CHAPTER - 3
LITERATURE REVIEW

3.1 Introduction ........................................................................................................... 22

3.2 Need and Scope .................................................................................................... 22
   3.2.1 Development, Constructions and Structure of Thesaurus .......................23
   3.2.2 Automated, Networked and Relational Database Thesauri .......... 26
   3.2.3 Use of SKOS and other Standards in Creation of Automated
       Thesauri e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e e
CHAPTER – 3

LITERATURE REVIEW

3.1 Introduction

Literature review is an important pre-research practice followed before starting actual research work. It is a process of research, which is necessary to get relevant information on the research area/topic. It highlights on the scattered literature published in the past and co-relates with the current research.

Literature review is the first and foremost step for any research. An effort has been made in this chapter to review an important works undertaken in the past on the development of thesauri, use of relational database management system to automate thesauri, standards for creation of automated thesauri in India and abroad.

3.2 Need and Scope

A literature review of thesaurus construction and automated thesaurus provides valuable knowledge for the understanding of the theoretical and practical base of research area. It provides fare idea for construction of thesaurus using standard methodology. Literature on various aspects of thesaurus development has been searched using various indexing, abstracting tools and full text databases, such as Library and Information Science Technology Abstract (LISA) and Library Information Science & Technology Abstracts (LISTA). Numerous researches have been done all over the world on the development of thesaurus, computerized thesaurus and multilingual thesaurus, some of the automated thesauri are also
available online such as ERIC thesaurus, AGROVOC, UK archival thesaurus etc. The survey includes the literature published during 1970 to 2015. Some of the important websites which are providing authentic information on the related subject has also been covered in the literature search.

During literature search process, it has been observed that the literature related to the development of thesaurus using relational database management system and Simple Knowledge Organization System (SKOS) is published in various sources of information and some of the open source software is also developed by freelancers and organizations to automate thesaurus. For easy understanding of the research area, literature has been reviewed under following five major categories for the development of thesaurus using relational database management system.

1. Development, construction and structure of thesaurus
2. Automated, networked and relational database thesauri
3. Use of SKOS and other standards in creation of automated thesauri
4. Thesauri on different subjects and general literature
5. Special reference to Social Science Thesaurus

A compendious of the literature survey, noticing the prominent findings and citations to the literature have been provided at the end of the review.

3.2.1 Development, Constructions and Structure of Thesaurus

According to Kim (1973) there were no theoretical rules for thesaurus construction and maintenance. The study provides detailed information on construction and updating of thesaurus and concludes that updating thesauri is largely
done by the indexer himself at the time of indexing and no attempts were made to develop a theory for thesaurus construction and updating.

Wall (1975) observed that retrieval of documents for financial information would have not been possible without thesaurus. The study listed out six major applications for which thesaurus is necessary and also discussed about DISCLOSURE thesaurus which is especially dedicated for financial information.

Wang & Vandendorpe (1985) discussed about development of a relational thesauri based on lexical semantic relations. They have constructed thesauri based on term classification and found that its maintenance is relatively simple. They have also done some experiments to evaluate thesauri using information retrieval system (IRS) at Illinois Institute of Technology. The analysis revealed that the individual relations are more useful in an interactive retrieval system which displays indexing terms and allows users to choose most relevant ones.

Harter & Cheng (1996) introduced a new concept and technique for information retrieval called co-linked descriptor. The idea is basically derived from co-cited references, which provide theory and method to identify search terms. The result of the study shows that the approach is effective, although methodological problem in idea is reported. However, colinked descriptor is a strong argument for building richer and more complex thesauri and can improve the vocabulary selection as it is done by the end user.

Construction of a thesaurus is a time consuming task; one has to keep all the aspects in mind and follow the standard rules at the time of compilation. Miller (1997) has pointed out general theoretical and practical problems of thesaurus.
construction. Aitchison et al. (2000) have provided detailed information on construction of thesauri. They have also discussed about the web based thesauri and its implications in organizing documents on the network. The concept of thesauri, design and its compilation has also been discussed.

Construction of a thesaurus through knowledge representation using semantic web technology has now become much popular. According to Bechhofer & Goble (2001) semantic metadata to describe subject content of the document, plays vital role in indexing and retrieval. Construction of a large thesaurus is difficult but, it can be facilitated through knowledge representation technique. They have described a scheme called Description Logic (DL) to construct thesaurus relationships for semantic web.

Thesauri, taxonomies and ontologies are frequently used terms by, among information scientists and AI practitioners. Gilchrist (2003) has explained these three terms in detail with examples. An overview on designing and construction of thesaurus has been provided by McCulloch (2005). The essay provides some basic guidelines for construction of thesaurus.

Broughton (2006) has provided general as well as technical information on thesaurus construction. She has discussed about nature, functions, features and types of thesauri used for indexing and retrieval for online information sources. Further, she has thoroughly described relationships used in the thesauri. Based on the retrospective measurement and estimates Losee (2007) described a decision method for term selection and subdivision of term. He has also explained how these decision-theoretic techniques may be used in construction of thesaurus.
Hedden (2008a) described meaning of controlled vocabularies, thesauri and taxonomies. With rising usage of automated thesauri in indexing and retrieval of information, whole concept of thesauri has been changed. Hedden (2008b) studied three software programme namely Multi Tes, Tem Tree 2000 and WebChoir TCS-10 which, can be used for the creation and editing thesauri. The study reveals that, there is no best or worst thesaurus software out of these three and each one of them fulfills the basic requirements.

3.2.2 Automated, Networked and Relational Database Thesauri

According to Jones (1970) retrieval of keywords from the classification system would be the most preferred way to create automated thesaurus. It has been also observed that human generated thesaurus is more relevant compare to system generated thesaurus. Further, study reveals that the automatic word classification is based on the distributional behavior and occurrence of the word in the document.

Rolling (1970) pointed out that there are three types of thesauri, based on the terms covered in the thesauri, namely uniterm, uniconcepts and subject headings. Rolling considered three English language thesauri compiled by ASM, CSM and ISCOR; a French thesaurus developed by CNRS; a number of encyclopedias; the Metals handbook and the subject indexes of major abstracting journals and started compiling process using computer. The result shows, out of total 10,000 terms originating from three thesauri, 1100 terms were duplicate and 900 terms were triplicate which has been later deleted and alphabetically merged, leaving a collection of 8000 terms.
Rada & Bicknell (1989) have conducted an exploratory study to evaluate and improve thesaurus for effective retrieval. They have developed a formula called DISTANCE and applied it to the documents indexed using MeSH (Medical Subject Heading) for assessment and conceptual distance between query and documents. The study reveals that power of computer simulation stems from the tendency of people to rely heavily on broader term relations in making conceptual distance from MeSH. It has been also observed that adding non-BT relations to MeSH could improve document ranking and the DISTANCE ware also revised accordingly to treat non-BT relations differently.

According to Crouch (1990) automatic thesaurus generation in particular has proven to be difficult task. The study examines both early and current approach for automatic thesaurus construction to generate automated global thesauri using term discrimination value model of Salton, Yang and Yu on clustering algorithm. The results of the study indicate 10 to 15 percent improvements in retrieval performances using said model in the test collection.

Milstead (1991) has suggested specifications of software for thesaurus maintenance and information retrieval, which are intended to support both existing packages and also for design of custom software. She has also suggested that the software should be developed based on ANSI standards for relationship to avoid conflict between terms further, wide variety of online displays and printed reports are required but not limited to alphabetical and hierarchical has also been suggested.

Pollard (1993) has proposed a hypertext based thesaurus to use as navigational aid for particular subject domain. He has designed experimental hypertext based
browsing interface and implemented on Thesaurus of ERIC descriptors. Strategy for linking thesaurus database and relationships has also been discussed.

An electronic community system encodes research community’s information, knowledge and provides an online support for their research in knowledge manipulation. According to Chen et al. (1995) an automatic algorithmic approach required to generate automatic thesauri for electronic community system. They have considered Worm Community system which contains comprehensive data and literature on molecular biology. The study revealed that thesaurus was an excellent “memory-jogging” devise for learning as well as browsing and it can become a useful tool for researchers in exploring complex electronic community system.

Jones et al. (1995) discusses about building intelligence rules for interactive thesaurus navigation. They have investigated weight based algorithms in general purpose software for interactive thesaurus navigation and found some evidence of patterns which could be used to justify strong rules, indeed the basic assumption that thesaurus based query expansion will improve retrieval performance has come under question.

Thesaurus is one of the most important tools for indexing documents and it is also important for effective retrieval of documents. According to López-huertas (1997) there is a need for thesauri to help users in searching their information from online information system. In view of this, it has been felt necessary to create structure of interactive thesaurus which can help users in effective information retrieval. The study reveals how the author’s writings can be used to identify subject domain and later it can be inserted to thesaurus structure.
Network based thesauri have become important tool for indexers as it can be accessed without geographical limitation. Kramer, Nikolai & Habeck (1997) discusses about the integration of various thesauri for document retrieval in network based environment in Internet age. They have provided an approach to integrate multiple thesaurus databases using CORBA (Common Object Request Broker Architecture) and Java technologies. Further, it has been also demonstrated to integrate distributed and heterogeneous thesaurus databases and integration of multilingual and monolingual thesauri using CORBA.

Semantic Index System (SIS) is a product of the Institute of Computer Science – FORTH and it is an object oriented semantic network of databases used to store and maintain knowledge representation applications. Doerr & Fundulaki (1998) have presented a method called SIS-TMS for thesaurus management, which is specially designed to implement necessary data structures and management functions. The system allows integrating different databases on a central system which provides consistency and easy access on the network.

According to Plaunt & Norgard (1998) there are two stage algorithms, lexical relations in a document representation and subjects assigned from controlled vocabulary. They have examined INSPEC collection of 4626 documents and created dictionary of associations between lexical terms and controlled vocabulary (subject headings assigned to the documents by human). The study unveils that the algorithm which have been used to compare terms between automatic and human generated subject headings is consistent.
Shiri & Revie (2000) have discussed on the development of web based thesauri. They have provided detailed information about some thesaurus projects undertaken to facilitate information resources on the Internet. Different types of thesauri, integrated thesauri and multiple thesaurus system have also been considered to generate networked thesauri.

Laurin system is a software package for digital clipping archives at University of Innsbruck, Austria. To index newspapers clippings, the Department of German Language of the university had developed Laurin thesaurus. Retti & Stehno (2004) have provided brief information about the whole process of thesaurus construction. They have also discussed how the relationships between subject keywords have been knitted with SQL statements for easy and effective retrieval of the keywords. A study on integrated thesaurus and relationship with searching and browsing of online collection of photographs has been undertaken by the Dalmau et al. (2005). They had developed a database using Oracle Programming Language and SQL as well as web interface to fetch appropriate keywords form the thesaurus.

Generation of bilingual automatic thesauri has been described by Tsuji & Kageura (2006). They have proposed a method for automatic generation of Japanese-English bilingual thesauri based on bilingual corpora. Further, they have provided justification for using bilingual corpora in generation of automatic thesauri. Automatic thesaurus development using term extraction from title metadata has been described by Wang (2006). It has been observed that, in networked environment lots of new concepts are emerging and, it is very critical to enhance retrieval of formal thesauri frequently, therefore, Wang has highlighted the relevancy of term extraction from title metadata.
Shiri & Revie (2006) investigated query expansion behavior of end user who is interacting with thesaurus on the web. They have gathered data from 90 searches performed by 30 users using questionnaire, screen capturing and interview technique. The study reveals that in complex topics, interaction with thesaurus required more query expansion. It has been also observed that during search process, 50% additional search terms were suggested from the thesaurus.

Application of thesauri in networked environment is seriously challenging task for enhancement and retrieval. Wang (2006) described enhancement of networked thesauri using term extraction from title metadata. The term extraction process has been conducted in three steps using which automated thesaurus can be developed. Araujo & Pérez-Agüera (2006) proposed pattern matching method to identify relationships between terms for automatic thesauri generation. They have used two different kinds of patterns namely structural patterns and key patterns for experiments. The experiment revealed a clear improvement in the performance.

Usually, a web is referred as a repository of knowledge scattered in different literature, language and culture. To manage knowledge, retrieved from different languages and from different geographic locations is an important chore. Yang, Wei, & Li (2008) discussed about cross-lingual semantic interoperability in multilingual knowledge domain. To solve the problem of cross-lingual semantic interoperability they have constructed an automated thesaurus using benchmarking and forward algorithm for evaluation. The investigation shows that the benchmark algorithm is more efficient than the forward algorithm.
Mayr & Petras (2008) have provided systematic explanation of terminology mapping of various thesauri and its relationship with other controlled vocabularies and created cross-concordances. The study reveals that the 64 crosswalks and 5 lakh relations were established successfully.

According to Lacasta et al. (2010) maintenance of different types of terminological ontologies requires great amount of time and infrastructure. They have analyzed the need for providing single infrastructure to all terminological ontologies under one roof. They have also proposed architecture of Ontology Manager to manage all the terminological ontologies for easy access to all ontologies.

PoolParty is a thesaurus management tool available in the market at cost. The software allows users to manage their thesaurus in a systematic way which can be easily retrieved whenever needed. Schandl & Blumauer (2010) have discussed about PoolParty thesaurus management software/tool to manage linked data on the semantic web using SKOS framework. They have also provided brief information on technology used for the development of PoolParty and its capabilities to handle linked open data.

Use of uncontrolled vocabulary for indexing and organization knowledge is simple and cost effective compare to controlled vocabulary but, when it comes to retrieval the recall ratio might be reduced. Nagaya et al. (2011) have introduced a JavaScript library called covo.js which enable users to make a controlled vocabulary as metadata. They have written a programme using Java Script but, it does not include hierarchical relations of controlled vocabulary as of now.
Shiri et al. (2011) conducted a comparative user evaluation of two multilingual thesaurus-enhanced visual interface namely T-saurus and Searchling developed for digital libraries. They have studied 25 academic users, UNESCO’s digital portal using questionnaire, interview and observation to investigate user’s evaluation for multilingual digital libraries. The study unveils that users were able to carry out search task using enhanced thesaurus search interface. It has been also observed that the Searchling thesaurus is more flexible and user friendly.

3.2.3 Use of SKOS and other Standards in Creation of Thesauri

Simple Knowledge Organization system (SKOS) is one of the standards developed by WWW (World Wide Web) for semantic web. It provides guidelines to encode thesaurus in RDF format (Resource Description and Framework), which can be shared across the network. Greenberg (2001) discusses about query expansion method for semantically encoded thesauri and explored that, the thesauri becomes more useful for online indexing and information retrieval. The study reveals that the query expansion processing methods for semantically encoded thesauri are effective in automatic or interactive environment.

Referring traditional thesaurus is little time consuming, due to which automated thesauri have become more popular. SKOS provides RDF framework using the said framework one can create automated thesaurus. Assem et al. (2006) have provided detailed information on conversion of manual thesauri into SKOS format. They have done case study on Integrated Public Sector Vocabulary (IPSV) thesaurus, GTAA thesaurus and MeSH thesaurus and described conversion process of thesauri to SKOS. Cantara (2006) discussed that the controlled vocabularies do
not guarantee flawless searching across multiple collections but it can be more effective if, it is encoded using SKOS over semantic web.

Rowley (2007) elaborated about the British standard BS 8723 which have been developed to standardize structured vocabularies and thesauri for information retrieval. The review provides brief information on the standard BS 8723 for creation of thesauri and interoperability between thesauri. ISO 2788 and ISO 5964 are the international standards for monolingual and multilingual thesauri respectively. These two standards are identical to British standard BS 5723 and BS 6723. Clarke (2008) mentions that the evolution of ISO 25964 is based on the review of ISO 2877 and ISO 5964 during 2007.

Pastor et al. (2009) have provided the basic structure and relationship for various thesauri to convert it in SKOS. The detailed information on semantic relations and labels used for mapping relationship has also been discussed.

According to Weinberg (2009) an accurate and exhaustive guidelines needed for controlled vocabulary. It has been suggested that the NISO has to make some changes in existing guidelines developed in 2005 for construction of monolingual controlled vocabularies as it have inconsistencies between definition and text. With an advent of WWW, Clarke (2010) have also observed that it is very much necessary to revise existing standards and introduce new one for controlled vocabulary.

As SKOS provides platform to host data over semantic web and allows interchange of data between thesauri. Fugazza et al. (2010) have discussed importance of SKOS for developing thesauri for Spatial Data Infrastructures and presented as a prototype application for the said domain. They have also provided
solution to approach the semantic heterogeneity problem in the field of geographic information among distributed spatial data infrastructures.

Zeng & Zumer (2010) have discussed to map Functional Requirement for Subject Authority Data (FRSAD) with SKOS to interchange data between different formats. They have provided information for mapping FRSAD data into SKOS for interoperability in different formats such as OWL, DCMI etc.

Kubik (2011) has discussed about the thesauri available on geospatial domain and its role in the information management and retrieval. It has been also observed that the implementation of widely acceptable standards (SKOS) for sharing and exchanging data will provide great scope to access thesaurus in geospatial domain.

Resource Description Framework is a framework using which one can host relational data on the semantic web for the purpose of interchange and easy access. Yu (2011) provided detailed information on structure of RDF with simple example. He has also differentiated thesauri and ontology and explained how it can be converted to SKOS using RDF.

Will (2012) discusses about ISO 25964 thesauri model for information retrieval which is now, more extensive and intended to provide rigorous presentation of the entities and relationship to put thesaurus data on the semantic web using SKOS.

Lacasta et al. (2013) described that the knowledge is used in context of digital libraries to improve data sharing and information retrieval. They believe, to increase use of the data, it needs to be converted to a standard interchange format. They have observed that, SKOS is the best option for creating data in sharable form. They
have chosen ThManager open source software to create data in SKOS interchange format and shown the process of converting thesauri.

3.2.4 Thesauri on Different Subjects and General Literature

According to Eichhorn & Reinecke (1970) a thesaurus for visual sciences has been developed to serve as the integrating unit for computerized information storage and retrieval system for Vision Information Centre. They have also described that the thesaurus plays vital role in indexing and it uses numerical coding system which is necessary for in-depth indexing and it is highly relevant to the search requests in the field.

According to Rorvig (1999) NASA (National Aeronautics and Space Administration) has a huge collection of images and it is maintained by Photography and Television and Technology Division at the Johnson Space Center in Huston. It has been felt by the indexer that the traditional thesaurus would be difficult for easy and effective retrieval and therefore they have designed “The Visual Thesaurus” for indexing images of NASA. They have also discussed technical and practical problems occurred during the construction process.

A thesaurus in the area of management has been constructed by Paricharak (2004). It covers subject headings of all the aspects management. It was constructed as a Ph.D research work and later on published in the form book.

Stroud (2008) discussed about AIATSIS an international thesaurus which has been approved by Library of Congress for indexing documents from January 2008. The thesauri originally developed by Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) and it covers subject headings for language
groups and people, subjects relating to Australian Aboriginal and Torres Strait Islander studies.

Broughton (2010) describes the methodology of constructing a thesaurus for law literature from a faceted classification (Bliss Bibliographic Classification 2nd ed.). The relationships between terms and process for deriving descriptor and non-descriptor from faceted classification have also been demonstrated.

Use of controlled vocabulary and thesauri in UK online finding aids surveyed by Fenton (2010). The study reveals that the number of finding aids in US is growing fast and number of free text terms used in finding aids is increasing, therefore, consistent indexing and retrieval by subject becomes extremely unmanageable and time consuming.

According to Marley (2010) indexing team of the Parliament of UK has developed a thesauri to index parliamentary debates and questions, legislation, Parliamentary papers etc., to maintain uniformity in indexing process. It has been also described that House of Commons Library’s books and journals as well as web content is also indexed using Parliamentary thesauri.

Saarti & Hypén (2010) discusses the challenges for indexing and retrieval of fictions. They have also discussed about the need for development of web based tool or thesaurus for indexing fiction content and described how Kaunokki the Finnish fiction thesaurus has been created in 1996 and later its development into Kirjasampo SAHS web service to resolve the problem.

AGROVOC is a multilingual thesaurus developed by Food and Agricultural Organization (FAO) of the United Nations. It covers around 40000 terms and
available in 19 languages. Caracciolo et al. (2011) discusses about the advantages to convert AGROVOC data onto the widely used and implemented SKOS standard. They have also described process for publishing AGROVOC as linked data for semantic web which can allow easy maintenance and interoperability. The AGROVOC thesaurus is widely used by the various institutions. Fumani (2011) explored AGROVOC thesaurus in Persian language as indexer and observed that there are some problem exist in English version and in the Persian translation.

Maghsoodi & Homayounpour (2011) have proposed a SVM (Support Vector Machine) based approach for automatic classification of Persian text. The proposed approach aims to enhance accuracy of classification by using Farhang-e Teyfi thesaurus. The study reveals that the SVM based approach improves performance of classification.

Bartol (2012) assessed the utility of non-agriculture-specific information systems databases and thesauri in organizing and retrieval of agricultural information. The study reveals that some of the literature published in agricultural domain with specific aspect may not be covered in the specific databases of agriculture as it is not indexed by major databases and falls in some other area (i.e. Agricultural economics related information). It has been also observed that all the thesauri containing agriculture based subject headings relationships is different and when it is searched using subject heading, abstract, title the recall ratio improves by up to 60 per cent. Rajbhandari & Keizer (2012) described an evolution of AGROVOC thesaurus to AGROVOC concept scheme based on OWL (web ontology language) and SKOS model. They have explained key features of both the models and system architecture of the thesaurus.
Sharma (2015) has given brief information about controlled vocabulary followed by the process of constructing thesaurus in the area of nanotechnology. He has elaborated the steps of deriving descriptors and non-descriptors. He has also observed that any thesaurus needs constant updation and the thesaurus developed for nanotechnology will be online so that changes could be made easily and updated version could be accessed to all users.

3.2.5 Special Reference to Social Science Thesaurus

One of the most important thesaurus related to education in early 70's have been developed by Education Resources Information Centre (ERIC). Wright (1973) described the problem of indexing and thesaurus structure which have been developed by ERIC for indexing documents. Earlier, Case Western Reserve University has developed information retrieval thesaurus for education and it was arranged in the faceted manner, but never used by ERIC system. Instead ERIC developed its own thesaurus using ERIC descriptor and it had been evaluated by subject experts. After this process, rules for thesaurus construction and maintenance were developed.

Thesaurus is a bunch of terms and phrases with possible representation and the relationship between them. A structured thesaurus in the field of social sciences was developed by UNESCO in 1977. According to Aitchison (1977) the thesaurus has been bifurcated in two parts, classified thesaurus and alphabetical thesaurus respectively. The thesaurus contains descriptor and non-descriptors along with relationships and necessary explanation.

Kim (1982) had carried out a statistical investigation on the retrieval language of social sciences and natural sciences. The study examined the occurrence of the
words in the natural sciences and social sciences. As per the statistical investigation, in the area of social sciences, frequently occurring descriptors are indeed semantically ambiguous whereas, no theoretical justification was found in case of natural sciences. Therefore, the statistical investigation shows that the updating practice will not keep up-to-date descriptors in the area of natural sciences.

Over a period of time there was significant development noticed in the area of social sciences. The concept of the social science has grown and social scientists have also expressed their conceptual base in a particular discipline and therefore, compilers have developed subject headings and thesauri. Roberts (1985) has examined 18 subject headings and 20 thesauri in the area of social sciences and concluded that the distinctiveness of social sciences tools is little different from science and technologies.

Di Lauro et al. (1998) have developed a thesaurus for indexing information emanating in the field of economic and social development. They have structured thesaurus as per ISO Guidelines of establishment and development of monolingual thesauri. The thesaurus contains 5178 descriptors and 2057 nonDescriptors.

Thesaurus is useful tool for indexers, but contemporary thesaurus of search terms and synonyms is designed as a tool for searcher; suggesting alternative terms for searching is discussed in the review undertaken by Smith (2000).

A bilingual Russian – English Socio-Political thesaurus has been developed in 1994. It covers terminology of different sub domains of social sciences with around 85 thousand English terms. Ageev et al. (2006) has discussed about the structure and concept of the Russian – English bilingual thesaurus as well as its role in the indexing and retrieval of social science literature.
Pinto (2008) surveyed a Spanish professional user group to identify quality of thesauri used in social science databases. She has applied factor analysis methodology to the data gathered from the opinion on restricted sample of social science databases. The study reveals that the social science database thesauri require considerable improvement in the conceptual framework.

African Studies thesaurus, developed by African Studies Centre, Leiden is one of the thesauri available for indexing document in social sciences. Doorn & Polman (2010) provided detailed information on its evolution and the process of deriving descriptor and non-descriptor as well as its role in document retrieval.

Legal advice and Metadata Content Classification Scheme is devised to index contents of website generating information in the field of Law. Jones (2010) described that the special classification scheme and thesaurus developed for indexing and retrieval of legal advice on free websites and also explained how it is implemented and maintained.

3.3 Critical Appraisal of Literature Review

The review of literature attempted on various aspects of thesaurus construction, automated thesaurus, use of SKOS in generation of automated thesaurus as well as use of relational database with special reference to social sciences. The study of literature reveals that creating subjective thesaurus is not a new phenomenon, it exist since evolution of classification system.

The majority of the studies on construction of thesaurus are based on the experiments and experiences. However, some of the study provides practical guidelines on construction and updating. Before invention of the Information and Communication Technology (ICT), it was very difficult to index and retrieve
documents in specific time as it was done manually. Now, the time has changed and lots of applications have been evolved as part of World Wide Web (WWW). One of the most important applications of ICT is automatic indexing using term extraction from title and abstract.

Review of literature shows that considerable studies and guidelines have been published on Simple Knowledge Organization System (SKOS) which enables creator to host and share their data over semantic web. The study of literature on SKOS revealed that, an attempt was made to convert major thesauri on SKOS such as GESIS thesaurus, AGROVOC etc. It has been also observed that some of the studies explained mapping of thesaurus terms and its relationships in SKOS. A few studies discussed use of ISO & BS standards in construction of thesauri.

The critical analysis of the literature unveils that, UNESCO, GESIS and OECD have developed a thesaurus on social sciences. Out of these three thesauri, only GESIS thesaurus has been constructed on relational database management system. Moreover, GESIS has used proper nouns as descriptors due to which it has become bulky. It has also been observed that the coverage in GESIS is more country specific where it has been developed. Once again it is observed that all these three thesauri are not suitable to index Indian social science literature. Therefore, an exhaustive thesaurus to organize knowledge emanating from Indian social sciences is very much necessary.

References


(47)


