AuTopicGen: Rule based Positional Pattern Approach for Topic Collection in IR

Payal Joshi
M.Sc. (IT) Program,
Veer Narmad South Gujarat University,
Surat

S. V. Patel
Department of Computer Science,
Veer Narmad South Gujarat University,
Surat

ABSTRACT
IR systems consist of phases like document preprocessing, indexing, query expansion, query matching, ranking etc. The document preprocessing phase is the most important phase to parse the document and collect keywords. Relevance of overall IR system improves if main topics of document are perfectly identified during this phase. It is a known fact that Topics are mostly phrase based. Existing phrase search methods like n-grams or positional indexes are quite complex and also suffer from problems of inaccuracy, requirement of large storage space etc. Moreover, IR system like digital library may consist of eBooks on one or more subjects. So for phrase collection, one may have to use appropriate ontology to retrieve phrases or topics. This paper presents a new approach called AuTopicGen (Automatic Topic Generator) that automatically collects most relevant topics of eBooks from its contents and indexes using rule based positional patterns approach. From the collected topics, we create topic hierarchy that can work as light weight ontology to improve overall performance of information retrieval system especially for phrase based queries and to assist user with query recommendation. Further this will be useful as topic maps, mind maps, to improve user interface to help user navigate through topics, for categorization, query expansion and ranking algorithms. We have successfully implemented the approach for topics collection practically on eBooks and presented in this paper.

Categories and Subject Descriptors
Document preprocessing, Topics collection

General Terms
Algorithms, Performance

Keywords
Topic Collection, IR System

1. INTRODUCTION
Digital world of 21st century is bundled with enormous information and it is growing tremendously day by day. To make desired information available to the user, Information Retrieval systems are developed. Major stages of IR systems are document preprocessing, indexing, query expansion, query matching, ranking etc. Digital library and search engines are examples of IR systems.

Phrase search is a major research issue in IR with the aim to retrieve most relevant documents. Collection of most relevant phrases or topics of document are very useful to find exact documents that match it, to assist user with different other topics related to user search, to use in query recommender and to provide visual browsing user interface to make user aware with contents available in document.

Identification of phrases is done at document preprocessing stage. It is evident that performance and efficiency of IR systems heavily depends on work done in document preprocessing phase. For this IR systems make use of bi-word or tri-word indexes, TF-IDF, or word position indexes to find documents that match user query [1]. Modern IR systems make use of existing ontology or topic models. Further IR system may consist of documents on different subjects, therefore for phrase collection we may use appropriate ontology depending on domain. These methods are quite complex, time consuming and also suffer from problems of inaccuracy, requirement of large storage space etc.

In view of the above, an automated approach is required for topic collection from documents on different subjects.

This paper presents a new approach called AuTopicGen (Automatic Topic Generator) which automatically collects most relevant topics of documents from contents using rule based positional pattern approach. We collect topics from Contents and Index of eBooks because they contain major technical phrase based topics related to the document. From the collected topics, we create topic hierarchy that can work as light weight ontology to improve overall performance of information retrieval system for phrase based query and search. This will also be useful as topic maps, mind maps, to improve user interface to help user navigate through topics, categorization, query expansion and ranking algorithms. This approach has been successfully implemented for topics collection on eBooks and presented in the paper.

Rest of the paper is organized as follows: Section II describes related work. Section III describes methodology of topics collection. Section IV shows results and section V concludes the paper.

2. RELATED WORK
Phrase or topic based query must be matched with relevant document that actually contain its words in its exact sequence. Using bag of words approach for this mostly retrieves irrelevant documents because most document contain these keywords more number of times but keywords may not exist in same order in documents. For phrase search, IR systems use one or combination of following approaches:

2.1 Use of Readymade Ontology or Topic Maps
For phrase search in [3] OntoRo and OntoCorp ontology are used. In [11] use of WordNet is done for the same. It helps to retrieve documents that contain exact words or synonyms of it. But it is quite complex and takes more time to process large amount of documents with bulky sized ontology. Apart from this disadvantage, use of more than one domain specific ontology makes operations complex for document collection of diverse domains.

2.2 Phrase Collection
Various techniques like bi-word, tri-word or n-word index and positional indexes are used in [1]. N-word index require a large amount of storage and majority of phrases may not be of use. In

[2][7][8][9][10][12] authors have removed irrelevant phrases but they have used different word level positional indexes which require storing positional location of each word in document which is complex, time consuming and occupies a large storage space.

2.3 Manually Built Topic Maps

Topic maps contain list and hierarchy of topics. Manually-built topic maps are developed as stated [4] for IR systems. Usefulness of such maps is described [5][6]. Manually built topic maps may generate more relevant results but it requires a lot of time and expert assistance for its development.

Considering all these issues, we have developed an automated approach to collect topics or topic maps. In this approach we automatically collect most relevant topics of documents from contents and indexes based on positional patterns that work effectively for collection of documents of any domain or subject or mix of various subjects.

3. AuTopicGen: AUTOMATED TOPICS COLLECTION APPROACH

AuTopicGen i.e Automatic Topics Generator has been practically implemented using Java to programmatically collect topics and topics hierarchy from contents and indexes of eBooks collection. We collect topics from Contents and Index of eBooks because they contain major technical phrase based topics related to the document and contain text of highly unstructured nature. We have observed format of various books and its patterns for contents and indexes. Few of them are shown here (as shown in Figure 1).

Some eBooks contains "Preface" as a title for contents. In certain eBooks, first list of chapters are given and then detailed contents list is provided (Figure 1[e]). Similarly different formats are adapted for index also. Due to such heterogeneity complexity increases.

Considering such issues we implemented rule based decision tree approach. We have made use of regular expression and fuzzy approach to find similar strings and patterns rather than matching exact strings and patterns. Overview of our methodology is given below.

1. Read document line by line. Find lines with words like "index", "contents", "table of contents", "Preface" etc. Also check lower neighboring lines. If lines contain topic numbers and/or corresponding page numbers, then we found an exact location to start collecting topics from contents.
2. Check lower neighboring lines. If lines contain topic numbers and/or corresponding page numbers, then it is a valid member to be part of our collection.
3. Based on chapter number and sub topics numbering patterns build up hierarchy of topics.
4. Clean the pattern. (Remove topic number, page numbers and punctuation marks) In this step also different patterns are followed in different eBooks. Like "2.4 Topic Name ……..36", "Topic Name 20" etc. While removing topic number, page numbers and punctuation marks, we have to take care of preserving punctuation marks used in topic itself.

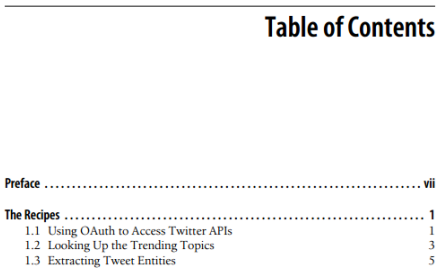
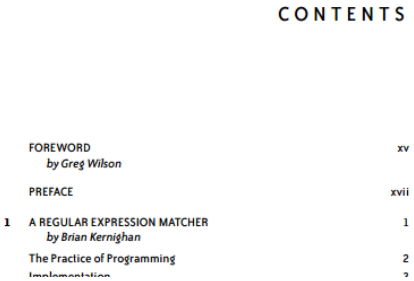


Table of Contents

Preface vii

The Recipes 1
1.1 Using OAuth to Access Twitter APIs 1
1.2 Looking Up the Trending Topics 3
1.3 Extracting Tweet Entities 5

Figure 1. (a) eBook showing title of contents as "Table of Contents"

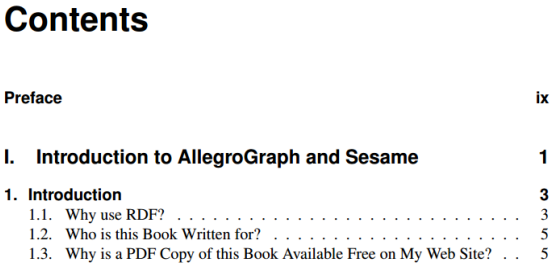

CONTENTS

FOREWORD xv
by Greg Wilson

PREFACE xvii

1 A REGULAR EXPRESSION MATCHER 1
by Brian Kernighan
The Practice of Programming 2

Figure 1. (b) eBook showing title of contents as "C O N T E N T S"

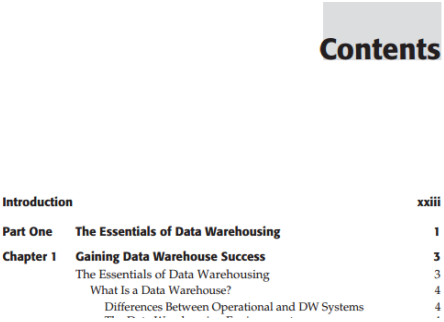

Contents

Preface ix

I. Introduction to AllegroGraph and Sesame 1

1. Introduction 3
1.1. Why use RDF? 3
1.2. Who is this Book Written for? 5
1.3. Why is a PDF Copy of this Book Available Free on My Web Site? . . 5

Figure 1. (c) eBook showing title of contents as "Contents"

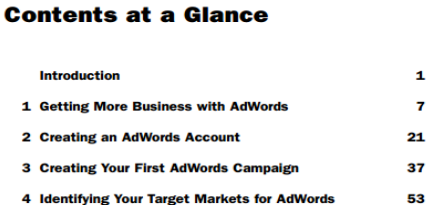

Contents

Introduction xxiii

Part One The Essentials of Data Warehousing 1

Chapter 1 Gaining Data Warehouse Success 3
The Essentials of Data Warehousing 3
What Is a Data Warehouse? 4
Differences Between Operational and DW Systems 4

Figure 1. (d) eBook showing title of contents as "Contents" on right side


Contents at a Glance

Introduction 1
1 Getting More Business with AdWords 7
2 Creating an AdWords Account 21
3 Creating Your First AdWords Campaign 37
4 Identifying Your Target Markets for AdWords 53

Figure 1. (e) eBook showing "Contents at a Glance"

5. Filter topic if it is one of the member stop list. We also filter one word topics. In this collection, certain words repeat and they are not useful to improve relevance. We have made a stop word list of total 42 such terms. Few of such terms are "FIGURE", "Introduction", "Summary", "Conclusion", "Q&A", "Exercise", etc. While collecting topics we filter it and store only useful topics.
6. Find location of line with word "Index". Some eBooks may not contain index.
7. Check lower neighboring lines and collect topics. If lines topic names and page numbers, then it is a valid member to be part of our collection. In INDEX, main topic does not contain line number and its sub topic contains line number. Based on this build up hierarchy of topics.
8. Clean the pattern. (page numbers and punctuation marks) In this step also different patterns are followed in different eBooks. Like "Topic Name ……..36", "Topic Name 20" etc. Here also while removing topic number, page numbers and punctuation marks, we have to take care of preserving punctuation marks used in topic itself.
9. Filter topic if it is one of the member stop list.
10. After topics identification, undergo with usual keywords collection approach.

In this approach, we have not applied stemming and removing common stop word process on collected topics to maintain originality of topics.

4. RESULTS AND ANALYSIS

We have parsed eBooks and collected topics. Here we have shown sample of our result with 10 eBooks and number of topics collected from them (Table 1).

Table 1. Outcomes of 10 sample e-books for collected topics

No.	Book	No. of Topics Found	Relevant Topics
1	21 recipes for mining twitter	21	21
2	A managers guide to data warehousing	2442	2238
3	Data Mining, Practical Machine Learning Tools and Techniques 2nd.Ed (Morgan Kaufmann 2005)WEKA	1418	1374
4	Manning Schuetze StatisticalNLP	188	187
5	Practical Artificial Intelligence in prolog-lisp-java	85	81
6	Practical Semantic Web and Linked Data Applications	124	119
7	Think Python	2377	1938
8	The myths of security	403	385
9	ThinkStats	367	356
10	WebManual	206	204
Total		7631	6903
Accuracy		90.45%	

A short sample of topic hierarchy obtained is as given below:

1Gaining Data Warehouse Success

1.1 The Essentials of Data Warehousing

1.2 What Is a Data Warehouse?

..

..

2 Understanding Where You Are and Finding Your Way

2.1Assessing Your Current State

2.2 What Is Your Company's Strategic Direction?

2.3 What Are the Company's Top Initiatives?

2.4 How Healthy Is Your Data?

..

..

Topic hierarchy that we created can also work as light weight ontology to improve semantic relevance of IR system. Its accuracy is 90.45% for collection shown in this paper, however it may vary for collection of different number and types of documents. In Table-1, relevant topics means correctly extracted topics. Correctness of topics is checked by manually reading them.

Rather than making use of domain level knowledge to collect topics, we have used a positional pattern based approach hence it can work well with document collection of any type of domain.

5. CONCLUSION

Topic collection is useful to improve overall performance of IR system. This paper introduces first practical version of an automated topics collection approach AuTopicGen which stands for Automatic Topic Generator which automatically collects most relevant topics of eBooks from contents and indexes using rule based positional pattern approach. Its accuracy is 90.45% for collection shown in this paper and it may vary for collection of different number and types of documents. We are working to enhance it and to utilize it for our digital library as part of our research work. In future in our digital library, this topic collection will be used to improve performance in querying, making navigational topic interface, query recommender and ranking algorithm.

Topic collection enhances relevance of result for phrase queries. This approach is purely positional pattern based. So it is useful to be applied on eBooks from any type of domain. Using our topic collection, easy to use browsing interface can be made for users which helps them to search topic even if they do not know them. (rather than just typing query terms blindly). It can also improve user interface to help user navigate through topics.

Interactive query expansion and query recommender systems can be enhanced from topic collection. Improvement can be done in performance of categorization, and ranking algorithms stages of IR system. Our approach is quite useful and efficient to collect topics of any domain or subject.

6. REFERENCES

[1] Christopher D. Manning. Prabhakar Raghavan An.Introduction to. Information. Retrieval, Online edition (c) 2009 Cambridge UP. 1 (Aug. 2006). DOI= http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

[2] Dongdong Shan, Wayne Xin Zhao, Jing He, Rui Yan, Hongfei Yan, Xiaoming Li, 2011, Efficient phrase querying with flat position index. CIKM 2011. In Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2001-2004. ISBN: 978-1-4503-0717-8

[3] Rossitza M. Setchi, Qiao Tang. 2007. Concept Indexing using Ontology and Supervised Machine Learning. In International Journal of Computer, Information, Systems and Control Engineering Vol:1 No:1.

[4] Xing Wei & W. Bruce Croft. 2007. Investigating Retrieval Performance with Manually-Built Topic Models. RIAO'07.

[5] Beel, J., Gipp, B., Stiller, J.-O. 2009. Information retrieval on mind maps - what could it be good for?, 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2009. CollaborateCom 2009. 11-14 (Nov. 2009), 1-4. ISBN: 978-963-9799-76-9, DOI= http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8298

[6] Joeran Beel, Stefan Langer1, Marcel Genzmehr, Bela Gipp. 2014. Utilizing Mind-Maps for Information Retrieval and User Modeling. 14 (Apr. 2014), UMAP 2014.

[7] Paolo Rossol, Edgardo Ferretti, Daniel Jiménez, and Vicente Vidal. Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.). 2004. Text Categorization and Information Retrieval Using WordNet Senses. In Proceedings of GWC-2004, 299-304.

[8] Kiril Panev and Klaus Berberich.B. Benatallah et al. (Eds.). 2014. Phrase Queries with Inverted + Direct Indexes. WISE 2014 . 156-169. Springer International Publishing Switzerland 2014.

[9] Jinru He, Torsten Suel. 2012. Optimizing Positional Index Structures for Versioned Document Collections. In SIGIR'12, 12-16 (Aug. 2012), Portland, Oregon, USA.

[10] Manish Patil, Sharma V Thankachan, Rahul Shah, Wing-Kai Hon, Jeffrey Scott Vitter, and Sabrina Chandrasekaran. Inverted indexes for phrases and strings. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 555–564. ACM, 2011.

[11] Shuang Liu, Fang Liu, Clement Yu, Weiyi Meng. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in Information Retrieval, pages 266-272. ACM, 2004.

[12] Shashank Gugnani, Rajendra Kumar Roul. Triple Indexing: An Efficient Technique for Fast Phrase Query Expansion. In International Journal of Computer Applications (0975 8887) Volume 87 - No 13, February 2014.