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: The work 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.] has indeed been discussed and authors offered a convincing argument. I think however, that a similar attitude could have been adopted for the works of [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.]

Answer 1: Based on the comment, we have carefully read the above paper, and introduce the main idea in our paper. The detail description is in Section 2, Page 5 as follows:

And some learning approaches are proposed. Zhu et al. (2014) use structural SVM to learn to identify a document subset with maximum word coverage, but they just learn the maximum word coverage and do not mine the aspects underlying the query. Xia et al. (2015) utilize both positive and negative ranking documents to train a maximal marginal relevance model for ranking, Xia et al. (2016) propose a neural tensor network to learn a nonlinear novelty function to select document, but they may pay much attention on novelty and less on relevance. Our approach 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, which are all positive differing from (Xia et al. (2015)), using Markov Random Field. 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: Regarding the novelty of MRF, I think that authors should be more sensitive to the general picture. Obviously the proposed approach suggests a novel integration of MRF, however, I think that authors would agree that this is not the first time that MRF have been used for feature extraction. This paper is meant to be read by the general audience of ESWA, so novelty should be interpreted and presented for this audience. It is a matter of positioning the contributions of the paper. In other words, if I have understood correctly, in section 3, the novelty of the paper exists in §3.5.

Answer 2: The novelty of this paper is that we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation. Firstly, we describe the each part in this paper including mining aspects, features extraction and loss function from Section 1 to Section 4 respectively, and in Section 5, we show all algorithm in this paper. So the novelty of the paper exists in Section 3.5.

Comment 3: Peer reviewers have suggested to add more updated references (and I totally support this recommendations), but there were added just one reference from 2015, one from 2016, and one from 2017.

Answer 3: Based on the comment, we have introduced five new references 2015(2), 2016(2), 2017(1). The content is shown as follows:

In Section 2, Page 3, Line 55: Search result diversification has a wide range of applications, such as patent search (Kim & Croft (2015)), legal information retrieval (Koniaris et al. (2017)) and so on.

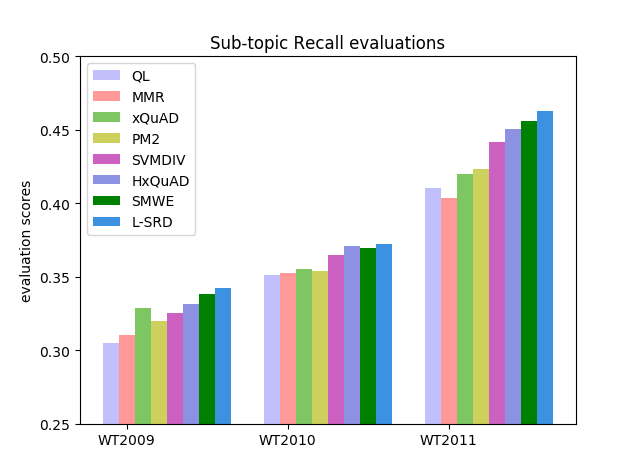
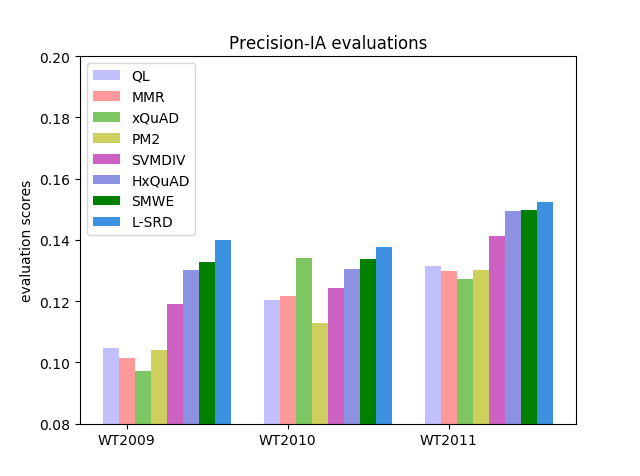
In Section 2, Page 5,Line 90: And some learning approaches are proposed. Zhu et al. (2014) use structural SVM to learn to identify a document subset with maximum word coverage, but they just learn the maximum word coverage and do not mine the aspects underlying the query. Xia et al. (2015) utilize both positive and negative ranking documents to train a maximal marginal relevance model for ranking, Xia et al. (2016) propose a neural tensor network to learn a nonlinear novelty function to select document, but they may pay much attention on novelty and less on relevance.

In Section 4.2, Page 15, Line 250: **SMWE** mines query subtopic by exploiting the word embedding and short-text similarity measure. (Ullah et al. (2016))

Comparing with the original version, we totally add 7 updated references, 2015(3), 2016(3), 2017(2).

Comment 4: In Figure 3, please try to modify (trim) the scale on the y-axis. I think that this way the differences in the performances will be better exposed.

Answer 4: Based on the comment, we have modified the pictures in Figure 3. The useful part can be enlarged by trimming y-axis. The modified figures are as follows:



Comment 5: Please check again the references. There seem to be some misses (e.g., Hu, S., & et al. (2015), Marden, J.I. (1996) )

Answer 5: Based on the comment, we recheck the paper, and the reference Marden, J.l.(1996) in Section 3.4,Page 10. But we do not introduce Hu,S.,& et al. (2015) even in the original version.

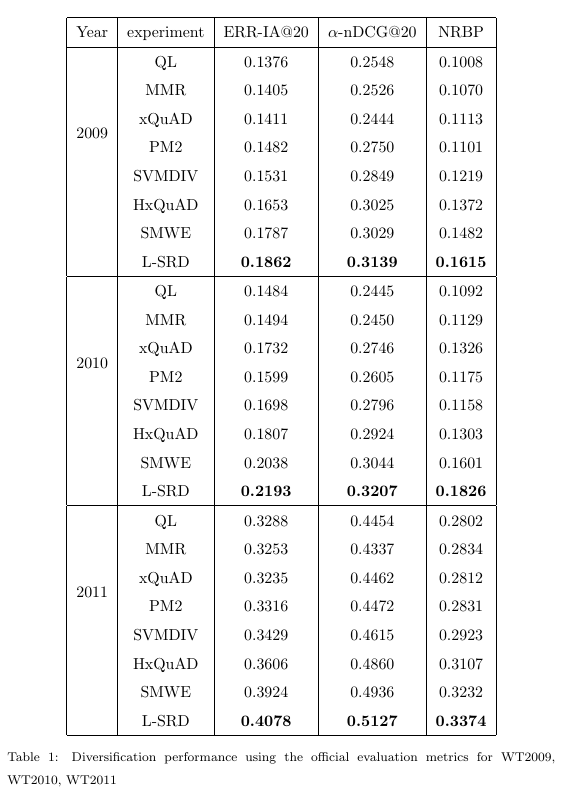
Reviewer #2: The comparison with methods published before 2015 may make the results obtained little credible. Why is no included the comparison against the new reference of 2017? or another one from 2016/2017.

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

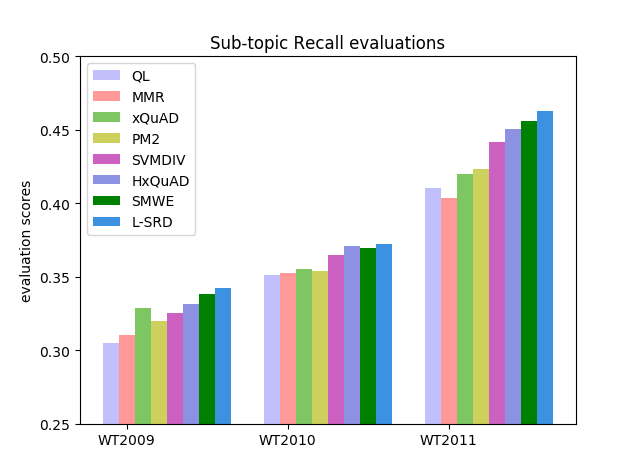
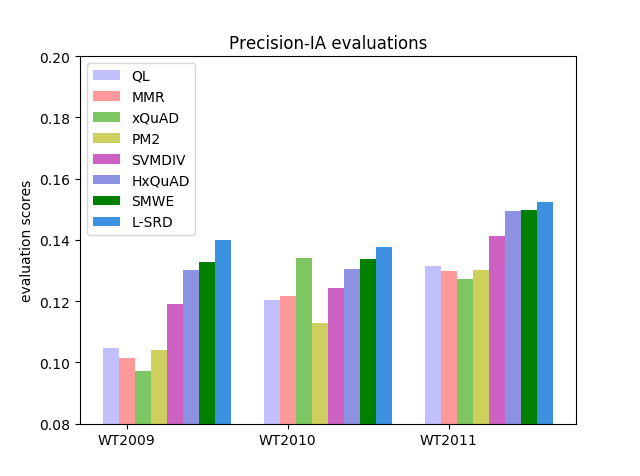
In Section 4.2, Page 15, Line 250: **SMWE** mines query subtopic by exploiting the word embedding and short-text similarity measure. (Ullah et al. (2016))

In Section 4.3 is the analysis of result as follows:

The result shows that L-SRD always performs best in terms of all metrics. It consistently improves the initial retrieval ranking method with gains up to 23.19%, 31.17%, 15.11% in terms of α-nDCG on WT2009, WT2010, WT2011 respectively. It indicates that our learning approach tackles the diversity measurement problem more effectively with the consideration of integrate different features. The reason is that features such as query-aspects relevance and information richness conform to the property of diversity. Furthermore, comparing with the explicit diversification models in terms of the evaluation of α-nDCG, the improvement of L-SRD over the xQuAD is up to 28.44%, 16.87%, 14.90% on WT2009, WT2010, WT2011 respectively, and the improvement of L-SRD over the PM2 is up to 14.15%, 23.11%, 14.65% on WT2009, WT2010, WT2011 respectively. Previous explicit diversifications use a predefined function to calculate the diversity score, which cannot reach an optimal result from the overall situation. A learnable approach to solve the diversity measurement and parameter tuning problem is significant. 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. Comparing with SWME in terms of the evaluation of α-nDCG, the improvement of L-SRD over the HxQuAD is up to 3.63%, 5.35%, 3.86% on WT2009, WT2010, WT2011 respectively. SMWE mines enough subtopics, but it cannot learn enough features to represent the document. Besides the non-learning model, the improvement of L-SRD over the SVMDIV is up to 10.18%, 14.70%, 11.09% on WT2009, WT2010, WT2011 respectively. It shows that considering relevance and different types of features in diversity measurement is helpful in the learning approach. That is the reason why our model wins. Therefore, L-SRD shows better understanding on the diverse ranking and leads to a better result. So utilizing learning mechanism indeed promotes the performance of search result diversification.



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 reranking, 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.



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, HxQuAD and SMWE is 1.31, 1.46, 1.52, 1.77, 2.51, 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, HxQuAD and SMWE is up to 102.28%, 81.51%, 74.34%, 49.72%, 5.58 %, 5.58% respectively. And it confirms the overall performance of our model is not restricted to a small subset, it still works in the whole dataset for three years data. Our different types of features and learning approach address this problem well.

