, an attraction is considered relevant using the same criteria used for P@5. A MRR score is calculated as $\frac{1}{k}$, where k is the rank of the first relevant attraction found. If there are no relevant attractions in the first 5 attractions in the ranked list a score of 0 is given.

4.3 TBG

In an effort to develop a metric better suited to evaluating this task the organizers of this track developed a metric based on TBG metric introduced by Smucker and Clarke[2]. The modified version of TBG is calculated by the equation described by Dean-Hall, et al.[1]:

$$\sum_{k=1}^{5} D(T(k))A(k)(1-\Theta)^{\sum_{j=1}^{k-1} Z(j)}$$

- D is a decay function.
- T(k) is how long it took the user to reach rank k, calculated using the following two rules:
  - The user reads every description which takes time T_{desc}.
  - If the description judgement is 2 or above then the user reads the document which takes time T_{doc}.
- A(k) is 1 if the user gives a judgement of 2 or above to the description and 3 or above to the document, otherwise it is 0.
- Z(k) is 1 if the user gives a judgement of 1 or below to either the description or the document, otherwise it is 0.

Note that, for this metric, the user always gives a rating of 0 to the document if the document has a geographical rating of 0. The four parameters for this metric are taken from Dean-Hall et al. [1]: $\Theta = 0.5$, $T_{desc} = 7.45s$, and $T_{doc} = 8.49s$, and the half-life for the decay function $H = 224$.

5 Participant Approaches

There were 34 runs submitted by 19 groups, 7 of these were ClueWeb12 runs and 27 were open web runs. 14 groups provided descriptions of their runs which are included below. One of the groups consisted of the two baseline runs which are described above in section 3.3.

5.1 PITT at TREC 2013 Contextual Suggestion Track

Rundids: ming_1, ming_2

Authors: Ming Jiang and Daqing He

This system made used of data from Yelp for creating candidate suggestion and supplementing user profiles. The system used vector space models to compute the similarity between candidates and examples and linear regression models to combine multiple attributes of candidate profiles into the calculations. The system was trained and tested using 5-fold cross validation on 2012 track data.

5.2 An Opinion-aware Approach to Contextual Suggestion

Runids: UDInfoCS1, UDInfoCS2

Authors: Peilin Yang and Hui Fang

This system set out to evaluate the effectiveness of (1) an opinion-based method to model user profiles and rank candidate suggestions; and (2) a template-based summarization method that leverages the information from multiple resources to generate the description of candidate suggestion.

Given a user and context pair, this system gathered candidate suggestions from Yelp and then ranked the candidate suggestions based on their similarity with the user profile. This system estimated the user profile based on the reviews of the candidate suggestions in contrast to using the description or category information of the suggestions.

5.3 Boosting Venue Page Positions for Contextual Retrieval InfoSense at TREC 2013 Contextual Suggestion

Runids: BOW_V17, BOW_V18

Authors: Jiyun Luo and Grace Hui Yang

This system makes ClueWeb12 suggestions and uses two main approaches. The first approach extracts venue names from WikiTravel and then formulates queries to search the collection. The similarity between venue name and anchor text for pages is used to locate the most relevant URL for the venue, as opposed to non-relevant documents, for example "yellow-page"-like list pages. The second approach divides venues into a two-level venue categorization, for example "landmark" or "amusement park", then the system creates a language model for each category. The category-specific language models are used to perform the retrieval for each individual category mentioned in a user's profile.

Suggestions are personalized by making distinctions between "major", "minor", and "negative" personal interests. The system creates ranked suggestions lists by merging venues from multiple categories while favouring venues that the user has a "major" personal interest in.

5.4 Exploiting Location-based Social Networks for Contextual Recommendations

Runids: uogTrCFX, uogTrCFP

Authors: M-Dyaa Albakour, Nut Limsopatham, Craig Macdonald, and Iadh Ounis

This system uses Location-based Social Networks (LSNs), such as FourSquare and Facebook Places, to gather data on both venues and users. First, the similarity between the venue descriptions and a textual representation of the user's interests is calculated. Venue descriptions are extracted from their web pages or their profiles on LSNs. In the uogTrCFP run, on top of this similarity, the system focuses on using the social

aspect in the ranking by incorporating an estimation of the popularity of the venue based on the number of previous interactions of the users on LSNs. The second run, uogTrCFX, also uses this popularity but focuses on using a personalisation model based on the XQuAD diversity framework, which allows the system to cover multiple categories of interest to the user.

5.5 DUTH Team

Runids: DuTH_A and DuTH_B

Authors: George Drosatos, Giorgos Stamatelatos, Avi Arampatzis and Pavlos S. Efraimidis

This system collects attractions from three commercial search engines, Google Places, Foursquare, and Yelp. The attractions returned by these services are enhanced by adding snippets from the Google and Bing search engines using crowdsourcing techniques. The first run submits each candidate place as a query in an index of examples and scores it based on the top-k users' preferences. The second run is based on Rocchio's algorithm and uses the examples per profile to generate a personal query which is then submitted to the index of attractions.

5.6 CWI team

Runid: IBCosTop1

Authors: Thaer Samar, Arjen de Vries, Alejandro Bellogin, Jimmy Lin, and Alan Said

This system uses the full ClueWeb12 dataset as a source. For each context a sub-collection of the most relevant documents is formed. Once this is done the system then ranks these subcollections based on the profiles. User profiles are formed by taking user preferences for the sample attractions and descriptions for those attractions. Personalized rankings are generated by computing the cosine similarity between the document and these user profiles.

5.7 University of Lugano at the TREC 2013 Contextual Suggestion Track

Runids: complexScore, simpleScore

Authors: Andrey Rikitianskiy, Morgan Harvey, and Fabio Crestani

This system uses the Google Places API to obtain an initial list of suggestions. These were grouped into 27 different types. Description snippets were generated using the Yandex Rich Content API and Google Custom Search APIs. For each user a positive and negative model is generated, these models were based on the example descriptions which the system then expanded. These models were then used to rank suggestions lists for each profile-context pair.

5.8 A Simple Context Dependent Suggestion System

Runid: isirun

Authors: Dwaipayan Roy, Ayan Bandyopadhyay, and Mandar Mitra

This system groups attractions into categories and determines which attraction categories users prefer by using the attraction ratings given in the user profiles. Suggestions are fetched using the Google Places API, passing the context location as a parameter. These suggestions are then ranked based on the distance between the suggestion and the context location and the level of preference the user has for the suggestion category. Suggestion descriptions are result snippets returned from Google when the suggestion named is passed as a query.

5.9 IRIT Team

Runids: IRIT.ClueWeb, IRIT.OpenWeb

Authors: Guillaume Cabanac, et. al.

The "IRIT.OpenWeb" run ranks suggestions based on how close it is to the context and by considering the categories matched between the user and the suggestions. The system assigns users one or more categories from WordNet and Google Places based on which suggestions users liked. Attractions are retrieved for the 50 contexts from Google Places and also tagged with the same categories. Descriptions for suggestions were fetched using Yahoo! BOSS Placefinder and Bing.

The "IRIT.ClueWeb" run is made up of documents retrieved using Terrier with queries composed of the users' categories. Documents are then ranked according to their retrieval score and similarity to the profile. Users are assigned to categories as in the "IRIR.OpenWeb" run and descriptions are also generated in the same way.

5.10 Diversifying Contextual Suggestion Search Based on User Profiles

Runids: udel_run_D, udel_run_SD

Authors: Karankumar Sabhnani and Ben Carterette

This system starts by analysing user profiles and generating bags of keywords depicting the user interests. Then, given a context, they query the Google Places API with each bag, retrieving lists containing places in that context which fit in the specified genre. Their first run creates suggestions for each profile-context combination by iterating through the lists in a round-robin fashion, selecting one place at a time to create a list of 50. Their second run sorts the retrieved lists by average user ratings before iterating round-robin.

5.11 A Nearest-Neighbor Approach to Contextual Suggestion

Runids: uncsils_base, uncsils_param

Authors: Sandeep Avula, John O'Connor, and Jaime Arguello

This system gathers a candidate set of venues from the Yelp API. In the first run, uncsils_base, for each context-profile pair, the candidate venues are scored using the weighted average rating associated with the venues in the profile. For this calculation, each profile venue was given a weight based on the cosine similarity between the candidate venue and profile venue. The goal with this approach is to score each candidate venue based on the rating associated with the most similar venues in the profile. The second run, uncsils_param, boosted the contribution from the profile venue with the greatest similarity with the candidate venue and rating.

5.12 University of Amsterdam Team

Runids: UAmsTF30WU

This systems extracts suggestions for sightseeing, shopping, eating, and drinking from Wikitravel pages dedicated to US cities. Descriptions from positive examples in the user profiles are used as queries to rank suggestions. The system then merges the per-query rankings of positive examples into a single result list. These ranked suggestions are then filtered based on the context.

5.13 University of Indonesia Team

Runids: csui01, csui02

This system gathers a list of candidate venues by issuing queries against Yelp and Foursquare that combined context and preference information. Venues were then ranked giving a high amount of weight to the rating and number of people who rated. Reviews/comments which both were liked by users and gave the venue a high rating were used as suggestion descriptions. The first run considered only the venue rating whereas the second run also employed diversity based on venue categories.

5.14 National University of Ireland, Galway Team

Runids: CIRG_IRDISCOA, CIRG_IRDISCOB

This systems finds candidate places from Google Places and WikiTravel. For each candidate place a corresponding description is extracted from the Google Places API or the Bing API. Related Wikipedia articles are also found for both example suggestions and candidate places based on the place's description. The system calculates an intersection of these Wikipedia articles between example suggestions and candidate places, these Wikipedia articles are then used to extract Wikipedia categories. A score based on the similarity of categories between candidate places and examples suggestions is calculated by the system, places with the highest score are then returned.

6 Results

Table 1 lists the scores for all open web runs for all three metrics and table 2 lists the scores for all ClueWeb12 runs. Both of these tables are sorted by their P@5 rankings (our main metric). We do not compare open web and ClueWeb12 runs against each other as part of this track. Figure 2 compares the three metrics against each other for all runs, note that there is a high amount of agreement between the three metrics. Also note that the best two performing runs are ranked the same regardless of the metric used.

7 Conclusions

We plan to continue this track for TREC 2014. Task details should remain essentially the same. Removing the temporal aspect of contexts lead to a more focused task and we have no plans of bringing it back. The ClueWeb12 collection allows easier reuse of data, however we plan to keep the open web option available. We are happy with the success of using crowdsourcing and plan to move to a fully crowdsourced approach to generating user profiles.

For next 2014 we plan to continue to use the same metrics however we will consider using contexts that are based on locations outside of the US.

References

[1] Adriel Dean-Hall, Charles LA Clarke, Jaap Kamps, and Paul Thomas. Evaluating contextual suggestion. 2013.

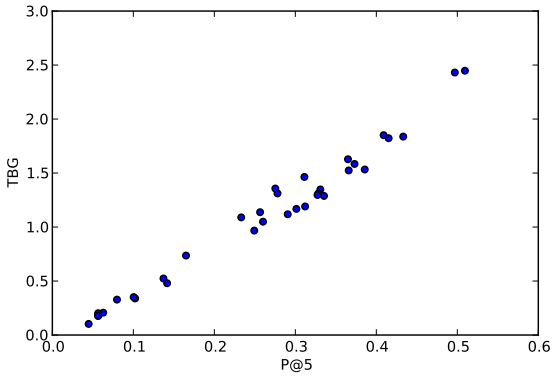
[2] Mark D Smucker and Charles LA Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 95–104. ACM, 2012.

Run	P@5 Rank	P@5 Score	TBG Rank	TBG Score	MRR Rank	MRR Score
UDInfoCS1	1	0.5094	1 (-)	2.4474	1 (-)	0.6320
UDInfoCS2	2	0.4969	2 (-)	2.4310	2 (-)	0.6300
simpleScore	3	0.4332	4 (Down 1)	1.8374	4 (Down 1)	0.5871
complexScore	4	0.4152	5 (Down 1)	1.8226	6 (Down 2)	0.5777
DuTH_B	5	0.4090	3 (Up 2)	1.8508	3 (Up 2)	0.5955
1	6	0.3857	8 (Down 2)	1.5329	7 (Down 1)	0.5588
2	7	0.3731	7 (-)	1.5843	5 (Up 2)	0.5785
udel_run_D	8	0.3659	9 (Down 1)	1.5243	8 (-)	0.5544
isirun	9	0.3650	6 (Up 3)	1.6278	9 (-)	0.5165
udel_run_SD	10	0.3354	16 (Down 6)	1.2882	10 (-)	0.5061
york13cr2	11	0.3309	12 (Down 1)	1.3483	15 (Down 4)	0.4637
DuTH_A	12	0.3283	14 (Down 2)	1.3109	12 (-)	0.4836
york13cr1	13	0.3274	15 (Down 2)	1.2970	14 (Down 1)	0.4743
UAmsTF30WU	14	0.3121	17 (Down 3)	1.1905	13 (Up 1)	0.4803
IRIT.OpenWeb	15	0.3112	10 (Up 5)	1.4638	11 (Up 4)	0.4915
CIRG_IRDISCOA	16	0.3013	18 (Down 2)	1.1681	16 (-)	0.4567
CIRG_IRDISCOB	17	0.2906	20 (Down 3)	1.1183	19 (Down 2)	0.4212
uncsils_param	18	0.2780	13 (Up 5)	1.3115	18 (-)	0.4271
uogTrCFP	19	0.2753	11 (Up 8)	1.3568	17 (Up 2)	0.4327
ming_1	20	0.2601	22 (Down 2)	1.0495	22 (Down 2)	0.3816
uncsils_base	21	0.2565	19 (Up 2)	1.1374	20 (Up 1)	0.4136
ming_2	22	0.2493	23 (Down 1)	0.9673	23 (Down 1)	0.3473
uogTrCFX	23	0.2332	21 (Up 2)	1.0894	21 (Up 2)	0.4022
run01	24	0.1650	24 (-)	0.7359	24 (-)	0.2994
baselineA	25	0.1372	25 (-)	0.5234	25 (-)	0.2316
csui02	26	0.0565	26 (-)	0.1785	26 (-)	0.1200
csui01	27	0.0565	27 (-)	0.1765	27 (-)	0.1016

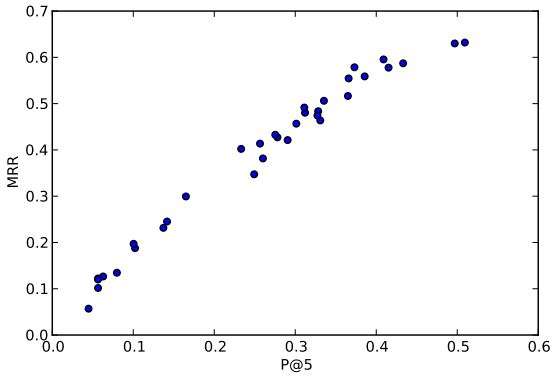
Table 1: P@5, TBG, and MRR rankings for all open web runs.

Run	P@5 Rank	P@5 Score	TBG Rank	TBG Score	MRR Rank	MRR Score
baselineB	1	0.1417	1 (-)	0.4797	1 (-)	0.2452
BOW_V17	2	0.1022	3 (Down 1)	0.3389	3 (Down 1)	0.1877
BOW_V18	3	0.1004	2 (Up 1)	0.3514	2 (Up 1)	0.1971
IRIT.ClueWeb	4	0.0798	4 (-)	0.3279	4 (-)	0.1346
RUN1	5	0.0628	5 (-)	0.2069	5 (-)	0.1265
RUN2	6	0.0565	6 (-)	0.2020	6 (-)	0.1223
IBCosTop1	7	0.0448	7 (-)	0.1029	7 (-)	0.0569

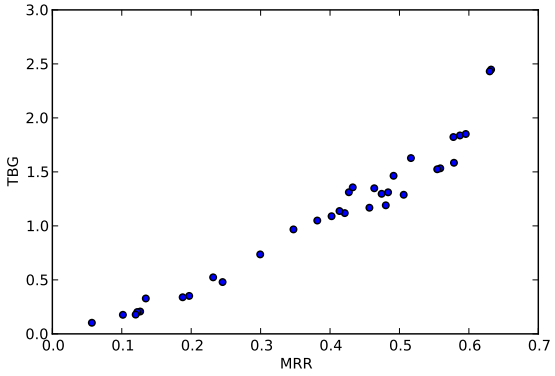
Table 2: P@5, TBG, and MRR rankings for all ClueWeb12 runs.



(a) P@5 vs TBG $\tau = 0.8160$



(b) P@5 vs MRR $\tau = 0.8959$



(c) MRR vs TBG $\tau = 0.8632$

Figure 2: Comparisons between P@5, MRR, and TBG.