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 carefully read the above papers, and introduce the key idea in our paper. The detail description is in Section 2, Page 5, as follows:

Some learning approaches are also proposed for search result diversification. For example, 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. However, different from the existing approaches, we use a learnable process to identify features from documents 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: In this paper, we treat the query aspect diversification as a learning problem and propose a Learnable Search Result Diversification (L-SRD) method. We incorporate various features into diversity measurement based on the Markov Random Field (MRF), 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.

The novelty of the paper exists in Section 3.5. In Section 3.5, given the precise definition of loss function, the next step is minimizing the loss function to get the best performance. First, we generate the training data and apply the optimization method. Next, we use our ranking function to predict the final diverse ranking result.

Based on the comment, we revised the paper in Section 1, Page 3, as follows:

The main contributions of our work are listed as follows:

1. L-SRD introduces the learning mechanism to the query aspect diversification model. We conduct inference for the loss function based on its sequential selection model, which solves the parameters tuning problem automatically at the same time

2. We utilize the Markov Random Field to integrate different types of features to address the diversity measurement problem for query aspect search result diversification.

3. We propose a sequential prediction method, which selects the best document from candidate set by maximizing ranking score.

4. We conduct extensive experiments to verify L-SRD achieve better performance comparing with the existing diversification methods.

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 add 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: Some learning approaches are also proposed for search result diversification. For example, 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. However, different from the existing approaches, we use a learnable process to identify features from documents 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.

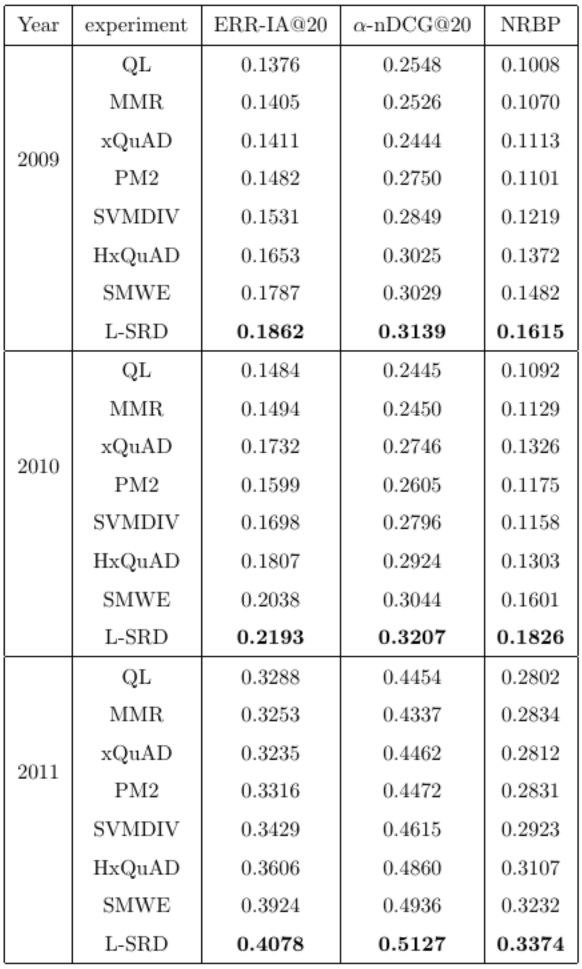
In Section 2, Page 5, Line 85: Ullah et al. (2016) mine query subtopic by exploiting the word embedding and short-text similarity measure.

Besides, we use the method as a baseline method (Ullah et al. (2016)). L-SRD performs best in terms of ERR-IA@20, α-nDCG@20 and NRBP. The detail analysis of result is as follows:

We add the baseline method in Section 4.2, Page 15 as follows: **SMWE** mines query subtopic by exploiting the word embedding and the short-text similarity measure. (Ullah et al. (2016)).

We show the diversification analysis of result in Section 4.3.1 as follows:

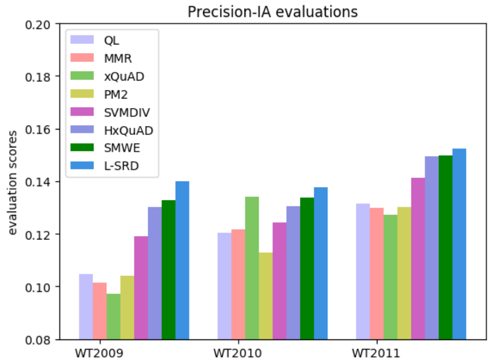
The result shows that L-SRD always performs best in terms of all metrics as shown in Table 1. It consistently improves the initial retrieval ranking method with gains up to 23.19%, 31.17%, 15.11% in terms of α-nDCG on WT2009, WT2010, WT2011 respectively. It indicates that our learning approach tackles 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 we find that utilizing learning mechanism indeed promotes the performance of search result diversification.

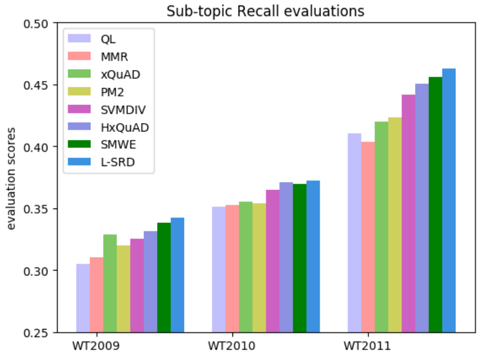


Year	experiment	ERR-IA@20	α-nDCG@20	NRBP
2009	QL	0.1376	0.2548	0.1008
	MMR	0.1405	0.2526	0.1070
	xQuAD	0.1411	0.2444	0.1113
	PM2	0.1482	0.2750	0.1101
	SVMDIV	0.1531	0.2849	0.1219
	HxQuAD	0.1653	0.3025	0.1372
	SMWE	0.1787	0.3029	0.1482
	L-SRD	**0.1862**	**0.3139**	**0.1615**
2010	QL	0.1484	0.2445	0.1092
	MMR	0.1494	0.2450	0.1129
	xQuAD	0.1732	0.2746	0.1326
	PM2	0.1599	0.2605	0.1175
	SVMDIV	0.1698	0.2796	0.1158
	HxQuAD	0.1807	0.2924	0.1303
	SMWE	0.2038	0.3044	0.1601
	L-SRD	**0.2193**	**0.3207**	**0.1826**
2011	QL	0.3288	0.4454	0.2802
	MMR	0.3253	0.4337	0.2834
	xQuAD	0.3235	0.4462	0.2812
	PM2	0.3316	0.4472	0.2831
	SVMDIV	0.3429	0.4615	0.2923
	HxQuAD	0.3606	0.4860	0.3107
	SMWE	0.3924	0.4936	0.3232
	L-SRD	**0.4078**	**0.5127**	**0.3374**

Table 1: Diversification performance using the official evaluation metrics for WT2009, WT2010, WT2011

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.





We show the robustness analysis of result in Section 4.3.2 as follows:

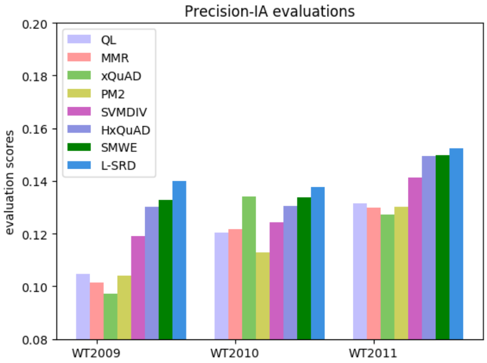
From table 2, we find that L-SRD model performs best with its ratio of 2.65. While the Win/Loss ratio of MMR, xQuAD, PM2, SVMDIV, HxQuAD and SMWE is 1.31, 1.46, 1.52, 1.77, 2.51, 2.44 respectively. It reflects the remarkable robustness of L-SRD model comparing with other outstanding diversification models. 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 %, 8.61 % 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.

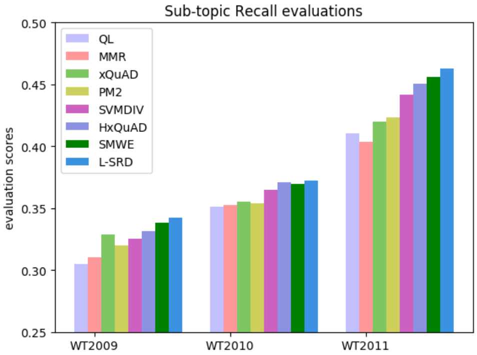
experiment	WT2009	WT2010	WT2011
MMR	20/18	24/17	23/16
xQuAD	23/18	23/16	24/14
PM2	22/20	26/14	25/14
SVMDIV	24/18	27/13	27/13
HxQuAD	27/15	30/10	31/10
SMWE	27/14	29/10	32/11
L-SRD	**28/14**	**30/10**	**32/10**

Table 2: Win/Loss ratio

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 is 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 find these references still in our paper.

In Section 2, Page 4: Wang et al. (2013) and Hu & et al. (2015) think the aspects underlying the query should be hierarchical, and propose some hierarchical measures to find the relationships among aspects. Ullah et al. (2016) mine query subtopic by exploiting the word embedding and short-text similarity measure.

In Section 3.4, Page 10: On the basis of the Plackett-Luce Model (Marden (2016)), we derive the steps in our generation process shown as follows:

$$P(Y \mid D) = \prod_{i=1}^{n} P(y_i \mid D \setminus S_{i-1}) = \prod_{i=1}^{n} \frac{\exp(f(y_i, D \setminus S_{i-1}))}{\sum_{k=i}^{n} \exp(f(y_k, D \setminus S_{k-1}))} \tag{10}$$

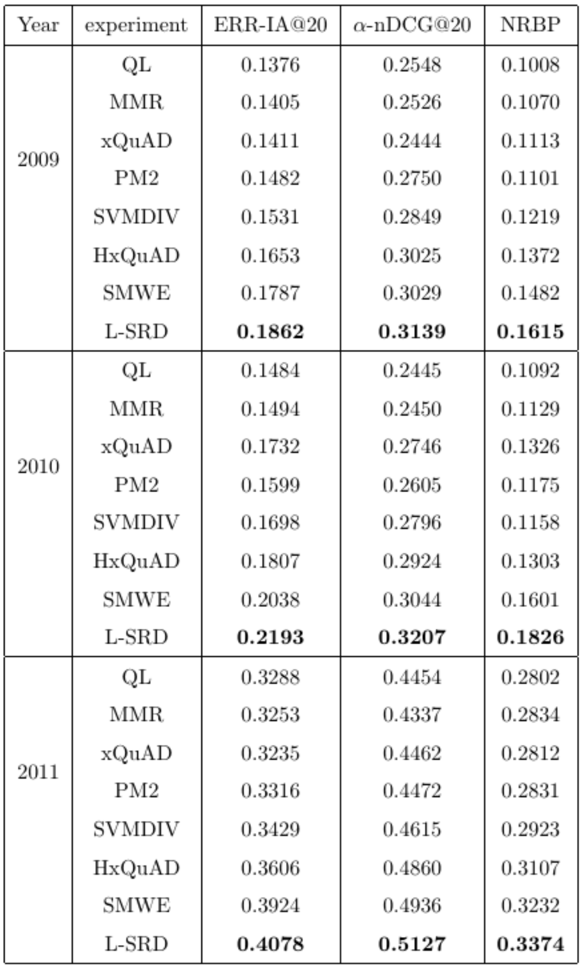
Reviewer #2: The comparison with methods published before 2015 may make the results obtained little credible. Why is no included the comparison against the new reference of 2017? or another one from 2016/2017.

Answer: Based on the comment, we introduce a new method called **SMWE** (Ullah et al. (2016)) for comparison and revise the paper as follows:

We add the baseline method in Section 4.2, Page 15 as follows: **SMWE** mines query subtopic by exploiting the word embedding and the short-text similarity measure. (Ullah et al. (2016))

We show the diversification analysis of result in Section 4.3.1 as follows:

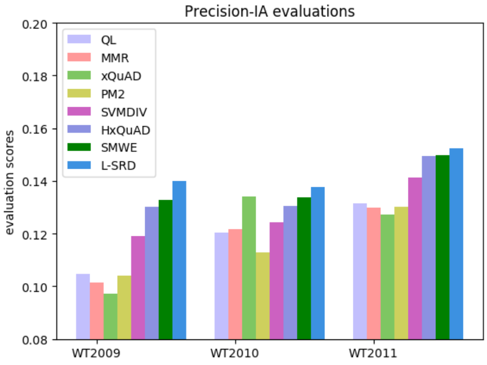
The result shows that L-SRD always performs best in terms of all metrics as shown in Table 1. It consistently improves the initial retrieval ranking method with gains up to 23.19%, 31.17%, 15.11% in terms of α-nDCG on WT2009, WT2010, WT2011 respectively. It indicates that our learning approach tackles 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 we find that utilizing learning mechanism indeed promotes the performance of search result diversification.

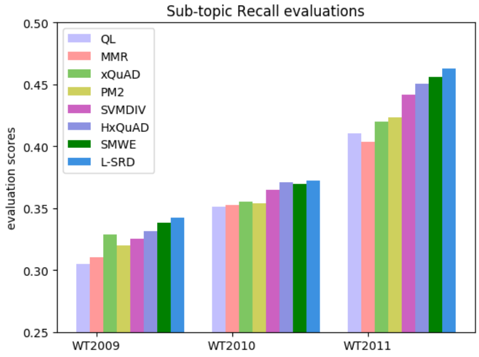


Year	experiment	ERR-IA@20	α-nDCG@20	NRBP
2009	QL	0.1376	0.2548	0.1008
	MMR	0.1405	0.2526	0.1070
	xQuAD	0.1411	0.2444	0.1113
	PM2	0.1482	0.2750	0.1101
	SVMDIV	0.1531	0.2849	0.1219
	HxQuAD	0.1653	0.3025	0.1372
	SMWE	0.1787	0.3029	0.1482
	L-SRD	**0.1862**	**0.3139**	**0.1615**
2010	QL	0.1484	0.2445	0.1092
	MMR	0.1494	0.2450	0.1129
	xQuAD	0.1732	0.2746	0.1326
	PM2	0.1599	0.2605	0.1175
	SVMDIV	0.1698	0.2796	0.1158
	HxQuAD	0.1807	0.2924	0.1303
	SMWE	0.2038	0.3044	0.1601
	L-SRD	**0.2193**	**0.3207**	**0.1826**
2011	QL	0.3288	0.4454	0.2802
	MMR	0.3253	0.4337	0.2834
	xQuAD	0.3235	0.4462	0.2812
	PM2	0.3316	0.4472	0.2831
	SVMDIV	0.3429	0.4615	0.2923
	HxQuAD	0.3606	0.4860	0.3107
	SMWE	0.3924	0.4936	0.3232
	L-SRD	**0.4078**	**0.5127**	**0.3374**

Table 1: Diversification performance using the official evaluation metrics for WT2009, WT2010, WT2011

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.





We show the robustness analysis of result in Section 4.3.2 as follows:

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, 2.44 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 %, 8.61 % 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.

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
SMWE	27/14	29/10	32/11
L-SRD	**28/14**	**30/10**	**32/10**

Table 2: Win/Loss ratio