A Dissertation Proposal

On

Keyword Based Identification of Thrust Area: A KDD Approach

Submitted To

LOVELY PROFESSIONAL UNIVERCITY

In partial fulfillment of the requirement for the award of degree of

MASTER IN PHILOSAPHY (M.Phil)

In

COMPUTER SCIENCE

Submitted by:

Ms. Nirmal Kaur

Reg.No. 11512292

Supervised by:

Dr. Manmohan Sharma

SCHOOL OF COMPUTER APPLICATION

LOVELY PROFESSIONAL UNIVERCITY

PHAGWARA (PUNJAB)

ACNKOWLEDGEMENTS

I really would like to thanks to God for giving me the honor of writing this dissertation with enormous of devotion and profound sense of appreciation. With an overwhelming sense of desire and self-importance, I express my external appreciation and immensity to my esteemed guide, Dr. Manmohan Sharma, Department of Computer Application, Lovely Professional University, Phagwara, Whose cordial guidance, motivating ideas and reassurance helped me in all the time of research and writing of my thesis.

I really feel short words to express my heartiest gratitude and sincere thanks to my parents, my professors and my friends for their support and all they have comprised for me during tenure of my work.

Lastly and most importantly, I wish to thanks to all those who were directly and indirectly involved in the completion of this work thereby making it a total success.

Signature of Candidate

Nirmal Kaur

DECLARATION

I hereby declare that the dissertation entitled **"Keyword Based Identification of Thrust Area: A KDD Approach"** is an authentic record of my own original work carried out for the award of degree of M.Phil (Computer Application) and all ideas and references have been duly acknowledged. The matter presented in the dissertation has not been submitted in part or full to any other universities and institute for the award of any degree.

Date: 09 Dec 2016

Signature of the student

Nirmal Kaur

Reg. No: 11512292

CERTIFICATE OF THE SUPERVISOR

This is to certify that the work **"Keyword Based Identification of Thrust Area : A KDD Approach"** is a section of research work done by **Ms. Nirmal Kaur** under my guidance and supervision for the degree of Master of Philosophy in computer Science of Lovely Professional University, Phagwara, Punjab, India. To the best of my knowledge, the present work is the result of his original analysis and study. No part of the project report has ever been submitted for any other degree or diploma. The dissertation is apt for the submission for the partial fulfillment of the conditions for the award of M.Phil in computer science.

Date: 09 Dec 2016

Signature of Supervisor

Dr. Manmohan Sharma

Table of Contents

Page No

List of Figures

Page No.

List of Tables

Page No.

Chapter 1

Knowledge Discovery from Databases

1.1 Introduction

Knowledge Discovery from Databases is an advanced process to identify new from the existing data. Data are stored in databases in a very large amount so that knowledge discovery processes fetch great attention for extracting meaningful data from the database. According to knowledge discovery, it may be grouped into information generalization, information retrieval, association, and classification and clustering of data. There are various techniques are used to identify the appropriate data from the databases. For example Medical Databases, Libraries Databases, Educational Databases, Financial Databases etc. The concept of knowledge discovery is used to find motivating and essential useful designs patterns or procedures in data. Knowledge discovery briefly produce the complete information extraction method, as well as however data are stored and retrieved, it produce cost-effective and efficient algorithms to explore huge information, it shows how to visualize and interpret the results, and also to building and support the communication between machine and human. It conjointly considers support for knowledge and analyzing the application domain [1]. Knowledge Discovery approaches lead to discovering the knowledge about inductive learning, semantic query optimization, Bayesian statistics, and knowledge extraction for expert systems. KDD processes include the characteristics of data integration, data exploration, and a frequency distribution by matching some keywords and find relevant information the relevancy of data. Data mining is the essential process of Knowledge discovery. The investigation in databases and data origination has presented ascend to a approach to compact with store and control this useful information for further decision making [2]. Data mining uses the information from data warehouse and develops a knowledge system by using AI and statics related techniques to find out useful patterns that can be further used by the user [3]. With the help of Data Mining to find the knowledge from two datasets by factor and comparison the classification technique is used to classify the information or data by taking prediction. Compare data in groups by using cluster technique. Segmentation

of data by pattern evaluation and visualization of data by taking alternatives and data modeling used to make a model by prediction and apply regression and dependency of variables. Knowledge discovery plays an important role in information retrieval system for information extraction by using searching algorithms and to fulfill the searching mechanism the algorithm should be efficient. In order to measure the efficiency of an algorithm, the complexity should be polynomial with respect to time and space [4]. These two factors are most important to make an efficient algorithm. Knowledge discovery in textual data is victimization the best kind of data extraction, specifically the categorization of a topic of a text by significant ideas [4]. Text categorization used in textual databases for divide the text into different classes, such as text classification. Data mining is a region of strong activity for developing new algorithms and techniques as a research topic but additionally an application area wherever realistic advantage are to be created [5].

Figure 1: Knowledge Discovery Process

1.2 Knowledge Discovery Process

The process "starts" with determinant the Knowledge discovery goals, and "ends" with the execution of the revealed data [6].

- **Goal: -** This step includes the goal of Knowledge discovery from databases and understanding the application domain with preprocessed data.

- **Data Selection: -** This includes the selection of data from the application domains and selects the required knowledge then set data as target data set on which discovery will be performed.
- **Data Cleaning: -** This step includes the data cleaning process and preprocesses the data by determining useful approaches to filling missing fields and uses the data as per requirements needed in data cleaning.
- **Data Transformation: -** This step includes the process of data transportation and converts the data into useful knowledge by using data mining techniques.
- **Pattern Evaluation: -**In this Step, data mining techniques identify the useful pattern and apply the algorithm for evaluation.
- **Knowledge Representation: -** In this step by evaluating patterns useful and understandable knowledge will be gained from unstructured data.

1.3 Traditional Approaches of Knowledge Discovery

- **Database Query-** To access a database for gain interesting knowledge which presents by a well-defined query language like SQL. These queries processed the data from the database and give results to the user. The results usually a subcategory of the database.
- **Data Query or Data Mining –** In Data query the knowledge represented by four types that can be defined data mining techniques on appropriate data, such as "Shallow Knowledge, Multidimensional Knowledge, Hidden Knowledge, Deep Knowledge".
  - **Shallow Knowledge-** This type of knowledge can be simply manipulated and stores in the database. In query language, SQL provides extraction of shallow knowledge from a database that is factual knowledge.
  - **Multidimensional Knowledge-** This type of knowledge is also factual and Online Analytical Processing (OLAP) used in this type of data.
  - **Hidden Knowledge-** This type of knowledge cannot found by SQL query language easily. There is a need to apply some data mining algorithms to discovering the hidden knowledge with ease.
  - **Deep Knowledge-** This type of knowledge only found when we are given some clue or direction about the appropriate data.

1.4 Role of Data Mining in Knowledge Discovery

The process of data mining used to "Extract, transform, and load transaction" from the data warehouse system by evaluate the raw data. By using data mining techniques we can manage and store the data in a multidimensional database systems [7]. Data mining contains set of algorithms, tools and methods used for evaluating and analyzing the data or match the useful pattern from data. Merriam-Webster said about data mining "*It is the process of searching huge amount of digital or computerized data to find interesting results or trends"*. Data mining comprises of applying knowledge enhancement and discovery algorithm by using inductive and statistical data modeling to harvest a particular record or patterns or results over the data [7, 8]. There are a number of major data mining techniques that have been used in data mining projects recently including association, clustering, classification, prediction and sequential patterns and regression models.

- **Association Rule**

It is a pattern discovered method in data mining by established on a relationship of a specific item on other items in the identical transaction.

- **Classification**

The classification method based on machine learning used in data mining technique. Mostly classification methods are used to classify every item in a set of predefined classes on the basis of similarity.

- **Clustering**

It is the process of establishing the data into clusters with common characteristics whose supporters are related to each other in a specific manner. In this method, we make new classes according to data requirements.

- **Prediction**

Prediction method is one of a data mining procedures that is used to recognize the relationship between independent variables and also identify the relationship in dependent and independent variables over the data.

- **Sequential Patterns**

Sequential patterns analysis used to discover statistically relevant patterns in data transaction. It is a subsequence that acts in a number of sequences of the database.

1.5 Keyword Matching Mechanism

In the field of databases, various keywords matching and pattern matching algorithm developed by these techniques such as Frequent item set, Sequential pattern searching, probability based keyword searching, n-gram etc. keywords are most important terms in document or text fields. In every class containing keywords in C_1 (k1, k2, k3,…….kn). We can calculate the occurrence frequency of any keywords in database by Calculate "Term Frequency and Inverted Document Frequency (TF-IDF)" of those particular keyword. Various methods developed in text mining by using joining tree to match keywords from the database. Keyword searching mechanism also used in term-weighting and topic identification.

Matching is a very useful operation in the ancient application, like data integration, data repositing, distributed query process and so on [9]. Keyword search concluded structured data such as relational databases is a progressively important capability [10.11.12] taking benefits of an arrangement of database and information retrieval techniques. One of the most significant problems for facilitating such as query facility is to be choosing the most convenient data sources related to the keyword query. Mostly Information retrieval systems used keyword-frequency statistics for textual databases [13].

Data mining techniques have previously been utilized for text exploration by extracting co-occurring relationships as descriptive phrases from document collection [14]. A keyword based approach presents an important role in database query to give useful results that can be used in any database for the purpose of retrieval of data from the database based on keyword search using different technologies such as CGI (Common Gateway Interface), JavaScript and Servlet, Java Swing etc. these programs accept query from user and processed query with keyword matching mechanism and give appropriate result to the user [15]. In information retrieval systems keyword based searching plays an important role. Keyword based model use the idea of semantic search in information retrieval system. Knowledge representation in documents by term vectors leads to high dimensionality problem in index terms in textual data so clustering methods can be used to overcome these problems. [16]

1.6 Keyword based Querying methods

These methods are effective for keyword-based querying mechanism in database systems that can be generally categorized as given below:

Graph Based Query Systems: These type of systems, such as the BANKS, the database is converted into an illustration level data presented by graph where authorities specify the several ways in which tuples or records are linked.

SQL-Query Based Systems: These type of systems, each keyword query is transformed into a regular set of SQL queries and performs a joint operation on the basis of foreign key concepts.

Composite Systems: These types of systems try to influence benefits of both graph based and SQL based approaches, e.g. the ESKO system [17].

In this thesis we are using MapReduce algorithm for keyword searching by some modification according to our work. So we introduce about MapReduce algorithm work in keyword searching mechanism.

1.7 MapReduce Algorithm

MapReduce is a model used for programming to produce huge data sets and an application for processing and generating clusters from data sets. In this model the written programs are parallelized and executes naturally on wide clusters. [18]. MapReduce algorithm provides the mechanism for Sorting, Searching, calculate TF-IDF, Breath-first search, Page Rank and some more advanced features. It represents a data flow more than a procedure. It checks the key value line by line then sort the mapper value according to the contents. The main concept behind map reduce is "mapping" your data set into a collection of (key, value) pairs, and then it reducing overall pairs with the same key. We use recognized map reduce for counting word frequencies in a large text files.

Workflow of map reduce: Map reduce work flow starts with the framework that split the data inputs into segments then passing each segment into a different machine. The script takes input data and maps the data into key and value < key, value> pairs according to our requirements [19]. Mapper are used to Map the key in different input sources and then reducer creates the blocks of keyword by apply sorting and then blocks are divided into single block by counting key value as per their frequency.

Chapter 2

Literature Review

2.1 Related Work

This chapter describes the related work of this thesis, also providing some related concepts. More specifically, the following topics are discussed:

- Existing Relation of Knowledge Discovery with Data Mining, Information Retrieval System and Relational Databases.
- Existing Pattern Matching discovery techniques.
- Existing techniques for classification and clustering for Knowledge Discovery.
- Existing Keyword Extraction Mechanism for Knowledge Discovery
- Existing Keyword based searching techniques for knowledge discovery.

Matheus et al. [20] described a model construct for knowledge discovery from real-world databases. They compare three models such as EXPLORA, CoverStory, and knowledge discovery Workbench on these model they evaluates some components upon these models like controller, Database Interface, knowledge base, focus, pattern extraction, evaluation. EXPLORA is an integrated system for conceptually analyzing data and searching for interesting relationships. They have argued that autonomy requires domain knowledge where versatility implies domain knowledge. CoverStory is a commercial system developed by information resources and knowledge workbench is a collection of tools for the interactive analysis of large databases.

Usama et al. [21] described in brief all the steps of knowledge discovery process. They provide the overview of relationship between Data mining and Knowledge discovery in growing applications. Also examined the existing data mining tools and represented the work of data mining algorithms in the area of Artificial Intelligence, Neural Network, Machine Learning, Statistic and Databases. They discussed the issues for deploying successful applications. They also highlighted the

facts of describing the role of data mining algorithms in different disciplines in growing areas.

Gouda and Cheng [22] describe about the human behavior and describe how the human can solve the problem by using relevant methods for finding relevancy problem. The new discoveries are based on the human thinking and the process which is made by human, the new algorithm for knowledge discovery is based on human intelligence and the knowledge of human. In this they are using a reasoning process to solve this problem.

Kumar, and Rathee [23] shows the combination of classification and clustering techniques and gives the more accurate result than simple exiting classification techniques. This technique helps to classify the data set into useful information and convert into different attributes and classes. This technique develops rules for data set which contains missing values of data. A cluster technique used to build the learning technique in the form of numbers that defines the instances for classification. Also defines the decision rules which are applying on the both clustering and classification techniques. These decision rules helps in integrated the data from the data sets and apply the integration mechanism.

Ning Zhong et al. [24] describe about the various association rule mining techniques by using text mining. This mechanism provide the method for finding frequent data set, sequential pattern, close pattern, maximum pattern and apply rule mining. By this we find the pattern in textual data. It is quite difficult but tries to solve in this paper to discover the knowledge. They used two methods pattern deploying and pattern evolving to define the pattern in text documents.

Menaka and Radha [25] (2013) compares the three machine learning techniques such as Naïve Bayes, KNN and Decision tree. And shows that Decision tree are more efficient for keyword attraction among these three techniques in text classification. They preprocess the document by remove stop words, stemming, feature extraction and count tf-idf for keyword extraction.

S. M. Kamruzzaman [26] describes the three techniques to classify text first using Association rule with Naïve Bayes classifier which helps to collect the relevance text data. Second using Association rule with Decision tree which cleaning the text and implement the Apriori Algorithm to generate frequent item set and decision tree helps to find the associated class. Third is genetic algorithm which used Roulett Wheel selection method. They used Apriori Algorithm for finding the text from related table by using joins concepts and also count the words appearing in different tables.

Rekha Baghel et al. [27] proposed technique, FCDC (Frequent Concepts based document clustering), a clustering algorithm that works with frequent concepts slightly than frequent items used in traditional text mining techniques. They found that "FCDC technique" more accurate, effective and scalable and when compared with existing algorithms. Clustering algorithms like Bisecting UPGMA, K-means, and FIHC.

Seung-Shik Kang [28] proposed a content-based systematic method to enhance the similarity calculation and propose an algorithm for keyword-based clustering algorithm. They performed an experiment for the clustering of similar documents and the results showed that keyword-based weighting scheme is better than the frequency-based method.

Bei Yu et al. [29] Study the database selection problem for relational data sources, and propose a method that effectively summarizes the relationships between keywords in a relational database based on its structure. They develop effective ranking methods based on the keyword relationship summaries in order to select the most useful databases for a given keyword query. They have implemented their system on PlanetLab.

Pattan Kalesha [30] proposed work for specific document placement before their pattern Discovery and describe about the work of document preprocessing for extraction of keywords and also apply the tokenization method on extracted keywords from documents. They describe the work for probability calculation and document

clustering after pattern matching and also describe some rules and show accuracy of proposed techniques.

Patil and Uddin [31] proposed an ontology based text mining framework with clustering of documents of research papers. They proposed a framework that stores the last five years publications data and also upload new papers then classify the paper in respective Domain Area and apply clustering technique of k-means on classified document and distributed the paper according to their experts.

Akshita Thakkar et al. [32] address the problem of extracting learning objects or keywords from bunch of Documents, with the goal of use this objects for various purpose like Resume filtering, Email filtering, content classification etc. Keyword extraction, concept finding are in learning objects is very important subject in today's eLearning environment. In this proposed System they Calculate the TF-IDF of each word, then Decision tree algorithm is used for feature selection process using wordnet dictionary. WordNet is a lexical database of English which is used to find similarity from the candidate words.

Sarda and Jain [33] proposed a system for keyword based searching in databases and describe a prototype for searching in database named Mragyati. This approach is very scalable for searching relationships among the objects and user query. They implemented a data retrieved system on the web using common gateway interface and java script and shows the result by ranking on query in database and also provide the answer of foreign key concept in database. They optimized search query by query representation, query classification and query processing and result ordering for final implementation. They provided the prototype designing structure for semantics and intelligently answer for user's query.

Xiaofeng Li et al [34] proposed a search engine called sphinx to integrate the approach of full-text search from character and words. This engine optimized the search queries by matching segmentation algorithms. They apply queries for search full-text from database in the IPTV network server with input to improve the efficiency. They described the analysis of the information, content input method. In

this engine the searching work divided into three categories such as program name, description and several servers respectively.

Luping Li et al [35] overcome two problems in relational databases to improve the effectiveness and efficiency for keyword searching. They proposed an efficient ranking algorithm and a ranking model and perform the experiments on real data sets. They designed a new index which provide efficiency in time and size factors. Then they found top-k-results based on aggregate keyword query by ranking algorithm in performance evolution. The ranking algorithm takes less time for existing algorithms.

Arvind hulgery et al [36] presented a survey of keyword querying in database and describe about the BANKS which is an integrated database for keyword querying and interactive browsing. They provide the overview of model and their prototypes. Framework of this model describes informal description, formal database model with vertices, edges and edge weight and querying formulations. They implemented their approach based on Dijkstra shortest path algorithm. The BANKS system contribute a rich interface in database to browse the data. BANKS system provides schemas to query and integrated browsing with ease.

Zhaoxin Fan et al. [37] used the Neural Network Model with Bag of words technique to divided items in a bag and bag represents the text. Creates vectors of text before apply text clustering. k-means clustering used on vectors which contain Term frequency and Inverted Document frequency (TF-IDF) value of each word. Also used LDA (Latent Dirichlet allocation) for establishing the connection between document, themes and word.

Sonia Bergamaschi et al [38] proposed a system for keyword based searching in database with keymantic features that answers the keyword queries by depends on purposeful knowledge. Keymantic guess the intention of user by producing the possible understandings in database structures. Keymantic contains three modules such as wrapper, keyword mapper and interpretation generation where wrapper used for extracting the meta-data source, keyword mapper generates the configurations and their weights and interpretation generation submit the final SQL queries.

Vagelis Hristidis et al [39] proved that DISCOVER is useful for finding relevant candidate networks without redundancy in data. DISCOVER provide a simple interface for querying keyword on a search engine that associated with relationships and attached via more than one tuple. They provide a candidate network generation algorithm with efficiency and define greedy algorithm for creating a near optimal execution plan.

Qi Su and Jennifer Widom [40] made and architecture which supports highly efficient keyword based search. They made a system named ESKO which creates virtual documents from database by joining tuples and for indexing and keyword search processing used DB2 net search extender. Experiment results shows keyword queries response time is collaborative for representative queries.

Tru H.Cao et al. [41] shows the work of vector space model which are used for making different vectors of documents. They perform linear clustering and hard clustering which are used for clusters the document by Named entity, type and identifiers respectively. They shows and count no of named entity and shows documents which contain places, people, organization and exchange keywords and their alias. Also performed document indexing based on keyword occurring in text, and make a search engine named VN-KIN which support the hierarchical clustering.

2.2 Research Gap

When we talk about the relationship among knowledge discovery, information retrieval system, databases and ontologies for textual information there are various techniques developed.

- From 1989 knowledge discovery starts with artificial intelligence and machine learning techniques used for extract useful information from data.
- Mostly used classification and clustering techniques from data mining to classify the text or assigning the keywords to new cluster.
- In keyword matching different model were developed such as BANKS, DISCOVER, ESKO etc for keyword searching from database by generalized the query in keymantic and Semantic based approaches.

- Developed interfaces for querying keyword from user and classify these keywords based on frequent item sets, indexing and ranking of documents.

2.3 Problem Statement of Our Approach

Identification of the Thrust Area in Computer Science on the basis of Keyword searching with Knowledge Discovery approach.

2.4 Objectives of the Study

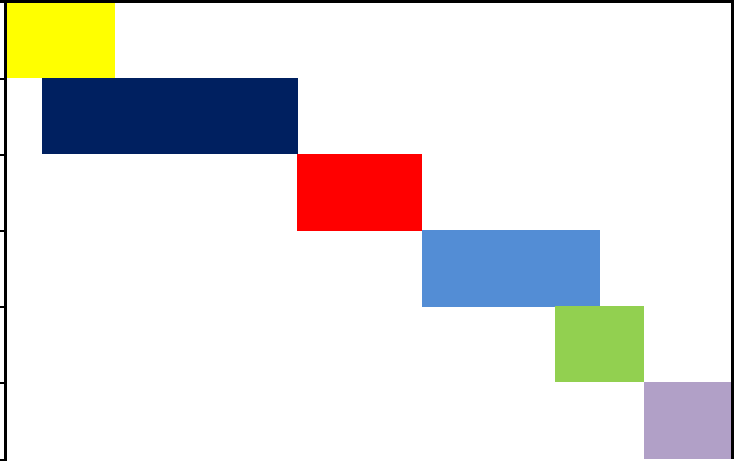
1. To develop a model for keyword searching and find out the thrust area on the basis of that keyword from database.
2. To calculate efficiency of an algorithm with respect to size of data.

2.5 Organization of Thesis

The complete thesis is organized in the form of different chapters. Chapter 1 describe about Knowledge Discovery from Databases and Keyword Searching Mechanism and keyword counting algorithm. Chapter 2 describes Literature Review, problem Formulation and Objectives of Thesis. Chapter 3 defines Methodology and System model of the work. Chapter 4 deals with implementation and evaluates the results of our work. Chapter 5 deals with Conclusion and Future work.

2.6 Timelines

S. No	ACTIVITIES	DURATION IN MONTHS
1	Dataset creation	
2	Literature Review	
3	Building of model	
4	Coding	
5	Implementation	
6	evaluation	



7	Writing the research paper and communicate the result		

Table 1. Timelines for Thesis

Chapter 3

Research Methodology and System model

3.1 Overview of our Framework

To fulfill the goal of our work we build a framework for this that describes the overall progress of our thesis work.

- First we define the problem.
- Second create Datasets required for our work.
- Third do literature survey. It is a continuous process.
- Four designs a model that provide interface for keyword searching.
- Five do experiment on build model.
- Six evaluate the results and provide some knowledge after successful evaluation.

This figure describes the flow of overall progress of our work.

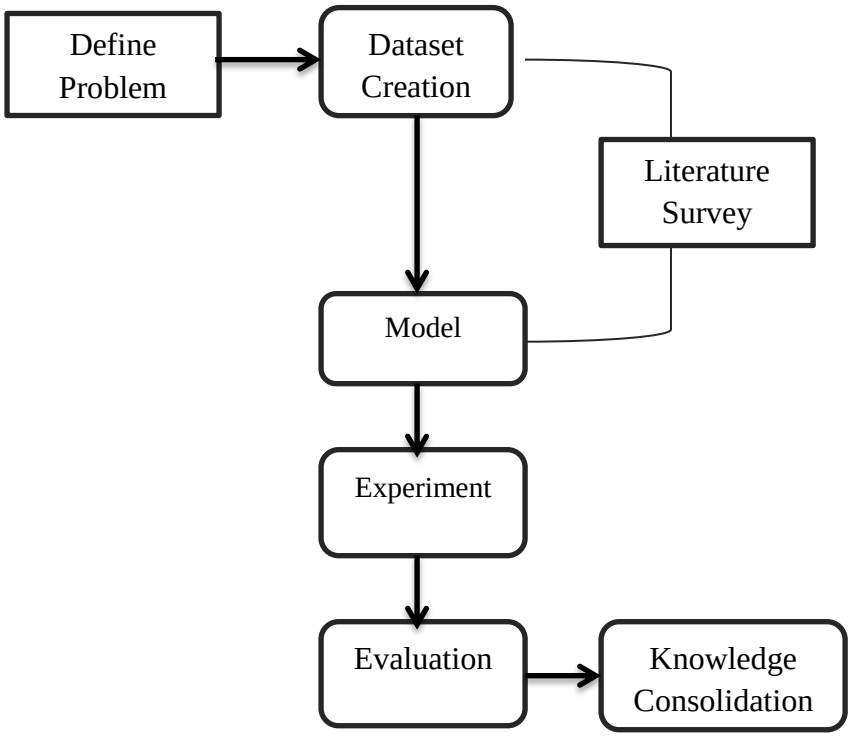


Figure 2. The overall Research Progress

3.2 Methodology

Step 1: In first step we create a pool of Dump files which contain the keywords data of various disciplines in computer science and make different classes according to their disciplines.

Step 2: Build a framework for searching keywords for their respective disciplines.

Step 3: Apply MapReduce algorithms with some modifications which counts the keywords frequency appeared in different files and sort the keyword by its high frequency then match that searched keyword which has high frequency from database.

Step 4: When Keyword matches from the database then specify the thrust area based on that keyword.

Step 5: Calculate the time taken by each input source and define the no of clusters made by that keyword.

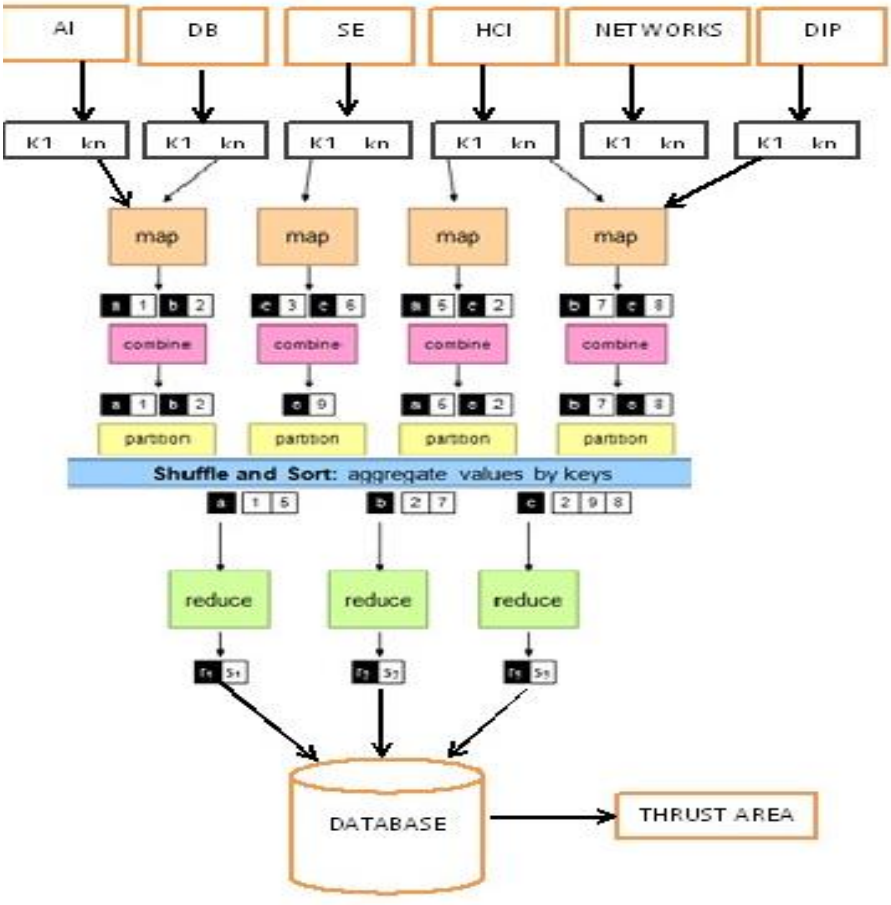
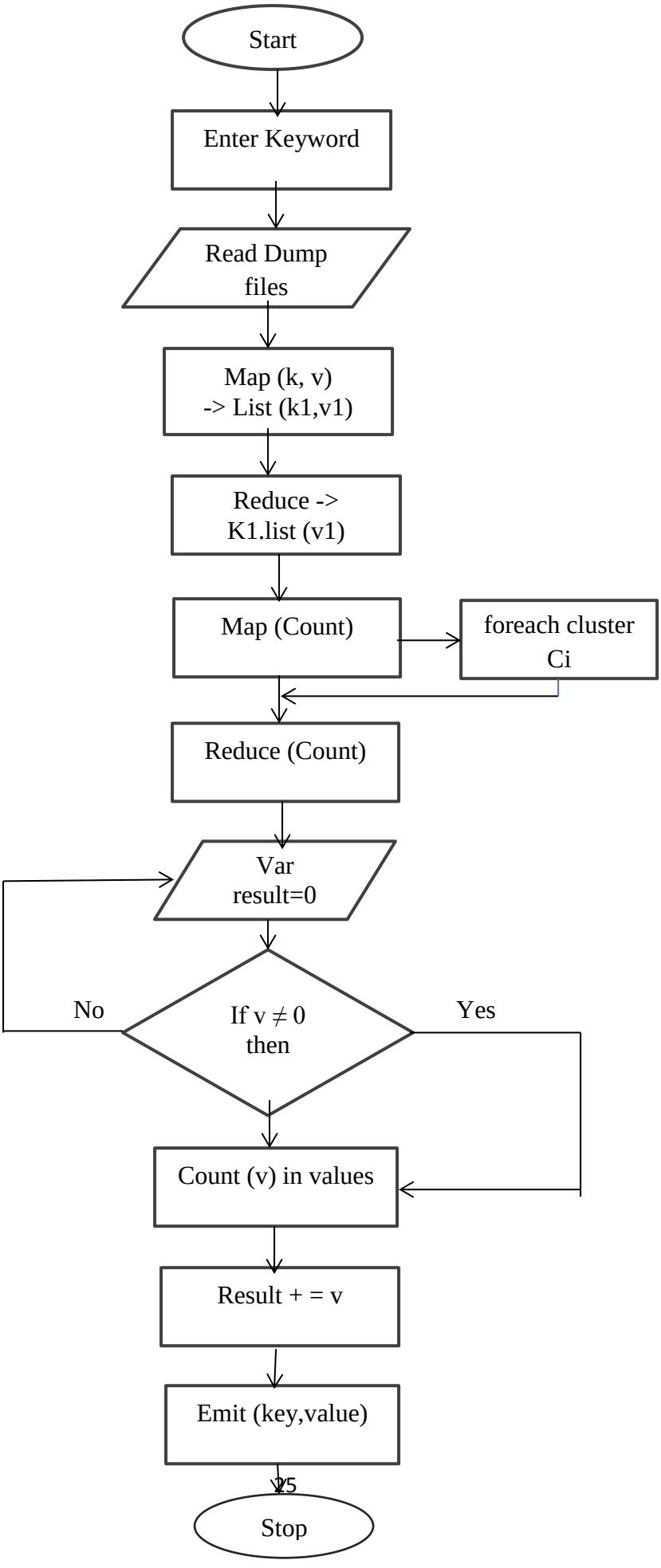


Figure-3: Overview of Model

In our modified algorithm sorting is not w.e.f keywords as available in MapReduce algorithm instead do sort using count value so that frequency of keywords can be found for further Thrust Area.

3.3 Flow Chart



3.4 Pseudocode

Input: a set of key/value pairs

User supplies two functions:

map(k,v) -> list(k1,v1)

reduce(k1, list(v1)) -> v2

(k1,v1) is an intermediate key/value pair

Output is the set of (k1,v2) pairs

Mapper (Keyword Count)

//input clusters stored in the form of text file

For each cluster Ci

map(key, value):

// key: document name; value: text of document

For each word w in value:

emit(w, 1)

endfor

endfor

Reducer (Keyword Count)

reduce(key, values):

// key: a word; values: an iterator over counts

result = 0

for each count v in values:

result += v

emit(key,result)

end for

This algorithm finds the thrust area for the highest frequency keyword searched from the input clusters connect database and search thrust area for the keyword and corresponding area name.

Chapter 4

Result and Discussion

In this chapter we show the results of our implementation according to querying keywords and also time taken by each input sources. Results show the Thrust Area of particular entered keyword by its frequency.

4.1 Outcomes of Research

1. Time taken for such data source and cumulative time to process all input data sources.
2. Cluster formulations
   n + m
   where n = No of Data sources and m = no of sub data sources.
3. Thrust Area Identification.

First we show the table that shows the time consumption in seconds on various trials from input resources taken by us.

S No.	Trial No.	Time Taken (seconds)
1.	First	28.9
2.	Second	34.1
3.	Third	17.9
4.	Fourth	24.5
5.	Fifth	30.1

Table 2: Time consumption in seconds on various trials (input resources taken 7)

Second we show the graph of time calculations per unit or dataset of input files.

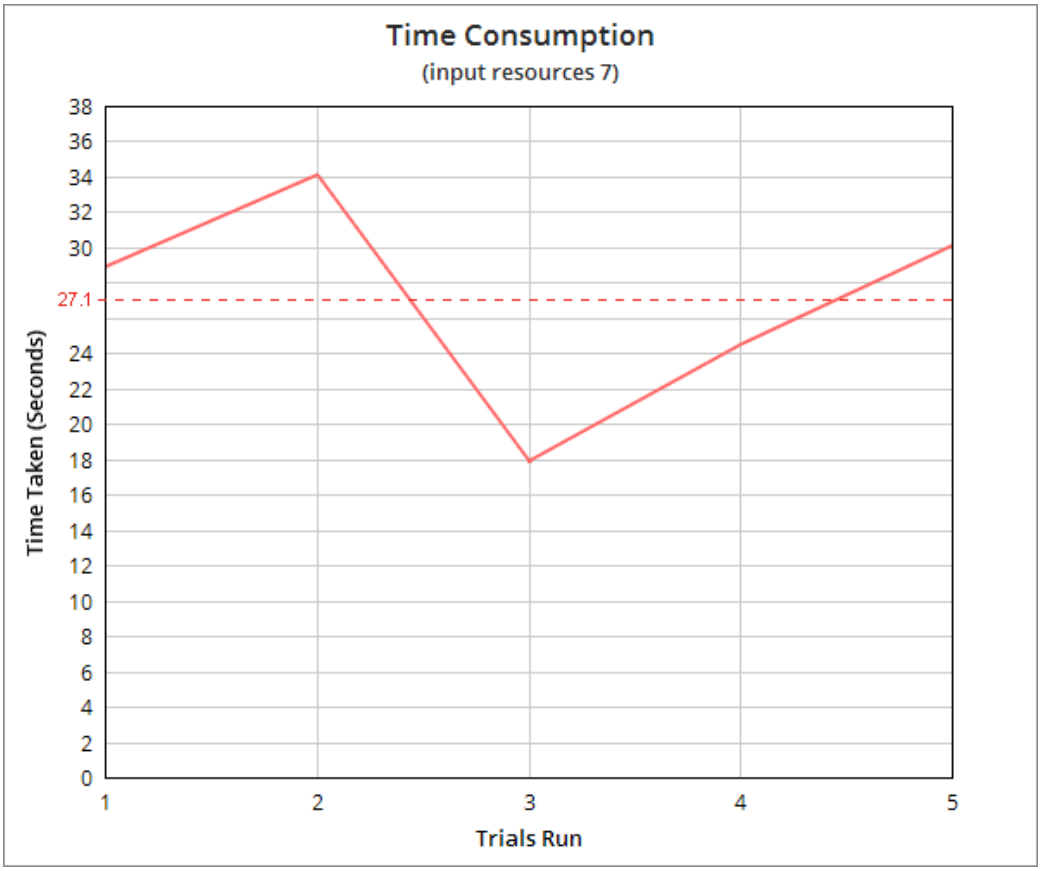


Fig 4: Time consumption with same no. of resources for different trials

After calculating the time consumption by trials then we calculate the time taken by each input resources and also describe clusters for each input resources.

S No.	Input Resources	Clusters	Time Taken (Sec)
1.	1 domain	5	0.06
2.	2 domain	23	0.72
3.	3 domain	30	2.53
4.	4 domain	47	4.57
5.	5 domain	65	7.69
6.	6 domain	80	12.65
7.	7 domain	99	17.42

Table 3: Clusters Formation and Time Consumption for increasing input resources

This graphs shows the work of cluster formulations for each domain area and calculating the time taken by each domain

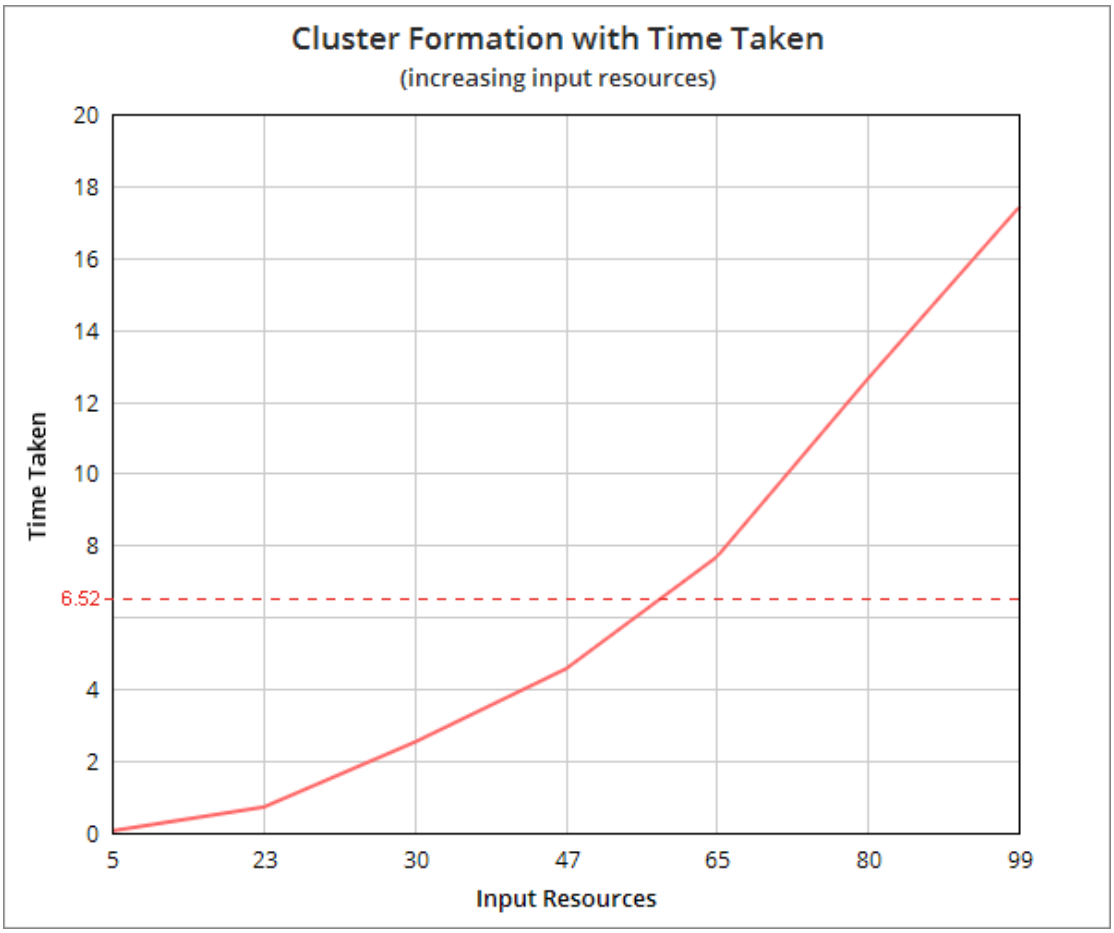


Fig 5: Cluster Formation with increasing input resources time consumption

This graph shows results on the basis of searched keyword and shows the frequency of that keyword according to all disciplines and specifies the Thrust Area for that keyword. In this we take disciplines files as input 1, input 2……input n.

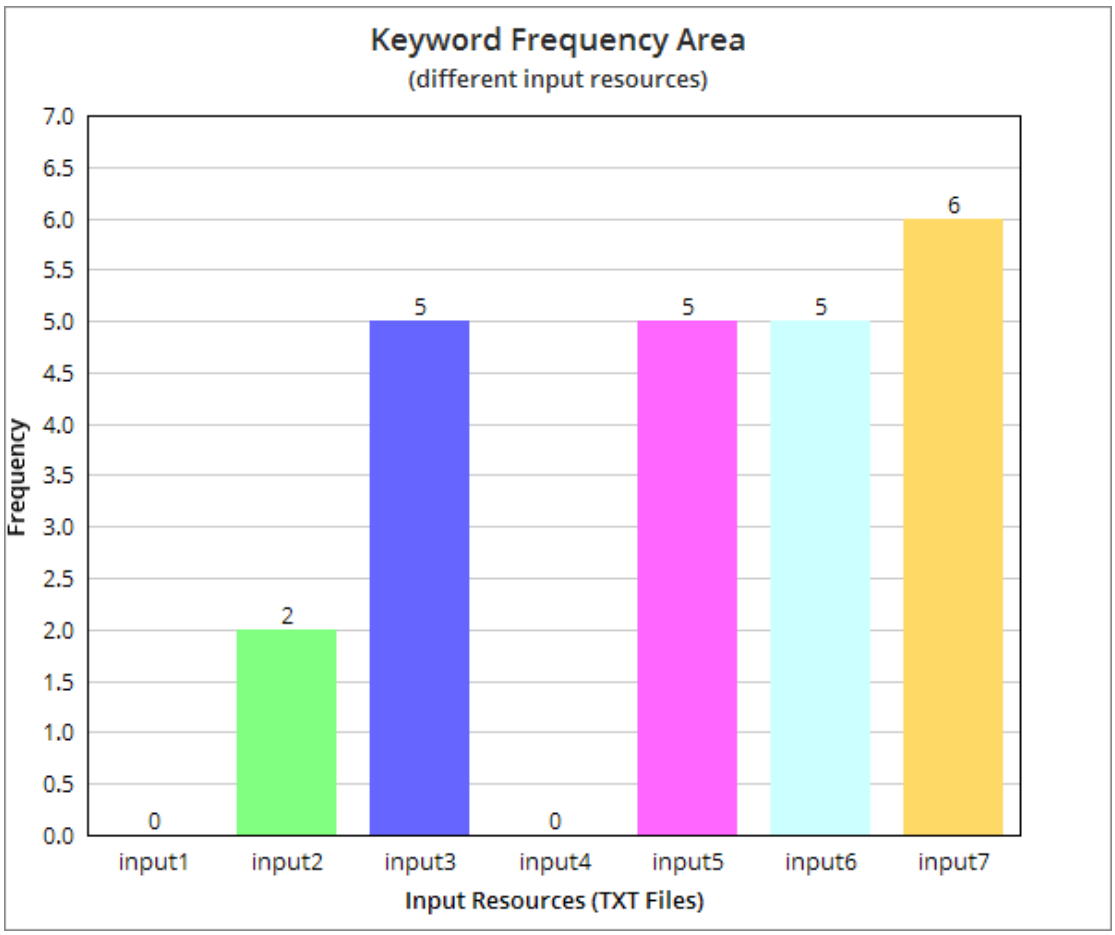


Fig 6: Keyword Frequency Measurement

This work defines knowledge about particular searched keyword and shows the relationship of that keyword to the particular area according to keyword frequency.

For building this model we used MapReduce algorithm with some modification and after apply counting approach from this algorithm. We match the keyword from our database and specify the Thrust Area in computer science disciplines based on keyword identification.

This work fulfills the goal of keyword based identification of thrust area. There are various techniques available in keyword based searching from database and we used the MapReduce algorithm which used in Google ranking and also used in Hadoop for big data. This technique is used for ranking, sorting and counting word frequencies for textual databases.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis we build a framework for keyword searching and counting their frequencies according to input datasets and make clusters based on particular keyword and if keyword found in database then specify the thrust area for that particular keyword and check the frequency of keywords in respective areas.

5.2 Future Work

In the future we will add an interface in it that will upload pdf file from user and store it in database then extract keyword from pdf file and store the keywords in database. We will expand this work with other respective domains such as Medical domain, other engineering domain etc.

References

[1] K.J. Cios, W. Pedrycz, R.W. Swiniarski, L. Kurgan, “Data mining: A Knowledge Discovery Approach”, http://w.w.w.springer.com, 2007.

[2] Bharti M. Ramagari, “Data Mining Techniques and Application”, Indian journal of computer science and engineering, 2011.

[3] Joyce Jackson, “Data Mining: A Conceptual Overview”, communications for the association for information systems”, 2002.

[4] Ronen Feldman and Ido Dagan, “Knowledge Discovery in Textual Databases (KDT)”, proceedings of KDD-95, 1995.

[5] Ken Mcgarry, “A Survey of interestingness Measures for Knowledge Discovery”, The knowledge engineering review, Cambridge University Press, 2005.

[6] Oded Maimon and Liar Rokach, “Introduction to Knowledge Discovery in Databases”, from Data Mining and Knoledge Discovery Handbook.

[7] C. Priyadharsini and Dr. Antony, “An overview of Knowledge Discovery Databases and Data Mining Techniques”, International Journal of Innovative Resesarch in computer and communication engineering, 2012.

[8] Octavian Rusu, Ionela Halcu et al., “Converting Unstructured and Semi-Structured data into knowledge”,

[9] Pavel Shivaiko and Jeroma Euzenat, “A survey of Schema-based Matching approaches”, Journal on Data Semantics, Springer, 2005.

[10] S. Agrawal, S. Chaudhuri, and G. Das, “DBXplorer: A system for keyword-based search over relational databases”, In ICDE, 2002.

[11] A. Balmin, V. Hristidis, and Y. Papakonstantinou. “ObjectRank: Authority-based keyword search in databases”, In VLDB, 2004.

[12] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. “Keyword searching and browsing in databases using BANKS”, In ICDE, 2002.

[13] Bei Yu, Guoiang Li and Karen Sollins,"Effective Keyword based selection of Relational database", SIGMOD,2007.

[14] Pattan Kalesha, M.Babu Rao and Ch. Kavitha," Efficient preprocessing and patterns identification Approach for text Mining", International Journal of Computer Trends and Technology, 2011

[15] N.L Sarda and Ankur jain,"A system of Keyword based and searchingin databases",

[16] Surajit chandhari and Gautam Das," Keyword Querying and Ranking in Databases", VLDB endowment.