EFFECTIVE & EFFICIENT METHODS FOR WEB SEARCH RESULT DIVERSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET MURAT ÖZDEMIRAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SEPTEMBER 2015

Approval of the thesis:

EFFECTIVE & EFFICIENT METHODS FOR WEB SEARCH RESULT DIVERSIFICATION

submitted by **AHMET MURAT ÖZDEMIRAY** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** ______________

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** ______________

Assist. Prof. Dr. İsmail Sengör Altıngövde
Supervisor, **Computer Engineering Department, METU** ______________

Examining Committee Members:

Prof. Dr. Özgür Ulusoy
Computer Engineering Department, Bilkent University ______________

Assist. Prof. Dr. İsmail Sengör Altıngövde
Computer Engineering Department, METU ______________

Prof. Dr. Ahmet Coşar
Computer Engineering Department, METU ______________

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU ______________

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU ______________

Date: ______________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: AHMET MURAT ÖZDEMIRAY

Signature :

ABSTRACT

EFFECTIVE & EFFICIENT METHODS FOR WEB SEARCH RESULT DIVERSIFICATION

Özdemiray, Ahmet Murat
Ph.D., Department of Computer Engineering
Supervisor : Assist. Prof. Dr. İsmail Sengör Altıngövde

September 2015, 124 pages

Search result diversification is one of the key techniques to cope with the ambiguous and/or underspecified information needs of the web users. In this study we first extensively evaluate the performance of a state-of-the-art explicit diversification strategy and pin-point its weaknesses. We propose basic yet novel optimizations to remedy these weaknesses and boost the performance of this algorithm. Secondly, we cast the diversification problem to the problem of ranking aggregation and propose to materialize the re-rankings of the candidate documents for each query aspect and then merge these rankings by adapting the score(-based) and rank(-based) aggregation methods. As a third contribution, for the first time in the literature, we propose using post-retrieval query performance predictors (QPPs) to estimate, for each aspect, the retrieval effectiveness on the candidate document set, and leverage these estimations to set the aspect weights. In addition to utilizing well-known QPPs from the literature, we also introduce three new QPPs that are based on score distributions and hence, can be employed for online query processing in real-life search engines. For the last contribution, we use retrieval performance predictions of query aspects to selectively expand those aspects that perform below some threshold, using the top retrieved documents of the aspect's own results.

Our extensive experimental evaluations show that, despite having lower computational complexity than the state-of-the-art diversification strategies, certain ranking aggregation methods are superior to the existing explicit diversification strategies in terms of the diversification effectiveness. Furthermore, using QPPs for aspect weighting improves almost all state-of-the-art diversification algorithms in comparison to using a uniform weight estimator and also the proposed QPPs are comparable or superior to the existing predictors in the context of aspect weighting. Lastly, using QPP methods to selectively expand the query aspects provide better diversification performance compared to unexpanded or fully expanded aspects, for most of the diversification strategies.

Keywords: Web Search Systems, Search Result Diversification, Ranking Aggregation, Query Performance Prediction, Query Expansion

ÖZ

WEB ARAMA CEVAPLARININ ÇEŞİTLENDİRİLMESİNDE ETKİN VE VERİMLİ YÖNTEMLER

Özdemiray, Ahmet Murat
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. İsmail Sengör Altıngövde

Eylül 2015 , 124 sayfa

Arama sonuçlarının çeşitlendirilmesi, web kullanıcılarının muğlak veya eksik belirtilmiş bilgi ihtiyaçlarıyla baş edilmesi için kullanılan anahtar tekniklerden biridir. Son yıllarda, sorgu cephelerinin açıkça bilinmesine dayanan stratejiler, sorgu sonuçlarının çeşitlendirilmesinde çok etkili yöntemler olarak kullanılmaya başlamıştır. Bu çalışmada, öncelikle açıkça bilinen sorgu cephelerine dayanan modern çeşitlendirme stratejilerinden birini detaylı bir şekilde değerlendirerek onun zayıf noktalarını tespit ediyoruz. Bu zayıflıklara çözüm getirmek ve algoritmanın performansını artırmak için basit ama daha önce uygulanmamış optimizasyonlar öneriyoruz. İkinci katkı olarak, mevcut çeşitlendirme stratejilerinin aday dokümanların sorgu cephelerine yakınlığından faydalanmasından ilham alarak, çeşitlendirme problemini sıralama birleştirme problemine benzeştiriyoruz. Bu amaçla, aday dokümanların her bir sorgu cephesi için oluşturulmuş sıralamasını kullanmayı ve bu sıralamaları skor tabanlı ve sıra tabanlı birleştirme yöntemlerini adapte ederek birleştirmeyi öneriyoruz. Üçüncü olarak, literatürde ilk defa sorgu sonrası performans tahmincileri (QPP) kullanarak, her sorgu cephesi için aday doküman kümesinin performansını kestirip, bu bilgiyi kullanarak sorgu cephelerinin ağırlıklarını belirliyoruz. Literatürde iyi bilinen QPP'lerin kullanımının yanında, gerçek arama motorları tarafından çevrimiçi sorgu işleme sırasında kullanılabilecek skor dağılımına dayalı üç yeni QPP daha tanımlıyoruz. Son katkı olarak da, performans tahminleri belirli eşiğin altında olan sorgu cephelerini, sorgu cephesinin kendi sonuçlarını kullanarak genişletiyoruz.

Yoğun deneysel değerlendirmelerimiz gösteriyor ki, bakığında belirli sıralama birleştirme yöntemleri, açıkça bilinen sorgu cephelerine dayanan modern çeşitlendirme stratejilerinden çeşitlendirme etkinliği açısından daha iyi performans sağlıyor. Ayrıca, bu sıralama birleştirme yöntemleri, mevcut çeşitlendirme yöntemlerinden daha az işlem güçlüğü gerektiriyor. Ayrıca, QPP'lerin sorgu cephelerinin ağırlığını bulmak için kullanılması neredeyse tüm modern çeşitlendirme stratejilerinde eşit ağırlıklandırmaya nazaran daha iyi sonuç veriyor. Bunun yanında, önerilen QPP'ler de aspekt ağırlıklandırma açısından mevcut QPP'lerle kıyaslandığında benzer ya da daha iyi sonuç veriyorlar. Son olarak, genişletilecek sorgu cephelerinin QPP yöntemleri ile belirlenmesi ile elde edilen sonuçlar genişletilmemiş veya tamamı genişletilmiş sorgu cephelerine göre daha iyi çeşitlendirme performansı sunuyor.

Anahtar Kelimeler: Web Arama Sistemleri, Arama Sonuçlarını Çeşitlendirme, Sıralama Birleştirme, Sorgu Performans Tahmini, Sorgu Genişletme

To my dearest wife

and my kids Yağmur and Furkan

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Professor İsmail Sengör Altıngövde for his suggestions, efforts, support and guidance. He made this work possible by opening a new path for me, when all directions were blocked and he always encouraged me to complete this work.

I acknowledge with thanks and appreciation to the thesis monitoring committee members, Prof. Dr. Özgür Ulusoy and Prof. Dr. Ahmet Coşar for their support and constructive comments on my thesis work. They always gave valuable feedback for the progress of this work, and were not hesitant to warn me of the shortcomings or risks of my work.

I also would like to thank to the thesis defense jury members, Prof. Dr. İsmail Hakkı Toroslu and Assoc. Prof. Dr. Pınar Karagöz for reviewing and evaluating my thesis.

I would like to express my appreciation for TÜBİTAK BİLGEM İLTAREN for their support during my academic studies. I also would like to thank my colleagues for creating a peaceful yet competitive working environment and especially I would like to thank to ILTAREN director Dr. Nedim Alpdemir for his support and guidance during my professional and academic studies.

I thank to my parents, for bringing up me as I am. They taught me to give my best in every circumstance and I always tried harder to make them more proud.

I also thank to aunt Fatma for taking care of us and our children during me and my wife's extensive studies.

Lastly, I thank to my wife for being supportive and thoughtful during my studies. She had taken care of our children in the absence of me and carried the burden of our home.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xvii

LIST OF FIGURES xxv

LIST OF ABBREVIATIONS xxvi

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Contributions 2

2 OPTIMIZATIONS ON EXPLICIT DIVERSIFICATION METHODS 5

2.1 Introduction 6

2.2 Related Work 7

2.2.1 Implicit Search Result Diversification 7

2.2.2 Explicit Search Result Diversification 8

2.2.3 Score Normalization 9

2.3 xQuAD Framework: Potential Weaknesses and Extensions . 10

2.3.1 xQuAD Framework 11

2.3.1.1 Potential Weaknesses of xQuAD . . . 11

2.3.1.2 Relevance Score Normalization for xQuAD 14

2.3.1.3 Document Novelty Estimation for xQuAD 15

2.4 Experimental Setup 16

2.4.1 Collection, Queries and Aspects 16

2.4.2 Initial Retrieval Model 16

2.4.3 Baseline Diversification Strategies and Evaluation Metrics 17

2.4.3.1 Intent Aware (IA)-select 17

2.4.3.2 org_xQuAD 17

2.4.3.3 PM2 18

2.4.3.4 Evaluation metrics 19

2.5 Evaluation Results 20

2.5.1 Performance of the Score Normalization Techniques 20

2.5.2 Performance of xQuAD variants 21

2.5.3 Summary of the Main Findings 23

2.6 Conclusion 24

3 RANKING AGGREGATION METHODS FOR DIVERSIFICATION 25

3.1 Introduction 25

3.2 Related Work 26

3.3 Ranking Aggregation Methods for Diversification 27

3.3.1 Score-based Aggregation Methods 28

3.3.1.1 CombSUM (mix_CombSUM). 29

3.3.1.2 CombMNZ (mix_CombMNZ) 30

3.3.2 Rank-based Aggregation Methods 30

3.3.2.1 Simple voting (mix_SV). 31

3.3.2.2 Borda voting (mix_BV). 31

3.3.2.3 Markov chain based methods. 31

3.4 Experiments and Results 33

3.4.1 Evaluation Results 33

3.4.2 Impact of the Components and Parameters 37

3.4.2.1 Impact of the aspect representation. . 37

3.4.2.2 Impact of the initial retrieval model. . 38

3.4.2.3 Other score normalization techniques . 39

3.4.2.4 Impact of the probability mixture model in ranking aggregation. 40

3.5 Conclusion 41

4 QUERY PERFORMANCE PREDICTION FOR ASPECT WEIGHTING IN DIVERSIFICATION 43

4.1 Introduction 43

4.2 QPPs for Aspect Weighting 45

4.2.1 Baseline QPPs for Aspect Weighting 46
4.2.2 Proposed QPPs for Aspect Weighting 47
4.3 Experimental Evaluation 48
4.3.1 Explicit diversification methods 49
4.3.2 Experimental Results 49
4.4 Conclusion 52
5 ASPECT EXPANSION FOR EXPLICIT SEARCH RESULT DIVERSIFICATION 53
5.1 Introduction 54
5.2 Related Work 55
5.3 Selectively Expanding Aspect Queries 56
5.3.1 Aspect Expansion 57
5.3.1.1 Term Scoring Functions 57
5.3.2 Selecting Aspects to Expand 58
5.4 Experiments and Results 58
5.4.1 Experimental Setup 58
5.4.1.1 Explicit diversification methods 58
5.4.1.2 Sub-topic Query Expansion 59
5.4.1.3 Selective Sub-topic Query Expansion . 59
5.4.2 Evaluation Results 60
5.4.2.1 Expansion of official sub-topics using candidate re-rankings 60

5.4.2.2 Expansion of official sub-topics using own ranking . . . 63

5.4.2.3 Expansion of suggested topics using own ranking . . . 65

5.4.2.4 Selective expansion of official sub-topics using own rankings . . . 67

5.5 Conclusions and Future Work . . . 69

6 CONCLUSION . . . 71

REFERENCES . . . 73

APPENDICES

A ADDITIONAL EXPERIMENTS . . . 81

A.1 BM25 retrieval model and query aspects obtained from the suggestions . . . 81

A.2 LM retrieval model and query aspects obtained from official sub-topics . . . 84

A.3 LM retrieval model and query aspects obtained from the suggestions . . . 87

A.4 Aspect Weighting using Query Performance Predictions . . . 90

A.4.1 2009 official sub-topics . . . 90

A.4.2 2009 sub-topics from suggestions . . . 92

A.4.3 2010 suggested subtopics . . . 93

A.5 Query expansion tables . . . 94

A.5.1 2009 official sub-topics . . . 94

A.5.2 2010 official sub-topics . . . 101

A.5.3 2009 sub-topics from suggestions 108

A.5.4 2010 sub-topics from suggestions 115

CURRICULUM VITAE 123

LIST OF TABLES

TABLES

Table 2.1 Diversification performance w.r.t. the relevance normalization techniques for different retrieval models using the query aspects obtained from the official sub-topics. The highest scores are shown in boldface. 21

Table 2.2 Diversification performance of the xQuAD variants using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface. 22

Table 3.1 Diversification performance of the score aggregation methods using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface. 34

Table 3.2 Diversification performance of the rank aggregation methods using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface. 35

Table 3.3 Comparison of the best cases for the baseline and proposed methods using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface. 36

Table 3.4 Comparison of the best cases for the baseline and proposed methods using the query aspects obtained from the suggestions. The highest scores across all methods are shown in boldface. 37

Table 3.5 Comparison of the best cases for the baseline and proposed methods using the LM for the initial retrieval and query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface. 39

Table 3.6 Comparison of the best cases for the baseline and proposed methods using the LM for the initial retrieval and query aspects obtained from the suggestions. The highest scores across all methods are shown in boldface. 40

Table 4.1 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the baseline QPPs for the query aspects obtained from the official sub-topics. The highest score is boldfaced. 50

Table 4.2 Diversification performance (α-NDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the best baseline QPPs and proposed QPPs for the query aspects obtained from the official sub-topics. The highest score in each group is bold, the overall winner is underlined. 51

Table 5.1 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced. 61

Table 5.2 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced. 62

Table 5.3 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced. 63

Table 5.4 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced. 64

Table 5.5 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankingsas PRF. The highest score is boldfaced. 65

Table 5.6 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankingsas PRF. The highest score is boldfaced. 66

Table 5.7 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 67

Table 5.8 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 68

Table A.1 Diversification performance w.r.t. the relevance normalization techniques using the query aspects obtained from the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface. 81

Table A.2 Diversification performance of the xQuAD variants using the query aspects obtained from the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface. 82

Table A.3 Diversification performance of the score aggregation methods using the query aspects obtained from the the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface. 82

Table A.4 Diversification performance of the rank aggregation methods using the query aspects obtained from the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface. 83

Table A.5 Diversification performance w.r.t. the relevance normalization techniques using the query aspects obtained from the official sub-topics and LM as the retrieval model. The highest scores are shown in boldface. . . . 84

Table A.6 Diversification performance of the xQuAD variants using the query aspects obtained from the official sub-topics and LM as the retrieval model. The highest scores are shown in boldface. 85

Table A.7 Diversification performance of the score aggregation methods using the query aspects obtained from the official sub-topics and LM as the retrieval model. The highest scores are shown in boldface. 85

Table A.8 Diversification performance of the rank aggregation methods using the query aspects obtained from the official sub-topics and LM as the retrieval model.. The highest scores are shown in boldface. 86

Table A.9 Diversification performance w.r.t. the relevance normalization techniques using the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface. 87

Table A.10 Diversification performance of the xQuAD variants using the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface. 88

Table A.11 Diversification performance of the score aggregation methods the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface. 88

Table A.12 Diversification performance of the rank aggregation methods the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface. 89

Table A.13 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the baseline QPPs for the query aspects obtained from the official sub-topics. The highest score is boldfaced. 90

Table A.14 Diversification performance (α-NDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the best baseline QPPs and proposed QPPs for the query aspects obtained from the official sub-topics. The highest score in each group is bold, the overall winner is underlined. 91

Table A.15 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the proposed QPPs for the query aspects obtained from the suggestions. The highest score is boldfaced. 92

Table A.16 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the baseline QPPs. The highest score is boldfaced. 93

Table A.17 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced. 94

Table A.18 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced. 95

Table A.19 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 96

Table A.20 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 97

Table A.21 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 98

Table A.22 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the official sub-topics and their selectiveexpansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 99

Table A.23 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the official sub-topics and their selectiveexpansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 100

Table A.24 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced. 101

Table A.25 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced. 102

Table A.26 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 103

Table A.27 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 104

Table A.28 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 105

Table A.29 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 106

Table A.30 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 107

Table A.31 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankings as PRF. The highest score is boldfaced. 108

Table A.32 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankings as PRF. The highest score is boldfaced. 109

Table A.33 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankingsas PRF. The highest score is boldfaced. 110

Table A.34 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 111

Table A.35 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 112

Table A.36 Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their selectiveexpansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 113

Table A.37 Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their selectiveexpansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 114

Table A.38 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankingsas PRF. The highest score is boldfaced. 115

Table A.39 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankingsas PRF. The highest score is boldfaced. 116

Table A.40 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankingsas PRF. The highest score is boldfaced. 117

Table A.41 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 118

Table A.42 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methodsfor the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 119

Table A.43 Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their selectiveexpansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 120

Table A.44 Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their selectiveexpansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced. 121

LIST OF FIGURES

FIGURES

Figure 2.1 Percentage of the eliminated aspects after choosing the documents for each rank in τ_q^*. 13

Figure 2.2 Diversification performance of the xQuAD variants vs. trade-off parameter λ. 23

LIST OF ABBREVIATIONS

xQuAD	e**X**plicit **Qu**ery **A**spect **D**iversification
IA-Select	**I**ntent-**A**ware **Select**
α-nDCG	α weighted **n**ormalized **D**iscounted **C**umulative **G**ain
CombSUM	**Comb**ination by **SUM**mation of similarities
CombMNZ	**Comb**ination by summations **M**ultiplied by number of **N**on-**Z**ero similarities
BV	**B**orda **V**oting
SV	**S**imple **V**oting
MC	**M**arkov **C**hain
QPP	**Q**uery **P**erformance **P**rediction
TREC	**T**ext **RE**trieval **C**onference
WIG	**W**eighted **I**nformation **G**ain
NQC	**N**ormalized **Q**uery **C**ommitment
ScrAvg	**Sc**o**r**e **Av**era**g**e
ScrDev	**Sc**o**r**e **Dev**iations
VScrFirst	**V**irtual **Sc**o**r**e **First**
VScrAvg	**V**irtual **Sc**o**r**e **Av**era**g**e
ScrRatio	**Sc**o**r**e **Ratio**
KLD	**K**ullback **L**eibler **D**istance
PRF	**P**seudo-**R**elevance **F**eedback

CHAPTER 1

INTRODUCTION

1.1 Motivation

With the proliferation of the digital age, Internet became the main source of information. The text content hosted in the Web covers a broad range, including but not limited to academic, educational, entertaining, informational, navigational and social material. Search engines are the main tools to access those broad range of content on the Web. In year 2014, one of the popular search engines, Google, responded more that 2 trillion web searches [1].

More than half of the web searches consist of at most two terms [2], most of which are ambiguous or underspecified, making it a challenge for search systems to determine the intention of the user. For example, when a single term query, say 'Jordan', is issued to the search engine; the user's search intention may be to get information about the country Jordan, or the basketball legend Michael Jordan, which makes this query ambiguous. Furthermore, the user may also want to know some demographic or welfare information of people living in country Jordan, or the contact information of the Jordan embassy in his country (or Jordan brand shoes, or the career stats of "Michael Jordan") which also makes this query underspecified. For such queries, the search engine can provide a result set that can cover possible different interpretations of the query to satisfy the user.

Search result diversification methods try to improve user satisfaction in case of am-

[1] http://www.statisticbrain.com/google-searches
[2] http://www.keyworddiscovery.com/keyword-stats.html

biguous or underspecified queries, either by implicitly discovering the query aspects using the contents of the candidate documents, or by explicitly by using the previously obtained query aspects through some mechanisms.

In this thesis, we presume that the query aspects are explicitly known during query execution and we propose effective and efficient strategies to diversify web search results, while improving state-of the-art explicit search result diversification methods also.

1.2 Contributions

The contributions in this thesis can be divided into four parts. Firstly, in Chapter 2, we extensively evaluate one of the better performing state-of-the-art explicit search result diversification methods (i.e. xQuAD [51]) and pin-point some of its weaknesses. Being a probabilistic framework, xQuAD uses a greedy algorithm to construct the query result, by choosing a candidate document at each iteration which maximize the relevance to the original query and novelty among other selected documents. While examining this algorithm, we noticed that for some queries, if a document which fully represents a query aspect is selected to the result set, that query aspect is neglected and the algorithm only diversifies the rest of the result set using other query aspects. We called this problem "aspect elimination problem". Secondly, we realized that, after selecting a few documents to the final result set, the novelty component's weight becomes negligible compared to the relevance component in the xQuAD mixture model, making the algorithm choose rest of the result set based on the relevance to the original query.

To overcome these weaknesses of xQuAD algorithm, we first apply some score normalization methods in the literature to estimate the probability of aspect's satisfaction by choosing the document. We also propose a novel normalization algorithm which depends on a virtual document to approximate the upper-bound of the document's relevance score. In order to mitigate the second issue, we propose to utilize some aggregate functions to model the novelty component of xQuAD.

In Chapter 3, we present our second contribution, which is motivated by the observation that, computing the relevance of candidate documents to query aspects play a central role in current explicit search result diversification strategies. Inspired by this finding, we exploit the re-rankings of the candidate documents according to the query aspects and merged these re-rankings using score-based and rank-based ranking aggregation algorithms. The work reported in Chapter 2 and Chapter 3 was published in:

- A. M. Ozdemiray and I. S. Altingovde. Score and rank aggregation methods for explicit search result diversfication. Technical Report METU-CENG-2013-01, Middle East University, Computer Engineering Department, September 2013.
- A. M. Ozdemiray and I. S. Altingovde. Explicit search result diversification using score and rank aggregation methods. *Journal of the Association for Information Science and Technology*, 66(6):1212-1228, 2015.

During our extensive evaluations, we observed that the weighting of the query aspects play an important role in the success of the diversification mechanism. In Chapter 4, for the first time in the literature, we propose using post-retrieval query performance predictors to estimate the retrieval effectiveness of each query aspect on the candidate document set to find the relative weights of the query aspects. In addition to utilizing the well-known post retrieval QPPs from the literature, we introduce three new QPPs that are based on the score distributions of the candidate documents in the re-rankings. The work presented in Chapter 4 was published in:

- A. M. Ozdemiray and I. S. Altingovde. Query Performance Prediction for Aspect Weighting in Search Result Diversification. In: *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM 2014, pages 1871-1874. ACM, 2014.

Inspired by the success of query expansion and re-writing techniques applied in ad hoc retrieval, we propose to expand the query aspects in Chapter 5. In particular, we use pseudo-relevance feedback (PRF) methods on the top-k results retrieved for each query aspect to find expansion terms to better represent the aspect. Moreover, we introduce a novel selective strategy, based on the findings of the previous chapter, to expand those aspects that are likely to benefit from the expansion. Specifically, we use the proposed QPP methods to predict the performance of the aspects and expand the aspect queries if necessary.

We conclude and point some future work directions in Chapter 6.

CHAPTER 2

OPTIMIZATIONS ON EXPLICIT DIVERSIFICATION METHODS

Search result diversification methods try to satisfy user information needs in case of ambiguous or underspeficied user queries. Some of these strategies assume that query aspects are gathered through some mechanism and try to diversify the initial query using these aspects. In this chapter [1], we pin-point some of the weaknesses of one of the best performing state-of-the-art explicit diversification methods and propose some optimizations to remedy these weaknesses. We also applied one of these optimizations to some state-of-the-art explicit diversification methods to observe its behavior.

In Sections 2.1 we provide an introduction to the problem at hand and in Section 2.2 an overview of the related studies in the literature are given. We identify two potential weaknesses of a state-of-the-art explicit diversification framework, xQuAD, and introduce our solutions in Section 2.3. In the next two sections, we describe our experimental setup and present the evaluation results, respectively. The last section provides the conclusion.

[1] A. M. Ozdemiray and I. S. Altingovde. "Explicit search result diversification using score and rank aggregation methods", *Journal of the Association for Information Science and Technology*, 66(6):1212-1228. ©2015 John Wiley and Sons. http://dx.doi.org/10.1002/asi.23259. Reprinted by permission with license number 371383091410

2.1 Introduction

Search result diversification is a popular problem that receives attention from both academia and industry. At the heart of the problem lies the fact that a large fraction of web queries are vaguely specified and/or ambiguous, making it very hard (if not impossible) for a search system to figure out the underlying search intent of the users. For such queries, it seems to be a good compromise to provide a result set that can cover possible different interpretations of the query and, thus, try to minimize the risks of disappointing the users (e.g., [70]).

A number of result diversification strategies in the literature assume that potential query aspects can be explicitly identified (say, by categorizing the queries according to a taxonomy [1] or mining query logs [51, 9]), and aim to diversify the initial retrieval results (candidate documents) of a query based on these already known aspects. In this study, we also assume the availability of explicit query aspects and propose new strategies for result diversification in this setup.

In this chapter, we extensively evaluate the performance of a state-of-the-art explicit diversification strategy, namely, xQuAD ([51]), and pin-point some of its weaknesses. xQuAD is a probabilistic framework that constructs the final query result in a greedy manner, by choosing the candidate document d that maximizes the *relevance* (based on the likelihood of observing d for the query) and *diversity* (based on the relevance of d to each query aspect, and the *novelty* of d with respect to the documents that are already selected into the result) at each iteration. We identify two issues, so-called "aspect elimination problem" and "aspect fading problem", that may arise due to the ways the relevance and novelty probabilities are computed and/or estimated in this framework. In essence, both of these problems are related to having some query aspects that end up with a negligible or no impact during the early stages of the diversification process; i.e., after selecting a few documents into the final result set.

To remedy the former problem, we explore a variety of relevance score normalization methods and also propose a normalization strategy based on the upper-bound score estimated for a given query and retrieval model. To address the second problem, we propose to employ alternative functions while computing the novelty component of xQuAD.

We evaluate the performance of the xQuAD variants, ranking aggregation methods and QPPs in the context of aspect weighting using the standard TREC datasets and explicit aspects discovered from different sources, and report the results for a number of well-known metrics. We compare the proposed diversification methods to three state-of-the-art explicit diversification strategies, namely IA-Select ([1]), xQuAD (as originally proposed in ([51]), and PM2 strategy in ([20]). Our experiments show that the xQuAD variants with the new score normalization and novelty components outperform the original algorithm as well as the other baselines.

2.2 Related Work

Generating diverse/novel results is a hot topic with the potential of application in various contexts, ranging from web search engines (e.g., [50])) to recommenders (e.g., [59]) and topic tracking systems (e.g., [2]). In this study, we focus on the search result diversification problem that aims to provide both relevant and diverse results for the ambiguous or underspecified web queries. In the literature, the approaches that address this problem are broadly categorized as either *implicit* or *explicit* ([51]).

2.2.1 Implicit Search Result Diversification

The strategies in this category assume no prior knowledge of the query aspects; so they either exploit the inter-similarity of the documents in the candidate set or attempt to discover the underlying query aspects in an unsupervised manner ([50]). A pioneering example of the former approach is the Maximum Marginal Relevance (MMR) strategy that constructs the final ranking in a greedy manner ([10]). In each iteration, a document's score is computed by the difference of its relevance to the original query and similarity to the documents that are selected into the final ranking up to this point; and the document with the highest score is selected. Various strategies in the literature adapt this greedy algorithm, yet differ in the way they compute the inter-document similarities. For instance, Zhai et al. ([67]) utilize unigram language models for representing the individual documents as well as the set of documents that are already selected into the final ranking at any point during the greedy iterations. In contrast, Zuccon and Azzopardi ([74]) make use of the quantum probability ranking principle while modeling the interference among the ranked documents. Two independent works in the literature propose to adapt the modern portfolio theory to the result diversification problem ([44], [63]). In this case, the inter-document similarities are modeled based on the variance of the relevance among the ranked documents.

Gollapudi and Sharma ([25]) identify the connection between the result diversification problem and facility dispersion optimization problems, and adapt some approximate solutions (namely, Max-Sum and Max-Min algorithms) from the operations research field to the diversification context. Minack et al. employ these algorithms and improve their efficiency for diversifying continuous data streams ([36]). A comparative analysis of various implicit diversification algorithms using five different datasets (other than standard TREC collections) is provided by Vieira et al. ([62]). More recently, Zuccon et al. introduce an alternative perspective and model the diversification problem within the desirable facility placement (DES) framework ([75]).

Different from the above approaches, some other implicit diversification strategies (so-called coverage based methods in ([50])) attempt to model the underlying query aspect from the initial retrieval results. For instance, Carterette and Chandar ([14]) identify the aspects (facets) using relevance modeling and topic models, and then constructs the final ranking in a round-robin fashion, i.e., by choosing the best document for each facet. He et al. ([26]) also use topic models to partition the candidate documents into clusters; but they only consider the most relevant clusters to the query for the subsequent diversification stages where well-known strategies such as the MMR and round-robin are applied.

2.2.2 Explicit Search Result Diversification

In the explicit diversification methods, query aspects are modeled explicitly, i.e., by exploiting the query labels, which are assigned either manually or automatically, or from the reformulations of the query. IA-Select approach adopts the former option and assumes that both queries and documents are associated with some categories from a taxonomy ([1]). The diversification is achieved by favoring documents from different categories and penalizing the documents that fall into already covered categories. Alternatively, Radlinski and Dumais ([43]) use a given query and its reformulations to obtain a candidate result set; which is then re-ranked and personalized for a given user. Capannini et al. ([9]) employ query logs to decide when/how query results should be diversified, and propose a new algorithm based on the popularity of query reformulations in the log.

xQuAD is one of the most effective diversification strategies that also exploit query reformulations obtained from TREC subtopics and search engines to model the query aspects [51]. In a follow-up work, Santos et al. [52] employ both xQuAD and IA-Select to achieve result diversification for the queries with navigational, informational, or transactional intents. Vallet and Castells [58] incorporate a personalization component into both of the latter algorithms by explicitly introducing the user as a random variable. In another study, Vargas et al. again employ these two algorithms, xQuAD and IA-Select, and propose to model their relevance models explicitly, i.e., using the relevance judgments or, more practically, click statistics [60]. Finally, Zheng et al. propose a coverage based diversification framework where they experiment with several coverage functions [72]. While these latter works also improve or build on xQuAD, none of them focus on its components in a way similar to ours. Different from the previous studies, we propose optimizations for the relevance score normalization and novelty estimation components of xQuAD.

2.2.3 Score Normalization

The problem of score normalization is often tackled in the context of score-based ranking aggregation. In one of the earliest works, Lee [31] employs `MinMax` normalization (see Equation 2.3) to combine the retrieval scores of different systems. Montague and Aslam ([37]) identify the desirable properties of the score normalization techniques for meta-search and propose two new techniques, namely Sum and ZMUV (zero-mean, unit-variance). A more detailed comparison of the latter techniques is provided by Sever and Tolun ([53]). Fernandez et al. propose a probabilistic normalization strategy for score-based aggregation ([24]). Arampatzis and Kamps ([4]) propose a normalization approach based on the assumption that the retrieval scores are composed of a signal and a noise component. In a rather different context, Ravana and Moffat ([46]) investigate the score aggregation techniques for summarizing the performance of a retrieval system over a set of queries. To the best of our knowledge, none of the previous studies explore the impact of score normalization on the explicit result diversification.

2.3 xQuAD Framework: Potential Weaknesses and Extensions

Preliminaries

Assume a query q is processed over a collection C and retrieves a ranked list of documents τ_q, where $|\tau_q| = N$.

Result Diversification Problem: Construct a ranked list τ_q^* of k documents $(k < N)$ such that τ_q^* maximizes both the relevance and diversity among all possible rankings $\tau_i(|\tau_i| = k)$ of τ_q.

A particular case of this general problem is the explicit result diversification problem, where there is a set of explicitly identified query aspects (a.k.a., sub-topics, interpretations, sub-queries) denoted as $T = \{q_1, ..., q_m\}$ associated with the original query q. Then, the objective function is finding a top-k ranking τ_q^* that maximizes the overall relevance to multiple query aspects and at the same time, minimizes its redundancy with respect to these aspects ([25]).

It can be shown that the general form of this problem is an instance of the maximum coverage problem and thus, it is NP-hard (e.g., see [51]). A large number of diversification strategies based on the approximation algorithms, heuristics and/or meta-heuristics are proposed in the literature (as briefly reviewed in the previous section). In what follows, we describe one of the most effective strategies, xQuAD, that is investigated and extended in more depth in the following sections.

2.3.1 xQuAD Framework

xQuAD is a probabilistic framework ([51]) that constructs the ranking τ_q^* in a greedy manner, by choosing the document $d_i \in \tau_q$ that maximizes the following probability mixture model at each iteration:

$$(1-\lambda)P(d|q) + \lambda \sum_{q_i \in T} P(q_i|q)P(d|q_i)P(\bar{\tau}_q^*|q_i), \tag{2.1}$$

where $P(d|q)$ denotes the relevance (i.e., likelihood of observing d for the query q) whereas the summation captures the diversity. In particular, $P(q_i|q)$ denotes the likelihood of the aspect (sub-query) q_i for the query q (referred to as sub-query importance in [51]), $P(d|q_i)$ is the likelihood of observing d for the aspect q_i and finally $P(\bar{\tau}_q^*|q_i)$ denotes the probability of q_i not being satisfied by the documents that are already in τ_q^*. The latter probability, which indeed captures the novelty, can be represented as the product of the probabilities of each document in τ_q^* for not satisfying q_i:

$$P(\bar{\tau_q^*}|q_i) = \prod_{d_j \in \tau_q^*} (1 - P(d_j|q_i)). \tag{2.2}$$

2.3.1.1 Potential Weaknesses of xQuAD

xQuAD is one of the most successful strategies for the explicit result diversification and placed among the top-performers in the diversity tasks of both TREC 2009 and 2010 ([16], [17]). However, we still identify two problems that can significantly diminish the performance of xQuAD, as follows.

Aspect elimination problem In the above model, a key component is the relevance computation of a document d to the query q and its aspects (sub-queries) q_i, denoted as $P(d|q)$ and $P(d|q_i)$, respectively. In previous works, these probabilities are usually based on the popular weighting models like BM25, language models, etc. (e.g., [51]). Typically, the scores produced by these methods are normalized to $[0, 1]$ range at the query-level, so that they can be employed in the xQuAD's mixture model. While no details are provided on the exact procedure employed in previous works, a practical and tempting approach is using the `MinMax` score normalization, where the score range of a query is mapped to the range $[0, 1]$; i.e., the top-ranked document in a list having the score 1. `MinMax` normalization can be formally expressed as ([31], [47]):

$$P(d|q) = \frac{s(d,q) - \min_{d_i \in \tau_q} s(d_i,q)}{\max_{d_i \in \tau_q} s(d_i,q) - \min_{d_i \in \tau_q} s(d_i,q)}, \tag{2.3}$$

where τ_q is the ranked retrieval result for q, $s(d,q)$ is the score generated by the retrieval model and $P(d|q)$ is the normalized relevance probability.

However, we realize that `MinMax` and other normalization techniques that set the $P(d|q)$ (or, $P(d|q_i)$) value to 1 for the highest scoring documents for q (or, q_i) cause a deficiency in the model. Once the top-scoring document d^* for an aspect q_i is selected for τ_q^*, for all following iterations, the impact of covering this aspect will be nullified. That is, as $P(d^*|q_i) = 1$ using, say, `MinMax` normalization, the probability $\prod_{d_j \in \tau_q^*}(1 - P(d_j|q_i))$ will be 0 once d^* is selected for τ_q^*. Therefore, the algorithm will not care covering aspect q_i from this point on. Even worse, for a query with just a few aspects, if the documents with the highest scores for each aspect are selected at the early stages of the algorithm, then diversification part of the xQuAD will be totally neglected, and all remaining documents will be selected solely based on $P(d|q)$.

The problem is more pronounced for the queries with a few aspects and when the diversified set size is relatively large; i.e., $k \geq 20$. In Figure 2.1, we show the number of *eliminated* aspects after choosing the documents for each rank position i $(1 \leq i \leq 20)$ using xQuAD on TREC 2009 diversity task setup for the λ that yields the highest α-nDCG@20 score (see the section Experimental Setup for the details). The figure shows that even after selecting the first two documents into τ_q^*, 23% of the query aspects (i.e., 56 out of 241 aspects specified for the 50 topics in TREC 2009) are neglected, which is clearly not helpful for the diversification purposes.

Finally, the aspect the aspect elimination problem can be further harmful for the informational queries, for which the users usually need more than one document (per aspect) to satisfy their information needs. Within this latter context, Welch et al. ([64]) report the existence of the aspect elimination problem for another diversification strategy, namely, IA-Select ([1]). Note that, since the IA-Select strategy in its original setup employs the scores obtained from a classifier, the problem in their case is not directly related to the normalization techniques. Nevertheless, in this chapter, we include IA-Select among our baseline strategies (replacing the classifier scores with $P(d|q_i)$ scores as in [51]), and evaluate the impact of the relevance score nor-

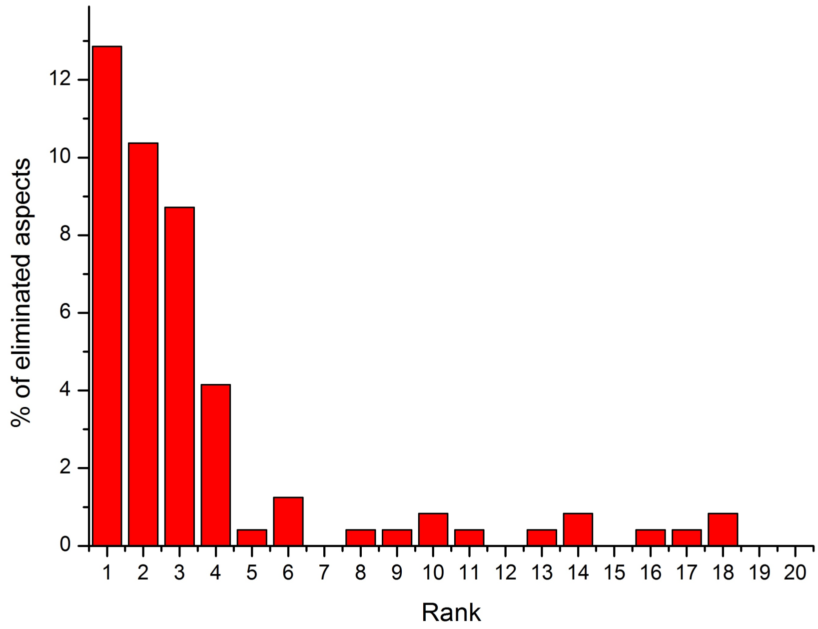

Figure 2.1: Percentage of the eliminated aspects after choosing the documents for each rank in τ_q^*.

malization techniques (described in the next section) also for IA-Select.

Aspect fading problem: Even when the top-scoring document of an aspect is not selected for τ_q^*, the impact of the aspect q_i fades away after choosing, say, a couple of documents with high $P(d|q_i)$ values; as the novelty component is based on the product of $(1 - P(d|q_i))$ scores. For instance, if only two documents with 0.9 coverage probability of the aspect q_1 are in τ_q^*, for all the remaining documents, their $P(d|q_1)$ scores will be multiplied with 0.01, rendering this aspect practically useless. Furthermore, for the queries with a small number of aspects, the novelty scores computed for the remaining documents would be numerically very small after selecting the first few documents into τ_q^*; and from this point on, the selection process would be essentially guided by the relevance scores $P(q|d)$[2]. In the following sub-sections we discuss solution methods for each of these problems.

[2] The λ parameter can help to remedy the situation if the numerical differences are small; but it is still useless when the relevance and diversity scores vary in the order of magnitudes.

2.3.1.2 Relevance Score Normalization for xQuAD

While the problem of retrieval score normalization is investigated on its own in previous works and especially in the context of score-based ranking aggregation in meta-search (e.g., [37], [47], [24]), to the best of our knowledge, its impact on the result diversification is not yet addressed[3]. To remedy the aspect elimination problem discussed in the previous section, a straightforward solution can be using a normalization that does not map the top-ranked document relevance to 1 for a given list τ. To this end, a practical approach is using `Sum` normalization, defined as follows ([24]):

$$P(d|q) = \frac{s(d,q)}{\sum_{d_i \in \tau_q} s(d_i,q)} \tag{2.4}$$

Our problem at hand is different than the traditional ranking aggregation problem for meta-search engines in that the diversification is usually applied by the party that actually generates the initial retrieval scores for τ_q; i.e., the system does not only know the scores but also knows how they are computed. Exploiting this information, we propose an alternative normalization based on the highest possible score that can be generated for a given query and retrieval model. In this chapter, we employ two weighting models for initial retrieval, namely, a variant of Okapi-BM25 ([48]) and the query-likelihood language model with Dirichlet smoothing ([69]) as implemented in the Zettair text retrieval system ([66]).

For each retrieval model, we define a virtual best score that would be generated by a virtual document that is supposed to include each query term in the document with the frequency of the document length, i.e., as if the document is only composed of the query terms[4]. We set this virtual document's length to the average document length in the collection. While this is an unrealistically high upper-bound, our experiments reveal that it serves quite well for the purposes of this study. Therefore, we normalize the scores in τ_q by dividing each score by the virtual best score obtained for q using the same retrieval function that generatedτ_q. Note that, the same procedure is also applied

[3] Note that, Vargas et al. ([60]) recently proposed using the number of clicks instead of the retrieval scores for estimating the relevance probabilities. This is a viable though orthogonal approach to what we propose here.

[4] This is similar to computing an upper-bound for the relevance scores in dynamic pruning strategies, e.g., see [34].

while normalizing the scores for $P(d|q_i)$. We call this normalization `Virtual`:

$$P(d|q) = \frac{s(d,q)}{s(d^V,q)} \tag{2.5}$$

where $s(d^V, q)$ is the upper-bound score computed for the virtual best document d^V.

2.3.1.3 Document Novelty Estimation for xQuAD

As discussed above, the aspect fading problem arises as xQuAD computes the novelty of a document d for an aspect q_i by multiplying the dissatisfaction probability of q_i by the documents in the current set τ_q^*, as follows:

$$P(\bar{\tau}_q^*|q_i) = \prod_{d_j \in \tau_q^*}(1 - P(d_j|q_i)).$$

To avoid the negligible document novelty estimations (in comparison to the relevance scores), instead of taking the product of probabilities in $P(\bar{\tau}_q^*|q_i)$, we propose to use either arithmetic mean or geometric mean of the aspect dissatisfaction probabilities (as shown in Equations 2.6 and 2.7, respectively). This is a simple yet effective optimization to make the relevance and diversity sides of the mixture model comparable to each other in terms of their numerical values. Furthermore, by this optimization, λ can be determined more accurately among various queries, as it would serve only as a trade-off parameter as intended, but not for the purposes of remedying the gap between the numerical scores.

$$P(\bar{\tau}_q^*|q_i) = \frac{\sum_{d_j \in \tau_q^*}(1-P(d_j|q_i))}{|\tau_q^*|} \tag{2.6}$$

$$P(\bar{\tau}_q^*|q_i) = \sqrt[|\tau_q^*|]{\prod_{d_j \in \tau_q^*}(1 - P(d_j|q_i))} \tag{2.7}$$

The xQuAD versions that employ the arithmetic and geometric means of the probabilities in the novelty estimation component are referred to as `art_xQuAD` and `geo_xQuAD` in the rest of this study.

2.4 Experimental Setup

2.4.1 Collection, Queries and Aspects

We use the standard framework of "Diversity Task" as described in the TREC Web Track. In particular, we employ ClueWeb09 collection Part-B that includes around 50 million English web documents. The collection is initially parsed and indexed using the publicly available Zettair IR system ([66]). During the indexing, Zettair is executed with the "no stemming" option, yielding a vocabulary of 163,629,158 terms.

We report our results for TREC 2009 and 2010 topic sets that include 50 and 48 query topics, respectively[5]. For each topic in these sets, a number of sub-topics (up to 8) are described and the relevance judgments are provided at the sub-topic level. In the following experiments, we generate the query aspects in two ways. First, following the common practice in the previous works (e.g., [20], [51]), we use the "query" field of each topic as the initial query and generate its aspects (sub-queries) using the official sub-topic descriptions provided in the TREC topic sets. This case represents the idealistic scenario with the perfect knowledge of the query aspects. Secondly, we simulate a more realistic scenario and use top-10 query suggestions (auto-completions) collected from Google search engine as the aspects of each query, as first proposed in [51].

2.4.2 Initial Retrieval Model

For the initial retrieval runs, we used our homemade IR system with two popular retrieval models, namely, a variant of the well-known Okapi BM25 metric ([48]) and the query-likelihood language model with Dirichlet smoothing ([69]). For BM25 we set k_1 to 1.2 and b to 0.50, and for the language model (LM) we set $\mu = 2000$.

We first retrieve top-N candidate documents (τ_q) using one of these weighting models, and then run the diversification strategies to obtain the final top-k results, i.e, τ_q^*. Unless stated otherwise, for all the experiments we set $N = 100$ and $k = 20$. During

[5] Note that, we prefer to report evaluations separately on each topic set (but not their union) for the sake of comparability with the previous works.

retrieval, standard stopwords are removed.

Previous studies that experimented with the ClueWeb09 collection report that applying spam filtering can considerably improve the initial retrieval performance. Therefore, we also employ the spam filtering technique in ([18]). In particular, we utilize the publicly available Waterloo Spam Rankings[6] that assigns a spam percentile score to each document in the ClueWeb09 collection. During the initial retrieval, we set the relevance scores of the documents with spam score of less than 60 to $-\infty$ (as in [20]), so that these documents are eliminated from the top-N candidate documents.

2.4.3 Baseline Diversification Strategies and Evaluation Metrics

We have three strategies that serve as the diversification baselines. All of these strategies are greedy in nature and differ in the scoring function that is used to select the best document at each iteration, until all k documents are selected into τ_q^*. We briefly summarize these strategies as follows:

2.4.3.1 Intent Aware (IA)-select

This strategy aims to choose the document with the highest probability of satisfying the user given that all previously selected ones fail to do so [1]. The scoring function of IA-Select is as follows:

$$\sum_{q_i \in T} P(q_i|q) V(d|q, q_i) \prod_{d_j \in \tau_q^*} (1 - V(d_j|q, q_i)). \tag{2.8}$$

where $V(d|q, q_i)$ is the likelihood of d satisfying q for the underlying aspect q_i. As there is no strict enforcement on the implementation of this latter component in ([1]), it is replaced by $P(d|q_i)$ in our experiments (as in [51]).

2.4.3.2 org_xQuAD

This is the original xQuAD algorithm ([51]) as elaborated in the previous sections. Its scoring function, which is basically the combination of Equations 2.1 and 2.2, is

[6] http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

as follows:

$$(1-\lambda)P(d|q) + \lambda \sum_{q_i \in T} \left[P(q_i|q)P(d|q_i) \prod_{d_j \in \tau_q^*} (1 - P(d_j|q_i)) \right]. \tag{2.9}$$

2.4.3.3 PM2

In [20], two strategies, namely PM1 and PM2, are proposed within a proportionality-based diversification framework. The authors report that PM2 outperforms both its simpler predecessor PM1 and the original xQuAD for several evaluation metrics. Therefore, we include PM2 strategy as our third diversification baseline.

The intuition for this strategy is that, in a similar manner to allocation of seats to party representatives in some election systems, the ranks in τ_q^* should be allocated to documents that satisfy the query aspects in proportion to the popularity of these aspects in τ_q. At a given iteration p, first the *winner* aspect q_{i^*} is determined by the popularity of the aspect in τ_q and number of positions in τ_q^* that are allocated to this aspect up to iteration p (i.e., referred to as *quotient score*). Next, for this winner aspect q_{i^*}, PM2 selects the document d that maximizes the following score function:

$$\lambda \times qt[i^*] \times P(d|q_{i^*}) + (1-\lambda) \sum_{i \neq i^*} qt[i] P(d|q_i) \tag{2.10}$$

where $qt\,[i]$ is the quotient score and λ is the trade-off parameter between the relevance to the winner aspect and other aspects. Since the selected document in PM2 is expected to satisfy not only the winner aspect but also some other aspects, the number of positions allocated to each aspect is also updated accordingly (see [20] for details).

In all of these diversification baselines, we compute the relevance of the candidate documents to query aspects, i.e., $P(d|q_i)$, using the same model employed for the initial retrieval. While doing so, standard stopwords are removed from the aspect descriptions. Following the practice in [51], aspect probabilities $P(q_i|q)$ are computed uniformly as $1/|T|$, where T is the set of aspects $\{q_1, ..., q_m\}$ for a given query q.

For the strategies xQuAD and PM2, we test all values of the trade-off parameter λ in [0,1] range with a step size of 0.01, and the best λ values obtained on one of the topic sets (say, TREC 2009) is employed to obtain the reported results on the other topic set (say, TREC 2010).

2.4.3.4 Evaluation metrics

To evaluate the diversification performance, we compute most common measures, namely, α-nDCG, ERR-IA and Precision-IA, at the cut-off value of 20, using ndeval software[10]. For α-nDCG, α is typically set to 0.5, i.e., relevance and diversity are equally weighted.

Reproducibility of the results.

For search result diversification, a standard evaluation framework, namely "Diversity Task" in TREC Web Track, is available, which allows the use of a common dataset, queries and relevance judgments. Still, we identified some issues that complicate, or occasionally, make it impossible to make direct comparison of the results in different studies. First, even when the same document collection is employed (usually ClueWeb09 in the last years), the software used for indexing (e.g., Zettair, Terrier (e.g., [51]), Lemur/Indri (e.g., [20]), etc.) and choice of the parameters (list of stop-words, stemming options, handling various HTML tags during the parsing, spam filtering, etc.) can considerably alter the final results. Secondly, the retrieval models and their parameters can differ. A third issue that complicates comparing the results in our case is the list of query aspects. Even when the original TREC sub-topics are used for generating the aspects, there might be subtle differences in parsing the sub-topic descriptions. Obviously, if Web search engine suggestions are used to this end, the aspects employed by the works conducted at different times would differ significantly, making the results even less comparable.

In the light of above discussion, we provide the following data items to allow other researchers compare and contrast their findings with ours[11]. First, we provide the initial retrieval results, i.e., top-100 document identifiers, obtained over the ClueWeb09 Part-B collection. This would allow researchers to start with the same basis, i.e., candidate document set, to apply their own diversification strategies. Secondly, we provide the list of query aspects generated for each topic using TREC sub-topics and search engine suggestions.

2.5 Evaluation Results

In this section, we seek answers to the following research questions:

1. What is the impact of the score normalization techniques on the performance of the baseline diversification strategies, especially xQuAD and IA-Select that can suffer from the aspect elimination problem?
2. Can the xQuAD variants with the new relevance normalization and novelty estimation components outperform the original xQuAD strategy and other baselines?

In the following experiments, we essentially report our results using the BM25 model for the initial retrieval stage and official TREC sub-topics for representing the query aspects. In the next chapter, we will provide additional experiments where we explore the impact of the alternative retrieval models and aspect representations.

2.5.1 Performance of the Score Normalization Techniques

We begin with comparing the performance of the baseline diversification strategies on TREC 2009 and 2010 topic sets and using the aspects obtained from the official sub-topics and BM25 as the retrieval model (Table 2.1). For each diversification strategy, we normalize the relevance scores using the `MinMax` and `Sum` methods from the literature ([24]), as well as the virtual best score (denoted as `Virtual`) as we describe in this study. We also report the trade-off parameter λ employed in each case.

The following findings can be observed from Table 2.1. First, all diversification methods perform better than the non-diversified retrieval for both BM25 and LM models as shown in the literature. Secondly, the results show that the score normalization component affect all diversification methods; which is a justification for our interest in the normalization techniques in this study.

Table 2.1: Diversification performance w.r.t. the relevance normalization techniques for different retrieval models using the query aspects obtained from the official subtopics. The highest scores are shown in boldface.

	Relevance	2009				2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25	-	-	0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
org_xQuAD	MinMax	1.00	0.2242	0.3240	0.0769	0.99	0.2372	0.3281	0.1256
	Sum	0.15	0.2181	0.3154	**0.0902**	0.77	0.2506	**0.3570**	**0.1589**
	Virtual	1.00	**0.2318**	**0.3263**	0.0802	0.95	**0.2634**	0.3509	0.1315
IA-Select	MinMax	-	0.2242	0.3240	0.0769	-	0.2445	0.3386	0.1252
	Sum	-	0.2141	0.3162	**0.0929**	-	0.2529	0.3568	**0.1592**
	Virtual	-	**0.2318**	**0.3263**	0.0802	-	**0.2681**	**0.3660**	0.1334
PM2	MinMax	0.40	0.2233	0.3271	0.0899	0.57	0.2477	0.3576	0.1515
	Sum	0.57	0.2233	0.3266	0.0898	0.62	0.2571	0.3651	0.1555
	Virtual	0.52	**0.2328**	**0.3330**	**0.0932**	0.46	**0.2675**	**0.3713**	**0.1601**

Third, for the org_xQuAD and IA-Select strategies, the normalization schemes Virtual and/or Sum yield a better performance than the MinMax (especially on TREC 2010), which demonstrates that they can help in remedying the aspect elimination problem for these two diversification strategies. In particular, the org_xQuAD strategy with Virtual yields the highest ERR-IA scores (i.e., with a relative improvement of 3% and 5% over the second-best normalization technique for TREC 2009 and 2010 sets, respectively) and α-nDCG score (i.e., with a relative improvement of 1% over the MinMax on TREC 2009 set). Similarly, IA-Select achieves its best performance with the normalization techniques Sum (yielding an up to 19% relative improvement for the Precision-IA metric) and Virtual (yielding an up to 2% relative improvement for the ERR-IA and α-nDCG metrics). Finally, Virtual is the best technique also for PM2, as for all the reported evaluation metrics it yields a relative improvement that ranges from 2% to 4% over the second-best normalization technique.

2.5.2 Performance of xQuAD variants

In Table 2.2, we compare the diversification performance of the original xQuAD to the variants that use arithmetic and geometric means for the novelty estimation components, namely, art_xQuAD and geo_xQuAD, respectively. For the ease of comparison, we repeat the results for org_xQuAD from Table 2.1. As before, each strategy is combined with three different normalization techniques.

Our findings in Table 2.2 reveal that the novelty estimation methods proposed in this study considerably improve the org_xQuAD. The highest scores for all of the evaluation metrics (as shown in boldface in Table 2.2) are produced by the art_xQuAD and geo_xQuAD strategies that usually employ Virtual method for the relevance score normalization. For instance, using the TREC2010 topics, the art_xQuAD variant with Virtual normalization scheme provides a relative improvement of around 7% for both ERR-IA and α-nDCG metrics over the best-performing configuration of the original xQuAD strategy.

Table 2.2: Diversification performance of the xQuAD variants using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface.

	Relevance	2009				2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25			0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
org_xQuAD	MinMax	1	0.2242	0.3240	0.0769	0.99	0.2372	0.3281	0.1256
	Sum	0.15	0.2181	0.3154	0.0902	0.77	0.2506	0.3570	0.1589
	Virtual	1	0.2318	0.3263	0.0802	0.95	0.2634	0.3509	0.1315
geo_xQuAD	MinMax	0.92	0.2305	0.3301	0.0857	0.97	0.2494	0.3461	0.1418
	Sum	0.15	0.2174	0.3134	0.0892	0.75	0.2495	0.3515	0.1571
	Virtual	0.56	0.2333	0.3292	0.0905	0.86	**0.2842**	**0.3876**	0.1606
art_xQuAD	MinMax	0.91	0.2326	**0.3374**	0.0912	0.92	0.2629	0.3732	0.1578
	Sum	0.15	0.2174	0.3134	0.0892	0.75	0.2495	0.3515	0.1571
	Virtual	0.57	**0.2338**	0.3301	**0.0918**	0.86	0.2835	0.3868	**0.1609**

We further investigate the impact of the trade-off parameter λ on the performance of xQuAD using the union of topics from TREC 2009 and 2010. In Figure 2.2, we report the α-nDCG@20 scores for org_xQuAD using all three normalization methods, and for our geo_XQuAD and art_XQuAD only with the best-performing normalization, Virtual (to simplify the plot). The trade-off parameter λ is varied in the range [0, 1] with a step size of 0.01. Our findings reveal that both Sum and Virtual normalization techniques outperform MinMax for the entire range of values for the org_xQuAD strategy. Furthermore, while Sum reaches the peak effectiveness score when λ is around 0.15, the other two techniques perform better as we increase the λ; and the overall best performance for `org_xQuAD` is obtained with Virtual for $\lambda = 1$. Vargas et al. ([60]) and Zheng et al. ([71]) independently report a similar finding; i.e., the best λ value being 1 for xQuAD, and the latter work attributes this due to the use of real sub-topics from TREC as the query aspects. Nevertheless, our `art_xQuAD` and `geo_xQuAD` strategies with Virtual normalization yield the best effectiveness results and outperform `org_xQuAD` coupled with any of these normalization techniques.

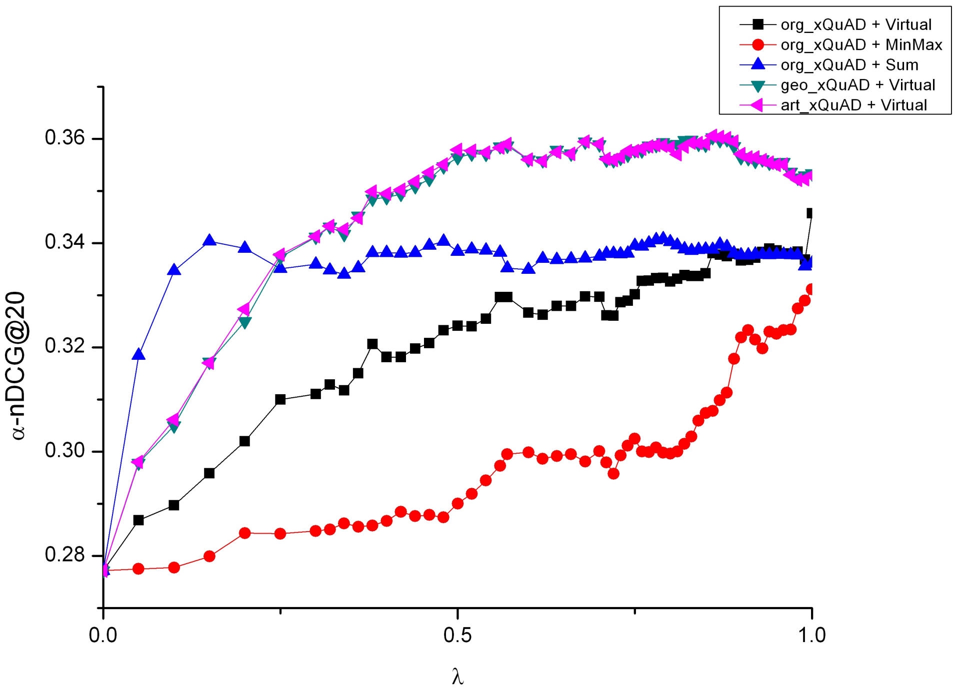

Figure 2.2: Diversification performance of the xQuAD variants vs. trade-off parameter λ.

2.5.3 Summary of the Main Findings

Our experimental evaluations reveal that the new xQuAD variants `art_xQuAD` and `geo_xQuAD` (coupled with the Virtual normalization technique) considerably improve the performance of the original strategy. We further show that the score and rank aggregation methods adapted for the result diversification problem are quite effective. In particular, we find that `mix_CombSUM` and `mix_MC2` are the best-performing representatives of the score and rank aggregation methods, respectively.

Overall, the proposed xQuAD variants and certain ranking aggregation methods (especially mix_CombSUM) consistently outperform all three diversification baselines for most of the cases and evaluation metrics (as shown in Tables 3.3, 3.4, 3.5, and 3.6). The success of `mix_CombSUM` is remarkable as its computational complexity is less than the baseline diversification strategies and xQuAD variants, as we discuss in the section Score-based Aggregation Methods. This finding further justifies the use of the ranking aggregation methods in the context of search result diversification, as we propose in this study.

2.6 Conclusion

In this chapter, we improved the state-of-the-art in explicit search result diversification. Namely, we proposed optimizations for the relevance score normalization and novelty estimation components of xQuAD, a top-performing approach for the explicit result diversification. We showed that the new xQuAD variants outperform the original strategy and normalization methods improve xQuAD and other diversification baselines employed in our study.

CHAPTER 3

RANKING AGGREGATION METHODS FOR DIVERSIFICATION

In this chapter [1], we inspired from the success of the current diversification strategies that exploit the relevance of the candidate documents to individual query aspects, and propose to use ranking aggregation methods to diversify search results. In Section 3.1 we provide our motivation and in Section 3.2 we give some information about related work in ranking aggregation methods. In Section 3.3 we explain score-based and rank-based aggregation methods and our proposed adaptations to these methods to be used in diversification framework. In the next section, we evaluate the ranking aggregation methods in diversification domain and compare the effectiveness of ranking aggregation methods to baseline diversification methods described in previous chapter in different setup configurations. We conclude the chapter with Section 3.5.

3.1 Introduction

Our second contribution is motivated by the observation that computing the relevance of the candidate documents to each query aspect plays a central role in the success of the current explicit diversification strategies, such as xQuAD. Encouraged by this finding, we propose to materialize the re-rankings of the candidate documents for each query aspect and then merge them by adapting the score(-based) and rank(-based) aggregation methods that are widely applied in the meta-search scenario. In

[1] A. M. Ozdemiray and I. S. Altingovde. "Explicit search result diversification using score and rank aggregation methods", *Journal of the Association for Information Science and Technology*, 66(6):1212-1228. ©2015 John Wiley and Sons. `http://dx.doi.org/10.1002/asi.23259`. Reprinted by permission with license number 371383091410

other words, we cast the diversification problem to the problem of *aggregating* the re-rankings per query aspect. We hypothesize that if each of these re-rankings can place the most relevant documents for their respective aspects in their top-k results, then the aggregation of these rankings would be both relevant and diverse in terms of the coverage of these aspects, as required.

To the best of our knowledge, we are the first to propose to model and solve the result diversification problem using the score and rank aggregation methods. For the purposes of score aggregation, we adapt two traditional methods, namely, CombSUM and CombMNZ ([54], [31]), and investigate their performance employing various score normalization techniques. We show that the normalization strategy proposed for xQuAD proves to be useful for the score aggregation methods, as well. For the rank aggregation, we adapt the classical methods like simple voting and Borda voting ([21]) as well as the Markov chain based approaches ([23]). We extend both the score and rank aggregation methods by weighting the initial ranking and aspect rankings within the classical probability mixture framework of the diversification approaches, for the purposes of balancing the relevance and diversity in the final result.

We further find that, for various parameter configurations and evaluation metrics, certain ranking aggregation methods as adapted here are also superior to all of the baseline strategies. This is a remarkable finding as these ranking aggregation methods can be computed more efficiently than the baseline diversification strategies and our xQuAD variants.

3.2 Related Work

In real life, a common use of ranking aggregation (a.k.a. ranking fusion, result merging/fusion) methods is the election systems that allow voters to rank the candidates in the order of preference[2]. In computer science, score(-based) and/or rank(-based) aggregation methods are investigated for and applied to various research problems, such as meta-search ([5], [23], [47]), federated search ([55]), spam detection ([23]), word association ([23]), search quality evaluation ([38]) and result generation from search

[2] http://en.wikipedia.org/wiki/Voting_system

engine caches ([8]). However, to the best of our knowledge, no previous study proposes to adapt such methods for the result diversification task (We discuss the details of these methods in the section Ranking Aggregation Methods for Diversification.).

Note that, while the proportionality framework of Dang and Croft ([20]) also has its roots in the voting systems; their approach is different than ours. More specifically, their diversification strategies are based on the votes given to the *aspects* whereas here we focus on the votes given to the *documents* by each aspect.

3.3 Ranking Aggregation Methods for Diversification

A key component of the xQuAD framework discussed in the previous chapter is $P(d|q_i)$, i.e., the likelihood of observing d for the aspect q_i (see Equations 2.1 and 2.2). In practice, this component computes the relevance of candidate documents to each query aspect using a retrieval model. Indeed, such a computation is not only involved in xQuAD, but also included in two other competing strategies, namely, IA-Select ([1]) and PM2 ([20]). Encouraged by the success of all these explicit diversification strategies demonstrated in the earlier works, we propose an alternative perspective to exploit this key component.

In this study, we materialize the *re-rankings* of the candidate documents for each query aspect and then tackle the result diversification problem from a *ranking aggregation* perspective. In the classical ranking aggregation context, the goal is producing a merged list τ from the given full or partial rankings $\{\tau_1, ..., \tau_m\}$ so that the final list τ has the minimal distance from each individual list τ_i. In our case, for a given query q with the set of aspects $T = \{q_1, ..., q_m\}$ and initial retrieval result $\tau_q(|\tau_q| = N)$, let's assume that τ_{q_i} denotes the re-ranking of the documents in τ_q with respect to the relevance probabilities $P(d|q_i)$ for the aspect q_i, and $\tau_{q_i}^k$ denotes the top-k documents in τ_{q_i}. We hypothesize that if each ranking τ_{q_i} places the most relevant documents higher for the corresponding aspect q_i, then the aggregation of these top-k rankings would be both relevant and diverse; i.e., cover as many diverse aspects as possible.

In the context of ranking aggregation described above, it is tempting to optimize the Kendall tau distance, which counts the number of pairwise disagreements between two lists, as a typical measure of distance between two rankings. However, Dwork et al. ([23]) show that computing the aggregation that optimizes the Kendall distance, so-called *Kemeny optimal aggregation*, is NP-hard even for four different rankings. Fortunately, there are various sub-optimal methods that are shown to serve well in real life applications, such as building meta-search engines and combating spam results (see the section Related Work for other examples). Such ranking aggregation methods in the literature are categorized based on the type of information used during the fusion process. Score-based aggregation methods exploit the relevance scores associated with each document in each ranking, whereas rank-based aggregation methods only rely on the document's position in the list. In the rest of this section, we adapt a number of representative methods from each category for the purposes of result diversification.

An important difference of our problem from the rank aggregation in meta-search is that in our setup, there exists an initial ranking τ_q, and all τ_{q_i} lists are basically re-rankings of the former[3]. In the ranking aggregation methods employed in this study, we exploit both τ_q and τ_{q_i} rankings to generate the final diversified ranking τ_q^*. To emphasize this mixture of the initial and aspect rankings, the abbreviations of the method names are prefixed with mix in the following discussions.

3.3.1 Score-based Aggregation Methods

One of the well-known approaches for ranking aggregation in the context of meta-search is combining the normalized relevance scores with various functions, such as min, max, median and sum ([54], [31]). Among these variants, CombSUM and CombMNZ are the most effective ones that are widely employed in the subsequent works (e.g., [47], [5]).

[3] Since our aggregation methods operate on the re-rankings of the initial ranking τ_q, the missing document problem usually encountered in meta-search (e.g., see [22]) is not a concern for the result diversification framework.

3.3.1.1 CombSUM (mix_CombSUM).

This method computes the overall score of d for the query q by simply adding up the document's scores in each ranking τ_{q_i}. For the purposes of diversification, we also incorporate the initial ranking τ_q using a mixture model as typical in all diversification frameworks and come up with the following formula:

$$S(q,d) = (1-\lambda)P(d|q) + \lambda \sum_{q_i \in T} P(q_i|q)P(d|q_i), \tag{3.1}$$

where $P(q_i|q)$ denotes the aspect likelihood that is typically included in most of the explicit diversification strategies. A similar notion of associating priorities to the rankings has also been employed for the score aggregation methods in the meta-search context ([47]). The final ranking τ_q^* includes the top-k documents (computed using a heap of size *k*) in descending order of $S(q,d)$ values (ties are broken randomly).

Notice that the formula is indeed quite similar to that of xQuAD (and IA-Select method defined in [1]) with one crucial difference: the latter strategy constructs the final ranking in a greedy manner and takes into account the novelty with respect to the documents that are already selected in τ_q^* while computing the score $S(q,d)$. In contrast, `mix_CombSUM` applies a linear weighted summation of the scores for every aspect as well as the initial results, which is cheaper in terms of the computational cost. In particular, `mix_CombSUM` needs to make a single pass over the candidate documents to compute the scores, and then constructs the final ranking τ_q^* using a heap of size *k* (e.g., see [65]), which implies an overall complexity of $O(N \log k)$[4]. In contrast, since xQuAD compares every candidate document to those already selected into the τ_q^* for each iteration, its overall complexity is $O(Nk)$ ([9]). Therefore, `mix_CombSUM` is more efficient than xQuAD, as well as the other diversification baselines IA-Select and PM2, which actually have the same computational complexity as xQuAD (see the section Experimental Setup for the details of the baseline strategies).

[4] Following the practice in the literature ([9]), we neglect the number of query aspects, $|T|$, from the complexity analysis of the methods presented in this study, as it is assumed to be a small constant.

3.3.1.2 CombMNZ (mix_CombMNZ)

. This method is similar to the previous one, but the score of d is weighted by the sum of the votes for d given by each $\tau^k_{q_i}$, as follows:

$$S(q,d) = (1-\lambda)P(d|q) + \lambda \sum_{q_i \in T} v(d,\tau^k_{q_i}) \sum_{q_i \in T} P(q_i|q)P(d|q_i). \tag{3.2}$$

In Equation 3.2, $v(d,\tau^k_{q_i})$ denotes the number of rankings $\tau^k_{q_i}$ where d appears, and it is computed as

$$v(d,\tau^k_{q_i}) = \begin{cases} 1, & if\ d \in \tau^k_{q_i}, \\ 0, & otherwise. \end{cases} \tag{3.3}$$

Similar to the `mix_CombSUM` method, the final ranking τ^*_q includes the top-k documents (computed using a heap) in descending order of $S(q,d)$ values (ties are broken randomly). Thus, the computational complexity of `mix_CombMNZ` is $O(N \log k)$. This overall complexity subsumes the cost of generating the top-*k* rankings per aspect ($\tau^k_{q_i}$), which is again $O(N \log k)$, given that the number of aspects is a small constant that can be neglected, as in the previous analysis.

Note that, the relevance probabilities $P(d|q)$ and $P(d|q_i)$ in Equations 3.1 and 3.2 should be normalized, as in the case of xQuAD. As we mention before, the diversification scenario allows us to employ the `Virtual` technique that makes use of the actual retrieval model and the collection statistics, in addition to the traditional `MinMax` and `Sum` normalization schemes. In our experimental evaluations, we consider all three normalization techniques along with the `mix_CombSUM` and `mix_CombMNZ` techniques.

3.3.2 Rank-based Aggregation Methods

In rank(-based) aggregation methods, the relevance scores are not taken into account and the final ranking is obtained by only using the order of documents in each aspect ranking.

3.3.2.1 Simple voting (mix_SV).

In this method (e.g., see [8]), we assume that each document $d \in \tau_q$ receives a vote from each ranking[5] $\tau_{q_i}^k$ weighted with the aspect likelihood $P(q_i|q)$; i.e, the vote count per document is computed as:

$$C(q,d) = (1-\lambda)v(d,\tau_q^k) + \lambda \sum_{q_i \in T} P(q_i|q)v(d,\tau_{q_i}^k) \tag{3.4}$$

where vote $v(d,\tau_{q_i}^k)$ is computed as in Equation 3.3 and τ_q^k denotes the top-k documents of the initial ranking τ_q.

The final ranking τ_q^*includes the top-k documents in descending order with respect to the vote counts $C(q,d)$ (ties are broken using $P(d|q)$ values). As we discussed for the mix_CombMNZ method, the worst-case complexity of `mix_SV` is also $O(N \log k)$.

3.3.2.2 Borda voting (mix_BV).

This is based on Borda's classical method ([21]) that also takes the position of the documents in the ranked lists into account while computing the vote counts, as follows:

$$C(q,d) = (1-\lambda)\tau_q(d) + \lambda \sum_{q_i \in T} P(q_i|q)\tau_{q_i}(d) \tag{3.5}$$

where $\tau_q(d)$ is the rank position of d in some list τ_q. The final ranking τ_q^* is constructed in ascending order with respect to the vote count (again, ties are broken using $P(d|q)$ values). Note that, since this method requires computing the ranking τ_{q_i}per aspect (but not only top-*k* re-rankings $\tau_{q_i}^k$as in the previous methods), its overall complexity is $O(N \log N)$.

3.3.2.3 Markov chain based methods.

Dwork et al. ([23]) have proposed using Markov chains for aggregating ranked partial lists and described four different variants. In what follows we discuss this approach

[5] We consider only $\tau_{q_i}^k$ lists for this method, as using τ_{q_i}'s would result in the same vote count for all the documents.

using our own problem setup and notation, please refer to [23] for basics and adaptation to the general ranking aggregation problem.

For this case, we define the document space U as the union of the documents in τ_q^k as well as the all top-k re-rankings per aspect ($\tau_{q_i}^k$), as follows:

$$U = \bigcup_{q_i \in T} \tau_{q_i}^k \bigcup \tau_q^k \tag{3.6}$$

Note that, the document space is limited to top-k documents from each ranked list, as otherwise the number of states and size of the transition matrix would be too large for on-the-fly-computation of the diversified results. In this approach, each document $d \in U$ is considered as a state in the Markov chain. A non-negative stochastic matrix M (of size $|U| \times |U|$) defines the probability of the systems' transitions from one state to another. In our case, these probabilities are based on the positions of the documents in various ranked lists. Once the system starts on some state probability distribution (typically, the uniform distribution), it eventually reaches to a unique fixed point where the state distribution does not change. This is called the stationary distribution and for our purposes, the stationary probabilities of the states at this point are used to sort the documents (states) and obtain the final τ_q^*.

Dwork et al. ([23]) define four different Markov chains by describing four different ways of constructing the transition matrix, as follows:

1. *MC1*: If the current state (document) is d_i, the next state is chosen uniformly from the multiset of all documents d_j such that both d_i and d_j appear in some ranking τ and d_j is ranked higher; i.e., $\tau(d_j) \leq \tau(d_i)$.

2. *MC2*: If the current state (document) is d_i, then first pick a ranking τ uniformly from all rankings that include d_i, and then choose a document d_j uniformly that is ranked higher than d_i in τ, i.e., $\tau(d_j) \leq \tau(d_i)$.

3. *MC3*: If the current state (document) is d_i, then first pick a ranking τ uniformly from all rankings that include d_i, and then choose a document d_j uniformly from τ. If $\tau(d_j) < \tau(d_i)$ then go to d_j else stay in the state d_i.

4. *MC4*: If the current state (document) is d_i, then first pick a document d_j uniformly from U. If $\tau(d_j) < \tau(d_i)$ for the majority of the lists τ that ranked both

d_i and d_j, then go to d_j, else stay in the state d_i.

Some nice theoretical intuitions for constructing these particular Markov chains are provided in [23], and a set of example transition matrices for the ranking aggregation in meta-search scenario is given in [47]. Following the practice in the latter work, we computed the stationary distribution using the simple power-iteration method. That is, we start the iteration by a state vector where each state has $1/|U|$ probability and repetitively multiply it with the transition matrix M till the state probabilities are stabilized, i.e., converge to the stationary distribution.

The computational complexity of computing (or, more precisely, sampling) the stationary distribution is $O(|U|)$, as shown by Dwork et al. ([23]). Given that the input top-k rankings per aspect ($\tau_{q_i}^k$) (see Equation 13) can be constructed in $O(N \log k)$ time, as well as the final ranking (using a heap of size k as before), the overall complexity becomes $O(|U| + \ N \log k) \approx O(N \log k)$. Therefore, the methods based on the Markov chains are still more efficient than the diversification baselines (i.e., xQuAD, IA-Select and PM2), which have the complexity of $O(Nk)$.

Note that, as we use both the initial ranking τ_q and aspect rankings τ_{q_i} while constructing the document space U, we again prefix the names of these methods with mix, hereafter.

3.4 Experiments and Results

In this chapter, we used the same experimental setup as the previous chapter.

3.4.1 Evaluation Results

In this section we first evaluate the performance of the score aggregation methods and then we report our results for rank aggregation methods. Finally we seek answer to the following research questions

1. Can proposed diversification methods (i.e. xQuAD variants, score and rank aggregation methods) outperform the diversification baselines

Table 3.1: Diversification performance of the score aggregation methods using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface.

	Relevance	2009				2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25		-	0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
mix_CombMNZ	MinMax	0.45	0.2188	0.3191	0.0915	0.85	0.2510	0.3536	0.1563
	Sum	0.3	0.2135	0.3142	0.0950	0.05	0.2448	0.3502	0.1541
	Virtual	0.75	0.2209	0.3216	0.0958	0.25	0.2450	0.3487	0.1557
mix_CombSUM	MinMax	0.7	0.2230	0.3255	0.0923	0.95	0.2599	0.3599	**0.1638**
	Sum	0.15	0.2174	0.3134	0.0892	0.75	0.2495	0.3511	0.1569
	Virtual	0.55	**0.2370**	**0.3320**	**0.0975**	0.9	**0.2719**	**0.3712**	0.1609

2. How does retrieval model and aspect representations affect our proposed methods' effectiveness

In the following experiments, we essentially report our results using the BM25 model for the initial retrieval stage and official TREC sub-topics for representing the query aspects. In the section Impact of the Components and Parameters, we provide additional experiments where we explore the impact of the alternative retrieval models and aspect representations.

Table 3.1 shows the performance of the score aggregation methods `mix_CombSUM` and `mix_CombMNZ` when coupled with each of the three relevance normalization schemes described in Section 2.3.1.2. Our findings reveal that both methods significantly outperform the non-diversified BM25 baseline. For both TREC 2009 and 2010 topic sets, `mix_CombSUM` coupled with the `Virtual` normalization technique outperforms all other configurations by 2% to 6% (relatively) for the majority of the metrics (see the boldfaced cells in Table 3.1). This is a further evidence for the robustness and usability of the `Virtual` technique in the context of result diversification.

Next, we report our results for the rank aggregation methods, namely, Simple Voting (`mix_SV`), Borda Voting (`mix_BV`) and Markov chain based models (`mix_MC1`, `mix_MC2`, `mix_MC3`, and `mix_MC4`). Table 3.2 reveals that, `mix_MC2` outperforms both the other Markov chain based strategies and the relatively simplistic methods `mix_SV` and `mix_BV` (with a relative improvement of more than 3% over the

Table 3.2: Diversification performance of the rank aggregation methods using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface.

	2009				2010			
	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25		0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
mix_SV	0.9	0.2077	0.3094	**0.0954**	0.9	0.2327	0.3381	0.1524
mix_BV	0.85	0.2140	0.3135	0.0910	1	0.2437	0.3475	**0.1743**
mix_MC1	-	0.2129	0.3182	0.0915	-	0.2234	0.3328	0.1460
mix_MC2	-	**0.2249**	**0.3307**	0.0888	-	**0.2559**	**0.3645**	0.1404
mix_MC3	-	0.2183	0.3204	0.0914	-	0.2275	0.3367	0.1462
mix_MC4	-	0.2177	0.3157	0.0878	-	0.2489	0.3505	0.1390

second-best method for the ERR-IA and α-nDCG metrics). In contrary, the latter methods perform well for the P-IA metric. A further comparison of Tables 3 and 4 shows that score aggregation methods are usually superior to Simple Voting and Borda Voting. However, rank aggregation methods based on the Markov chains perform comparable to the score based methods. These findings confirm the previous results reported in the context of meta-search ([47]).

Finally, in Table 3.3 we make an overall comparison of the best-performing configurations (determined based on the α-nDCG@20 scores) of the state-of-the-art diversification baselines (see Section 2.4.3 to those representing each class of the strategies proposed in Section 2.3.1.3, namely, xQuAD variants, and the score and rank aggregation methods. From Table 3.3, we first observe that `Virtual` turns out to be the most effective normalization technique for the majority of the diversification strategies. More crucially, the score aggregation method `mix_CombSUM` and xQuAD variants are always the best performers for different evaluation metrics on both TREC 2009 and 2010 topic sets (see the boldfaced cells in Table 3.3). Given that we have three strong diversification strategies that are presented in their best configurations, our improvements are remarkable. For instance, on TREC 2010, our `geo_XQuAD` variant provides a relative improvement of 6% and 4% for the ERR-IA and α-nDCG metrics, respectively, over the best diversification baseline (PM2 with the Virtual normalization).

Table 3.3: Comparison of the best cases for the baseline and proposed methods using the query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface.

		Rel. norm.	λ	ERR-IA	α-nDCG	P-IA
TREC 2009						
Baseline	BM25	-	-	0.1878	0.2757	0.0760
	IA-Select	Virtual	-	0.2318^{B}	0.3263^{B}	$0.0802^{P,C}$
	org_xQuAD	Virtual	1.00	0.2318^{B}	0.3263^{B}	$0.0802^{P,C}$
	PM2	Virtual	0.52	0.2328^{B}	0.3330^{B*}	$0.0932^{X,I}$
Proposed	mix_CombSUM	Virtual	0.55	$\mathbf{0.2370}^{B*}$	0.3320^{B*}	$\mathbf{0.0975}^{B,X}_{I}$
	mix_MC2	-	-	0.2249^{B*}	0.3307^{B*}	0.0888^{B}
	art_xQuAD	MinMax	0.91	0.2326^{B*}	$\mathbf{0.3374}^{B*}$	0.0912^{B*}
TREC 2010						
Baseline	BM25	-	-	0.1947	0.2788	0.1254
	IA-Select	Virtual	-	$0.2681^{B*,X_g}$	$0.3660^{B*,X_g*}$	$0.1334^{X*,P*}_{C*,X_g*}$
	org_xQuAD	Sum	0.77	0.2506^{B}	$0.3570^{B*,X_g}$	$0.1589^{B*,I*}$
	PM2	Virtual	0.46	$0.2675^{B,X_g*}$	$0.3713^{B*,X_g*}$	0.1601^{B*I*}_{M}
Proposed	mix_CombSUM	Virtual	0.9	$0.2719^{B*,X_g}$	$0.3712^{B*,X_g}$	$\mathbf{0.1609}^{B*,I*}_{M}$
	mix_MC2	-	-	0.2559^{B*}	0.3645^{B*}	$0.1404^{B,P}_{C,X_g}$
	geo_xQuAD	Virtual	0.86	$\mathbf{0.2842}^{B*,P*}_{I,C}$	$\mathbf{0.3876}^{B*,X,P*}_{I*,C}$	$0.1606^{B*,I*}_{M}$

Note. The sub/superscripts of a result denote a statistically significant difference from the BM25 (B), IA-Select (I), org_xQuAD (X),PM2 (P), mix_MC2 (M), mix_CombSUM (C) or geo_xQuAD (X_g) at 0.05 level. The sub/superscripts with a star denote a statistically significant difference at 0.01 level.

We also conducted an analysis of the statistical significance of our findings using Wilcoxon signed-rank test at the 95% and 99% confidence levels. We found that while all the diversification strategies significantly outperform the non-diversified baseline for most of the cases, the results are mixed among the diversification strategies. However, recent works in the literature also present similar findings. For instance, Dang and Croft report that none of the improvements of PM2 over the original xQuAD strategy are indeed statistically significant on TREC 2009 topics; and their results are also mixed on TREC 2010 (see Table 2 in [20]). We also observed a larger number of statistically significant cases on TREC 2010 topic set, which is possibly due to the much larger differences among the actual effectiveness scores of the strategies (e.g., see Table 3.3).

3.4.2 Impact of the Components and Parameters

3.4.2.1 Impact of the aspect representation.

In this experiment, for each query in our topic files, we obtain the top-10 query suggestions (auto-completions) from Google search engine to represent the aspects, as in [51]. Some of these suggestions include terms that are not in the collection vocabulary, and after filtering the suggestions with such terms, we ended up with 9 aspects per query, on the average.

Table 3.4: Comparison of the best cases for the baseline and proposed methods using the query aspects obtained from the suggestions. The highest scores across all methods are shown in boldface.

		Rel. norm.	λ	ERR-IA	α-nDCG	P-IA
TREC 2009						
Baseline	BM25	-	-	0.1878	0.2757	0.0760
	IA-Select	MinMax	-	0.1778	0.2814	0.0783
	org_xQuAD	MinMax	0.83	0.1884^{C}	0.2801^{X_g}	0.0757^{X_g}
	PM2	MinMax	0.25	0.1937	0.2891	0.0840
Proposed	mix_CombSUM	Virtual	0.25	$0.2004^{B,X}$	0.2913	0.0847
	mix_MC4		-	**0.2014**	0.2937^{B}	0.0801
	geo_xQuAD	MinMax	0.86	0.1938	$\mathbf{0.2948}^{X}$	$\mathbf{0.0868}^{B,X}$
TREC 2010						
Baseline	BM25	-	-	0.1947	0.2788	0.1254
	IA-Select	Virtual	-	0.2028	0.2966	$0.1129^{X,C*}$
	org_xQuAD	Sum	0.1	0.2041^{C}	0.2963	$0.1369^{B,I}_{C}$
	PM2	Sum	0	0.2145	0.3028	0.1297^{C*}
Proposed	mix_CombSUM	Virtual	0.3	0.2161^{X}	**0.3123**	$\mathbf{0.1499}^{B*,X,P*}_{I*,S*}$
	mix_SV	-	0.85	**0.2271**	0.3027	0.1277^{C*}
	art_xQuAD	Virtual	0.38	0.2070	0.3008	0.1360^{B}

Note. The sub/superscripts of a result denote a statistically significant difference from the BM25 (B), IA-Select (I), org_xQuAD (X), PM2 (P), mix_SV (S), mix_CombSUM (C) or geo_xQuAD (X_g) at 0.05 level. The sub/superscripts with a star denote a statistically significant difference at 0.01 level.

In Table 3.4, we present the best-performing configurations for the sake of brevity[6].

[6] The detailed results are at the Appendix

We first notice that the effectiveness scores are considerably lower than those presented in Table 5. This is expected and confirms the previous findings (e.g., see [51]), as the suggestions cannot perfectly represent the query aspects as the actual sub-topics from TREC. As a further difference, for the baseline strategies, there are cases where MinMax outperform the others. This is because in this setup, we have a far larger number of aspects per query as mentioned above, and this probably makes the aspect elimination problem less of a concern.

Nevertheless, the trends in Table 3.4 are still similar to our previous results, as the xQuAD variants and/or rank and score aggregation methods are superior to all the traditional baselines. In particular, geo_xQuAD (mix_CombSUM) achieves the highest P-IA and α-nDCG scores on TREC 2009 (2010) sets, respectively. Remarkably, mix_CombSUM provides a relative improvement of 15.3% over the best-performing baseline strategy, PM2 with the Sum normalization, in terms of the P-IA metric on TREC 2010 topics. In this latter case, the differences between the mix_CombSUM and all other strategies (except art_xQuAD) are found to be statistically significant at 95% confidence level.

3.4.2.2 Impact of the initial retrieval model.

In order to investigate the impact of the initial retrieval model, we repeated all the experiments using the query-likelihood language model (LM) with Dirichlet smoothing ([69]). Table 3.5 shows the best-performing configurations when the query aspects are based on the TREC sub-topics. As before, the proposed methods perform quite well and for the majority of the evaluation metrics, the score aggregation methods mix_CombSUM and mix_CombMNZ outperform the best-performing baseline methods by 1% to 11% (relatively). Note that, the second best-performer is usually an xQuAD variant, either art_xQuAD or geo_xQuAD.

In Table 3.6, we continue with the best-performing configurations for the experiments that employ the search engine suggestions as the query aspects. As before, the actual scores are lower for all metrics in comparison to Table 3.5, but the trends are similar in that the score aggregation method mix_CombSUM yields the best diversification performance for the majority of the cases, especially on TREC 2009.

Table 3.5: Comparison of the best cases for the baseline and proposed methods using the LM for the initial retrieval and query aspects obtained from the official sub-topics. The highest scores across all methods are shown in boldface.

		Rel. norm.	λ	ERR-IA	α-nDCG	P-IA
TREC 2009						
Baseline	LM			0.0877	0.1895	0.0798
	IA-Select	MinMax	-	0.2240^{B*}	0.3311^{B*}	0.0920
	org_xQuAD	MinMax	1.00	0.2240^{B*}	0.3311^{B*}	0.0920
	PM2	MinMax	0.66	0.2160^{B*},	0.3259^{B*}	$0.0923^{C,X_a}$
Proposed	mix_CombMNZ	MinMax	0.45	$\mathbf{0.2343}^{B*}$	$\mathbf{0.3334}^{B*}$	$\mathbf{0.1022}^{P}$
	mix_MC2		-	0.2222^{B*}	0.3282^{B*}	0.0944
	art_xQuAD	Virtual	0.95	0.2240^{B*}	0.3284^{B*}	0.1006^{P}
TREC 2010						
Baseline	LM	-	-	0.1959	0.2842	0.1406
	IA-Select	Virtual	-	0.2631^{B*}	0.3624^{B*}	$0.1291^{X*,C*}_{X_g*}$
	org_xQuAD	Sum	0.56	0.2634^{B*}	0.3689^{B*}	$0.1562^{P*,I*}_{M}$
	PM2	MinMax	0.76	0.2679^{B*}	0.3751^{B*}	$0.1314^{X*,C,}_{X_g*}$
Proposed	mix_CombSUM	Virtual	-	$\mathbf{0.2740}^{B*}$	$\mathbf{0.3805}^{B*}$	$0.1519^{P,I*}$
	mix_MC2	-	-	0.2645^{B*}	0.3714^{B*}	0.1394^{X}
	geo_xQuAD	Virtual	0.78	0.2721^{B*}	0.3792^{B*}	$\mathbf{0.1614}^{P*,I*}$

Note. The sub/superscripts of a result denote a statistically significant difference from LM (B), IA-Select (I), org_xQuAD (X),PM2 (P), mix_MC2 (M), mix_CombSUM (C), art_xQuAD (X_a) or geo_xQuAD (X_g) at 0.05 level. The sub/superscripts with a star denote a statistically significant difference at 0.01 level.

3.4.2.3 Other score normalization techniques

In addition to those discussed in the previous sections, we also repeat our experiments using another normalization technique, namely, z-score normalization ([47]). This technique subtracts the mean score of τ from each score, and then divides them by the standard deviation of the ranking. Since the resulting score values do not fall into [0, 1] range, they are further normalized using the `MinMax` method. In our experiments, we find that the z-score normalization does not yield better results than `MinMax` when coupled with our diversification strategies, and thus the results are not reported here.

Table 3.6: Comparison of the best cases for the baseline and proposed methods using the LM for the initial retrieval and query aspects obtained from the suggestions. The highest scores across all methods are shown in boldface.

		Rel. norm.	λ	ERR-IA	α-nDCG	P-IA
TREC 2009						
Baseline	LM	-		0.0877	0.1895	0.0798
	IA-Select	MinMax	-	0.1913^{B*}	0.2916^{B*}	$0.0908^{X,X_g}$
	org_xQuAD	Sum	0.57	0.1928^{B*}	0.2929^{B*}	$0.1014^{I,M}$
	PM2	MinMax	0.74	0.1891^{B*}	0.2870^{B*}	0.0921^{X_g}
Proposed	mix_CombSUM	MinMax	0.85	$\mathbf{0.1992}^{B*}$	$\mathbf{0.2967}^{B*}$	0.0965
	mix_MC2	-	-	0.1955^{B*}	0.2917^{B*}	$0.0907^{X,X_g}$
	geo_xQuAD	Virtual	0.76	0.1938^{B*}	0.2941^{B*}	$\mathbf{0.1001}^{I,M}$
TREC 2010						
Baseline	LM	-	-	0.1959	0.2842	0.1406
	IA-Select	Virtual	-	0.2106	0.3078^{B}	$0.1195^{X*,C}_{X_g}$
	org_xQuAD	Sum	0.46	0.2164^{B}	0.3147^{B*}	$0.1486^{I*,M}$
	PM2	MinMax	0.46	0.2039	0.3033	$0.1344^{X*,C}_{X_g}$
Proposed	mix_CombSUM	Sum	0.50	$\mathbf{0.2149}^{B}$	0.3124^{B}	$\mathbf{0.1480}^{I,M}$
	mix_MC1	-	-	0.2124	$\mathbf{0.3165}^{B}$	$0.1329^{X,C}_{X_g}$
	geo_xQuAD	Virtual	0.62	0.2146	0.3122^{B}	$0.1464^{I,M}$

Note. The sub/superscripts of a result denote a statistically significant difference from LM (B), IA-Select (I), org_xQuAD (X), PM2 (P), mix_CombSUM (C), mix_MC1 (M), mix_MC2 (M), or geo_xQuAD (X_g) at 0.05 level. The sub/superscripts with a star denote a statistically significant difference at 0.01 level.

3.4.2.4 Impact of the probability mixture model in ranking aggregation.

For all the score and rank aggregation methods considered in this study, we also experimented with the versions that do not take the initial ranked list τ_q into account during the diversification process. Our results reveal that, for almost all cases and evaluation metrics, the versions with the probability mixture model are superior to their counterparts without the model.

3.5 Conclusion

In this chapter, we adapted various score and rank aggregation strategies that are used in meta-search scenarios in the literature to the diversification problem. Our experiments revealed that some of these strategies, despite their simplicity, also serve well for the diversification purposes and outperform three state-of-the-art baselines from the literature. This is an especially important finding given that these ranking aggregation methods can be computed more efficiently than the baseline diversification strategies and our xQuAD variants.

CHAPTER 4

QUERY PERFORMANCE PREDICTION FOR ASPECT WEIGHTING IN DIVERSIFICATION

Explicit search result diversification strategies depend on the availability of potential query aspects and exploit them to diversify the initial retrieval results using a weighted mixture model. Accurate estimation of query aspect weights is an important issue to improve the performance of explicit search result diversification algorithms. In this chapter [1], for the first time in the literature we propose using post-retrieval query performance predictors (QPPs) to estimate the relative weights of the query aspects. In addition to utilizing well-known QPPs from the literature, we also introduce three new QPPs that are based on score distributions.

The rest of the chapter is organized as follows. In Section 4.1 we provide the motivation of our work. In the next section, we describe the QPPs from the literature and introduced our proposed QPPs which will be used to weight query aspects. In Section 4.3, we describe the experimental setup and the results of the proposed weighting methods. The conclusion is provided in Section 4.4.

4.1 Introduction

Explicit diversification methods directly model the query aspects, exploiting manually or automatically assigned query labels in a taxonomy [1], or query reformulations

[1] A. M. Ozdemiray and I. S. Altingovde. "Query Performance Prediction for Aspect Weighting in Search Result Diversification", Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pages 1871-1874, ©2014 ACM, Inc. http://dx.doi.org/10.1145/2661829.2661975. Reprinted by permission with license number 3713830639102.

in a search log [51]. In the latter case, aspects weights that can represent the importance [51], popularity [9] or likelihood [1] of each aspect for a given query is of utmost importance to optimize the quality of the final result.

In this chapter, we put a new perspective on aspect weighting to improve the performance of explicit search result diversification. The weight to be assigned to an aspect in a diversification method should not only depend on the aspects' intrinsic properties (such as those exemplified above), but it should better reflect the expected retrieval effectiveness of the top-ranked results (in the candidate set) that match to this aspect. We explain the underlying intuition as follows. In a typical explicit diversification framework, the relevance score of candidate documents for each explicit aspect is computed (using some retrieval model); and each aspect contributes *its* highest scoring documents to the final query result, which is typically of size 10 or 20. Thus, given an aspect (regardless of how important or likely it is for a given query), if the candidate documents with the highest matching scores to this aspect are indeed irrelevant, such an aspect cannot help improving the final result quality, and may even degrade it.

In this light, we propose leveraging query performance predictors (QPPs) to estimate the retrieval effectiveness of the query aspects over the candidate documents. To this end, we employ post-retrieval QPPs that are based on score distribution analysis, namely, weighted information gain (WIG) [73], normalized query commitment (NQC) [56] and their variants presented in [35]. The choice of these QPPs is intentional, to satisfy the demanding efficiency requirements of online query processing. As mentioned above, the state-of-the-art explicit diversification algorithms [1, 20, 51] compute the relevance of aspects to candidate documents, and hence, the input to these predictors will be created for free, without any additional cost or effort. To the best of our knowledge, no previous work employs QPPs for weighting query aspects in the context of search result diversification.

We also introduce three new predictors that are again based on the score distribution analysis and hence, directly applicable in aspect weighting scenario. The first one is a simple yet effective QPP that is based on the score ratios. The other two predictors are novel in that their performance estimations are based on a virtual document that yields the best possible relevance score for a given query aspect.

We evaluate the existing and proposed QPPs in the context of aspect weighting using the standard TREC Diversity Task framework. Our experiments include a wide range of explicit diversification methods, namely, IA-Select [1], xQuAD [51] (and its variants proposed in Chapter 2), PM2 [20], and a well known score-based aggregation strategy CombSUM, which is adapted to diversification problem in Chapter 3. Our findings show that, performance based weighting of query aspects consistently improves the result quality for these algorithms. Furthermore, the proposed predictors are superior to the existing QPPs when applied in the context of aspect weighting.

4.2 QPPs for Aspect Weighting

Let's assume that a given query q retrieves an initial set of N documents, i.e., so-called the candidate set D_q, over a corpus C. The goal of result diversification is constructing a ranking D_q^k of k documents that maximizes both relevance and diversity. In case of the explicit result diversification, it is assumed that there is a set of explicitly identified query aspects (a.k.a., sub-topics, interpretations, sub-queries) denoted as $T = \{q_1, ..., q_m\}$ associated with the original query q. These aspects are usually obtained from external resources, such as a taxonomy or query log.

In most explicit diversification methods (as discussed in the next section), there is an aspect weight component, which may represent the likelihood, popularity or importance of a given aspect q_i for the query q. This aspect weight can be assigned in various ways. For instance, Agrawal et al. employ a classifier trained on the ODP taxonomy to associate categories (as aspects) to the queries along with the class likelihood scores (as weights) [1]. Santos et al. apply three different methods to compute aspect weights, the simplest being the uniform probability assigned as a weight to each aspect [51]. They also suggest weighting methods based on the number of results retrieved by the query aspects from an external collection (e.g., using a search engine) and the local corpora C (in a similar manner to resource selection methods employed in distributed retrieval systems). In their work, the simple uniform estimator is reported to yield the best performing aspect weights, and hence, it is also adopted in the succeeding works by others (like [20]).

In this thesis, we propose a novel perspective for aspect weighting that is different from all the aforementioned approaches. Our proposal is based on the observation that the most successful explicit diversification methods (such as [20, 51]) compute and exploit the relevance $rel(d, q_i)$ of each candidate document $d \in D_q$ to each aspect q_i during the diversification process. Furthermore, since the ultimate goal is coming up with a final ranking D_q^k and there may be several aspects of a query, only the highest scoring documents for an aspect can have a chance to be selected into this final ranking. Subsequently, an aspect q_i can improve the quality of the final result only if its top-p documents over the candidate set, $D_{q_i}^p$, is highly relevant to q_i. This suggests that the effectiveness of $D_{q_i}^p$ for the aspect q_i is a natural indicator of the weight that should be assigned to q_i during diversification. Hence, in this thesis, we propose using QPPs to assign weights to query aspects in result diversification algorithms.

Since the rankings $D_{q_i}^p$ per aspect are typically computed by the state-of-the-art diversification methods, it is a natural choice to employ post-retrieval QPPs that rely on the score distribution analysis for aspect weighting task. By doing so, we avoid additional costs that may be incurred by the predictors and can satisfy the demanding requirements of online query processing in large-scale search engines. In what follows, we describe these baseline QPPs (in addition to simple uniform estimator) adopted for query aspect weighting. Next, in Section 2.2, we introduce our own QPPs that are again based on score distributions.

4.2.1 Baseline QPPs for Aspect Weighting

Uniform predictor. This is the straightforward approach employed in several earlier works [20, 51]. For a query with the set of aspects $T = \{q_1, ..., q_m\}$, the aspect weights are computed as $W(q_i) = 1/m$.

Weighted Information Gain (WIG). This predictor is originally proposed to capture the divergence between the mean retrieval score of top ranked documents and that of the entire corpus [73]. To compute WIG, we use Eq. 4.1 presented in [11]. Note that,

$rel(C, q_i)$ represents the relevance score of the corpus C to the aspect q_i, and it further helps to make different aspect weights comparable, i.e., serves as a normalization factor.

$$W(q_i) = \frac{1}{p\sqrt{|q_i|}} \left(avg_{d \in D^p_{q_i}} (rel(d, q_i)) - rel(C, q_i) \right) \tag{4.1}$$

Normalized Query Commitment (NQC). Shtok et al. propose that the mean retrieval score for the top-ranked results of a query represents the score of a possible misleader (as the result list would include some irrelevant documents besides the relevant ones) [56]. Therefore, NQC computes the standard deviation of the relevance scores over the list $D^p_{q_i}$ and again normalizes the result value by the relevance score of the corpus (Eq. 4.2).

$$W(q_i) = \frac{\sqrt{\frac{1}{p} \sum_{d \in D^p_{q_i}} (rel(d, q_i) - avg_{d \in D^p_{q_i}} (rel(d, q_i))^2}}{|rel(C, q_i)|} \tag{4.2}$$

ScoreAvg. Markovits et al. employ a simpler variant of WIG in a data fusion setting [35]. In this variant, called here ScoreAvg, instead of using $rel(C, q_i)$ for normalization as in WIG, the relevance scores $rel(d, q_i)$ are sum normalized to [0, 1] before computing their average.

ScoreDev. This method [35] is a variant of NQC, and applies Eq. 4.2 without the normalization factor $rel(C, q_i)$. Note that, there are other works [42, 19, 56] that again make use of the standard deviation of the document scores in various ways, and not considered here for the sake of space.

4.2.2 Proposed QPPs for Aspect Weighting

ScoreRatio. This predictor is motivated by the intuition that as the gap between the scores of the documents in a ranking widens, the likelihood of seeing irrelevant documents also increases. Thus, the ScoreRatio predictor computes the ratio of the scores of the first and last documents in $D^p_{q_i}$.

VScoreAvg. In Chapter /refchapter:xquad, we have shown that explicit diversification algorithms are quite sensitive to techniques that are employed for normalizing the relevance scores between documents and query aspects. Furthermore, we have proposed an effective score normalization technique, so-called Virtual, which we adapt here for the purposes of query performance prediction.

Our virtual-score based predictors differ from the previously described QPPs in the following way. Instead of considering the score of the entire corpus (as a huge single document) for normalization (as in WIG or NQC), we consider a virtual document that can yield the highest possible relevance score for a query aspect q_i on a given corpus. More specifically, for a given aspect q_i, we assume a virtual document d^V that includes each term in the aspect with the frequency of the document length and no other terms, i.e., as if the document is only composed of the query terms. The length of the virtual document is set to the average document length in the corpus. Then, we compute the relevance score of this virtual document d^V to q_i as an upper-bound value, i.e., the score of an imaginary perfect match for this aspect. Assume that for a given q_i, the virtual(-normalized) scores for each d in $D^p_{q_i}$ are defined as follows:

$$rel_{Virtual}(d, q_i) = \frac{rel(d, q_i)}{rel(d^V, q_i)} \tag{4.3}$$

Then, VScoreAvg predictor computes the weight of an aspect q_i as shown in Eq. 4.4

$$W(q_i) = \frac{1}{k} \sum_{d \in D^k_q} rel_{Virtual}(d, q_i) \tag{4.4}$$

VScoreFirst. Inspired from the earlier approaches that use highest retrieval score as an indicator of the query performance [57], for each aspect q_i, we use the virtual score of the top-ranked document in $D^p_{q_i}$.

4.3 Experimental Evaluation

We used the same dataset, query topics and initial retrieval models as in Chapter 2.

4.3.1 Explicit diversification methods

In this study, we employ various explicit diversification methods that can be broadly categorized as greedy approaches and aggregation-based approaches. While outlining these methods we conform to their original descriptions that are typically based on a probabilistic mixture model, where $P(d|q)$ ($P(d|q_i)$) represents the likelihood of a document for a given query (aspect), respectively; and $P(q_i|q)$ corresponds to the aspect weight. In our experiments, for the former probability, we employ $rel(d, q)$ and $rel(d, q_i)$ scores that are computed by BM25 retrieval model, after normalizing them with one of the techniques discussed later in this section. For the latter probability, aspect weight, we use the baseline and proposed QPP strategies described in the previous section. While doing so, the weights computed for the aspects of a query are sum normalized to [0, 1] so that they can replace $P(q_i|q)$ in the explicit diversification methods described in Section 2.4.3 and CombSUM method ([39, 41]) described in Section 3.3.1.1.

In our experiments, for all the diversification strategies that employ the trade-off parameter λ, we test all values in $[0, 1]$ range with a step size of 0.01, and report the test results for the λ values that maximize the α-nDCG@20 scores. We also employ three normalization techniques described in Section 2.3.1.2, namely MinMax, Sum and Virtual, to normalize the relevance scores generated by BM25, so that these scores can replace the corresponding probabilities in the diversification methods. Our results are reported for all three techniques, as diversification algorithms are shown to be sensitive to the applied normalization in previous chapters.

4.3.2 Experimental Results

We evaluate the baseline and proposed QPPs by incorporating the predicted aspect weights into each of the seven diversification algorithms. Note that, for all QPPs, we set the parameter p as 10, i.e., we obtain top-10 documents (out of a candidate set of 100 documents) for each aspect and provide their scores to the performance predictors. Since every query in TREC topic set has more than one aspect and the final ranking has size 20, we believe setting p as 10 would be adequate (as will be

Table 4.1: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the baseline QPPs for the query aspects obtained from the official sub-topics. The highest score is boldfaced.

Div. method	Relevance norm.	Uniform	Baseline QPPs			
			WIG	NQC	ScrAvg	ScrDev
IA-Select	MinMax	0.3386	0.3291	0.3430	0.3452	**0.3543**
	Sum	0.3568	**0.3490**	0.3350	0.3447	0.3488
	Virtual	0.3660	**0.3568**	0.3464	0.3555	0.3485
xQuAD	MinMax	0.3386	0.3308	0.3430	0.3452	**0.3543**
	Sum	0.3664	**0.3681**	0.3440	0.3546	0.3586
	Virtual	0.3660	**0.3568**	0.3464	0.3555	0.3485
art_xQuAD	MinMax	0.3751	0.3755	0.3717	0.3671	**0.3810**
	Sum	0.3612	**0.3622**	0.3417	0.3519	0.3525
	Virtual	**0.3892**	0.3802	0.3670	0.3753	**0.3808**
geo_xQuAD	MinMax	0.3581	0.3602	0.3523	0.3539	**0.3646**
	Sum	0.3612	**0.3622**	0.3417	0.3519	0.3525
	Virtual	**0.3890**	**0.3796**	0.3671	0.3746	0.3802
PM2	MinMax	0.3705	0.3707	0.3645	0.3664	**0.3754**
	Sum	0.3669	**0.3657**	0.3524	0.3578	0.3591
	Virtual	0.3756	0.3666	0.3593	0.3588	**0.3726**
mix_CombSUM	MinMax	0.3662	0.3674	0.3658	0.3635	**0.3813**
	Sum	0.3613	**0.3610**	0.3410	0.3516	0.3531
	Virtual	**0.3811**	0.3747	0.3634	0.3702	**0.3806**

justified by the results). We report α-NDCG@20 scores computed with `ndeval` software.

From our results shown in Table 4.1 and 4.2, we draw the following conclusions: (i) We see that using QPPs for aspect weighting improves almost all the diversification methods (15 out of 18 cases) in comparison to assigning uniform weights to each aspect. The absolute improvements in α-NDCG scores reach up to 2%, whereas the relative improvements are up to 6% (e.g., for the xQuAD method with `Sum` normalization). (ii) Considering the baseline predictors, WIG and ScoreDev are the most effective ones. Among the proposed QPPs, the ScoreRatio predictor outperforms the other two. (iii) Comparing baseline predictors to the proposed ones, we observe that the latter are more effective as they (especially the ScoreRatio predictor) yield the

Table 4.2: Diversification performance (α-NDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the best baseline QPPs and proposed QPPs for the query aspects obtained from the official sub-topics. The highest score in each group is bold, the overall winner is underlined.

Div. method	Relevance norm.	Uniform	**Baseline QPPs**		**Proposed QPPs**		
			WIG	ScrDev	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.3386	0.3291	**0.3543**	0.3400	0.3385	**0.3442**
	Sum	0.3568	**0.3490**	0.3488	0.3574	0.3565	**0.3682**
	Virtual	0.3660	**0.3568**	0.3485	0.3633	0.3632	**0.3636**
xQuAD	MinMax	0.3386	0.3308	**0.3543**	0.3400	0.3385	**0.3442**
	Sum	0.3664	**0.3681**	0.3586	0.3655	0.3711	**0.3804**
	Virtual	0.3660	**0.3568**	0.3485	0.3633	0.3632	**0.3636**
art_xQuAD	MinMax	0.3751	0.3755	**0.3810**	**0.3818**	0.3777	0.3716
	Sum	0.3612	**0.3622**	0.3525	0.3579	0.3645	**0.3758**
	Virtual	**0.3892**	0.3802	**0.3808**	0.3858	**0.3878**	0.3805
geo_xQuAD	MinMax	0.3581	0.3602	**0.3646**	**0.3637**	0.3602	0.3594
	Sum	0.3612	**0.3622**	0.3525	0.3579	0.3645	**0.3758**
	Virtual	**0.3890**	**0.3796**	0.3802	0.3863	**0.3886**	0.3798
PM2	MinMax	0.3705	0.3707	**0.3754**	0.3728	**0.3691**	0.3664
	Sum	0.3669	**0.3657**	0.3591	0.3722	0.3732	**0.3755**
	Virtual	0.3756	0.3666	**0.3726**	0.3776	0.3732	**0.3778**
mix_CombSUM	MinMax	0.3662	0.3674	**0.3813**	**0.3758**	0.3728	0.3632
	Sum	0.3613	**0.3610**	0.3531	0.3580	0.3621	**0.3720**
	Virtual	**0.3811**	0.3747	**0.3806**	0.3761	**0.3788**	0.3742

highest α-nDCG scores in 9 cases (covering all algorithms and most normalization techniques), whereas uniform estimator is the most effective estimator in 3 cases and baseline estimators yield the best effectiveness for a total of 6 cases. Note that, VScoreFirst (VScoreAvg) predictor outperforms all baseline QPPs in 9 (10) out of 18 cases, respectively.

4.4 Conclusion

For the first time in the literature, we used post-retrieval QPPs in the context of aspect weighting in explicit search result diversification. To this end, we introduced three new QPPs that are based on score distributions, as well as using several others from the literature. Through extensive experiments, we showed that predicting the retrieval effectiveness of each individual aspect on the candidate document set is a good indicator of an aspect's contribution to the quality of the final result.

CHAPTER 5

ASPECT EXPANSION FOR EXPLICIT SEARCH RESULT DIVERSIFICATION

In this chapter, we propose to use query expansion techniques to represent the query aspects better, and hence improve the overall effectiveness of explicit search result diversification methods. As our first contribution, we exploit the results of each query aspect itself for the purposes of expansion, and show that this is better than expanding the aspects based on the results of the main query, as well as expanding the main query itself. Secondly, we propose a novel selective approach that only expands certain query aspects based on their retrieval performance, which is obtained using post-retrieval query performance predictors (QPPs). Our experiments reveal that selective expansion of aspects is better than expanding all the aspects blindly.

The rest of the chapter is organized as follows. In the next section, we provide the motivation for proposing aspect expansion in the context of explicit result diversification. In Section 5.2, we review earlier works that essentially focus on employing query expansion techniques only for the main query. In Section 5.3, we first define a general aspect expansion strategy based on the retrieval results of each aspect when executed on the collection, and then propose a selective approach that only expands certain aspects. Section 5.4 devoted to experimental results. Finally, we conclude and point future work directions in Section 5.5.

5.1 Introduction

Earlier works on explicit diversification methods rely on the assumption that aspects of a given query can be obtained a priori from various resources, and aim to exploit these aspects to the greatest extent to obtain the highest diversification effectiveness. In this setup, the query aspects are typically obtained from some external resources, like some taxonomies (such as ODP), Wikipedia, or query logs ([51, 9]), and once they are obtained, they are fed to diversification strategies without any further processing. However, in many cases, it is possible that the aspect terms extracted from such external corpora do not include all the terms to represent the aspect ideally for the collection on which the diversification will take place; and such aspects may not be as useful as they could be for the diversification algorithms, or may even mislead the algorithm.

In this chapter, inspired by the success of query expansion and re-writing techniques applied in ad hoc retrieval ([13]), we propose to expand the query aspects based on the documents they retrieve on the target collection. In particular, we apply typical pseudo-relevance feedback (PRF) methods on the top-k results retrieved for each query aspect. To the best of our knowledge, all the previous work in the literature either use a given set of aspects, or aim to expand the main query itself (in a way that will introduce diverse terms to the query). In contrast, for the first time in the literature, we expand the query aspects using the feedback from the target collection.

We believe that our proposal fits well to practical retrieval systems, and in particular, search engines, due to their very large result caches. More specifically, assuming the aspects extracted from various resources (like query logs and Wikipedia) for a given query, it is very likely that such aspects (or, at least, most of them) have been submitted to the search engine as independent queries, and hence, their results would be available in the result caches. For instance, assume the infamous "jaguar" query. The search engine, once having discovered the query's mot probable intents (say, "jaguar car pictures", "jaguar branches", "jaguar animal", etc.) from its logs and other external resources, would most likely find the result of these query aspects in its result cache, which can be large enough to store several millions of queries and their results in these days ([3]). This means that aspect expansion in practical settings may not require executing each aspect as a separate query on the collection, and its overhead for the system would only be running an expansion algorithm on the results, which can be done even offline.

As our second contribution in this chapter, we introduce a novel selective strategy that expands only those aspects that are likely to benefit from the expansion. This idea is inspired by our findings in the previous chapter, where we have shown that an aspect's weight should be proportional to its predicted retrieval performance on the candidate result set of the main query. In this case, if an aspect's retrieval performance suffers over the candidate documents, then it is more likely that expanding this aspect using its own retrieval results (as described before) will be useful. Therefore, we again leverage query performance predictors to estimate the retrieval effectiveness of the query aspects over the candidate documents, and then selectively expand certain aspects based on their estimated performance.

In our experiments, we use the standard TREC Diversity Task setup (as described in the previous chapters) and several baselines, namely, expanding only the main query and expanding the main query and diverse aspects (to serve as an upperbound for the approaches discussed in [7]), as well as a naïve no-expansion baseline. Our findings reveal that aspect expansion usually improves the diversification performance of allmost all state-of-the-art explicit diversification methods. Moreover, selectively expanding particular aspects of a query yields higher diversification performance than that of blindly expanding all the aspects of the query.

5.2 Related Work

Automatic query expansion is used to improve the precision of the search results by embedding new terms to usually short user queries. The expansion terms can be selected based on either the original query terms (i.e. term-based) or the top-retrieved documents of the initial search results (i.e. result-based) [13]. In term-based approaches, the expansion terms are selected by linguistic techniques like using stemmers or external sources like thesaurus, ConceptNet, WordNet or Wikipedia [28, 29, 32]; or by analyzing the co-occurrence of the terms in the corpus [13]. In result-based approaches, either the terms in the top-retrieved documents are analyzed to find new expansion terms [49, 12], or a statistical language model is built using the top-retrieved documents to assign probabilities to expansion terms [30, 68].

Vargas et. al. ([61]), proposed to use query expansion in Search Result Diversification framework, by selecting expansion terms for original query to improve the diversity of the results. Similar to explicit search result diversification methods, they assumed the explicit knowledge of the aspects of a query and used assessed documents of each aspect as the feedback documents to find the candidate expansion terms for each aspect. After finding the candidate terms for each aspect, they select the expansion terms by using a procedure inspired from the xQuAD algorithm ([51]).

In [6], Bouchoucha et. al. utilize ConceptNet to find candidate terms for the given query and calculate the similarity between the terms to be used in MMRE, MMR-based Expansion algorithm [10], to select the most diverse expansion terms that cover multiple aspects implicitly. In a following work [7], they integrate multiple resources, namely ConceptNet, Wikipedia, query logs and PRF to diversify the search results. In the first phase, they find expanded queries for each resource using a generalized version of MMRE algorithm. In the second phase, they retrieve the top documents from the collection for each expanded query to construct the candidate document set. Finally they iteratively select the final result set by applying MMR principle.

5.3 Selectively Expanding Aspect Queries

In explicit search result diversification the aspects q_i of a query q are assumed to be known during the query execution. After q is executed on the collection C, and $top-k$ documents D_q^k are retrieved, diversification methods calculate the relevance $rel(d, q_i)$ of each document in the candidate set, $d \in D_q^k$, and generate are ranking lists, $D_{q_i}^k$ (i.e. re-rankings) for each aspect. In the final phase, the ranking lists are aggregated using the diversification method.

5.3.1 Aspect Expansion

Although, the query aspects are explicitly known, the terms defining the aspect may be inadequate to represent the aspect successfully. In order to prevent from possible shortcomings, we propose a novel approach to search result diversification by expanding aspect queries. In particular, we use pseudo-relevance feedback from top retrieved documents of a document set, and use the following simplified version of Rocchio's formula [49] defined in [13]:

$$w'_{t,q'} = (1 - \lambda) \cdot w_{t,q} + \lambda \cdot score_t \tag{5.1}$$

to find new expansion terms for the aspect query, where $w_{t,q}$ is the weight of the term t for original query q and $w'_{t,q'}$ is the weight of t for expanded query w' and λ is the weighting factor between original terms and the expansion terms. Although Equation 5.1 can be used to reweight the terms in the query ([12, 13]), we just use the equation to pick the expansion terms from the top-scoring terms with respect to $w'_{t,q'}$, and leave the rest to the retrival model.

5.3.1.1 Term Scoring Functions

In Rocchio's original formula ([49]), the term score is the sum of term weights in the top-retrieved documents. Instead of a simple proportion of term frequencies, in Equation 5.2 we use Okapi BM25 ([48]) term scores as term weights to reuse already calculated scores during re-ranking of candidate documents for each query aspect.

$$score(t) = \sum_{d \in R} \frac{(k_1 + 1) \cdot f_{d,t}}{k_1 \cdot [(1 - b) + b.\frac{d_{len}}{avr\_d_{len}}] + f_{d,t}} \tag{5.2}$$

In [12], in addition to Rocchio's original term scoring function, Carpineto et. al. used different term scoring functions to find expansion terms and showed that Kullback-Leibler distance (i.e. KLD) based following function generate better expansion terms:

$$score(t) = [p_R(t)] \cdot log[p_R(t)/p_C(t)] \tag{5.3}$$

where $p_X(t)$ is the ratio of occurrence of term t in document set X where $X = R$ for result set, and $X = C$ for the whole collection.

5.3.2 Selecting Aspects to Expand

Since diversification methods use mixture models (using relevance or ranking of the documents), the diversification performance depend on aspects' individual performance. The expansion of all aspects may give harm to the overall effectiveness of the diversification. On the one hand, expansion of an aspect may decrease the retrieval performance of that aspect, which may affect the recall of that aspect in the final list. On the other hand, expansion of an aspect may boost the performance of that aspect so that the documents relevant to that aspect may dominate the final list.

Therefore, we propose to select the aspects to be expanded using two of our post-retrieval QPP methods ([40]) described in Section 4.2.2, namely VScoreFirst and VScoreAvg. In particular, after an explicit diversification algorithm generate the re-ranking of the candidate set D_{qi} for an aspect q_i, we use the top-k documents to predict the sub-query performance and decide to expand the query if the prediction score is below some threshold.

5.4 Experiments and Results

We used the same dataset, query topics and initial retrieval models as in Chapter 2.

5.4.1 Experimental Setup

5.4.1.1 Explicit diversification methods

In this study, we employ various explicit diversification methods that can be broadly categorized as greedy approaches and aggregation-based approaches. While outlining these methods we conform to their original descriptions for which that are typically based on a probabilistic mixture model, where $P(d|q)$ ($P(d|q_i)$) represents the likelihood of a document for a given query (aspect), respectively; and $P(q_i|q)$ corresponds to the aspect weight. In our experiments, for the former probability, we employ $rel(d, q)$ and $rel(d, q_i)$ scores that are computed by BM25 retrieval model, after normalizing them with one of the techniques discussed later in this section. For the latter probability, aspect weight, we use the baseline and proposed QPP strategies described in the previous chapter. While doing so, the weights computed for the aspects of a query are sum normalized to [0, 1] so that they can replace $P(q_i|q)$ in the explicit diversification methods described in Section 2.4.3 and CombSUM method described in Section 3.3.1.1.

In our experiments, for all the diversification strategies that employ the trade-off parameter λ, we test all values in $[0, 1]$ range with a step size of 0.01, and report the test results for the λ values that maximize the α-nDCG@20 scores. We also employ three normalization techniques described in Section 2.3.1.2, namely MinMax, Sum and Virtual, to normalize the relevance scores generated by BM25, so that these scores can replace the corresponding probabilities in the diversification methods. Our results are reported for all three techniques, as diversification algorithms are shown to be sensitive to the applied normalization in previous chapters.

5.4.1.2 Sub-topic Query Expansion

The employed query expansion methods generate a ranking for the terms used in the documents in a document set. We both used sub-topic's own ranking, S_{qi}, and top-m documents from the re-ranking of D_q according to sub-topic q_i as the document sets. In the experiments, we either add 5 expansion terms to a sub-topic, or we fix the number of terms of a sub-topic to 10.

5.4.1.3 Selective Sub-topic Query Expansion

In order to select the aspects that needs expansion, we used VScoreAvg and VScore-First QPPs. Empirically we set the threshold to 0.7 for official sub-topics and expand the sub-topics whose QPP score is below the threshold, based on the observation in Chapter 4 that if the performance of the sub-topic is not good enough then it may not have the relevant documents to improve the diversification performance of the final result set. On the other hand, we set the threshold to 0.6 and expand the sub-topics that perform better than others for aspects generated from query suggestions. This controversy stems from the observation that, most of the aspects in query suggestions does have relevant documents and therefore do not contribute to the final result. In that sense, we try to improve the quality of the results of the promising aspects instead of dealing with the aspects that perform badly.

5.4.2 Evaluation Results

5.4.2.1 Expansion of official sub-topics using candidate re-rankings

We firstly used the candidate re-rankings for the pseudo-relevance feedback documents to find the expansion terms. In Table 5.1 we see that in TREC 2009 diversification task, expanding the sub-topics using candidate documents did not provide better diversification results than original sub-topics. However, as seen in Table 5.2, expansion with candidate documents actually improve diversification performance of some methods for TREC 2010 topics.

Table 5.1: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	**0.3240**	0.3019	0.2782	0.2850	0.2795
	Sum	**0.3162**	0.2914	0.2738	0.2872	0.2979
	Virtual	**0.3263**	0.3016	0.2800	0.2883	0.3009
xQuAD	MinMax	**0.3298**	0.3086	0.2964	0.3058	0.3095
	Sum	**0.3238**	0.3181	0.3024	0.3133	0.3208
	Virtual	**0.3268**	0.3108	0.2966	0.3137	0.3096
art_xQuAD	MinMax	**0.3387**	0.3222	0.3155	0.3216	0.3256
	Sum	**0.3238**	0.3176	0.3006	0.3124	0.3215
	Virtual	**0.3354**	0.3170	0.3037	0.3200	0.3213
geo_xQuAD	MinMax	**0.3420**	0.3184	0.3093	0.3195	0.3207
	Sum	**0.3238**	0.3176	0.3006	0.3124	0.3215
	Virtual	**0.3341**	0.3152	0.3048	0.3201	0.3204
PM2	MinMax	**0.3334**	0.3075	0.2865	0.2942	0.2960
	Sum	**0.3310**	0.3007	0.2827	0.2917	0.2996
	Virtual	**0.3360**	0.3056	0.2859	0.2963	0.2953
mix_CombSUM	MinMax	**0.3323**	0.3183	0.3133	0.3125	0.3149
	Sum	**0.3235**	0.3176	0.3004	0.3128	0.3213
	Virtual	**0.3339**	0.3212	0.3027	0.3147	0.3181
mix_Borda		**0.3273**	0.3094	0.3074	0.3055	0.3086
mix_SV		**0.3094**	0.3026	0.3088	0.2848	0.2967
mix_MC2		**0.3307**	0.3276	0.3115	0.3181	0.3106

Table 5.2: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.3386	0.3576	**0.3612**	0.3349	0.3426
	Sum	0.3568	**0.3618**	0.3516	0.3484	0.3447
	Virt	0.3660	**0.3773**	0.3715	0.3627	0.3578
xQuAD	MinMax	0.3386	0.3576	**0.3612**	0.3360	0.3426
	Sum	0.3664	0.3683	0.3700	0.3520	0.3472
	Virtual	0.3660	**0.3773**	0.3715	0.3627	0.3583
art_xQuAD	MinMax	0.3751	**0.3850**	0.3844	0.3645	0.3563
	Sum	0.3612	0.3662	**0.3672**	0.3506	0.3468
	Virtual	**0.3892**	0.3769	0.3755	0.3574	0.3558
geo_xQuAD	MinMax	0.3581	**0.3735**	0.3706	0.3524	0.3487
	Sum	0.3612	0.3662	**0.3672**	0.3506	0.3468
	Virtual	**0.3890**	0.3765	0.3755	0.3584	0.3551
PM2	MinMax	0.3705	**0.3751**	0.3673	0.3507	0.3519
	Sum	0.3669	**0.3740**	0.3683	0.3548	0.3538
	Virtual	0.3756	**0.3767**	0.3659	0.3578	0.3526
mix_CombSUM	MinMax	**0.3662**	0.3539	0.3599	0.3477	0.3555
	Sum	0.3613	0.3653	**0.3669**	0.3500	0.3468
	Virtual	**0.3811**	0.3616	0.3619	0.3568	0.3536
mix_Borda		0.3542	0.3661	**0.3679**	0.3620	0.3561
mix_SV		0.3381	**0.3506**	0.3486	0.3434	0.3504
mix_MC2		0.3645	**0.3684**	0.3604	0.3418	0.3523

5.4.2.2 Expansion of official sub-topics using own ranking

Then we used sub-topic's own rankings to expand the query. As Table 5.3 and Table 5.4 shows, expanding all sub-topics improve the diversification result of some methods in 2009 and most of the methods in 2010. Please note that using BM25 term-ranking function provide better results than using KLD.

Table 5.3: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is bold-faced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	**0.3240**	0.3151	0.3078	0.2966	0.3097
	Sum	**0.3162**	0.3082	0.2881	0.2985	0.3036
	Virtual	**0.3263**	0.3238	0.3126	0.2967	0.3109
xQuAD	MinMax	0.3298	**0.3302**	0.3216	0.3167	0.3195
	Sum	0.3238	0.3217	0.3149	0.3194	**0.3248**
	Virtual	0.3268	**0.3273**	0.3207	0.3151	0.3198
art_xQuAD	MinMax	0.3387	**0.3442**	0.3347	0.3318	0.3346
	Sum	**0.3238**	0.3212	0.3143	0.3181	0.3234
	Virtual	**0.3354**	0.3345	0.3295	0.3259	0.3308
geo_xQuAD	MinMax	0.3420	**0.3469**	0.3353	0.3287	0.3311
	Sum	**0.3238**	0.3212	0.3143	0.3181	0.3234
	Virtual	0.3341	**0.3345**	0.3292	0.3262	0.3313
PM2	MinMax	0.3334	**0.3367**	0.3195	0.3138	0.3161
	Sum	**0.3310**	0.3273	0.3146	0.3130	0.3249
	Virtual	**0.3360**	0.3286	0.3139	0.3151	0.3180
mix_CombSUM	MinMax	0.3323	**0.3377**	0.3263	0.3189	0.3285
	Sum	**0.3235**	0.3200	0.3136	0.3180	0.3225
	Virtual	**0.3339**	0.3269	0.3236	0.3220	0.3298
mix_Borda		**0.3273**	0.3119	0.3127	0.3144	0.3191
mix_SV		0.3094	0.3218	**0.3260**	0.2949	0.3160
mix_MC2		0.3307	**0.3318**	0.3223	0.3216	0.3269

Table 5.4: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.3386	**0.3636**	0.3616	0.3411	0.3448
	Sum	0.3568	**0.3823**	0.3749	0.3486	0.3416
	Virtual	0.3660	0.3810	**0.3836**	0.3549	0.3542
xQuAD	MinMax	0.3386	**0.3636**	0.3616	0.3411	0.3448
	Sum	0.3664	**0.3941**	0.3858	0.3618	0.3529
	Virtual	0.3660	0.3810	**0.3836**	0.3610	0.3542
art_xQuAD	MinMax	0.3751	**0.3900**	0.3866	0.3741	0.3719
	Sum	0.3612	**0.3898**	0.3813	0.3564	0.3500
	Virtual	0.3892	0.3898	**0.3987**	0.3658	0.3661
geo_xQuAD	MinMax	0.3581	0.3789	**0.3802**	0.3654	0.3677
	Sum	0.3612	**0.3898**	0.3813	0.3564	0.3500
	Virtual	0.3890	0.3921	**0.3991**	0.3674	0.3671
PM2	MinMax	0.3705	**0.3833**	0.3794	0.3592	0.3667
	Sum	0.3669	**0.3919**	0.3877	0.3633	0.3743
	Virtual	0.3756	0.3853	**0.3864**	0.3561	0.3640
mix_CombSUM	MinMax	**0.3662**	0.3629	0.3652	0.3625	0.3602
	Sum	0.3613	**0.3886**	0.3809	0.3565	0.3500
	Virtual	**0.3811**	0.3725	0.3786	0.3490	0.3517
mix_Borda		0.3542	0.3691	**0.3728**	0.3577	0.3641
mix_SV		0.3381	0.3731	**0.3799**	0.3590	0.3626
mix_MC2		0.3645	0.3733	**0.3845**	0.3708	0.3773

5.4.2.3 Expansion of suggested topics using own ranking

We also applied the same expansion methodology to suggested sub-topics. In Table 5.5 and Table 5.6, it is shown that sub-topic expansion improves suggested sub-topics also.

Table 5.5: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25		KLD	
			Add 5	Fix 10	Add 5	Fix 10
IA-Select	MinMax	0.2814	**0.3089**	0.2997	0.2530	0.2756
	Sum	0.2688	**0.2883**	0.2793	0.2635	0.2786
	Virtual	0.2675	**0.3020**	0.2842	0.2658	0.2800
xQuAD	MinMax	0.2960	**0.3102**	0.2997	0.2975	0.2897
	Sum	0.2846	**0.3116**	0.3056	0.2936	0.2945
	Virtual	0.2887	**0.3231**	0.3196	0.3027	0.3115
art_xQuAD	MinMax	0.3049	0.3071	**0.3111**	0.3040	0.3000
	Sum	0.2860	**0.3090**	0.3061	0.2909	0.2938
	Virtual	0.2948	0.3162	**0.3177**	0.3081	0.3015
geo_xQuAD	MinMax	0.3019	**0.3131**	0.3125	0.3097	0.3037
	Sum	0.2860	**0.3090**	0.3061	0.2909	0.2938
	Virtual	0.2957	0.3160	**0.3188**	0.3090	0.3014
PM2	MinMax	0.2935	0.2928	**0.2972**	0.2728	0.2857
	Sum	0.2775	**0.3007**	0.2876	0.2716	0.2801
	Virtual	0.2889	**0.2994**	0.2965	0.2777	0.2868
mix_CombSUM	MinMax	**0.3043**	0.2958	0.2960	0.2930	0.2956
	Sum	0.2860	**0.3092**	0.3063	0.2912	0.2934
	Virtual	0.2959	**0.3079**	0.3020	0.2998	0.2977
mix_Borda		0.2894	**0.2936**	0.2894	0.2859	0.2908
mix_SV		0.2757	**0.2983**	0.2757	0.2757	0.2891
mix_MC2		0.2858	**0.2919**	0.2879	0.2818	0.2864

Table 5.6: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.2952	**0.3099**	0.3056	0.2847	0.3074
	Sum	0.3043	0.3154	**0.3157**	0.2907	0.2857
	Virtual	0.3046	**0.3223**	0.3211	0.2872	0.3114
xQuAD	MinMax	0.3072	0.3240	**0.3319**	0.3090	0.3166
	Sum	0.3215	**0.3245**	0.3226	0.3104	0.3125
	Virtual	0.3090	0.3392	**0.3397**	0.3120	0.3247
art_xQuAD	MinMax	0.3225	0.3465	**0.3484**	0.3205	0.3282
	Sum	0.3225	**0.3243**	0.3209	0.3051	0.3078
	Virtual	0.3210	**0.3386**	0.3354	0.3125	0.3285
geo_xQuAD	MinMax	0.3228	0.3368	**0.3466**	0.3192	0.3249
	Sum	0.3225	**0.3243**	0.3209	0.3051	0.3078
	Virtual	0.3194	**0.3387**	0.3369	0.3128	0.3288
PM2	MinMax	0.3129	**0.3333**	0.3312	0.3078	0.3136
	Sum	0.3129	0.3226	**0.3380**	0.2998	0.3109
	Virtual	0.3107	0.3277	**0.3400**	0.3045	0.3160
mix_CombSUM	MinMax	0.3145	0.3291	**0.3295**	0.3135	0.3215
	Sum	0.3224	**0.3245**	0.3200	0.3049	0.3076
	Virtual	0.3256	**0.3273**	0.3229	0.3122	0.3257
mix_Borda		0.3190	**0.3425**	0.3408	0.3360	0.3309
mix_SV		0.3179	**0.3464**	0.3546	0.3347	0.3349
mix_MC2		0.3081	0.3328	**0.3441**	0.3141	0.3244

5.4.2.4 Selective expansion of official sub-topics using own rankings

Since adding 5 terms to the query using BM25 term weights generate the better results in previous experiments, we applied selective expansion using that setup. We can see from Table 5.7 and Table 5.8 selectively expanding sub-topics improved the diversification performance compared to expanding all sub-topics blindly.

Table 5.7: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by**	
				VScrAvg	VScrFirst
IA-Select	MinMax	**0.3240**	0.3151	0.2970	0.3097
	Sum	**0.3162**	0.3082	0.3064	0.3050
	Virtual	0.3263	0.3238	0.3127	**0.3295**
xQuAD	MinMax	0.3298	**0.3302**	0.3142	0.3251
	Sum	**0.3238**	0.3217	0.3208	0.3229
	Virtual	0.3268	0.3273	0.3164	**0.3295**
art_xQuAD	MinMax	0.3387	**0.3442**	0.3392	0.3421
	Sum	**0.3238**	0.3212	0.3196	0.3220
	Virtual	0.3354	0.3345	0.3316	**0.3415**
geo_xQuAD	MinMax	0.3420	**0.3469**	0.3324	0.3405
	Sum	**0.3238**	0.3212	0.3196	0.3220
	Virtual	0.3341	0.3345	0.3308	**0.3407**
PM2	MinMax	0.3334	0.3367	0.3280	**0.3432**
	Sum	0.3310	0.3273	0.3167	**0.3311**
	Virtual	0.3360	0.3286	0.3281	**0.3384**
mix_CombSUM	MinMax	0.3323	0.3377	0.3425	**0.3441**
	Sum	**0.3235**	0.3200	0.3193	0.3218
	Virtual	0.3339	0.3269	0.3279	**0.3389**
mix_Borda		**0.3273**	0.3119	0.3182	0.3205
mix_SV		0.3094	**0.3218**	0.3159	0.3182
mix_MC2		0.3307	0.3318	0.3308	**0.3388**

Table 5.8: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by** VScrAvg	**Selective by** VScrFirst
IA-Select	MinMax	0.3386	0.3636	**0.3650**	0.3388
	Sum	0.3568	**0.3823**	0.3721	0.3702
	Virtual	0.3660	**0.3810**	0.3638	0.3529
xQuAD	MinMax	0.3386	0.3636	**0.3650**	0.3388
	Sum	0.3664	**0.3941**	0.3874	0.3734
	Virtual	0.3660	**0.3810**	0.3638	0.3529
art_xQuAD	MinMax	0.3751	0.3900	**0.3944**	0.3727
	Sum	0.3612	**0.3898**	0.3823	0.3690
	Virtual	0.3892	0.3898	**0.3961**	0.3748
geo_xQuAD	MinMax	0.3581	0.3789	**0.3797**	0.3560
	Sum	0.3612	**0.3898**	0.3823	0.3690
	Virtual	0.3890	0.3921	**0.3973**	0.3737
PM2	MinMax	0.3705	**0.3833**	0.3747	0.3701
	Sum	0.3669	**0.3919**	0.3797	0.3771
	Virtual	0.3756	**0.3853**	0.3802	0.3804
mix_CombSUM	MinMax	0.3662	0.3629	**0.3684**	0.3639
	Sum	0.3613	**0.3886**	0.3803	0.3674
	Virtual	0.3811	0.3725	**0.3796**	0.3784
mix_Borda		0.3542	**0.3691**	0.3662	0.3608
mix_SV		0.3381	0.3731	**0.3815**	0.3693
mix_MC2		0.3645	0.3733	**0.3799**	0.3707

5.5 Conclusions and Future Work

In this chapter, we used PRF to expand aspect queries to improve the search result diversification. To this end, we used top-k documents from candidate documents' re-ranking and subtopic's own ranking which probably reside in the search engine's cache. Furthermore, we applied QPPs to select the subtopics that take benefit from the expanded terms. Through extensive experiments, we showed that selecting the sub-topics which will be expanded using PRF as subtopic's own ranking improve the performance of the diversification procedure. As a future work, we plan to use other metrics to select the aspects that need expansion.

CHAPTER 6

CONCLUSION

Search result diversification strategies try to provide a result set that covers the different interpretations of an ambiguous query to satisfy the user intentions. In this thesis, we evaluate state-of-the-art explicit search result diversification methods, find some weaknesses and propose methods to overcome these weaknesses. Our experiments showed that the new xQuAD variants outperform both the original xQuAD strategy and other better performing state-of-the-art diversification baselines.

We are also inspired from the success of explicit diversification methods which utilize the relevance of the candidate documents for each query aspect to propose to adapt score and rank based ranking aggregation methods to search result diversification domain. Our experiments revealed that some of these strategies, also serve well for the diversification purposed and outperform some of the state-of-the-art baselines from the literature. This is an especially important finding given that these methods can be computed more efficiently than the baseline diversification strategies and our xQuAD variants.

For the first time in the literature we proposed to use post-retrieval query performance predictors to estimate the query aspect weights and introduced 3 new QPP strategies while using several other strategies from the literature. The extensive experiments showed that predicting the retrieval effectiveness of each individual aspect on the candidate document set is a good indicator of an aspect's contribution to the quality of the final result.

Lastly, we used PRF to from candidate re-rankings and subtopics' own ranking from cache to expand the subtopic, and used QPPs to select the aspects that require expansion. Our experiments showed that expanding all sub-topics using sub-topics own results from the cache yield better diversification performance than unexpanded sub-topics. Furthermore selecting the aspects that require expansion also improve the diversification performance.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In R. A. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu, editors, *WSDM*, pages 5–14. ACM, 2009.

[2] C. Aksoy, F. Can, and S. Kocberber. Novelty detection for topic tracking. *JASIST*, 63(4):777–795, 2012.

[3] S. Alici, I. S. Altingövde, R. Ozcan, B. B. Cambazoglu, and Ö. Ulusoy. Timestamp-based result cache invalidation for web search engines. In Ma et al. [33], pages 973–982.

[4] A. Arampatzis and J. Kamps. A signal-to-noise approach to score normalization. In Cheung et al. [15], pages 797–806.

[5] J. A. Aslam and M. H. Montague. Models for metasearch. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 275–284. ACM, 2001.

[6] A. Bouchoucha, J. He, and J. Nie. Diversified query expansion using conceptnet. In Q. He, A. Iyengar, W. Nejdl, J. Pei, and R. Rastogi, editors, *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1861–1864. ACM, 2013.

[7] A. Bouchoucha, X. Liu, and J. Nie. Integrating multiple resources for diversified query expansion. In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, volume 8416 of *Lecture Notes in Computer Science*, pages 437–442. Springer, 2014.

[8] B. B. Cambazoglu, I. S. Altingövde, R. Ozcan, and Ö. Ulusoy. Cache-based query processing for search engines. *TWEB*, 6(4):14, 2012.

[9] G. Capannini, F. M. Nardini, R. Perego, and F. Silvestri. Efficient diversification of web search results. *PVLDB*, 4(7):451–459, 2011.

[10] J. G. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336. ACM, 1998.

[11] D. Carmel and O. Kurland. Query performance prediction for IR. In Hersh et al. [27], pages 1196–1197.

[12] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, Jan. 2001.

[13] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1, 2012.

[14] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In Cheung et al. [15], pages 1287–1296.

[15] D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors. *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*. ACM, 2009.

[16] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-278. National Institute of Standards and Technology (NIST), 2009.

[17] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2010 web track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*, volume Special Publication 500-294. National Institute of Standards and Technology (NIST), 2010.

[18] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.*, 14(5):441–465, 2011.

[19] R. Cummins, J. M. Jose, and C. O'Riordan. Improved query performance prediction using standard deviation. In Ma et al. [33], pages 1089–1090.

[20] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In Hersh et al. [27], pages 65–74.

[21] J. C. de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1784.

[22] E. D. Diaz, A. De, and V. Raghavan. A comprehensive owa-based framework for result merging in metasearch. In D. Slezak, J. Yao, J. F. Peters, W. Ziarko, and X. Hu, editors, *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference, RSFDGrC 2005, Regina, Canada, August 31 - September 3, 2005, Proceedings, Part II*, volume 3642 of *Lecture Notes in Computer Science*, pages 193–201. Springer, 2005.

[23] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In V. Y. Shen, N. Saito, M. R. Lyu, and M. E. Zurko, editors, *WWW*, pages 613–622. ACM, 2001.

[24] M. Fernández, D. Vallet, and P. Castells. Probabilistic score normalization for rank aggregation. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *ECIR*, volume 3936 of *Lecture Notes in Computer Science*, pages 553–556. Springer, 2006.

[25] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 381–390. ACM, 2009.

[26] J. He, E. Meij, and M. de Rijke. Result diversification based on query-specific cluster ranking. *JASIST*, 62(3):550–571, 2011.

[27] W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors. *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*. ACM, 2012.

[28] M.-H. Hsu, M.-F. Tsai, and H.-H. Chen. Query expansion with conceptnet and wordnet: An intrinsic comparison. In H. Ng, M.-K. Leong, M.-Y. Kan, and D. Ji, editors, *Information Retrieval Technology*, volume 4182 of *Lecture Notes in Computer Science*, pages 1–13. Springer Berlin Heidelberg, 2006.

[29] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: Leveraging conceptnet to improve search results for difficult queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 403–412, New York, NY, USA, 2012. ACM.

[30] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.

[31] J.-H. Lee. Analyses of multiple evidence combination. In *SIGIR*, pages 267–276. ACM, 1997.

[32] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using wikipedia asexternal corpus. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 797–798, New York, NY, USA, 2007. ACM.

[33] W.-Y. Ma, J.-Y. Nie, R. A. Baeza-Yates, T.-S. Chua, and W. B. Croft, editors. *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. ACM, 2011.

[34] C. Macdonald, I. Ounis, and N. Tonellotto. Upper-bound approximations for dynamic pruning. *ACM Trans. Inf. Syst.*, 29(4):17, 2011.

[35] G. Markovits, A. Shtok, O. Kurland, and D. Carmel. Predicting query performance for fusion-based retrieval. In X. wen Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors, *CIKM*, pages 813–822. ACM, 2012.

[36] E. Minack, W. Siberski, and W. Nejdl. Incremental diversification for very large sets: a streaming-based approach. In Ma et al. [33], pages 585–594.

[37] M. H. Montague and J. A. Aslam. Relevance score normalization for metasearch. In *CIKM*, pages 427–433. ACM, 2001.

[38] R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614, 2006.

[39] A. M. Ozdemiray and I. S. Altingovde. Score and rank aggregation methods for explicit search result diversification. Technical Report METU-CENG-2013-01, Middle East University, Computer Engineering Department, September 2013.

[40] A. M. Ozdemiray and I. S. Altingovde. Query performance prediction for aspect weighting in search result diversification. In J. Li, X. S. Wang, M. N. Garofalakis, I. Soboroff, T. Suel, and M. Wang, editors, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1871–1874. ACM, 2014.

[41] A. M. Ozdemiray and I. S. Altingovde. Explicit search result diversification using score and rank aggregation methods. *Journal of the Association for Information Science and Technology*, 66(6):1212–1228, 2015.

[42] J. Pérez-Iglesias and L. Araujo. Standard deviation as a query hardness estimator. In E. Chávez and S. Lonardi, editors, *SPIRE*, volume 6393 of *Lecture Notes in Computer Science*, pages 207–212. Springer, 2010.

[43] F. Radlinski and S. T. Dumais. Improving personalized web search using result diversification. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors, *SIGIR*, pages 691–692. ACM, 2006.

[44] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In Rappa et al. [45], pages 781–790.

[45] M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors. *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*. ACM, 2010.

[46] S. D. Ravana and A. Moffat. Score aggregation techniques in retrieval experimentation. In A. Bouguettaya and X. Lin, editors, *ADC*, volume 92 of *CRPIT*, pages 59–67. Australian Computer Society, 2009.

[47] M. E. Renda and U. Straccia. Web metasearch: Rank vs. score based rank aggregation methods. In *SAC*, pages 841–846. ACM, 2003.

[48] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC*, pages 0–, 1994.

[49] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[50] R. L. T. Santos, P. Castells, I. S. Altingövde, and F. Can. Diversity and novelty in information retrieval. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai, editors, *SIGIR*, page 1130. ACM, 2013.

[51] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In Rappa et al. [45], pages 881–890.

[52] R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In Ma et al. [33], pages 595–604.

[53] H. Sever and M. R. Tolun. Comparison of normalization techniques for metasearch. In T. M. Yakhno, editor, *ADVIS*, volume 2457 of *Lecture Notes in Computer Science*, pages 133–143. Springer, 2002.

[54] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2*, pages 243–252, 1994.

[55] M. Shokouhi and L. Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

[56] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2):11, 2012.

[57] S. Tomlinson. Robust, web and terabyte retrieval with hummingbird searchserver at TREC 2004. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-261. National Institute of Standards and Technology (NIST), 2004.

[58] D. Vallet and P. Castells. Personalized diversification of search results. In Hersh et al. [27], pages 841–850.

[59] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In B. Mobasher, R. D. Burke, D. Jannach, and G. Adomavicius, editors, *RecSys*, pages 109–116. ACM, 2011.

[60] S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In Hersh et al. [27], pages 75–84.

[61] S. Vargas, R. L. T. Santos, C. Macdonald, and I. Ounis. Selecting effective expansion terms for diversity. In J. Ferreira, J. Magalhães, and P. Calado, editors, *Open research Areas in Information Retrieval, OAIR '13, Lisbon, Portugal, May 15-17, 2013*, pages 69–76. ACM, 2013.

[62] M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. T. Jr., and V. J. Tsotras. On query result diversification. In S. Abiteboul, K. Böhm, C. Koch, and K.-L. Tan, editors, *ICDE*, pages 1163–1174. IEEE Computer Society, 2011.

[63] J. Wang and J. Zhu. Portfolio theory of information retrieval. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *SIGIR*, pages 115–122. ACM, 2009.

[64] M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, editors, *WWW*, pages 237–246. ACM, 2011.

[65] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*. Morgan Kaufmann, 1999.

[66] Zettair. The Zettair search engine. http://www.seg.rmit.edu.au/zettair/. Last accessed on September 2013.

[67] C. Zhai, W. W. Cohen, and J. D. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17. ACM, 2003.

[68] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410, New York, NY, USA, 2001. ACM.

[69] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[70] C. Zhai and J. D. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.

[71] W. Zheng and H. Fang. A comparative study of search result diversification methods. In *Proceedings of the ECIR 2011 Workshop on Diversity in Document Retrieval*, 2011.

[72] W. Zheng, X. Wang, H. Fang, and H. Cheng. Coverage-based search result diversification. *Inf. Retr.*, 15(5):433–457, 2012.

[73] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 543–550. ACM, 2007.

[74] G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. M. Rüger, and K. van Rijsbergen, editors, *ECIR*, volume 5993 of *Lecture Notes in Computer Science*, pages 357–369. Springer, 2010.

[75] G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-k retrieval using facility location analysis. In R. A. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, and F. Silvestri, editors, *ECIR*, volume 7224 of *Lecture Notes in Computer Science*, pages 305–316. Springer, 2012.

APPENDIX A

ADDITIONAL EXPERIMENTS

A.1 BM25 retrieval model and query aspects obtained from the suggestions

Table A.1: Diversification performance w.r.t. the relevance normalization techniques using the query aspects obtained from the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25		-	0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
org_xQuAD	MinMax	0.83	0.1884	**0.2801**	0.0757	0.60	0.1921	0.2779	0.1235
	Sum	0.2	**0.1902**	0.2792	**0.0822**	0.10	**0.2041**	**0.2963**	**0.1369**
	Virtual	0.97	0.1797	0.2737	0.0763	0.38	0.2012	0.2883	0.1241
IA-Select	MinMax	-	0.1778	**0.2814**	0.0783	-	0.1863	0.2815	0.1103
	Sum	-	**0.1806**	0.2688	**0.0796**	-	**0.2035**	**0.2962**	**0.1300**
	Virtual	-	0.1744	0.2675	0.0779	-	0.2028	0.2966	0.1129
PM2	-	0.25	**0.1937**	**0.2891**	**0.0840**	0.34	0.2021	0.3014	**0.1318**
	Sum	0.25	0.1809	0.2710	0.0798	0	**0.2145**	**0.3028**	0.1297
	Virtual	0.64	0.1692	0.2636	0.0791	0.05	0.2118	0.3006	0.1302

Table A.2: Diversification performance of the xQuAD variants using the query aspects obtained from the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25		-	0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
	MinMax	0.83	0.1884	0.2801	0.0757	0.60	0.1921	0.2779	0.1235
org_xQuAD	Sum	0.2	0.1902	0.2792	0.0822	0.10	0.2041	0.2963	0.1369
	Virtual	0.97	0.1797	0.2737	0.0763	0.38	0.2012	0.2883	0.1241
		0.86	0.1938	**0.2948**	0.0868	0.95	0.2025	0.2971	0.1234
geo_xQuAD	Sum	0.2	0.1913	0.2828	0.0829	0.1	0.2022	0.2921	**0.1396**
	Virtual	0.5	0.1938	0.2904	0.0860	0.38	0.2068	0.3005	0.1359
	MinMax	0.82	**0.1954**	0.2936	**0.0887**	0.66	**0.2070**	0.2968	0.1328
art_xQuAD	Sum	0.2	0.1913	0.2828	0.0829	0.1	0.2022	0.2921	**0.1396**
	Virtual	0.5	0.1924	0.2854	0.0836	0.38	**0.2070**	**0.3008**	0.1360

Table A.3: Diversification performance of the score aggregation methods using the query aspects obtained from the the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25		-	0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
	MinMax	1	0.1906	0.2811	0.0800	0.1	0.2012	0.2838	0.1257
mix_CombMNZ	Sum	1	0.1817	0.2735	0.0819	0	0.1947	0.2788	0.1254
	Virtual	0.6	0.1879	0.2795	0.0818	0.1	0.2224	0.3073	0.1303
	MinMax	0.9	0.1953	0.2871	**0.0881**	0.6	0.1919	0.2840	0.1300
mix_CombSUM	Sum	0.2	0.1914	0.2829	0.0829	0.1	0.2033	0.2942	0.1400
	Virtual	0.25	**0.2004**	**0.2913**	0.0847	0.3	**0.2161**	**0.3123**	**0.1499**

Table A.4: Diversification performance of the rank aggregation methods using the query aspects obtained from the suggestions and BM25 as the retrieval model. The highest scores are shown in boldface.

	2009				2010			
	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
BM25		0.1878	0.2757	0.0760	-	0.1947	0.2788	0.1254
mix_SV	0.95	0.1738	0.2671	0.0808	0.85	**0.2271**	**0.3027**	0.1277
mix_BV	1	0.1932	0.2851	**0.0844**	0.9	0.2127	0.2991	**0.1374**
mix_MC1	-	0.1873	0.2815	0.0795	-	0.2059	0.2902	0.1243
mix_MC2	-	0.1914	0.2858	0.0799	-	0.2060	0.2976	0.1214
mix_MC3	-	0.1911	0.2854	0.0798	-	0.2016	0.2885	0.1252
mix_MC4	-	**0.2014**	**0.2937**	0.0801	-	0.2068	0.2904	0.1286

A.2 LM retrieval model and query aspects obtained from official sub-topics

Table A.5: Diversification performance w.r.t. the relevance normalization techniques using the query aspects obtained from the official sub-topics and LM as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
LM		-	0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
org_xQuAD	MinMax	1	0.2240	**0.3311**	0.0920	1.00	0.2517	0.3499	0.1496
	Sum	0.44	0.2078	0.3065	0.0939	0.56	**0.2634**	**0.3689**	**0.1562**
	Virtual	0.92	**0.2242**	0.3283	**0.0940**	0.79	0.2584	0.3514	0.1409
IA-Select	MinMax	-	**0.2240**	**0.3311**	**0.0920**	-	0.2517	0.3499	**0.1496**
	Sum	-	0.2113	0.3096	0.0874	-	0.2547	0.3618	0.1451
	Virtual	-	0.2143	0.3148	0.0774	-	**0.2631**	**0.3624**	0.1291
PM2	MinMax	0.66	**0.2160**	**0.3259**	**0.0923**	0.76	**0.2679**	**0.3751**	0.1314
	Sum	0.44	0.2110	0.3111	0.0896	0.71	0.2469	0.3547	**0.1326**
	Virtual	0.1	0.2094	0.3076	0.0888	0.8	0.2662	0.3674	0.1331

Table A.6: Diversification performance of the xQuAD variants using the query aspects obtained from the official sub-topics and LM as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
LM		-	0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
org_xQuAD	MinMax	1	0.2240	**0.3311**	0.0920	1.00	0.2517	0.3499	0.1496
	Sum	0.44	0.2078	0.3065	0.0939	0.56	0.2634	0.3689	0.1562
	Virtual	0.92	**0.2242**	0.3283	0.0940	0.79	0.2584	0.3514	0.1409
geo_xQuAD	MinMax	0.96	0.2163	0.3238	0.0918	1	0.2521	0.3530	0.1478
	Sum	0.4	0.2048	0.3038	0.0940	0.79	0.2568	0.3527	0.1453
	Virtual	0.95	0.2238	0.3281	**0.1006**	0.78	**0.2721**	**0.3792**	**0.1614**
art_xQuAD	MinMax	0.97	0.2166	0.3235	0.0978	0.91	0.2604	0.3687	0.1576
	Sum	0.4	0.2048	0.3038	0.0940	0.79	0.2568	0.3527	0.1453
	Virtual	0.95	0.2240	0.3284	**0.1006**	0.78	**0.2721**	**0.3792**	**0.1614**

Table A.7: Diversification performance of the score aggregation methods using the query aspects obtained from the official sub-topics and LM as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
LM		-	0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
mix_CombMNZ	MinMax	0.45	**0.2343**	**0.3334**	0.1022	0.45	0.2552	0.3604	0.1549
	Sum	0.55	0.2138	0.3077	0.0948	0.15	0.2597	0.3618	0.1560
	Virtual	0.8	0.2273	0.3242	0.1007	0.25	0.2506	0.3533	0.1564
mix_CombSUM	MinMax	0.85	0.2193	0.3207	**0.1036**	0.8	0.2639	0.3682	**0.1648**
	Sum	0.4	0.2047	0.3037	0.0943	0.8	0.2546	0.3508	0.1449
	Virtual	0.95	0.2232	0.3253	0.1033	0.75	**0.2712**	**0.3780**	0.1639

Table A.8: Diversification performance of the rank aggregation methods using the query aspects obtained from the official sub-topics and LM as the retrieval model.. The highest scores are shown in boldface.

	2009				2010			
	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
LM		0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
mix_SV	0.8	0.2119	0.3137	**0.0983**	0.7	0.2421	0.3452	0.1535
mix_BV	0.85	0.2066	0.3071	0.0967	0.8	0.2362	0.3421	**0.1608**
mix_MC1	-	**0.2222**	0.3242	0.0950	-	0.2525	0.3606	0.1483
mix_MC2	-	**0.2222**	**0.3282**	0.0944	-	**0.2645**	**0.3714**	0.1394
mix_MC3	-	0.2185	0.3213	0.0952	-	0.2591	0.3677	0.1479
mix_MC4	-	0.2134	0.3117	0.0921	-	0.2484	0.3594	0.1497

A.3 LM retrieval model and query aspects obtained from the suggestions

Table A.9: Diversification performance w.r.t. the relevance normalization techniques using the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface.

	Relevance	TREC 2009				TREC 2010			
	norm.	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
LM		-	0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
org_xQuAD	MinMax	0.97	0.1923	**0.2929**	0.0941	0.99	0.2057	0.3020	0.1398
	Sum	0.57	**0.1928**	**0.2929**	**0.1014**	0.46	**0.2164**	**0.3147**	**0.1486**
	Virtual	0.97	0.1895	0.2905	0.0945	0.78	0.2127	0.2997	0.1393
IA-Select	MinMax	-	0.1913	**0.2916**	0.0908	-	0.1997	0.2956	0.1294
	Sum	-	**0.1985**	0.2881	**0.0953**	-	0.2021	0.3020	**0.1424**
	Virtual	-	0.1890	0.2847	0.0882	-	**0.2106**	**0.3078**	0.1195
PM2	MinMax	0.74	**0.1891**	**0.2870**	0.0921	0.46	**0.2039**	**0.3033**	0.1344
	Sum	0.9	0.1721	0.2702	0.0893	0	0.1956	0.2913	0.1353
	Virtual	0.87	0.1697	0.2670	**0.0950**	0.2	0.1954	0.2867	**0.1367**

Table A.10: Diversification performance of the xQuAD variants using the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface.

	Relevance norm.	TREC 2009 λ	ERR-IA	α-nDCG	P-IA	TREC 2010 λ	ERR-IA	α-nDCG	P-IA
LM		-	0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
org_xQuAD	MinMax	0.97	0.1923	0.2929	0.0941	0.99	0.2057	0.3020	0.1398
	Sum	0.57	0.1928	0.2929	**0.1014**	0.46	**0.2164**	**0.3147**	0.1486
	Virtual	0.97	0.1895	0.2905	0.0945	0.78	0.2127	0.2997	0.1393
geo_xQuAD	MinMax	0.85	0.1934	0.2939	0.0986	0.94	0.2064	0.3091	0.1422
	Sum	0.57	0.1913	0.2871	0.0992	0.46	0.2160	0.3130	**0.1488**
	Virtual	0.76	**0.1938**	**0.2941**	0.1001	0.62	0.2146	0.3122	0.1464
art_xQuAD	MinMax	0.97	0.1923	0.2927	0.0957	0.84	0.2123	0.3090	0.1441
	Sum	0.57	0.1913	0.2871	0.0992	0.46	0.2160	0.3130	**0.1488**
	Virtual	0.77	0.1929	0.2931	0.0992	0.62	0.2139	0.3117	0.1469

Table A.11: Diversification performance of the score aggregation methods the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface.

	Relevance norm.	TREC 2009 λ	ERR-IA	α-nDCG	P-IA	TREC 2010 λ	ERR-IA	α-nDCG	P-IA
LM		-	0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
mix_CombMNZ	MinMax	0.65	0.1941	0.2860	0.0963	0.5	0.2033	0.2983	0.1343
	Sum	0.35	0.1838	0.2763	0.0987	0.1	0.2117	0.3062	0.1326
	Virtual	0.15	0.1929	0.2865	0.0973	0.1	0.2001	0.2967	0.1341
mix_CombSUM	MinMax	0.85	**0.1992**	**0.2967**	0.0965	0.85	0.2116	0.3075	0.1452
	Sum	0.7	0.1913	0.2821	0.0983	0.5	**0.2149**	**0.3124**	0.1480
	Virtual	0.75	0.1871	0.2857	**0.1001**	0.55	0.2127	0.3114	**0.1485**

Table A.12: Diversification performance of the rank aggregation methods the query aspects obtained from the suggestions and LM as the retrieval model. The highest scores are shown in boldface.

	2009				2010			
	λ	ERR-IA	α-nDCG	P-IA	λ	ERR-IA	α-nDCG	P-IA
LM		0.1738	0.2645	0.0930	-	0.1959	0.2842	0.1406
mix_SV	0.95	0.1790	0.2711	**0.0944**	0.65	**0.2098**	0.2963	0.1370
mix_BV	0.35	0.1773	0.2682	0.0923	0.7	0.2063	0.2981	**0.1382**
mix_MC1	-	0.1872	0.2808	0.0918	-	0.2039	0.2936	0.1286
mix_MC2	-	**0.1932**	**0.2887**	0.0913	-	0.2024	**0.2984**	0.1230
mix_MC3	-	0.1873	0.2816	0.0929	-	0.2023	0.2953	0.1286
mix_MC4	-	0.1887	0.2786	0.0900	-	0.1889	0.2828	0.1294

A.4 Aspect Weighting using Query Performance Predictions

A.4.1 2009 official sub-topics

Table A.13: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the baseline QPPs for the query aspects obtained from the official sub-topics. The highest score is boldfaced.

Div. method	Relevance norm.	Uniform	**Baseline QPPs**			
			WIG	NQC	ScrAvg	ScrDev
IA-Select	MinMax	0.3250	**0.3286**	0.3040	0.3224	0.3119
	Sum	0.3179	0.3182	0.2935	**0.3202**	0.3103
	Virtual	**0.3263**	0.3095	0.3038	0.3144	0.3096
xQuAD	MinMax	0.3308	**0.3313**	0.3118	0.3291	0.3184
	Sum	0.3255	0.3223	0.3068	**0.3284**	0.3159
	Virtual	**0.3268**	0.3154	0.3045	0.3260	0.3147
art_xQuAD	MinMax	0.3391	**0.3403**	0.3208	0.3387	0.3284
	Sum	0.3255	0.3230	0.3082	**0.3275**	0.3149
	Virtual	0.3354	0.3304	0.3162	**0.3359**	0.3203
geo_xQuAD	MinMax	**0.3430**	0.3426	0.3176	0.3395	0.3261
	Sum	0.3255	0.3230	0.3082	**0.3275**	0.3149
	Virtual	0.3341	0.3300	0.3183	**0.3348**	0.3202
PM2	MinMax	0.3322	**0.3345**	0.3253	0.3339	0.3282
	Sum	**0.3328**	0.3293	0.3269	0.3272	0.3282
	Virtual	0.3360	**0.3372**	0.3243	0.3308	0.3287
mix_CombSUM	MinMax	0.3323	**0.3335**	0.3102	0.3275	0.3202
	Sum	0.3249	0.3226	0.3083	**0.3271**	0.3129
	Virtual	**0.3338**	0.3314	0.3211	0.3313	0.3184

Table A.14: Diversification performance (α-NDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the best baseline QPPs and proposed QPPs for the query aspects obtained from the official sub-topics. The highest score in each group is bold, the overall winner is underlined.

Div. method	Relevance norm.	Uniform	Baseline QPPs		Proposed QPPs		
			WIG	ScrAvg	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.3250	**<u>0.3286</u>**	0.3224	0.3233	0.3203	**0.3236**
	Sum	0.3179	0.3182	**0.3202**	0.3169	**<u>0.3223</u>**	0.3208
	Virtual	**<u>0.3263</u>**	0.3095	**0.3144**	0.3161	**0.3252**	0.3213
xQuAD	MinMax	0.3308	**<u>0.3313</u>**	0.3291	0.3261	0.3247	**0.3309**
	Sum	0.3255	0.3223	**0.3284**	0.3260	**<u>0.3301</u>**	0.3283
	Virtual	0.3268	0.3154	**0.3260**	0.3274	**<u>0.3312</u>**	0.3287
art_xQuAD	MinMax	0.3391	**0.3403**	0.3387	0.3312	0.3330	**<u>0.3414</u>**
	Sum	0.3255	0.3230	**0.3275**	0.3238	**<u>0.3302</u>**	0.3282
	Virtual	0.3354	0.3304	**0.3359**	0.3351	**<u>0.3383</u>**	0.3357
geo_xQuAD	MinMax	**<u>0.3430</u>**	**0.3426**	0.3395	0.3379	0.3363	**0.3400**
	Sum	0.3255	0.3230	**0.3275**	0.3238	**<u>0.3302</u>**	0.3282
	Virtual	0.3341	0.3300	**0.3348**	0.3344	**<u>0.3389</u>**	0.3346
PM2	MinMax	0.3322	**0.3345**	0.3339	0.3329	0.3333	**<u>0.3360</u>**
	Sum	0.3328	**0.3293**	0.3272	0.3340	0.3289	**<u>0.3345</u>**
	Virtual	0.3360	**0.3372**	0.3308	0.3343	0.3334	**<u>0.3388</u>**
mix_CombSUM	MinMax	0.3323	**<u>0.3335</u>**	0.3275	0.3247	**0.3293**	0.3290
	Sum	0.3249	0.3226	**0.3271**	0.3234	**<u>0.3298</u>**	0.3277
	Virtual	**<u>0.3338</u>**	0.3314	0.3313	0.3316	**0.3328**	0.3327

A.4.2 2009 sub-topics from suggestions

Table A.15: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the proposed QPPs for the query aspects obtained from the suggestions. The highest score is boldfaced.

Div. method	Relevance norm.	Uniform	**Proposed QPPs**		
			VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.2814	0.2875	**0.2913**	0.2892
	Sum	0.2688	**0.2729**	0.2728	0.2689
	Virtual	0.2675	0.2711	0.2716	**0.2716**
xQuAD	MinMax	0.2960	**0.2971**	0.2967	0.2951
	Sum	0.2846	0.2854	0.2862	**0.2922**
	Virtual	0.2887	0.2930	0.2931	**0.2956**
art_xQuAD	MinMax	0.3049	**0.3059**	0.3044	0.3030
	Sum	0.2860	0.2824	0.2831	**0.2911**
	Virtual	0.2948	0.3000	0.3003	**0.3009**
geo_xQuAD	MinMax	0.3019	**0.3055**	0.3051	0.3023
	Sum	0.2860	0.2824	0.2831	**0.2911**
	Virtual	0.2957	**0.3026**	0.2999	0.3008
PM2	MinMax	0.2935	0.2988	**0.3016**	0.2959
	Sum	0.2775	**0.2949**	0.2855	0.2789
	Virtual	0.2889	**0.2986**	0.2869	0.2800
mix_CombSUM	MinMax	0.3043	**0.3056**	0.3031	0.3024
	Sum	0.2860	0.2823	0.2833	**0.2902**
	Virtual	0.2959	0.2963	0.2974	**0.2990**

A.4.3 2010 suggested subtopics

Table A.16: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the baseline QPPs. The highest score is boldfaced.

Div. method	Relevance norm.	Uniform	Proposed QPPs		
			VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.2952	**0.3022**	0.2939	0.2967
	Sum	**0.3043**	0.3000	0.3021	0.3037
	Virtual	0.3046	0.3066	0.3044	**0.3092**
xQuAD	MinMax	0.3072	0.3077	0.3090	**0.3113**
	Sum	0.3215	**0.3238**	0.3221	0.3196
	Virtual	0.3090	0.3082	0.3059	**0.3108**
art_xQuAD	MinMax	0.3225	0.3258	0.3250	**0.3277**
	Sum	0.3225	**0.3238**	0.3235	0.3197
	Virtual	**0.3210**	0.3201	0.3195	0.3209
geo_xQuAD	MinMax	0.3228	0.3212	0.3248	**0.3260**
	Sum	0.3225	**0.3238**	0.3235	0.3197
	Virtual	0.3194	**0.3201**	0.3186	0.3200
PM2	MinMax	0.3129	0.3153	0.3113	**0.3175**
	Sum	0.3129	0.3094	0.3063	0.3074
	Virtual	**0.3107**	0.3077	0.3063	0.3061
mix_CombSUM	MinMax	**0.3145**	0.3131	0.3125	0.3098
	Sum	0.3224	**0.3244**	0.3235	0.3197
	Virtual	0.3256	**0.3275**	0.3262	0.3260

A.5 Query expansion tables

A.5.1 2009 official sub-topics

Table A.17: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25		KLD	
			Add 5	Fix 10	Add 5	Fix 10
IA-Select	MinMax	**0.2242**	0.2060	0.1805	0.1912	0.1854
	Sum	**0.2141**	0.1975	0.1783	0.1924	0.2000
	Virt	**0.2318**	0.2024	0.1852	0.1883	0.1956
xQuAD	MinMax	**0.2326**	0.2108	0.2071	0.2092	0.2145
	Sum	0.2218	**0.2256**	0.2036	0.2132	0.2221
	Virt	**0.2329**	0.2220	0.1992	0.2179	0.2150
art_xQuAD	MinMax	**0.2342**	0.2235	0.2134	0.2158	0.2206
	Sum	0.2218	**0.2264**	0.2032	0.2143	0.2224
	Virt	**0.2355**	0.2237	0.2030	0.2181	0.2229
geo_xQuAD	MinMax	**0.2355**	0.2210	0.2125	0.2156	0.2180
	Sum	0.2218	**0.2264**	0.2032	0.2143	0.2224
	Virt	**0.2350**	0.2227	0.2045	0.2183	0.2218
PM2	MinMax	**0.2325**	0.2047	0.1950	0.1895	0.1901
	Sum	**0.2272**	0.2018	0.1860	0.1857	0.1981
	Virt	**0.2347**	0.2031	0.1906	0.1885	0.1888
mix_CombSUM	MinMax	**0.2362**	0.2214	0.2136	0.2106	0.2146
	Sum	0.2215	**0.2265**	0.2040	0.2146	0.2225
	Virt	**0.2353**	0.2309	0.2104	0.2167	0.2199
mix_Borda		**0.2307**	0.2150	0.2102	0.2105	0.2117
mix_SV		0.2077	0.2005	**0.2078**	0.1831	0.1968
mix_MC2		0.2249	**0.2282**	0.2083	0.2148	0.2056

Table A.18: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25		KLD	
			Add 5	Fix 10	Add 5	Fix 10
IA-Select	MinMax	**0.2242**	0.2146	0.2095	0.1907	0.2101
	Sum	**0.2141**	0.2017	0.1849	0.1913	0.1964
	Virt	**0.2318**	0.2209	0.2067	0.1912	0.2022
xQuAD	MinMax	**0.2326**	0.2309	0.2198	0.2237	0.2252
	Sum	**0.2218**	0.2213	0.2133	0.2173	0.2183
	Virt	**0.2329**	0.2299	0.2251	0.2105	0.2197
art_xQuAD	MinMax	0.2342	**0.2365**	0.2270	0.2281	0.2302
	Sum	**0.2218**	0.2210	0.2130	0.2094	0.2180
	Virt	**0.2355**	0.2278	0.2267	0.2178	0.2219
geo_xQuAD	0.2355	**0.2379**	0.2265	0.2276	0.2293	
	Sum	**0.2218**	0.2210	0.2130	0.2094	0.2180
	Virt	**0.2350**	0.2279	0.2260	0.2181	0.2213
PM2	MinMax	**0.2325**	0.2297	0.2185	0.2130	0.2095
	Sum	**0.2272**	0.2184	0.2113	0.2066	0.2154
	Virt	**0.2347**	0.2221	0.2118	0.2103	0.2102
mix_CombSUM	MinMax	**0.2362**	0.2350	0.2260	0.2178	0.2235
	Sum	**0.2215**	0.2211	0.2125	0.2175	0.2154
	Virt	**0.2353**	0.2263	0.2240	0.2159	0.2223
mix_Borda		**0.2307**	0.2138	0.2146	0.2137	0.2201
mix_SV		0.2077	0.2179	**0.2277**	0.1958	0.2143
mix_MC2		0.2249	**0.2272**	0.2197	0.2112	0.2206

Table A.19: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	**0.3240**	0.3151	0.3204	0.3195	0.3154
	Sum	**0.3162**	0.3082	0.3012	0.3011	0.3035
	Virtual	**0.3263**	0.3238	0.3162	0.3154	0.3198
xQuAD	MinMax	0.3298	0.3302	0.3318	**0.3333**	0.3235
	Sum	0.3238	0.3217	**0.3264**	0.3243	0.3258
	Virtual	0.3268	0.3273	0.3274	0.3192	**0.3281**
art_xQuAD	MinMax	0.3387	0.3442	0.3398	0.3436	**0.3445**
	Sum	0.3238	0.3212	**0.3253**	0.3241	0.3245
	Virtual	0.3354	0.3345	0.3351	0.3335	**0.3365**
geo_xQuAD	MinMax	0.3420	**0.3469**	0.3384	0.3446	0.3463
	Sum	0.3238	0.3212	**0.3253**	0.3241	0.3245
	Virtual	0.3341	0.3345	0.3349	0.3336	**0.3359**
PM2	MinMax	0.3334	**0.3367**	0.3298	0.3261	0.3253
	Sum	**0.3310**	0.3273	0.3155	0.3162	0.3164
	Virtual	**0.3360**	0.3286	0.3247	0.3239	0.3169
mix_CombSUM	MinMax	0.3323	**0.3377**	0.3347	0.3304	0.3352
	Sum	0.3235	0.3200	**0.3258**	0.3239	0.3250
	Virtual	**0.3339**	0.3269	0.3268	0.3290	0.3334

Table A.20: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	**0.2242**	0.2146	0.2222	0.2205	0.2132
	Sum	**0.2141**	0.2017	0.1973	0.1976	0.2002
	Virtual	**0.2318**	0.2209	0.2130	0.2083	0.2150
xQuAD	MinMax	0.2326	0.2309	0.2313	**0.2356**	0.2223
	Sum	0.2218	0.2213	**0.2271**	0.2240	0.2261
	Virtual	**0.2329**	0.2299	0.2298	0.2234	0.2312
art_xQuAD	MinMax	0.2342	0.2365	0.2315	0.2366	**0.2377**
	Sum	0.2218	0.2210	**0.2268**	0.2263	0.2255
	Virtual	**0.2355**	0.2278	0.2297	0.2293	0.2303
geo_xQuAD	MinMax	0.2355	0.2379	0.2304	0.2370	**0.2389**
	Sum	0.2218	0.2210	**0.2268**	0.2263	0.2255
	Virt	**0.2350**	0.2279	0.2303	0.2294	0.2298
PM2	MinMax	**0.2325**	0.2297	0.2223	0.2183	0.2141
	Sum	**0.2272**	0.2184	0.2105	0.2078	0.2117
	Virt	**0.2347**	0.2221	0.2142	0.2154	0.2100
mix_CombSUM	MinMax	**0.2362**	0.2350	0.2327	0.2294	0.2356
	Sum	0.2215	0.2211	**0.2271**	0.2261	0.2263
	Virt	**0.2353**	0.2263	0.2258	0.2295	0.2301

Table A.21: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by** VScrAvg	VScrFirst
IA-Select	MinMax	**0.2242**	0.2146	0.2007	0.2114
	Sum	**0.2141**	0.2017	0.2014	0.1995
	Virt	**0.2318**	0.2209	0.2143	0.2304
xQuAD	MinMax	**0.2326**	0.2309	0.2240	0.2249
	Sum	0.2218	0.2213	**0.2219**	0.2214
	Virt	**0.2329**	0.2299	0.2244	0.2304
art_xQuAD	MinMax	0.2342	**0.2365**	0.2325	0.2335
	Sum	**0.2218**	0.2210	0.2215	0.2209
	Virt	0.2355	0.2278	0.2298	**0.2394**
geo_xQuAD	MinMax	0.2355	**0.2379**	0.2274	0.2302
	Sum	**0.2218**	0.2210	0.2215	0.2209
	Virt	0.2350	0.2279	0.2295	**0.2364**
PM2	MinMax	0.2325	0.2297	0.2260	**0.2333**
	Sum	**0.2272**	0.2184	0.2119	0.2240
	Virt	**0.2347**	0.2221	0.2220	0.2325
mix_CombSUM	MinMax	0.2362	0.2350	**0.2371**	0.2362
	Sum	0.2215	0.2211	**0.2215**	0.2209
	Virt	0.2353	0.2263	0.2277	**0.2376**
mix_Borda		**0.2307**	0.2138	0.2238	0.2187
mix_SV		0.2077	**0.2179**	0.2179	0.2147
mix_MC2		0.2249	0.2272	0.2318	**0.2322**

Table A.22: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**by VScrFirst < 0.7**			
				Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.3240	0.3151	0.3097	0.3144	0.3173	**0.3204**
	Sum	0.3162	0.3082	0.305	0.3152	**0.3232**	0.3119
	Virtual	0.3263	0.3238	**0.3295**	0.3205	0.3251	0.3262
xQuAD	MinMax	0.3298	**0.3302**	0.3251	0.3259	0.3234	0.3267
	Sum	0.3238	0.3217	0.3229	0.3244	0.3325	**0.3342**
	Virtual	0.3268	0.3273	0.3295	0.3278	0.3315	**0.3335**
art_xQuAD	MinMax	0.3387	**0.3442**	0.3421	0.3418	0.3414	0.3435
	Sum	0.3238	0.3212	0.3220	0.3245	0.3330	0.3339
	Virtual	0.3354	0.3345	0.3415	0.3410	0.3399	**0.3432**
geo_xQuAD	MinMax	0.3420	**0.3469**	0.3405	0.3426	0.3383	0.337
	Sum	0.3238	0.3212	0.3220	0.3245	0.3330	**0.3339**
	Virtual	0.3341	0.3345	0.3407	0.3398	0.3403	**0.3424**
PM2	MinMax	0.3334	0.3367	**0.3432**	0.3398	0.3341	0.3406
	Sum	0.3310	0.3273	0.3311	**0.3330**	0.3284	0.3285
	Virtual	0.3360	0.3286	0.3384	**0.3397**	0.3319	0.3365
mix_CombSUM	MinMax	0.3323	0.3377	0.3441	0.3398	0.3386	**0.3450**
	Sum	0.3235	0.3200	0.3218	0.3242	0.3327	**0.3336**
	Virtual	0.3339	0.3269	**0.3389**	0.3294	0.3338	0.3332

Table A.23: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	by VScrFirst < 0.7			
				Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	**0.2242**	0.2146	0.2114	0.2151	0.2156	0.2210
	Sum	0.2141	0.2017	0.1995	0.2112	**0.2221**	0.2075
	Virt	**0.2318**	0.2209	0.2304	0.2186	0.2228	0.2251
xQuAD	MinMax	**0.2326**	0.2309	0.2249	0.2272	0.2233	0.2280
	Sum	0.2218	0.2213	0.2214	0.2247	0.2318	**0.2334**
	Virt	0.2329	0.2299	0.2304	0.2292	0.2331	**0.2367**
art_xQuAD	MinMax	0.2342	**0.2365**	0.2335	0.2336	0.2327	0.2336
	Sum	0.2218	0.2210	0.2209	0.2249	0.2323	**0.2327**
	Virt	0.2355	0.2278	**0.2394**	0.2348	0.2359	0.2388
geo_xQuAD	MinMax	0.2355	**0.2379**	0.2302	0.2337	0.2295	0.2324
	Sum	0.2218	0.2210	0.2209	0.2249	0.2323	**0.2327**
	Virt	0.2350	0.2279	0.2364	0.2342	0.2335	**0.2392**
PM2	MinMax	0.2325	0.2297	**0.2333**	0.2331	0.2293	0.2315
	Sum	**0.2272**	0.2184	0.2240	0.2224	0.2249	0.2207
	Virt	0.2347	0.2221	0.2325	**0.2362**	0.2285	0.2339
mix_CombSUM	MinMax	0.2362	0.2350	0.2362	0.2346	0.2339	**0.2370**
	Sum	0.2215	0.2211	0.2209	0.2248	0.2322	**0.2326**
	Virt	0.2353	0.2263	**0.2376**	0.2263	0.2301	0.2322

A.5.2 2010 official sub-topics

Table A.24: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.2445	**0.2714**	0.2707	0.2476	0.2541
	Sum	0.2529	**0.2732**	0.2667	0.2628	0.2586
	Virt	0.2681	**0.2804**	0.2736	0.2664	0.2638
xQuAD	MinMax	0.2445	**0.2714**	0.2707	0.2512	0.2541
	Sum	0.2643	0.2739	**0.2802**	0.2651	0.2592
	Virt	0.2681	**0.2804**	0.2736	0.2664	0.2639
art_xQuAD	MinMax	0.2652	0.2853	**0.2892**	0.2653	0.2616
	Sum	0.2629	0.2736	**0.2787**	0.2643	0.2603
	Virt	0.2799	0.2744	**0.2832**	0.2678	0.2656
geo_xQuAD	MinMax	0.2571	0.2794	**0.2839**	0.2607	0.2581
	Sum	0.2629	0.2736	**0.2787**	0.2643	0.2603
	Virt	0.2799	0.2739	**0.2822**	0.2615	0.2651
PM2	MinMax	0.2702	**0.2800**	0.2733	0.2622	0.2582
	Sum	0.2597	0.2787	**0.2818**	0.2661	0.2501
	Virt	0.2722	**0.2802**	0.2732	0.2701	0.2653
mix_CombSUM	MinMax	0.2576	0.2688	**0.2693**	0.2601	0.2651
	Sum	0.2629	0.2728	**0.2784**	0.2637	0.2603
	Virt	**0.2761**	0.2706	0.2721	0.2710	0.2621
mix_Borda		0.2474	0.2805	**0.2834**	0.2748	0.2664
mix_SV		0.2327	0.2510	0.2538	0.2531	**0.2571**
mix_MC2		0.2559	0.2552	**0.2723**	0.2687	0.2722

Table A.25: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.2445	**0.2681**	0.2645	0.2428	0.2470
	Sum	0.2529	**0.2874**	0.2770	0.2469	0.2363
	Virt	0.2681	**0.2831**	0.2827	0.2541	0.2470
xQuAD	MinMax	0.2445	**0.2681**	0.2645	0.2428	0.2470
	Sum	0.2643	**0.2961**	0.2881	0.2560	0.2467
	Virt	0.2681	**0.2831**	0.2827	0.2581	0.2470
art_xQuAD	MinMax	0.2652	**0.2839**	0.2788	0.2663	0.2645
	Sum	0.2629	**0.2949**	0.2871	0.2578	0.2499
	Virt	0.2799	0.2895	**0.2967**	0.2653	0.2692
geo_xQuAD	MinMax	0.2571	**0.2838**	0.2794	0.2637	0.2680
	Sum	0.2629	**0.2949**	0.2871	0.2578	0.2499
	Virt	0.2799	0.2902	**0.2968**	0.2665	0.2691
PM2	MinMax	0.2702	**0.2820**	0.2774	0.2512	0.2567
	Sum	0.2597	0.2836	**0.2886**	0.2556	0.2616
	Virt	0.2722	**0.2837**	0.2782	0.2521	0.2572
mix_CombSUM	MinMax	0.2576	**0.2698**	0.2693	0.2600	0.2526
	Sum	0.2629	**0.2914**	0.2846	0.2576	0.2501
	Virt	0.2761	0.2809	**0.2878**	0.2541	0.2595
mix_Borda		0.2474	0.2742	**0.2862**	0.2600	0.2669
mix_SV		0.2327	0.2718	**0.2821**	0.2531	0.2630
mix_MC2		0.2559	0.2628	**0.2720**	0.2569	0.2662

Table A.26: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.3386	0.3636	0.3630	**0.3765**	0.3679
	Sum	0.3568	**0.3823**	0.3792	0.3710	0.3777
	Virtual	0.3660	0.3810	0.3835	**0.3892**	0.3857
xQuAD	MinMax	0.3386	**0.3636**	0.3630	**0.3765**	0.3679
	Sum	0.3664	**0.3941**	0.3877	0.3799	0.3862
	Virtual	0.3660	0.3810	0.3835	**0.3917**	0.3857
art_xQuAD	MinMax	0.3751	0.3900	**0.3946**	0.3899	0.3790
	Sum	0.3612	**0.3898**	0.3826	0.3760	0.3834
	Virtual	0.3892	**0.3898**	0.3896	0.3877	0.3857
geo_xQuAD	MinMax	0.3581	0.3789	**0.3814**	0.3743	0.3648
	Sum	0.3612	**0.3898**	0.3826	0.3760	0.3834
	Virtual	0.3890	**0.3921**	0.3899	0.3876	0.3861
PM2	MinMax	0.3705	**0.3833**	0.3779	0.3790	0.3757
	Sum	0.3669	**0.3919**	0.3892	0.3851	0.3842
	Virtual	0.3756	**0.3853**	0.3823	0.3807	0.3808
mix_CombSUM	MinMax	**0.3662**	0.3629	0.3604	0.3608	0.3605
	Sum	0.3613	**0.3886**	0.3814	0.3746	0.3830
	Virtual	**0.3811**	0.3725	0.3728	0.3728	0.3730

Table A.27: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.2445	0.2681	0.2665	**0.2789**	0.2717
	Sum	0.2529	**0.2874**	0.2862	0.2795	0.2829
	Virt	0.2681	0.2831	0.2834	**0.2906**	0.2885
xQuAD	MinMax	0.2445	0.2681	0.2665	**0.2789**	0.2717
	Sum	0.2643	**0.2961**	0.2913	0.2863	0.2917
	Virt	0.2681	0.2831	0.2834	**0.2936**	0.2885
art_xQuAD	MinMax	0.2652	0.2839	**0.2877**	0.2852	0.2757
	Sum	0.2629	**0.2949**	0.2892	0.2840	0.2907
	Virt	0.2799	0.2895	0.2882	**0.2916**	0.2865
geo_xQuAD	MinMax	0.2571	**0.2838**	0.2802	0.2709	0.2716
	Sum	0.2629	**0.2949**	0.2892	0.2840	0.2907
	Virt	0.2799	0.2902	0.2886	**0.2908**	0.2868
PM2	MinMax	0.2702	**0.2820**	0.2751	0.2781	0.2801
	Sum	0.2597	0.2836	**0.2912**	0.2861	0.2892
	Virt	0.2722	0.2837	0.2872	**0.2873**	0.2865
mix_CombSUM	MinMax	0.2576	**0.2698**	0.2626	0.2618	0.2660
	Sum	0.2629	**0.2914**	0.2885	0.2838	0.2902
	Virt	0.2761	0.2809	0.2809	**0.2854**	0.2817

Table A.28: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by** VScrAvg	**Selective by** VScrFirst
IA-Select	MinMax	0.2445	0.2681	**0.2714**	0.2412
	Sum	0.2529	**0.2874**	0.2652	0.2703
	Virt	0.2681	**0.2831**	0.2684	0.2608
xQuAD	MinMax	0.2445	0.2681	**0.2714**	0.2412
	Sum	0.2643	**0.2961**	0.2860	0.2725
	Virt	0.2681	**0.2831**	0.2684	0.2608
art_xQuAD	MinMax	0.2652	0.2839	**0.2901**	0.2648
	Sum	0.2629	**0.2949**	0.2822	0.2720
	Virt	0.2799	0.2895	**0.2899**	0.2652
geo_xQuAD	MinMax	0.2571	0.2838	**0.2854**	0.2591
	Sum	0.2629	**0.2949**	0.2822	0.2720
	Virt	0.2799	**0.2902**	0.2902	0.2656
PM2	MinMax	0.2702	**0.2820**	0.2715	0.2607
	Sum	0.2597	**0.2836**	0.2729	0.2761
	Virt	0.2722	**0.2837**	0.2750	0.2736
mix_CombSUM	MinMax	0.2576	**0.2698**	0.2697	0.2600
	Sum	0.2629	**0.2914**	0.2799	0.2706
	Virt	0.2761	**0.2809**	0.2798	0.2674
mix_Borda		0.2474	**0.2742**	0.2665	0.2588
mix_SV		0.2327	0.2718	**0.2809**	0.2656
mix_MC2		0.2559	0.2628	**0.2673**	0.2566

Table A.29: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	by VScrAvg < 0.7			
				Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.3386	0.3636	0.3650	**0.3747**	0.3626	0.3493
	Sum	0.3568	**0.3823**	0.3721	0.3748	0.3744	0.3715
	Virtual	0.3660	**0.3810**	0.3638	0.3570	0.3561	0.3529
xQuAD	MinMax	0.3386	0.3636	0.3650	**0.3747**	0.3626	0.3493
	Sum	0.3664	**0.3941**	0.3874	0.3906	0.3847	0.3850
	Virtual	0.3660	**0.3810**	0.3638	0.3600	0.3618	0.3559
art_xQuAD	MinMax	0.3751	0.3900	0.3944	**0.3997**	0.3888	0.3826
	Sum	0.3612	**0.3898**	0.3823	0.3854	0.3819	0.3802
	Virtual	0.3892	0.3898	**0.3961**	0.3904	0.3836	0.3807
geo_xQuAD	MinMax	0.3581	0.3789	0.3797	**0.3866**	0.3782	0.3717
	Sum	0.3612	**0.3898**	0.3823	0.3854	0.3819	0.3802
	Virtual	0.3890	0.3921	**0.3973**	0.3896	0.3820	0.3800
PM2	MinMax	0.3705	**0.3833**	0.3747	0.3751	0.3691	0.3600
	Sum	0.3669	**0.3919**	0.3797	0.3822	0.3821	0.3851
	Virtual	0.3756	**0.3853**	0.3802	0.3794	0.3814	0.3815
mix_CombSUM	MinMax	0.3662	0.3629	0.3684	**0.3708**	0.3632	0.3653
	Sum	0.3613	**0.3886**	0.3803	0.3824	0.3801	0.3799
	Virtual	0.3811	0.3725	**0.3796**	0.3694	0.3668	0.3682

Table A.30: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the official sub-topics and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**by VScrAvg < 0.7** Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.2445	0.2681	0.2714	**0.2783**	0.2654	0.2517
	Sum	0.2529	**0.2874**	0.2652	0.2678	0.2722	0.2723
	Virt	0.2681	**0.2831**	0.2684	0.2588	0.2590	0.2528
xQuAD	MinMax	0.2445	0.2681	0.2714	**0.2783**	0.2654	0.2517
	Sum	0.2643	**0.2961**	0.2860	0.2885	0.2798	0.2864
	Virt	0.2681	**0.2831**	0.2684	0.2635	0.2645	0.2610
art_xQuAD	MinMax	0.2652	0.2839	0.2901	**0.2942**	0.2846	0.2811
	Sum	0.2629	**0.2949**	0.2822	0.2872	0.2855	0.2810
	Virt	0.2799	0.2895	**0.2899**	0.2816	0.2756	0.2737
geo_xQuAD	MinMax	0.2571	0.2838	0.2854	**0.2880**	0.2804	0.2786
	Sum	0.2629	**0.2949**	0.2822	0.2872	0.2855	0.2810
	Virt	0.2799	**0.2902**	0.2902	0.2811	0.2728	0.2731
PM2	MinMax	0.2702	**0.2820**	0.2715	0.2667	0.2602	0.2586
	Sum	0.2597	**0.2836**	0.2729	0.2741	0.2759	0.2773
	Virt	0.2722	**0.2837**	0.2750	0.2680	0.2723	0.2760
mix_CombSUM	MinMax	0.2576	0.2698	0.2697	**0.2725**	0.2651	0.2695
	Sum	0.2629	**0.2914**	0.2799	0.2859	0.2848	0.2808
	Virt	0.2761	**0.2809**	0.2798	0.2666	0.2675	0.2709

A.5.3 2009 sub-topics from suggestions

Table A.31: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.2814	**0.2877**	0.2777	0.2530	0.2475
	Sum	**0.2688**	0.2497	0.2462	0.2319	0.2447
	Virtual	0.2675	**0.2723**	0.2550	0.2437	0.2482
xQuAD	MinMax	**0.2960**	0.2877	0.2956	0.2849	0.2923
	Sum	0.2846	**0.2911**	0.2857	0.2828	0.2900
	Virtual	0.2887	0.2892	**0.2948**	0.2859	0.2916
art_xQuAD	MinMax	**0.3049**	0.2921	0.2996	0.2905	0.2995
	Sum	0.2860	0.2894	0.2842	0.2821	**0.2908**
	Virtual	0.2948	**0.2965**	0.2962	0.2905	0.2963
geo_xQuAD	MinMax	0.3019	0.2987	**0.3050**	0.2890	0.2956
	Sum	0.2860	0.2894	0.2842	0.2821	**0.2908**
	Virtual	0.2957	0.2964	0.2965	0.2916	**0.2965**
PM2	MinMax	**0.2935**	0.2638	0.2691	0.2458	0.2507
	Sum	**0.2775**	0.2710	0.2676	0.2436	0.2471
	Virtual	**0.2889**	0.2650	0.2560	0.2382	0.2446
mix_CombSUM	MinMax	**0.3043**	0.2950	0.2928	0.2885	0.2954
	Sum	0.2860	0.2889	0.2840	0.2817	**0.2901**
	Virtual	**0.2959**	0.2832	0.2878	0.2870	0.2924
mix_Borda		**0.2894**	0.2757	0.2796	0.2824	0.2861
mix_SV		0.2757	**0.2845**	0.2757	0.2777	0.2773
mix_MC2		**0.2858**	0.2657	0.2706	0.2531	0.2569

Table A.32: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25		KLD	
			Add 5	Fix 10	Add 5	Fix 10
IA-Select	MinMax	0.1778	**0.1829**	0.1774	0.1544	0.1508
	Sum	**0.1806**	0.1627	0.1558	0.1477	0.1593
	Virt	0.1744	**0.1786**	0.1646	0.1510	0.1571
xQuAD	MinMax	**0.2081**	0.1829	0.2015	0.1927	0.1998
	Sum	0.1941	**0.1976**	0.1924	0.1879	0.1911
	Virt	0.1985	0.1928	**0.2038**	0.1939	0.1989
art_xQuAD	MinMax	**0.2088**	0.1950	0.2016	0.1951	0.2036
	Sum	0.1946	**0.1967**	0.1927	0.1880	0.1915
	Virt	0.2010	0.2023	**0.2034**	0.1966	0.2018
geo_xQuAD	MinMax	0.1968	0.1915	**0.2042**	0.1946	0.2023
	Sum	0.1946	**0.1967**	0.1927	0.1880	0.1915
	Virt	0.2016	0.2021	**0.2040**	0.1971	0.2019
PM2	MinMax	**0.1984**	0.1745	0.1748	0.1563	0.1579
	Sum	**0.1862**	0.1753	0.1731	0.1556	0.1591
	Virt	**0.1955**	0.1697	0.1650	0.1501	0.1539
mix_CombSUM	MinMax	**0.2100**	0.2026	0.2011	0.1940	0.2000
	Sum	0.1946	**0.1966**	0.1926	0.1877	0.1911
	Virt	**0.2022**	0.1920	0.1965	0.1952	0.1985
mix_Borda		**0.1960**	0.1878	0.1806	0.1927	0.1949
mix_SV		0.1878	**0.1925**	0.1878	0.1818	0.1804
mix_MC2		**0.1914**	0.1689	0.1759	0.1621	0.1645

Table A.33: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25		KLD	
			Add 5	Fix 10	Add 5	Fix 10
IA-Select	MinMax	0.1778	**0.2030**	0.1981	0.1520	0.1721
	Sum	0.1806	**0.1988**	0.1872	0.1666	0.1829
	Virt	0.1744	**0.2100**	0.1923	0.1720	0.1856
xQuAD	MinMax	0.2081	**0.2084**	0.1981	0.1979	0.1931
	Sum	0.1941	**0.2153**	0.2082	0.1926	0.1954
	Virt	0.1985	**0.2249**	0.2218	0.2011	0.2082
art_xQuAD	MinMax	**0.2088**	0.2072	0.2062	0.2052	0.1970
	Sum	0.1946	**0.2144**	0.2078	0.1920	0.1947
	Virt	0.2010	0.2219	**0.2224**	0.2080	0.2003
geo_xQuAD	MinMax	0.1968	**0.2078**	0.2066	0.2041	0.1989
	Sum	0.1946	**0.2144**	0.2078	0.1920	0.1947
	Virt	0.2016	0.2219	**0.2231**	0.2084	0.2008
PM2	MinMax	0.1984	0.2012	**0.2048**	0.1720	0.1861
	Sum	0.1862	**0.2078**	0.1936	0.1737	0.1835
	Virt	0.1955	**0.2128**	0.2024	0.1835	0.1953
mix_CombSUM	MinMax	**0.2100**	0.2055	0.2049	0.2019	0.1958
	Sum	0.1946	**0.2146**	0.2078	0.1930	0.1946
	Virt	0.2022	**0.2156**	0.2081	0.2009	0.1973
mix_Borda		0.1960	**0.1998**	0.1957	0.1947	0.1932
mix_SV		0.1878	**0.2057**	0.1878	0.1878	0.1877
mix_MC2		0.1914	**0.1974**	0.1905	0.1791	0.1834

Table A.34: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.2814	0.3089	0.3120	**0.3134**	0.3090
	Sum	0.2688	0.2883	0.2883	0.2848	**0.2888**
	Virtual	0.2675	0.3020	0.3022	**0.3026**	0.3025
xQuAD	MinMax	0.2960	0.3102	0.3146	**0.3155**	0.3101
	Sum	0.2846	**0.3116**	0.3088	0.3018	0.3008
	Virtual	0.2887	**0.3231**	0.3223	0.3202	0.3182
art_xQuAD	MinMax	0.3049	**0.3071**	0.3063	0.3045	0.3023
	Sum	0.2860	**0.3090**	0.3064	0.3034	0.2981
	Virtual	0.2948	**0.3162**	0.3145	0.3103	0.3074
geo_xQuAD	MinMax	0.3019	0.3131	0.3178	**0.3193**	0.3093
	Sum	0.2860	**0.3090**	0.3064	0.3034	0.2981
	Virtual	0.2957	**0.3160**	0.3151	0.3109	0.3085
PM2	MinMax	0.2935	0.2928	**0.3084**	0.2976	0.2906
	Sum	0.2775	0.3007	**0.3065**	0.2915	0.3001
	Virtual	0.2889	0.2994	**0.3048**	0.2922	0.2950
mix_CombSUM	MinMax	**0.3043**	0.2958	0.2948	0.2950	0.2956
	Sum	0.2860	**0.3092**	0.3064	0.3033	0.2983
	Virtual	0.2959	**0.3079**	0.2996	0.2941	0.2970

Table A.35: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.1778	0.2030	0.2083	**0.2102**	0.2055
	Sum	0.1806	0.1988	0.1990	0.1969	**0.2043**
	Virt	0.1744	**0.2100**	0.2076	0.2077	0.2090
xQuAD	MinMax	0.2081	0.2084	0.2140	**0.2151**	0.2087
	Sum	0.1941	**0.2153**	0.2148	0.2112	0.2095
	Virt	0.1985	**0.2249**	0.2179	0.2168	0.2153
art_xQuAD	MinMax	0.2088	0.2072	0.2075	**0.2110**	0.2001
	Sum	0.1946	**0.2144**	0.2141	0.2105	0.2086
	Virt	0.2010	**0.2219**	0.2175	0.2175	0.2112
geo_xQuAD	MinMax	0.1968	0.2078	0.2125	**0.2171**	0.2077
	Sum	0.1946	**0.2144**	0.2141	0.2105	0.2086
	Virt	0.2016	**0.2219**	0.2180	0.2179	0.2120
PM2	MinMax	0.1984	0.2012	**0.2207**	0.2120	0.2000
	Sum	0.1862	0.2078	**0.2185**	0.2067	0.2144
	Virt	0.1955	0.2128	**0.2185**	0.2034	0.2082
mix_CombSUM	MinMax	**0.2100**	0.2055	0.2047	0.2047	0.2050
	Sum	0.1946	**0.2146**	0.2141	0.2104	0.2078
	Virt	0.2022	**0.2156**	0.2077	0.2045	0.2048

Table A.36: Diversification performance (α-nDCG@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by** VScrAvg	**Selective by** VScrFirst
IA-Select	MinMax	0.2814	**0.3089**	0.3085	0.3083
	Sum	0.2688	**0.2883**	0.2830	0.2852
	Virtual	0.2675	0.3020	0.3026	**0.3057**
xQuAD	MinMax	0.2960	**0.3102**	0.3085	0.3083
	Sum	0.2846	**0.3116**	0.3013	0.3013
	Virtual	0.2887	**0.3231**	0.3121	0.3165
art_xQuAD	MinMax	0.3049	0.3071	0.3078	**0.3124**
	Sum	0.2860	**0.3090**	0.2993	0.3011
	Virtual	0.2948	**0.3162**	0.3091	0.3109
geo_xQuAD	MinMax	0.3019	0.3131	0.3114	**0.3167**
	Sum	0.2860	**0.3090**	0.2993	0.3011
	Virtual	0.2957	**0.3160**	0.3093	0.3118
PM2	MinMax	0.2935	0.2928	**0.3017**	0.3000
	Sum	0.2775	**0.3007**	0.2918	0.2919
	Virtual	0.2889	**0.2994**	0.2932	0.2938
mix_CombSUM	MinMax	**0.3043**	0.2958	0.2954	0.2946
	Sum	0.2860	**0.3092**	0.2993	0.3010
	Virtual	0.2959	**0.3079**	0.3019	0.3037
mix_Borda		0.2894	**0.2936**	0.2878	0.2918
mix_SV		0.2757	0.2983	**0.3037**	0.2965
mix_MC2		0.2858	0.2919	**0.3007**	0.2945

Table A.37: Diversification performance (ERR-IA@20) of the algorithms on TREC 2009 topics using original query aspects obtained from the suggestions and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by** VScrAvg	VScrFirst
IA-Select	MinMax	0.1778	0.2030	**0.2060**	0.2052
	Sum	0.1806	**0.1988**	0.1948	0.1984
	Virt	0.1744	**0.2100**	0.2043	0.2090
xQuAD	MinMax	0.2081	**0.2084**	0.2060	0.2052
	Sum	0.1941	**0.2153**	0.2110	0.2111
	Virt	0.1985	**0.2249**	0.2145	0.2191
art_xQuAD	MinMax	0.2088	0.2072	0.2118	**0.2154**
	Sum	0.1946	**0.2144**	0.2102	0.2107
	Virt	0.2010	**0.2219**	0.2132	0.2126
geo_xQuAD	MinMax	0.1968	0.2078	0.2132	**0.2168**
	Sum	0.1946	**0.2144**	0.2102	0.2107
	Virt	0.2016	**0.2219**	0.2120	0.2135
PM2	MinMax	0.1984	0.2012	0.2043	**0.2065**
	Sum	0.1862	**0.2078**	0.2030	0.2040
	Virt	0.1955	**0.2128**	0.2040	0.2071
mix_CombSUM	MinMax	**0.2100**	0.2055	0.2042	0.2023
	Sum	0.1946	**0.2146**	0.2102	0.2106
	Virt	0.2022	**0.2156**	0.2090	0.2106
mix_Borda		0.1960	**0.1998**	0.1916	0.1991
mix_SV		0.1878	0.2057	**0.2115**	0.2056
mix_MC2		0.1914	0.1974	**0.2040**	0.2029

A.5.4 2010 sub-topics from suggestions

Table A.38: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.2952	**0.3109**	0.3009	0.2722	0.2842
	Sum	**0.3043**	0.2909	0.2864	0.2568	0.2682
	Virtual	**0.3046**	0.2973	0.2952	0.2831	0.2751
xQuAD	MinMax	0.3072	**0.3225**	0.3116	0.2924	0.3045
	Sum	**0.3215**	0.3051	0.2990	0.2810	0.2857
	Virtual	0.3090	**0.3161**	0.3130	0.2959	0.3127
art_xQuAD	MinMax	0.3225	0.3246	**0.3261**	0.2957	0.3049
	Sum	**0.3225**	0.3042	0.2981	0.2788	0.2809
	Virtual	**0.3210**	0.3145	0.3114	0.2935	0.3005
geo_xQuAD	MinMax	0.3228	**0.3305**	0.3246	0.2947	0.3083
	Sum	**0.3225**	0.3042	0.2981	0.2788	0.2809
	Virtual	**0.3194**	0.3138	0.3108	0.2931	0.3015
PM2	MinMax	0.3129	**0.3169**	0.3107	0.2828	0.2944
	Sum	0.3129	**0.3186**	0.3119	0.2900	0.3000
	Virtual	0.3107	**0.3148**	0.3065	0.2847	0.2888
mix_CombSUM	MinMax	**0.3145**	0.3024	0.3007	0.2807	0.2857
	Sum	**0.3224**	0.3007	0.2981	0.2788	0.2815
	Virtual	**0.3256**	0.3064	0.2987	0.2839	0.2874
mix_Borda		**0.3190**	0.3184	0.3161	0.3015	0.3107
mix_SV		**0.3179**	0.3176	0.3083	0.3058	0.3079
mix_MC2		0.3081	**0.3105**	0.3064	0.3001	0.2831

Table A.39: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using candidate re-rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25		KLD	
			Add 5	Fix 10	Add 5	Fix 10
IA-Select	MinMax	0.1978	**0.2223**	0.2103	0.1843	0.1946
	Sum	0.2116	**0.2163**	0.2069	0.1715	0.1892
	Virt	0.2094	**0.2150**	0.2107	0.1923	0.1941
xQuAD	MinMax	0.2172	**0.2378**	0.2303	0.2039	0.2155
	Sum	**0.2301**	0.2247	0.2187	0.1948	0.2032
	Virt	0.2143	**0.2324**	0.2267	0.2032	0.2190
art_xQuAD	MinMax	0.2272	0.2348	**0.2384**	0.2050	0.2167
	Sum	**0.2306**	0.2244	0.2179	0.1947	0.1967
	Virt	0.2227	**0.2318**	0.2260	0.2071	0.2170
geo_xQuAD	MinMax	0.2251	**0.2435**	0.2325	0.2042	0.2166
	Sum	**0.2306**	0.2244	0.2179	0.1947	0.1967
	Virt	0.2228	**0.2313**	0.2253	0.2073	0.2177
PM2	MinMax	0.2134	**0.2350**	0.2233	0.1984	0.2138
	Sum	0.2220	**0.2380**	0.2203	0.2047	0.2124
	Virt	0.2150	**0.2363**	0.2216	0.2025	0.2039
mix_CombSUM	MinMax	0.2196	0.2235	**0.2237**	0.1946	0.2073
	Sum	**0.2304**	0.2239	0.2178	0.1947	0.1966
	Virt	**0.2316**	0.2279	0.2178	0.2003	0.2054
mix_Borda		0.2269	**0.2399**	0.2341	0.2185	0.2249
mix_SV		**0.2348**	0.2346	0.2328	0.2259	0.2272
mix_MC2		0.2153	**0.2251**	0.2220	0.2168	0.1924

Table A.40: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their expansions using sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	BM25 Add 5	BM25 Fix 10	KLD Add 5	KLD Fix 10
IA-Select	MinMax	0.1978	0.2175	**0.2219**	0.1950	0.2155
	Sum	0.2116	0.2277	**0.2346**	0.2056	0.2016
	Virt	0.2094	0.2315	**0.2316**	0.1978	0.2141
xQuAD	MinMax	0.2172	0.2314	**0.2468**	0.2195	0.2254
	Sum	0.2301	0.2374	**0.2378**	0.2174	0.2250
	Virt	0.2143	0.2411	**0.2519**	0.2137	0.2347
art_xQuAD	MinMax	0.2272	0.2459	**0.2577**	0.2258	0.2358
	Sum	0.2306	**0.2380**	0.2377	0.2160	0.2165
	Virt	0.2227	0.2484	**0.2493**	0.2143	0.2370
geo_xQuAD	MinMax	0.2251	0.2415	**0.2554**	0.2241	0.2336
	Sum	0.2306	**0.2380**	0.2377	0.2160	0.2165
	Virt	0.2228	0.2481	**0.2496**	0.2146	0.2359
PM2	MinMax	0.2134	**0.2394**	0.2381	0.2146	0.2217
	Sum	0.2220	0.2318	**0.2406**	0.2110	0.2176
	Virt	0.2150	0.2298	**0.2447**	0.2148	0.2257
mix_CombSUM	MinMax	0.2196	**0.2402**	0.2386	0.2236	0.2309
	Sum	0.2304	**0.2380**	0.2372	0.2166	0.2163
	Virt	0.2316	**0.2433**	0.2366	0.2188	0.2355
mix_Borda		0.2269	**0.2569**	0.2564	0.2456	0.2379
mix_SV		0.2348	0.2547	**0.2651**	0.2449	0.2410
mix_MC2		0.2153	0.2368	**0.2487**	0.2224	0.2287

Table A.41: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.2952	0.3099	0.3044	0.3108	**0.3120**
	Sum	0.3043	0.3154	0.3143	0.3080	**0.3176**
	Virtual	0.3046	0.3223	**0.3248**	0.3197	0.3228
xQuAD	MinMax	0.3072	0.3240	0.3216	0.3225	**0.3244**
	Sum	0.3215	0.3245	0.3266	0.3220	**0.3267**
	Virtual	0.3090	**0.3392**	0.3367	0.3294	0.3369
art_xQuAD	MinMax	0.3225	**0.3465**	0.3387	0.3361	0.3421
	Sum	0.3225	0.3243	0.3261	0.3197	**0.3278**
	Virtual	0.3210	**0.3386**	0.3323	0.3264	0.3339
geo_xQuAD	MinMax	0.3228	**0.3368**	0.3333	0.3318	0.3352
	Sum	0.3225	**0.3243**	0.3261	0.3197	0.3278
	Virtual	0.3194	**0.3387**	0.3331	0.3270	0.3336
PM2	MinMax	0.3129	**0.3333**	0.3258	0.3283	0.3327
	Sum	0.3129	**0.3226**	0.3175	0.3181	0.3207
	Virtual	0.3107	**0.3277**	0.3185	0.3196	0.3181
mix_CombSUM	MinMax	0.3145	**0.3291**	0.3174	0.3182	0.3238
	Sum	0.3224	0.3245	0.3262	0.3196	**0.3278**
	Virtual	0.3256	**0.3273**	0.3235	0.3179	0.3218

Table A.42: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using the aspect weights assigned by the QPP methods for the original query aspects obtained from the suggestions and their expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expansion using BM25 Add 5 terms			
			Uniform	VScrFirst	VScrAvg	ScrRatio
IA-Select	MinMax	0.1978	0.2175	0.2107	0.2171	**0.2180**
	Sum	0.2116	0.2277	0.2317	0.2260	**0.2323**
	Virt	0.2094	0.2315	**0.2332**	0.2264	0.2301
xQuAD	MinMax	0.2172	0.2314	0.2299	0.2320	**0.2328**
	Sum	0.2301	0.2374	**0.2440**	0.2375	0.2400
	Virt	0.2143	**0.2411**	0.2372	0.2301	0.2368
art_xQuAD	MinMax	0.2272	**0.2459**	0.2428	0.2445	0.2447
	Sum	0.2306	0.2380	**0.2435**	0.2372	0.2424
	Virt	0.2227	**0.2484**	0.2451	0.2338	0.2443
geo_xQuAD	MinMax	0.2251	**0.2415**	0.2390	0.2399	0.2408
	Sum	0.2306	0.2380	**0.2435**	0.2372	0.2424
	Virt	0.2228	**0.2481**	0.2456	0.2330	0.2438
PM2	MinMax	0.2134	0.2394	0.2316	0.2355	**0.2396**
	Sum	0.2220	0.2318	0.2301	0.2335	**0.2347**
	Virt	0.2150	0.2298	0.2320	**0.2338**	0.2302
mix_CombSUM	MinMax	0.2196	**0.2402**	0.2294	0.2288	0.2335
	Sum	0.2304	0.2380	**0.2436**	0.2371	0.2424
	Virt	0.2316	**0.2433**	0.2389	0.2338	0.2360

Table A.43: Diversification performance (α-nDCG@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their selective expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by**	
				VScrAvg	VScrFirst
IA-Select	MinMax	0.2952	0.3099	**0.3123**	0.3100
	Sum	0.3043	0.3154	0.3132	**0.3205**
	Virtual	0.3046	0.3223	**0.3316**	0.3281
xQuAD	MinMax	0.3072	**0.3240**	0.3141	0.3125
	Sum	0.3215	0.3245	0.3244	**0.3257**
	Virtual	0.3090	**0.3392**	0.3316	0.3331
art_xQuAD	MinMax	0.3225	**0.3465**	0.3346	0.3303
	Sum	0.3225	0.3243	0.3211	**0.3270**
	Virtual	0.3210	**0.3386**	0.3292	0.3309
geo_xQuAD	MinMax	0.3228	**0.3368**	0.3253	0.3234
	Sum	0.3225	0.3243	0.3211	**0.3270**
	Virtual	0.3194	**0.3387**	0.3296	0.3309
PM2	MinMax	0.3129	**0.3333**	0.3262	0.3278
	Sum	0.3129	0.3226	0.3153	**0.3249**
	Virtual	0.3107	**0.3277**	0.3196	0.3253
mix_CombSUM	MinMax	0.3145	**0.3291**	0.3185	0.3156
	Sum	0.3224	0.3245	0.3211	**0.3269**
	Virtual	0.3256	0.3273	0.3248	**0.3276**
mix_Borda		0.3190	0.3425	0.3347	**0.3484**
mix_SV		0.3179	**0.3464**	0.3235	0.3330
mix_MC2		0.3081	**0.3328**	0.3241	0.3308

Table A.44: Diversification performance (ERR-IA@20) of the algorithms on TREC 2010 topics using original query aspects obtained from the suggestions and their <u>selective</u> expansions by adding 5 expansion terms calculated with BM25 term ranking function on sub-topic's own rankings as PRF. The highest score is boldfaced.

Div. method	Relevance norm.	Original	Expand All	**Selective by** VScrAvg	**Selective by** VScrFirst
IA-Select	MinMax	0.1978	**0.2175**	0.2158	0.2153
	Sum	0.2116	0.2277	0.2256	**0.2347**
	Virt	0.2094	0.2315	0.2298	**0.2328**
xQuAD	MinMax	0.2172	**0.2314**	0.2234	0.2241
	Sum	0.2301	0.2374	0.2373	**0.2377**
	Virt	0.2143	**0.2411**	0.2298	0.2382
art_xQuAD	MinMax	0.2272	**0.2459**	0.2374	0.2322
	Sum	0.2306	0.2380	0.2364	**0.2394**
	Virt	0.2227	**0.2484**	0.2388	0.2418
geo_xQuAD	MinMax	0.2251	**0.2415**	0.2313	0.2299
	Sum	0.2306	0.2380	0.2364	**0.2394**
	Virt	0.2228	**0.2481**	0.2389	0.2418
PM2	MinMax	0.2134	**0.2394**	0.2280	0.2281
	Sum	0.2220	0.2318	0.2238	**0.2354**
	Virt	0.2150	0.2298	0.2276	**0.2390**
mix_CombSUM	MinMax	0.2196	**0.2402**	0.2278	0.2234
	Sum	0.2304	0.2380	0.2365	**0.2394**
	Virt	0.2316	**0.2433**	0.2352	0.2396
mix_Borda		0.2269	0.2569	0.2470	**0.2609**
mix_SV		0.2348	**0.2547**	0.2369	0.2418
mix_MC2		0.2153	**0.2368**	0.2288	0.2305

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Özdemiray, Ahmet Murat

Nationality: Turkish (TC)

Date and Place of Birth: 22.09.1982, Konya

Marital Status: Married

Phone: +905326680775

EDUCATION

Degree	**Institution**	**Year of Grad.**
M.S.	Dept. of Computer Eng., Bilkent University	2008
B.S.	Dept. of Computer Eng., Bilkent University	2005
High School	Konya Meram Anadolu Lisesi	2000

PROFESSIONAL EXPERIENCE

Year	**Place**	**Enrollment**
2005 - Present	TUBITAK BILGEM ILTAREN	Chief Researcher

PUBLICATIONS

Journal Publications

A. M. Ozdemiray and I. S. Altingovde. Explicit search result diversification using score and rank aggregation methods. *Journal of the Association for Information Sci-*

ence and Technology, 66(6):1212-1228, 2015.

Conference Publications

A. M. Ozdemiray and I. S. Altingovde. Query Performance Prediction for Aspect Weighting in Search Result Diversification. In: *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM 2014, pages 1871-1874. ACM, 2014.