Evaluating Contextual Suggestion

Adriel Dean-Hall
Waterloo

Charles L. A. Clarke
Waterloo

Jaap Kamps
Amsterdam

Paul Thomas
CSIRO

ABSTRACT

As its primary evaluation measure, the TREC 2012 Contextual Suggestion Track used precision@5. Unfortunately, this measure is not ideally suited to the task. The task in this track is different from IR systems where precision@5, and similar measures, could more readily be used. Track participants returned travel suggestions that included brief descriptions, where the availability of these descriptions allows users to quickly skip suggestions that are not of interest to them. A user's reaction to a suggestion could be negative ("dislike"), as well as positive ("like") or neutral, and too many disliked suggestions may cause the user to abandon the results. Neither of these factors are handled appropriately by traditional evaluation methodologies for information retrieval and recommendation. Building on the *time-biased gain* framework of Smucker and Clarke, which recognizes time as a critical element in user modeling for evaluation, we propose a new evaluation measure that directly accommodates these factors.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

geotemporal recommendation, time-biased gain

1. INTRODUCTION

Due to serious limitations in the evaluation methodology for the TREC Contextual Suggestion Track [3], significant differences in system performance may be missed. The track imagines a traveler in a new city. Given a set of the traveler's preferences for places and activities in their home city, participating systems suggested places and activities in the new city. For example, given that the traveler likes the Underground Garage and the Berlin Nightclub in Toronto, a system might suggest the Arrow Bar in New York.

Track participants were given profiles for thirty-four potential travelers, each a twenty-something student from the the greater Toronto area. Each profile indicates the traveler's opinion (like, dislike, or neutral) regarding fifty places and activities in and around the city of Toronto. Track participants were also given fifty geotemporal contexts, each indicating a city in the United States, a day of the week, a time of the day, and a season of the year. For each profile+context pair, participating systems produced a ranked list of fifty suggestions tailored to the traveler's preferences and the geotemporal context. Each suggestion included the name of the place or activity, a brief description, and a URL referencing a webpage providing more information.

Suggestions were then returned to the potential travelers for judging. Judgments (either like, dislike or neutral) were given both after viewing the name/description and after viewing the full webpage. Trained TREC assessors separately judged the appropriateness of the suggestions with respect to the geotemporal contexts.

This contextual suggestion task falls somewhere between traditional information retrieval and traditional recommendation. Unlike traditional information retrieval, the query is fixed ("entertain me" [1]), with the search results varying only to reflect the traveler's profile and the geotemporal context. Unlike traditional recommendation, the range of suggestions is completely open, with the quality of the description forming an important aspect of the user experience. Ideally, this description would be tailored to reflect the preferences of the individual traveler.

Evaluation was based on precision@5. As the primary basis for evaluation, a suggestion was counted as "relevant" if the suggestion was geotemporally appropriate, and if both the name/description and the webpage were liked by the traveler. All other suggestions were counted as "non-relevant". Using these definitions for relevant and non-relevant, precision@5 was computed for each profile+context pair, and then averaged across all pairs. Since resources were not available to judge all profile+context pairs, only forty-four were fully judged and included in this average. This averaged precision@5 value formed the primary basis for inter-system comparisons.

Along with this primary measure, the track reported a number of secondary measures. Some of these secondary measures were also based on precision@5, but used other definitions of relevant and non-relevant, e.g., relevance based only on geotemporal appropriateness. Other secondary measures were based on the mean reciprocal rank (MRR) of the

EVIA'13, June 18, 2013, Tokyo, Japan.

first relevant suggestion, again using a number of definitions for relevant.

Unfortunately, neither precision@5 nor MRR are ideally suited to contextual suggestion. Precision@5 implicitly assumes the user always views exactly five suggestions, never more or less. MRR assumes that the user stops at the first suggestion they like. Both measures ignore the impact of descriptions and negative judgments.

To create a measure that appropriately accommodates these factors, we turn to the time-biased gain (TGB) framework proposed by Smucker and Clarke [8]. This framework uses time-based calibration to account for the impact of user choices and actions. After specializing the framework to create a version of TGB specifically geared to our contextual suggestion task, we apply the measure to re-evaluate experimental runs submitted to the track.

2. EVALUATION MEASURES

Traditionally, the evaluation of recommender systems is based on homogeneous data sets, for example movies, where items in the data set are given ratings by users [4]. Many evaluation techniques do not account for unstructured data from a variety of sources, sometimes without ratings attached to them, where the system makes recommendations based more heavily on the content and less so on ratings and relationships between users.

Traditionally, information retrieval evaluation is based on judgments of document relevance. These judgments are used to compute standard measures such as precision@k, MRR, discounted cumulative gain [5], rank biased precision [6], expected reciprocal rank [2], and many others. All of these measures implicitly assume that the user works their way down a ranked search result list at a fixed rate, eventually stopping, perhaps due to boredom, tiredness, or because they found what they are seeking [2, 6, 8, 9]. None of these measures appropriately account for document length, duplicate documents, and snippets (i.e., short captions describing a document, which may allow non-relevant results to be quickly skipped). Relevance is generally viewed in positive terms only, indicating the degree to which a user likes a document.

3. TIME-BIASED GAIN

To accommodate the limitations of traditional information retrieval evaluation measures, Smucker and Clarke [8] introduced time-biased gain (TBG). A general form of TBG may be written as the Riemann-Stieltjes integral:

$$\int_0^\infty D(t)dG(t). \tag{1}$$

This equation assumes the user is working through a ranked list of retrieval results, reading documents, viewing videos, considering suggestions, or performing whatever other actions are appropriate to the retrieval task at hand. The function $G(t)$ represents the *cumulative gain*, or benefit, received by the user as time passes.

The *decay* function $D(t)$ indicates the probability that the user continues until time t. This function represents the possibility that the user will stop at some point due to factors such as tiredness or boredom, rather than due to the influence of the results themselves. Based on an analysis of a log from a commercial search engine, Smucker and Clarke suggest an exponential decay function with a half-life of 224 seconds. In the absence of other information, we adopt the same decay function for our version of TBG.

When gain is realized as a step function, e.g., increasing by a fixed amount when the the user views a suggestion they like, Equation 1 may be re-expressed as sum over documents, suggestions, or other discrete retrieval items:

$$\sum_{k=1}^{\infty} g_k D(T(k)). \tag{2}$$

In this equation, g_k represents the gain realized from the kth item. In the case of contextual suggestion, we measure gain as the number of suggested webpages the user views and likes. The function $T(k)$ represents the time it takes the user to reach rank k. The decay function is applied to this time to determine the proportion of users who reach rank k.

In the next two subsections, we provide simple estimates for g_k and $T(k)$. While both estimates are based on relatively crude user models, they illustrate reasonable methods for accommodating the impact of descriptions and negative judgments. Extension and further validation of these models is left for future work.

3.1 Likes and Dislikes

To estimate g_k we borrow an idea from the *cascade model* of browsing behavior for search results [2, 9]. Under the cascade model, the gain realized at rank k depends on the relevance of documents appearing at ranks 1 to $k-1$. As more relevant documents are seen by the user, the more likely they are to stop browsing, since their information need may be satisfied.

For contextual suggestion, we use a cascade-like model to account for disliked suggestions. As more disliked suggestions are seen, the more likely the user stops browsing. This stopping probability is operationalized by attenuating the gain obtained from a liked suggestion as more and more disliked suggestions are seen.

We define a function indicating if the user *likes* the suggestion at rank k as follows:

$$A(k) = \begin{cases} 1, & \text{if the user likes or is neutral about the description at rank } k \text{ and also likes the (geotemporally appropriate) webpage at rank } k \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the user likes a suggestion only if they don't dislike the description, and after clicking through to the webpage, they like it. Geotemporal appropriateness is considered only at the webpage level: the user never likes a webpage unless it is geotemporally appropriate. We define a function indicating if the user *dislikes* the suggestion at rank k as follows:

$$Z(k) = \begin{cases} 1, & \text{if the user dislikes the description at rank } k \text{ or if ther user likes or is neutral about the description but dislikes the webpage} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the user dislikes a suggestion if they dislike the description, but also can dislike a suggestion if they don't dislike the description, but after clicking through to the webpage, they end up disliking the webpage.

We now define g_k in terms of the user's likes and dislikes as they browse a ranked list of suggestions, as:

$$g_k = A(k)(1-\theta)^{\sum_{j=1}^{k-1} Z(j)}. \quad (3)$$

If the user views and likes the suggestion at rank k they receive a gain of 1, but this gain is attenuated according to the number of disliked suggestions seen at ranks 1 to $k-1$. The parameter θ ($0 < \theta < 1$) indicates the probability that the user will stop browsing after viewing a disliked suggestion. Note that under this model neutral documents have no impact nor any gain. In the absence of other information, we adopt a value of $\theta = 0.5$.

3.2 Time to Reach a Given Rank

The time to reach rank k, $T(k)$ may be estimated from actual user behaviour captured during the judging process. Using timing logs from potential travelers, we compute the mean time it takes for users to read a description (T_D) and the mean time it takes users to examine a webpage (T_W). Often the suggested webpage is the front page of a larger site describing the suggestion, and may contain Flash, banners, etc., with little detailed information. While examining webpages, potential travelers may have looked only at the webpage suggested by the system, or may have clicked through to additional linked pages. Examination of these additional pages is included in the times to examine the suggested webpage.

In estimating the time taken to examine a document, Smucker and Clarke [8] consider the document's length. Since users are allowed to click through to other web pages, we do not parameterize by document length. In addition, since descriptions were limited, by the task, to 512 characters, and are generally close to that length, the time to judge descriptions was also not parameterized by length.

As part of the TREC judging process, users clicked through to every website regardless of whether or not they liked the description. In building our model, we assume real users would be exhibit different behaviour, only clicking through to pages with a description they like. Under our model, users read every description, and if they like a description they will click through to the website and examine it. Thus, the time to reach rank k may be expressed as:

$$T(k) = \sum_{j=1}^{k-1} T_D + l_j T_W. \quad (4)$$

where l_j is 1 if the user likes the description at rank j, and therefore examines the webpage, and 0 otherwise.

We estimate the mean time it takes users to judge a description as $T_D = 7.45$ sec, and the mean time it takes to judge a website as $T_W = 8.49$ sec. These estimates were calculated by excluding the slowest 10% of judgments. Removing the slowest judgments from our calculation considerably tightened the standard deviation around the mean due, in part, to the elimination of situations such as the assessor taking a break from judging (thus taking a long time to move from one judgment to the next).

4. EXPERIMENTAL COMPARISON

The goal of our experiment is to compare precision@5 used in the TREC track to our modified version of TBG. Using the description and website judgments from the TREC track, we gave each of the 27 experimental runs submitted to the track a rank based on their mean score on TBG across all the topics that were judged. For both the TBG and precision@5 rankings we used the judgments given for the description and website, as well as the judgments by the trained TREC accessors given for the geographical and temporal relevance. The primary measure used in the TREC track was precision@5, where a document was considered relevant if it was judged relevant on both the website and geotemporal ratings. Note that judgments are only available up to rank 5 so the final TBG for each profile+context pair is a sum from rank 1 to 5 using equation 2.

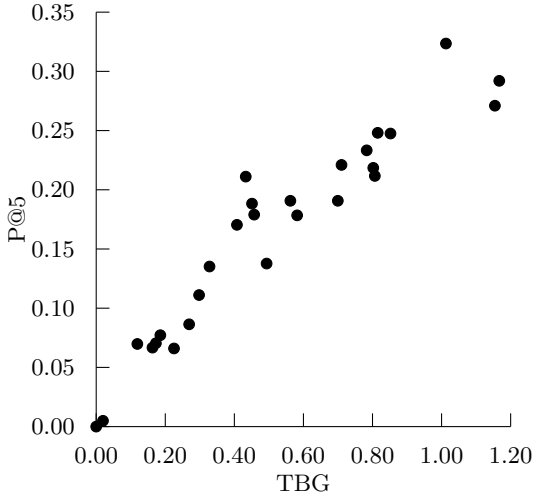

Figure 1: Time Biased Gain vs. P@5. $\tau = 0.85$

Table 1 shows the difference in rankings between TBG and precision@5. Figure 1 shows a scatter plot of the rankings plotted against each other, the Kendall tau ranking correlation is $\tau = 0.85$. The majority of the runs (63%) stayed in the same position or shifted by one position, however there were some large shifts as well with the largest change being a move down by 6 positions by the udelp run.

We also compare the discriminative power of the two methods. Sakai [7] proposed a method for calculated discriminative power where the the number of pairs that are significant at a certain significance level for a given significance test is used. In our case, we used the paired t-test. The discriminative power of precision@5 which is 58.4% (using a significance level of 0.05). We compare this to the discriminative power of TBG which is 59.8%, a slight improvement.

For the TREC 2012 results we can see the gain we (calculated by TBG) users see on average for the best run is about 1.2, which is very low considering that 5 results were shown to users. P@5 also has low results. For the results 56% of the ranked lists have no gain in TBG and 53% have a P@5 of 0. In P@5 is 0 that means the suggestion was not geotemporally appropriate and the website was not liked by the user. In this case the TBG will also be 0, so anytime P@5 is 0 TBG is also 0. Note that when TBG is 0 P@5 is not necessarily 0 because TBG also has an extra restriction that the description cannot be disliked. Due to the low number of relevant suggestions in the dataset the P@5 and TBG scores often match, which brings the Kendal tau and discriminative power numbers for the two measures close together.

Run	TBG Score	P@5 Score	Diff
guinit	1.1670	0.2920	1
gufinal	1.1544	0.2710	1
iritSplit3CPv1	1.0126	0.3235	-2
PRISabc	0.8521	0.2475	1
UDInfoCSTc	0.8151	0.2481	-1
hplcrating	0.8068	0.2117	3
run02K	0.8022	0.2185	1
hplcranking	0.7832	0.2333	-2
UDInfoCSTdc	0.7103	0.2210	-2
run01TI	0.6996	0.1907	1
baselineA	0.5818	0.1784	4
ICTCONTEXTRUN2	0.5622	0.1907	0
waterloo12a	0.4934	0.1377	4
iritSplit3CPv2	0.4574	0.1790	0
udelnp	0.4511	0.1883	-2
udelp	0.4330	0.2111	-6
baselineB	0.4075	0.1704	-1
UAmsCS12wtSUM	0.3281	0.1352	0
ICTCONTEXTRUN1	0.2979	0.1111	0
waterloo12b	0.2691	0.0864	0
FASILKOMUI01	0.2253	0.0660	4
csiroth	0.1857	0.0772	-1
UAmsCS12wtSUMb	0.1728	0.0704	-1
FASILKOMUI02	0.1629	0.0667	0
csiroht	0.1191	0.0698	-2
watcs12a	0.0196	0.0049	0
watcs12b	0.0000	0.0000	0

Table 1: Ranking of the TREC runs ordered by TBG scores and compared to precision at 5 scores.

5. CONCLUSIONS

Time-biased gain [8] (TBG) provides a general framework for information retrieval evaluation, allowing evaluation measures to better reflect the impact of user interfaces and user behavior. We propose a model for users of a contextual suggestion system, and apply this model to create a version of time-biased gain [8] for these system. Our version of TBG accounts for the impact of descriptions and disliked suggestions, which are ignored by the official track measures. The model assumes a user working their way through a ranked list of suggestions, pausing to investigate the webpages associated with descriptions they like. Gain — or benefit to the user — is recognized after the user views an geotemporally appropriate webpage they like. Disliked suggestions may cause the user to stop browsing, attenuating the gain for later suggestions. Following Smucker and Clarke [8], we adopt an exponential decay function to model the possibility that the user will stop due to factors unrelated to the suggestions themselves, such as boredom or tiredness on the part of the user.

The model has four parameters. Two parameters, associated with the time to read descriptions (T_D) and the time to read Web pages (T_W) are estimated from user data collected during the TREC judging process. The half-life for the decay function is taken from Smucker and Clarke. Only the parameter θ, representing attenuation due to disliked suggestions, is completely arbitrary. We adopt a value of $\theta = 0.5$, but other values are reasonable. Lower values would reflect a more tolerant user; higher values would reflect a less tolerant user.

Our user model is essentially the simplest possible model that can reasonably accommodate the impact of descriptions and disliked suggestions. However, when incorporated into time-biased gain, our model is sufficient to identify differences between runs that were missed by precision@5, and similar measures. In future work, we plan to extend the model, improving its calibration and validating it against additional user data.

6. ACKNOWLEDGEMENTS

We thank track participants for making the track a success. We thank NIST, and particularly Ellen Voorhees, for their help and resources. This research was funded by Google, the GRAND Network of Centres of Excellence, and the Natural Sciences and Engineering Research Council of Canada.

7. REFERENCES

[1] Nicholas J. Belkin, Charles L.A. Clarke, Ning Gao, Jaap Kamps, and Jussi Karlgren. Report on the SIGIR workshop on "entertain me": Supporting complex search tasks. *SIGIR Forum*, 45(2):51–59, December 2012.

[2] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *18th ACM Conference on Information and Knowledge Management*, pages 621–630, Hong Kong, 2009.

[3] Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Paul Thomas, and Ellen Voorhees. Overview of the TREC 2012 contextual suggestion track. In *21st Text REtrieval Conference*, Gaithersburg, Maryland, 2012.

[4] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[5] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[6] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):2:1–2:27, December 2008.

[7] Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 525–532, New York, NY, USA, 2006. ACM.

[8] Mark D. Smucker and Charles L.A. Clarke. Time-based calibration of effectiveness measures. In *35th International ACM SIGIR Conference on Research and Development in information retrieval*, pages 95–104, Portland, Oregon, 2012.

[9] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected browsing utility for web search evaluation. In *19th ACM International Conference on Information and Knowledge Management*, pages 1561–1564, Toronto, 2010.