Overview of the TREC 2012 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 Voorhees
NIST

1 Introduction

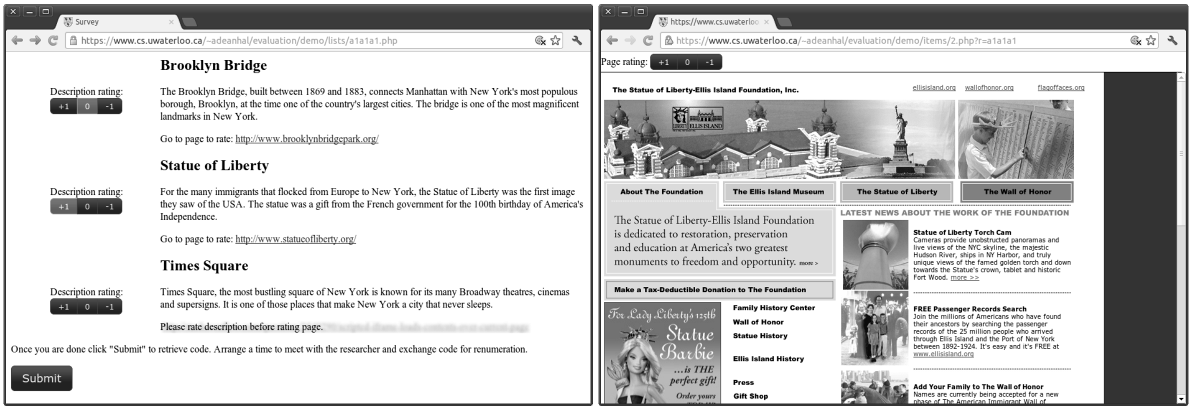
The contextual suggestion track investigates search techniques for complex information needs that are highly dependent on context and user interests. According to a report from the Second Strategic Workshop on Information Retrieval in Lorne [1]: "Future information retrieval systems must anticipate user needs and respond with information appropriate to the current context without the user having to enter an explicit query... In a mobile context such a system might take the form of an app that recommends interesting places and activities based on the user's location, personal preferences, past history, and environmental factors such as weather and time... In contrast to many traditional recommender systems, these systems must be open domain, ideally able to make suggestion and synthesize information from multiple sources..."

For example, imagine a group of information retrieval researchers with a November evening to spend in beautiful Gaithersburg, Maryland. A contextual suggestion system might recommend a beer at the Dogfish Head Alehouse (www.dogfishalehouse.com), dinner at the Flaming Pit (www.flamingpitrestaurant.com), or even a trip into Washington on the metro to see the National Mall (www.nps.gov/nacc). As its primary goal, the contextual suggestion track seeks to develop evaluation methodologies for such systems.

TREC 2012 is the first year for the track. For this first year, we introduced a single task to evaluate contextual suggestion from the open Web. As input to the task participants were given a set of example suggestions, a set of user preference profiles, and a set of geotemporal contexts. The task was to take the profiles and contexts and to produce up to 50 ranked suggestions for each combination of profile and context. Participants gathered suggestions from the open Web.

Each profile corresponds to a single user, and indicates that user's preference with respect to each sample suggestion. For example, one suggestion might be to have a beer at the Dogfish Head Alehouse, and the profile might include a negative preference with respect to this suggestion. Each suggestion includes a title, description, and an associated URL. Each context corresponds to a particular geotemporal location, including city, day of the week, time of day, and season. For example, the context might be Gaithersburg, Maryland, on a weekday evening in the fall. The geographical contexts are very coarse-grained (i.e., an entire city) to help simplify the task.

A total of 14 groups submitting 27 runs participated in this first year of the track, this includes the 2 baseline submissions from CSIRO and 2 baseline submissions from the University of Waterloo. These baselines will be discussed later in this report. Given the newness of the track, participants were given the option of basing their suggestions solely on the user profiles (returning suggstions appropriate to any place and time) or solely on geotemporal context (returning suggestions appropriate to a generic user). Only one group, from the University of Delaware, took advantage of this option, submitting runs based solely on user profile.



Survey
https://www.cs.uwaterloo.ca/~adeanhal/evaluation/demo/lists/a1a1a1.php
Description rating:
+1 0 -1
Submit
https://www.cs.uwaterloo.ca/~adeanhal/evaluation/demo/items/2.php?r=a1a1a1
Page rating:
+1 0 -1
ellisisland.org
wallofhonor.org
flagoffaces.org
About The Foundation
The Ellis Island Museum
The Statue of Liberty
The Wall of Honor
Make a Tax-Deductible Donation to The Foundation
Family History Center
Wall of Honor
Statue History
Ellis Island History
Press
Gift Shop
For Lady Liberty's 125th Statue Barbie ...is THE perfect gift! Order yours

Brooklyn Bridge

The Brooklyn Bridge, built between 1869 and 1883, connects Manhattan with New York's most populous borough, Brooklyn, at the time one of the country's largest cities. The bridge is one of the most magnificent landmarks in New York.

Go to page to rate: http://www.brooklynbridgepark.org/

Statue of Liberty

For the many immigrants that flocked from Europe to New York, the Statue of Liberty was the first image they saw of the USA. The statue was a gift from the French government for the 100th birthday of America's Independence.

Go to page to rate: http://www.statueofliberty.org/

Times Square

Times Square, the most bustling square of New York is known for its many Broadway theatres, cinemas and supersigns. It is one of those places that make New York a city that never sleeps.

Please rate description before rating page.

Once you are done click "Submit" to retrieve code. Arrange a time to meet with the researcher and exchange code for renumeration.

The Statue of Liberty-Ellis Island Foundation, Inc.

The Statue of Liberty-Ellis Island Foundation is dedicated to restoration, preservation and education at America's two greatest monuments to freedom and opportunity. more >

LATEST NEWS ABOUT THE WORK OF THE FOUNDATION

Statue of Liberty Torch Cam
Cameras provide unobstructed panoramas and live views of the NYC skyline, the majestic Hudson River, ships in NY Harbor, and truly unique views of the famed golden torch and down towards the Statue's crown, tablet and historic Fort Wood. more >>

FREE Passenger Records Search
Join the millions of Americans who have found their ancestors by searching the passenger records of the 25 million people who arrived through Ellis Island and the Port of New York between 1892-1924. It's easy and it's FREE at www.ellisisland.org

Add Your Family to The Wall of Honor
Names are currently being accepted for a new phase of The American Immigrant Wall of

Figure 1: When judgeing contexts users were presented with (a) A list with descriptions and links to website; after clicking on a link they were presented with (b) the website with a rating bar at the top

2 Detailed Task Description

Detailed formats for the sample suggestions, profiles, and contexts are given below. These descriptions are followed by a brief discussion of the result suggestions returned by the participants. An experimental run consists of a single result suggestion file generated automatically from the sample suggestions, profiles, and contexts. Each participating group was permitted to return up to two experimental runs.

2.1 Sample Suggestions

Suggestions are descriptions of attractions that are intended to be recommended to a user as something they would find interesting. Suggestions consist of a title, short description and a website URL. Fifty sample suggestions were gathered manually to be used to create profiles. However, due to problems described in the next section, one sample was excluded from the distributed sample suggestion file. The sample suggestions were all attractions within the Toronto region. The URL for these sample attractions were the attraction's main page and the descriptions were taken from the website's meta tag description or, if that was unavailable, from a description provide by a third-party website.

```
<example number="1">
    <title>Fresh on Bloor</title>
    <description>Our vegan menu boasts an array of exotic
     starters, multi-layered salads, filling wraps, high
     protein burgers and our signature Fresh bowls.</description>
    <url>http://www.freshrestaurants.ca</url>
</example>
 ...
```

Listing 1: Example of a sample suggestion

2.2 Profiles

Profiles indicate a user's preference for a particular attractions. The profiles distributed to participants consist of preferences for the attractions in the sample suggestions list. In order to build profiles for the task, we surveyed University of Waterloo students, creating a profile from the preferences of each participating student.

```
<context number=”1”>
    <city>New York City</city>
    <state>NY</state>
    <lat>40.71427</lat>
    <long>−74.00597</long>
    <day>weekday</day>
    <time>afternoon</time>
    <season>summer</season>
</context>
...
```

Listing 3: Example of a context.

In the survey, sample suggestions were presented to users in a random order. Users were asked to give two ratings for each attraction, one for the attraction's description and one for the attraction's website. Each rating given was one of three levels of interest in the attraction. The levels were: +1 (looks interesting), 0 (indifferent), and -1 (looks boring). Examples of the survey interface are given in figure 1.

In total 34 results from survey participants were released as profiles. A few users responded to the survey but did not complete it. Preferences from these users were not were not included in the released profiles.

```
<profile number=”1”>
    <example number=”1” initial=”−1” final=”0”>
    <example number=”2” initial=”1” final=”1”>
        ...
    <example number=”49” initial=”1” final=”1”>
</profile>
...
```

Listing 2: Example of part of a profile.

Survey participants were given 50 example suggestions. However during the course of the survey one of the participants brought to our attention that one of the example suggestions had a URL that did not match with the title and description. This error was caused by a incorrect listing for the attraction on a tourist website. This example suggestion was removed from the profiles distributed to the track participants.

2.3 Contexts

Contexts describe where the user physically is located and what time of day, week, and year it is when the user is searching. Each context consists of a city in the United States, a part of the week (weekday or weekend), a time of day (morning, afternoon, or evening), and a season (spring, summer, fall, or winter). For the contexts distributed to the participants, the part of week, the time of day, and the season were chosen at random, with each option in each field having an equal chance of being selected. The cities were chosen from the geonames.org list of US cities, cities with populations less than 100,000 were filtered out. The likelihood of being chosen was proportionate to the population of the city. If this random process generated the same context more than once, the duplicate context was discarded, and a new context was generated.

2.4 Result Suggestions

Each submitted run consists of 50 ranked suggestions for each profile/context pair, with formatting similar to that of the sample suggestions, including a title, description, and url. In generating the suggestions, task participants could use whatever resources they wished, including review Websites. The goal was that each suggestion should be appropriate to the profile (based on the user's preferences) and the context (according to the geotemporal location). Ideally, the description of the suggestion would be tailored to reflect the preferences of that user. For the purposes of generating suggestions, participating groups could assume that the users were persons of legal drinking age in their early-to-mid twenties. Participating groups could also assume that a user has up to five hours available to follow a suggestion and has access to appropriate transportation (e.g., a car).

3 Judging

The task of judging the result suggestions was split up into two independent tasks: judging the suggestions with regard to the profile and judging the suggestions with regard to the context.

3.1 Profile Relevance

In order to judge the relevance of suggestions with respect to a profile, we conducted a second survey. In this second survey, users from the first survey were invited to return to judge suggestions from task participants. Returning users were given the option to judge up to three sessions, with a session consisting of suggestions from a single geotemporal context. Each session was made up of all returned suggestions up to rank 5 for all runs for a particular profile-context pair (giving at most 135 suggestions per session).

As we did for the first survey, suggestions were ordered randomly and users were asked to give one rating for the suggestion's description and one for the suggestion's website. Scores given by users were incremented by one before being distributed to task participants (i.e., from -1, 0, or 1 to 0, 1, or 2). This renumbering was done for consistency with the geographical and temporal judgment scores described below.

All 34 users who had a profile corresponding to them distributed for this task were invited to participate in the second survey, 17 students responded to this invitation and completed the second survey. A total of 44 profile-context pairs were judged.

3.2 Context Relevance

Assessors at NIST judged how relevant a suggestion was with regard to a context by visiting the attraction's website and looking for where the attraction is located and what time the attraction is open. 15 contexts for all profiles up to suggestions at rank 5 were judged this way. Each judged suggestion was given a temporal and a geographical score.

- If the suggestions was appropriate it was given a score of 2.
- If the suggestions was marginally appropriate it was given a score of 1.
- If the suggestions was not appropriate it was given a score of 0.
- If the website for the suggestion did not load at the time of assessment it was given a score of -2.

3.3 Baseline Runs

There were four baseline runs submitted as part of the track. The first baseline (waterloo12a) consist of the top 50 attractions from each city according to the ranking given by tripadvisor.com. Only the geographical aspect of the context is taken into accout for this run, the profiles and temporal ascpect of the context are ignored.

The second baseline (waterloo12b) also uses tripadvisor.com, incorporating the profiles by using the site's search tool to search for attractions similar to the ones the user gave a high rating to in their profile. The search terms used were manually generated based on the sample suggestions. Again the attractions were sorted according to the rating given on the website.

The third baseline (baselineA) is output from the Google Places API, a commercial database of places and reviews. It allows us to compare TREC runs with simply using a commercial API, admittedly simple-mindedly. Each context was split into day, time, and place. The API was queried for the place at a number of days and times to cover the context: for example, a "weekend morning" would see queries issued from 0800 to 1045, local time, on a Saturday and Sunday. The union of all results was taken, deduplicated, and the top-rated results were kept as suggestions. Descriptions were generated by feeding the URLs for each suggestion to the Yahoo BOSS search API and taking Yahoo's snippets. Profile information wasn't used in this run since commercial APIs do not expose this for general use.

The final baseline is the "pub run" baseline (baselineB), it is the same as the commercial baseline, with candidates restricted to pubs, restaurants, and cafes. The intuition is that a pub will likely be a popular choice, but the run serves as a test of diversity: if baselineB does well despite only ever recommending pubs, we may need to consider an explicit diversity measure of some sort in future.

4 Participant Approaches

4.1 CSIRO

Runs: csiroht and csiroth

CSIRO's two runs used the intersection of candidate venues provided by Google Places and Foursquare. Venues were ranked according to a linear combination of two scores: one that captures the venue's appropriateness at the given time, and another that captures its appropriateness given the user's prior likes and dislikes.

For time-of-day scores, CSIRO sampled a large number of locations from Foursquare and noted the number of people "checked in" at regular intervals. By aggregating over locations and time it was possible to learn for example that throughout the week movie theatres are popular on evenings; while theme parks are popular throughout the day on weekends, but only on afternoons during the week.

For matches to profiles, the description of each example was first reduced to a BM25-weighted vector. Vectors were summed, with fixed multipliers of 0.75 for positive examples and -0.25 for negative examples. The resulting term weights were used to score the description of each candidate suggestion.

Run csiroth mixed the text-based scores with the time-of-day scores with ratio 7:3. Run csiroht mixed the scores in the ratio 3:7.

4.2 Fasilkom UI from Universitas Indonesia

Runs: FASILKOMUI01 and FASILKOMUI02

Fasilkom UI (GroupID: FASILKOMUI) from Universitas Indonesia submitted two runs for the TREC 2012 Contextual Suggestion tracks. Their contextual suggestion system combines the user model and the geolocation from the context information to generate the place suggestions. The system gathers place data from various web services such as Yelp, Google Place, and Trip Advisor. The suggestion search is based on the Yelp's place category list. Thus, given a user model and a context, the system will find places that match the categories listed in the user model and also located in the proximity of the current location. It will also expand the search to the similar categories if it cannot find a particular category in the current location. As for the scoring, they use the review rating for the places to produce the ranked results. They also try to apply diversity to the suggestion results to make the suggestion varied and more interesting for the users. Both submitted runs namely FASILKOMUI01 and FASILKOMUI02 use same techniques, the difference is that FASILKOMUI01 uses diversity while FASILKOMUI02 does not apply diversity to its suggestion results.

4.3 HP Labs China

Runs: hplcrating and hplcranking

HPLC uses a context-aware recommendation approach to produce the ranked list of suggestions. In their approach, they crawled all places in all related cities. They employed Matrix Factorization for collaborative filtering to learn the latent factor of each user in profiles and Yelp. In detail, hplcrating uses SVD++ to predict the scores of all suggestions for a user, while hplcranking uses pairwise ranking model to rank a list of suggestions for a user. In addition, their approach makes use of category information in the matrix factorization model to learn user's general preference. They assume that if a person likes "Hammam Spa", he or she probably prefers similar places in the category "Beauty and Spas". Due to lack of contextual information in profile and dataset of users' preferences crawled from Yelp, they use the contextual post-filtering approach to adjust the resulting set of suggestions. They filter out suggestion according to the category feature and the information extracted from Yelp. For example, "Bars" usually do not open in the morning. Some places extracted from Yelp contain the opening and closing time which can be used to filter unmatched context. Website of each suggestion is extracted from Yelp in their approach. The description for each suggestion is generated by a human defined template which includes location, category and many other features about the suggestion.

4.4 IRIT Lab

Runs: iritSplit3CPv1 and iritSplit3CPv2

The IRIT Lab defined a retrieval framework that combines two modules:

- Context processing. IRIT used the Google Places API to retrieve a list of places. This takes as input geographic coordinates (those defined in each track context) and a set of place types. Different sets of place types were intuitively defined to match the combinations of temporal parts provided in the track contexts.

- Preference processing i.e., result personalization according to user interests. IRIT used vector-based profiles relying on the Vector Space Model. For each profile, the positive user preferences were based on the positively rated examples. In the same way, the negative preferences were based the negatively rated examples. Two approaches were defined to build and use positive and negative user preferences.

The coarse-grained approach consisted in defining a positive preference vector and a negative preference vector for each user profile. The fine-grained approach consisted in defining positive and negative preferences as sets of positive and negative preference examples (one vector per example). A similarity score between each place vector (from Google Places) and each preference vector based on the cosine measure was then computed.

The two submitted runs built on the same approach of context processing. The iritSplit3CPv1 run featured the coarse-grained approach of preference processing, while the iritSplit3CPv2 run featured the fine-grained approach.

4.5 TNO RaboudUniv

Runs: run01TI and run02K

First, initial recommendations were collected by using the location context as search query in Google Places. The recommendations were ranked by their textual similarity to the user profiles. In run01TI this similarity was based on a TF-IDF measure. For each term the term frequency was based on the number of occurrences of a term in the textual content of examples from the profile that were rated positively. The inverse document frequency was based on all the terms in all the documents. The cosine similarity of an initial recommendation to the positive profile determined the ranking. In run02K, the textual similarity was based on a point-wise Kullback Leibler divergence score. This measure is based on the probability of observing a term in the set of examples that were rated positively, compared to the probability of observing it in the set as a whole. The sum of Kullback Leibler scores for the terms that occur in the initial recommendations determined the rank. In order to improve the ranking of generally popular sights, the resulted ranking was combined with a number of other rankings based on Google Search, popularity and categories. As a final step, items that did not match the temporal context were filtered out.

4.6 InfoLab from University of Delaware

Runs: UDInfoCSTc and UDInfoCSTdc

The InfoLab from University of Delaware (i.e., udel_fang) submitted two runs for the contextual suggestion task. Their goal is to evaluate (1) an information gathering strategy that crawls and integrates the information about candidate places from multiple sources, such as Yelp, Google, Foursquare etc.; and (2) a ranking-based method that computes the similarity between a candidate place and a user profile based on both descriptions and categories of the candidate and example places. In UDInfoCSTc, they use categories only to compute the similarities between each potential suggestion/example suggestion pair. For each user, positive and negative similarities are combined using optimal coefficients learned from examples. Final ranking is based on combined similarities scores. In UDInfoCSTdc, they used the similar strategy with UDInfoCSTc but when computing similarities they used categories together with descriptions. For both runs, they post-process the results to meet the geographic and temporal requirements.

5 Measures

A total of 12 different measures were used in this task. There were 4 scores as output from the judgments: description rating (D), website rating (W), geographical relevance (G), and temporal relevance (T). For each score the mean reciprocal rank (MRR) up to rank 5 and precision at rank 5 (P@5) were calculated for each profile-context pair. A perfect score (i.e., a 2) was treated as a 1, otherwise the score was treated as a 0 for these measures. There were also two combinations of scores computed: WGT combined and GT combined. In order to score a 1 in these combined measures, a perfect score was required for each of the individual scores in the combination. Again for these combinations MRR and P@5 were calculated. MRR and P@5 for each of the 4 individual scores as well as the 2 combination scores makes for a total of 12 measures. In addition to calculations for each profile-context pairs there are also calculations for a mean over all contexts for each profile, a mean over all profiles for each context and an overall mean across all profile-context pairs. The overall means across all profile-context pairs are the ones reported in the results section.

6 Results

Table 1 presents all runs ordered by P@5 on the WGT score; table 2 presents all runs ordered by MRR on the WGT score. All runs submitted were judged according to the judging criteria of both the profile scores and the context scores. As mentioned early, the University of Delaware submitted runs based solely on the user profile. However, their suggestions generally appeared to be geotemporally appropriate, so we have also included their full results in the table.

Figure 2 shows the correlation between 4 different pairs of measures. All pairs of MRR and P@5 for the same score or score combination had a correlation similar to figure 2a with the mean Kendall's tau coefficient among these 6 pairs being 0.8267 (the lowest being at 0.7361). The correlation between P@5 WGT and P@5 W shown in figure 2c is not as strong between the correlation between P@5 WGT and P@5 GT shown in figure 2b. Figure 2d shows somewhat of a correlation between the description and website P@5 measures.

7 Conclusions

At the time of writing, we plan to continue the track for TREC 2013. Task details should remain essentially the same, but suggestions will not be taken from the open Web. Instead, we plan to use the new ClueWeb12 collection, developed at CMU. This collection includes approximately a billion Web pages crawled from the general Web in mid-2012. The crawl was seeded (in part) with a large number of travel and review sites, which we hope will make it appropriate for use in the track.

Another limitation of the track this year was the low number of profiles available, moving from a survey based method to a crowdsourcing based method will allow us to gather more profiles. Instead of tens of profiles our goal for TREC 2013 is to have hundreds of profiles made available to track participants.

Finally, NIST assessors found that judging the temporal aspects of the context difficult and ended up labelling most suggestions as temporally relevant. For TREC 2013 we plan to re-access our use of the temporal aspects of the context and we may remove them from the track.

References

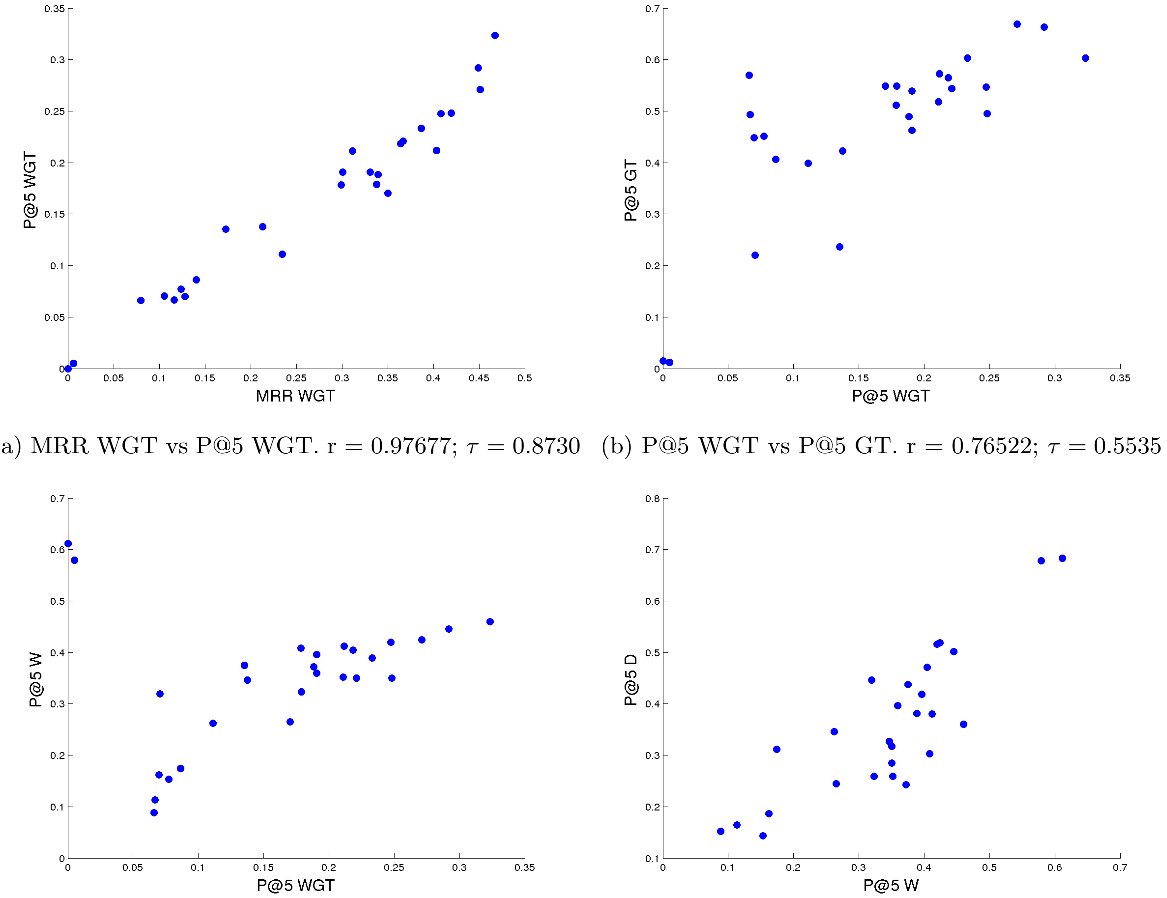
[1] Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum*, 46(1), June 2012.

Run	**P@5 WGT**	P@5 GT	P@5 G	P@5 T	P@5 W	P@5 D
iritSplit3CPv1	0.3235	0.6027	0.8930	0.6156	0.4599	0.3605
guinit	0.2920	0.6635	0.8802	0.6997	0.4451	0.5019
gufinal	0.2710	0.6689	0.8852	0.7031	0.4241	0.5191
UDInfoCSTc	0.2481	0.4950	0.7565	0.5794	0.3500	0.2852
PRISabc	0.2475	0.5464	0.9036	0.5510	0.4198	0.5160
hplcranking	0.2333	0.6032	0.8148	0.6147	0.3889	0.3815
UDInfoCSTdc	0.2210	0.5442	0.7939	0.6210	0.3500	0.3173
run02K	0.2185	0.5649	0.9034	0.5839	0.4049	0.4710
hplcrating	0.2117	0.5725	0.8815	0.5833	0.4124	0.3802
udelp	0.2111	0.5181	0.8530	0.5365	0.3519	0.2593
run01TI	0.1907	0.5392	0.8934	0.5598	0.3963	0.4185
ICTCONTEXTRUN2	0.1907	0.4624	0.8274	0.4955	0.3593	0.3969
udelnp	0.1883	0.4893	0.8414	0.5049	0.3722	0.2432
iritSplit3CPv2	0.1790	0.5486	0.8466	0.5580	0.3235	0.2593
baselineA	0.1784	0.5114	0.7908	0.5694	0.4086	0.3031
baselineB	0.1704	0.5482	0.8060	0.5883	0.2654	0.2444
waterloo12a	0.1377	0.4229	0.8230	0.4451	0.3463	0.3272
UAmsCS12wtSUM	0.1352	0.2363	0.4011	0.4335	0.3753	0.4377
ICTCONTEXTRUN1	0.1111	0.3986	0.8045	0.4055	0.2623	0.3463
waterloo12b	0.0864	0.4065	0.6827	0.4988	0.1741	0.3117
csiroth	0.0772	0.4516	0.7579	0.4734	0.1531	0.1438
UAmsCS12wtSUMb	0.0704	0.2202	0.4145	0.4040	0.3198	0.4463
csiroht	0.0698	0.4483	0.7573	0.4712	0.1623	0.1864
FASILKOMUI02	0.0667	0.4935	0.7770	0.5243	0.1136	0.1648
FASILKOMUI01	0.0660	0.5701	0.7894	0.6154	0.0883	0.1519
watcs12a	0.0049	0.0120	0.0134	0.6967	0.5790	0.6784
watcs12b	0.0000	0.0147	0.0187	0.5365	0.6117	0.6833

Table 1: All 6 P@5 measures sorted by WGT.

Run	**MRR WGT**	MRR GT	MRR G	MRR T	MRR W	MRR D
iritSplit3CPv1	0.4675	0.7585	0.9480	0.7634	0.6493	0.5461
gufinal	0.4514	0.8068	0.9108	0.8589	0.5684	0.6048
guinit	0.4492	0.7889	0.9040	0.8456	0.5299	0.6201
UDInfoCSTc	0.4195	0.6176	0.8110	0.7121	0.5283	0.4133
PRISabc	0.4086	0.7040	0.9521	0.7048	0.5451	0.6083
hplcrating	0.4037	0.7373	0.9473	0.7494	0.5930	0.5947
hplcranking	0.3868	0.7450	0.8857	0.7562	0.5193	0.5499
UDInfoCSTdc	0.3668	0.6713	0.8596	0.7490	0.5011	0.4979
run02K	0.3643	0.7422	0.9410	0.7556	0.5681	0.6409
baselineB	0.3504	0.7470	0.9274	0.7817	0.4384	0.3951
udelnp	0.3395	0.6557	0.9108	0.6636	0.6291	0.3736
iritSplit3CPv2	0.3377	0.6795	0.9072	0.6853	0.4500	0.3841
run01TI	0.3307	0.7136	0.9465	0.7233	0.5692	0.5908
udelp	0.3118	0.6607	0.9131	0.6698	0.5242	0.4067
ICTCONTEXTRUN2	0.3010	0.6579	0.9042	0.6854	0.5748	0.5092
baselineA	0.2993	0.6447	0.8906	0.7002	0.5366	0.4632
ICTCONTEXTRUN1	0.2346	0.5411	0.8514	0.5515	0.4134	0.4655
waterloo12a	0.2130	0.5703	0.8615	0.6119	0.3859	0.4183
UAmsCS12wtSUM	0.1727	0.3140	0.4868	0.5900	0.5374	0.5438
waterloo12b	0.1404	0.5304	0.7447	0.6149	0.2775	0.4467
csiroht	0.1281	0.5719	0.8096	0.6000	0.2774	0.3033
csiroth	0.1237	0.5760	0.8312	0.6121	0.2385	0.2175
FASILKOMUI02	0.1163	0.5789	0.7893	0.6125	0.1865	0.2608
UAmsCS12wtSUMb	0.1058	0.2646	0.4476	0.5813	0.5213	0.6196
FASILKOMUI01	0.0800	0.6561	0.7979	0.7055	0.1093	0.1777
watcs12a	0.0062	0.0395	0.0424	0.8669	0.6543	0.8009
watcs12b	0.0000	0.0416	0.0510	0.6930	0.6421	0.8246

Table 2: All 6 MRR measures sorted by WGT.



(a) MRR WGT vs P@5 WGT. r = 0.97677; $\tau = 0.8730$

(b) P@5 WGT vs P@5 GT. r = 0.76522; $\tau = 0.5535$

(c) P@5 WGT vs P@5 W. r = 0.27899; $\tau = 0.4429$

(d) P@5 W vs P@5 D. r = 0.84979; $\tau = 0.6143$

Figure 2: Scatter plots for different measures plotted against each other. Pearson's r coefficients and Kendall's tau coefficient are also included.