The Impact of Pre-Processing On Automated Taxonomy Generation

By

Rida Hafeez

NUST201463929MSEECS60014F

Supervisor

Dr. Sharifullah Khan

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.
(December 2017)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS thesis written by Ms. RIDA HAFEEZ, (Registration No. NUST201463929MSEECS60014F), of SEECS-NUST has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: ________________________________

Name of Supervisor: _______________________

Date: ___________________________________

Signature (HOD): __________________________

Date: ___________________________________

Signature (Dean/Principal): __________________

Date: ___________________________________

Approval

It is certified that the contents and form of the thesis entitled “Impact of Pre-Processing on Automated Taxonomy Generation" submitted by Rida Hafeez have been found satisfactory for the requirement of the degree.

Advisor: Dr. Sharifullah Khan

Signature:_______________

Date:___________________

Committee Member 1: Dr. Anis ur Rahman

Signature:_________________________

Date:_____________________________

Committee Member 2: Dr. Muneeb ullah

Signature:_________________________

Date:_____________________________

Committee Member 3: Ms. Sana Khalique

Signature:_________________________

Date:_____________________________

Abstract

Preprocessing is an essential and primary step in automatic taxonomy generation for text documents because text data is unstructured; and more inconsistent and noisy than structured data. Different taxonomy generation systems involve different preprocessing steps during generation. However, there is no existing benchmark mark to analyze the impact of preprocessing techniques to improve the quality of taxonomy. To overcome this deficiency, a new methodology is proposed to study the comparative analysis of various preprocessing techniques and to evaluate the quality of generated taxonomy. Different combinations of preprocessing techniques have been selected and applied in generating taxonomy to amplify pertinent information for further analysis and processing. This research investigates the impact of various preprocessing techniques on the quality of the generated taxonomy and proposed a comparative analysis on the basis of various evaluation matrices. Various combinations of preprocessing techniques have been applied in taxonomy generation on two text data sets, selected from different domains i.e., ACM and MEDLINE. The experimental results revealed that selecting a suitable combination of preprocessing techniques can improve the quality of automated taxonomy. However applying all preprocessing techniques in the generation process does not guarantee high quality. The experiments were conducted on document based taxonomy however, in future, the scope of research can be extended to concept based taxonomy as well.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: RIDA HAFEEZ

Signature: _____________

Dedication

This thesis is dedicated to my parents and my husband Mr. Jawad Arshad, who never lost faith in me and supported me unconditionally throughout this effort and made it possible. I could not have done it without them.

Acknowledgments

First of all, I would like to thank ALLAH Almighty for providing me with the insight and knowledge to fulfill this task and pursue my dreams; nothing would‘ve been possible without HIS blessings and benevolence. I would like to express my deep and sincere gratitude to my supervisor, Dr. Sharifullah Khan, Department of Computing, SEECS, NUST. His wide knowledge and his logical way of thinking have been of great value for me. His understanding of the topic, persistent support and constant encouragement has inspired and motivated me as a researcher and a student.

I am also extremely thankful to my committee members Dr. Anis-ur-Rehman, Ms. Sana Khalique and Dr. Muhammad Muneebullah for taking out time from work to give their valuable suggestions and kind feedback. I would also like to thank Irfan Ali Khan, Maliha Rafi and Rabia Irfan for their constant support and help. In the end I would like to thank everybody who was important to the successful realization of my thesis.

I would specially like to thank Dr. Azeem Abbas from Arid Agriculture University for his unconditional help and support in this research.

Table of Contents

Contents

1. Introduction 2
  1.1. Motivation 2
  1.2. Automatic Taxonomy Generation 2
  1.3. Problem statement 3
  1.4. Research Objectives 4
  1.5. Proposed Solution 4
  1.6. Thesis Outline 4
  1.7. Summary 5
2. BACKGROUND 7
  2.1. Preprocessing 7
    2.1.1. Preprocessing Types 7
    2.1.2. Importance of Text Preprocessing 8
    2.1.3. General Pre-Processing Techniques 8
  2.2. Taxonomy 10
    2.2.1. Applications of Taxonomy 10
    2.2.2. Automatic Taxonomy Generation Process 10
  2.3. Summary of Chapter 16
3. Related Work 17
  3.1. Impact on Classification Techniques 18
  3.2. Impact on Clustering Techniques 19
  3.3. Comparison of This Study with the Previous Ones: 19
  3.4. Summary of Chapter 20
2. Proposed Methodology 22
  2.1. General Overview of the System: 22
  2.2. Data Collection 24
  2.3. Pre-processing Combinations 26
  2.4. Taxonomy Generation Process 28
    2.4.1. Data Pre-processing Stage 28

2.4.2. Data Modeling Stage 30

2.4.3. Hierarchy Formation Stage 31

2.4.4. Node Labeling Stage 32

2.5. Evaluation Metrics 32

2.6. Summary of chapter 33

3. Experiments and Result Evaluation: 34

3.1. Experimental Setup: 35

3.1.1. Dataset Specification 35

3.1.2. Method 36

3.1.3. Execution 36

3.2. Results & Discussion 37

3.2.1. Impact of different Dataset Size 40

3.2.2. Impact of different combinations 40

3.3. Summary 41

4. Conclusion: 43

4.1. Research Summary 43

4.2. Outcome of research 43

4.3. Future Work 43

4.3.1. Limitations 43

4.3.2. Future work 44

Following are the recommendations to further extend this research: 44

REFRENCES: 45

List of Figures

Figure 1: Stemming Process 9
Figure 2: Sample Dendrogram for Cluster Representation 14
Figure 3: System’s Flow Chart 24
Figure 4: Taxonomy Generation Steps 28
Figure 5: Division of Datasets 35
Figure 6: Experiments on data sets 36
Figure 7: Experimental results for ACM 224 documents 38
Figure 8: Experimental results for ACM 492 documents 38
Figure 9: Experimental results for MEDLINE 138 documents 39
Figure 10: Experimental results for MEDLINE 242 documents 39

List Of Tables

Table 1: COMPARISON OF STUDIES 20
Table 2: Medline Dataset Categories 25
Table 3: Acm Dataset Categories 26
Table 4: Combinations of the preprocessing techniques 27
Table 5: Experimental Results 37

CHAPTER 1:

INTRODUCTION

1. Introduction

This chapter gives an introduction to this thesis i.e. what this thesis is all about and what section it covers. It explains the reason, motivation and the purpose behind conducting the research. It also discusses briefly the problem statement, taxonomy generation steps and applications of taxonomy. It also introduces the other chapters included in the thesis.

1.1. Motivation

The digital data on the internet is growing drastically now a days, hence it is very difficult to retrieve desirable information from the pool of data. For the retrieval of most relevant data, it is important to process it and arrange it in such a structure which makes its access feasible. Taxonomy is a solution to this structural need as it is defined as a hierarchical arrangement of concepts in a dataset [1]. Various researchers elaborated different applications of taxonomy i.e. it used for data categorization and data organization [2] , standardization [3], data and knowledge management [4], data search, and it is also used in data mining techniques [5] .

Categorization of concepts/things in a hierarchy is not a new area rather it is an old convention to arrange things into categories. Before the advent of computers, the taxonomic categorization related to different domains was done by the domain experts of the specific field. It was considered with time that manual taxonomy generation is a cumbersome process, as it requires expert human beings of a particular field which are rare and it is costly to heir them. Moreover, with the progressive growth of internet, there is a huge amount of fast growing data available so it is very laborious to construct manual taxonomy of online data now a days. Due to above mentioned facts; the researchers felt the need of automatic or semi-automatic taxonomy about ten years ago and a lot of work has already been done in this field in recent years [4,6]. In the beginning, even automatic taxonomy required a lot of manual input and they were not considered as automatic rather semi-automatic. However, with the progress of computer technologies, many researchers have made the attempts to generate automatic taxonomies [6,7]. But there is a trade off in using automatic taxonomy generation techniques as they require less effort and cost but deteriorates the quality of taxonomy. Therefore, human involvement is still needed to generate an accurate taxonomy.

1.2. Automatic Taxonomy Generation

There are many taxonomy generation tools and techniques available now a days which can generate taxonomy with minimum or no human involvement, namely, “Inxight”, “Autonomy”, “Stratigy”, “Verity” [8]. These tools include different aspects of taxonomy generation i.e. domain independence [9], language independence [1] , semantics & proficiency [7], and accuracy [6]. However, different technologies adopt slight variations while generating taxonomy. Some of them find semantic relations only using dataset while some use sources like WordNet and Wikipedia [10]. Similarly, some techniques use only keywords from the dataset [11] while some use whole data set for taxonomy generation [12].

Despite of the above mentioned differences, all of the taxonomy generation techniques involve four main steps:

1. **Data Preprocessing:**
   The real world data contain noise, inconsistent and incomplete values therefore it cannot be used directly for processing. The data is cleaned and made ready for further processing in this step.

2. **Data Modeling**
   The data is still not ready for actual processing even after first step. In this step the data is converted for processing by extracting weighted terms and concepts from the data and arranged it into computational form.

3. **Hierarchy Formation**
   The hierarchical or parent child relationships are identified in this step using clustering techniques. Clustering combines the relevant terms or concepts together in the form of clusters.
4. **Nodes Labeling**
   Hierarchy formation stage generates unlabeled clusters as there are no predefined labels involved in unsupervised learning. Therefore, this step labels the clusters formed in the previous stage.

1.3. Problem statement

The huge amount of documents available on internet makes it very difficult for users to retrieve relevant documents belonging to one topic. Automatic taxonomy generation makes it easier to find documents of relevant topic. Preprocessing is one of the key components in a taxonomy generation framework. Various techniques of automated taxonomy generation have been proposed up till now which use different datasets and preprocessing techniques. However, there is no benchmark for evaluation and comparison of the applied techniques to check the accuracy and performance of taxonomy. We will focus on the impact of preprocessing on taxonomy generation in terms of various aspects such as accuracy and complexity. We will use different preprocessing techniques on a bulk of documents and then evaluate the quality and accuracy of generated taxonomy using various evaluation matrices.

1.4. Research Objectives

The objective of this research is to design a taxonomy generation and evaluation criteria to set a benchmark for preprocessing technique i.e. techniques which improve the quality of taxonomy. This research evaluates the taxonomies generated as a result of using various preprocessing techniques hence gives a comparative analysis of using these technique. It also highlights the advantages and disadvantages of using different preprocessing techniques.

1.5. Proposed Solution

To study the impact of various pre-processing techniques, we have analyzed the preprocessing techniques used in the previous studies and created five combinations of different techniques.

We have applied those techniques to two different datasets i.e. ACM and MEDLINE individually and generated taxonomy after applying each combination.

Furthermore, we have evaluated the taxonomy using evaluation matrices with reference to gold standard taxonomy, and analyzed the results based on the evaluation conducted.

1.6. Thesis Outline

- **Chapter 2: Background**
  In this chapter we have explained the background of preprocessing and taxonomy in detail. The preprocessing techniques and their applications are discussed. Moreover, taxonomy generation process steps are also discussed in detail to provide theoretical knowledge to understand the thesis.

- **Chapter 3: Related Work**
  In this chapter we have discussed related work and identified the loop holes in the previous studies which make the basis of establishing this thesis.

- **Chapter 4: Proposed Methodology**
  This chapter explains the system made to address the research problem. It also discusses the data sets and step by step process to achieve the results.

- **Chapter 5: Experiments and Result evaluation**
  This chapter discusses and analyzes the results which are generated using the proposed methodology. It also discusses the evaluation matrices on the basis of which the results are compared and analyzed.

- **Chapter 6: Conclusion**
  This chapter provides the summary of the research conducted. It also explains deficiencies in this work and the future work which can be done to improve this work.

1.7. Summary

This chapter gave the brief overview of what this research is all about. It also describes the motivation behind conducting this research, applications, problem statement and proposed solution. A brief outline of the thesis is also discussed in this chapter.

CHAPTER 2

BACKGROUND

2. BACKGROUND

This chapter gives the background of the relevant terms and concepts used in the thesis. These concepts are explained in detail with reference to the back ground in the following chapter.

2.1. Preprocessing

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text. NLP researchers aim to collect knowledge on how human beings understand and use language so that fitting tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the preferred tasks [14]. The basics of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems and so on [15,16].

2.1.1. Preprocessing Types

Preprocessing stage in taxonomy generation process performs cleansing of raw data. The methods used in this stage mostly depend upon the nature of data. Based on methods commonly adopted to perform data pre-processing activities, we have divided this category further into two subcategories: NLP based and non-NLP based approaches.

(a) NLP based:

Natural Language Processing (NLP) techniques are important as they help machine to understand language written and spoken by humans. They are mostly applicable in a data set dealing with long and descriptive text data. In taxonomy generation process, before extracting concepts and their relationships from a data set, basic NLP techniques like tokenization, stemming, part of speech tagging and parsing are used. The works [6,9] have applied basic NLP techniques to identify noun phrases from the given text data. Noun phrase refers to those phrases whose head or principal phrase is a noun. Noun usually contains the important concepts related to a particular domain such as, information of people, place, organization, location. Authors in [7] have also applied these basic NLP techniques to identify noun phrases. However, they have observed that a noun phrase can have more than one sense and should be processed further for finding their true context.

(b) Non-NLP based:

Non-NLP techniques are useful in modeling data which involves less descriptive text data i.e. tags or data type other than text such as image, audio etc. Some past works have organized a linked data set in the form of hierarchy by utilizing the linked data set structure. Linked data set structure is based on Resource Description Framework (RDF) 8 tuples providing knowledge about instance types and object types. In data pre-processing stage, they have filtered the linked data set utilizing class types to which they belong and object types from which their attributes are obtained. [46]

2.1.2. Importance of Text Preprocessing

Pre-processing impacts the text mining results in a lot of ways and can improve the output as explained below in detail.

1. Pre-processing is used to decrease file size by removing noise and unnecessary characters from text documents
    i. Size of text document can be reduced by 20-30% by removing stop words.
    ii. Stemming can reduce indexing or file size by 40- 50%

2. Pre-processing can also improve the efficiency of the information retrieval (IR) system.
    i. As stop words are not meaningful so they are not useful for searching and improving the efficiency of retrieval system
    ii. However, stemming is used for identifying same words in a document. [16]

2.1.3. General Pre-Processing Techniques

Pre-processing is a pre-step of data analysis process for any document corpus. Following are some of the common pre-processing techniques [17]:

- **Tokenization**, it is the process of splitting text on the basis of a given token. A token can be any character i.e. space, comma, colon etc.
- **Lowercase conversion** this technique converts all the letters to lowercase.
- **Special character removal:** it excludes all the numbers and special characters i.e.+, -, !, ?, ., ,, ;, :, =, &, #, %, $, [,], /, <, >, n, \, " etc.
- **Stop Word Removal:** it removes all the meaningless words that are not important for the classification of a text documents e.g. prepositions articles etc. (e.g., a, the, at, etc.).
- **Stemming:** it rounds off the word to its stem or root. It reduces the forms of verbs and plurals to the original singular word as shown below.

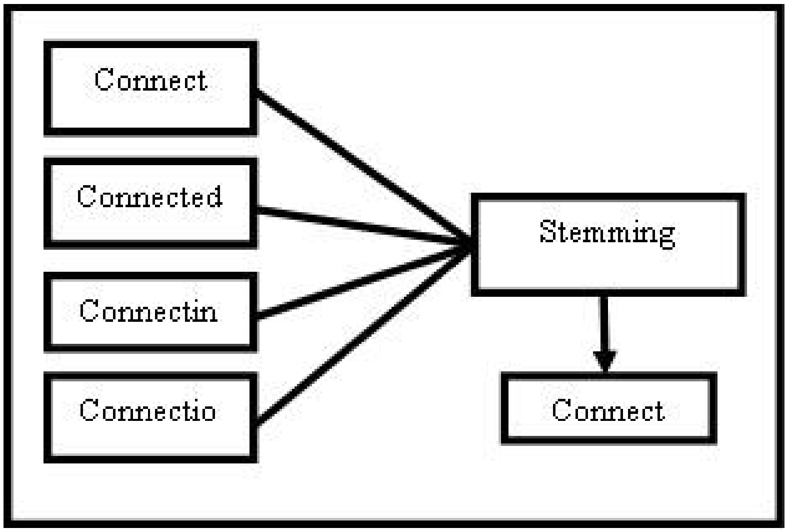


Figure 1: Stemming Process

- **Pruning** it removes the highly occurring and the least occurring terms from the document. These terms are discarded on the basis of their TF/IDF (term frequency/ document frequency) score as highly frequent or least frequent terms do not help in identifying the topic of a document.
- **Treating synonyms:** this technique replaces the different words with the same meaning. If two words have the same semantic meaning then one will be replaced by the other. This improves the frequency of that word in a document.
- **Document representation**: After applying all the pre-processing techniques, the document is represented as a vector with weight terms.

There are different techniques to represent term weightage for each document. There are two most commonly used measures for that purpose i.e. TF (Term Frequency) and TF/IDF (Term Frequency Inverse Term Frequency). Following definitions for these measures are provided in [18].

The above listed pre-processing techniques are very important as they help in reducing the noise and meaningless characters from the documents, eventually reduce the "curse of dimensionality". Stop word removal, pruning and stemming increase the quality of text classification/clustering and decrease the number of dimensions by removing useless words from the documents. The techniques reduce the huge number of terms in a document which help in the Document representation stage and deal with identifying the importance of each term in a document. [17,19]

2.2. Taxonomy

The word “Taxonomy” is derived from two Greek words “taxis” means “arrangement” or order and “nomos” means “law or science [20]. So taxonomy is the science of ordering a data collection. By definition, taxonomy is the hierarchical (i.e., parent/child) organization of concepts present in a data collection. There are many types of relationships that can exist among the terms/concepts inside a data collection, such as hierarchical and associative. These relationships are represented and organized by a knowledge organization structure, so that a data collection can be utilized for an effective understanding and reasoning task.

Information seeking in World Wide Web through powerful search engines is time taking and most of the time confusing, if an information seeker is not very clear about what he wants to search [21]. This is the drawback of Web search engines that they provide results on the basis of matching and ranking. Taxonomy is structured, hierarchical and an effective way of browsing over the information which an information seeker is looking for. The hierarchical structure gives the ability to an information seeker to understand the relationship between various concepts and terms he is looking for, in an efficient and effective manner. This way an information seeker can save a lot of valuable time, which was otherwise wasted due to vagueness of search results.

2.2.1. Applications of Taxonomy

The authors in [22,23] have described the use of taxonomy to facilitate browsing and searching. With the help of taxonomy, a user can look over important concepts and understand them with the help of hierarchical relationships that they possess with other concepts.

Organizations can adopt taxonomy for knowledge management purposes [24]. Taxonomy can help an organization in categorizing their goals, objectives, policies and strategies, so that various objectives can be met effectively. The structure of taxonomy can enable the people in an organization to understand and share important contents in a standardize way, that can result in continuous improvement of the organization.

2.2.2. Automatic Taxonomy Generation Process

The existing automated taxonomy generation techniques involve three major steps. These steps include:

1. **Generation**: This stage deals with the steps which are involved in the generation of and automated taxonomy.
2. **Evaluation**: This stage involves the procedure and matrices to evaluate the quality of and automated taxonomy.

3. **Representation**: This stage involves the steps which are required to represent an automated taxonomy.

2.2.2.1. GENERATION

It is observed that all the existing techniques for automatic/ semi-automatic taxonomy generation may have different aspects, goals and objectives to achieve from taxonomy, but all of these tools follow same basic steps to generate the taxonomy. These basic steps are applied with minor changes to generate the taxonomy automatically in different techniques. Generally, the generation process of taxonomy adopted by existing automatic taxonomy generation techniques involves four main steps:

i. Data Pre-processing

Data pre-processing stage involves those activities that are related to raw data collection and its initial cleansing, so that it becomes ready for processing. The kind of pre-processing that needs to be applied on data, depends upon data type and type of application for which data is getting ready for processing. In data mining applications, for numerical data, the data pre-processing activities comprises of finding missing attribute values, removing outliers and discretization [25]. In text mining applications, textual data pre-processing comprises different natural language processing (NLP) techniques such as: tokenization, lemmatization and stemming [26]. Similarly for taxonomy generation, it also depends upon the type of data and end objective of taxonomy that what kind of pre-processing needs to be applied on data. Some of the basic NLP techniques [14] that are applied in taxonomy generation applications are tokenization, stemming, part of speech tagging and parsing. The data pre-processing stage can be defined as follows: Data pre-processing stage involves all those activities that are performed on a raw data, so that it becomes clean and ready for undergoing further processing.

ii. Data Modeling

Mostly the data, even after applying the data pre-processing techniques, is not ready for actual processing to be performed on it. Data modeling stage involves all those activities that are performed on the preprocessed data, in order to find a computable representation of the data. The computable form should reflect the context and true sense of important concepts present in the data. The computable form should also be free from complicated and complex details present in the data and should be precise, so that actual processing can be applied efficiently and accurately. These activities initially extract relevant terms/concepts out of the data. These relevant terms/concepts are a set of features that reflect properties or characteristics of a data set [27]. They can also be referred as dimensions of a data set. Data modeling activities give a model that expresses the data in a form suitable for computation. Vector Space Modeling (VSM) is one of the most widely used data modeling techniques.

In natural language, a word can have different meanings in different domains. In order to perform computation on data accurately, it is important to determine the true context and semantics of concepts present in a data set. To determine semantics of concepts present in a data set, external knowledge sources like WordNet, Wikipedia5, Freebase, DBbase are usually involved in this stage. During taxonomy generation process, in order to determine hierarchical relationships that exist among the important concepts, it is required to clearly determine the true context in which a concept is used in a data set. Advanced natural language processing techniques [14] like word sense disambiguation along with external knowledge sources can be applied here. Furthermore, domain specific terminologies can be determined by applying various domain specific statistical measures. Dimensionality reduction techniques that can reduce dimensions of data can also be applied in this stage. Dimensionality reduction involves mathematical mapping or transformation techniques to select a minimum set of features and hence reduce the computational complexity.

iii. Hierarchy Formation

The computable form of data is now ready to undergo actual processing. In case of taxonomy generation, this actual processing refers to two sub-steps. The first step is the determination of parent-child relationships that exist among concepts identified in the data modeling stage. We can call this sub-step as a hierarchical relationship identification step. Arrangement of these relationships in the form of a hierarchical structure is the next sub-step after relationship identification.

For the arrangement of concepts and their hierarchical relationships in the form of a hierarchy, different taxonomy generation techniques have adopted different approaches. Most of the reviewed techniques have adopted hierarchical clustering techniques in this stage. The hierarchical clustering techniques combine the above mentioned sub-stages i.e., hierarchical relationship identification and hierarchy generation. Clustering is the division of similar objects into groups. Those objects that are similar to each other in some ways are kept together to form a cluster and dissimilar objects are kept away into other clusters [28]. This is the unsupervised form of learning where no prior information (i.e., training date) about cluster labels or classes is provided.

The clusters are formed by measuring similarity and dissimilarity distances between objects [29]. Hierarchical clustering organizes clusters in the form of hierarchy, unlike the flat partitioning based approach where, clusters exist standalone without any explicit structure relating them to each other [30]. Apart from clustering approach, approaches based on graph theory and rules based on parent-child relationships among contents in a data set can also be utilized in the hierarchy formation stage to produce the taxonomy. The most commonly used technique for clustering is Hierarchical Agglomerative Clustering (HAC) which is described below:

Hierarchical Agglomerative Clustering (HAC)

HAC (hierarchical agglomerative clustering) approach is the most famous approach for hierarchical clustering in which we takes each data point as a cluster and then combine them in each propagative step. There are following different flavors of doing it. [47]

- **Single Link:** According to this method we take the distance between two clusters as the distance between two nearest data points in the clusters. This is the minimum distance between two the clusters. According to this definition, on each progressive step we combine those two clusters which have the minimum single link distance between them.

- **Complete Link:** According to this method we take the distance between two clusters as the distance between two farthest data points in the clusters. This is the maximum distance between two the clusters. According to this definition, on each progressive step we combine those two clusters which have the maximum complete link distance between them.

- **Average Link:** According to this method, we take the distance as the average of all the distances of data points of one cluster to all the data point of the other cluster. After calculating the distance we combine those two clusters which have the minimum average link distance between them.

- **Centroid Method:** According to this method, we take distance as the mean of the vectors of the clusters. After calculating the distance we combine those two clusters which have the minimum centroid distance between them.

- **Dendrogram:** The agglomerative hierarchical clustering algorithms build a cluster hierarchy that is commonly displayed as a tree diagram called a “dendrogram”. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated.

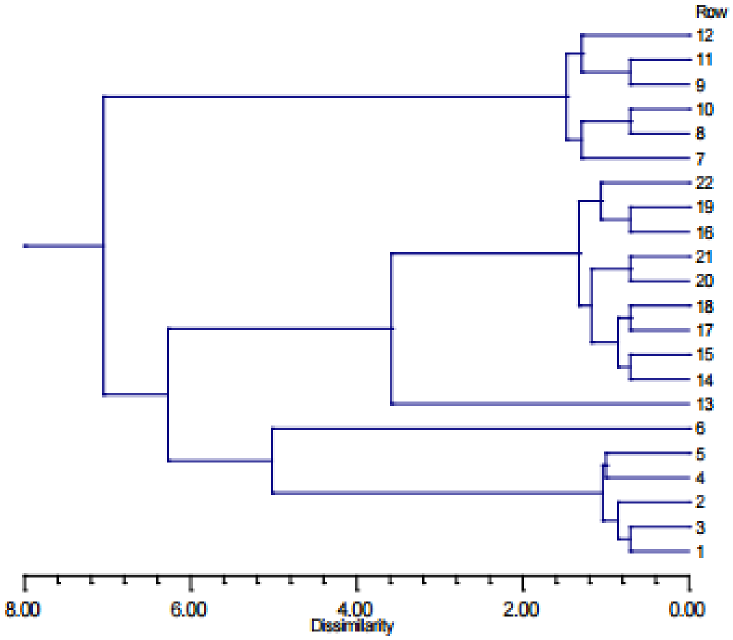


Figure 2: Sample Dendrogram for Cluster Representation

The horizontal axis of the “Dendrogram” represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The “Dendrogram” is fairly simple to interpret. Remember that our main interest is in similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters.

iv. Nodes Labeling

The hierarchical structure formed as a result of the hierarchy formation stage, is unlabeled and is not a taxonomy in its real sense at this stage. Some more processing is needed in order to convert the hierarchical structure in the form of labeled taxonomy. Nodes labeling is more appropriate to apply in clustering based techniques. Since clustering is unsupervised and no labels are assigned before actual clustering has been done. So nodes labeling is most appropriate to apply in those taxonomy generation techniques that use clustering based approach in the hierarchy formation stage. In clustering based approaches usually centroids of clusters are involved in finding labels for taxonomic nodes. Labeling techniques are mostly combined with rules and heuristics in order to find appropriate labels for taxonomy. External sources like WordNet can also be utilized for finding appropriate labels for taxonomic nodes. According to [30] feature selection techniques e.g., (i) Mutual Information (ii) χ^2 test, can be utilized in the

nodes labeling stage. These techniques can differentiate a label suitable for one cluster from others. Attempts were made to explore labels for hierarchical structure automatically in the works [31]. They are statistical techniques that rely on frequently occurring top k terms in a cluster to identify its label. The nodes labeling stage can be defined as follows:
Nodes labeling stage involves all those activities that are performed to assign appropriate labels to unlabeled nodes of taxonomy formed in the hierarchy formation stage.

2.2.2.2. Evaluation

This stage involves the process of verification of quality of taxonomy once it is generated. It is very vital to check the accuracy of automated taxonomy whether it is able to achieve the improved quality or not. To check the efficiency of generated taxonomy, various taxonomy generation techniques verify the taxonomy using different evaluation measures. These measures can be fully automated or semi-automated i.e. may need human involvement. Evaluation techniques are normally accompanied with a comparison with gold standard taxonomy. The evaluation techniques can be divided into various categories as discussed below:

- This technique involves human judges and experts in the process of evaluation. Domain experts are hired which can interpret the semantics of the concepts involved in the generated taxonomy. These experts can judge whether the taxonomy is made accurately according to the human interpretation and knowledge of the domain or not [32].
- In this technique, gold standard taxonomy is involved as a reference to evaluate the generated taxonomy. There is some data set for which gold taxonomy is available for example MeSH is a gold taxonomy for Medline document data set. If gold taxonomy is not available for comparison, manual gold standard taxonomy can be generated [7].
- This evaluation criterion is called qualitative measure as it evaluates the taxonomy on the basis of non-numeric measures. It includes features like hierarchy of taxonomy, depth of taxonomy and output results [9].
- This measure is called quantitative measure as it involves some numeric measures to compare the generated taxonomy with some gold taxonomy. These evaluation techniques include measure like precision, recall and F-measure [33]. More over some of the measures like average, mean and probability distribution are also used to evaluate the generated taxonomy [6].

2.2.2.3. Representation

This stage deals with the arrangement or representation of data set in the form of hierarchy to construct taxonomy. Taxonomy is a hierarchy of concepts and terms and this stage models the concepts in hierarchical structure. Different concepts have different meaning in context of different domain same word has multiple meanings associated to multiple domain. Representation stage deals with the relationship between dataset and taxonomy i.e. the ways in which dataset can be modeled. Following are the subcategories:

- This category is call single view model, in which only a single taxonomy is generated out of the given dataset despite of the end user's requirements. In this category only a single taxonomy is generated no matter what the data set and end requirements are.
- In this category, multiple taxonomies are generated out of a single data set, that is why it is called multiple view model. This technique can generate multiple taxonomies from a single data set. Therefore, it can cater the needs of different end users with different taxonomies.
- Another form of taxonomy is called static taxonomy. This type of taxonomy does not change or evolve with the change of data. Many existing taxonomy generation techniques generate static taxonomy [7,9,6]. If some change appears in the data then it is required to run the whole process of taxonomy generation to update the taxonomy which is not suitable and very costly in case of continuously changing data.
- There is another type of taxonomy representation which is called dynamic or evolving taxonomy in which a taxonomy changes with any change in data. This technique is suitable in case of rapidly changing data in today's digital world. Work is being done on evolving taxonomies to meet the modern requirements [34].

2.3. Summary of Chapter

This chapter explains the back ground of concepts used in this research. This chapter explained the concept preprocessing, taxonomy and clustering in detail. It discussed preprocessing, its importance, it types and its common techniques. It also explains the taxonomy, its applications, generation and representation.

CHAPTER 3

RELATED WORK

3. Related Work

Data preprocessing is not a new field of study. Many researchers have been working on it for many years and they studied the impact of different preprocessing technique on different data sets under various system setups. There are two dimensions involved in a machine learning algorithm i.e. classification and clustering. In classification, we have predefined labels/classes and we categorize the data according to those labels therefore we call it supervised learning. On the other hand, clustering is unsupervised learning in which there are no prior labels and we assign labels to data after making clustering.

We have discussed the various research papers discussing the impact of preprocessing on both classification and clustering techniques. Clustering is used in taxonomy and it is discussed specified to this research.

3.1. Impact on Classification Techniques

- The researchers in [19] have studied the impact of pre-processing techniques, such as, tokenization, stemming, stop word removal and lowercase conversion on news-papers and email datasets in Turkish and English languages. They have concluded that different preprocessing combinations have different impacts depending on the domain and enabling or disabling all the techniques may degrade the results. However, lowercase conversion improved the results regardless of the domain and language.
- In [35,17] authors applied tokenization, stemming, stop word removal, pruning and synonym check on the Medline dataset. The results suggest that by applying stemming, stop word removal and pruning, the results of classification, i.e., accuracy can be improved.
- In [36] the analysis of using stemming, stop word removal and different schemes of tokenization were done to filter spam emails. The analysis shows that stemming and stop word removal decreases the performance of support vector machine (SVM). However, it is studied that there are a few stop words which are rarely used in spam emails and they should not be removed to get more accurate results.

Furthermore, selecting different tokenization schemes may have different effects on the performance of spam email filtering.

- There is another study discussed in [37] on the influence of tokenization, stemming and stop word removal on different data sets, i.e., springer News groups and Reuters. The study concludes that use of stemming and stop word removal techniques has not noticeable impact on classification results.
- The authors in [38] have discussed stemming with the purposes to reduce different words like nouns, adverbs, verbs, and adjectives. They have discussed different methods of stemming and they have drawn a picture of their comparisons in terms of advantages, usage as well as limitations. It was concluded that different stemming algorithms are good in different situations and their performance cannot be generalized.
- The authors in [39] have investigated that the effectiveness of classification of twitter sentiment can be improved by removing stop words. The authors applied six different stop word identification methods to twitter data from six different datasets and show that how removing using precompiled lists of stop words negatively impacts the performance of twitter

sentiment analysis. While on contrary, the dynamic generation of stop word list increases the performance of sentiment analysis. The impact of preprocessing is specifically analyzed on taxonomy by following authors.

3.2. Impact on Clustering Techniques

- In [6] the authors have discussed the impact of noun phrase extraction on the taxonomy of Medline documents. They have measured the content and structural quality of taxonomy using various experiments. Content quality was measured using content quality measure-precision (CQM-P) and content quality measure-recall (CQM-R), and structural quality was measured using structural quality measure-precision (SQM-P) and structural quality measure-recall (SQM-R). It is concluded that applying noun phrase extraction gives better results for CQM-P, but poor results for CQM-R because applying noun phrase extraction may have removed useful labels which were actually present in the gold taxonomy. This paper did not conclude any results for the impact of using noun phrase extraction on structural quality of the taxonomy.
- Impact of preprocessing on clustering of Slovak dataset has been discussed in [40]. The data set consists of 30,000 newspaper articles and 10,000 blogs in Slovak language. Firstly, the data was converted into English language using Google translator and then preprocessing techniques, i.e., tokenization, stop word removal and stemming, were applied. The results for clustering were evaluated using F-measure and it was concluded that stemming improved results in the dataset.
- Automated taxonomy of 70,000 English news documents of one year was designed and discussed in [9]. For document preprocessing, they have applied stop word removal, stemming and extracted frequently used noun phrases and nouns. They have observed that stop word removal are helpful for further document processing; noun phrases give better results as compared to single word nouns; and stemming improves the recall and provide more meaning full terms for taxonomy.

3.3. Comparison of This Study with the Previous Ones:

There are a lot of researchers who have studied the impact of preprocessing on various data sets as discussed in literature review. Following table shows a brief comparison of this study with the previous ones.

Table 1: COMPARISON OF STUDIES

	TK	LC	SP	SW	St	PR	NE	NP	Multiple collection
[6]								✓	
[7]							✓	✓	
[40]	✓			✓	✓				
[9]				✓	✓		✓	✓	
[19]	✓	✓		✓	✓				
[17]	✓	✓	✓	✓	✓	✓			
[36]	✓			✓	✓				
[37]	✓			✓	✓				✓
[38]					✓				
[39]				✓					
Proposed work	✓	✓	✓	✓	✓	✓	✓	✓	✓

This research discusses the impact of different well known preprocessing techniques such as tokenization, stop-word removal, lowercase conversion, stemming, special-character removal, pruning, noun extraction and noun phrase extraction. This study is different as compared to the previous studies mentioned as we have applied different combinations of the preprocessing techniques on two different data collections, such as MEDLINE and ACM. In this way, contribution of the regarding preprocessing tasks to the quality of taxonomy at various feature dimensions, possible interactions among these tasks, and also the dependency of these tasks to the domain studied on are extensively assessed. In order to clarify the differences of this work from the previous ones, a comparison is presented in the table 2. Tokenization, lowercase conversion, stop-word removal, stemming, special character removal, pruning, noun extraction and noun phrase extraction are abbreviated as TK, SR, LC, ST, SC, PR, NE and NP respectively. The experimental settings include multiple collections, clustering, and feature selection.

3.4. Summary of Chapter

In this chapter we have highlighted different studies related to impact of preprocessing techniques on both classification and clustering. A comparison of our research with the previous studies is also discussed to identify the loop holes in the precious studies.

CHAPTER 4:

PROPOSED METHODOLOGY

2. Proposed Methodology

In this chapter, a flow chart of general methodology and system design is proposed for checking the impact of preprocessing on the generated taxonomy using various evaluation matrices. Moreover, the collection of dataset, preprocessing combination, taxonomy generation and evaluation matrices are also discussed one by one in detail in the following sections.

2.1. General Overview of the System:

For checking the impact of pre-processing on the quality of generated taxonomy, we have collected the dataset comprising of two domains i.e. ACM and Medline documents. For conducting the relevant experiments on the available data set, a system design is proposed which combines different pre-processing techniques to make various five combinations. These combinations will then be applied on the text documents data to pre-process them.

Moreover, this cleaned and pre-processed data will then be passed to taxonomy generation algorithm to generate different taxonomies using data modeling, hierarchy formation and node labeling steps. The number of generated taxonomies will be equal to the number of combinations used so that we can evaluate the taxonomy results generated for each pre-processing technique. After generating the taxonomies, results will be evaluated, using precision, recall and f1-measure, to check the quality of generated taxonomy like which pre-processing technique is responsible to improve or which is responsible to decrease the quality of results. There are different techniques and tools used for carrying out the whole process which will be discussed in details in later sections as follows:

- **Data Collection**
- **Pre-Processing Combinations**
- **Taxonomy Generation**
- **Evaluation Matrices**

Following is the flow chart describing the broader picture of the system.

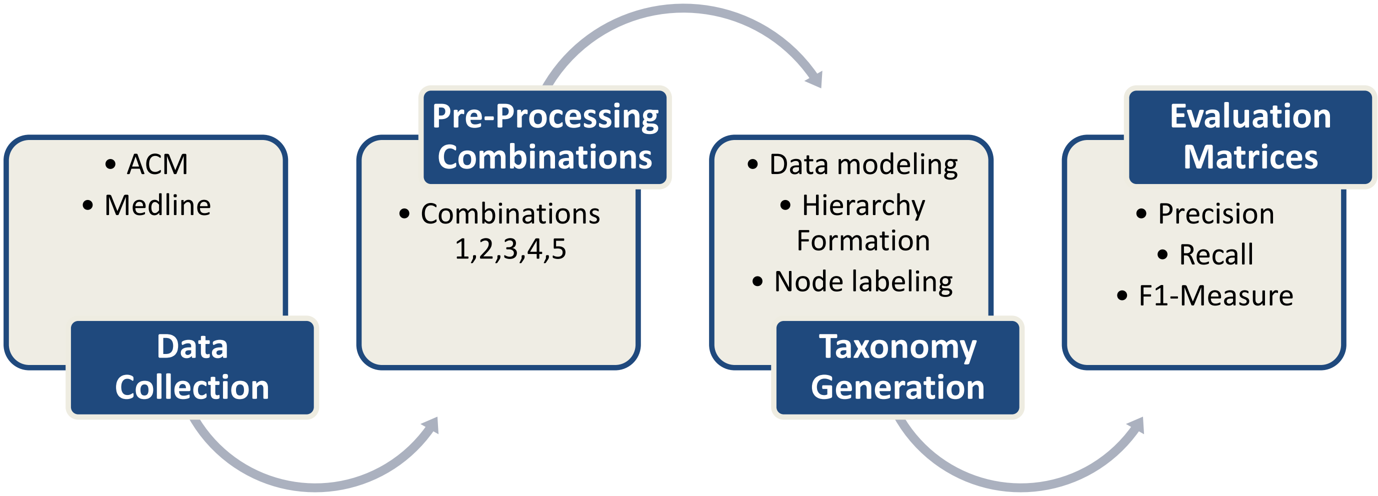


Figure 3: System's Flow Chart

2.2. Data Collection

Data sets were selected from two domains, where gold taxonomies are available, i.e., Medical - Medical Subject Headings (MeSH) [41] and Computer Science - ACM Computing Classification System (ACM CCS) [42]. Both the gold taxonomies are available in RDF (SKOS) format. They are very comprehensive. Medical documents: 242 were collected from MEDLINE digital library that were indexed on MeSH taxonomy's node, i.e., *neoplasms* and its sub nodes. Similarly computer Science documents: 492 were collected from ACM digital library that were indexed on nodes *computer systems organization*, *hardware*, *software and its engineering*, *networks* and their sub-nodes of ACM CCS. Table 1 shows the sub division of MEDLINE dataset and table 2 shows the sub division of ACM dataset.

Table 2: Medline Dataset Categories

Domain	Category	No of Subcategory	No of Documents
Medical	Cysts	17	43
	Hamartoma	03	13
	Neoplasms by Histologic Type	0	02
	Neoplasms by Site	0	03
	Neoplasms, Experimental	05	14
	Neoplasms, Hormone-Dependent	0	04
	Neoplasms, Multiple Primary	02	10
	Neoplasms, Post-Traumatic	0	04
	Neoplasms, Radiation-Induced	01	07
	Neoplasms, Second Primary	0	05
	Neoplastic Processes	04	16
	Neoplastic Syndromes, Hereditary	07	34
	Paraneoplastic Syndromes	03	20
	Precancerous Conditions	08	37

	Pregnancy Complications, Neoplastic	01	30
Total			242

Table 3: Acm Dataset Categories

Domain	Category	No of Subcategory	No of Documents
Computing	Computer Systems Organization	3	128
	Hardware	7	191
	Software and its Engineering	2	119
	Networks	5	54
Total:			492

2.3. Pre-processing Combinations

Data preprocessing techniques that have been applied for taxonomy generation in this work are briefly explained in this section.

- Tokenization (TK) is used to split the text words in a sentence on the basis of a delimiter, i.e., space, comma or any such character.
- Lower case conversion (LC) method converts all alphabets to lower case so that “WORLD” and “world” are considered as the same terms.
- Special character (SC) removal removes all non-alphabetic characters from a text dataset. These characters are considered as noise data.

- Stop word (SW) removal is another very important preprocessing technique, which removes frequently used meaningless words of a language, i.e., prepositions, articles, pronouns, e.g., the, in, a, an, with. The list of stop words can be customized to any group of words that can be considered as stop words for a particular purpose.
- Stemming (ST) plays a very important role in preprocessing of documents. It is a process of reducing word into their stem. For example, agreed, agreeing and agreement all are stemmed to agree. It is a technique which is used widely in text mining/analysis.
- Pruning is a technique, which discards the terms that appear too rare or too frequent in a dataset.
- Noun extraction (NE) step extracts single nouns or uni- grams from a text using parts of speech (PoS) tagging.
- Similarly Noun phrase extraction (NP) extracts nouns which consist of more than one word. Noun phrases are more meaningful as compared to uni-grams.

In order to study the impact of various preprocessing techniques on the quality of the generated taxonomy, five different combinations of the above mentioned preprocessing techniques that have been identified in the literature review chapter, were considered in this research for conducting experiments. The combinations are shown in following table. Preprocessing techniques: lowercase conversion, Tokenization, special character removal, stop-word removal, stemming, noun extraction and noun phrase extraction are abbreviated as LC, TK, SC, SW, ST, NE and NP respectively. Value 1 in the table represents a preprocessing techniques is present in a combination and value 0 shows a technique is absent in a combination.

Table 4: Combinations of the preprocessing techniques

No.	**LC**	**TK**	**SC**	**SW**	**ST**	**NE**	**NP**	**Ref:**
Combination 1	1	1	1	1	0	1	0	[43]
Combination 2	1	1	1	1	1	1	0	[40]
Combination 3	1	0	0	1	0	0	1	[7]
Combination 4	1	0	1	1	0	0	1	[6]
Combination 5	1	0	1	1	1	0	1	[9]

After applying one or multiple preprocessing techniques, each concept or term of a document is represented in vector space model. The occurrence of terms in a document is represented with tf-idf (i.e., term frequency - inverse term frequency) and only tf (i.e., term frequency). We have used tf-idf in this work

2.4. Taxonomy Generation Process

There are many well-known techniques and tools for generating taxonomy automatically or semi-automatically and we have observed that the basic steps for generating taxonomy remain the same, though the primary goals and usage of their generated taxonomy differs. These steps are applied with minor changes to generate the taxonomy automatically in different techniques. Basically there are four main steps in the existing taxonomy generation techniques as shown in the figure below. In the following section, we have discussed these steps one by one specific to our research.

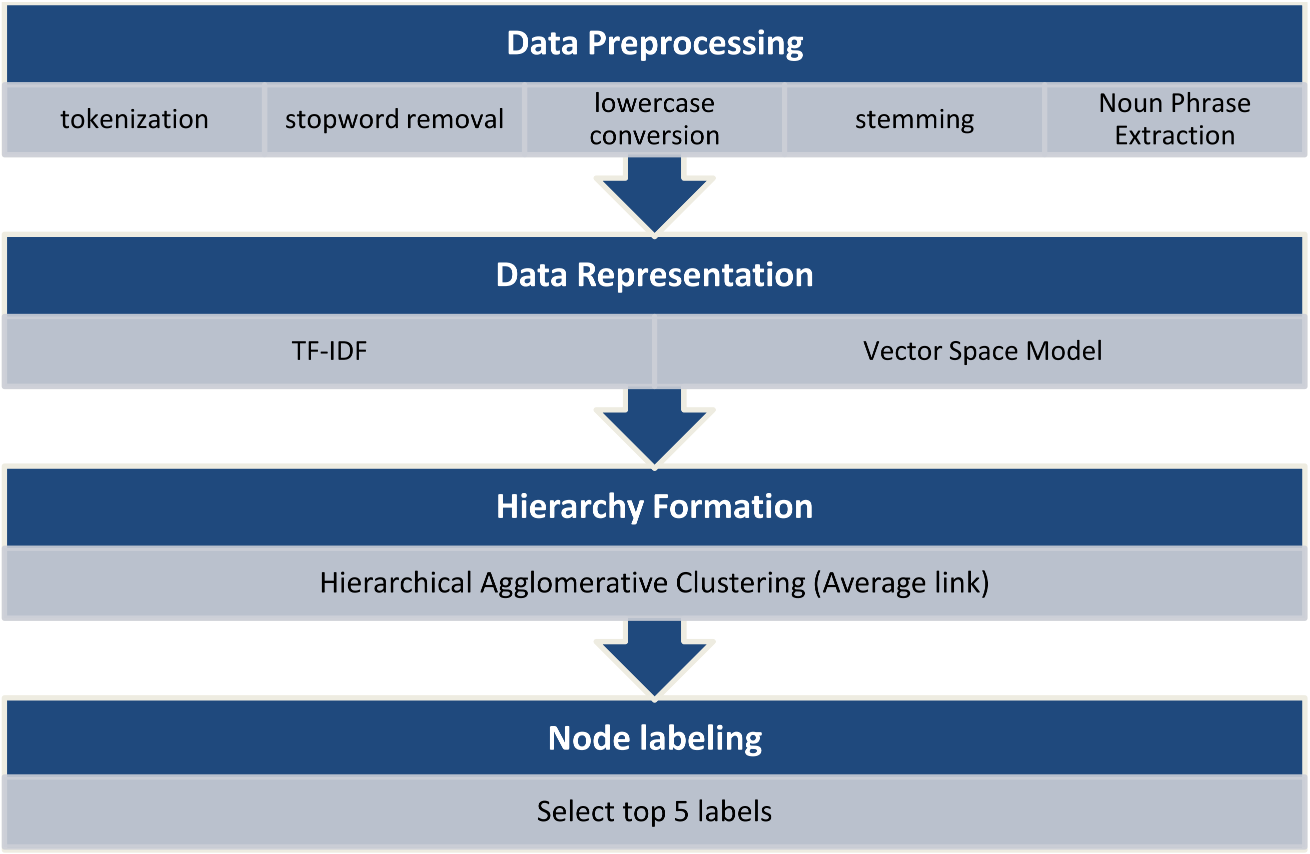


Figure 4: Taxonomy Generation Steps

2.4.1. Data Pre-processing Stage

We will focus on the first step i.e. "Data Pre-Processing" involved in the taxonomy generation process. This step is very important as cost of processing increases exponentially with the size of input data [13]. Therefore, it is not possible to generate an accurate taxonomy by inputting noisy or irrelevant data. Data pre-processing is the stage which involves all those set of activities that are performed on a raw data, so that it gets ready for further processing. The kind of preprocessing steps depends on the nature of data. For text mining application the pre-processing techniques include tokenization, lowercase conversion, special character removal, stop word removal, lemmatization, pruning, parts of speech tagging i.e. Noun and Noun phrase extraction. In order to get relevant information out of dataset, we have applied these seven commonly used pre-processing techniques on two text data sets I.e. ACM and Medline.

We have implemented all the above mentioned preprocessing techniques in the form of combinations as listed in sections 4.3 above. There are three tools used i.e. built in java language functions, Stanford Natural Language Processing tool[1] in java and Natural Language ToolKit (NLTK) library in python language[2].

- ***Tokenization***

Tokenization (TK) is used to split the text words from a sentence on the basis of a delimiter i.e. space, comma or any such character. We have implemented this technique using built in string function in java programming language. The delimiter in our application was “space”, thus, text sentences were converted into tokens of words on the basis of space characters.

- ***Special Character Removal***

Special character removal technique removes all the non-alphabetic characters i.e numeric and special characters from the dataset. These characters are considered as meaning less and noisy in text dataset domain. We have accomplished the removal of these characters using regular expressions in java.

- ***Stop Word Removal***

Stop word removal is another very important preprocessing technique which removes frequently used meaningless words of a language i.e. prepositions, conjunctions, adverbs, articles, pronouns e.g. the, in, a, an, with etc. we can customize the list of stop words according to our need as any group of words can be considered as stop words for a particular purpose.

We have used java language function for the removal of those words present in the standard list of stop words in English language. It can also be done using Stanford NLP by applying parts of speech tagging and then removing those words which are tagged as prepositions, conjunctions, adverbs, articles, pronouns.

[1] https://stanfordnlp.github.io/CoreNLP

[2] http://www.nltk.org

- ***Stemming***

Stemming plays a very important role in preprocessing of text documents. It is the process of reducing words into their stem. For example agreed, agreeing and agreement all are stem to agree. It is a technique which is used widely in text mining/analysis as it reduces the number of discrete words by classifying them as one. We have implemented this technique using Stanford NLP “lemmatize” function.

- ***Noun Extraction***

Noun extraction step extracts single nouns or uni- grams are single nouns in the text data. We have extracted nouns from text using parts of speech (PoS) tagging in NLTK, then separate those words which are tagged as noun.

- ***Noun Phrase Extraction***

Noun phrase extraction extracts nouns which consist of more than one word i.e. bi-grams and tri-grams. Noun phrases are more meaningful as compared to uni-grams. We have extracted noun phrases using NLTK in Python, and then used those words in java application.

- ***Lowercase Conversion***

Lower case conversion (LC) method converts all the alphabets to lower case so that “WORLD” and “world” are considered as same terms. We have applied lowercase conversion using string functions in java.

2.4.2. Data Modeling Stage

Data modeling stage involves all those activities that are performed on the preprocessed data like extracting terms/concepts, in order to find a computable representation of the data. This stage models data in a computational form and makes it ready for actual processing. Vector Space Modeling (VSM) [45] is one of the most widely used data modeling techniques. According to this model each document is represented in the form of a vector. The relation between documents and terms is represented in the form of a matrix in which most vectors are sparse because majority entries are zero. The weight of terms can be defined in binary form i.e. 1 means term is present in the document 0 indicated absence of a term. The weights can be non-binary such that based on frequency of a term in a document. We have used a weighing measure call TF/IDF score which is a statistical value to measure the importance of a term in a document. The higher the frequency of the term in a document the more important it is, but it is less important if its occurrence is more frequent in other documents of the corpus. TF/IDF can be calculated as follows:

- **Term Frequency: TF**

It is a measure to determine the frequency of a term in a document. To normalize this measure the frequency is divided by the number of terms in a document as it is possible that a term occur more frequently in a long document as compared to a short document. Suppose t is term and d is document then:

$$\mathbf{TF(t,d)} = \frac{\textbf{Frequency of a term in document}}{\boldsymbol{Total\ no.of\ terms\ in\ document}} \qquad (1)$$

- **Inverse Document Frequency: IDF**

This measures is used to find the importance of a term. It is possible that most frequent term is of less importanct i.e. common words which occure frequently. Therefore occurrence of a term within a corpus is taken into account.

$$\mathrm{IDF(t,d)} \ = \log(\frac{\text{No.of documents in a collection}}{No.of\ documents\ in\ which\ a\ term\ occurs}) \qquad (2)$$

- **TF-IDF Measure:**

It is a product of two above mentioned measures.

$$TF\_IDF = \mathrm{TF} \ \times \mathrm{IDF} \qquad (3)$$

Once we have calculated the TF/IDF score, we have a matrix of weight of each term against each document.

2.4.3. Hierarchy Formation Stage

Hierarchy Formation is also known as clustering, it is the stage in which the objects are organized in the form of a hierarchical structure based on the identified hierarchical relationships among them. In hierarchy formation phase, Hierarchical Agglomerative Clustering Algorithm (HAC) using Average Link [44] was used in the experiment for generating unlabeled hierarchy. This is two sub-steps process, first step is the determination of parent-child relation also known as hierarchy relationship identification step and second steps is an arrangement of these relationships in the form of hierarchical structure (hierarchical clustering), cosine similarity and Euclidian distance. We have used cosine similarity in our work for the calculation of similarity. We can calculate the similarity between two documents using the magnitude and direction of their vectors. Cosine similarity is the measure which is used to calculate the similarity between two vectors using dot product. Similar documents have value of dot product 1 and dissimilar documents give 0. Following is the formula for calculating cosine similarity where d1 and d2 are two documents. Numerator is the dot product of two vectors whereas denominator is the product of magnitude of those vectors.

$$\text{Cosine Similarity (d1, d2)} = \frac{\text{Dot product(d1,d2)}}{|\text{d1}| \times |\text{d2}|} \qquad (4)$$

For the arrangement of concepts and their hierarchical relationships in the form of a hierarchy, different taxonomy generation techniques have adopted different approaches. The hierarchical clustering techniques combine the above mentioned sub-stages i.e., hierarchical relationship identification and hierarchy generation. Clustering is the division of similar objects into groups. Documents those are similar to each other in some ways are kept together to form a cluster and dissimilar objects are kept away into other clusters. This is the unsupervised form of learning where no prior information (i.e., no training data) about cluster labels or classes is provided, we have used hierarchical agglomerative clustering (HAC) algorithm with three different average link as mean of arrangement of hierarchy. In average link we define the distance between two clusters as the average distance between the data point in first cluster and data points in the second cluster, and then we combine those two clusters with the minimum average link distance between them i.e. two most similar documents will be combined. This will give us the unlabeled hierarchy of clusters at this stage.

2.4.4. Node Labeling Stage

The hierarchical structure formed as a result of the hierarchy formation stage, is unlabeled and is not taxonomy in its real sense at this stage. Some more processing is needed in order to convert the hierarchical structure in the form of labeled taxonomy. Nodes labeling is more appropriate to apply in clustering based techniques. Since clustering is unsupervised and no labels are assigned before actual clustering has been done. So nodes labeling is most appropriate to apply in those taxonomy generation techniques that use clustering based approach in the hierarchy formation stage. In clustering based approaches centroids of clusters are involved in finding labels for taxonomic nodes. We have used top five terms approach in our research which combines 5 most relevant terms in a cluster which can identify a node and represent a document.

2.5. Evaluation Metrics

To compare the generated taxonomy with the gold taxonomy (i.e., manually constructed), we have used simple and pragmatic metric for evaluation, i.e., Content Quality Metric (CQM) [6]. It measures the quality of the content (i.e., labels). In other words, it measures overlap in the labels present in a generated taxonomy with that of a gold taxonomy. The labels of a generated taxonomy *taxLabels(T_{gen})* are compared to the labels of the golden taxonomy *taxLabels(T_{gold})* to measure the overlap in the labels for measuring the quality of the content. The quality of content is measured through precision CQM-P, recall CQM-R and F1-measure. CQM-P is the percentage of labels in a generated taxonomy (T_{gen}) that appear in the gold taxonomy (T_{gold}) and CQM-R: is the percentage of labels in T_{gold} that appear in T_{gen}. They are computed as follow.

$$CQM\text{–}P = \frac{|taxLabels(Tgen) \cap taxLabels(Tgold)|}{|taxLabels(Tgen)|} \qquad (5)$$

$$CQM\text{–}R = \frac{|taxLabels(Tgen) \cap taxLabels(Tgold)|}{|taxLabels(Tgold)|} \qquad (6)$$

$$F1\text{–}measure = \frac{2 \times (CQM\text{–}P \times CQM\text{–}R\ )}{CQM\text{–}P + CQM\text{–}R} \qquad (7)$$

2.6. Summary of chapter

This chapter proposes a methodology to perform comparative analysis of impact of using various preprocessing techniques in taxonomy generation process. It also explained the evaluation metrics to measure the quality of automatic taxonomy generation systems

CHAPTER 5:

EXPERIMENTS AND RESULTS EVALUATION

3. Experiments and Result Evaluation:

In this chapter, we have discussed dataset, its division and impact on results. Various experiments have been carried out which are discussed in details. Evaluation matrices and results are discussed in detail in the following sections. The results are presented both in tabular and graphical forms.

3.1. Experimental Setup:

This section describes the details of the experiment that have been carried out in this research. The experimental setup involves following stages:

- Dataset Specification
- Method
- Execution

3.1.1. Dataset Specification

Data sets were selected from two domains, where gold taxonomies are available which are required evaluation process, i.e.,

- Medical - Medical Subject Headings (MeSH) [41]
- Computer Science - ACM Computing Classification System (ACM CCS) [42]

Total 242 medical documents were collected from MEDLINE digital library and 492 computer science documents were collected from ACM digital library. The above mentioned datasets are divided in two subgroups to check the impact of data size. The experiments are performed first on full data set of ACM and Medline and then on a sample of half data set to measure the effect of full population and the sample of population of documents. The data set is divided into two groups as shown below.

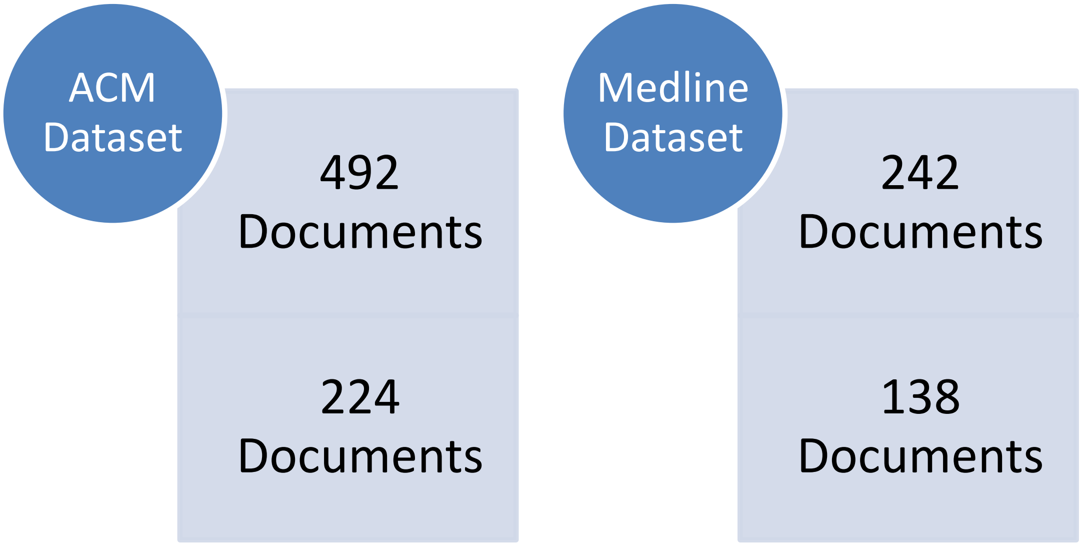
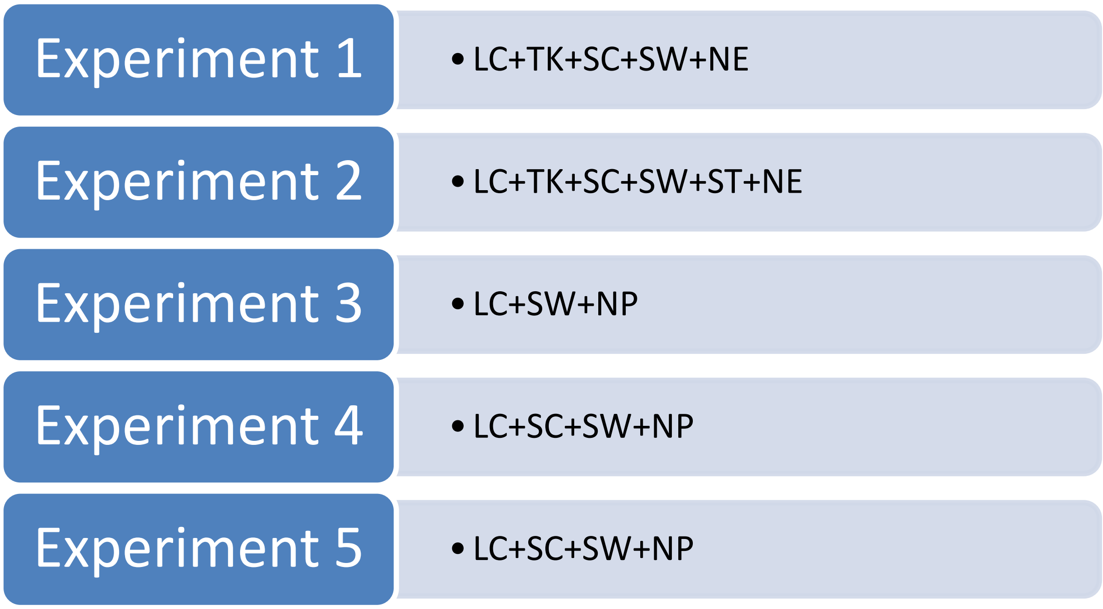


Figure 5: Division of Datasets

3.1.2. Method

In order to study the impact of various preprocessing techniques on the quality of the generated taxonomy, five different combinations of preprocessing techniques that have been used in the literature, were considered in this research for experiment. The combinations are discussed in section 4.3. Preprocessing techniques: lowercase conversion, Tokenization, special character removal, stop-word removal, stemming, noun extraction and noun phrase extraction are used as explained before. Five different experiments are performed on each data set including specific preprocessing techniques, as a result each data set will give five taxonomies i.e. total 20 taxonomies are generated.



Experiment 1	• LC+TK+SC+SW+NE
Experiment 2	• LC+TK+SC+SW+ST+NE
Experiment 3	• LC+SW+NP
Experiment 4	• LC+SC+SW+NP
Experiment 5	• LC+SC+SW+NP

Figure 6: Experiments on data sets

3.1.3. Execution

Five combinations of different preprocessing techniques are executed individually on four different sized data sets. This results in total of 20 taxonomies after overall execution of experiments.. Development of the system was performed in Java and python languages, and Netbeans IDE 8.1 and MySQL Workbench 6.3C were used as development environment. First of all, preprocessing was applied on the datasets using java and python scripts. Then, those short listed terms were used for the generation of taxonomy

Due to the requirement of extensive computing, the experiments were performed on super computer machine located in Research Center for Modeling & Simulation (RCMS) National University of Sciences & Technology (NUST), having following hardware specifications.

- Two 2.27 GHz 64bit Intel 4-core Xeon E5520 processors
- 8 physical cores (16 logical cores if using Hyper-Threading)
- 24GB DDR3 RAM
- 2 x 250GB SATA Hard Drives

And following software specifications

- JDK-8
- MySQL-5.5
- python-2.7

3.2. Results & Discussion

To evaluate the impact of pre-processing, the present work assess the evaluation metrics (CQM-P, CQM-R and F1 measure) against different combinations of pre-processing techniques (refers to Table 4) over two different datasets (492 computing and 242 medical documents). In addition to aforementioned evaluation metrics, the impact of the size of a given dataset is also evaluated. Following is the tabular representation of the results:

Table 5: Experimental Results

Combinations	**Data sets**											
	ACM 492			**ACM 224**			**Medline 242**			**Medline 138**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Combination 1	6.25	6.54	6.39	6.05	6.04	6.09	3.61	4.05	3.81	3.66	4.25	4.11
Combination 2	10.79	11.3	11.03	11.81	11.24	11.23	4.8	5.4	5.09	4.92	5.7	5.29
Combination 3	18.75	19.64	19.18	19.25	20.74	20.38	13.2	14.86	14	14.25	15.86	15.2
Combination 4	25.0	26.19	25.58	25.9	27.29	26.78	18.07	20.27	19.1	17.07	19.17	18.2
Combination 5	31.25	32.73	31.97	29.25	30.73	30.52	27.71	31.08	29.29	29.71	32.08	30.69

Figure 7 shows the results of different combination of pre-processing techniques performed on ACM dataset having 224 documents, while Figure 8 shows the results for 492 ACM documents. Similarly Figure 9 & 10 show the results of different pre-processing combinations on Medline 138 and 242 documents respectively. The data is shown in both tabular and graphical form to give the clear understanding of the results.

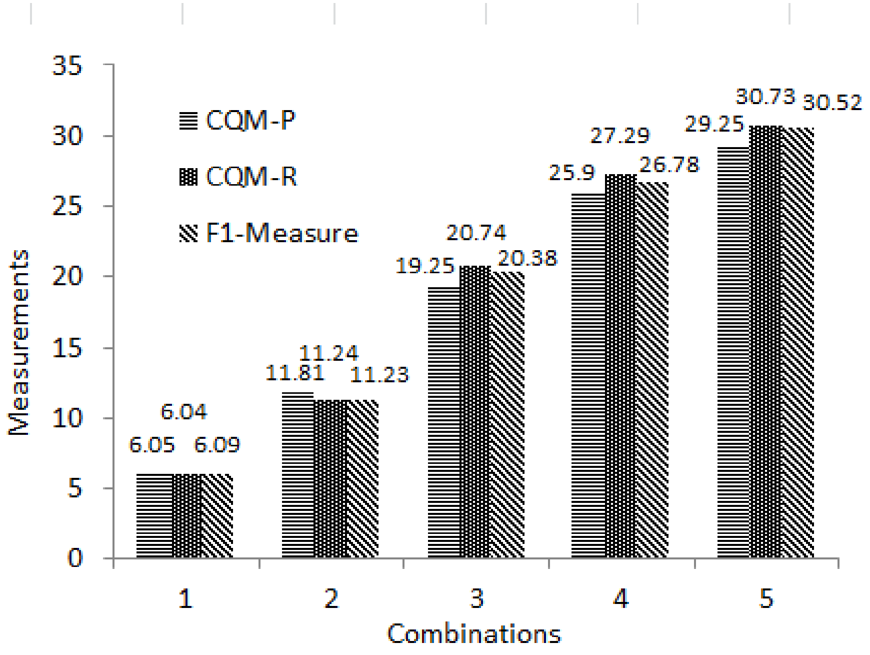


Figure 7: Experimental results for ACM 224 documents

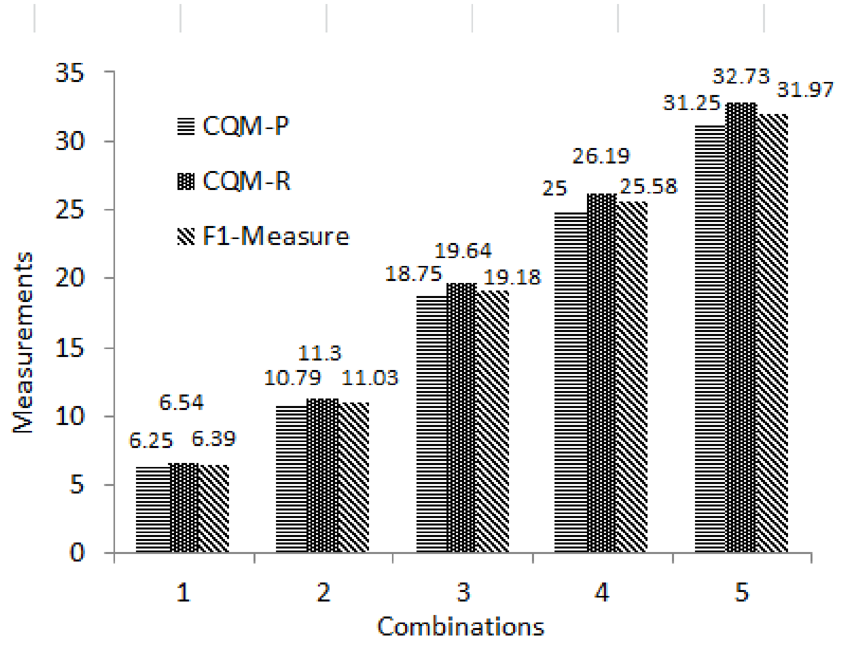


Figure 8: Experimental results for ACM 492 documents

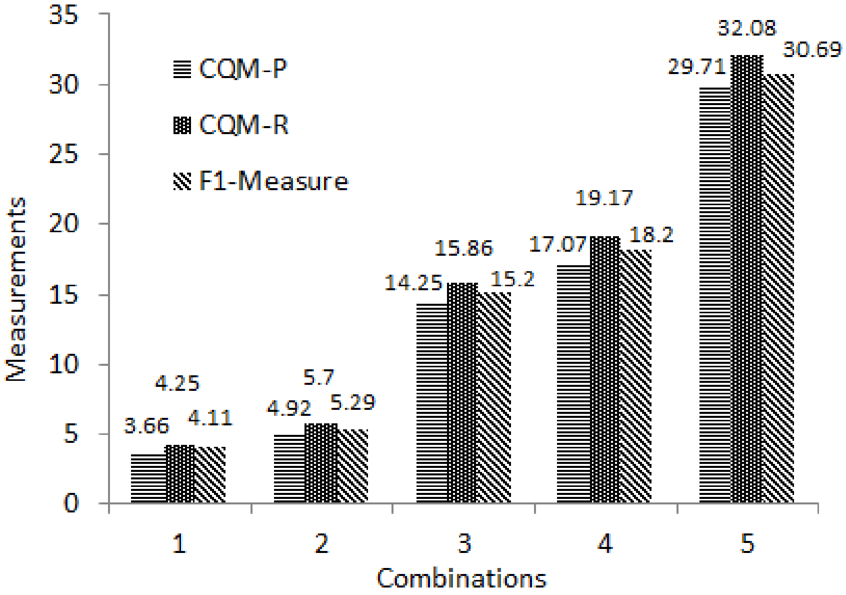


Figure 9: Experimental results for MEDLINE 138 documents

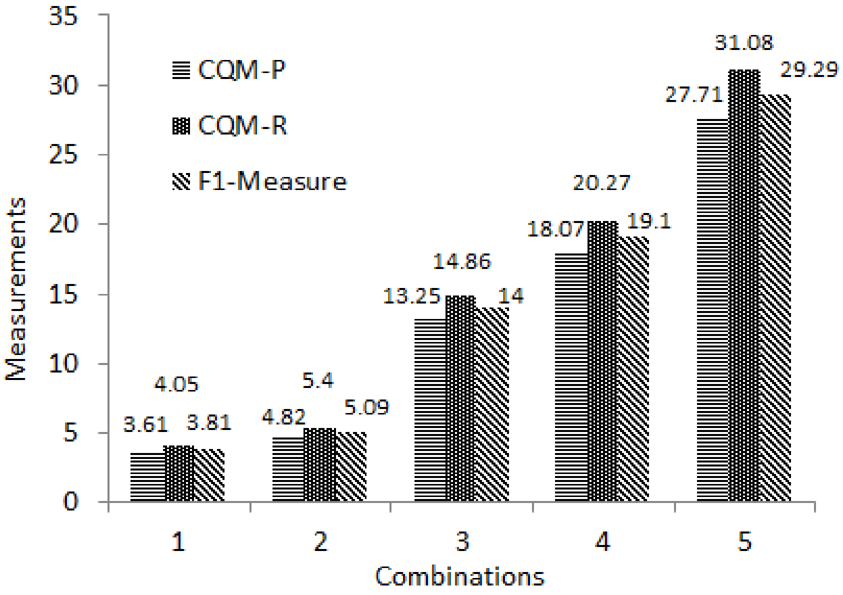


Figure 10: Experimental results for MEDLINE 242 documents

3.2.1. Impact of different Dataset Size

From the results shown above, it is apparent that there is no significant difference among the results performed on different size of datasets. It reflects that pre-processing techniques has a similar behavior for varying number of documents, hence size of data set do not impact the results in any way.

3.2.2. Impact of different combinations

Strong evidence is visible from the results that the difference among combinations was primarily due to their technique selection. Combinations 3, 4, 5 with Noun Phrase (NP) technique (i.e., as shown in Table 5) have high F1 measure value (Figure 7, 8, 9 and 10) over combinations without NP technique, i.e., Combination 1 and 2. This clearly shows the advantage of using NP pre-processing technique for auto-generation of taxonomy. Moreover, a careful investigation for this high F1-measure value of combinations with NP technique reveals that taxonomies heavily relies on multi-gram noun phrases instead of uni-gram nouns, for example, *computer System Organization* and *neoplastic processes* (i.e., as shown in Table 2 & 3). Hence, these pre-processing combinations (i.e., Combination 1 and 2) rely only on noun extraction (NE) fails to achieve high results. A restriction is visible that combinations using NP do not use tokenization, whereas tokenization is used with NE technique. Another reason is that tokenization may eliminate the semantics, whilst taxonomy primarily focuses on semantics of phrases, e.g., *computer architecture*.

The only difference between combination 4 and 5 is the use of stemming technique. The results of these two combinations have a noticeable difference. The combination 5 involving stemming has high F1-measure value as compared to combination 4. This shows the benefit of using stemming for auto-generation of taxonomies. The difference in results based on usage of the stemming technique is even higher in medical documents. Furthermore, combination 3 and 4 uses similar pre-processing techniques except the special character removal (SC). A higher result has been observed for combination 4 containing SC technique, whereas the combination 3 has lower results without including SC technique. Overall, the impact of SC technique is the same for both computing and medical domain documents. The benefit of stop word removal (SW) is observable by comparing combination 1 and 2. The use of SW has increased the results of combination 2 as compared to combination 1 without SW. The difference in results due to SW technique is higher in computing documents.

Among all combinations, the minimalistic use of pre-processing techniques was by the combination 3. Consequently, combination 3 has the average result that shows a combination using all pre-processing techniques does not guarantee a high F1 measure value. Combination 2 has involved nearly all pre-processing techniques, but its results are no better than the lowest one. From the results, it can be concluded that appropriate selection of the pre-processing techniques is vital for any given domain.

3.3. Summary

In this chapter, we have discussed the experimental setup, execution and results. The chapter included the detailed discussion on the data set division on the basis of size, the preprocessing combination performed in each experiment, hardware and software specifications of the system and experimental results. We have included the results both in tabular as well as graphical form above. After analyzing the results, it is concluded that performing all the preprocessing techniques at once can decrease the results, however, performing selective techniques can improve the results. It is concluded that extracting multi-gram noun phrases can significantly improve the taxonomy results.

CHAPTER 6:

CONCLUSION

4. Conclusion:

4.1. Research Summary

In this research, the impact of various preprocessing techniques on the quality of the generated taxonomy was investigated. Five different combinations of preprocessing techniques that have been used in the literature were considered in this research for experiments. Two data sets from two different domains: medical and computer science were selected because both the domains have their gold standard taxonomies that were used in evaluation. Content Quality Metric (CQM) was adopted for evaluation to measure the quality of the content of generated taxonomies. The experiment was run with different data sizes in each data set to know the impact of data size on quality of generated taxonomies. However data size has no significant impact on quality of generated taxonomies. The results of various combinations of preprocessing techniques can impact the quality of the generated taxonomy. It has been observed that the selection of multi-gram noun phrases from text documents can generate taxonomy of reasonable quality in Medical and Computer Science domains because their taxonomies contain mostly multi-gram phrases. Moreover applying all preprocessing techniques in a generation process does not guarantee high quality taxonomy because more essential information may be lost in this way. Further study can be carried out in domains other than Medical and Computer Science to know the impact of preprocessing on the quality of generated taxonomy

4.2. Outcome of research

The main purpose of this research was to set a benchmark of preprocessing techniques and their impact on taxonomy generation, i.e., which technique gives best quality of taxonomy.

However, the difference among combinations was primarily due to their preprocessing technique selection. Combinations 3,4 and 5 with Noun Phrase (NP) technique (i.e., as shown in table 4) have high F1 measure value over combinations without NP technique, i.e., Combination 1 and 2. This clearly shows the advantage of using NP pre-processing technique for auto-generation of taxonomy. Moreover, a careful investigation for this high F1-measure value of combinations with NP technique reveals that taxonomies heavily relies on multi-gram noun phrases instead of uni-gram nouns. Therefore, using NP extraction gives best quality of taxonomy as compared to using other techniques.

4.3. Future Work

4.3.1. Limitations

- The preprocessing techniques used in this study are evaluated on the hierarchy generated on the dataset of ACM Digital Library and Medline therefore it works well for computing and medical domain only.

- The taxonomies evaluated only on the precision recall and f-measure metrics. They can also be evaluated on the basis of hierarchy structure based evaluation metrics e.g. hierarchal tree levels (depth and spread).

4.3.2. Future work

Following are the recommendations to further extend this research:

- Various preprocessing techniques and their combinations can be used to enhance the scope of research.
- All three flavors of HAC i.e. single link and complete link and average link can be used to analyze the impact of various clustering techniques on generated taxonomy.
- The experiments can be conducted on other datasets as well, for which golden taxonomy is available. The datasets should not be limited to computing and medical domain only as results can vary by changing datasets.
- In this research, only document based taxonomy has been discussed. In future the impact of preprocessing techniques can be analyzed on concept based taxonomies.
- Taxonomy can be evaluated on the basis of other evaluation matrices as well.

REFRENCES:

[1] M.-S. Paukkeri, A. P. Garc'\ia-Plaza, V. Fresno, R. M. Unanue, and T. Honkela, "Learning a taxonomy from a set of text documents," *Appl. Soft Comput.*, vol. 12, no. 3, pp. 1138–1148, 2012.

[2] R. Sujatha, R. Bandaru, and R. Rao, "Taxonomy construction techniques--issues and challenges," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 5, 2011.

[3] W. Engel, C. Pryde, and P. Sappington, "Method and system for enhanced taxonomy generation." Google Patents, 2010.

[4] H. Hedden, *The accidental taxonomist*. Information Today, Inc., 2016.

[5] T. Li and S. Anand, "Exploiting domain knowledge by automated taxonomy generation in recommender systems," *E-Commerce Web Technol.*, pp. 120–131, 2009.

[6] V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth, "TaxaMiner: an experimentation framework for automated taxonomy bootstrapping," *Int. J. Web Grid Serv.*, vol. 1, no. 2, pp. 240–266, 2005.

[7] E.-A. Dietz, D. Vandic, and F. Frasincar, "Taxolearn: A semantic approach to domain taxonomy learning," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, 2012, pp. 58–65.

[8] R. Blumberg and S. Atre, "The problem with unstructured data," *Dm Rev.*, vol. 13, no. 42–49, p. 62, 2003.

[9] A. Muller, J. Dorre, P. Gerstl, and R. Seiffert, "The TaxGen framework: Automating the generation of a taxonomy for a large document collection," in *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*, 1999, p. 9--pp.

[10] S. P. Ponzetto and M. Strube, "Deriving a large scale taxonomy from Wikipedia," in *AAAI*, 2007, vol. 7, pp. 1440–1445.

[11] Y. Song, S. Liu, H. Wang, Z. Wang, and H. Li, "Automatic taxonomy construction from keywords." Google Patents, 2016.

[12] S. A. Caraballo, "Automatic construction of a hypernym-labeled noun hierarchy from text," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 120–126.

[13] V. Srividhya and R. Anitha, "Evaluating preprocessing techniques in text categorization," *Int. J. Comput. Sci. Appl.*, vol. 47, no. 11, pp. 49–51, 2010.

[14] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *J. Am. Med. Informatics Assoc.*, vol. 18, no. 5, pp. 544–551, 2011.

[15] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[16] S. Kannan and V. Gurusamy, "Preprocessing Techniques for Text Mining." 2014.

[17] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, "The Impact of Pre-processing on the Classification of MEDLINE Documents.," in *PRIS*, 2010, pp. 53–61.

[18] W. Zhou, N. R. Smalheiser, and C. Yu, "A tutorial on information retrieval: basic terms and concepts," *J. Biomed. Discov. Collab.*, vol. 1, no. 1, p. 2, 2006.

[19] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.

[20] M. C. Velilla, “Taxonomies for categorisation and organisation in Web sites,” *Hipertext. net Anu. Acad{é}mico sobre Doc. Digit. y Comun. Interactiva*, no. 3, 2005.

[21] J. Vernau, “The Business Benefits of Taxonomy,” 2005.

[22] G. Sacco, “Dynamic taxonomy process for browsing and retrieving information in large heterogeneous data bases.” Google Patents, 2004.

[23] G. M. Sacco, “Dynamic taxonomies and guided searches,” *J. Assoc. Inf. Sci. Technol.*, vol. 57, no. 6, pp. 792–796, 2006.

[24] M. Earl, “Knowledge management strategies: Toward a taxonomy,” *J. Manag. Inf. Syst.*, vol. 18, no. 1, pp. 215–233, 2001.

[25] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[26] A. Hotho, A. Nürnberger, and G. Paaß, “A brief survey of text mining.,” in *Ldv Forum*, 2005, vol. 20, no. 1, pp. 19–62.

[27] S. Kashyapi and M. Kumari, “RESEARCH ISSUES IN TEXT CATEGORIZATION BASED ON MACHINE LEARNING: A REVIEW.”

[28] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[29] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Trans. neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[30] I. C. Mogotsi, “Christopher d. manning, prabhakar raghavan, and hinrich sch{ü}tze: Introduction to information retrieval.” Springer, 2010.

[31] A. Popescul and L. H. Ungar, “Automatic labeling of document clusters,” *Unpubl. manuscript, available http//citeseer. nj. nec. com/popescul00automatic. html*, 2000.

[32] O. Medelyan, S. Manion, J. Broekstra, A. Divoli, A.-L. Huang, and I. H. Witten, “Constructing a focused taxonomy from a document collection,” in *Extended Semantic Web Conference*, 2013, pp. 367–381.

[33] S.-L. Chuang and L.-F. Chien, “Taxonomy generation for text segments: A practical web-based approach,” *ACM Trans. Inf. Syst.*, vol. 23, no. 4, pp. 363–396, 2005.

[34] B. Cui, J. Yao, G. Cong, and Y. Huang, “Evolutionary Taxonomy Construction from Dynamic Tag Space.,” in *WISE*, 2010, pp. 105–119.

[35] M. Yetisgen-Yildiz and W. Pratt, “The effect of feature representation on MEDLINE document classification,” in *AMIA annual symposium proceedings*, 2005, vol. 2005, p. 849.

[36] J. R. Méndez, E. L. Iglesias, F. Fdez-Riverola, F. D’\iaz, and J. M. Corchado, “Tokenising, stemming and stopword removal on anti-spam filtering domain,” in *Conference of the Spanish Association for Artificial Intelligence*, 2005, pp. 449–458.

[37] J. Pomikálek and R. Rehurek, “The Influence of preprocessing parameters on text categorization,” *Int. J. Appl. Sci. Eng. Technol.*, vol. 1, pp. 430–434, 2007.

[38] A. G. Jivani and others, “A comparative study of stemming algorithms,” *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[39] H. Saif, M. Fernández, Y. He, and H. Alani, “On stopwords, filtering and data sparsity for sentiment analysis of twitter,” 2014.

[40] T. Kuzar and P. Navrat, “Preprocessing of Slovak Blog Articles for Clustering,” in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International*

Conference on, 2010, vol. 3, pp. 314–317.

[41] "Medical Subject Headings," 2017. .

[42] "ACM Computing Classification System," 2017. .

[43] M. Y. Dahab, H. A. Hassan, and A. Rafea, "TextOntoEx: Automatic ontology construction from natural English text," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 1474–1480, 2008.

[44] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval Cambridge University Press, 2008," *Ch*, vol. 20, pp. 405–416.

[45] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[46] N. Zong, D. Hyuk Im, S. Yang, H. Namgoon, and H. Gee. Kim. Dynamic generation of concepts hierarchies for knowledge discovering in bio-medical linked data sets. ICUIMC '12, pages 12:112:5, New York, NY, USA, 2012. ACM.

[47] G. Punj and D. W. Stewart Cluster Analysis in Marketing Research: Review and Suggestions for Application.