UNIVERSITÀ DEGLI STUDI DI UDINE
Dipartimento di Scienze Matematiche, Informatiche e Fisiche
ALPEN-ADRIA-UNIVERSITÄT KLAGENFURT
Faculty of Technical Sciences
Corso di Laurea Magistrale in Informatica Internazionale
Joint Degree in International Informatics

Master Thesis

Improving the Efficiency of Retrieval Effectiveness Evaluation: Finding a Few Good Topics with Clustering and Dimensionality Reduction

Supervisor:
Prof. Stefano MIZZARO

Candidate:
Kevin ROITERO

Co-supervisor:
Prof. Klaus SCHOEFFMANN

Academic year 2015-2016
Abstract

We consider the issue of using fewer topics in the effectiveness evaluation of information retrieval systems. Previous work has shown that using fewer topics is theoretically possible: there exist smaller topic sets that evaluate a population of systems in almost the same way as a full and larger topic set. One of the main issues that remains to be solved is how to find such a small set of a few good topics. To this aim, in this thesis we use a novel approach based on clustering of topics together with dimensionality reduction.

We present various approaches: we consider an a posteriori approach, i.e., we build clusters of topics only after the evaluation exercise (and the relevance assessments) has already been performed; an a priori approach, i.e., before any human assessment has been performed, and thus without using the features produced during assessment; and some approaches that occur during the evaluation process.

We show that clustering is effective in the topic set reduction problem, if the a posteriori approach is considered together with dimensionality reduction (i.e., principal component analysis). We provide furthermore a statistical significance analysis of the obtained result.

We also consider a priori features and features that occur during the evaluation process. The results show that the a priori approaches and the approaches that occur during the evaluation approaches considered are comparable to the random choice of topics. Finally, we provide convincing evidence that straightforward clustering (i.e., without the dimensionality reduction) is not effective in the topic reduction problem.
# Contents

1 Introduction .................................................. 1
   1.1 Information Retrieval ........................................ 1
   1.2 Aims of the Thesis ........................................... 4
   1.3 Structure of the Thesis ....................................... 5
   1.4 Publications ................................................. 6

I Background ................................................. 7

2 Fundamentals ................................................. 9
   2.1 The Relevance Concept ....................................... 9
   2.2 Metrics in IR ................................................ 13
      2.2.1 Precision ............................................... 14
      2.2.2 Recall ................................................ 14
      2.2.3 Precision Recall Curve ................................ 15
      2.2.4 Average Precision ...................................... 15
      2.2.5 Mean Average Precision (MAP) .......................... 16
      2.2.6 Geometric Mean Average Precision (GMAP) .......... 16
      2.2.7 Logistic Average Precision (logitAP) ................. 17
      2.2.8 Yet Another Average Precision (yaAP) ............... 18
      2.2.9 P@n ................................................ 18
      2.2.10 R-Precision ............................................ 19
   2.3 Evaluation of Effectiveness By Means of User Studies .... 19
   2.4 Evaluation of Effectiveness By Means of Test Collections .. 20
   2.5 Conclusions .................................................. 22

3 Fewer Topics ................................................. 23
   3.1 Background .................................................. 23
   3.2 First Work – A Few Good Topics ............................ 24
   3.3 Second Work – On Using Fewer Topics ..................... 29
   3.4 Conclusions ................................................... 31
# II  Clustering of Topics – A Posteriori

4  Aims and Methodology
   4.1  Terms and Notation ............................... 35
   4.2  Research Questions ............................... 36
   4.3  Data ............................................. 38
      4.3.1  Ad Hoc Track (TREC-8) ................... 39
      4.3.2  Terabyte Track (TREC 2006) ............. 39
      4.3.3  Robust Track (TREC 2004) ............... 41
      4.3.4  Million Query Track (TREC 2007) ....... 42
      4.3.5  Document Pool ............................. 42
   4.4  Conclusions .................................. 43

5  First Experiment: Clustering Using System Effectiveness
   5.1  Intuition and Description of the Experiment .... 45
   5.2  Coverage Measure ............................... 48
      5.2.1  Definition ................................ 48
      5.2.2  Coverage of best, worst, and average topic sets .... 49
   5.3  Cluster of Topics vs. Random Choice ............ 52
   5.4  Discussion and Conclusions .................... 54

6  Curse of Dimensionality
   6.1  Introduction .................................. 57
   6.2  Curse of Dimensionality Using Space Comparison .. 57
   6.3  Curse of Dimensionality Using Vector Comparison .. 60
   6.4  The Effect of Curse of Dimensionality .......... 62
   6.5  Conclusions .................................. 62

7  Dimensionality Reduction
   7.1  Principal Component Analysis (PCA) ............. 63
   7.2  Principal Component Analysis with System Effectiveness .... 68
   7.3  Discussion and Conclusions .................... 71

8  Statistical Significance
   8.1  Sampling Distribution ........................... 79
   8.2  Statistical Significance – Mann Whitney U test .... 80
   8.3  Discussion and Conclusions .................... 83
CONTENTS

III Clustering of Topics – A Priori 87

9 Second Experiment: Clustering Using Topic Title 89
  9.1 Intuition and Description of the Experiment ............... 89
  9.2 Topic Structure ........................................ 91
  9.3 Term Frequency and Inverse Document Frequency ........... 92
  9.4 Cluster of Topics vs. Random Choice ...................... 94
  9.5 Discussion and Conclusions ............................ 96

10 Third Experiment: Clustering Using Full Topic Narration 99
  10.1 Intuition and Description of the Experiment ............... 99
  10.2 Cluster of Topics vs. Random Choice ...................... 100
  10.3 Discussion and Conclusions ............................ 102

11 Dimensionality Reduction – A Priori Approach 105
  11.1 Intuition and Experiment Description ...................... 105
  11.2 Discussion and Conclusions ............................ 107

12 Further Experiments with Clustering – Variation of Cluster
   Parameters and Features 109
  12.1 Varying Clustering Parameters ........................ 109
    12.1.1 The Linkage Method .............................. 110
    12.1.2 The Cluster Algorithm ........................... 110
    12.1.3 The Distance and the Similarity Measure .......... 110
  12.2 More Clustering Features ............................... 111
    12.2.1 Identifiers of Documents in the Pool ............... 111
    12.2.2 Identifiers of Relevant Documents in the Pool ...... 111
    12.2.3 Number of Documents in the Pool .................. 112
    12.2.4 Number of Relevant Documents in the Pool .......... 112
    12.2.5 Percentage of Relevant Documents in the Pool ...... 112
    12.2.6 Final Set of Experiments .......................... 113
  12.3 Discussion and Conclusions ............................ 113

IV Discussion and Future Developments 115

13 Future Developments 117
  13.1 Vary the Number of Clusters ............................ 117
  13.2 More Samples for Each Cardinality ....................... 117
  13.3 Consider Centroids and Medoids ........................ 118
  13.4 Cluster on the basis of HITS Algorithm .................. 118
  13.5 Entropy of the Documents .............................. 118
13.6 Hyper-Graphs ........................................ 119
13.7 Classifier ........................................... 119
13.8 Generalization ....................................... 119
13.9 Discussion and Conclusions ....................... 119

14 Summary ................................................. 121
  14.1 Overview .......................................... 121
  14.2 The Answer to the Research Questions .......... 122
  14.3 Conclusions ....................................... 123

Acknowledgments ......................................... 125

Bibliography ................................................ 127