Chapter Two

"The underdeveloped nations of the present are those who came late in industrial revolution, and the underdeveloped nations of the future will be those who have already come late in information revolution" (Escorsia, G)

LITERATURE REVIEW

The ancient civilization was sustained on agro-based society, where plough and soil were the main driving forces. The industrial revolution formed an industry-based society, where machinery was the main driving force behind all planning and implementations. The onward trend of industry-base remained persistent upto the end of First World War, and gradually changed orientations towards information-base since the beginning of Second World War, which is an epoch-making landmark in the records of historical events of the last century. The data and information were the main driving force of the information-based society. Another new era prevailed over the information-based society to take a new turn towards the post-modern society since late eighties of the last century, which is utterly knowledge-based. It is knowledge, which is the main driving force of this post-modern, post-information-based society. Knowledge is the power and fuel of any sort of functional engines within the domain of current socio-economic infrastructure. Knowledge accumulates information; and information, in turn broaden the spectrum of knowledge. The outlook towards wealth and poverty in this context has also been somehow molded. The financial poverty is rather preferable to info-poverty and the info-richness is preferable to financial-richness. Information is an omnipresent resource today, but it is a resource only when processed rationally and scientifically; and organized in a logical sequence and used profitably. Unorganized information causes

information overload and documentary chaos. Unprocessed information gives rise to information loss, and unorganized information allows retrieval of irrelevant information that causes noise in an information system. Unorganized and unprocessed information just add junk to the multitudes of documents. The processing and organisation of information is a crucial function.

The information for finding the research trend of a subject is best obtainable from the journal articles of that subject. The information about the content of a journal article is reflected from the keywords of the concerned article. Keywords are most dynamic and brawny hauler of information. An analysis of information, from this standpoint effectively turns into nothing but scrutiny of keywords occurred in journal articles. The earliest use of keywords has been found in 1975 in the *Journal of Applied Behavior Analysis*, as pointed out by Hartley and Kostoff (2003). The significance of keywords has been described by them. Keywords are selective words occur in texts to throw light on the core theme discussed therein. The list of keywords inform about the principal domains of a subject. The groups of keywords form a subset to the set of all words. Models for the distribution of words in a text are used to estimate the performance of algorithms for text compression and for full text retrieval from databases. Many of these models, such as the Zipf and lognormal distributions, are used to predict the distribution of word frequencies (Carroll, 1967; Witten & Bell, 1990; Zipf, 1936, 1949). But such models ignore the effect of clustering. Clustering arises from the fact that words tend to be repeated a number of times in the same piece of text, even words that are quite rare. There are also some accurate models for studying distribution of words in text, which consider clustering effects (Witten & Bell, 1990) and (Thom & Zobel, 1992). But these models include all words occur in a text, not a limited number of selective keywords. The

thrust areas of research and core theme of the subject are not reflected by such models.

The keywords are extracted from titles and abstracts of journal articles. The relevance of titles as source of keywords is discussed by Bottle (1970), Hansen (1972), Kraft (1964), Lancaster (1972), Olive (1973) and Ruhl (1964). The relative merits of using title, subject heading and abstract as sources of keywords is discussed by Byrne (1975). The comparative layout between title keywords and subject descriptors is discussed by Voorbij (1998). The subject- descriptors comprise controlled terms which are required for subject indexing, while title keywords are available directly without any such intellectual activities. Dubois (1987) and Taylor (1995) summarized the advantages and disadvantages of both approaches. Studies on indexing show significant variation in the use of keywords selected by different indexers to represent the same topic or document (Bertrand and Cellier, 1995). Suraud et al. (1995) observed the non-existence of well-defined keywords in newly-emerging subject areas, which makes bibliographic searches difficult. Bates et al. (1993) discussed about development in the structures of thesauri and in the designs of the online information systems. Hurt (1997) emphasized on renewal and expansion of indexing and classification systems. Soergel et al. (2004) also point out that existing classification schemes and thesauri lack well-defined semantics and structural consistency. Juvan et al. (2005) executed keyword analysis to identify narrower research fields within the broader scientific field. They proposed a bibliometric methodology that was based on keyword analysis and the structuring of data into hierarchical tree system and could be used for the assessment of bibliographic databases and the identification of research trends.

Many researchers have used co-word analysis as an important method to explore the concept network in different fields, for instance, software engineering (Coulter, Monarch, & Konda, 1998), polymer chemistry

(Callon et al., 1991), scientometrics (Courtial, 1993, 1994), information retrieval (Ding, Chowdhury & Foo, 2001), neural network research (Noyons & van Raan, 1998; van Raan & Tijssen, 1993), chemical engineering (Peters & van Raan, 1993a and 1993b), biological safety (Cambrosio, Limoges, Courtial, & Laville, 1993), acidification research (Law & Whittaker, 1992), patents (Courtial, Callon, & Sigogneau, 1993), optomechanotronics (Noyons & van Raan, 1994), bioelectronics (Hinze, 1994), medicine (Rikken, Kiers, & Vos, 1995), biology (Rip & Courtial, 1984; Looze & Lemarie, 1997), condensed matter physics (Bhattacharya & Basu, 1998), and so on.

The subject area covered in this study is Fermi liquid. In all of these studies keywords are treated on an identical perspective, but not classified according to its attributes. The keywords are classified here from different dimensions of its modes of occurrences within the database. The methodology of classification involved here is multidimensional classification that describes a dynamic taxonomy (Sacco, 2006, 2000).

This study also includes investigation of transience and continuance of the keywords. The study of transience and continuance in scientific authorship was carried out by Gupta & Kumar (2001); and Price & Gursey (1976). These studies described various demographic characteristics of the scientists, i.e. the study of emergence, survival and disappearance of authors within a scientific speciality. The statistical report on emergence, survival and disappearance of keywords is presented in this study. Different mathematical models are framed to describe the trends of temporal changes take place on emergence, survival and disappearance of keywords.

Keywords tell about emerging trends and abrupt changes in the scientific literature that come across due to new discoveries and scientific breakthroughs such as the discovery of a supermassive black hole in

astrophysics. There are also other external effects which provoke scientists to study a subject matter from new perspectives. For example, the September 11, 2001, terrorist attacks have raised a variety of new issues to be addressed by researchers in national security, health care, post-traumatic stress disorder (PSTD) research and many other areas. The research trend of a subject is clearly recognizable from the set of keywords. The concept of “Research front” was originally introduced by Price (1965) to characterize the transient nature of a research field in terms of citations and number of journal articles. Price (1965) didn’t consider the occurrence of keywords in a research front. A research front has been studied in at least three forms as pointed out by Chen (2006): (a) a cluster of co-cited articles (Chen & Morris, 2003; Small & Griffith, 1974), (b) a cluster of co-cited articles and all articles that cite the cluster (Garfield, 1994; Garfield & Torpie, 1964), and (c) a cluster of articles that cite a common group of articles (Morris, Yen, Wu & Asnake, 2003).

The above literature survey indicates that the existing systems do not necessarily allow for sufficiently accurate and correct keyword cluster analysis, and that some further development and improvement in the analytical method of keyword clusters might be useful. Newly emerging scientific fields, in particular, require more accurate classification of keywords and greater inclusion of relevant keywords in bibliographic subject access tools.

In this study, we endeavor to develop a new taxonomy of keywords taking sample from a dynamic subject field (Fermi liquid). We identify and assess the existing keywords and its semantic relationship with others. We have analyzed the temporal variation trends of different keyword-clusters and developed various mathematical models to describe transience and continuance. Further, the design of information retrieval thesaurus at micro-level and development of a new subject-access tool based on statistics of the parameters of the keyword have been suggested.

References

Bates, M.J., Wilde, D.N. and Siegfried, S. 1993. An analysis of search terminology used by humanities scholars: the *Getty Online Searching Project Report*, No.1. *Library Quarterly* 63 (1): 1-39.

Bertrand, A. and Cellier, J.M. 1995. Psychological approach to indexing effects of the operator's expertise upon indexing behaviour. *Journal of Information Science* 21 (6): 459-472.

Bhattacharya, S. & Basu, R. K. 1998. Mapping a research area at the micro level using co-word analysis. *Scientometrics* 43 (3): 359-372.

Bottle, R.T. 1970. The information content of titles in engineering literature. *IEEE Transactions on engineering Writing*, EWS-13: 41-45.

Byrne, J.R. 1975. Relative Effectiveness of Titles, Abstracts, and Subject Headings for Machine retrieval from the COMPENDEX services. *Journal of the American Society for Information Science (pre-1986)* 26 (4): 223-229.

Callon, M., Courtial, J. P., & Laville, F. 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics* 22 (1): 153-205.

Cambrosio, A., Limoges, C., Courtial, J. P., & Laville, F. 1993. Historical scientometrics? Mapping over 70 years of biological safety research with co-word analysis. *Scientometrics* 27 (2): 119-143.

Carroll, J. 1967. On sampling from a lognormal model of word frequency distribution. In H. Kiicera and W. Francis (eds.), *Computational analysis of present-day American English* (pp. 406-424). Providence, RI: Brown University Press.

Chen, C. & Morris, S. 2003. Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. Proceedings of IEEE Symposium on Information Visualization". (67-74). Seattle, WA: IEEE Computer Society Press.

Chen, C. 2006. Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 57 (3): 359-377.

Coulter, N., Monarch, I., & Konda, S. 1998. Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science* 49 (13): 1206-1223.

Courtial, J. P. 1994. A co-word analysis of scientometrics. *Scientometrics* 31 (3): 251-260.

Courtial, J. P., Cahlik, T., & Callon, M. 1994. A model for social interaction between cognition and action through a keyword simulation of knowledge growth. *Scientometrics* 31 (2): 173-192.

Courtial, J. P., Callon, M. & Sigogneau, A. 1993. The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics* 26 (2): 231-242.

Ding, Y., Chowdhury, G.G. & Foo, S. 2001. Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management* 37: 817-842.

Dubois, C.P.R. 1987. Free text vs. controlled vocabulary; a reassessment. *Online Review* 11: 243-253.

Garfield, E. 1994. Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioral sciences* 7 (45): 5-10.

Garfield, E., Sher, I.H. & Torpie, R.J. 1964. *The use of citation data in writing the history of science.* Philadelphia: Institute for Scientific Information.

Gupta, B.M. & Kumar, S. 2001. Demographic profile of authors in theoretical population genetics speciality. *ILA Bulletin* 37 (4): 127-131.

Hansen, I. 1972. Evaluation of the database CA condensates compared with chemical titles. *Journal of Chemical Documentation* 12: 101-109.

Hartley, J and Kostoff, R.N. 2003. How useful are ‘keywords’ in scientific journals? *Journal of Information Science* 29 (5): 433–438.

Hinze, S. 1994. Bibliographical cartography of an emerging interdisciplinary discipline: The case of bio-electronics. *Scientometrics* 29 (3): 353-376.

Hurt, C.D. 1997. Classification and subject analysis: looking to the future at a distance. *Cataloging and Classification Quarterly* 24 (1-2): 97-112.

Juvan, S., Bartol, T. and Boh, B. 2005. Data structuring and classification in newly-emerging scientific fields. *Online Information Review* 29 (5): 483-498. (Available at: www.emeraldinsight.com/1468-4527.htm)

Kraft, D.H. 1964. Comparison of keyword-in-context indexing of titles with a subject-heading classification system. *American Documentation*, 15: 48-52.

Lancaster, F.W. 1972. *Vocabulary control for information retrieval*. Washington DC: Information Resources Press.

Law, J. & Whittaker, J. 1992. Mapping acidification research: A test of the co-word method. *Scientometrics* 23 (3): 417-461.

Looze, M. D. & Lemarie, J. 1997. Corpus relevance through co-word analysis: An application to plant proteins. *Scientometrics* 39 (3): 267-280.

Morris, S.A., Yen, G., Wu, Z. & Asnake, B. 2003. Timeline visualization of research front. *Journal of the American Society for Information Science and Technology* 55 (5): 413-422.

Noyons, E. C. M. & van Raan, A. F. J. 1994. Bibliometric cartography of scientific and technological development of an R & D field. *Scientometrics* 30 (1): 157-173.

Noyons, E. C. M., & van Raan, A. F. J. 1998. Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science* 49 (1): 68-81.

Olive, G., Terry, J.E. and Datta, S. 1973. Studies to compare retrieval using titles with that using index terms. *Journal of Documentation* 29: 169-191.

Peters, H. P. F. & van Raan, A. F. J. 1993a. Co-word based science maps of chemical engineering, Part I: Representations by direct multidimensional scaling. *Research Policy* 22: 23-45.

Peters, H. P. F. & van Raan, A. F. J. 1993b. Co-word based science maps of chemical engineering. Part II: Combined clustering and multidimensional scaling. *Research Policy* 22: 47-71.

Price, D. & Gursey, S. 1976. Studies in Scientometrics, part 1: Transients and continuance in scientific authorship. *International Forum of Information & Documentation* 1 (2): 17-24.

Price, D.D. 1965. Networks of scientific papers. *Science* 149: 510-515.

Rikken, P., Kiers, H. A. L. & Vos, R. 1995. Mapping the dynamics of adverse drug reactions in subsequent time-periods using INDSCAL. *Scientometrics* 33 (3): 367-380.

Rip, A., & Courtial, J. P. 1984. Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics* 6: 381-400.

Ruhl, M.J. 1964. Chemical documents and their titles: Human concept indexing vs. KWIC machine indexing. *American Documentation* 15: 136-141.

Sacco, G.M. 2000. Dynamic taxonomies: A model for large information bases. *IEEE Transactions on Knowledge and Data Engineering* 12 (2): 468-479.

Sacco, G.M. 2006. Dynamic taxonomies and guided searches. *Journal of the American Society for Information Science and Technology* 57 (6): 792-796.

Small, H.G. & Griffith, B.C. 1974. The structure of scientific literatures: Identifying and graphing specialities. *Science Studies*, 4: 17-40.

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J. and Katz, S. 2004. Re-engineering thesauri for new applications: the AGROVAC example. *Journal of Digital Information*, 4 (4). (Available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/)

Suraud, M.G., Quoniam, L., Rostaing, H. and Dou, H. 1995. On the significance of databases keywords for a large-scale bibliometric investigation in fundamental physics. *Scientometrics*, 33 (1): 41-63.

Taylor, A.G. 1995. On the subject of subjects. *Journal of Academic Librarianship* 21: 484-491.

Thorn, J.A. and Zobel, J. 1992. A Model for Word Clustering. *Journal of the American Society for Information Science* 43 (9): 616-622.

van Raan, A. F. J. & Tijssen, R. J. W. 1993. The neural net of neural network research. *Scientometrics* 26 (1): 169-192.

Voorbij, H.J. 1998. Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation* 54 (4): 466-476.

Witten, I., & Bell, T. 1990. Source models for natural language text. *International Journal of Man-Machine Studies* 32: 545-579.

Zipf, G. 1936. *The psycho-biology of language: An introduction to dynamic philology*. London: Routledge.

Zipf, G. 1949. *Human Behaviour and the Principle of Least Effort*. Cambridge: Addison-Wesley.