UFMG at the TREC 2016 Dynamic Domain track

Felipe Moraes[1], Rodrygo L. T. Santos[1], Nivio Ziviani[1,2]
{felipemoraes, rodrygo, nivio}@dcc.ufmg.br

[1]Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil

[2]Kunumi
Belo Horizonte, MG, Brazil

ABSTRACT

In TREC 2016, we focus on tackling the challenges posed by the Dynamic Domain (DD) track. The goal of the TREC DD track is to support research in dynamic, exploratory search within a complex domain. To this end, our participation investigates the suitability of multiple diversification approaches for dynamic information retrieval. In particular, based on fine-grained real-time feedback obtained from a simulated user, we apply diversification strategies that make use of single-source as well multi-source information for mining flat or hierarchical query aspects.

1. INTRODUCTION

The TREC 2016 Dynamic Domain (DD) track [5] consists of performing an interactive search task in a specific domain. In particular, the goal of this track is to cover multiple aspects of the user's information need as early as possible. Therefore, in this track, each participant relies on a simulated user (called "the jig") to get real-time feedback on a short list of documents that the participant's system returns. This process repeats until the participant's system decides that the user has fulfilled his or her information need.

The central theme of our participation in the TREC 2016 DD track is to investigate the suitability of search result diversification as a mechanism to increase the coverage of the possible aspects underlying the user's query while requiring as little as possible user interaction. We make use of the exploratory and explicit nature of the user's inputs, whereby he or she directly assists our system in assigning aspects to documents passages. In particular, we investigate approaches for diversification that leverage either flat or hierarchical aspects from a single or from multiple sources. Along with the diversification approaches, we investigate three strategies proposed in the literature for reducing the number of iterations of the retrieval-feedback cycle needed to fulfill the aspects that are of interest to the user.

The remainder of this paper is structured as follows. In Section 2, we briefly formalize the task tackled in the TREC 2016 DD track and provide an overview of our system. In Section 3, we discuss the results of our official and unofficial runs. In Section 4, we present our conclusions.

2. INTERACTIVE DIVERSIFICATION

The TREC DD track invites participants to propose search systems capable of interactively satisfying the multiple aspects associated with the information need of a (simulated) target user. Given a query q posed by the target user, a search system may return a batch $\mathcal{D}$ of documents to the user. For each document $d \in \mathcal{D}$, the user provides the system with a feedback set $\mathcal{F}$. Each feedback $f \in \mathcal{F}$ is defined as a tuple $f = \langle p, a, g \rangle$ comprising a passage p that the user deemed relevant to aspect a at a given relevance level g. The search system may then return another batch of documents for more feedback or decide to stop the search process.

Our approach has three components: (i) a baseline ranker, which retrieves a candidate set of documents given a query q; (ii) a diversification strategy, which interactively re-ranks the candidate set to improve the coverage of the possible aspects underlying q; and (iii) a stopping strategy, which determines when to stop the interactive process.

2.1 Baseline Ranker

In our experiments, we use Elasticsearch[1] for both indexing and retrieval, after applying Krovetz stemmer and removing standard English stopwords. In particular, we index the title, content, and anchor-text of each document. As a baseline ranker, we use a field-based query likelihood model with Dirichlet smoothing. We set $\mu = 2,500$, and the weights for the title, content, and anchor-text fields to 0.3, 0.7, and 0.1, respectively. For each query, we retrieve the top 1,000 documents as a candidate set $\mathcal{R}$.

2.2 Diversification Strategies

Given a query q and an initial ranking $\mathcal{R}$ produced for this query, we build a new ranking $\mathcal{D}$ by iteratively selecting the highest scored document $d \in \mathcal{R} \setminus \mathcal{D}$ according to:

$$d^* = \underset{d \in \mathcal{R} \setminus \mathcal{D}}{\arg\max}\ (1 - \lambda) rel(q, d) + \lambda div(q, d, \mathcal{D}), \qquad (1)$$

where $rel(q, d)$ and $div(q, d, \mathcal{D})$ are the estimated relevance of d given q and the estimated diversity of d given the already selected documents in $\mathcal{D}$. While $rel(q, d)$ is estimated by the aforementioned relevance ranker, we experiment with multiple diversification strategies for estimating $div(q, d, \mathcal{D})$.

2.2.1 Flat Diversification

Santos et al. [4] introduced the xQuAD framework, which estimates $div(q, d, \mathcal{D})$ as the probability that document d covers explicitly identified aspects $\mathcal{S}$ underlying the query q that are not well covered by the already selected documents

[1] https://www.elastic.co/products/elasticsearch

in $\mathcal{D}$. Precisely, xQuAD defines $div(q, d, \mathcal{D})$ as follows:

$$div_X(q, d, \mathcal{D}) = \sum_{s \in \mathcal{S}} P(s|q)\, P(d|q, s) \prod_{d_j \in \mathcal{D}} (1 - P(d_j|q, s)), \tag{2}$$

where $P(s|q)$ denotes the relative importance of aspect s given q, $P(d|q, s)$ denotes the coverage of document d with respect to this aspect, and the rightmost product denotes the novelty of any document covering this aspect, based upon how badly this aspect is covered by documents in $\mathcal{D}$.

2.2.2 Hierarchical Diversification

Hu et al. [2] proposed HxQuAD as an extension of the xQuAD framework to support diversification using hierarchically organized aspects. Such a hierarchy can be modeled as a tree in which each node represents an aspect and the set of aspects at the i-th level of the tree is denoted $\mathcal{S}_i$. Likewise, we can define the diversity of a document d with respect to aspects at the i-th level according to:

$$div_i(q, d, \mathcal{D}) = \sum_{s \in \mathcal{S}_i} P(s|q)\, P(d|q, s) \prod_{d_j \in \mathcal{D}} (1 - P(d_j|q, s)), \tag{3}$$

where both $P(s|q)$ and $P(d|q, s)$ are defined recursively, so that, for any non-leaf aspect s with children $\mathcal{C}$, we have:

$$P(s|q) = \sum_{c \in \mathcal{C}} P(c|q) \quad \text{and} \quad P(d|q, s) = 1 - \prod_{c \in \mathcal{C}} (1 - P(d|q, c)). \tag{4}$$

Given these definitions, HxQuAD estimates the overall diversity of document d by linearly combining its diversity estimates at multiple hierarchy levels, according to:

$$div_H(q, d, \mathcal{D}) = \alpha div_1(q, d, \mathcal{D}) + (1 - \alpha) div_2(q, d, \mathcal{D}) + \frac{(1 - \alpha)^2}{\alpha} div_3(q, d, \mathcal{D}) + ... + \frac{(1 - \alpha)^{n-1}}{\alpha^{n-2}} div_n(q, d, \mathcal{D}), \tag{5}$$

where the α hyperparameter controls the influence of different hierarchy levels in the final estimation, with $\alpha = 0.5$ indicating that all levels are equally weighted.

2.2.3 Multi-Dimensional Diversification

Both xQuAD (Equation (2)) and HxQuAD (Equation (5)) assume that query aspects are mined from a single source. Inspired by Dou et al. [1], we extend both frameworks to leverage aspects from multiple sources $k \in \mathcal{K}$, such that:

$$div_M(q, d, \mathcal{D}) = \sum_{k \in \mathcal{K}} \theta_k\, div_k(q, d, \mathcal{D}), \tag{6}$$

where $div_k(q, d, \mathcal{D})$ denotes either Equation (2) or Equation (5) leveraging aspects from source k with corresponding weight θ_k. In particular, in our experiments, we consider two sources of query aspects, as described next.

2.2.4 Aspect Mining

Similarly to previous work [4], we use query suggestions provided by a major Web search engine as a source of query aspects. Following Hu et al. [2], we only consider two-level hierarchical aspects. In particular, for each query, we collect a first level of aspects by mining suggestions for the initial query. Then, we generate a second level of aspects by mining query suggestions for each first-level aspect.

In addition to query suggestions, which provide a static surrogate for the actual aspects of interest to the user, we consider a second aspect source built by directly leveraging the user's feedback. In particular, recall that, for each retrieved document d, a user's feedback $f = \langle p, a, g \rangle$ includes a passage p that the user deemed relevant to aspect a at a given relevance level g. For the flat diversification performed by xQuAD (Equation (2)), because multiple passages can be deemed relevant to the same aspect, we estimate the coverage of document d with respect to aspect a as follows:

$$P(d|q, a) = \max_{p \in a} P(d|p), \tag{7}$$

where $P(d|p)$ denotes the coverage of passage p by document d. Before the user's feedback is received, this probability is estimated proportionally to the cosine between tf-idf representations of both the passage and the document. Afterwards, this probability is estimated proportionally to the relevance level g directly assigned by the user. In contrast, for HxQuAD (Equation (5)), we directly leverage the hierarchical relationship between the aspect a and each of its associated passages $p \in a$ as contributed by the user.

2.3 Stopping Strategies

Based on stopping strategies investigated by Maxwell et al. [3], we evaluate three strategies in our participation:

Fixed count. Our system will stop once we have returned x_1 documents to the simulated user. We set $x_1 = 50$, which corresponds to 10 iterations.

Cumulative off-topic count. Our system will stop once we have returned a total of x_2 off-topic documents to the simulated user. We set $x_2 = 10$, which corresponds to off-topic documents worth of 2 iterations.

Windowed off-topic count. Our system will stop once we have returned x_3 off-topic documents in a contiguous window. We set $x_3 = 10$, which corresponds to a window of 2 iterations.

3. EXPERIMENTS

3.1 Runs Summary

We produced a total of 12 runs in our participation in the TREC DD track, five of which were officially submitted:

- ufmgXS1 (unofficial) applies flat diversification with single-source aspects and the fixed count stopping condition. We used real-time user feedback as aspects source and xQuAD diversification parameter $\lambda = 0.8$.

- ufmgXS2 (submitted) is similar to ufmgXS1, except that it applies the cumulative stopping condition.

- ufmgXS3 (unofficial) is similar to ufmgXS1, except that it applies the windowed stopping condition.

- ufmgXM1 (unofficial) applies flat diversification with multi-source aspects and fixed count stopping condition. We use xQuAD diversification parameter $\lambda = 0.8$. At the first iteration, we used only search engine's suggested aspects and in the remaining iterations we used real-time user feedback aspects.

Table 1: Results of our runs on TREC 2016 Dynamic Domain track's official measures

Run	Submitted	Description			Avg Cube Test			Cube Test		
		Model	Source	Stop	@1	@2	@10	@1	@2	@10
TREC avg	N/A	N/A	N/A	N/A	0.1472	0.1361	0.1045	0.2049	0.1388	0.0849
TREC median	N/A	N/A	N/A	N/A	0.1516	0.1352	0.0985	0.2174	0.1281	0.0801
ufmgXS1	✗	X	F	1	**0.1751**	0.1587	0.0877	0.2309	0.1506	0.0452
ufmgXS2	✓	X	F	2	**0.1751**	0.1587	0.1188	0.2309	0.1506	0.0786
ufmgXS3	✗	X	F	3	**0.1751**	0.1587	0.1146	0.2309	0.1506	0.0764
ufmgXM1	✗	X	F+S	1	0.1750	**0.1612**	0.0882	**0.2474**	0.1574	0.0452
ufmgXM2	✓	X	F+S	2	0.1750	**0.1612**	**0.1237**	**0.2474**	0.1574	**0.0852**
ufmgXM3	✗	X	F+S	3	0.1750	**0.1612**	0.1169	**0.2474**	0.1574	0.0797
ufmgHS1	✗	H	F	1	**0.1751**	0.1601	0.0892	0.2309	**0.1578**	0.0457
ufmgHS2	✓	H	F	2	**0.1751**	0.1601	0.1219	0.2309	**0.1578**	0.0801
ufmgHS3	✗	H	F	3	**0.1751**	0.1601	0.1172	0.2309	**0.1578**	0.0763
ufmgHM1	✗	H	F+S	1	**0.1751**	0.1601	0.0893	0.2309	0.1577	0.0463
ufmgHM2	✓	H	F+S	2	**0.1751**	0.1601	0.1219	0.2309	0.1577	0.0842
ufmgHM3	✓	H	F+S	3	**0.1751**	0.1601	0.1181	0.2309	0.1577	0.0808

Table 2: Results of our runs on TREC 2016 Dynamic Domain track's unofficial measures

Run	α-nDCG			nERR-IA			P@R		
	@1	@2	@10	@1	@2	@10	@1	@2	@10
TREC avg	0.2999	0.3339	0.3552	0.2821	0.2985	0.3048	0.3486	0.3387	0.2957
TREC median	0.2952	0.3142	0.3142	0.2691	0.2777	0.2778	0.3208	0.3038	0.2811
ufmgXS1	0.3516	0.3987	0.4852	**0.3383**	0.3622	0.3849	0.4000	**0.4283**	0.3547
ufmgXS2	0.3516	0.3987	0.4324	**0.3383**	0.3622	0.3723	0.4000	**0.4283**	**0.3871**
ufmgXS3	0.3516	0.3987	0.4458	**0.3383**	0.3622	0.3759	0.4000	**0.4283**	0.3787
ufmgXM1	**0.3559**	0.4028	0.4786	0.3355	0.3588	0.3786	**0.4226**	0.4038	0.3000
ufmgXM2	**0.3559**	0.4028	0.4265	0.3355	0.3588	0.3667	**0.4226**	0.4038	0.3504
ufmgXM3	**0.3559**	0.4028	0.4364	0.3355	0.3588	0.3698	**0.4226**	0.4038	0.3452
ufmgHS1	0.3516	**0.4079**	**0.4921**	**0.3383**	**0.3664**	**0.3887**	0.4000	0.4075	0.3479
ufmgHS2	0.3516	**0.4079**	0.4367	**0.3383**	**0.3664**	0.3758	0.4000	0.4075	0.3794
ufmgHS3	0.3516	**0.4079**	0.4504	**0.3383**	**0.3664**	0.3794	0.4000	0.4075	0.3714
ufmgHM1	0.3516	0.4055	0.4897	**0.3383**	0.3653	0.3877	0.4000	0.4075	0.3335
ufmgHM2	0.3516	0.4055	0.4367	**0.3383**	0.3653	0.3754	0.4000	0.4075	0.3716
ufmgHM3	0.3516	0.4055	0.4481	**0.3383**	0.3653	0.3784	0.4000	0.4075	0.3654

- ufmgXM2 (submitted) is similar to ufmgXM1, except that it applies the cumulative stopping condition.

- ufmgXM3 (unofficial) is similar to ufmgXM1, except that it applies the windowed stopping condition.

- ufmgHS1 (unofficial) applies hierarchical diversification with single-source aspects and the fixed count stopping condition. We used real-time user feedback as aspects source and HxQuAD diversification parameters $\lambda = 0.8$ and $\alpha = 1.0$.

- ufmgHS2 (submitted) is similar to ufmgHS1, except that it applies the cumulative stopping condition.

- ufmgHS3 (unofficial) is similar to ufmgHS1, except that it applies the windowed stopping condition.

- ufmgHM1 (unofficial) applies hierarchical diversification with multi-source aspects and the fixed count stopping condition. We used HxQuAD diversification parameters $\lambda = 0.8$ and $\alpha = 1.0$. At the first iteration, we used only search engine's suggested aspects and in the remaining iterations we used real-time user feedback aspects.

- ufmgHM2 (submitted) is similar to ufmgHS1, except that it applies the cumulative stopping condition.

- ufmgHM3 (submitted) is similar to ufmgHS1, except that it applies the windowed stopping condition.

3.2 Results

Table 1 shows the results of our unofficial as well as our officially submitted runs to this task for the official measures Average Cube Test at the 1st, 2nd and 10th iterations. Firstly, we organize the table into models with X and H standing for xQuAD and HxQuAD. Secondly, we divide the table into single-source and multi-source aspects in which F and S mean user real-time feedback and query suggestions, respectively. Finally, we distinguish between the three considered stopping condition strategies where 1, 2 and 3 indicate fixed, cumulative off-topic and windowed off-topic count, respectively. Table 2 shows the results of our runs for unofficial measures and is organized in the same

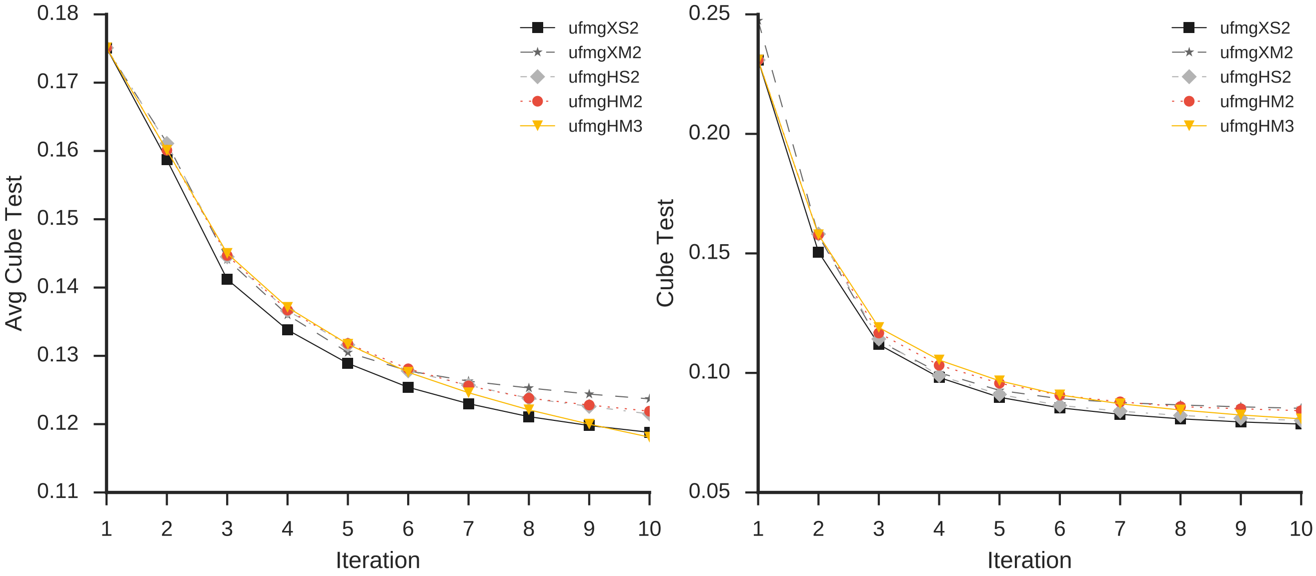

Figure 1: Cube Test and Average Cube Test over the iterations (averaged for all topics) for our official runs.

way as Table 2. Additionally, Figure 1 shows our system progress over the iterations for our officially submitted runs. We summarize our findings as follows:

- Our best result uses flat diversification with multi-source aspects for the Cube Test, α-nDCG and P@R measures at the 1st iteration and Average Cube Test and P@R at the 2nd iteration. We note improvements regarding Cube Test, α-nDCG and nERR-IA at the 2nd iteration with hierarchical diversification with single-source aspects.

- Our best result in terms of Average Cube Test and Cube Test at the 10th iteration uses hierarchical diversification with multi-source aspects and a cumulative stopping condition. However, in terms of P@R, our best result at the 10th iteration uses flat diversification with single-source aspects with cumulative stopping condition. In terms of nERR-IA and α-nDCG, our best result at the 10th iteration uses hierarchical diversification with single-source aspects with a fixed count stopping condition.

- Based on the official measures, our best officially submitted run was ufmgXM2. In Figure 1, we show how our officially submitted runs perform over all 10 iterations. For the Average Cube Test measure, ufmgXM2 performs distinctively well toward the last iterations, whereas for Cube Test, it outperforms the other alternatives toward the first iterations.

4. CONCLUSIONS

In TREC 2016, we participated in the Dynamic Domain track. Our participation focused around diversification approaches, as well as stopping strategies to fulfill the information need of the user as early as possible. Overall, our results on diversification with multi-source aspects in an interactive search process ranked high across iterations compared with the runs submitted by other participants in this track.

5. ACKNOWLEDGMENTS

This work was partially funded by projects InWeb (MCT/CNPq 573871/2008-6) and MASWeb (FAPEMIG/PRONEX APQ-01400-14), and by the authors' individual grants from CNPq and FAPEMIG.

6. REFERENCES

[1] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proc. of WSDM*, pages 475–484, 2011.

[2] S. Hu, Z. Dou, X. Wang, T. Sakai, and J.-R. Wen. Search result diversification based on hierarchical intents. In *Proc. of CIKM*, pages 63–72, 2015.

[3] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. Searching and stopping: An analysis of stopping rules and strategies. In *Proc. of CIKM*, pages 313–322, 2015.

[4] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proc. of WWW*, pages 881–890, 2010.

[5] H. Yang, J. Frank, and I. Soboroff. Trec 2015 dynamic domain track overview. In *Proc. of TREC*, 2015.