SEMANTIC INFORMATION RETRIEVAL BASED ON TOPIC MODELING AND COMMUNITY INTERESTS MINING

R.P. Minuri Chathurika Rajapaksha

(158734M)

Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

March 2019

SEMANTIC INFORMATION RETRIEVAL BASED ON TOPIC MODELING AND COMMUNITY INTERESTS MINING

R.P. Minuri Chathurika Rajapaksha

(158734M)

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

March 2019

DECLARATION

I declare that this dissertation does not incorporate, without acknowledgment, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name of Student	Signature of Student
Minuri Rajapaksha	Date:

The above candidate has carried out research for the Masters Dissertation under my supervision.

Supervised by	Signature of Supervisor
Dr. Thushari Silva	Date:

Abstract

Search engines or localized software systems developed for information searching, play an important role in knowledge discovery. Proliferation of data in the web and social media has posed significant challenges in finding relevant information efficiently even using those search engines or other software systems. Moreover, those systems or engines tend to collect massive number of data, which could be useful for humans in various ways but overlook the meaning of the search phrases, hence generate irrelevant search results. A unit level searching i.e. searching information within a website or page is also not effective as they follow exact keyword matching techniques and ignore the semantic level matching of search phrases. In order to address those deficiencies, this research proposes a hybrid approach which use the semantics of data, community preferences as well as collaborative filtering techniques for semantic information retrieval. More specifically, Topic modeling based on Latent Dirichlet Allocation together with topic-driven based community detection methods are applied for identifying personalized search results and hence improve the relatedness of the research results. Based on the proposed hybrid approach a framework for semantic search that can easily be integrated to a software application has been implemented. The evaluation results confirm the effectiveness of search results which outperform benchmark approaches that follow traditional keyword search algorithms.

TABLE OF CONTENTS

DECLARATION iii

ABSTRACT iv

TABLE OF CONTENTS v

LIST OF FIGURES viii

LIST OF TABLES x

CHAPTER 1 1

1. INTRODUCTION 1

1.1. Prolegomena 1

1.2. Objectives 2

1.3. Background and Motivation 3

1.4. Problem in Brief 3

1.5. Proposed Solution 3

1.6. Resource Requirements 5

1.7. Structure of the Thesis 5

1.8. Summary 5

CHAPTER 2 6

2. SEMANTIC INFORMATION RETRIEVAL ISSUES AND CHALLENGES .6

2.1. Introduction 6

2.2. Generation of Semantic Information Retrieval 6

2.3. State of the Art of Personalized Semantic Information Retrieval 8

2.4. Future Trends 16

2.5. Problem Definition 16

2.6. Summary 17

CHAPTER 3 18

3. TECHNOLOGIES ADAPTED 18

3.1. Introduction 18

3.2. Discover Topics Based on Optimized LDA Model 18

3.3. Community Detection Based on Topic-Driven Approach ... 21
3.4. Collaborative Filtering Module ... 22
3.5. Search Optimization Module ... 26
3.6. Summary ... 27
CHAPTER 4 ... 28
4. APPROACH ... 28
4.1. Introduction ... 28
4.2. Hypothesis ... 28
4.3. Process ... 30
4.4. Input and Output ... 31
4.5. Users ... 31
4.6. Features ... 32
4.7. Summary ... 32
CHAPTER 5 ... 33
5. DESIGN ... 33
5.1. Introduction ... 33
5.2. Discover Topics Module based on Optimized LDA Model ... 34
5.3. Community Extraction Module based on Topic-Driven Model ... 35
5.4. Item-based Collaborative Filtering Module ... 38
5.5. Search Optimization Module ... 39
5.6. Summary ... 39
CHAPTER 6 ... 40
6. IMPLEMENTATION ... 40
6.1. Introduction ... 40
6.2. Data Preprocessing ... 42
6.3. Discover Topics Based on Topic Modeling ... 43
6.3. Community Detection Based on Topic-Driven Approach ... 44
6.4. Item-Based Collaborative Filtering Module ... 46
6.5. Search Optimization Module ... 47

6.6. Summary 52
CHAPTER 7 53
7. EVALUATION 53
7.1. Introduction 53
7.2. Experimental Setup 53
7.3. Participants 53
7.4. Test Cases 53
7.5. Testing Strategies 54
7.6. Training Data 54
7.7. LDA Model Performance 54
7.8. Community Detection 57
7.9. Search Optimization 60
CHAPTER 8 62
8. CONCLUSION AND FURTHER WORK 62
8.1. Introduction 62
8.2. Conclusion 63
8.3. Further Work 65
REFERENCES 66
Appendix A – Python Code for Data Preprocessing 69
Appendix B – Calculate Distance between users Using Jensen Shannon Divergence 73
Appendix C – Construct the Communities using Girwan Newman Algorithm 76
Appendix D – Item Based Collaborative Filtering which returns Recommended Articles for Particular User 80
Appendix E – Integrate with Lucene 87

LIST OF FIGURES

Figure 3.1 - LDA model..20

Figure 3.2 - Lucene in a search system 26

Figure 4.1 - Design diagram (Top level architecture)..31

Figure 5.1 - High level architecture of semantic information retrieval...............33

Figure 5.2 - LDA model..34

Figure 6.1 - Semantic Information Retrieval System..41

Figure 6.2 - Word frequency dictionary...42

Figure 6.3 - LDA Model 43

Figure 6.4 - Sample Graph.text file content 45

Figure 6.5 - Constructed communities 46

Figure 6.6 - User Sign-up page 48

Figure 6.7 - User sign-in page...48

Figure 6.8 - List of user uploaded files and user can edit uploaded files.............49

Figure 6.9 - File up loader..49

Figure 6.10 - Create new file..50

Figure 6.11 - Search results for text "carrom" 50

Figure 6.12 - Search results for text "mom" 50

Figure 6.13 - Search results for text "county"..51

Figure 6.14 - Configuration set-up page 51

Figure 6.15 - Uploaded file content 52

Figure 7.1 - Choosing optimal model with coherence scores 56

Figure 7.2 - Topic-word distribution..56

Figure 7.3 - Word count and importance of topic keywords 57

Figure 7.4 - Constructed graph..58

Figure 7.5 - Topics treated in community1 58

LIST OF TABLES

Table 2.1 - Topic Modeling Current Practices and Issues 9

Table 2.2 - Community Detection Current Practices And Issues 10

Table 7.1 - Evaluate the model basaed on different topic numbers 55

Table 7.2 - Distance between users 57

Table 7.3 - Closeness between users for selected topic 58

Table 7.4 - Evaluate the community detection based on different number of topics for trained model 59

Table 7.5 - Filter Articles User Uploaded Wise 60

Table 7.6 - Filter articles based on item based collaborative filtering 61

Table 7.7 - Prioritize search results 61