Multi-Document Text Summarization for Competitor Intelligence: A Framework

A THESIS
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE **FELLOW PROGRAMME IN MANAGEMENT (INDUSTRY)**,
INDIAN INSTITUTE OF MANAGEMENT, INDORE

By

Swapnajit Chakraborti
FPM (Industry) 2013
IIM Indore, India
MARCH, 2017

Thesis Advisory Committee

1. -- [Chair]
[Prof. Shubhamoy Dey]

2. -- [Member]
[Prof. D. L. Sunder]

3. -- [Member]
[Prof. Rajhans Mishra]

Abstract

Proliferation of web as an easily accessible information resource has led many business organizations to gather competitor intelligence (CI) directly from various resources on the internet, namely, news, blogs, reports, reviews etc. While collection of such information is easy from the internet, the collation and structuring of them for perusal of business decision makers, is not very straight-forward. This work is an endeavor towards exploring the effectiveness and feasibility of text clustering followed by text summarization as a technique for comprehending competitor intelligence and thereby use it as an input for strategic decision making by the managers. Although applications of various text clustering and text summarization techniques are abundant in state-of-the-art literature, there have been very few authentic research on understanding their impact in organizing competitor intelligence for managerial decision making. This research contributes towards developing an integrated framework which supports analysis of competitor information by clustering the corpus collected from internet using a variation of K-means, called Multi-level K-means clustering, and then generating the extractive summaries of these topical clusters by applying global optimization techniques, such as, Artificial Bee Colony (ABC) optimization, Differential Evolution (DE) etc. The results of analysis of the quality of these system generated summaries, measured w.r.t. human generated golden summaries, are also presented in this thesis. As effectiveness of the system generated summaries towards competitor intelligence analysis, can be best judged by the practicing managers at various organizations, a study is also undertaken to obtain managerial feedback on these automatically generated system summaries and corresponding findings are reported.

List of Figures

Figure Detail	Page Number
Fig. 1. Growth in Total Number of Websites over Last Two Decades	20
Fig. 2. Porter’s Five Forces for Industry Analysis for Strategic Decision Making	22
Fig. 3. Value Chain Framework of Organizational Activities for Delivery of Consumer Value	24
Fig. 4. Various Aspects of Information required in Competitive Intelligence Process	26
Fig. 5. Classification According to Information Types (Rouach, 2001)	28
Fig. 6. Competitor Intelligence System	36
Fig. 7: Motivating factors for Research on CI Framework	67
Fig. 8: Architecture of the Proposed Summarization Framework for CI	71
Fig. 9: User view of the proposed framework	72
Fig. 10: Text Processing Stages of proposed framework	78
Fig. 11: Zipf’s Law on CI corpus	85
Fig. 12: Log-log plot of rank vs frequency of words of CI corpus	85
Fig. 13: Output of Elbow Method	88
Fig. 14: Gap Analysis Output	89
Fig. 15: Comparison of average APS (CI corpus)	100
Fig. 16: Number of levels for ML-KM with CI corpus	102
Fig. 17: Comparison of average APS (Reuters)	107
Fig. 18: Algorithm for Computing Total Penalty of a Candidate Summary	127
Fig. 19: Integer vector of a candidate summary	131

Fig. 20: MMR Computation formula for selecting sentences	132
Fig. 21: Basic algorithm of ABC optimization	133
Fig. 22: Initialization step of ABC algorithm (sn is population size & D is solution size)	134
Fig. 23: Employed Bee Phase	135
Fig. 24: Onlooker Bee Phase	136
Fig. 25: Formula for trial vector creation in DE	138
Fig. 26: Basic stages of Genetic Algorithm	140
Fig 27: Plot of Recall scores for ABC	151
Fig 28: Plot of Recall scores for DE	152
Fig 29: Plot of Recall scores for GA	152
Fig 30: Plot of Recall scores for MMR	153
Fig 31: Plot of Precision scores for ABC	165
Fig 32: Plot of Precision scores for DE	165
Fig 33: Plot of Precision scores for GA	166
Fig 34: Plot of Precision scores for MMR	166
Fig. 35: Analysis of Cohort-1 Responses	206
Fig. 36: Analysis of Cohort-2 Responses	208
Fig. 37: Analysis of Cohort-3 Responses	209
Fig. 38: Analysis of Cohort-1 Responses	211
Fig. 39: Analysis of Cohort-2 Responses	212
Fig. 40: Analysis of Cohort-3 Responses	213
Fig. 41: Analysis of Overall Responses	216
Fig. 42: Analysis of Overall Responses	219

Fig. 43: Directory structure after ML-KM run on a corpus showing the topical clusters	237
Fig. 44: Creation of "summary" folder after Summarization run	237
Fig. 45: Content of "summary" folder created by Summarization Process	238
Fig. 46: Snippet of the summary generated by optimization process	238
Fig. 47: Wordmap of Topical Cluster TOPIC_150_Cluster_1 generated by ML-KM	242
Fig. 48: Wordmap of System Summary of Topical Cluster TOPIC_150_Cluster_1 generated by ABC optimization	243

List of Tables

Table detail	Page No
Table 1: Definitions of Competitive Intelligence and Competitor Intelligence	43
Table 2: Definitions of TS or ATS	50
Table 3: Major Text Summarization Techniques	58
Table 4: Major Text Clustering Techniques	62
Table 5: Corpora derived using various values of number of latent topics in LSA	83
Table 6: List of corpus specific clusters generated by ML-KM Clustering	97
Table 7. Result of ML-KM Clustering on CI corpus	99
Table 8. Initial and Final number of clusters in ML-KM	101
Table 9. Number of small clusters with CI corpus	103
Table 10. Comparison of Average APS for 10 runs for 10-topic representation	104
Table 11: Result of Independent Sample t-test	104
Table 12: Result of Paired sample t-test	105
Table 13. Result of ML-KM Clustering on Reuters corpus	106
Table 14. Number of small clusters with Reuters corpus	108
Table 15. Initial and Final number of clusters in ML-KM	108
Table 16. Comparison of Average APS for 10 runs for 10-topic representation	109
Table 17: Results of Independent Sample t-test	109

Table 18: Result of Paired sample t-test	110
Table 19: APS vs Human Cluster Scores	112
Table 20: Results of Correlation Analysis of APS vs Human Cluster Score	114
Table 21: Results of Correlation Analysis of APS vs Human Cluster Score	114
Table 22: Parameter Values used in ABC optimization	137
Table 23: Parameter Values used in DE optimization	138
Table 24: Parameter Values used in GA optimization	140
Table 25: ROUGE-L Recall Scores for System Summaries of Topical Clusters	149
Table 26: Average ROUGE-L Recall Scores	153
Table 27: One-sample t-test for ROUGE-L Recall for ABC	154
Table 28: One-sample t-test for ROUGE-L Recall for DE	154
Table 29: One-sample t-test for ROUGE-L Recall for GA	155
Table 30. Paired Sample t-test Result for ROUGE-L Recall of ABC against MMR	156
Table 31. Paired Sample t-test Result for ROUGE-L Recall of DE against MMR	156
Table 32. Paired Sample t-test Result for ROUGE-L Recall of GA against MMR	157
Table 33: Result of Paired-sample t-test: ABC vs DE	158
Table 34: Result of Paired-sample t-test: ABC vs GA	158
Table 35: Result of Paired-sample t-test: DE vs ABC	158
Table 36: Result of Paired-sample t-test: DE vs GA	159
Table 37: Result of Paired-sample t-test: GA vs ABC	159

Table 38: Result of Paired-sample t-test: GA vs DE	159
Table 39: Correlation of ABC and DE recall scores	160
Table 40: Correlation of ABC and GA recall scores	161
Table 41: Correlation of DE and GA recall scores	161
Table 42: ROUGE-L Precision Scores for System Summaries of Topical Clusters	162
Table 43: Average ROUGE-L Precision Scores	167
Table 44: One-sample t-test for ROUGE-L Precision for ABC	168
Table 45: One-sample t-test for ROUGE-L Precision for DE	168
Table 46: One-sample t-test for ROUGE-L Precision for GA	168
Table 47: Paired Sample t-test Result for ROUGE-L Precision of ABC against MMR	170
Table 48: Paired Sample t-test Result for ROUGE-L Precision of DE against MMR	170
Table 49: Paired Sample t-test Result for ROUGE-L Precision of GA against MMR	171
Table 50: Result of Paired-sample t-test: ABC vs DE	172
Table 51: Result of Paired-sample t-test: ABC vs GA	172
Table 52: Result of Paired-sample t-test: DE vs ABC	173
Table 53: Result of Paired-sample t-test: DE vs GA	173
Table 54: Result of Paired-sample t-test: GA vs ABC	173
Table 55: Result of Paired-sample t-test: GA vs DE	174
Table 56: Correlation of ABC and DE precision scores	175
Table 57: Correlation of ABC and GA precision scores	175
Table 58: Correlation of ABC and DE precision scores	176
Table 59: APS and Recall Scores for ABC	177
Table 60: SPSS output of Correlation of APS and Recall	178

Table 61: Clusters with High APS and Low Recall	178
Table 62: Updated list of Clusters for Analysis of APS vs ABC Recall	179
Table 63: Result of Correlation Analysis of APS and Recall for ABC	180
Table 64: Clusters for Analysis of APS vs DE Recall	181
Table 65: Result of Correlation Analysis of APS and Recall for DE	182
Table 66: Clusters for Analysis of APS vs GA Recall	182
Table 67: Result of Correlation Analysis of APS and Recall for GA	183
Table 68: Clusters for Analysis of APS vs MMR Recall	184
Table 69: Result of Correlation Analysis of APS and Recall for MMR	185
Table 70: Clusters for Analysis of APS vs ABC Precision	186
Table 71: Result of Correlation Analysis of APS and Precision for ABC	187
Table 72: Clusters for Analysis of APS vs DE Precision	187
Table 73: Result of Correlation Analysis of APS and Precision for DE	188
Table 74: Clusters for Analysis of APS vs GA Precision	188
Table 75: Result of Correlation Analysis of APS and Precision for GA	189
Table 76: Clusters for Analysis of APS vs MMR Precision	190
Table 77: Result of Correlation Analysis of APS and Precision for MMR	191
Table 78: Clusters for Analysis Human Cluster Score vs ABC Recall Scores	192
Table 79: Result of Correlation Analysis	193
Table 80: Clusters for Analysis Human Cluster Score vs DE Recall Scores	193
Table 81: Result of Correlation Analysis	194
Table 82: ABC Recall vs Summary Scoring Function (Optimization Function)	195
Table 83: Result of Correlation Analysis	197

Table 84: DE Recall vs Summary Scoring Function (Optimization Function)	197
Table 85: Result of Correlation Analysis	199
Table 86: ROUGE-L Recall Scores for Long Runs	201
Table 87: Average values of Items for Cohort-1	207
Table 88: Average values of Items for Cohort-2	209
Table 89: Average values of Items for Cohort-3	210
Table 90: Results of Cohort-1 Analysis	214
Table 91: Results of Cohort-2 Analysis	214
Table 92: Results of Cohort-3 Analysis	215
Table 93: Average Values of Items	217
Table 94: SPSS output for One-sample t-test for Q1	218
Table 95: SPSS output for One-sample t-test for Q2	218
Table 96: SPSS output for One-sample t-test for Q3	218
Table 97: SPSS output for One-sample t-test for Q7	218
Table 98: SPSS output for One-sample t-test for Q4(A)	220
Table 99: SPSS output for One-sample t-test for Q4(B)	220
Table 100: Summary Scores assigned by all Managers	221
Table 101: SPSS output for One-sample t-test for Summary_A Score	222
Table 102: SPSS output for One-sample t-test for Summary_B Score	223
Table 103: SPSS output for paired-sample t-test	223

Glossary of Terms

Term	Meaning	Chapter Reference
LSA (Latent Semantic Analysis)	Finding out the hidden/latent topics within a corpus by applying matrix factorization	2
Centroid	Central theme of a text corpus represented by important words in the corpus	2
Topic Identification	Finding the topics from a corpus; done by clustering as well as probabilistic modelling	5
Corpus	Collection of text documents	4
TF-IDF	Term Frequency*Inverse Document Frequency	4
S-KM	Standard K-means (K-means++) clustering technique used as building block for Multi-level K-Means	5
APS	Average Pairwise Similarity; used as metric for cluster quality	5
AAPS	Average of APS of all clusters generated from a corpus	5
Extractive Summary	Summary generated by extracting sentences as it is from the corpus	6
System Summary	Summary generated automatically by applying optimization techniques	6
Golden Summary	Summary generated by human volunteer from a corpus	6
Total Penalty	Loss of similarity with central theme of a cluster computed by removing one sentence at a time from candidate system summary	6
ABC Optimization	Artificial Bee Colony Optimization; latest swarm optimization based on foraging behaviour of honey bees	6
DE Optimization	Differential Evolution based optimization	6
ROUGE	Recall Oriented Understudy of Gisting Evaluation; summary quality measurement tool	6
Recall Score	Fraction of golden summary that matches with system summary; measured using sequence overlap	6
Precision Score	Fraction of system summary that matches with golden summary; measured using sequence overlap	6
ML-KM	Multi-level K-means Clustering technique introduced for clustering CI corpus	5
CI	Competitor Intelligence	1
ATS	Automatic Text Summarization	1

TSS	Total Summary Score; Optimization function used to evaluate a candidate summary generated during optimization process	6

Table of Contents

1 Introduction 19

1.1 Growth of Internet 19

1.2 Competitor Intelligence & Strategic Decision Making 21

1.3 Information for Competitor Intelligence: Types and Sources 25

1.4 Process of Gathering Competitor Intelligence 31

1.5 Building an Effective Competitor Intelligence Information System 37

1.5.1 Creating a framework for CI Analysis 38

1.5.2 Adaption of existing clustering techniques 39

1.5.3 Application of Text Summarization 39

2 Literature Review 42

2.1 Competitor Intelligence 42

2.1.1 Definition of Competitor Intelligence 43

2.1.2 Types and Sources of Competitor Intelligence 45

2.1.3 Process of Gathering Competitor Intelligence 47

2.1.4 Information Systems used for CI process by Organizations 48

2.2 Text Summarization 49

2.2.1 Definition of Automatic Text Summarization (ATS) 49

2.2.2 Technologies used for ATS 51

2.2.3 Applications of ATS 58

2.3 Topic Identification using Text Clustering 61

2.4 Overall Learnings from Literature Review 63

3 Background of Proposed Work 65

3.1 Research Problem 68

3.1.1 Research Problem Definition 68

3.1.2 Research Objectives 68

3.2 Research Methodology 69

3.3 Architecture of Proposed Framework 71

4 Corpus Description and Pre-processing 73

4.1 Creation of Competitor Intelligence (CI) Corpus 73

4.2 Text/Corpus Pre-processing 75

5 Topic Identification Stage using Clustering 79

5.1 Corpus Representation for Clustering 81

5.2 Determination of Clustering Tendency of Corpora 84

5.2.1 Assessing Reliability of CI corpus 84

5.3 Multi-level K-means (ML-KM) Clustering Technique 86

5.4 Determining Number of clusters for ML-KM Clustering process 86
5.5 Determining Validity/Quality of Clusters 90
5.6 Multi-level K-means (ML-KM) Clustering Algorithm 91
5.7 Overall Algorithm for Topic Identification 94
5.8 Experimental Results 95
5.8.1 ML-KM Clustering Results for Samsung CI Corpus 98
5.8.2 ML-KM Clustering Results for Reuters Corpus 106
5.8.3 Generation of Cluster Quality Scores by Human Evaluators 111
5.8.4 Analysis of Relationship of APS Scores with Human Assigned Cluster Scores 112
6 Summary Generation Stage 116
6.1 Generation of System Summaries: Formulation of Summarization as Single-Objective Global Optimization Problem 119
6.1.1 Summary as a Single Unit (SaaU) 120
6.1.2 Summary Scoring Function 121
6.1.3 Description of the Optimization Problem for Summary Generation 129
6.1.4 Solution of the Optimization Problem using Stochastic Algorithms 129
6.2 Generation of Golden Summaries by Human Participants 141
6.2.1 Selection of Human Experts/Participants 142
6.2.2 Creation of a subset of topical clusters 143
6.2.3 Process of Summary Generation by Human Participants 144
6.2.4 Tasks performed by Human Participants 145
6.3 Experimental Results 146
6.3.1 ROUGE Formula used for Comparing System vs Golden Summaries 147
6.3.2 Analysis of ROUGE-L Recall Scores of System Summaries 149
6.3.3 Comparison of Recall Scores of Global Optimization based Summarization with Greedy (MMR) Summarization 155
6.3.4 Comparison of Recall Scores of ABC, DE and GA based Summarization Techniques 157
6.3.5 Analysis of ROUGE-L Precision Scores of System Summaries 161
6.3.6 Comparison of Precision Scores of Global Optimization Based Summarization with Greedy (MMR) based Summarization 169
6.3.7 Comparison of Precision Scores of ABC, DE and GA based Summarization Techniques 171
6.3.8 Analysis of Relationship of APS Scores of Clusters with Recall and Precision Scores of Corresponding Summaries 176
6.3.9 Analysis of Relationship of ROUGE-L Scores and Human Cluster Scores 191
6.3.10 Analysis of Relationship of Summary Scoring Function (Optimization Function) and ROUGE-L Recall Scores of System Summaries 195

6.3.11 Results of Long Runs 200
6.3.12 Example of Output Generated by the Proposed Framework 201
6.4 Evaluation of System Summaries by Managers 202
6.4.1 Selection of Respondents 202
6.4.2 Analysis of Manager Evaluations of System Summaries 204
7 Contributions, Conclusions & Future Work 224
7.1 Contributions of this research 224
7.2 Conclusions and Managerial Implications 225
7.3 Future Work 227