

Information Retrieval(IR) systems are gaining importance due to wide range of applications like recommender systems, search engines, etc., however, most of the IR systems use statistical methods built on top of bag-of-words approach for text retrieval. Graph-of-words approach is an alternative to bag-of-words approach that uses graph theoretic methods to rank keywords and related documents. We represent text documents as graphs whose vertices correspond to the unique terms belonging to the document. The edges represent co-occurrences between the terms. The underlying assumption is that the terms that co-occur have some sort of semantic relationship that can be harnessed for IR systems. The significant terms can be extracted using graph centrality measures. In this book, we have proposed a novel graph-of-words indexing technique using eigenvector scores that uses case separation for Gujarati language. We compared the performance of IR systems of our approach over the classical bag-of-words approach, mean average precision (MAP) values obtained in our experiments show that our approach has shown significant improvement over classical approaches.

Significant Keywords from Gujarati Text

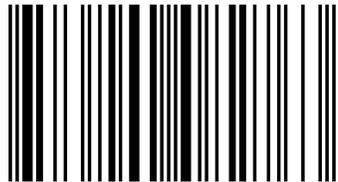


Dr. Hardik Joshi



Dr. Hardik Joshi is an Assistant Professor with the Dept. of Computer Sc., Gujarat University, India. His research interests include Natural Language Processing and Information Retrieval.

# Identification of Significant Keywords from Gujarati Text Documents



978-620-0-08276-3

Joshi

 **LAMBERT**  
Academic Publishing

**Dr. Hardik Joshi**

**Identification of Significant Keywords  
from Gujarati Text Documents**

FOR AUTHOR USE ONLY

**LAP LAMBERT Academic Publishing**

**Imprint**

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Cover image: [www.ingimage.com](http://www.ingimage.com)

Publisher:

LAP LAMBERT Academic Publishing

is a trademark of

International Book Market Service Ltd., member of OmniScriptum Publishing Group

17 Meldrum Street, Beau Bassin 71504, Mauritius

Printed at: see last page

**ISBN: 978-620-0-08276-3**

Copyright © Dr. Hardik Joshi

Copyright © 2019 International Book Market Service Ltd., member of  
OmniScriptum Publishing Group

FOR AUTHOR USE ONLY

# Contents

<b>Abstract</b>	iii
<b>Acknowledgements</b>	iv
<b>Contents</b>	vi
<b>List of Figures</b>	xi
<b>List of Tables</b>	xiii
<b>List of Abbreviations</b>	xv
<b>1 Introduction</b>	1
1.1 Architecture of an IR System	2
Indexing . . . . .	4
Retrieval . . . . .	4
1.2 Evaluation Workshops	5
1.2.1 Text Retrieval Conference	6
1.2.2 Conference and Labs of the Evaluation Forum	6
1.2.3 NII Test Collection for IR Systems . . . . .	7
1.2.4 Forum for Information Retrieval and Evaluation .	8
1.3 Evaluation Metrics . . . . .	8
1.4 Classification of Evaluation Metrics	9
1.4.1 Metrics based on Binary Relevance	10
Precision and Recall	10
F Measure . . . . .	12

	Average Precision . . . . .	12
	Mean Average Precision . . . . .	12
	Precision @ N . . . . .	14
	R-Precision . . . . .	14
	Geometric MAP . . . . .	15
1.4.2	Metrics based on Graded Relevance . . . . .	15
	DCG and NDCG . . . . .	15
1.4.3	Metrics used in dissertation . . . . .	16
	Tool for Evaluation . . . . .	16
1.5	Scope . . . . .	17
1.6	Software and Libraries . . . . .	18
1.7	Acronyms, Index and Notations . . . . .	18
1.8	Outline . . . . .	18
<b>2</b>	<b>Gujarati Language and its Linguistic Features</b>	<b>20</b>
2.1	Script, Symbols and Characters . . . . .	21
	2.1.1 Punctuation Marks . . . . .	21
2.2	Linguistic features . . . . .	23
	2.2.1 Nouns . . . . .	23
	Genders . . . . .	23
	Numbers . . . . .	23
	2.2.2 Adjectives . . . . .	24
	2.2.3 Verbs . . . . .	25
	2.2.4 Adverbs . . . . .	25
	2.2.5 Closed Class Words . . . . .	26
2.3	Harnessing Linguistic Features of Gujarati . . . . .	26
	2.3.1 Stopwords . . . . .	26
	2.3.2 Stems . . . . .	27
	2.3.3 Part-of-Speech Tagging . . . . .	28
	Steps to perform POS Tagging using CRFSuite . . . . .	29
	2.3.4 Thesaurus . . . . .	31
	2.3.5 Sentence Structure . . . . .	32
<b>3</b>	<b>Bag of Words Approach for Information Retrieval</b>	<b>33</b>
3.1	Text Representation . . . . .	33
	3.1.1 Terminology used in text representation . . . . .	34

3		3
3		3
3		3
3	The Boolean model	3
3	Probabilistic Model	3
3	Language Model	3
3	Vector space model	37
3		3
3	Widely used IR Systems	3
3		3
3		3
3	Zettair	4
3	Zebra	40
3	MeTA	4
3	Terrier	4
	Features Overview of Terrier	4
	Terrier architecture	4
3		4
3		4
3		4
3	Baseline of Gujarati IR tasks	4
3		4
3 7	Stemmers in Gujarati IR tasks	4
3 8	Frequent Case Generation	4
3 9	Query Expansion in Gujarati IR tasks	4
	3 9.1	51
3		52
<b>4</b>	<b>Graph Theoretic Approach for Information Retrieval</b>	<b>54</b>
4	Graph theory preliminaries	54
4		55
4		55
4	Edges	55
4	Walk Vs. Path	56
4	Shortest Path	56

4.1.6	Connectivity	57
4.1.7	Diameter	57
4.2	Graph Representation	57
4.2.1	Adjacency Matrix	57
4.2.2	Adjacency List and Edge List	58
4.3	Graphs as Networks	58
4.4	Information retrieval and Graphs	59
4.4.1	Early approaches in graph based IR	59
4.4.2	Recent work in graph based IR	60
4.5	Identification of Keywords	61
4.5.1	Overview	61
4.5.2	Keyword Identification methods	62
	Keyword Assignment	63
	Keyword Extraction	63
4.6	Graph based methods for Keyword Extraction	64
4.6.1	Graph Representations	64
4.6.2	Applications of Graph based representations	65
4.7	Centrality Measures and Keywords	65
4.8	Centrality Measures	66
4.8.1	Degree Centrality	66
	Normalizing Degree Centrality	68
4.8.2	Betweenness Centrality	68
4.8.3	Closeness Centrality	69
4.8.4	Eccentricity	70
4.8.5	Eigenvector Centrality	71
4.8.6	Katz Centrality	73
4.8.7	PageRank	74
4.8.8	Few other Centrality Measures	75
	HITS	75
	Stress Centrality	76
	Power Centrality	76
	Information Centrality	76
4.8.9	Comparison of Centrality Measures	76
4.9	Complexity of Centrality Algorithms	78
4.10	Experiments on Gujarati Dataset	78

4.10.1	Example of Gujarati Text Document . . . . .	78
4.10.2	Preprocessing of Gujarati Text Document . . . . .	81
4.10.3	Centrality Measures and its distribution . . . . .	82
4.10.4	Comparison of Graph Centrality Measures . . . . .	84
<b>5</b>	<b>Graph of Words approach for IR Systems</b>	<b>87</b>
5.1	Building graphs from text using RF_IDF Model . . . . .	87
5.1.1	Graph Generation . . . . .	88
5.2	Scoring using GOW Approach . . . . .	89
<b>6</b>	<b>Proposed Model and its Evaluation</b>	<b>92</b>
6.1	Motivation . . . . .	92
6.2	Query Expansion . . . . .	94
6.2.1	Word Embeddings . . . . .	94
6.3	Proposed Model using GOW approach . . . . .	95
6.3.1	Preserving semantics in GOW . . . . .	95
6.3.2	Architecture of proposed system . . . . .	97
6.4	Experimental Setup . . . . .	99
6.4.1	Test Collection . . . . .	100
6.4.2	Queries . . . . .	101
6.4.3	Pre-processing of Corpus . . . . .	102
6.5	Results of proposed system . . . . .	102
6.6	Analysis of Proposed Model . . . . .	103
<b>7</b>	<b>Conclusion and Future Work</b>	<b>107</b>
<b>A</b>	<b>Few Stopwords and Suffixes of Gujarati Language</b>	<b>108</b>
A.1	Few Stopwords of Gujarati Language . . . . .	108
A.2	Few Suffixes of Gujarati Language . . . . .	108
<b>B</b>	<b>PoS Tagset (Partial)</b>	<b>109</b>
<b>C</b>	<b>Queries for Evaluation</b>	<b>110</b>
<b>D</b>	<b>List of Publications</b>	<b>125</b>
	<b>Bibliography</b>	<b>126</b>

# List of Figures

1.1	IR System Architecture . . . . .	3
1.2	Indexing of Documents . . . . .	5
1.3	Calculation of Average Precision . . . . .	13
2.1	Gujarati Vowels and Consonants . . . . .	22
2.2	POS Tagging Process . . . . .	30
3.1	Sample Gujarati Text Document . . . . .	43
3.2	Relevance feedback mechanism . . . . .	51
4.1	Degree Centrality . . . . .	67
4.2	Betweenness Centrality . . . . .	69
4.3	Eigenvector Centrality . . . . .	72
4.4	Katz Centrality . . . . .	73
4.5	PageRank Centrality Example . . . . .	75
4.6	Example of All Centrality Measures . . . . .	77
4.7	Example of Unprocessed Gujarati Text Document . . . . .	79
4.8	Example of Text Graph (Unprocessed) with Random Place- ment of Nodes . . . . .	80
4.9	Example of Text Graph (Unprocessed) with Circular Layout of Nodes . . . . .	81
4.10	Example of POS Tagged Gujarati Text Document . . . . .	82
4.11	Example of Text Graph (after preprocessing) with Circular Layout of Nodes . . . . .	83
4.12	Example of Text Graph (after preprocessing) with Level Lay- out of Nodes . . . . .	84
4.13	Distribution of Eigenvector Scores of all documents . . . . .	85
5.1	Sample graph with window size=4 . . . . .	88

6.1	Knowledge Graph . . . . .	93
6.2	Sentence structure in English and Gujarati . . . . .	96
6.3	Separation of suffix of figure 6.2 . . . . .	97
6.4	Proposed graph based IR System . . . . .	98
6.5	Graph representation of 3 sentences with case separation . . . . .	99
6.6	Sample Gujarati Text Document . . . . .	101
6.7	Performance of queries for year 2011 . . . . .	105
6.8	Performance of queries for year 2012 . . . . .	105

FOR AUTHOR USE ONLY

# List of Tables

1.1	Confusion matrix . . . . .	11
2.1	Punctuation Marks in Gujarati Language . . . . .	21
2.2	Case Markers in Gujarati Language . . . . .	24
2.3	Genitive cases in Gujarati Language . . . . .	24
2.4	Inflections of Verbs in Gujarati Language . . . . .	25
3.1	Statistics of Gujarati Corpus . . . . .	44
3.2	Baseline Results . . . . .	46
3.3	Results after applying Stopword removal . . . . .	47
3.4	Results after applying Rule based stemmer . . . . .	48
3.5	Results after applying FCG technique . . . . .	50
3.6	Results of Query Expansion . . . . .	52
3.7	MAP values for BOW approach - 2011 queries . . . . .	53
3.8	MAP values for BOW approach - 2012 queries . . . . .	53
4.1	Comparison between Centrality Measures . . . . .	77
4.2	Top 10 keywords derived by applying Centrality Measures on unprocessed Gujarati text documents . . . . .	80
4.3	Top 10 keywords derived by applying Centrality Measures on POS Tagged Gujarati text documents . . . . .	83
4.4	Performance of each centrality measure in terms of precision, recall and F-measure at the top 10 keyterms . . . . .	85
5.1	Calculating node weights . . . . .	89
5.2	Graph based scoring for document d1 . . . . .	90
5.3	Graph based scoring for document d2 . . . . .	90
5.4	Graph based scoring for document d3 . . . . .	90
6.1	Statistics of Gujarati Corpus . . . . .	100

6.2	MAP values for the year 2011 Queries . . . . .	103
6.3	MAP values for the year 2012 Queries . . . . .	103
6.4	Improvement in Recall values . . . . .	104
6.5	Improvement in MAP score . . . . .	104

FOR AUTHOR USE ONLY