Textual and Visual Information Retrieval using Query Refinement and Pattern Analysis

S. G. Shaila · A. Vadivel

Textual and Visual Information Retrieval using Query Refinement and Pattern Analysis

Springer

S. G. Shaila
Department of Computer Science and Engineering
Dayananda Sagar University
Bangalore, India

A. Vadivel
Department of Computer Science and Engineering
SRM University AP
Amaravati, Andhra Pradesh, India

ISBN 978-981-13-2558-8 ISBN 978-981-13-2559-5 (eBook)
https://doi.org/10.1007/978-981-13-2559-5

Library of Congress Control Number: 2018955166

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

This book is dedicated to my guide, who is the co-author of the book, my parents and my family. Their encouragement and understanding helped me to complete this work. I thank my husband and children, Vinu and Shamitha, for their love and support to bring this book as a reality.

Foreword

I am extremely delighted to write the foreword for *Textual and Visual Information Retrieval using Query Refinement and Pattern Analysis*. The authors, Dr. S. G. Shaila and Dr. A. Vadivel, have disseminated their knowledge on information retrieval through this book. The content of the book is a very good educational and valuable resource for researchers in the domain of information retrieval.

The book includes various chapters on deep Web crawler, event pattern retrieval, Thesaurus generation and query expansion, CBIR applications and indexing and encoding to cover the whole concept of information retrieval. Each chapter contains the theoretical information with experimental results and is intended for information retrieval researchers. The layout of each chapter includes a table of contents, introduction, material content and experimental results with analysis and interpretation.

This edition of the book reflects new guidelines that have evolved in information retrieval in terms of text- and content-based information retrieval schemes. The indexing and encoding mechanism of the low-level feature vector is also presented with results and analysis.

It is my hope and expectation that this book will provide an effective learning experience and referenced resources for all information retrieval researchers, leading to advanced research.

Bangalore, Karnataka, India

Dr. M. K. Banga
Chairman
Department of Computer
Science Engineering

Dean (Research)
School of Engineering
Dayananda Sagar University

Preface

Multimedia information retrieval from the distributed environment is an important research problem. It requires an architecture specification for handling various issues such as techniques to crawl information from WWW, user query prediction and refinement mechanisms, text and image feature extraction, indexing and encoding, similarity measure. In this book, research issues related to all the above-mentioned problems are discussed and suitable techniques are presented in various chapters.

Both the text and images are presented in web documents. In a comprehensive retrieval mechanism, text-based information retrieval (TBIR) plays an important role. In Chap. 1, the text-based retrieval is used for retrieving relevant documents from the Internet by using a suitable crawler with the capability to crawl deep and surface web. The functional dependency of core and allied fields in HTML FORM is identified for generating rules using SVM classifier. The presented crawler fetches a large number of documents while using the values in most preferable class. This architecture has a higher coverage rate and reduces fetching time.

In recent times, information classification is very important for text-based information retrieval. In Chap. 2, the classification based on events is presented and also the event Corpus is discussed, which is important for many real-time applications. Event patterns are identified and extracted at the sentence level using term features. The terms that trigger events along with the sentences are extracted from web documents. The sentence structures are analysed using POS tags. A hierarchal sentence classification model is presented by considering specific term features of the sentence, and the rules are derived along with fuzzy rules to get the importance of all term features of the sentences. The performance of the method is evaluated for 'Crime' d 'Die' and found that the performance of this approach is encouraging.

In general, the retrieval system depends on the user query to retrieve the web documents. The user-defined queries should have sufficient relevant terms, since the retrieval set depends on the queries. The query refinement through query expansion mechanism plays an important role. In Chap. 3, the N-gram Thesaurus construction mechanism for query expansion is presented. The HTML TAGs in web documents are considered and their syntactical context is understood. Based on the significance

of the TAGs in designing the web pages, suitable weight is assigned for TAGs. The term weight is calculated using corresponding TAG weight and frequency of the term. The terms along with the TAG information are updated into an inverted index. The N-grams are generated using the term and term weights in the document and updated as N-grams in the Thesaurus. During the query session, the term is expanded based on the content in the Thesaurus and suggested to the user. It is found that while the selected query is submitted to the retrieval system, the retrieval set consists of a large number of relevant documents.

In Chap. 4, the issues related to content-based image retrieval (CBIR) are presented. The chapter presents a histogram based on human colour visual perception by extracting the low-level features. For each pixel, the true colour and grey colour proportion are calculated using a suitable weight function. During histogram construction, the hue and intensity values are iteratively distributed to the neighbouring bins. The NBS distance is calculated between the reference bin and their adjacent bins. The NBS distance provides the overlapping proportion of the colour from the reference bin to its adjacent bins, and accordingly, the weight is updated. The distribution makes it possible to extract the background colour information effectively along with the foreground information. The low-level features of all the images in the database are extracted and stored in a feature database, and the relevant images are retrieved based on the rank. The Manhattan distance is used as a similarity measure, and the performance of the histogram is evaluated on Coral benchmark dataset.

In Chap. 5, the issues of indexing and encoding of low-level features and a similarity measure are presented. In CBIR system, the low-level features are stored along with the images and require a large number of storage space along with increased search and retrieval time. The search time increases linearly with the database size, which reduces the retrieval performance. The colour histograms of images are considered as low-level features. The bin values are analysed to understand their contribution representing image colour. The trivial bins are truncated and removed, and only important bins are considered to have histograms with lesser number of bins. The coding scheme used GR coding algorithm, and the quotient and remainder code parts are evaluated. Since there is variation between the number of bins in the query and database histogram, *BOSM* is used as a similarity measure. The performance of all the schemes is evaluated in an image retrieval system. The retrieval time, number of bits needed for histogram construction and precision of retrieval are evaluated using benchmark datasets, and the performance of the presented approach is encouraging.

Finally, as a whole, the book presents various important issues in information retrieval research filed and will be very much useful for the postgraduates and researchers working in information retrieval problems.

Bangalore, India — S. G. Shaila
Amaravati, India — A. Vadivel

Acknowledgements

First and foremost, we thank the Almighty for giving the wisdom, health, environment and people to complete this book.

We express our sincere gratitude to Dr. Hemachandra Sagar and Premachandra Sagar, Chancellor and Pro-Chancellor, Dayananda Sagar University, Bangalore; Dr. A. N. N. Murthy, Vice Chancellor, Dayananda Sagar University, Bangalore; Prof. Janardhan, Pro-Vice Chancellor, Dayananda Sagar University, Bangalore; Dr. Puttamadappa C., Registrar, Dayananda Sagar University, Bangalore; Dr. Srinivas A., Dean, School of Engineering, Dayananda Sagar University, Bangalore; Dr. M. K. Banga, Chairman, Department of CSE, and Dean Research, Dayananda Sagar University, Bangalore, for providing an opportunity and motivation to write this book.

We express our sincere gratitude to Dr. P. Sathyanarayanan, President, SRM University, Amaravati, AP; Prof. Jamshed Bharucha, Vice Chancellor, SRM University, Amaravati, AP; Prof. D. Narayana Rao, Pro-Vice Chancellor, SRM University, Amaravati, AP; Dr. D. Gunasekaran, Registrar, SRM University, Amaravati, AP, for providing an opportunity and motivation to write this book.

We would like to express our sincere thanks to our parents, spouse, children and faculty colleagues for their support, love and affection. Their inspiration gave us the strength and support to finish the book.

Dr. S. G. Shaila
Dr. A. Vadivel

Contents

About the Authors

Dr. S. G. Shaila is an Associate Professor in the Department of Computer Science and Engineering in Dayananda Sagar University, Bangalore, Karnataka. She earned her Ph.D. in computer science from the National Institute of Technology, Tiruchirappalli, Tamil Nadu, for her thesis on multimedia information retrieval in distributed systems. She brings with her years of teaching and research experience. She worked for the DST project, Govt of India and Indo US based projects as Research Fellow. Her main areas of interest are information retrieval, image processing, cognitive science and pattern recognition. She published 20 international journals and confernce proceedings.

Dr. A. Vadivel received his master's in science from the National Institute of Technology, Tiruchirappalli (NITT), Tamil Nadu, before completing a master's in Technology (M.Tech.) and Ph.D. at the Indian Institute of Technology (IIT), Kharagpur, India. He has 12 years of technical experience as a network and instrumentation engineer at the IIT Kharagpur and 12 years of teaching experience at Bharathidasan University and NITT. Currently, he is working as Associate Professor at SRM University, Amaravati, AP. He has published papers in more than 135 international journals and conference proceedings. His research areas are content-based image and video retrieval, multimedia information retrieval from distributed environments, medical image analysis, object tracking in motion video and cognitive science. He received the Young Scientist Award from the Department of Science and Technology, Government of India, in 2007; the Indo-US Research Fellow Award from the Indo-US Science and Technology Forum in 2008; and the Obama-Singh Knowledge Initiative Award in 2013.

Abbreviations

ACE	Automatic Content Extraction
ALNES	Active Long Negative Emotional Sentence
ALNIES	Active Long Negative Intensified Emotional Sentence
ALPES	Active Long Positive Emotional Sentence
ALPIES	Active Long Positive Intensified Emotional Sentence
ANFIS	Artificial neuro-fuzzy inference system
ANN	Artificial neural network
ASNES	Active Short Negative Emotional Sentence
ASNIES	Active Short Negative Intensified Emotional Sentence
ASPES	Active Short Positive Emotional Sentence
ASPIES	Active Short Positive Intensified Emotional Sentence
Bo1	Bose–Einstein statistics model
BoCo	Bose–Einstein statistics co-occurrence model
BOSM	Bin overlapped similarity measure
CART	Classification and regression tool
CBIR	Content-based image retrieval
CF	Core field
Co	Co-occurrence
CRF	Conditional random fields
DCT	Discrete cosine transform
DOM	Document object model
EMD	Earth mover's distance
ET	Emotional triggered
FGC	Form Graph Clustering
FPN	Fuzzy Petri-Nets
GR	Golomb–Rice
HCPH	Human colour perception histogram
HiWE	Hidden Web Exposer
HQAOS	High-Qualified Active Objective Sentence
HQASS	High-Qualified Active Subjective Sentence

HQPOS	High-Qualified Passive Objective Sentence
HQPSS	High-Qualified Passive Subjective Sentence
HSV	Hue Saturation Value
HTML	Hyper-Text Markup Language
IBC	Iraq Body Count
InPS	Internal Property Set
IPS	Input Property Set
IR	Information retrieval
IRM	Integrated Region Matching
IWED	Integrated Web Event Detector
KDB	K-Dimensional B-tree
KLD	Kullback–Liebler divergence
KLDCo	Kullback–Liebler divergence co-occurrence model
LDA	*Latent Dirichlet allocation*
LP	Least Preferable
LVS	Label value set
MAP	Mean average precision
MCCM	Colour-based co-occurrence matrix scheme
ME	Mutually Exclusive
MEP	Minimum Executable Pattern
MHCPH	Modified human colour perception histogram
MP	Most Preferable
MRR	Mean reciprocal recall
MUC	Message Understanding Conference
NBS	National Bureau of Standards
NIST	National Institute of Standards and Technology
NLP	Natural language processing
NQAOS	Non-Qualified Active Objective Sentence
NQASS	Non-Qualified Active Subjective Sentence
NQPOS	Non-Qualified Passive Objective Sentence
NQPSS	Non-Qualified Passive Subjective Sentence
OPS	Output Property Set
PHOTO	Pyramid histogram of topics
PIW	Publicly Indexable Web
PLNES	Passive Long Negative Emotional Sentence
PLNIES	Passive Long Negative Intensified Emotional Sentence
PLPES	Passive Long Positive Emotional Sentence
PLPIES	Passive Long Positive Intensified Emotional Sentence
POS	Part of speech
PQ	Product quantization
PSNES	Passive Short Negative Emotional Sentence
PSNIES	Passive Short Negative Intensified Emotional Sentence
PSPES	Passive Short Positive Emotional Sentence
PSPIES	Passive Short Positive Intensified Emotional Sentence
QA	Question Answering

QAOS	Qualified Active Objective Sentence
QASS	Qualified Active Subjective Sentence
QPOS	Qualified Passive Objective Sentence
QPSS	Qualified Passive Subjective Sentence
RED	Retrospective new Event Detection
RS	Rule set
SCM	Sentence classification model
SOAP	Simple object access protocol
SQAOS	Semi-Qualified Active Objective Sentence
SQASS	Semi-Qualified Active Subjective Sentence
SQPOS	Semi-Qualified Passive Objective Sentence
SQPSS	Semi-Qualified Passive Subjective Sentence
SVM	Support vector machine
SWC	Surface web crawlers
SWR	Semantic Web Retrieval
TBIR	Text-based information retrieval
TDT	Topic Detection and Tracking
TS	Text Summarization
TSN	Term semantic network
TTW	TAG term weight
UMLS	Unified Medical Language System
URL	Uniform Resource Locator
ViDE	Vision-based data extraction
WaC	Web as Corpus
WSDL	Web Services Description Language
WWW	World Wide Web
XML	Extensible Markup Language

List of Figures

List of Tables