

Comparative Study Between Words in Titles and Keywords of Some Articles on Knowledge Organisation

Dines Ch. Maiti* and Bidyarthi Dutta**

Central Library, Indian Institute of Technology, Kharagpur-821 302
E-mail: dines.mati@gmail.com

Department of Library & Information Science, Vidyasagar University, Midnapore-721 102
E-mail: bidyarthi.bhaswati@gmail.com

ABSTRACT

This paper has presented research trends in the area of knowledge organisation since 2002 to 2011. In all, the bibliographic details of 285 research articles on some thrust research areas of knowledge organisation as represented by three keywords, viz., classification and indexing, faceted classification and folksonomy, have been collected from Emerald E-journal Consortia hosted by UGC-INFONET and analysed to study the research trend. The growth of literature in this area shows a logistic pattern over the said span. The distribution of articles over journals has been studied and *Journal of Documentation* ranked top followed by *Aslib Proceedings*, *The Electronic Library* and *Library Hi Tech* for three specific subject areas consecutively. The frequencies of assigned keywords have been analysed and core keywords have been identified. The words in titles of 285 articles have been analysed and compared with the words in assigned keywords. Two indicators have been defined, viz., Matching Coefficient (MC) and Title-to-Keyword Ratio (TKR) to study the state of matching between these two types of word clusters after comparative study. There are so many MC as found for different articles ranging from 0 to 0.5. The MCs have been graded in a scale set according to their numerical values. In all, seven grades have been set to study the state of matching between two types of words. In general, percentage of bad matching outnumbers the good matching. As titles are assumed as a most concise abstract of an article, it is customary to assume words in title as true reflector of central theme of the same. If words in assigned keywords poorly match the same in titles, then it may be logically inferred that keywords are incorrectly assigned for majority articles.

Keywords: Knowledge organisation, subject classification, subject indexing, faceted classification, folksonomy, ontology, title analysis, keyword analysis, information retrieval

1. INTRODUCTION

The term knowledge organisation (KO) indicates a field of study chiefly related to Library and Information Science (LIS). The scope of KO covers various functional activities such as content description of document, indexing and abstracting, document classification, cataloguing of information resources, database management, etc. These activities are done by librarians, archivists, subject specialists as well as by computer algorithms. The KO, as a field of study is concerned with the nature and quality of such knowledge organising processes (KOP) as well as the knowledge organising systems (KOS) used to organise documents, document representations, and concepts.

There exist different historical and theoretical approaches for organising knowledge, which are related to different views of knowledge, cognition, language, and social organisation. Library and information

service professionals have often concentrated on applying new technology and standards in this area of knowledge. Traditional human-based activities are increasingly challenged by computer-based retrieval techniques. Hjørland¹ provided an overview of different approaches to KO as:

(a) Traditional Approach

The approaches provided by Melvil Dewey (1851-1931) and Henry Bliss (1870-1955) are mainly classed under this category. Dewey's approach found an efficient way to manage library collections. An important characteristic in Henry Bliss' (and many contemporary thinkers of KO) was that the sciences tend to reflect the order of nature and that library classification should reflect the order of knowledge as uncovered by science:

Natural order --> Scientific classification --> Library classification (KO).

The principles, which may be attributed to the traditional approach to KO are:

- Principle of controlled vocabulary
- Cutter's rule about specificity
- Hulme's principle of literary warrant
- Principle of organising from the general to the specific

After more than 100 years of research and development in LIS, the 'traditional' approach still has a strong position in KO and in many ways its principles still dominate.

(b) Facet Analytic Approach

This approach was initiated with the publication of S.R. Ranganathan's Colon Classification in 1933, which has been further developed by, in particular, the British Classification Research Group. Its analytico-synthetic methodology. The meaning of the term 'analysis' is: Breaking down each subject into its basic concepts. The meaning of the term 'synthesis' is: Combining the relevant units and concepts to describe the subject matter of the information package in hand.

(c) Information Retrieval Approach (IR)

This approach was incepted with the Cranfield experiments, which were founded in the 1950s, and the TREC experiments (Text Retrieval Conferences) started in 1992. It was the Cranfield experiments, which introduced the famous measures "recall" and "precision" as evaluation criteria for systems efficiency. The Cranfield experiments found that classification systems like UDC and facet-analytic systems were less efficient compared to free-text searches or low level indexing systems ('UNITERM'). The Cranfield I test found the following results:

- UNITERM 82 % recall
- Alphabetical subject headings 81.5 % recall
- UDC 75.6 % recall
- Facet classification scheme 73.8 % recall

Although these results have been criticised and questioned, the IR-tradition became much more influential while library classification research lost influence. The dominant trend has been to regard only statistical averages.

(d) User-oriented and Cognitive Approach

User studies demonstrated very early that users prefer verbal search systems as opposed to systems based on classification notations. This is one example of a principle derived from empirical studies of users. Adherents of classification notations may, of course, still have an argument: That notations are well-defined and that users may miss important information by not considering them. Folksonomy is a recent kind of KO based on users' rather than on librarians' or subject specialists' indexing. The term

folksonomy has come as an amalgamation of two words folk and taxonomy. Concept of folksonomy embraces user-assigned or clientele-assigned keywords for the purpose of tagging. It is social tagging but not standardised tagging according to some pre-supposed subject access tool. The concept of keywords is more important here rather the concept of subject headings or subject descriptors. The subject heading is a standardised list while the keywords are personalised to some extent. There is no standard criterion for keyword selection in any subject. The subject-specific keywords may be classified on some chosen parameters². The keywords selected from any specific subject area show special characteristics as was observed in case of the subject domain Fermi liquid³ under the broad area of physics. The content management of any subject depends on categories of selected keywords as observed in case of physics⁴. Another notable feature in this approach of KO is recognising necessary and sufficient conditions under which a word may be regarded as keyword. Actually, it is true that all keywords are words but all words are not necessarily keywords. But the question is that under which circumstances a word may be recognised as a keyword. Is there any objective criteria that may chalk out the necessary and sufficient conditions to be fulfilled by a word for becoming a keyword? Also, a keyword may contain either one or several single words. The association between several single words generally creates keywords that was studied for the specific subject domain low temperature physics⁵.

(e) Bibliometric Approaches

These approaches are primarily based on using bibliographical references to organise networks of papers, mainly by bibliographic coupling or co-citation analysis. In recent years it has become a popular activity to construe bibliometric maps as structures of research fields. The bibliographic scattering is an important phenomenon in this approach, which was investigated from the viewpoint of generalised source approach by Sen⁶. The scattering phenomenon was also observed from the viewpoint of Bose-Einstein Statistics by Dutta & Sen⁷. The physicist and information scientist, Derek De Solla Price unified all informetric and scientometric laws in a logical way⁸. This approach however centrally speaks about networking of scholarly literature on various streams.

(f) Domain Analytic Approach

The indexing of a given document should reflect the needs of a given group of users or a given ideal purpose. In other words, any description or representation of a given document is more or less suited to the fulfillment of certain tasks. It is the

only approach to KO which has seriously examined epistemological issues in the field, i.e., comparing the assumptions made in different approaches to KO and examining the questions regarding subjectivity and objectivity in KO. Hjørland described this approach as a new horizon in information science^{9,10}. Subjectivity concerns about individual differences. A kind of subjectivity about many users is related to philosophical positions. In any field of knowledge different views are always at play. In arts, for example, different views of art are always present.

The different approaches of KO prevailing since long back and even today it is an interesting topic of active research. It is a crux area in library and information science. The subject basically deals with processing and organisation of information so that it can be promptly retrievable by information clientele. LIS is an interdisciplinary and multidisciplinary subject. Therefore this subject embraces so many other subject domains. In fact, multidisciplinary or interdisciplinary study of LIS or KO is another vast area of study. Understanding of subject is a core problem of LIS and KO as well, and attempts are continuously being taken since last hundred fifty years so that an objective or exact answer might be sought. But still it is perhaps an unsolved problem in the area of LIS. Hjørland¹¹ described in detail the history and evolution of the concept of subject in information science during last hundred years. Also, the paradigm shift of the concept of library classification in the internet era is nicely described by Hjørland¹². Some important areas of KO are, as also clear from Hjørland's description of different approaches as above, classification, indexing, folksonomy, ontology, etc. For this study three keywords have thus been chosen, i.e. classification and indexing, faceted classification and folksonomy, as they may be regarded as appropriate representative of the core area of knowledge organisation. The foremost keyword classification and indexing is highly contextual to first four approaches. Actually it a very general keyword of the subject KO. Actually here two separate keywords, classification and indexing are connected. These two keywords closely resemble trains of interlinked functions. Therefore these two keywords are jointly chosen here. The keywords faceted classification and folksonomy are contextual to second and fourth approaches mainly. The subject domains represented by these three keywords also indicate thrust areas of active research in KO.

2. OBJECTIVES

The objectives of this study are to:

- (a) Study research trend of some selective areas of knowledge organisation since the year 2002 to 2011

- (b) Observe growth of literature since the year 2002 to 2011
- (c) Find out core keywords reflecting the central theme of the said subject
- (d) Identify core journals in the said subject area
- (e) Draw a comparative study between the words in title and words in assigned keywords for collected 285 research articles, and
- (f) Measure the appropriateness of chosen keywords through analysis of a pre-defined parameter, viz., matching coefficient.

3. SCOPE AND METHODOLOGY

This study has been carried out over a sample of bibliographic details of 285 research articles distributed over 36 journals collected from the bibliographic database of Emerald accessed through e-journal consortium INFONET hosted by INFLIBNET at Vidyasagar University. In Emerald database research articles are spread over a time span of ten years ranging from the year 2002 to 2011. In all, three keywords have been selected as representative of thrust areas of knowledge organisation for this study. The keywords classification and indexing are crux areas of knowledge organisation in the context of LIS. The keyword faceted classification indicates a modern approach to classification incepted by Ranganathan in the third decade of last century in contrast to traditional enumerative approach introduced by Melvil Dewey in the seventh decade of nineteenth century.

The keyword Folksonomy is a comparatively new term. This term was coined by Thomas Van Der Waals in 2004¹³. This term specifically refers to subject indexing systems created within internet communities. A folksonomy may be viewed as a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorise content. The practices involved in folksonomy are also known as collaborative tagging, social classification, social indexing and social tagging.

In Emerald database the information is available from the following types of sources, i.e., journals, books, bibliographic database, case studies, and site pages. Only the journal has been considered for the present study as source of information. Since this study involves analysis of research trend, therefore journal articles have been taken under consideration. Three sources also feebly contain original research articles. It is to be noted that, journals also contain different types of writings other than research articles, for instance, short communication, news item, letters to editor, book review, editorial comment, etc., All those types of writings were not considered for this study except only the research articles.

Table 1. Keywords-wise classification of research papers

Keywords	No. of research articles retrieved	No. of journals covered	No. of authors involved	No. of assigned keywords
Classification and indexing	107	16	168	203
Faceted classification	125	34	222	280
Folksonomy	53	14	117	129
Total no. of research articles	285			

The numbers of research papers retrieved against each keyword along with total numbers of journals, authors and assigned keywords are given in Table 1.

Each bibliographic detail of papers downloaded contains fourteen metadata, which are listed in abbreviated form at the left side of the record. The abbreviations used for each metadata have been explained as:

TY - Type

T1 - Title of the article

AU - Author of the article

KW - Assigned keyword

JO - Name of the journal where the concerned article was published

PY - Year of publication

PB - Name of the publisher

VL - Volume no.

IS - Issue no.

SP - Starting page no. of the article in the concerned journal

EP - Ending page no. of the article in the same journal

SN - ISSN of the concerned journal

M3 - Digital Object Identification no.

UR - URL of the article

N2 - Abstract of the paper

A sample of one downloaded articles bibliographic details is :

TY - JOUR

T1 - Google Scholar as a tool for discovering journal articles in library and information science

AU - Dirk Lewandowski

KW - Information science

KW - Libraries

KW - Search engines

JO - Online Information Review

PY - 2010

PB - Emerald Group Publishing Limited

VL - 34

IS - 2

SP - 250

EP - 262

SN - 1468-4527

M3 - DOI: 10.1108/14684521011036972

UR - <http://dx.doi.org/10.1108/14684521011036972/>

N2 - Purpose – The purpose of this paper is to measure the coverage of Google Scholar for Library and Information Science (LIS) journal literature uavailability.

Out of fourteen only five metadata have been used for this study, i.e., T1 – Title of the article, AU – Author of the article, KW – Assigned keyword, JO – Name of the journal and PY – Year of publication.

The authorship pattern has been studied from the list of authors. The year of publication was used to study the year-wise distribution of articles. The journal-name was studied to trace the core journals in the field. The assigned keywords were studied from the metadata tagged as KW. The title was analysed to study the words in title. The assigned keywords were also analysed to study the words in keywords. After that, the words in title were compared with the words in assigned keywords. It is to be noted that only single words have been taken both from titles and assigned keywords. The phrases containing more than one word were split up in single words. For instance, the phrase 'electronic journal', has been split up in two different words, 'electronic' and 'journal'.

It this study two parameters are defined to carry out comparative study between words in titles and assigned keywords. Suppose, in a record the title consists of 'M' no. of single words and all assigned keywords consist of 'N' no. of single words. Of these say, 'W' no. of single words from title exactly alike to same no. of single words from assigned keywords. The quantity may thus be denoted as number of matching words between title and assigned keywords. It is to be noted that during selection of single words from titles, only the full words have been taken, but the form words like article, proposition, conjunction, interjection, etc., have been excluded. The full words indicate words whose meaning could be found in dictionaries, that is to say the full words carry some lexical meaning. While the form words belong to the grammar and carry only grammatical implications. Also in case of words with identical roots but different stems, the

roots are taken under consideration. For instance, the two words childhood and children have been considered as matched words, because these two words contain identical root, i.e. child, but different stems, i.e. hood and ren. The two words library and libraries have also been considered as matched word on the same logical ground.

Let us represent the ratio of words in title to words in assigned keyword bedefined as TKR or Title-Keyword-Ratio, so

$$\text{TKR} = M/N \quad (1)$$

Secondly, let the ratio of matching words to the total number of single words appearing in title and assigned keywords be defined as Matching Coefficient (MC) so:

$$\text{MC} = W / (M + N) \quad (2)$$

where, M = No. of single words in title N = No. of single words in assigned keywords; and W = No. of matching words

Now the following cases may occur for TKR,

Case 1: If $M > N$, then $\text{TKR} > 1$,

Case 2: If $M < N$, then $\text{TKR} < 1$,

Case 3: If $M = N$, then $\text{TKR} = 1$,

The following cases may occur for MC,

Case 1: If all single words in title become exactly alike with all words in assigned keywords, i.e. no. of words in title is equal to no. of words in assigned keywords, then a special phenomena will happen that may be termed as perfect matching. In case of perfect matching, we have,

$$M = N = W, \text{ as } W = \text{No. of matched words} \quad (3)$$

Substituting Eqn (3) in Eqn (2), we get,

$$\text{MC} = 0.5 \quad (4)$$

As this is the case of best matching, therefore the value of MC in Eqn (4) is the maximum probable value of the same. Thus it may be written as,

$$(\text{MC})_{\max} = 0.5 \quad (5)$$

Case 2: If no word in title matches with that in assigned keyword, then $W = 0$, which is the case of worst matching or no matching. Thus $\text{MC} = 0$, which is the minimum possible value of the same.

$$\text{Hence, } (\text{MC})_{\min} = 0 \quad (6)$$

Thus the range of values of MC is as follows:

$$0 \leq \text{MC} \leq 0.5 \quad (7)$$

In this study, the parameter MC has been segmentised in seven zones depending on different ranges of their values as listed in Table 2.

The method of comparison between words in

title and assigned keywords has been presented below for the bibliographic record given above:

Here the title is: Google Scholar as a tool for discovering journal articles in library and information science.

The single full words in this title are: Google, Scholar, tool, discovering, journal, articles, library, information, science

Form words in this title are: as, a, for, in, and

Assigned keywords in this title are: Information science, Libraries, Search engines

Single full words in the assigned keywords are: Information, science, Libraries, Search, engines

Now, in this record, $M = 9$, $N = 5$ and $W = 3$, as no. of matched words are three.

Thus $\text{TKR} = 9/5 = 1.8$ and $\text{MC} = 3 / (9 + 5) = 3 / 14 = 0.214$; this value of MC belongs to moderate matching zone at per Table 2.

4. RESULTS AND ANALYSIS

4.1 Growth of Literature

The year-wise distribution of research papers has been presented in Table 3. The steady growth of literature has been observed since the year 2002 to 2011. The no. of articles in the year 2002 was 17, while the same in the year 2011 was 44, i.e., almost 2.5 times more than 2002. The maximum no. of articles was observed in 2010, which was 45. The graphical presentation of the growth of literature has been shown in Fig. 1. The graph has been plotted using the open-source graphic software Findgraph. It has been found out that the logistic pattern was followed by growth of literature during the said time interval, i.e. from 2002 to 2011. The exact equation followed by the growth pattern may be written as:

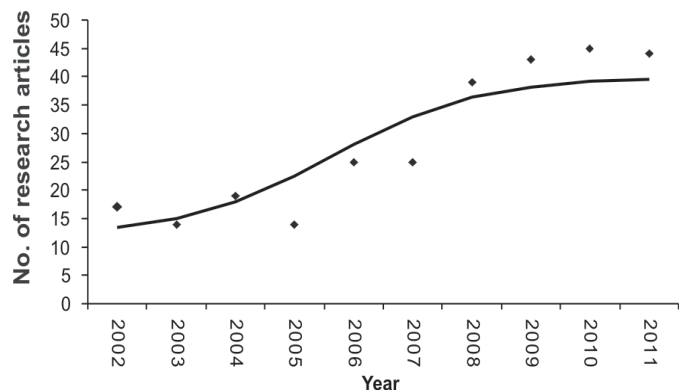


Figure 1. Growth of literature since 2002 to 2011.

Table 2. Scale fixing of matching coefficient

Type of matching	No matching	Very low matching	Low matching	Moderate matching	High matching	Very high matching	Perfect matching
Range of values of MC	0	0.01 - 0.1	0.11 - 0.2	0.21 - 0.3	0.31 - 0.4	0.41 - 0.49	0.5

$$y = 12 + 28 / (1 + \exp(3.7 - 0.8 \cdot x)) \quad (8)$$

where, x represents time variable, and y represents the no. of research articles.

Eqn (1) thus portrays the exact growth pattern of research articles in the concerned subject field.

The growth pattern presented by Fig. 1. shows the consolidated picture from three subject domains indicated by the keywords classification and indexing, faceted classification & folksonomy.

Table 3 reflects three distinct growth patterns for three different subject domains. For the subject domain, classification and indexing, no steady pattern has been observed. The growth of literature in this area followed a fluctuating pattern with the highest no. of publications in 2008 and lowest no. of publications in 2003. In the subject domain, faceted classification a steady growth pattern has been observed. The lowest no. of publications came in 2004, while highest no. of publications emerged in 2011. On the other hand, the growth process for the subject area folksonomy, started from the year 2006. Actually, the term folksonomy was coined in the year 2004, and therefore the first research activity initiated only after a year, i.e. since

in terms of research activities. The subject faceted classification, though newer compared to classification and indexing, but still it was incepted by Ranganathan in second decade of last century. The first edition of Colon Classification was published in 1933. Therefore this area also attained stability to some extent. But the folksonomy is a new area having full of unsolved problems that accelerates its high tune with research activities.

4.2 Distribution of Articles over Journals

The total number of article contributed in the subject Classification and Indexing is 107. The articles are published in 16 journals. The top five journals along with percentage of articles contained therein are presented in Table 4. It has been observed from Figure 2, that the core journal is *Journal of Documentation* followed by *The Electronic Library*, *Library Review*, *Online Information Review* and *Library Management*. These five journals contained 67 % of total contributions, i.e., nearly two-third of total number of articles. Therefore, these five journals may be regarded as top journals in this area.

Table 3. Distribution of research papers over the years since 2002 to 2011

Year	No. of articles for the keywords			Year-wise total no. of articles
	Classification and indexing	Faceted classification	Folksonomy	
2002	12	5	0	17
2003	6	8	0	14
2004	15	4	0	19
2005	9	5	0	14
2006	15	9	1	25
2007	7	15	3	25
2008	17	16	6	39
2009	9	18	16	43
2010	9	20	16	45
2011	8	25	11	44
Subject (keyword)-wise total no. of articles	107	125	53	285

2006. But just after inception, very rapid growth of research literature has been observed. The highest no. of output occurred in two consecutive years, i.e., 2009 and 2010. The steep rise pattern within a very short time is observed here unlike other two subject areas. This implies that the subject folksonomy is a budding area of active research. This is a promising area of knowledge organisation, where possibility of creation of new research scope is very high.

The other two subject domains are traditional areas; particularly the domain of classification and indexing is a subject having a rich historical tradition since Melvil Dewey's epoch around 1876, when the first edition of Decimal Classification was published. This area, therefore becomes more or less stabilised

Table 4. Top five journals for the subject domain classification and indexing

S. No.	Name of journal	Frequency (%)
1.	<i>Journal of Documentation</i>	19 (17.8 %)
2.	<i>Electronic Library</i>	14 (13.1 %)
3.	<i>Library Review</i>	14 (13.1 %)
4.	<i>Online Information Review</i>	14 (13.1 %)
5.	<i>Library Management</i>	9 (8.4 %)
6.	<i>Aslib Proceedings</i>	8 (7.5 %)
7.	<i>New Library World</i>	7 (6.5 %)
8.	<i>Library Hi Tech</i>	5 (4.7 %)
9.	<i>Bottom Line: Managing Library Finances</i>	4 (3.7 %)
10.	<i>Collection Building</i>	4 (3.7 %)
11.	Others	9 (8.4 %)

The total number of articles in the area of faceted classification is 125. The articles are published in 34 journals. The name of top five journals have been given in Table 5 along with number of articles contained therein. It has been observed that the top most journal is *Journal of Documentation* followed by *Aslib Proceedings*, *Online Information Review*, *Library Hi-tech*, *The Electronic Library* and *Library Review*. This top six journals (ranked one to five) contain 63 %, i.e., nearly two-third of total number of articles.

The total number of articles in the subject area folksonomy is 53. The articles are published in 15 journals. The Table 6 shows journals depicts the consolidated distribution of articles over journals. It has been observed that the top most journals is *Journal of Documentation* followed by *Library Hi Tech*, *Online Information Review*, *OCLC Systems and Services* and *Program: electronic library and information systems*. These five journals consist of 66 %, i.e., almost two-third of total number of articles.

Table 5. Top five journals for the subject domain faceted classification

Name of journal	Total
<i>Journal of Documentation</i>	22.4 %
<i>Aslib Proceedings</i>	9.6 %
<i>Online Information Review</i>	8.8 %
<i>Library Hi Tech</i>	8 %
<i>Electronic Library</i>	7.2 %
<i>Library Review</i>	7.2 %
<i>Reference Services Review</i>	5.6 %
<i>New Library World</i>	3.2 %
<i>Program: electronic library and information system</i>	3.2 %
<i>OCLS Systems & Services</i>	2.4 %
Others	22.4 %

Table 6. Top five journals for the subject domain folksonomy

Name of journal	Total
<i>Journal of Documentation</i>	18.9 %
<i>Library Hi Tech</i>	17 %
<i>Online Information Review</i>	15.1 %
<i>OCLC Systems & Services</i>	7.6 %
<i>Program: Electronic library and information systems</i>	7.6 %
<i>Aslib Proceedings</i>	5.66 %
<i>New Library World</i>	5.66 %
<i>VINE</i>	5.66 %
<i>Electronic Library</i>	3.77 %
<i>Library Management</i>	3.77 %
Others	9.43 %

4.3 Assigned Keyword Pattern

Total number of assigned keywords is 203 with total frequency of occurrence 438. The top fifteen keywords along with their frequencies are presented in Table 7. The total frequency of occurrence of top fifteen keywords is 143, almost one-third of the total frequency of 203 keywords. These top fifteen keywords may therefore be regarded as core keywords in this subject domain. It is clear that the more researches are done in this area than the other fields of knowledge. It is thus clear from Table 7 that the thrust research interest in this subject centre around the terms like libraries, information retrieval, academic libraries, information management and information services. It is clear from Table 7, that the words like electronic, information, research and library occurred maximum times, because these words seldom occurred alone but with other associated words.

Table 7. Top keywords in the subject domain classification and indexing

List of keywords	Frequency
Libraries	17
Information retrieval	15
Academic libraries	12
Information management	11
Information services	11
Digital libraries	10
Research	10
Databases	8
Electronic journals	8
Librarians	8
Internet	7
University libraries	7
Worldwide web	7
Electronic publishing	6
India	6

Total number of assigned keywords in this subject domain is 279 with total frequency of occurrence 528. The top eighteen core keywords along with their frequencies are shown in Table 8. The total frequency of top eighteen core keywords is 174, i.e., nearly one-third of total frequency of occurrence of all 280 keywords. It is evident from Table 8 above that the principal research interest in this area centre around the keywords like Information retrieval, classification, digital libraries, academic libraries, internet, online catalogue, semantics, classification schemes, controlled language construction, user interface etc. (Table 9). Some very broad umbrella terms also occurred here like internet, knowledge organisation, information science, information management, etc. The keywords like library, information, electronic,

Table 8. Top keywords in the subject domain faceted classification

List of keywords	Total
Information retrieval	22
Classification	17
Digital libraries	15
Knowledge management	12
Academic libraries	11
Internet	11
Worldwide web	10
Libraries	9
Online catalogues	8
Semantics	8
Classification schemes	7
Controlled language construction	7
Information systems	7
User interfaces	7
Information management	6
Information science	6
United Kingdom	6
Communication technologies	5

Table 9. Top keywords in the subject domain folksonomy

List of keywords	Total
Internet	11
Tagging	11
World wide web	9
Academic libraries	6
Information retrieval	6
Knowledge management	6
Learning	6
Social networks	6
Digital libraries	5
Communication technologies	4
Indexing	4

knowledge, research, online, organisational, etc., hardly occurred alone, but with other words.

4.4 Words in Titles and Assigned Keywords: A Comparative Study

A comparative study between words in titles and words in assigned keywords for classification and indexing is shown in Table. The values of Matching Coefficients have been graded into seven ranges in order of quality of matching. The value of matching co-efficient varies from 0 to 0.5. The zero value indicates no matching, while the highest value of 0.5 indicates perfect matching. The intermediate steps from no matching to perfect matching are: very low matching, low matching, moderate matching, high matching and very high matching. It is evident that 43 % of total no. of articles showed low matching. For 13.1 % of articles there was no matching between words in title and assigned keywords, while for 15.9 % of total no. of articles the matching profile was very low. Only 21.5 % articles showed moderate matching (Table 10). Article was found with very high and perfect matching profile. From the above categorisation it is clear that the words in titles and assigned keywords do not match in most cases. As title is the most concrete and compact abstract of an article, therefore it may be safely assumed that the title is always assigned in right way. The title reflects the central theme of an article in briefest way. The assigned keywords are generally supplied by the authors and/or by editors sometimes. Now, there should be tally between words in title and assigned keywords, so that the keywords may help in information retrieval properly. As in this case very poor tally or matching is observed, therefore it may be assumed that in most cases the keywords are not assigned in the right way. The assigned keywords thus rarely reflect the central theme or core area of research from an article as observed here. In about 73 % articles, i.e., cases of no matching, very low matching and low matching, the keywords may be presumed as wrongly assigned, while in remaining 27 % articles, i.e. cases of moderate matching and high matching, the keywords are more or less correctly assigned.

Also, it is clear from Table 11, that for 36 % of total articles, no. of words in title is smaller than no. of words in assigned keywords (i.e., $TKR < 1$),

Table 10. Distribution of articles over different scales of matching for classification and indexing

Matching profile	No matching	Very low matching	Low matching	Moderate matching	High matching	Very high matching	Perfect matching
Value range	0	0.01 - 0.1	0.11 - 0.2	0.21 - 0.3	0.31 - 0.4	0.41 - 0.49	0.5
No. of research articles with different matching profiles	14	17	46	23	7	0	0
Percentage of research articles with different matching profiles	13.1 %	15.9 %	43.0 %	21.5 %	6.5 %	0	0

Table 11. Title keyword ratio for classification and indexing

Relative size of M & N	No. of articles (%)	TKR=M/N
M<N	39 (36 %)	< 1
M=N	7 (7 %)	= 1
M>N	61 (57 %)	> 1

while for 57 % of articles the case is reverse (i.e., TKR > 1). For only 7 % of total articles no. of words in title exactly equals to the no. of words in assigned keywords, i.e., TKR = 1. In majority cases thus, no. of words in title is greater than so in assigned keywords, i.e., title is generally longer than all assigned keywords. Very few articles contained exactly equal no. of words in both titles and assigned keywords.

A comparative study between words in titles and assigned keywords is shown in Table 12 for 'faceted classification'. It is evident that 40 % of total no. of articles showed low matching. For 13 % of articles there was no matching between words in title and assigned keywords, while for 18 % of total no. of articles the matching profile was very low. No article was found with very high and perfect matching profile. From the above classification it is clear that the words in titles and assigned keywords do not match in most cases for faceted 'classification' also. In this case also very poor matching is observed; therefore it may be assumed that in most cases the keywords are not assigned in the right way. The assigned keywords thus rarely reflect the central theme or core area of research from an article as observed in this subject area too. In about 71 % articles, i.e., cases of no matching (NM), very low matching (VLM) and low matching (LM), the keywords may be presumed as

wrongly assigned, while in remaining 29 % articles, i.e., cases of moderate matching (MM) and high matching (HM), the keywords are more or less correctly assigned.

Also, it is clear from Table 13 that for 30 % of total articles, no. of words in title is smaller than no. of words in assigned keywords (i.e., TKR < 1), while for 55 % of articles the case is reverse (i.e., TKR > 1). For only 15% of total articles no. of words in title exactly equals to the no. of words in assigned keywords, i.e., TKR = 1. In majority cases thus, no. of words in title is greater than so in assigned keywords, i.e. title is generally longer than all assigned keywords. Very few articles contained exactly equal no. of words in both titles and assigned keywords as usual. Though in this case, the no. of articles with TKR = 1, is double than that of previous one.

A comparative study between words in titles and assigned keywords is shown in Table 14 36 % of total no. of articles showed low matching. For 9 % of articles there was no matching between words in title and assigned keywords, while for 9% of total no. of articles the matching profile was very low. No article was found with very high and perfect matching profile. From the above classification it is clear that the words in titles and assigned keywords do not match in most cases. In this case also very poor matching profile is observed; therefore it may be assumed that in most cases the keywords are not assigned in the right way. The assigned keywords thus rarely reflect the central theme or core area of research from an article as observed in this subject area too. In about 55 % articles, i.e., cases of no matching (NM), very low matching (VLM) and low matching (LM), the keywords may be presumed as

Table 12. Distribution of articles over different scales of matching for faceted classification

No matching	Very low matching	Low matching	Moderate matching	High matching	Very high matching	Perfect matching
0	0.01 - 0.1	0.11 - 0.2	0.21 - 0.3	0.31 - 0.4	0.41 - 0.49	0.5
16	22	50	27	10	0	0
13 %	18 %	40 %	22 %	8 %	0 %	0 %

Table 13. Title keyword ratio for faceted classification

Relative size of M & N	No. of articles (%)	TKR=M/N
M<N	37 (30 %)	< 1
M=N	19 (15 %)	= 1
M>N	69 (55 %)	> 1

wrongly assigned, while in remaining 45 % articles, i.e., cases of moderate matching (MM) and high matching (HM), the keywords are more or less correctly assigned. In case of folksonomy, however the percentage of good matching (cases of MM & HM) is relatively higher compared to other two subject areas, i.e., 45 %, while the same was 29 %

Table 14. distribution of articles over different scales of matching for folksonomy

No matching	Very low matching	Low matching	Moderate matching	High matching	Very high matching	Perfect matching
0	0.01 - 0.1	0.11 - 0.2	0.21 - 0.3	0.31 - 0.4	0.41 - 0.49	0.5
5	5	19	17	7	0	0
9 %	9 %	36 %	32 %	13 %	0 %	0 %

and 30 % respectively for other two subject areas. It has thus been observed that there is an increase in good matching and consecutively decrease in bad matching in case of 'folksonomy'.

Also, it is clear from Table 15 that for 40 % of total articles, no. of words in title is smaller than no. of words in assigned keywords (i.e., $TKR < 1$), while for 45 % of articles the case is reverse (i.e., $TKR > 1$). For only 15% of total articles no. of words in title exactly equals to the no. of words in assigned keywords, i.e. $TKR = 1$. In majority cases thus, no. of words in title is greater than so in assigned keywords, i.e. title is generally longer than all assigned keywords. Very few articles contained exactly equal no. of words in both titles and assigned keywords. Though in this case, the no. of articles with $TKR = 1$, is double than the subject of 'classification and indexing'.

Table 15. Title keyword ratio for folksonomy

Relative size of M & N	No. of articles (%)	TKR=M/N
M<N	21 (40 %)	< 1
M=N	8 (15 %)	= 1
M>N	24 (45 %)	> 1

5. CONCLUSIONS

This study has been carried over the bibliographic details of 285 research papers in all, collected from Emerald database. The findings that have been observed after thorough inspection of all research articles are listed below:

- (a) The growth pattern for three different subject domains have been studied in a consolidated way that has been presented in Table 3. It has been noticed that the consolidated growth pattern follows a mathematically systematic pattern that can be represented through a logistic function as expressed by Eqn 7. But when the growth patterns for three subject domains are studied separately, then three distinct patterns were followed. The subject domain classification and indexing shows no regular growth pattern, but a zig-zag one fluctuating over different years. The subject domain 'faceted classification' shows a steady growth pattern over the entire time span of ten years. The subject domain 'folksonomy' shows a steep growth pattern. As this concept was incepted in the year 2004, research about this topic was started in 2006. But since after beginning the growth occurred very quickly.
- (b) From the study of journal pattern it has been observed that *Journal of Documentation*, *Library Hitech*, *Aslib Proceedings* and *The Electronic Library* are core journals in this area.
- (c) From the analysis of assigned keyword profile the core keywords, which have been found are

as follows: library, internet, information retrieval, academic library, electronic library, classification, etc. Some keywords like, information, research, electronic, knowledge, online etc hardly occurred once, but with other terms. It is therefore clear that these areas indicate potential topics for research.

- (d) From the comparative study of words in title and assigned keywords it is clear that 40% of total no. of articles showed low matching. No article was found with very high and perfect matching profile. From the above classification it is clear that the words in titles and assigned keywords do not match in most cases. As title is the most concrete and compact abstract of an article, therefore it may be safely assumed that the words in title are always assigned in right way. The title reflects the central theme of an article in briefest way. The assigned keywords are generally supplied by the authors and/or by editors sometimes. Now, there should be tally between words in title and assigned keywords. As in this case very poor tally or matching is observed, therefore it may be assumed that in most cases the keywords are not assigned in the right way. The assigned keywords here rarely reflect the central theme or core area of research from an article.

In the traditional subject domains, like 'classification & indexing' and 'faceted classification', the good matching and bad matching percentages were about 30 % and 70 % respectively, while in the newly budding subject domain like 'folksonomy', the good matching and bad matching percentages were 45 % and 55 % respectively. It means, the keywords were wrongly assigned in traditional subject domains, while more or less rightly assigned in new subject domain.

- (e) It is also clear that the value of TKR for majority of articles is greater than one. For few number of articles it is equal to one, and for little fare no. of articles the same is less than one.
- (f) Differences have been observed in research trends between traditional and emerging subject domains.

REFERENCES

- Hjørland, B. What is knowledge organisation (KO)? *Knowledge Organisation*, 2008, **35**(2/3), 86-101.
- Dutta, B.; Majumder, K.P. & Sen, B.K. Classification of keywords extracted from research articles published in science journals. *Annals Lib. Inf. Studies*, 2008, **55**(4), 317-33.
- Dutta, B.; Majumder, K.P. & Sen, B.K. An analytical model for investigation of some characteristics

- of the keywords of the subject Fermi liquid: A case study. *Annals Lib. Inf. Studies*, 2009, **56**(4), 273-90.
4. Dutta, B.; Majumder, K.P. & Sen, B.K. Content management by keywords: An analytical study. *SRELS J. Inf. Manag.*, 2010, **47**(6), 599-620.
5. Dutta, C. & Dutta, B. Mechanism of keyword formation in research articles on physics: An empirical study. *SRELS J. Inf. Manag.*, 2013, **50**(1), 3-22.
6. Sen, S.K. Bibliographic scattering: A generalised source approach, *Scientometrics*, 1989, **17**, 197-204.
7. Dutta, B. & Sen, B.K. The scattering of articles over a journal system as observed from the viewpoint of Bose-Einstein statistics. *Annals Lib. Inf. Studies*, 2005, **52**(2), 51-55.
8. Price, D.D.S. Bradford's law and related statistical patterns. *Current Contents*, 1980, **19**, 5-12
9. Hjørland, B. & Albrechtsen, H. Toward a new horizon in information science: Domain-analysis. *JASIS*, 1995, **46**(6), 400-25.
10. Hjørland, B. Domain analysis in information science: Eleven approaches—traditional as well as innovative. *Journal of Documentation*, 2002, **58** (4), 422-62.
11. Hjørland, B. The concept of 'subject' in Information Science. *Journal of Documentation*, 1992, **48**(2), 172-200.
12. Hjørland, B. Is classification necessary after google. *Journal of Documentation*, 2012, **68** (3), 299-317.
13. <http://www.wikipedia.org/>.