Journal of Engineering Sciences
Vol 11, Issue 4 ,April/2020
ISSN NO:0377-9254

Supervised diversification of the search results through subtopic attention

S. Aleem Basha[1], M. Kanchana[2], M.S. Pallavi[3], P. Mahendra[4], V.N. Anand Reddy[5], P. Jaswanth[6]

[1]Assistant Professor, Dept of CSE, AITS, Rajampet, AP, India

[2,3,4,5,6] Student, Dept of CSE, AITS, Rajampet, AP, India

ABSTRACT

Diversification of the search results is aimed at producing varied results to meet as many different needs for knowledge as possible. Recently, supervised methods have been proposed for learning ranking functions and they have been shown to yield superior results to unsupervised methods. These techniques however use tacit approaches based on the Maximum Marginal Relevance (MMR) concept. We suggest in this paper a learning system for explicit diversification of outcomes, where subtopics are specifically modeled. Based on the information found in the sequence of selected documents, we use the process of attention to capture the subtopics to be concentrated on when selecting the next document, which naturally matches our diversification task of document selection. We use recurrent neural networks and max pooling as a preliminary attempt to instance the system. We use both distributed representations and conventional pertinence features to model implementing documents. The architecture is versatile, in either a flat list or a hierarchy, to model query intent. Experimental findings indicate that the strategy proposed greatly outperforms all of the approaches to diversification of current search findings.

Keywords— search result, diversification, subtopics, attention.

1. INTRODUCTION

In real search situation, questioning is typically vague or multifaced. In addition to being applicable to the application, the documents obtained are required to be as diverse as possible to cover different information needs.When users issue "apple," for example, the underlying purpose may be the IT company or the fruit. The documents collected will cover both topics to improve the chances of satisfying users with specific information needs. Current approaches to diversification of the search results are typically unattended and take on manually defined functions with empirically tuned parameters. They can be divided into implied and explicit approaches depending on whether the underlying intents (or subtopics) are specifically modeled. Implicit approaches do not specifically model purpose. We emphasize novelty, i.e. the current document will be "different" from the previous ones based on certain measurements of similarity. Instead, clear approaches directly attempt (or subtopics) at modeling. They seek to enhance coverage of purpose, i.e. the following document will cover attempts that were not satisfied with previous ones. Intents or subtopics can be calculated using techniques such as query reformulation and database clustering based on database logs and other information forms. Existing studies showed that explicit approaches have better performance than implicit approaches due to several reasons. On the one hand, they have a more straightforward way of dealing with subtopics than with implicit approaches; on the other, their rating functions are similar to the diversity evaluation criteria, often based on explicit subtopics. Moreover, most similarity measures used in the implicit methods, e.g. those based on the language model or vector space model, are calculated globally over the entire text, regardless of potential search attempts. This may be problematic for diversification of the search results: two documents that contain similar terms and be considered globally identical, but this identical aspect may be unrelated to underlying search attempts. A new family of research work using supervised learning is being introduced to avoid heuristic and handcrafted functions and parameters. They try to automatically learn a ranking feature. Their key emphasis is on diversity modeling, including structural prediction, rewarding functions for new content, direct measure-based optimization, and neural network-based method. All of these approaches inherit the spirit of MMR, which is an implicit approach and do not take intents into consideration, irrespective of the diversity modeling and optimization methods. Although the learning methods that result in a better measure of similarity,

the gap between reducing redundancy of documents and improving coverage of purpose is hindering them. With tacit, unsupervised methods they suffer from similar problems. Without explicit modeling of subtopics, they can not explicitly boost coverage of purpose. Therefore, explicit subtopic modeling needs to be integrated into supervised methods of diversification. We propose to use recurrent neural networks (RNN) and max-pooling to exploit both distributed representations and specific features, which we call DSSA-RNMP. We expanded this model further to add hierarchical subtopics. The basic concept is that subtopics function inherently as a hierarchical system, where high-level subtopics reflect general user intents while low-level subtopics are more specific. Only the consideration of coarse or fine-grained subtopics can lead to under-optimal coverage of purpose. Particularly for subtopics, attention is measured on different rates. The final score is the weighted sum of the various-level scores. This is what we call the hierarchical model HDSSA-RNMP. Experimental findings on TREC Web Track data show that efficiency is further enhanced by DSSA-RNMP outperforms the existing methods significantly and HDSSARNMP.

To our knowledge, this is the first time a supervised learning system with attention mechanism has been used to specifically model subtopics for diversification of the search results.

TABLE-1
Subtopic relevance example.

doc*n*subtopic	$^{i}1$	$^{i}2$	$^{i}3$
^{1}	*p*	*p*	
d_2	*p*	*p*	
d_3			*p*
d_4		*p*	
d_5			*p*

2. RELATED WORK

2.1 Implicit Diversification Approaches:

The basic principle of implicit approaches to diversification is that dissimilar documents are more likely to fulfil the diversification approaches dif TABLE 2 categorisation. Unattended implicit supervised MMR SVM-DIV, R-LTR, PAMM, NTN explicit IA-Select, xQuAD, PM2, TxQuAD, TPM2, HxQuAD, HPM2, 0-1 MSKP DSSA (our approach) ferent knowledge requirements. A document should not only be important to achieving high ranking.score but should also be dissimilar from the selected documents. The definition of measures for relevance and document similarity is crucial, which is done manually in this approach. Based on the optimal facility placement theory, the applicant documents are first grouped, then the diverse result collection is assembled, which strikes a good balance between productivity and effectiveness. Machine learning approaches to learn score functions have recently been leveraged.

TABLE -2

Categorization of diversification approaches.

	unsupervised	supervised
implicit	MMR	SVM-DIV, R-LTR, PAMM, NTN
explicit	IA-Select, xQuAD, PM2, TxQuAD, TPM2, HxQuAD, HPM2, 0-1 MSKP	**DSSA** (our approach)

The most representative approach is MMR [3]:

$$SMMR(q,d,C) = (1-\lambda)S^{rel}(d,q) - \lambda \max_{dj \in C} S^{div}(d,dj),$$

2.2 Explicit Diversification Approaches:

Explicit approaches model subtopics that underlie a query, with the goal of returning as many subtopics as possible of documents. These approaches leverage external resources to specifically reflect the needs in subtopics for knowledge. IASelect makes use of categorized topical taxonomy groups based on ODP. xQuAD is a probabilistic paradigm that makes use of query reformulations as representations of purpose.

2.3 RNN with Attention Mechanism:

The interdependence between elements in a sequence can be captured by RNN. Attention mechanism, typically based on RNN, mimics at different times the action of human attention that

focuses on specific local region of the object (an image, a sentence, etc.) In computer vision RNN carefully used to extract information from an image by choosing a series of the most informative regions instead of the entire image. Mechanism of focus is usually used in neural machine translation (NMT) at NLP. Traditional encoder-decoder models encode the source sentence in a fixed-length vector from which it decodes the target sentence. The fixed-length vector may not be sufficiently powerful to represent all the source sentence information.

3 DOCUMENT SEQUENCE WITH SUBTOPIC ATTENTION FRAMEWORK

Given a query set Q, a document set Dq and a subtopic set Iq for each query q ∈ Q, the goal of explicit methods is to learn a ranking function f(q,Dq,Iq) which is expected to output a ranking of documents in Dq that is both relevant and diverse. where L measures the quality gap between the ranking outputted by f and the best ranking Yq. Different from traditional retrieval tasks, diversity has to be considered in the ranking and evaluation process.

The loss function could be written in the following general form:

$$\sum_{q \in Q} L(f(q,Dq,Iq),Yq), \qquad (2)$$

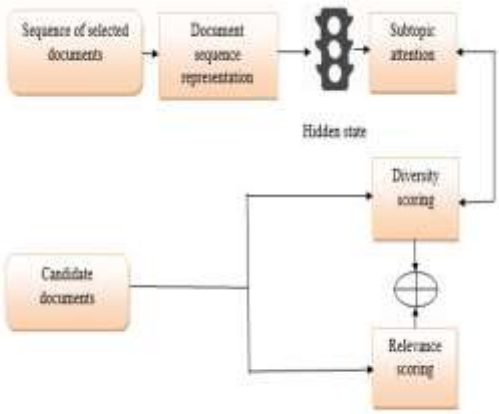

Fig-1. Illustration of DSSA framework.

4 RESULT DIVERSIFICATION USING DSSA:

In this section, we instantiate DSSA to a concrete form and articulate the training and prediction algorithms. The main idea of DSSA is to dynamically capture accumulative relevance information of previous document sequence, so as to calculate subtopic attention.

4.1 A Neural Network Implementation:

We first define the constitution of representations, namely vdt, vq, and vik, then expand on how we enforce the representation of document series, subtopic attention, and part scoring.

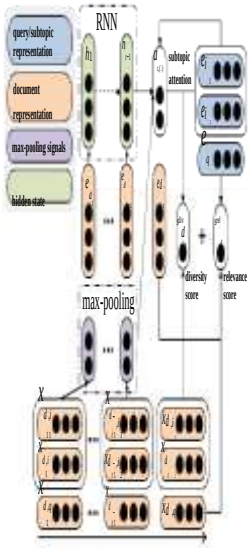

Fig-2. Architecture of DSSA-RNNMP

4.1.1 Document Sequence Representation:

H is instantiated using RNN to encode preceding document sequence information. It is possible to use different types of RNN cells, ranging from the plain vanilla cell, GRU cell, to the LSTM cell. We'll just display the vanilla cell here for simplicity. The cell converts past hidden layer ht−1 and current document distributed representation edt into another space where a bias bn is introduced and then a non-linear activation (i.e. tanh) occurs, creating the next hidden layer ht. H0 is initialized as a Zero vector.

4.1.2 Subtopic Attention:

By looking at ht−1 which stores the information of previous t−1 documents and ei(·) which represents the meaning of By looking at ht−1 which stores the knowledge found in previous t−1 documents and ei (·) which reflects the context of each subtopic, we are able to discover which attempts are not fulfilled and must therefore be emphasized at the tth position. To capture this concept, we use A′(ht−1,eik) at the t-th position to calculate the(unnormalized) importance of the k-thsub topic, which could be implemented in many ways.

4.1.3 Scoring:

The final score consists of relevance score srel dt and diversity score sdiv dt , which are combined by a coefficient λ. Where w^r and $w^r \in RR$ are the attention derived from the component of subtopic attention. The diversity score is calculated by attention distribution as a weighted combination of the pertinence of the document to each subtopic. We use the same approach to measure the relevance of a document to a query and its subtopics using both distributional representations and pertinence features, although different methods can be used.

4.2 Hierarchical Diversification:

We incorporate hierarchical subtopics into DSSA, motivated by the concept of organizing subtopics in a hierarchical system suggested by Hu et al. The theory behind hierarchical subtopics is that there are various granularities of consumer intents. It would be biased if we only find subtopics which are coarse or fine-. To better understand the shortcoming of using subtopics without understanding various granularities, we use the query "defender" (Query # 20) as an example, which shows the hierarchical subtopics extracted from the query suggestions of the search engine in Figure .

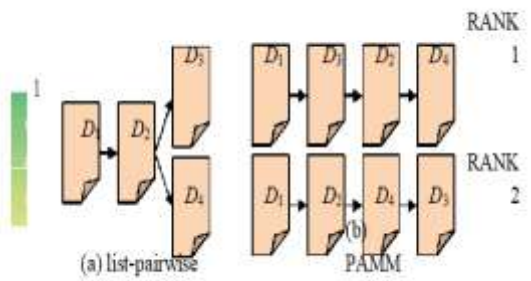

Fig. 3. Pair sample examples of (a) list-pairwise and (b) PAMM. Both samples are positive.

5 EXPERIMENTAL SETTINGS

5.1 Data Collections

We use the same dataset as which consists of Web Track dataset from TREC 2009 to 2012. There are 198 queries (query #95 and #100 are dropped because no diversity judgments are made for them),each of which includes 3to8 subtopics identified by TRECassessors. The relevance rating is given in a binary form at subtopic level. All experiments are conducted on ClueWeb09 collection.

5.2 Evaluation Metrics

We use ERR-IA , α-nDCG , and NRBP , which are official diversity evaluation metrics used in Web Track. They measure the diversity by explicitly rewarding novelty and penalizing redundancy. D#-measures, the primary metric used in NTCIR Intent and IMine task is also included. We also use traditional diversity measures Precision-IA (denoted as Pre-IA) and Subtopic Recall (denoted as S-rec) . Consistent with existing works and TREC Web Track, all these metrics are computed on top 20 results of a ranking. We use two-tailed paired t-test to conduct significance testing with p-value < 0.05.

5.3 Baseline Models We compare DSSA and HDSSA2 to various unsupervised and supervised diversification methods.

6 EXPERIMENTAL RESULTS

6.1 Overall Results

The overall results are shown in Table 7. We find that DSSA significantly outperforms all implicit and explicit baselines, including both unsupervised and supervised.

DSSA vs. unsupervised explicit methods. DSSA outperforms unsupervised explicit methods (xQuAD, PM2, TxQuAD, TPM2, HxQuAD, and HPM2) on all the measures.

DSSA vs. supervised implicit methods. DSSA also out performs supervised implicit methods (R-LTR, PAMM, R-LTR-NTN, and PAMM-NTN) by quite large margins. The improvement over R-LTR-NTN and PAMM-NTN, the best supervised implicit approaches is up to 9.9% and 9.4% respective lyonα-nDCG.This result demonstrates the utility of taking into account subtopics explicitly in supervised approaches.

HDSSA vs. DSSA. HDSSA outperforms DSSA on all the measures. Through paying attention to subtopics of different granularities, HDSSA has the potential to detect the most unsatisfied intents and keep balance between general and fine-grained intents.

TABLE -3

Performance comparison of all methods. The best result is in bold. Statistically significant differences between DSSA and baselines are marked with various symbols. H indicates significant improvement over all baselines ($p < 0.05$).

Methods	ERR-IA nDCG	-	NRBP	D?-nDCG	Pre-IA	S-rec
Lemur[A]	.271	.369	.232	.424	.153	.621
ListMLE[¶]	.287	.387	.249	.430	.157	.619
C-GLS	.288	.391	.246	.435	.153	.640
xQuAD	.317	.413	.284	.437	.161	.622
TxQuAD	.308	.410	.272	.441	.155	.634
HxQuAD	.326	.421	.294	.441	.158	.629
PM2	.306	.411	.267	.450	.169	.643
TPM2	.291	.399	.250	.443	.161	.639
HPM2	.317	.420	.279	.455	.172	.645
R-LTR	.303	.403	.267	.441	.164	.631
PAMM	.309	.411	.271	.450	.168	.643
R-LTR-NTN	.312	.415	.275	.451	.166	.644
PAMM-NTN	.311	.417	.272	.457	.170	.648
DSSA	.356[H]	.456[H]	.326[H]	.473[H]	.185[H]	.649[A]
HDSSA-B	.366[H]	.465[H]	.335[H]	.475[H]	.186[H]	.648[A]
HDSSA	**.369**[H]	**.467**[H]	**.337**[H]	**.478**[H]	**.187**[H]	**.653**[A]

6. CONCLUSION

In this paper, we propose a general learning framework DSSA to model subtopics explicitly for search result diversification. Based on the sequence of selected documents, unequal and varied subtopic attention is calculated, driving the model to emphasize different subtopics at different positions.This is the first time that attention mechanism isused to model the process. We further instantiate DSSA using RNN and max-pooling to handle both distributed representations and relevance features, which outperforms significantly the existing approaches. The results confirm that modeling subtopics explicitly in a learning framework is beneficial and effective and this also avoids heuristically defined functions and parameters. Through using hierarchical subtopics, performance is further improved because of the consideration of subtopics of different granularities. However, accurately modeling the interaction among documents and subtopics is still challenging. There are many other more complex implementations besides our particular way, which will be investigated in future work. The proposed model contains a number of parameters to be learned. This requires a large number of training data.

7. REFERENCES

[1]. 1.R. L. Santos, J. Peng, C. Macdonald, and I. Ounis, "Explicit search result diversification through sub-queries," in ECIR. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 87–99.

[2]. R. L. Santos, C. Macdonald, I. Ounis et al., "Search result diversification," Foundations and Trends® in Information Retrieval, vol. 9, no. 1, pp. 1–90, 2015.

[3]. J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in SIGIR. New York, NY, USA: ACM, 1998, pp. 335–336.

[4]. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in WSDM. New York, NY, USA: ACM, 2009, pp. 5–14.

[5]. R. L. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," in WWW. New York, NY, USA: ACM, 2010, pp. 881–890.

[6]. V. Dang and W. B. Croft, "Diversity by proportionality: An election-based approach to search result diversification," in SIGIR. New York, NY, USA: ACM, 2012, pp. 65–74.

[7]. V.DangandB.W.Croft,"Termlevelsearchresultdiversification," in SIGIR. New York, NY, USA: ACM, 2013, pp. 603–612.

[8]. S. Hu, Z. Dou, X. Wang, T. Sakai, and J.-R. Wen, "Search result diversification based on hierarchical intents," in CIKM, ser. CIKM '15. New York, NY, USA: ACM, 2015, pp. 63–72.

[9]. H.-T. Yu and F. Ren, "Search result diversification via filling up multiple knapsacks," in CIKM. New York, NY, USA: ACM, 2014, pp. 609–618.

[10]. J. Yi and F. Maghoul, "Query clustering using click-through graph," in WWW. New York, NY, USA: ACM, 2009, pp. 1055– 1056.

[11]. R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines," in

Current Trends in Database Technology EDBT 2004 Workshops, 2004, pp. 588–596.

[12]. Z. Zhang and O. Nasraoui, "Mining search engine query logs for query recommendation," in WWW. New York, NY, USA: ACM, 2006, pp. 1039–1040.

[13]. A. Herdagdelen, M. Ciaramita, D. Mahler, M. Holmqvist, K. Hall, S. Riezler, and E. Alfonseca, "Generalized syntactic and semantic models of query reformulation," in SIGIR. New York, NY, USA: ACM, 2010, pp. 283–290.

[14]. Y. Yue and T. Joachims, "Predicting diverse subsets using structural svms," in ICML. New York, NY, USA: ACM, 2008, pp. 1224–1231.

[15]. Y. Zhu, Y. Lan, J. Guo, X. Cheng, and S. Niu, "Learning for search result diversification," in SIGIR. New York, NY, USA: ACM, 2014, pp. 293–302.

[16]. L. Xia, J. Xu, Y. Lan, J. Guo, and X. Cheng, "Learning maximal marginal relevance model via directly optimizing diversity evaluation measures," in SIGIR. New York, NY, USA: ACM, 2015, pp. 113–122.

[17]. "Modeling document novelty with neural tensor network for search result diversification," in SIGIR. New York, NY, USA: ACM, 2016, pp. 395–404.

[18]. V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in NIPS, 2014, pp. 2204–2212.