Responses to the Reviewer Comments

Dear Editors and Reviewers,

Thank you for your letter and for the reviewers' comments concerning our manuscript entitled "A Learnable Search Result Diversification Method" (NO) ESWA-D-17-02634). Those comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our researches. We have studied the comments carefully and the responses to the reviewers' comments are listed as follows:

Reviewer 1#

Comment 1: I am cautious about the first highlight, given e.g., the works of [Zhu, Y., Lan, Y., Guo, J., Cheng, X. and Niu, S., 2014, July. Learning for search result diversification. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 293-302). ACM.] & [Xia, L., Xu, J., Lan, Y., Guo, J. and Cheng, X., 2015, August. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 113-122). ACM.] & [Xia, L., Xu, J., Lan, Y., Guo, J. and Cheng, X., 2016, July. Modeling Document Novelty with Neural Tensor Network for Search Result Diversification. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 395-404). ACM.].I think that claiming that this is the work that introduces "learning" to the field is not accurate.

Answer 1: In this paper, we treat the query aspect diversification as a learning problem and propose a Learnable Search Result Diversification method. We incorporate various features into diversity measurement based on the Markov Random Field, which enables the integration of various types of features. The values of parameters can be determined automatically, which saves the manual labour, and the parameters are more optimal. Based on the comment, we have revised the paper in Page 5, Line 85 as follows:

Our approach also accounts for the aspects of a query in an explicit way. However, differently from the existing approaches, we use a learnable process to identify features from documents using Markov Random Field.

It is other than using structural SVM to learn to identify a document subset with maximum word coverage (Zhu et al. (2014)). They just learn the maximum word coverage and do not mine the aspects underlying the query. However, we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation to resolve this bidirectional optimization problem.

Comment 2: In Page 3, last line "We categorize them into two groups...", I suggest that the categorizations originate from authors. I may recommend to rephrase. The related works sections does not deliver a comprehensible taxonomy of the existing approaches, hence the positioning of the paper in the field is not clear.

Answer 2: In this paper, we categorize the existing approaches into either implicit or explicit basing on whether they account for the query aspects or not. Based on the comment, we have revised the related word in Section 2, Page 3 as follows:

Search result diversification can be characterized as a bidirectional optimization problem, in which one seeks to maximize the overall relevance of a document ranking to multiple query aspects, while minimizing its redundancy (Santos et al. (2010b)). In particular, the existing approaches can be categorized as either implicit or explicit making a difference in how they account for the query aspects (Santos et al. (2010c)).

The basic assumption of implicit diversification approaches is that dissimilar documents are more likely to satisfy different information needs. The most representative approach in maximal marginal relevance (MMR) method and its probabilistic variants is shown as follows (Zhai et al. (2003)):

$$S_{MMR}(q,d,c) = (1-\lambda)S^{rel}(d,q) - \lambda \max_{d_j \in C} S^{div}(d,d_j) \qquad (1)$$

Where S^{rel} and S^{div} represents document $d's$ relevance to the query q and its similarity to a selected document d_j respectively. To gain high ranking score, a document should not only be relevant, but also be dissimilar from the selected documents. The special process of MMR proposed by Carbonell & Goldstein is selecting the document iteratively Carbonell & Goldstein (2009), and meanwhile, both content-based relevance and diversity relation between current selected document and the previously selected documents are considered. Yun et al. (2017) formulate this as a process of selecting and ranking k exemplar

documents and utilize linear programming to solve this problem. In summary, they are all implicit approaches without using aspects to mine the underlying aspects, besides, they are a low effective approaches (Santos et al. (2010); Drosou & Pitoura (2010)).

Explicit approaches make use of the aspects underlying the query to select documents that cover different aspects as far as possible. The algorithms such as IA-select, xQuAD and RxQuAD are proposed to reduce redundancy on the aspect levels (Agrawal et al. (2009);Santos et al. (2010); Vargas et al. (2012)). These methods select documents that cover more novel aspects. The PM-1 and PM-2 models pay more attention to maintain the proportionality of aspects (Dang & Croft (2012)). They produce the ranked result according to the proportionality of aspects. Intrinsic diversity products a series of successor queries to figure out the appropriate content to cover (Raman et al. (2013)). Wang et al. think the aspects underlying the query should be hierarchical, and propose some hierarchical measures to find the relationships among aspects (Wang et al. (2016)). To conclude, all existing explicit approaches are unsupervised, and the values of parameters need to be tuned by the experiment repeatedly without intention, causing a time-consuming optimizing problem to find the most suitable parameters.

Our approach also accounts for the aspects of a query in an explicit way. However, differently from the existing approaches, we use a learnable process to identify features from documents by Markov Random Field. It is other than using structural SVM to learn to identify a document subset with maximum word coverage (Zhu et al. (2014)). They just learn the maximum word coverage and do not mine the aspects underlying the query. Besides, we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation to resolve this bidirectional optimization problem.

Comment 3: Continuing with novelty, I am not sure about the differences of § 3.2 with xQuAD. Authors cite of course Santos et al., 2010, but it is not clear how original is this first stage of their approach. Is it all about the use of MRF of there are more? In addition, there are several works that have proposed MRF for feature extraction, so I'm not confident about the novelty here.

Answer 3: xQuAD is a greedy approximation. It sequentially selects the “local-best” document as each step. In this paper, we redefine the

diversity function and find a learnable process to determine the values of parameters. That is novelty. The detail description is shown in Section 3.2, Page 5 as follows:

Traditional topic diversity model is a greedy approximation Santos et al. (2010). It sequentially selects the "local-best" document from the candidate document set. The original function is formalized as follows:

$$f(d,\bar{S}) = (1-\lambda)P(d|q) + \lambda\sum_{q_i \in Q} P(q_i|q)P(d|q_i)P(\bar{S}|q_i) \quad (2)$$

where d denotes for the current document to be considered in the sequential process, $\bar{S}$ denotes for the unselected document set (equal to the $D \setminus S$ in Fig. (1)), q denotes for the query, λ is a balance parameter for a trade-off between relevance and diversity, q_i denotes for the aspects underlying the query q.

As for Eq. (2), the left part corresponds to the relevance score and the right part corresponds to the diversity score. We look forward to redefine the estimation of diversity score $P(\bar{S}|q_i)$. According to the conditional probabilistic formula, the task can be formalized as follows:

$$P(\bar{S}|q_i) = \frac{P(\bar{S},q_i)}{P(q_i)} \overline{\overline{rank}} P(\bar{S},q_i) \quad (3)$$

Where $P(q_i)$ denotes the occurrence rate of aspects q_i corresponding to query q, which is usually regard to be normalized as $1 \setminus n$ (n denotes the number of aspects) Santos et al. (2010). Because the values of $P(q_i)$ are equal and do not impact on the result of ranking, we neglect $P(q_i)$.

The main concern is how to define feature function for $P(\bar{S},q_i)$. There are many ways to integrate different features, just like linear regression, logistic regression and some other ways. Under our situation, we use Markov Random Field (MRF for short) to model $P(\bar{S},q_i)$. We can benefit from its convenient combination of different types of features and we can get its derivation easily.

Comment 4: Next, since authors give a lot of importance on comparisons, I have to note that comparisons with more contemporary approaches are missing. For example, comparisons with intent-aware search result diversification algorithms (e.g., Hu, S., Dou, Z., Wang, X., Sakai, T. and Wen, J.R., 2015, October. Search result diversification based on

hierarchical intents. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 63-72). ACM)

Answer 4: Based on the comment, we have introduced the HxQuAD for comparison. The results show that our method outperforms in terms of α-nDCG, ERR-IA, and other evaluation metrics. The detail results are shown as in Section 4.3.1 and Section 4.3.2. The compared description about HxQuAD is shown as follows:

In addition, comparing with the hierarchical diversification model in terms of the evaluation of α-nDCG, the improvement of L-SRD over the HxQuAD is up to 3.77%, 9.68%, 5.49% on WT2009, WT2010, WT2011 respectively. HxQuAD only use a predefined function to measure the diversity score, and the parameters may not be optimal because it needs to be tuned manually. Our learning model tackles the parameters tuning problem in an automatic fashion and reaches optimal result. As for robustness, the Win\Loss ratio of HxQuAD is only 2.51, but L-SRD is 2.65. Our different types of features and learning approach address optimal weights of parameters.

Reviewer #2:

Comment 1: Authors just need to review some sentences in lines 215, 245, and 320.

Answer 1: The sentences contain syntax problem. Based on the comment, we have revised the sentences as follows:

The sentence in line 215: The evaluation metrics are reported at different cutoffs. We use {5, 10, 20} as our cutoffs to set up experiments. These cutoffs focus on the evaluation at early ranking, which are particularly important in a Web search context Jansen et al. (1998). The α is set to 0.5 in our experiments for the reason it gives equal weight to both relevance and diversity.

The sentence in line 245: These baselines 2-4 (corresponding to MMR, xQuAD, PM2) possess a single parameter λ to tune, we perform a 5-fold cross validation to train λ through optimizing ERR-IA. In our model, we use a 5-cross validation with a ratio of 3:1:1 for training, validation and prediction for the test query on each year. The final results are calculated over all the folds.

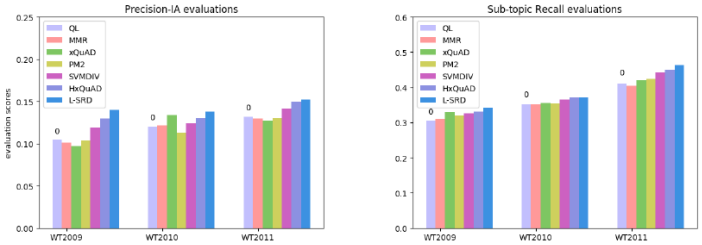
Comment 2: Experiments have to include comparison against (at least) one more recent method.

Answer 2: Based on the comment, we introduced a new approach called HxQuAD for comparison and revised the paper as follows:

In Section 4.3.1, line 280: In addition, comparing with the hierarchical diversification model in terms of the evaluation of α-nDCG, the improvement of L-SRD over the HxQuAD is up to 3.77%, 9.68%, 5.49% on WT2009, WT2010, WT2011 respectively. HxQuAD only use a predefined function to measure the diversity score, and the parameters may not be optimal because it needs to be tuned manually. Our learning model tackles the parameters tuning problem in an automatic fashion and reaches optimal result.

In Section 4.3.1, line 295: We consider not only the advanced diversity metrics, but also traditional diversity metrics, such as Precision-IA and Aspect Recall. The former indicates how many relevant documents for each aspect we have for ranking, the latter indicates how many of the aspects for which we have relevant documents. The result is shown in Fig. 3. MMR still underperforms all of them, as for Precision-IA, xQuAD

wins on WT2010 casually, while L-SRD performs more stable, even on WT2010, the gap is small. It proves that L-SRD outperforms others from different perspectives. Our learnable model solves the diverse ranking problem in a global perspective and always reaches prominent results.



(a) Precicison-IA evaluation (b) Aspect Recall evaluation

Figure 3: Performance comparison in WT2009, WT2010, WT2011 with Precision-IA and Aspect Recall

In Section 4.3.2, line 315: From Table 2, we find that L-SRD model performs best with its ratio of 2.65. While the Win/Loss ratio of MMR, xQuAD, PM2, SVMDIV and HxQuAD is 1.31, 1.46, 1.52, 1.77, 2.51 respectively. It reflects the remarkable robustness of L-SRD model comparing with other outstanding diversification models. The promotion of robustness over the MMR, xQuAD, PM2, SVMDIV and HxQuAD is up to 101.51%, 81.51%, 74.34%, 49.72%, 42.06% respectively. And it confirms the overall performance of our model is not restricted to a small subset, it still works in the whole data set for three years data. Our different types of features and learning approach address this problem well.

experiment	WT2009	WT2010	WT2011
MMR	20/18	24/17	23/16
xQuAD	23/18	23/16	24/14
PM2	22/20	26/14	25/14
SVMDIV	24/18	27/13	27/13
HxQuAD	27/15	30/10	31/10
L-SRD	**28/14**	**30/10**	**32/10**

Table 2: Win/Loss ratio

Comment 3: Authors have to include more recent references: 2014 (3), 2015 (0), 2016 (0), 2017 (0) and 2018 (0).

Answer 3: Based on the comment, we have introduced three new references 2015(1), 2016(1) and 2017(1). The content is shown as follows:

In page 4, line 80: Wang et al. and Hu et al. think the aspects underlying the query should be hierarchical, and propose some hierarchical measures to find the relationships among aspects (Wang et al. (2016);Hu & et al. (2015)).

In page 14, line 245: HxQuAD is a hierarchical diversification model based on xQuAD (Hu & et al. (2015)).

In page 4, line 65: Yun et al. formulate this as a process of selecting and ranking k exemplar documents and utilize linear programming to solve this problem (Yun et al. (2017)).

Reviewer #3:

Comment 1: In section, "Related work", the first paragraphs repeat concepts treated in the introduction.

Answer 1: Based on the comment, we have revised Related work and removed the repeat concepts as follows:

Search result diversification can be characterized as a directional optimization problem, in which one seeks to maximize the overall relevance of a document ranking to multiple query aspects, while minimizing its redundancy with respect (Santos et al. (2010b)). In particular, the existing approaches can be categorized as either implicit or explicit making a difference in how they account for the query aspects (Santos et al. (2010c)).

The basic assumption of implicit diversification approaches is that dissimilar documents are more likely to satisfy different information needs. The most representative approach in maximal marginal relevance (MMR) method and its probabilistic variants is shown as follows (Zhai et al. (2003)):

$$S_{MMR}(q,d,c) = (1-\lambda)S^{rel}(d,q) - \lambda \max_{d_j \in C} S^{div}(d,d_j) \qquad (1)$$

Where S^{rel} and S^{div} represents document $d's$ relevance to the query q and its similarity to a selected document d_j respectively. To gain high ranking score, a document should not only be relevant, but also be dissimilar from the selected documents. The special process of MMR proposed by Carbonell & Goldstein is selecting the document iteratively, and meanwhile, both content-based relevance and diversity relation between current selected document and the previously selected documents are considered (Carbonell & Goldstein (2009)). Yun et al. (2017) formulate this as a process of selecting and ranking k exemplar documents and utilize linear programming to solve this problem. In summary, they are all implicit approaches without using aspects to mine the underlying aspects, besides, they are a low effective approaches (Santos et al. (2010); Drosou & Pitoura (2010)).

Explicit approaches make use of the aspects underlying the query to select documents that cover different aspects as far as possible. The algorithms such as IA-select, xQuAD and RxQuAD are proposed to reduce redundancy on the aspect levels. These methods select documents that cover more novel aspects (Agrawal et al. (2009); Santos

et al. (2010); Vargas et al. (2012)). The PM-1 and PM-2 models pay more attention to maintain the proportionality of aspects (Dang & Croft (2012)). They produce the ranked result according to the proportionality of aspects. Intrinsic diversity products a series of successor queries to figure out the appropriate content to cover (Raman et al. (2013)). Wang et al. think the aspects underlying the query should be hierarchical, and propose some hierarchical measures to find the relationships among aspects (Wang et al. (2016)). To conclude, all existing explicit approaches are unsupervised, and the values of parameters need to be tuned by the experiment repeatedly without intention, causing a time-consuming optimizing problem to find the most suitable parameters.

Our approach also accounts for the aspects of a query in an explicit way. However, differently from the existing approaches, we use a learnable process to identify features from documents by Markov Random Field. It is other than using structural SVM to learn to identify a document subset with maximum word coverage (Hu et al. (2014)). They just learn the maximum word coverage and do not mine the aspects underlying the query. Besides, we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation to resolve this bidirectional optimization problem.

Comment 2: The section "Experiment" requires, for its understanding, a deep knowledge of the referenced works (in relation, for example, to the metrics used "to evaluate the performance"). We understand, however, that to evaluate, with the rigor that the authors claim, the results obtained, that the comparison, as it has been proposed by the authors, has been lost. Perhaps, it could be proposed, that the authors develop a summary of the main results of the analysis and comparisons made, more "accessible" for a reader without a thorough knowledge of the subject.

Answer 2: In order to understand the evaluation metrics well, we have revised the description of metrics, in which the usage and original for metrics are included. For better understanding the paper, we added a part to summarize our conclusion from the experimental results. The detail description is shown as follows:

For metrics in Page 12, line 205: There are three mainly evaluation metrics we use to evaluate the performance of our method: α-nDCG (Clarke et al. (2008)), ERR-IA (Chapelle et al. (2009)), NRBP (Clarke et al.

(2009b)). α-nDCG is used to balance both relevance and diversity of candidate documents. ERR-IA measures the expected effort required for a user to satisfy their information needs. And NRBP is a feasible metric to evaluate the balance between the complexity of needs and the query. These metrics penalize redundancy in a different degree for the document at each sorted position to maximize the aspects coverage. Additionally, we report our result using Precision-IA and Aspect Recall too (Agrawal et al. (2009)). To measure the robustness, we use Win/Loss ratio metrics (Yue & Joachims (2008); Dang & Croft (2012)). The Win/Loss ratio denotes whether the model improves or hurts the result when comparing with the basic relevant baseline QL in terms of evaluation metrics (Dang & Croft (2012)). Particularly in our experiment, we use ERR-IA to calculate the Win/Loss ratio.

For conclusion in Page 19, line 325: In this paper, we propose a Learnable Search Result Diversification model (L-SRD). We pay our attention to the explicit query aspect diversification models and introduce the learning approach to regard the model as a learning problem. Unlike the traditional explicit diversification models utilizing a predefined novelty measure function, we integrate different types of diversity features and estimate the weight with a learning approach. We derive our loss function as the likelihood of ground truth generation. Stochastic gradient descent algorithm is used to estimate the values of parameters. Benefiting from the learning approach, we can optimize the parameters in an automatic method. The prediction of our diversification model is provided by iterative maximizing the learned ranking function.

We have demonstrated the improvement of L-SRD comparing with other diversification models. We find L-SRD achieve considerable results in terms of official diversity metrics on three years in TREC web track data set. To prove its robustness, we set the experiment about Win/Loss ratio and usage of different retrieval algorithms. We believe L-SRD will play an important role to improve the Query Aspect Search Result Diversification by using a learning method.

Comment 3: Typos (in titles of sections)

Acknowledgements, reference -> References

Answer 3: Based on the comment, we have revised the References in line 335, and added Acknowledgements .