UNIVERSITI PUTRA MALAYSIA

MODELING LEXICAL SEMANTICS OF TERMS BASED ON SYNWORD IDENTIFICATION FOR IDEA MINING IN INFORMATION RETRIEVAL

MOSTAFA AHMED ALKSHER

FSKTM 2018 88

MODELING LEXICAL SEMANTICS OF TERMS BASED ON SYNWORD IDENTIFICATION FOR IDEA MINING IN INFORMATION RETRIEVAL

By

MOSTAFA AHMED ALKSHER

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

December 2018

DEDICATIONS

I would like to dedicate this thesis for the sake of Allah, my Creator and my Master
To my great teacher and messenger, Mohammed (MayAllah bless and grant him), who taught us the purpose of life.
To my wife and beloved kids, Ahmed, Faris, Rawa and Joury, whom I can't force myself to stop loving.

&

To All whom I love

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

MODELING LEXICAL SEMANTICS OF TERMS BASED ON SYNWORD IDENTIFICATION FOR IDEA MINING IN INFORMATION RETRIEVAL

By

MOSTAFA AHMED ALKSHER

December 2018

Chairman: Azreen Azman, PhD
Faculty: Computer Science and Information Technology

The exponential accumulation of digital information in the form of business or public data has brought with it great challenges about how to extract more value from data. Individuals and organizations can no longer rely on human review and extraction of useful data or ideas from huge volumes of digital data because it is time-consuming to identify useful ideas within a large amount of textual information. The idea is an important component in the information retrieval and plays a key role in Idea Mining (IM) from unstructured text. An idea has been defined as a pair of problem and solution (or a pair of mean and end) within the same context. IM is introduced as an automatic process of mining new and innovative ideas from unstructured text by using text-mining tools. Nowadays, many companies have invested in Text Mining (TM) technology to discover hidden valuable information from unstructured text, which is very important for decision-making.

Though there is no doubt about great ideas hidden within the huge public and business data, technically speaking, the major challenge is the idea characterization and reasoning. The traditional formation of ideas relies on identifying an individual idea either as a pair of the unknown solution to a known problem or known solution to an unknown problem. Then the idea mining identifier of this model makes a textual comparison between a new text (i.e., input query) and the collection documents. The output of the comparison should be in the form of *unknown* words and *known* words. *known* words refer to the terms that appear in both new text and collection documents. While the *unknown* words refer to the terms that only appear in the new text and has no matches in the document

collection. Identification of ideas is then made according to the balancing between *known* and *unknown* words.

However, this existing approach models the problem as an information retrieval problem, which relies on retrieving part of a text that potentially contains the pair of the unknown solution to a known problem (or known solution to the unknown problem). In other words, this existing approach of idea characterization is syntactical, and it lacks characterization of semantic relationships between terms in the new text and collection documents. We believe that considering the semantic dimension of examined words would contribute to improving the degree of balancing between *known* and *unknown* words. This is accomplished by the proposed balancing model that relies on characterizing the text as a triple of *known*, *SynWord*, and *unknown* terms.

The main aim of this research is to propose an idea mining model using a syntactic approach to extract the overlapping relations between terms that are not appearing in the matching process. It works by comparing part of the abstract with other text as a context text to find pairs of similar texts from the abstract and the context text. The (*known, unknown*, and *SynWord*) model is proposed to consider the semantic balancing between candidate text and description text. *SynWord* words in the proposed model refer to the terms that only existed in the query and not syntactically detected in the documents being searched, but there is a semantic relation between these words with the terms in the target documents. The processing of the standard idea mining framework is modified according to the new proposed balancing model. In contrast to the previous research, characterizing the *SynWord* attribute would help to characterize more candidate ideas effectively. The mean average precision is used in idea mining measurements and has achieved an overall MAP of (0.967) for identifying the idea which is comparatively better than the other approaches.

Furthermore, this research seeks to identify the pairs of text with similar and redundant content at higher ranks. Thus, this thesis attempts to improve the performance of the model by incorporating dissimilarity measure in the idea mining measurement to discriminate the redundancy in the text. The effectiveness of the measure is evaluated and the result is promising, showing that the proposed model can be more effective.

In addition, this research assumes that the text position within the abstract has a potential to be an effective feature for mining ideas. Therefore, this study investigates the impact of text position on the effectiveness of the idea mining method. In particular, modeling the text position measure is proposed by modifying the existing approaches to incorporate the weighting position method in the idea mining measurements. The proposed model enables calculating the importance of the position of the candidate idea based on the derived rules. Based on the observed results, applying rules in SynWord model achieved a MAP score of (0.967) which

showed that the conclusion section in the abstract has a higher chance to contain the idea as compared to the introduction and body sections.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PERMODELAN SEMANTIK LEKSIKEL ISTILAH BERDASARKAN KEPADA PENGENALAN SYNWORD UNTUK PERLOMBONGAN IDEA DALAM DAPATAN SEMULA MAKLUMAT

Oleh

MOSTAFA AHMED ALKSHER

Disember 2018

Pengerusi: Azreen Azman, PhD
Fakulti: Sains Komputer dan Teknolologi Maklumat

Pengumpulan data eksponen dalam maklumat digital yang berbentuk data perniagaan atau data awam telah membawa kepada cabaran yang besar tentang bagaimana cara untuk mengekstrak nilai yang lebih di dalam sesebuah data. Individu dan organisasi tidak lagi bergantung kepada semakan dan pengekstrakan data atau idea berguna daripada jumlah data digital yang besar kerana ianya mengambil masa yang agak lama untuk mengenal pasti idea-idea tersebut dalam jumlah maklumat teks yang banyak. Idea ini merupakan komponen yang penting dalam capaian maklumat dan memainkan peranan utama dalam Idea Perlombongan (IM) daripada teks yang tidak tersusun. Sesebuah idea ditakrifkan sebagai masalah dan penyelesaian (atau makna dan had) dalam konteks yang sama. IM telah diperkenalkan sebagai proses automatik bagi melancarkan idea-idea yang baru dan inovatif dari teks yang tidak berstruktur dengan menggunakan alat penambangan teks. Pada masa kini, banyak syarikat telah melabur dalam teknologi Perlombongan Teks (TM) untuk menemui maklumat berharga yang tersembunyi daripada teks yang tidak berstruktur di mana ianya sangat penting dalam pembuatan keputusan.

Walaupun semestinya tiada keraguan tentang idea-idea hebat yang tersembunyi dalam data awam dan data perniagaan yang besar ini, secara teknikalnya, cabaran utama adalah pencirian idea dan pemikiran. Pembentukan idea tradisional bergantung kepada pengenalpastian idea individu sama ada sebagai penyelesaian yang tidak diketahui kepada masalah yang diketahui atau penyelesaian yang diketahui kepada masalah yang tidak diketahui. Seterusnya pengecaman idea perlombongan dalam model ini membuatkan perbandingan tekstual antara teks yang baru (iaitu, permintaan input) dan pengumpulan dokumen. Hasil perbandin-

gan mestilah dalam bentuk perkataan yang tidak diketahui dan perkataan yang diketahui. Perkataan yang diketahui merujuk kepada terma-terma yang muncul dalam kedua-dua teks dan koleksi dokumen yang baru manakala perkataan yang tidak diketahui merujuk kepada istilah yang hanya muncul dalam teks baru dan tidak mempunyai padanan dalam pengumpulan dokumen. Pengenalpastian idea kemudian dibuat mengikut keseimbangan antara perkataan yang diketahui dan tidak diketahui.

Walau bagaimanapun, pendekatan yang sedia ada ini telah memodelkan masalah sebagai masalah pengambilan maklumat yang bergantung kepada pengambilan sebahagian teks yang berpotensi mengandungi pasangan penyelesaian yang tidak diketahui dengan masalah yang diketahui (atau penyelesaian yang diketahui kepada masalah yang tidak diketahui). Dalam erti kata yang lain, pendekatan yang sedia ada mengenai pencirian idea ini adalah sintaksis, dan ianya tidak mempunyai ciri-ciri hubungan semantik antara istilah dalam teks baru dan dokumen pengumpulan. Kami percaya dengan mengambil kira dimensi semantik dalam perkataan yang telah diperiksa akan menyumbang kepada peningkatan tahap pengimbangan antara perkataan yang diketahui dan tidak diketahui. Hal ini dapat dicapai dengan model keseimbangan yang telah dicadangkan dimana ianya bergantung kepada ciri teks iaitu diketahui, SynWord, dan istilah yang tidak diketahui.

Tujuan utama penyelidikan ini adalah untuk mencadangkan model perlombongan idea dengan menggunakan pendekatan sintaksis untuk mengekstrak pertindihan hubungan antara syarat yang tidak terdapat dalam proses yang sepadan. Ia berfungsi dengan membandingkan sebahagian daripada abstrak dengan teks lain sebagai teks konteks untuk mencari pasangan teks yang sepadan dari abstrak dan teks konteks. Model (yang diketahui, tidak diketahui, dan SynWord) dicadangkan untuk mempertimbangkan keseimbangan semantik di antara teks calon dan teks penerangan. Perkataan SynWord dalam model yang dicadangkan merujuk kepada istilah yang hanya wujud dalam pertanyaan dan tidak secara sintetik dikesan dalam dokumen yang dicari, tetapi terdapat hubungan semantik antara perkataan ini dengan istilah dalam dokumen sasaran. Proses rangka kerja perlombongan idea yang menepati piawaian diubah suai mengikut model keseimbangan yang baru. Berbeza dengan penyelidikan terdahulu, ciri-ciri SynWord akan membantu mencirikan lebih banyak idea lain. Purata ketepatan purata (MAP) telah digunakan dalam ukuran perlombongan idea dan mencapai MAP keseluruhan (0.967) untuk mengenal pasti idea yang lebih baik daripada pendekatan lain.

Selain itu, kajian ini bertujuan untuk mengenal pasti teks yang mengandungi kandungan yang sama dan berlebihan di peringkat yang lebih tinggi. Oleh itu, tesis ini dibuat untuk meningkatkan prestasi model dengan memasukkan ukuran ketidaksetaraan dalam ukuran perlombongan idea untuk membezakan kelebihan dalam teks. Keberkesanan langkah ini dinilai dan hasilnya menunjukkan bahawa model yang dicadangkan lebih efektif.

Di samping itu, kajian ini mengandaikan bahawa kedudukan teks dalam abstrak mempunyai potensi untuk menjadi ciri yang berkesan bagi idea perlombongan. Oleh sebab itu, kajian ini akan menyelidik kesan kedudukan teks dalam menguji keberkesanan kaedah perlombongan idea. Secara khususnya, model ukuran kedudukan teks telah dicadangkan dengan mengubahsuai pendekatan sedia ada untuk memasukkan kaedah kedudukan pemberat dalam ukuran perlombongan idea. Model yang dicadangkan ini membolehkan mengira kepentingan kedudukan idea berdasarkan syarat yang diperolehi. Berdasarkan hasil yang diperhatikan, peraturan yang digunakan dalam model SynWord mencapai skor MAP (0.967). Keputusan menunjukkan bahawa bahagian kesimpulan mempunyai peluang yang lebih tinggi untuk kandungan idea berbanding dengan pengenalan dan bahagian tubuh dalam abstrak.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Merciful, the Most Compassionate all praise be to Allah, the Lord of the worlds; and prayers and peace be upon (Mohamed) His servant and messenger.

First and foremost, I am grateful to my supervisor Dr. Azreen Azman,who worked hard with me from the beginning till the completion of the present research and for his patience, motivation, enthusiasm, and immense knowledge. His encouragement and help made me feel confident to overcome every difficulty I encountered in all the stages of this research. What I learned from him, however, is his attitude to work and life - always aiming for excellence.

I would like to extend my gratitude and thanks to the distinguished committee member, associate Professor Dr. Razali, Dr. Rabiah, and Dr. Abdulmajid Hussin for their encouragement and insightful comments.

I am very grateful to the Faculty of Computer Science and Information Technology and the staff of Postgraduate office, School of Graduate Studies, Library and Universiti Putra Malaysia, for providing me excellent research environment. Thanks to every person who has supported me to pursue and finish my Ph.D.

I am very grateful to my family, my father, my mother, my brothers, and my sisters for their dua'a and support throughout my life. I have no suitable words that can fully describe my everlasting love to them except, I love you all.

Words fail me to express my appreciation to my lovely wife, (Hajer), whose dedication, great sacrifice and persistent confidence in me helped me accomplish my degree. I owe her for being unselfishly let her intelligence, passions, and ambitions collide with mine. Special thank goes to my kids, Ahmed, Faris, Rawa and Joury, you have made me stronger, better and more fulfilled than I could have ever imagined. Thanks for giving me your valuable time through all this long process.

Last but not least, it gives me immense pleasure to express my deepest gratitude to my friends, colleagues and lab mates, especially Eissa Mohamed Alshari for being pillars of support, offering very crucially, together with guidance and direction on the research and dissertation in general whenever I was buffeted by the winds of doubt and uncertainty.

Finally, deepest thanks go to all people who took part in making this thesis real, as well as I express my apology that I could not mention them all personally.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Azreen Azman, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Razali Yaakob, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Rabiah Abdul Kadir, PhD
Research Fellow
Institute of Visual Informatics
Universiti Kebangsaan Malaysia
(Member)

Abdulmajid Mohamed, PhD
Faculty of Computer Science
Sebha University
Libya
(Member)

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: _____ Date: _____

Name and Matric No.: Mostafa Ahmed Alksher (GS36936)

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____
Name of Chairman
of Supervisory
Committee: Dr. Azreen Azman

Signature: _____
Name of Member
of Supervisory
Committee: Assoc. Prof. Dr. Razali Yaakob

Signature: _____
Name of Member
of Supervisory
Committee: Dr. Rabiah Abdul Kadir

Signature: _____
Name of Member
of Supervisory
Committee: Dr. Abdulmajid Mohamed

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Innovation has become the key to the success of many organizations or nations in order to be competitive in the real world. It is driven by the capability of its member or citizen to generate an interesting idea and making it work. Brainstorming has been used as an effective idea generation technique for decades (Kudrowitz and Wallace, 2013). However, it is both expensive and challenging creative process to discover interesting ideas in order to solve a problem or to assist in decision making. In the process, textual resources such as scientific publications and the Web have been utilized as the source for the idea (Thorleuchter et al., 2010c; Thorleuchter and Van den Poel, 2013b).

Idea mining (IM) is an interesting field in the areas of information retrieval (IR) due to the growing need to automatically extract information from text. It is introduced as an automatic mining of new and innovative ideas from unstructured text by using text-mining tools (Hotho et al., 2005). Nowadays, many companies have invested in text mining (TM) technology to discover hidden data from unstructured text. The goal of the text mining is to filter out meaningless terms, process the significant information, and extract the concept of terms latent within text or document using techniques from IR (Manning et al., 2008a), information extraction (IE) as well as natural language processing (NLP) (Aggarwal and Zhai, 2012). Moreover, text mining techniques can also assist researchers to do the analytical process, such as feature selection, text classification, summarization, clustering, the topic identification, and information mapping (Tseng et al., 2007).

Mining latent ideas from texts that might be able to solve existing problem are very important for decision-making. However, it is time-consuming to identify potential ideas from the huge textual information while only a few of them might be relevant to a current decision problem. The implication of the term's meaning and it's relation to other terms which represent the idea should be determined to improve the identification of new ideas. Efficient methods are required for mining and extracting ideas from documents and corpus. The main aim of this thesis is to propose an idea mining model using a syntactic lexicon approach to extract the overlapping relations between terms that are not appearing in the matching process.

The idea mining identifier compares a textual pattern from a new text to all

textual patterns from the given context information. It uses several parameters for the classification decision. This existing approach is taken over from psychology and cognitive science and follows how users create ideas. In order to realize the processing, methods from text mining and text classification (tokenization, term filtering methods, Euclidean distance measure, alpha cut method, etc.) are used and combined with a new proposed measure for mining ideas. The different measures that are proposed for the mining of the idea in contrast to the previous work will effectively characterize the idea based on the relationships between the selected terms.

The aim of idea mining (IM) is to identify and retrieve data from natural language text. Idea mining model is much more manageable that extracts the information from the text and presents them to the user instead of returning a link to a document. Idea mining is introduced by (Thorleuchter et al., 2010c) as an automatic process of mining ideas from the new text provided by the user and evaluate them concerning their ability to solve the problems described in the document collections. Therefore, specific idea mining measures model (*known/unknown*) terms as to be well balanced based on their co-occurrence in text pattern. The balancing measure of *known* and *unknown* terms is considered to be the backbone of extracting new idea field and idea mining is highly depending on the comparison between these terms. As shown in Figure 1.1, the overlapping between terms in *idea text* and *context text*.

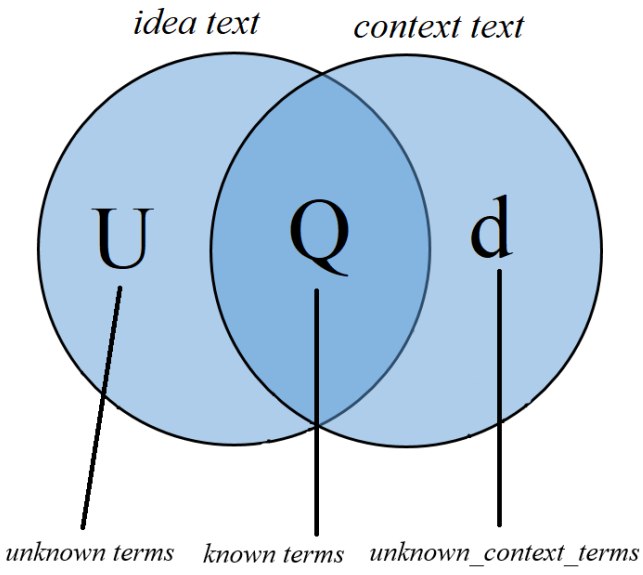

Figure 1.1 : The venn diagram of the overlapping of terms in *idea text* and *context text*

In this thesis, mining idea model using a syntactic lexicon approach to extract the

overlapping relations between terms is proposed. The new model will consider the semantic balancing between terms in both query and documents to characterize more attributes that would help to compose more candidate ideas effectively. On the other hand, new idea mining measurements will be used to measure the position of the idea from the text as an added value to improve the performance of idea mining process in data-intensive domains.

1.2 Research Motivation

Today, a huge amount of textual information is accessible which covered different topics, and it could be a valuable source for decision makers because it consists of many interesting ideas, which possibly are relevant to solve a current decision problem. However, it is time-consuming to identify these ideas within a large number of user's contents within the text for the human experts. This is because the expert has to scan all the relevant texts for the occurrence of interesting ideas.

This research aims to help users extract relevant information which is formed as an idea from large data collections. These ideas probably are valuable for producers as well as for researchers and developers. This is because they can lead to a new product development process or new research contribution. Another interesting motivation is to investigate the feasibility of employing text position as an additional feature for idea identification, which is not considered in the related works.

1.3 Problem Statement

Mining idea research is still in primary development; it still needs to develop new models to represent individual ideas that have features that make the mining process very effective and efficient. Features used in existing idea models are very limited, and it does not cater for mining ideas very accurately.

The traditional formulation of ideas is based on studies in psychology (Rohpohl, 1996), it depends on identifying an individual idea either as a pair of *mean* and *purpose* (Thorleuchter, 2008), *problem* and *solution* (Liu et al., 2015b), *request* and *known* terms (Dinh et al., 2015) or as *events* and their *relations* (Wang and Ohsawa, 2013). The existing idea mining identifier makes a textual comparison between input query and the collection documents to produce a text in the form of *unknown* words and *known* words (Thorleuchter et al., 2010c). Similar research in (Thorleuchter and Van den Poel, 2013b) proposed an idea web mining approach of finding dissimilarity of text terms between a problem description and a problem

solution idea. Identification of ideas is then made according to the balancing between *known* and *unknown* words.

Furthermore, these existing approaches model the problem as an information retrieval problem, which rely on retrieving part of a text that potentially contains the pair of unknown solution to a known problem (or known solution to unknown problem).

The research in (Thorleuchter and Van den Poel, 2015) combined the idea indication (weak signal analysis) with idea mining to filter the ideas using semantic clustering (LSI). However, this work is limited to recognize the low information content that neglected further text that could represent ideas. In other words, these existing approaches of idea characterization are syntactical, and they lack characterization of semantic relationships between terms in the new text and collection documents.

A study on idea discovery through data synthesis proposed a dynamic process of idea discovery to turn data into scenario maps by clustering the eliciting human insights (Wang and Ohsawa, 2013). Whereas Liu et al. (2015b) tended to extract noun-phrases within the titles and abstracts of the publications and claiming that the noun-phrases are sufficient to represent candidate ideas using n-grams model. Christensen et al. (2017) proposed a classification method to automatically detect text as an idea or none idea in online communities using machine learning and text mining techniques.

Typically, all approaches above are based on the classical Bag-of-Words model where each term is an independent feature. We believe that one of the problems with the existing models lie in the text pattern where the redundant text is most likely to emerge. The existing model retrieved a redundant text from *context text* where these redundancy affect the selection of the most appropriate candidate idea. Redundant text pattern would compromise the effectiveness of the ranking due to the repetition of the text patterns.

Furthermore, the model is considered noisy, and it would not lead to extracting enough ideas because it ignores the relationships between words in the extraction process of the *text pattern* set, which would be restricted to using limited text pattern set. Based on the literature, no attempt was made to exploit the overlapping relations between terms in the idea matching process.

In addition, one of the greater challenges of idea extraction is the huge number of *context text*, and the scientific paper's domain is one of the data-rich domains. The scientific papers have systematic structure, and that allows some parts to contain the idea. The existing of the idea can be resided in certain part within

the structure. However, most of the studies did not simultaneously examine the effect of the idea position within the text. The research problems are addressed as follows:

- The idea mining model is considered noisy, since the measurement depends on known and unknown words. Some of the known words are semantically related to some of the unknown words, however, the semantic relationships between (known & unknown) words were not considered (Thorleuchter et al., 2010c; Thorleuchter and Van den Poel, 2013b).
- The existing models depend on the related pair of text patterns by applying the Euclidean distance measure, as such, the redundant text is most likely to emerge on the top of the ranking results. However, the redundant text patterns would compromise the effectiveness of the ranking due to the redundancy of the selected pairs (Thorleuchter et al., 2010c,d).
- Based on the prior study, scientific papers have systematic structure and the idea could be located in certain position of texts (Alksher et al., 2018). The positions of the text patterns within the original text has not been considered as a feature in the existing models (Thorleuchter et al., 2010c). The characteristic terms values that represented by the sub measure m_c is introduced in existing models but has not been considered (Thorleuchter and Van den Poel, 2013b).

1.4 Research Objectives

The primary goal of this research work is to propose a new approach for improving the automatic idea mining from the text. The following objectives are set to accomplish this.

1. To incorporate lexical relation among *known* and *unknown* terms by proposing a novel idea measurement in idea mining.
2. To propose the dissimilarity model between *idea text* and *context text* for eliminating the effect of redundancy.
3. To incorporate the position of the text for idea identification to improve the effectiveness of idea mining measurement.

1.5 Research Scope

Nowadays, the scientific publication is attested an abundance, which would be a challenge for searchers to follow up on what is published. In this research, scientific corpus such as retrieval of scientific research papers, particularly abstracts are used. It is reasonable to focus on the abstract for identifying the idea from the text since it has always been an important part of science research and written systematically based on a certain structure that makes it easier for a human to identify the idea. The abstract is usually self-contained, and the most important information discussed in the paper are summarized in it. In reality, the majority of readers consider the abstract as the important part of the paper to be viewed when they search for potential idea (Andrade, 2011). Researchers publish their ideas and results by publishing their works, as well as consult the literature to keep them aware of what is going on in their field (Lawrence et al., 1999). The most attention in this research is to characterize the feature of the idea within abstracts since it performs extraction process at a finer level.

In this study, the maximum length of non-stopword to create the text pattern is sit to 8. For this, the text patterns should not be too small so that they contain all terms representing a new idea. Also, text patterns should not be too large so that only terms occur in the text patterns that are related to the new idea.

The limitation of this model is to focus on the English language because analyzing English texts with text mining methods is standard. Furthermore, The definition of the potential ideas are derived from the technique philosophy (Rohpohl, 1996) as consists of several terms that are not previously known by discovering the relationship between them and can be identified using text mining and text classification methods.

1.6 Significance of Study

The significant gain of this research is reflected in promoting assistance to establish a system that is useful in the following:

- Easily extracts embedded ideas from scientific publications.
- Automated extraction technique of identifying ideas and overcome the inconsistencies in the manual evaluation.
- Model the idea characteristic by considering the semantic balancing between query and document.

- Model the content of the idea semantically helps to mine the idea by formulating the textual pattern.
- Determine the implication of the term's meaning, and it's relation to other terms semantically to improve the measurements and identify new ideas with better performance.
- Detect relevant ideas among the large amount of textual information which helps strategic planners to consider future impacts on their strategic decision by time.

1.7 Research Contribution

The contributions of the proposed research are obvious as follows:

- Proposed an enhanced idea mining model that defines new attributes for lexically-active idea mining process.
- The construction of formally built dataset for idea mining from research publications.
- The development of a new mathematical model to improve the balancing complexity of the idea mining process.
- The development of idea positioning approach to improving the performance of IM process.

1.8 Thesis Organization

This thesis is divided into seven chapters as depicted in Figure 1.2. The detailed description of each chapter is presented as follows:

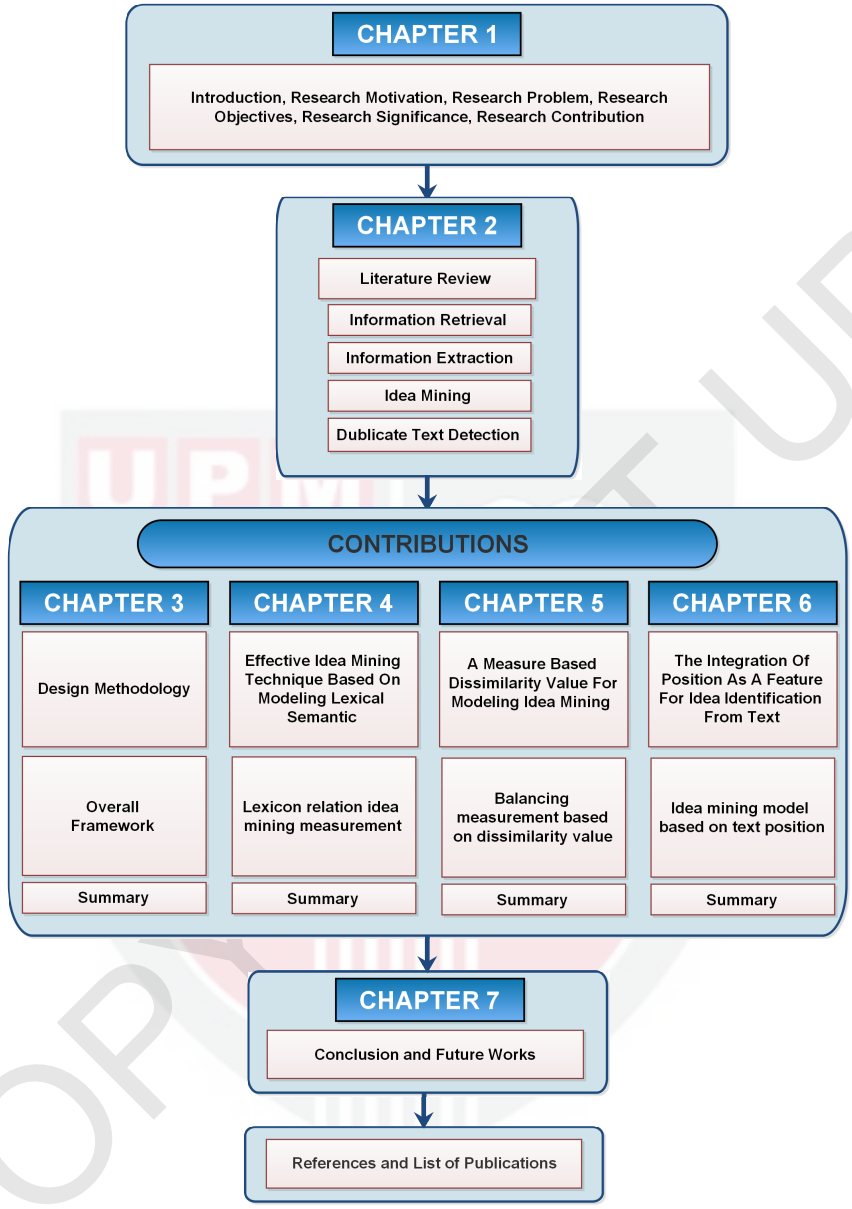

Figure 1.2 : Organization of the thesis

Chapter 1 introduced the introduction, motivation, problem statement, objectives, contribution to the research and concluded in the organization of the thesis.

Chapter 2 highlights the research motivation by giving an overview of related research and try to position this work therein. The important concepts and terminology in information retrieval that complete the understanding of idea mining and how others have approached it are reviewed in this chapter.

Chapter 3 provides the overall methodology with a detailed framework for the description of the proposed automated idea mining system.

Chapter 4 tackles the problem of characterizing the idea through defining new attributes by exploring a lexicon approach based on semantic relationships between words.

Chapter 5 highlights the mathematical model used to combine all the parameters of the idea mining measurements. This chapter focuses on the effects of the similarity attribute computed with the multi-balancing measure as to be included in the set of idea mining measurements. The cosine similarity is presented as an efficient attribute that will make use of the model parameters to improve the performance of the proposed model.

Chapter 6 focuses on the feasibility of employing position as an additional feature for idea identification to improve the performance of idea mining process.

Chapter 7 concludes the overall research work and followed by directions for future work.

1.9 Summary

This research is an integration of information retrieval and information extraction and highly contributes to identify the useful ideas based on lexical context. This chapter presents the essence of the thesis; issues faced the motivation for this work and the main contributions. In the next chapters, more details will be given to the techniques, models and the experimental analysis.

REFERENCES

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media. 1, 23

Alksher, M., Azman, A., Yaakob, R., Alshari, E. M., Rabiah, A. K., and Mohamed, A. (2018). Effective idea mining technique based on modeling lexical semantic. *Journal of Theoretical and Applied Information Technology*, 96(16):5350–5362. xvii, 5, 47

Alksher, M. A., Azman, A., Yaakob, R., Kadir, R. A., Mohamed, A., and Alshari, E. (2017). A framework for idea mining evaluation. In *16th International Conference on New Trends in Intelligent Software Methodology Tools, and Techniques, SoMeT 2017*. IOS Press. 61

Alksher, M. A., Azman, A., Yaakob, R., Kadir, R. A., Mohamed, A., and Alshari, E. M. (2016). A review of methods for mining idea from text. In *2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp. 88–93. IEEE. xvii, 34

Alrshah, M. A., Othman, M., Ali, B., and Hanapi, Z. M. (2014). Comparative study of high-speed linux tcp variants over high-bdp networks. *Journal of Network and Computer Applications*, 43:66–75. 81

Alshari, E. M. (2015). Semantic arabic information retrieval framework. *arXiv preprint arXiv:1512.03165*. 73

Alshari, E. M., Azman, A., Mustapha, N., Doraisamy, S. C., and Alksher, M. (2016). Prediction of rating from comments based on information retrieval and sentiment analysis. In *Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on*, pp. 32–36. IEEE. 41

Andrade, C. (2011). How to write a good abstract for a scientific paper or conference presentation. *Indian journal of psychiatry*, 53(2):172. 6, 42, 87

Andrews, N. O. and Fox, E. A. (2007). Recent developments in document clustering. *Computer Science Technical Reports*. 13, 15, 20

Aryal, S., Ting, K. M., Haffari, G., and Washio, T. (2014). mp-dissimilarity: A data dependent dissimilarity measure. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pp. 707–712. IEEE. 40, 82

Atanassova, I., Bertin, M., and Larivière, V. (2016). On the composition of scientific abstracts. *Journal of Documentation*, 72(4):636–647. 41

Azmi-Murad, M. and Martin, T. P. (2004). Using fuzzy sets in contextual word similarity. In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 517–522. Springer. 11

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York. xvii, 17

Bagga, A., Chai, J., and Biermann, A. (2001). Extracting information from text. *Computing with Words. New York: Wiley*, pp. 209–234. xvii, 23, 24

Baghela, V. and Tripathi, S. (2012). Text mining approaches to extract interesting association rules from text documents. *International Journal of Computer Science Issues*, 9(3):545–552. 30

Balwinder, S. and Vikram, S. (2014). An effective pre-processing algorithm for information retrieval systems. 16

Banko, M. and Etzioni, O. (2008). The tradeoffs between open and traditional relation extraction. *Proceedings of ACL-08: HLT*, pp. 28–36. 23

Barker, K. and Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 40–52. Springer. 37

Baumgartner, F. R. (2014). Ideas, paradigms and confusions. *Journal of European Public Policy*, 21(3):475–480. 26

Belkin, N. J. (1996). Intelligent information retrieval: whose intelligence? *ISI*, 96:25–31. 11

Bogdanova, D., Rosso, P., and Solorio, T. (2012). On the impact of sentiment and emotion based features in detecting online sexual predators. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 110–118. Association for Computational Linguistics. 14

Bora, N. N. (2012). Summarizing public opinions in tweets. *International Journal of Computational Linguistics and Applications*, 3(1):41–55. 14

Broder, A. Z. (2000). Identifying and filtering near-duplicate documents. In *Annual Symposium on Combinatorial Pattern Matching*, pp. 1–10. Springer. 40

Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 724–731. Association for Computational Linguistics. 38

Cai, Y., Zhang, Q., Lu, W., and Che, X. (2017). A hybrid approach for measuring semantic similarity based on ic-weighted path distance in wordnet. *Journal of Intelligent Information Systems*, pp. 1–25. 38, 39

Carus, A. B. (1999). Method and apparatus for improved tokenization of natural language text. US Patent 5,890,103. 14

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., and Quarteroni, S. (2013). An introduction to information retrieval. In *Web information retrieval*, pp. 3–11. Springer. 22, 73

Cerulo, L. and Canfora, G. (2004). A taxonomy of information retrieval models and tools. *Journal of Computing and Information Technology*, 12(3):175–194. 25, 46

Chan, Y. S. and Roth, D. (2010). Exploiting background knowledge for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 152–160. Association for Computational Linguistics. 23

Chang, L. and Kashyap, R. L. (2013). Evidence combination and reasoning and its application to real-world problem-solving. *arXiv preprint arXiv:1304.1125*. 95

Chowdhury, A., Frieder, O., Grossman, D., and McCabe, M. C. (2002). Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, 20(2):171–191. 39, 40

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. Facet publishing. 16

Christensen, K., Nørskov, S., Frederiksen, L., and Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creativity and Innovation Management*, 26(1):17–30. 4, 26, 31, 34, 38, 40, 87

Consoli, D. (2010). A new framework to extract knowledge by text mining tools. *Bilgi Ekonomisi ve Yönetimi Dergisi*, 5(2). 25

Coussement, K. and Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1):313–327. 62

Crawford, C. M. (2008). *New products management*. Tata McGraw-Hill Education. 26

Croft, W. B. (1993). Knowledge-based and statistical approaches to text retrieval. *IEEE Expert*, 8(2):8–12. xvii, 12, 13

Croft, W. B., Metzler, D., and Strohman, T. (2010). *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading. 73

Dhondt, J., Vertommen, J., Verhaegen, P.-A., Cattrysse, D., and Duflou, J. R. (2010). Pairwise-adaptive dissimilarity measure for document clustering. *Information Sciences*, 180(12):2341–2358. 40

Di Gangi, P. M. and Wasko, M. (2009). Steal my idea! organizational adoption of user innovations from a user innovation community: A case study of dell ideastorm. *Decision Support Systems*, 48(1):303–312. 25

Di Gangi, P. M., Wasko, M. M., and Hooker, R. E. (2010). Getting customers'ideas to work for you: Learning from dell how to succeed with online user innovation communities. *MIS Quarterly Executive*, 9(4). 87

Dinh, T.-C., Bae, H., Park, J., and Bae, J. (2015). A framework to discover potential ideas of new product development from crowdsourcing application. *arXiv preprint arXiv:1502.07015*. 3, 14, 26, 31, 33, 36, 38, 40, 78, 87

Doró, K. (2013). The rhetoric structure of research article abstracts in english studies journals. *Prague Journal of English Studies*, 2(1):119–139. 41

Driscoll, J. R. (1997). Method and system for searching for relevant documents from a text database collection, using statistical ranking, relevancy feedback and small pieces of text. US Patent 5,642,502. 10

Elabd, E., Alshari, E., and Abdulkader, H. (2015). Semantic boolean arabic information retrieval. *arXiv preprint arXiv:1512.03167*. 48

Emad, E., Alshari, E. M., and Abdulkader, H. (2013). Arabic vector space model based on semantic. *International journal of computer science (IJISI)*, 8(6):94–101. 20, 47

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 1535–1545. Association for Computational Linguistics. 23

Fairthorne, R. (1956). The patterns of retrieval. *Journal of the Association for Information Science and Technology*, 7(2):65–70. 11

Faloutsos, C. and Oard, D. W. (1998). A survey of information retrieval and filtering methods. Technical report. 55, 72

Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, 49(9):76–82. 13, 23, 45

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pp. 231–243. Springer. 37

Finzen, J., Kintz, M., and Kaufmann, S. (2012). Aggregating web–based ideation platforms. *International Journal of Technology Intelligence and Planning*, 8(1):32–46. 31

Gaeta, M., Loia, V., Mangione, G. R., Orciuoli, F., and Ritrovato, P. (2011). Social semantic web fosters idea brainstorming. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, volume 730, pp. 55–66. Citeseer. 26

Garcia, E. (2006). Description, advantages and limitations of the classic vector space model. *2006*. 21

Geum, Y. and Park, Y. (2016). How to generate creative ideas for innovation: a hybrid approach of wordnet and morphological analysis. *Technological Forecasting and Social Change*, 111:176–187. 37, 40

Giunchiglia, F., Kharkevich, U., and Zaihrayeu, I. (2010). Concept search: Semantics enabled information retrieval. Technical report, University of Trento. 11

Goby, N., Brandt, T., Feuerriegel, S., and Neumann, D. (2016). Business intelligence for business processes: the case of it incident management. In *ECIS*, p. ResearchPaper151. 10

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*. 12

Graetz, N. (1982). Teaching efl students to extract structural information from abstracts. *The International Symposium on Language for Special Purposes*. 41

Gu, Y., Storey, V. C., and Woo, C. C. (2015). Conceptual modeling for financial investment with text mining. In *International Conference on Conceptual Modeling*, pp. 528–535. Springer. 10

Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., and Stenius, U. (2010). Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp. 99–107. Association for Computational Linguistics. 42, 87

Haav, H.-M. and Lubi, T.-L. (2001). A survey of concept-based information retrieval tools on the web. In *Proceedings of the 5th East-European Conference ADBIS*, volume 2, pp. 29–41. 12

Hartley, J. (2003). Improving the clarity of journal abstracts in psychology: the case for structure. *Science Communication*, 24(3):366–379. 42

Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 415. Association for Computational Linguistics. 38

Herrera-Viedma, E. (2001). An information retrieval model with ordinal linguistic weighted queries based on two weighting elements. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(supp01):77–87. 16, 34

Herrera-Viedma, E. and López-Herrera, A. G. (2007). A model of an information retrieval system with unbalanced fuzzy linguistic information. *International Journal of Intelligent Systems*, 22(11):1197–1214. 12

Hiemstra, D. (2009). Information retrieval models. *Information Retrieval: searching in the 21st Century*, pp. 1–17. 47

Hofmann, T. (2017). Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, volume 51, pp. 211–218. ACM. 21

Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pp. 19–62. 1, 20, 46, 48

Hozack, W. J. (2016). Structured abstracts. *The Journal of arthroplasty*, 31(3):561. 41

Ibekwe-Sanjuan, F., Silvia, F., Eric, S., and Eric, C. (2011). Annotation of scientific summaries for information retrieval. *arXiv preprint arXiv:1110.5722*. 37

Ilmola, L. and Kuusi, O. (2006). Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision-making. *Futures*, 38(8):908–924. 10

Jessop, J. L. (2002). Expanding our students' brainpower: Idea generation and critical thinking skills. *IEEE Antennas and Propagation Magazine*, 44(6):140–144. 26

Jiang, J. (2012). Information extraction from text. In *Mining text data*, pp. 11–41. Springer. 23

Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938. 16

Karbasi, S. and Boughanem, M. (2006). Document length normalization using effective level of term frequency in large collections. In *European Conference on Information Retrieval*, pp. 72–83. Springer. 15

Khan, K., Baharudin, B., Khan, A., and Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. *Journal of King Saud University-Computer and Information Sciences*, 26(3):258–275. 10, 25

Kilgarriff, A. (2000). Wordnet: An electronic lexical database. 37

Kim, Y. (2006). Toward a successful crm: variable selection, sampling, and ensemble. *Decision Support Systems*, 41(2):542–553. 62

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information processing & management*, 41(3):433–455. 47

Ko, Y. and Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1):70–83. 31

Kontostathis, A. and Pottenger, W. M. (2006). A framework for understanding latent semantic indexing (lsi) performance. *Information Processing & Management*, 42(1):56–73. 49

Korobkin, D. M., Fomenkov, S. A., Kolesnikov, S. G., and Golovanchikov, A. B. (2016). Technical function discovery in patent databases for generating innovative solutions. In *Multi Conference on Computer Science and Information Systems*, volume 2016, p. 241. 30, 34

Kostoff, R. N. (2011). Literature-related discovery: Potential treatments and preventatives for sars. *Technological Forecasting and Social Change*, 78(7):1164–1173. 25

Kruse, P., Schieber, A., Hilbert, A., and Schoop, E. (2013). Idea mining–text mining supported knowledge management for innovation purposes. *Americas Conference on Information Systems (AMCIS), 2013 Nineteenth Americas Conference*. 10, 38, 87

Kudrowitz, B. M. and Wallace, D. (2013). Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design*, 24(2):120–139. 1, 26

Kumar, J. P. and Govindarajulu, P. (2009). Duplicate and near duplicate documents detection: A review. *European Journal of Scientific Research*, 32(4):514–527. 39

Lamiroy, B. and Sun, T. (2013). Computing precision and recall with missing or uncertain ground truth. In *Graphics Recognition. New Trends and Challenges*, pp. 149–162. Springer. 21, 35

Lancaster, F. W., Lancaster, F. W., Lancaster, F. W., and Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice*. Library Association London. 42

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211. 26

Lawrence, S., Bollacker, K., and Giles, C. L. (1999). Indexing and retrieval of scientific literature. In *Proceedings of the eighth international conference on Information and knowledge management*, pp. 139–146. ACM. 6

Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283. 12, 38

Lee, D. L., Chuang, H., and Seamons, K. (1997). Document ranking and the vector-space model. *IEEE software*, 14(2):67–75. 16, 47

Lee, T.-Y. (2012). A study on extracting ideas from documents and webpages in the field of idea mining. *Journal of the Korean Society for information Management*, 29(1):25–43. 15, 25

Leibovici, L. (2017). Structured abstracts for narrative reviews. *Clinical Microbiology and Infection*, 23(7):423. 41

Levi, K. (1989). Expert systems should be more accurate than human experts: evaluation procedures from human judgement and decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(3):647–657. 58, 59

Lin, Y.-S., Jiang, J.-Y., and Lee, S.-J. (2014). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, 26(7):1575–1590. 20, 35, 92

Litchfield, R. C., Gilson, L. L., and Gilson, P. W. (2015). Defining creative ideas: Toward a more nuanced approach. *Group & Organization Management*, 40(2):238–265. 28

Liu, B., An, X., and Huang, J. X. (2015a). Using term location information to enhance probabilistic information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 883–886. ACM. 41

Liu, H., Goulding, J., and Brailsford, T. (2015b). Towards computation of novel ideas from corpora of scientific text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 541–556. Springer. 3, 4, 26, 28, 29, 31, 33, 38, 64, 87

Liu, K., Chapman, W., Savova, G., Chute, C., Sioutos, N., and Crowley, R. S. (2011). Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. *Methods of information in medicine*, 50(05):397–407. 37

Ma, M. and An, J. (2015). Combination of evidence with different weighting factors: a novel probabilistic-based dissimilarity measure approach. *Journal of Sensors*, 2015. 79

Maimon, O. and Rokach, L. (2005). Decomposition methodology for knowledge discovery and data mining. *Data mining and knowledge discovery handbook*, pp. 981–1003. 87

Manning, C. D., Raghavan, P., and Schütze, H. (2008a). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. 1, 65

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008b). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge. 16, 21, 22

Matters, S. (2016). Intra-rater and inter-rater reliability of the aspca's behavior evaluation of fearful dogs. 59

Mendonça, S., Cardoso, G., and Caraça, J. (2012). The strategic strength of weak signal analysis. *Futures*, 44(3):218–228. 10

Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pp. 775–780. 38

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41. 37

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244. 39

Mooers, C. N. (1950). Information retrieval viewed as temporal signaling. In *Proceedings of the International Congress of Mathematicians*, volume 1, pp. 572–573. 11

Moohebat, M., Raj, R. G., Kareem, S. B. A., and Thorleuchter, D. (2015). Identifying isi-indexed articles by their lexical usage: A text analysis approach. *Journal of the Association for Information Science and Technology*, 66(3):501–511. 58

Müller, C. and Gurevych, I. (2009). A study on the semantic relatedness of query and document terms in information retrieval. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pp. 1338–1347. Association for Computational Linguistics. 12

Murad, M. A. A. and Martin, T. (2007). Similarity-based estimation for document summarization using fuzzy sets. *International Journal of Computer Science and Security*, 1(4):1–12. 18

Nijstad, B. A. and Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and social psychology review*, 10(3):186–213. 26

Ojo, A. and Adeyemo, A. (2012). Framework for knowledge discovery from journal articles using text mining techniques. *African Journal of Computing & ICT*, 5. 17, 25

Orasan, C. (2001). Patterns in scientific abstracts. In *Proceedings of Corpus Linguistics 2001 Conference*, pp. 433–443. 41, 42

Osborn, A. (2008). *Your creative power: How to use your imagination to brighten life, to get ahead*. University Press of America. 34

Osborn, A. F. (1953). Applied imagination, principles and procedures of creative thinking. 26

Özyirmidokuz, E. K. and Özyirmidokuz, M. H. (2014). Analyzing customer complaints: A web text mining application. *International Conference on Education and Social Sciences*. 31, 33

Paltoglou, G. and Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 1386–1395. Association for Computational Linguistics. 18

Piezunka, H. and Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, 58(3):856–880. 87

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137. 15, 45

Ramasubramanian, C. and Ramya, R. (2013). Effective pre-processing activities in text mining using improved porter's stemming algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(12):2278–1021. xvii, 14, 15

Razzak, F. (2012). Spamming the internet of things: A possibility and its probable solution. *Procedia computer science*, 10:658–665. 10

Ren, F. and Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. *Information Sciences*, 236:109–125. 18

Ren, K., Zhang, S., and Lin, H. (2012). Where are you settling down: Geolocating twitter users based on tweets and social networks. In *Asia Information Retrieval Symposium*, pp. 150–161. Springer. 17

Reynaud, C. and Safar, B. (2007). Exploiting wordnet as background knowledge. In *Proceedings of the 2nd International Conference on Ontology Matching-Volume 304*, pp. 291–295. CEUR-WS. org. 37

Richardson, R., Smeaton, A., and Murphy, J. (1994). Using wordnet as a knowledge base for measuring semantic similarity between words. 37

Riedl, C., May, N., Finzen, J., Stathel, S., Kaufman, V., Krcmar, H., et al. (2009). An idea ontology for innovation management. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(4):1–18. 28, 32

Rohpohl, G. (1996). Das ende der natur. *Naturauffassungen in Philosophie, Wissenschaft und Technik*, pp. 143–163. 3, 6, 29

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., et al. (2007). Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2):195–200. 91

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*. 54

Salton, G., Allan, J., and Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108. 17, 47, 48

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523. 17, 18, 40

Salton, G. and Harman, D. (2003). *Information retrieval*. John Wiley and Sons Ltd. 15

Salton, G. and McGill, M. (1983). Introduction to modern information philadelphia, pa. american association for artificial intelligence retrieval. 40

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620. 29, 39

Sanderson, M. (1997). Duplicate detection in the reuters collection. *" Technical Report (TR-1997-5) of the Department of Computing Science at the University of Glasgow G12 8QQ, UK"*. 39

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of extension*, 37(2):1–5. 61

Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43. 16

statistics, W. (2017). Wnstats(7wn) manual page. `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`. (Accessed on 11/03/2017). 37

Stoica, E. A. and Özyirmidokuz, E. K. (2015). Mining customer feedback documents. *International Journal of Knowledge Engineering*, 1(1):68–71. 47

Sun, Y., Deng, H., and Han, J. (2012). Probabilistic models for text mining. In *Mining text data*, pp. 259–295. Springer. 17

Swanson, D. (2008). Literature-based discovery? the very idea. *Literature-based discovery*, pp. 3–11. 28

Tabatabaei, N. (2011). Detecting weak signals by internet-based environmental scanning. Master's thesis, University of Waterloo. 20

Tami, G. and Gallagher, A. (2009). Description of the behaviour of domestic dog (canis familiaris) by experienced and inexperienced people. *Applied Animal Behaviour Science*, 120(3):159–169. 61

Theobald, M., Siddharth, J., and Paepcke, A. (2008). Spotsigs: robust and efficient near duplicate detection in large web collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 563–570. ACM. 39

Thomas, P. and Hawking, D. (2006). Evaluation by comparing result sets in context. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 94–101. ACM. 21

Thorleuchter, D. (2008). Finding new technological ideas and inventions with text mining and technique philosophy. In *Data Analysis, Machine Learning and Applications*, pp. 413–420. Springer. 3, 26, 32, 35, 44, 50, 64, 81, 92

Thorleuchter, D., Herberz, S., and Van den Poel, D. (2011). Mining social behavior ideas of przewalski horses. In *Advances in Computer, Communication, Control and Automation*, pp. 649–656. Springer. 25, 29, 44, 48, 54, 78

Thorleuchter, D., Scheja, T., and Van den Poel, D. (2014). Semantic weak signal tracing. *Expert Systems with Applications*, 41(11):5009–5016. 31

Thorleuchter, D. and Van den Poel, D. (2012). Extraction of ideas from microsystems technology. *Advances in Computer Science and Information Engineering*, pp. 563–568. 10, 25, 30, 32, 38, 48, 63, 64

Thorleuchter, D. and Van den Poel, D. (2013a). Analyzing website content for improved r&t collaboration planning. In *Advances in Information Systems and Technologies*, pp. 567–573. Springer. 14

Thorleuchter, D. and Van den Poel, D. (2013b). Web mining based extraction of problem solution ideas. *Expert Systems with Applications*, 40(10):3961–3969. 1, 3, 5, 10, 15, 17, 21, 29, 30, 33, 34, 36, 38, 40, 48, 49, 58, 63, 65, 66, 87, 93

Thorleuchter, D. and Van den Poel, D. (2015). Idea mining for web-based weak signal detection. *Futures*, 66:25–34. 4, 14, 18, 29, 31, 33, 64, 78, 87

Thorleuchter, D. and Van den Poel, D. (2016). Identification of interdisciplinary ideas. *Information Processing & Management*, 52(6):1074–1085. 26, 36, 47, 62, 63, 64, 65, 93

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010a). A compared r&d-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological forecasting and social change*, 77(7):1037–1050. 14, 25, 35

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010b). Extracting consumers needs for new products-a web mining approach. In *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*, pp. 440–443. IEEE. 15, 17, 25, 27, 47, 49, 78

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010c). Mining ideas from textual information. *Expert Systems with Applications*, 37(10):7182–7188. 1, 2, 3, 5, 15, 20, 25, 29, 30, 31, 32, 35, 36, 38, 40, 44, 46, 47, 49, 51, 52, 58, 64, 65, 69, 72, 76, 78, 79, 81, 82, 83, 87, 92, 93

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010d). Mining innovative ideas to support new product research and development. In *Classification as a Tool for Research*, pp. 587–594. Springer. 5

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010e). Mining innovative ideas to support new product research and development. In *Classification as a Tool for Research*, pp. 587–594. Springer. 25

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in b-to-b marketing. *Expert systems with applications*, 39(3):2597–2605. 31

Toubia, O. and Netzer, O. (2016). Idea generation, creativity, and prototypicality. *Marketing science*, 36(1):1–20. 26

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247. 1

Ufnalska, S. and Hartley, J. (2009). How can we evaluate the quality of abstracts. *European Science Editing*, 35(3):69–72. 42

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472. 26

Van de Cruys, T. (2010). Mining for meaning. the extraction of lexico-semantic knowledge from text. *Groningen Dissertations in Linguistics*, 82. 20

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, pp. 1–14. 11

Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 33(2):106–119. 11

Vijayarani, S. and Janani, R. (2016). Text mining: open source tokenization tools–an analysis. *Advanced Computational Intelligence*, 3(1):37–47. 14

Vlahovic, N. (2011). Information retrieval and information extraction in web 2.0 environment. *International Journal of Computers*, 5(1). 22

Walter, T. P. (2013). *Data Mining and Social Network Analysis of Ideation Contests: A Repeated Measures Design*. PhD thesis, University of St. Gallen. 25, 58

Wang, H. and Ohsawa, Y. (2013). Idea discovery: A scenario-based systematic approach for decision making in market innovation. *Expert Systems with Applications*, 40(2):429–438. xvii, 3, 4, 10, 31, 32, 34, 40, 64

Wang, W. (2013). *Unsupervised Information Extraction From Text–Extraction and Clustering of Relations between Entities*. PhD thesis, Paris 11. 38

Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*. 73

Ward, T. B. (1995). What's old about new ideas. *The creative cognition approach*, pp. 157–178. 26

Weihs, D. (2004). The hydrodynamics of dolphin drafting. *Journal of Biology*, 3(2):8. 42

White, R. W., Jose, J. M., and Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Information processing & management*, 42(1):166–190. 60

Wikipedia (2017). Abstract (summary) - wikipedia. `https://en.wikipedia.org/wiki/Abstract_(summary)`. (Accessed on 11/20/2017). xvii, 42, 43

Wong, S. M., Ziarko, W., and Wong, P. C. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 18–25. ACM. 15

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., Evans, A. C., et al. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73. 61

Wu, C. and Yan, M. (2017). Session-aware information embedding for e-commerce product recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2379–2382. ACM. 22, 57

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics. 38

Xiao, C., Wang, W., Lin, X., Yu, J. X., and Wang, G. (2011). Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems (TODS)*, 36(3):15. 39

Yauri, A. R., Kadir, R. A., Azman, A., and Murad, M. A. A. (2013). Ontology semantic approach to extraction of knowledge from holy quran. In *Computer Science and Information Technology (CSIT), 2013 5th International Conference on*, pp. 19–23. IEEE. 11

Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 603–610. ACM. 57

Yu, B., Xu, Z.-b., and Li, C.-h. (2008). Latent semantic analysis for text categorization using neural network. *Knowledge-Based Systems*, 21(8):900–904. 49

Zaghoul, F. A. and Al-Dhaheri, S. (2013). Arabic text classification based on features reduction using artificial neural networks. In *Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on*, pp. 485–490. IEEE. 14

Zeng, J., Duan, J., Cao, W., and Wu, C. (2012). Topics modeling based on selective zipf distribution. *Expert Systems with Applications*, 39(7):6541–6546. 15, 45

Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141. 18

Zhao, J., Huang, J. X., and Wu, S. (2012). Rewarding term location information to enhance probabilistic information retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 1137–1138. ACM. 41

Zhu, M. (2004). Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2:30. 22

BIODATA OF STUDENT

Mostafa Alksher was born on 20th January, 1975 in Zliten - Libya. He obtained his early education at his native town and finished higher secondary education from Zliten, Libya. Afterwards, he proceeded to get his degree from the Higher Institute of Qualifying Trainers in Computer science (H.D) in 1995. After completion of Higher diploma Degree, he worked as a trainer at the same institute for 5 years. Then, he had been sponsored by the Ministry of Higher Education and Scientific Research in Libya by offering a master scholarship. The first Master in computer science was obtained from Golden State University in Malaysia from 2002 to 2004.

During the year from 2004 to 2007 he has been working as a Lecturer in a few colleges in Libya. In 2007, he completed his Masters in Information Technology (M.Sc) Computer Science in University Utara Malaysia (UUM), in Malaysia and later in the year 2013 he joined Universiti Putra Malaysia (UPM) for doing his Doctor of Philosophy (PhD) programme in the field of Information Retrieval . The Author is married in 2005 and is blessed with four kids.

LIST OF PUBLICATIONS

International Refereed Journals

M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, E. M. Alshari, Effective Idea Mining Technique Based On Modeling Lexical Semantic, in: Journal of Theoretical and Applied Information Technology, vol. 96, no 16, 2018 **(Published 2018)**

International Refereed Conferences

M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, E. M. Alshari, A review of methods for mining idea from text, in: Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on, IEEE, 2016, pp. 88–93. **(Published 2016)**

M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, and E. Alshari, A framework for idea mining evaluation," in *16th International Conference on New Trends in Intelligent Software Methodology Tools, and Techniques, SoMeT 2017*, 2017, pp. 550–559. **(Published 2017)**

M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, E. M. Alshari, Feasibility of using the position as a feature for idea identification from text, in: Information Retrieval and Knowledge Management (CAMP), 2018 Fourth International Conference on, IEEE, 2018, pp. 114–119. **(Published 2018)**

UNIVERSITI PUTRA MALAYSIA

STATUS CONFIRMATION FOR THESIS / PROJECT REPORT AND COPYRIGHT

ACADEMIC SESSION : _____

TITLE OF THESIS / PROJECT REPORT :

MODELING LEXICAL SEMANTICS OF TERMS BASED ON SYNWORD IDENTIFICATION FOR IDEA MINING IN INFORMATION RETRIEVAL

NAME OF STUDENT: MOSTAFA AHMED ALKSHER

I acknowledge that the copyright and other intellectual property in the thesis/project report belonged to Universiti Putra Malaysia and I agree to allow this thesis/project report to be placed at the library under the following terms:

1. This thesis/project report is the property of Universiti Putra Malaysia.
2. The library of Universiti Putra Malaysia has the right to make copies for educational purposes only.
3. The library of Universiti Putra Malaysia is allowed to make copies of this thesis for academic exchange.

I declare that this thesis is classified as :

*Please tick (√)

☐	**CONFIDENTIAL**	(Contain confidential information under Official Secret Act 1972).
☐	**RESTRICTED**	(Contains restricted information as specified by the organization/institution where research was done).
☐	**OPEN ACCESS**	I agree that my thesis/project report to be published as hard copy or online open access.

This thesis is submitted for :

☐ **PATENT** Embargo from _____ (date) until _____ (date)

Approved by:

_____
(Signature of Student)
New IC No/ Passport No.:

Date :

_____
(Signature of Chairman of Supervisory Committee)
Name:

Date :

[Note : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization/institution with period and reasons for confidentially or restricted.]