Overview of the TREC 2014 Contextual Suggestion Track

Dean-Hall, A.; Clarke, C.L.A.; Kamps, J.; Thomas, P.; Voorhees, E.

Publication date
2014
Document Version
Final published version
Published in
Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014

Overview of the TREC 2014 Contextual Suggestion Track

Adriel Dean-Hall
University of Waterloo

Charles L. A. Clarke
University of Waterloo

Jaap Kamps
University of Amsterdam

Paul Thomas
CSIRO

Ellen Voorhes
NIST

1 Introduction

1.1 Summary for Previous Participants

For participants familiar with the 2013 Contextual Suggestion Track we have provided a list of the main changes to this year's track:

- Assessors were recruited only from a crowdsourcing service (Mechanical Turk) and not from any student bodies.
- Only CSV formatted files were available for profiles, contexts, and suggestions.
- Two seed cities were used instead of one (Chicago, IL and Santa Fe, NM) and the target cities were also changed.
- The number of ratings provided in profiles was changed from 50 to 70 or 100 (depending on the profile).
- 31 runs were submitted from 17 groups, 6 of these were ClueWeb12 runs and 25 were open web runs.

If you are already familiar with this track you can skip to Section 5 which gives an overview of the approaches participants used 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 third time as part of TREC 2014 after a positive response in previous years. This year participants were again given, as input, a set of 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 assessor and indicates that assessor'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 previous years each groups was allowed to submit up to two runs. A total of 17 groups submitting 31 runs participated in the track this year. 6 of these runs comprised suggestions from the ClueWeb12 dataset, the other 25 runs comprised suggestions from the open web.

2 Detailed Task Description

Profiles and contexts were distributed to participants in CSV formatted files. For this track we generated 299 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 an assessor's preferences to a list of 70-100 example suggestions within Chicago, IL and Santa Fe, NM. These profiles are built by conducting a survey advertised to crowdsourcing workers. These workers will be the same ones we use as assessors during evaluation.

Profiles are split into two files: examples2014.csv, profiles2014-70.csv, and profiles2014-100.csv. examples2014.csv contains a list of 100 suggestions which each consist of an id, a title, a description, and a url. Below are two suggestions from the file:

- **ID** 102
  Title Topolobampo/Frontera Grill
  Description Rick Bayless' innovative, refined Mexican cuisine served in an elegant, art-filled space.
  URL `http://www.rickbayless.com/`
- **ID** 185
  Title Santa Fe Chamber Music Festival
  Description Santa Fe Chamber Music Festival is music that moves you, played by artists who astound and dazzle you with their virtuosity and passion, playing chamber ...
  URL `http://www.sfcmf.org/`

The profiles contain ratings given by assessors for suggestions within examples2014.csv. Some assessors only gave 70 suggestions (all profiles with 70 ratings are the same subset of 70) and some gave all 100. Profiles with 70 ratings are in profiles2014-70.csv and profiles with 100 ratings are in profiles2014-100.csv. Below are a few lines from profiles-70.csv:

```
849,101,3,4
849,102,2,3
849,105,2,2
...
849,184,1,1
849,185,2,3
849,187,3,3
```

Listing 1: An excerpt from profiles2014-70.csv.

The first line means that the assessor with id 849 gave example suggestion number 101 a description rating of 3 and a website rating of 4 (on a scale from 0 to 4, see section 3.1).

2.1.1 Generating Example Suggestions

First we need to generate example suggestion which will rated by assessors in a survey. A suggestion consists of a title, short description, and a website URL. These 100 example suggestions were all from Chicago, IL and Santa Fe, NM. Example suggestions were selected from a commercial online listing of points-of-interest; example suggestions were chosen so that there was diversity in the types of attractions in our set.

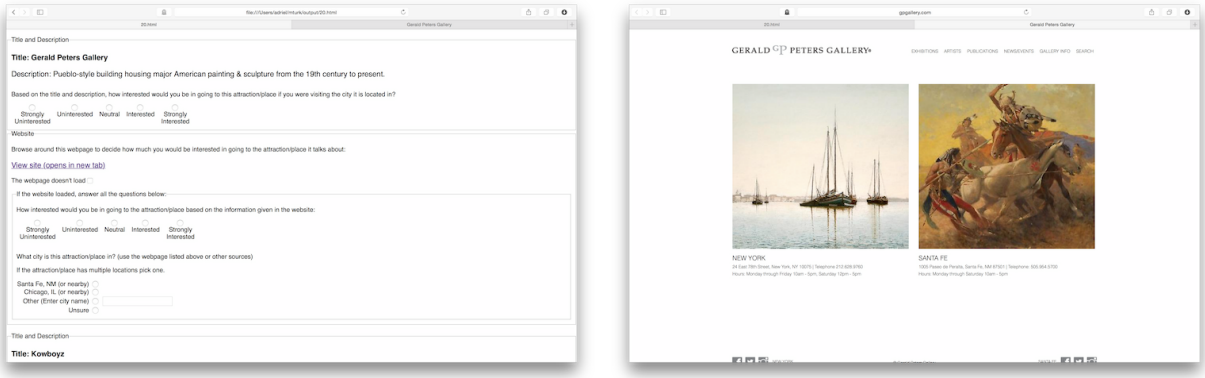


Figure 1: Screenshots of survey seen by assessors.

2.1.2 Gathering Attraction Preferences

Profiles distributed to participants indicated assessor's preference towards the example suggestions. In order to form the profiles, workers were recruited from Mechanical Turk and were asked to complete an online survey. In the survey sample suggestions were presented to assessors in a random order. Assessors were asked to give two 5-point ratings for each attraction, one for how interesting the attraction seemed to the assessor based on its description and one for how interesting the attraction seemed to the assessor based on its website. The survey interface, can be seen in figure 1. In total 299 crowdsourced assessors 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 cities used as contexts last year and the seed cities. The list of metropolitan areas was taken from Wikipedia[1].

Contexts are distributed to participants in the file contexts2014.csv.

```
...
117,Cumberland,MD,39.65287,-78.76252
118,Shreveport,LA,32.52515,-93.75018
119,Los Angeles,CA,34.05223,-118.24368
120,Portland,OR,45.52345,-122.67621
121,Grand Rapids,MI,42.96336,-85.66809
...
```

Listing 2: An excerpt from contexts2014.csv.

Here the first line means that context id 117 represents Cumberland (city), MD (state) with a latitude of 39.65287 and a longitude of -78.76252. For contexts the latitude and longitude are provided as a convenience and are synonymous with the city. They are not meant to represent the exact position of the user. Contexts represent locations at the granularity of a city-level.

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

2.3 Collections

Participants were able to gather suggestions from either the open web, ClueWeb12[2], 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** RAMA
  Run ID RAMARUN2
  Profile ID 843
  Context ID 118
  Rank 1
  Title Shreveport Railroad Museum
  Description The Shreveport Railroad Museum is on the grounds of the Shreveport Water Works Museum just outside of the Central Business District. Run and staffed by the...
  URL/Doc ID `http://www.yelp.com/biz/shreveport-railroad-museum-shreveport`
- **Group ID** udel_fang
  Run ID UDInfoCS2014_2
  Profile ID 843
  Context ID 120
  Rank 1
  Title Portland Art Museum
  Description Portland Art Museum is a museum. HERE ARE THE DESCRIPTIONS FROM ITS WEB SITE:The portland art museum is the seventh oldest museum in the united states and the oldest in the pacific northwest. HERE ARE REVIEWS FROM OTHER PEOPLE:The layout is excellent with certain special exhibits sectioned off to themselves. Each floor has just the right size exhibit space, which is perfect to appreciate the art presented.
  URL/Doc ID `http://portlandartmuseum.org/`

3 Judging

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

[2] `http://lemurproject.org/clueweb12/`

assessors.

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. Assessors 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, two, or three randomly chosen contexts.

The judgements given were one of:

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

Approximately half of our assessors responded to an invitation to the second survey. In total 299 context-profile pairs were judged by 144 assessors.

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

```
...
BJUTa 843 124 http://www.shamrockbrewing.com/ 2 2 6 2
BJUTa 843 124 http://www.outoftheblueskydiving.com/ 0 0 4 2
BJUTa 846 109 http://westmichiganbeertours.com/ 4 4 19 58
...
```

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

Here the last line means the the assessor (843) was strongly interested (4) in the point-of-interest based on both the description provided by run BJUTa from context 109 and the website. The last two numbers mean that the assessor took 19 sec. to make the judgement based on the description and 58 sec. to make the judgement based on 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 assessors 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.

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

Note that only NIST accessors explicitly made judgements of 1, crowdsourced assessors made judgements of either 0 or 2, however some of the user judgements are reported as 1 when crowdsourced assessors 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 crowdsourced assessors disagree on whether a suggestion is contextually appropriate the judgment given by NIST accessors is used.

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

```
...
102 http://www.adixiontours.com 0
102 http://www.aeropostale.com 2
102 http://www.ahsoaz.com 1
...
```

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

Here the first line means that for context 102 the website `http://www.adixiontours.com` is not geographically appropriate (0).

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

17 groups participated in the Contextual Suggestion track this year. An overview of the approaches used by groups that used the open web as their main datasource or ClueWeb12 as their main datasource is given in this section. More details about the approaches used by these teams are available in the individual team TREC reports.

5.1 Open Web Approaches

5.1.1 BJUT

This group used a combination of venue categories and venue ranking to make suggestions. Using the profile data they determined the probability the user would like the venue based on which of five categories it was in (attraction, activities, restaurant, shopping, and nightlife). The number of venues in the top 50 suggestions from each category is determined by this probability. These 50 suggestions are then ordered by either rank-only (BJUTa) or by both rank and user preferences (BJUTb).

5.1.2 BUW

This group used a ranking of venues that came from a commercial web service. Then then compared the difference between using a description generated from a different commercial web service (webis_1) and that same description with a sentence from an averagely ranked review prepended to it (webis_2).

5.1.3 eindhoven

This group used learning to rank techniques in order to make suggestions using two algorithms, RankNet (tueNet) and Random Forest (tueRforest). The set of attractions in the profile was used as training data. They used distance, scores from commercial web services, categories, description keywords and review keywords as features for their algorithm. Additionally they considered the impact of distance between training and testing venues and which commercial web services provided the most useful data.

5.1.4 Group.Xu

This team based their suggestions heavily on the venue's category. They used category information returned by a commercial web service. When multiple categories were returned they picked a core category using idf based heuristics. Here the core category is the one that appears in fewer venues. In addition to determining which categories each user liked they also determined which categories were more appropriate depending upon the population of the city suggestions were being made for. This team only had one run (dixlticmu).

5.1.5 ICTNET

This team compared generating feature vectors for users manually (cat) to generating them using LDA topic modeling (lda). In the first approach they manually annotate profiles and score candidate suggestions based on how many of these annotation appear in the venue information. In the second approach they use cosine similarity between the profile and candidate suggestion along with the ratings to score candidate suggestions.

5.1.6 RAMA

This team used three components to make up the scores given to candidate suggestions. They used a general user interest score based solely on the categories the user appeared to be interested in. They used a specific user interest score based on nouns extracted from venue titles and descriptions. Finally, they used a context score based on the user's distance from the venue, number of reviews, and average rating of the venue. These three scores were linearly combined. In both runs the context score was given a low weight, in one run (RAMARUN2) the general interest score was given a high weight and in the other run (RUN1) the specific interest score was given a high weight.

5.1.7 udel

This team generated terms that users were interested in based upon their profiles. A commercial web service is then queried with these terms, it returns venues which are incorporated into the final results. In contrast to a similar technique used by the group last year, this year the group ordered terms based on the number of likes were associated with each term. The run, run_DwD, didn't use this term prioritization whereas the second run, run_FDWD did.

5.1.8 udel_fang

This team uses positive and negative venue reviews to build positive and negative user models. Various representation of these models are experiments with: using full text, using unique terms, using only frequent terms (both unique and not unique), using only nouns, and using summaries generated using the Opinosis algorithm. Last year this team used linear interpolation of similarity scores between both positive and negative profiles and candidate reviews. This year they compared the approach with a learning to rank approach (using LambdaMART). The UDInfoCS2014_1 run uses this learning to rank approach and UDInfoCS2014_2 uses linear interpolation.

5.1.9 uogTr

This team used three different approaches to making suggestions. The first approach (uogTrCsLtr) used learning to rank (LambdaMART) to rerank venues that were already personalized using a technique similar to the one that this team used last year. This algorithm was passed 64 features including category information, city features, user features, and venue features. The second approach (uogTrBunSum) bundled venues together using venue popularity and similarity to the profiles as features. In order to ensure diversity only the most central venues in each bundle were suggested. The final approach was not submitted as a run and is similar to the approach this team used last year (which the learning to rank approach builds upon). It serves as a baseline for this team.

5.1.10 waterloo_clarke

This team clustered all the venues in a hierarchical fashion, so that at the top of the tree are the most general clusters of venues and lower in the tree are more specific clusters of venues. Terms from venue webpages and descriptions and normalized with K-L divergence and used as features for the clustering. In addition categories are also used as features. The clusters are then ordered by how many liked example attraction exist in them and attractions are picked from each cluster one at a time. The first run, waterlooA, used only positive ratings in the clustering ranking algorithm, and the second run, waterlooB, used both positive and negative ratings.

5.2 ClueWeb12 Approaches

5.2.1 CWI

This groups used standard vector space model techniques to compare the similarity between candidate venues and the user's profile. This similarity was used to rank the candidate venues being suggested. This group compared different techniques for generating the list of candidate suggestions being ranked using this technique. The first technique used to generate a set of candidate suggestion was the GeoFiltered technique where only venues that mentioned `City, State` (in that format) were considered (CWI_CW12_Full). The second technique, TouristFiltered, used a list of hand-picked domains that were geared towards tourist information and included all venues that were part of documents in those domains or documents that were linked to from documents in those domains (CWI_CW12.MapWeb). The third technique, that was not submitted as a run, used venues returned commercial web services to generate a set of candidate suggestions.

5.2.2 TJU_CS_IR

This team gathered attractions from Wikitravel and created vector representations of all the venues based on their titles and descriptions. In order to generate user profiles the ratings users gave for the example attractions along with the created vectors that represent each sample attractions are combined and passed to the Softmax algorithm. This outputs a user model which is then used to rank all candidate suggestions to porvide the final ranking. The output of this technique (RunA) is compared with using KNN instead of the Softmax algorithm (RunB).

5.2.3 UAmsterdam

This team scores venues based on three components: the prior probability of being relevant which is based on the PageRank algorithm (ensuring high quality pages are suggested), the probability of the suggestion being in the context, and the probability of the suggestion being in the user profile. Two approaches are used to develop the languages models to estimate these probabilities. The first approach (Model0) uses the full text of the documents in the index. The second approach (Model1) limits the terms in the documents to only terms that appear in anchor text linking to the document.

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

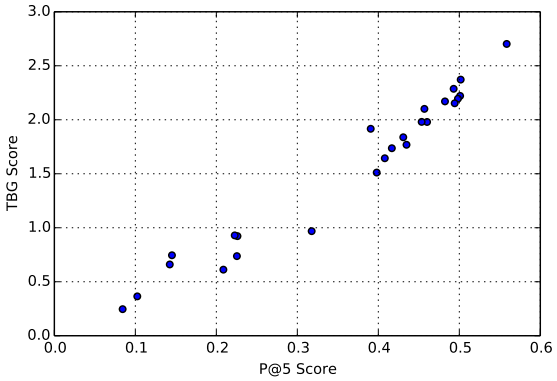
We are currently considering continuing this track as part of TREC 2015. The basic task will be similar to the one this year. However there are areas we will focus on improving next year. Firstly, ClueWeb12 provides a reusable test collection however most groups use the open web. Providing a reusable test collection that is more appropriate for this task and that participants are willing to use is important. Secondly, we are considering moving to suggestion and evaluation of itineraries rather than just individual points-of-interest.

Run	P@5 Rank	P@5 Score	TBG Rank	TBG Score	MRR Rank	MRR Score
UDInfoCS2014_2	1	0.5585	1 (-)	2.7021	1 (-)	0.7482
RAMARUN2	2	0.5017	2 (-)	2.3718	2 (-)	0.6846
BJUTa	3	0.5010	4 (Down 1)	2.2209	4 (Down 1)	0.6677
BJUTb	4	0.4983	5 (Down 1)	2.1949	6 (Down 2)	0.6626
uogTrBunSumF	5	0.4943	7 (Down 2)	2.1526	3 (Up 2)	0.6704
RUN1	6	0.4930	3 (Up 3)	2.2866	5 (Up 1)	0.6646
webis_1	7	0.4823	6 (Up 1)	2.1700	7 (-)	0.6479
simpleScoreImp	8	0.4602	10 (Down 2)	1.9795	8 (-)	0.6408
webis_2	9	0.4569	8 (Up 1)	2.1008	12 (Down 3)	0.5980
simpleScore	10	0.4535	9 (Up 1)	1.9804	9 (Up 1)	0.6394
run_FDwD	11	0.4348	13 (Down 2)	1.7684	13 (Down 2)	0.5916
waterlooB	12	0.4308	12 (-)	1.8379	10 (Up 2)	0.6244
waterlooA	13	0.4167	14 (Down 1)	1.7364	11 (Up 2)	0.6021
UDInfoCS2014_1	14	0.4080	15 (Down 1)	1.6435	14 (-)	0.5559
dixlticmu	15	0.3980	16 (Down 1)	1.5110	15 (-)	0.5366
uogTrCsLtrF	16	0.3906	11 (Up 5)	1.9164	16 (-)	0.5185
run_DwD	17	0.3177	17 (-)	0.9684	19 (Down 2)	0.3766
tueNet	18	0.2261	19 (Down 1)	0.9224	18 (-)	0.3820
choqrun	19	0.2254	21 (Down 2)	0.7372	23 (Down 4)	0.3412
tueRforest	20	0.2227	18 (Up 2)	0.9293	20 (-)	0.3604
cat	21	0.2087	23 (Down 2)	0.6120	21 (-)	0.3496
BUPT_PRIS_01	22	0.1452	20 (Up 2)	0.7453	17 (Up 5)	0.4475
BUPT_PRIS_02	23	0.1425	22 (Up 1)	0.6601	22 (Up 1)	0.3467
gw1	24	0.1024	24 (-)	0.3646	24 (-)	0.1649
lda	25	0.0843	25 (-)	0.2461	25 (-)	0.1564

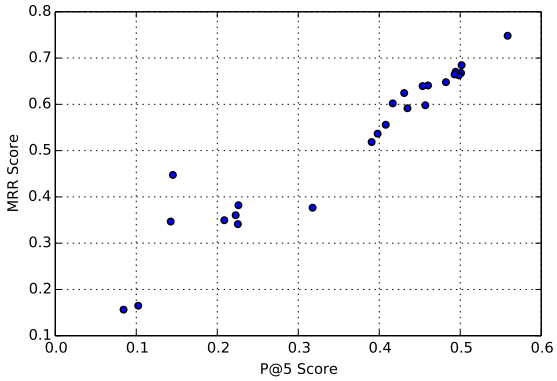
Table 1: P@5, TBG, and MRR rankings for all open web runs.

Run	P@5 Rank	P@5 Score	TBG Rank	TBG Score	MRR Rank	MRR Score
CWI_CW12.MapWeb	1	0.1445	1 (-)	0.6078	1 (-)	0.2307
Model1	2	0.0903	2 (-)	0.3411	2 (-)	0.1979
Model0	3	0.0582	3 (-)	0.1994	3 (-)	0.1023
runA	4	0.0482	4 (-)	0.1647	4 (-)	0.0856
CWI_CW12_Full	5	0.0468	5 (-)	0.1256	5 (-)	0.0767
runB	6	0.0254	6 (-)	0.0614	6 (-)	0.0552

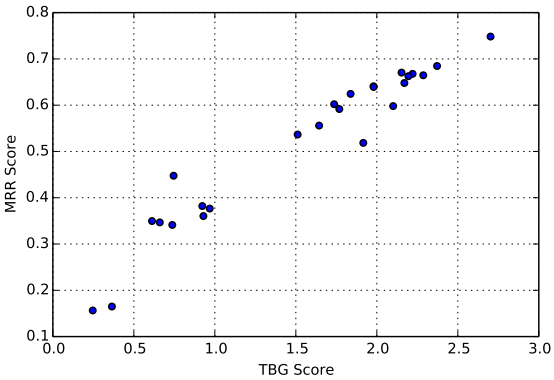
Table 2: P@5, TBG, and MRR rankings for all ClueWeb12 runs.



(a) P@5 vs TBG $\tau = 0.89$



(b) P@5 vs MRR $\tau = 0.89$



(c) MRR vs TBG $\tau = 0.84$

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

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.