International Journal of Information Science and Management

A Statistical Study on Persian Subject Headings Development

M. Tavakolizadeh Ravari, Ph.D.
Yazd University, I. R. of Iran
Email: tavakoli@yazduni.ac.ir

Abstract

Controlled vocabularies have been frequently used in information retrieval systems. Control of the vocabularies and evaluating the utility of their terms are two critical questions. This research aims at the development of Persian subject headings through statistical analyses. The current research was conducted on more than 450,000 records extracted from the electronic version of National Bibliography of Iran (NBI). Data has been processed through data mining techniques. The correlation analysis was performed to determine the relationship between the number of items in NBI and the number of Persian subject headings as well as the rank of each subject heading and its use frequency in NBI.The count of new subject headings vs. the count of new catalogued materials in NBI grew linearly at the beginning and increased logarithmically when the number of catalogued materials reached 3,200. The analysis of the use frequency of distinct headings within NBI resulted in three classes: most, frequent, and normal used subject headings. The findings partly agree with Lancaster's prediction, as he states that a controlled vocabulary will grow very fast in the beginning. It was also found that the majority of subject headings are rarely used by NBI. It is due to absence of a mechanism to control the building of new headings.

Keywords: National Bibliography of Iran, Persian Subject Headings, Controlled Vocabularies, Use Frequency.

Introduction

Controlled vocabularies have been developed and used in different fields in order to establish applications for language learning, classification and organization of information. How to control their size and how to evaluate the utility of their terms are two critical questions.

There is no agreement on how to control the size of controlled vocabularies. Svenonius (1986) says, "It is up to the designer of a controlled vocabulary to decide just how much control and what forms of control to incorporate in it". On the other hand, Lancaster (1986) states: "how large the vocabulary will be depends not only on the subject field but on the

specificity of the terms and type of terms used". Wurm (1964) believes that development of such vocabularies follows a mathematical function. He expresses: "the file growth [of controlled terms] follows a pattern which can be represented by a mathematical model. This seems to open up the possibility of estimating more exactly the total number of different terms … This is a probability problem and it seems reasonable to assume that within each category of terms the file size could be presented by combined sum of the sums of a number of geometrical progressions". It led him to offer a function to determine the size of controlled vocabularies.

A part of the current research focuses on the size of Persian Subject Headings which is a controlled vocabulary developed by the NBI as a tool for cataloguing Persian books. It is based on the hypothesis that the development of controlled vocabularies is normally supported by the publication materials that are indexed in their corresponded indexing or cataloguing system. It means that if a material is published for the first time and includes some new subjects, which have not been in the vocabulary before, we can say that something new should be added to it. Inferentially, we expect a relationship between the publication of new materials and inclusion of new terms to a controlled vocabulary.

In addition to the problem of vocabulary size, there is not a pervasive method to evaluate the utility of terms that have been added to the controlled vocabularies. We can suppose that the frequently used terms have been correctly created. Consequently, the rarely used subject terms should be replaced with proper ones or should be removed from their corresponded vocabulary. A term may also be used very frequently and result in information overload. Pratt (1999) states: "People become overwhelmed by the amount of information. They become frustrated when their searches yield tens or hundreds of relevant documents". These facts show that the efficiency of controlled terms will affect the retrieval performance.

Making use of the method that Zipf (1949) established to study the distribution of words in natural texts may help evaluate the utility of terms within the controlled vocabularies. He plotted the ranks of unique words against their use frequency and found out that there is a power-law (log-log) relationship between these two variables. Furthermore, he showed that the power of this distribution is equal to -1 in almost any English text. Since all the terms within the controlled vocabularies are content-bearing, their distribution type should be varied from Zipf's.

Review of Literature

The statistical works on the "controlled vocabulary size" and the "use frequency of index terms" flourished in 1960's. A number of them sought to find a model for distribution of index terms. The Zipf distribution has been noticed in many of these research works.

One of the earliest studies is that by Wurm (1964), who examined the relationship

between the number of “U.S. Pharmaceutical Patents” and its corresponded vocabulary. He found that it is not possible to obtain an accurate curve from relations between the number of documents and the number of terms.

Houston and Wall (1964) applied correlation analyses on nine indexes to predict the percentage of terms in a manipulative index vocabulary which will be used to index any given number of documents. They reported that a log-normal (logarithmic) relationship exists between total index entries and distribution of term usage, but Cleverdon et al (1966) in a work related to the factors determining the performance of indexing systems found a log-log (power-law) relationship between the number of documents and the number of index terms .

Bennett (1975) modeled index terms of two bibliographic databases using a Zipf distribution. He found a poor fit in the tail of curve and explained that this is due to the relatively small size of the index term set, but subsequent research by others would show this was often the case even with larger numbers of terms.

Burnett et al (1979) examined the effect of the size of controlled vocabularies on retrieval and found that the retrieval performance improved with an increase in the number of controlled vocabulary terms used to index documents.

Fedorowicz (1982) relied on different formulations of Zipf's Law to model the distribution of terms in the MEDLINE database. He used parameters describing the contents of MEDLINE and its inverted file and applied the form of Zipf's law that was developed by Booth (1967) which originally estimates the number of words of a particular frequency for a given author and text.

Umstätter (1986) plotted the use frequency of thesaurus terms in GEOLINE database against their rank and obtained a normal-log (exponential) relation.

Nelson (1989) used a probability model of the occurrence of index terms to derive discrete distributions which are mixtures of Poisson and negative binomial distributions. They found that these distributions, the generalized inverse Gaussian-Poisson and the Generalized Waring give better fits than the simpler Zipf distribution, particularly in the tails of the distribution where the high frequency terms are found. Wolfram (1992) found that a three-parameter Mandelbrot-Zipf have been shown to provide better fits than the traditional Zipf.

Spink et al (2001) analyzed the use frequency of terms that users had used in their search queries in Excite search engine. They followed the Zipf’s method and plotted the log term rank of queries against the log term frequency. They express that the resulting distribution is slightly unbalanced for the high and low ranking terms, indicating that, just as with database term distributions (which is the subject of the current research), a query term distribution may require a more sophisticated model to describe the relationship between the selection of terms and their frequency of appearance within queries.

Tavakolizadeh-Ravari (2007a) examined the relations between the size of Medical Subject Headings (MeSH) and the number of records in MEDLINE. He found: “MeSH has logarithmically developed through three different phases. The existence of each phase has been due to the need of optimizing the growth rate of thesaurus terms to cope with the exponential increase of indexed documents”.

Research Design

Problem: Following the development of retrieval technology, the automatic indexing is widely used now and free-text searching is a part of retrieval systems. Despite of this, subject indexing through controlled vocabularies has preserved its role in information retrieval systems. The problem in this way is the unavailability of a pervasive method to control their size and evaluate the utility of terms that have been added to these vocabularies. The same problem occurs in Persian Subject Headings (PSH), too.

Aim: The attempt is made to study the development of Persian subject headings through statistical analysis. In order to achieve this aim, the research seeks to determine:

1. The relationship between the inclusion of new items to NBI and the need for creation of further headings.

2. The frequency distribution of Persian subject headings in NBI.

Questions: The results were expected to find answers for the following questions:

1. How have the Persian subject headings expanded regarding the inclusion of new items to NBI?

2. How is the frequency distribution of main Persian subject headings in NBI?

Limitations: The research faced with four major limitations:

1. Hajizeynolabedini et al. (2000) reported a large number of errors in the electronic version of NBI. These errors were probably made during the data entry phase of the database creation. The lack of coherent general instruction to unify the subject headings led to these errors in that phase. In this research, a number of them were identified and corrected automatically but correction of the whole errors was impossible. Therefore, some headings actually were identical but the program could not make distinction between them. For example, wrong insertion of two spaces between the words “Persian” and “Poetry” makes this term different from that with one space. These errors give rise to a number of subject headings that actually are not available but incorrectly receive a frequency.

2. A number of items catalogued in NLI were not included in the electronic version of NBI. As a result, it was impossible to determine the use frequency of the whole Persian subject headings exactly.

3. Complex and specific Persian Unicode was used for producing the electronic version of NBI. Hence, the information processing took tens of hours by a high speed PC. This led to decide to limit the research to the main subject headings. Therefore, the

subheadings were eliminated despite their importance.

4. The automatic separation of “form headings” or “place names” was not possible. Thus, we see that the list of high frequent headings in NBI includes items such as English, Iran, and Quran which are not subject headings (see Table 2).

Materials and Methods

There is a relationship between the expansion of NBI items and the development of the Persian subject headings list. Two copies of all published materials in Iran should be delivered to the NLI. The cataloguing information of those materials is added to the NBI. This is a support for the development of Persian subject headings. It has the potential ability to show the time when a distinct subject heading is created and the number of times used. Therefore, the electronic version of NBI was applied as the main material for the current research. This included more than 450,000 records at the research time.

C-Sharp (C#) programming language was then chosen for the processing of information to achieve the data needed for conducting this research. The required fields for processing were “Publication Date”, and “Subject Headings”. Records were first transferred to a table in MS-SQL Server. The equivalent codes to the Persian characters were then identified to make the similar subject headings identical and to cope with the problem of complex Persian Unicode. This method enabled the correction of some common mistakes made by the typists. The process followed by eliminating subheadings through a query statement in MS-SQL. The personal names, those assigned as subjects, were ignored as well.

Two operations were performed on the data that were arranged in the tables. The first one focused on achieving data to find the relationship between the inclusion of items to NBI and the development of Persian subject headings. The records were sorted on the publication date in ascending order. The program read them one by one and looked for the subject headings used for cataloguing. If there were any subjects found for the first time, the number of their associated records and the sequence number of their appearances were noted. Table 1 is a cut of the obtained data through this process. Each row shows the processing result of fifty records.

Table 1

Number of new Persian subject headings vs. the number of items in NBI

No. of Items in NBI	No. of New Headings
50	39
100	62
150	90
200	118

No. of Items in NBI	No. of New Headings
250	149
300	178
350	210
400	240
450	271
...	...

The first row of the Table 1 expresses that thirty nine new subject headings were produced for the first 50 items in NBI. In the next one, it shows that sixty two subject headings were produced by the first hundred records. Finally, when the number of catalogued materials reached 458,250, the number of distinct subject headings amounted to 24,397.

The next operation was to determine the number of times that each of those 24,397 subject headings was assigned to the catalogued materials. To perform this, a relational table was created in MS-SQL. It consisted of two fields: “Subject Headings” and its associated “Record Number” in the main table. Each subject (without its subheadings) and its related record number transferred to that table. Through the SQL-Query Search Syntax, every distinct heading was searched in the relational table to count the number of times they were used in NBI. The results were sorted by the use frequency of the subject headings in descending order. The final result was organized as in the table below:

Table 2

The use frequency of each heading

Heading	Frequency	Rank
Persian Poetry	20772	1
Universities and Colleges	12003	2
English	8255	3
Iran	8040	4
Persian Fictions	7213	5
Mathematics	5391	6
Quran	5086	7
Fiction Stories	4115	8
Persian Language	3379	9
Examination, Graduate Education	3228	10
Animals, Stories and Legends	3175	11
Prayers	3012	12

Heading	Frequency	Rank
Children	2892	13
Education, Secondary	2852	14
Islam	2709	15
Religous Poetry	2647	16
Arabic Language	2639	17
Short Stories, Persian	2550	18
Physics	2540	19
Poetry	2464	20
Persian Literature	2378	21
Distance Education	2303	22
Medicine	2151	23
Social Fictions	2123	24
Painting	2116	25
Science	2058	26
Chemistry	1985	27
Ziyarat-Namah (Pray for visiting Shrines)	1966	28

For example, the first row shows that the first rank belongs to the “Persian Poetry” which is occurred 20,772 times in NBI. The next rows also reveal the rank number of each subject heading and its use frequency.

The correlation analysis was finally performed to determine the relationship between the number of items in NBI and the number of Persian subject headings as well as the rank of each subject heading and its use frequency in NBI.

Results

According to the research questions, results appear in two sections: 1. Production of main subject headings 2. Use distribution of main subject headings:

Production of main subject headings

This section shows how the Persian subject headings have been created and developed.

Figure 1 gives a quick look at the distribution of distinct headings in NBI from the beginning up to the year 2008.

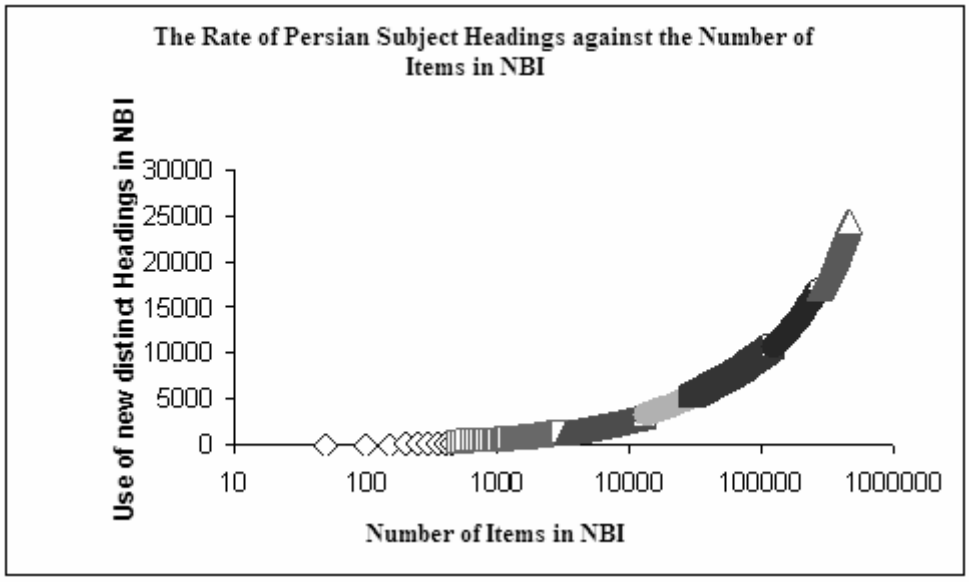

Figure 1: The rate of Persian subject headings against the number of items in NBI: A quick look

In the figure above, the number of items in NBI is plotted against the use rate of new distinct headings in NBI. The x-axis of the curve is scaled logarithmic to illustrate a better shape that discriminates several splits on the figure. Each one represents a particular phase during the creation and the development of PSH.

Breaking the curve into two separate figures sheds light on the detailed facts. The first two phases are displayed together and the other five as one:

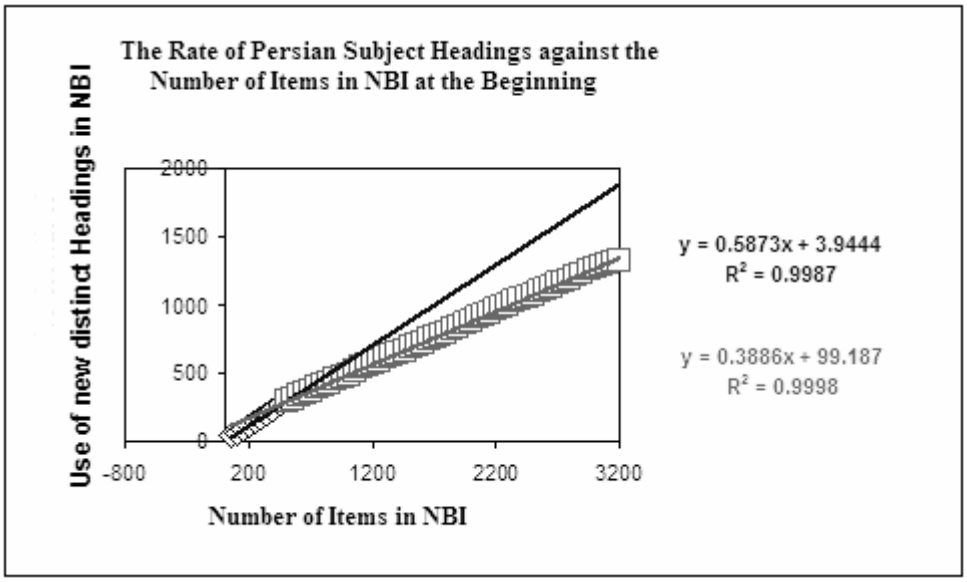

Figure 2: The rate of Persian subject headings against the number of items in NBI at the beginning

Figure 2 is a closer look at the two beginning phases. The fatter lines indicate the observed data and the thinner ones the trend lines. Distribution of new subject headings in

NBI was linear in these two phases. The coefficient of function decreases from 0.59 to 0.39 when the number of catalogued items reaches 450 titles. The linear distribution changes when the number of titles reaches 3,200. This above chunk can be called creation period of Persian subject headings.

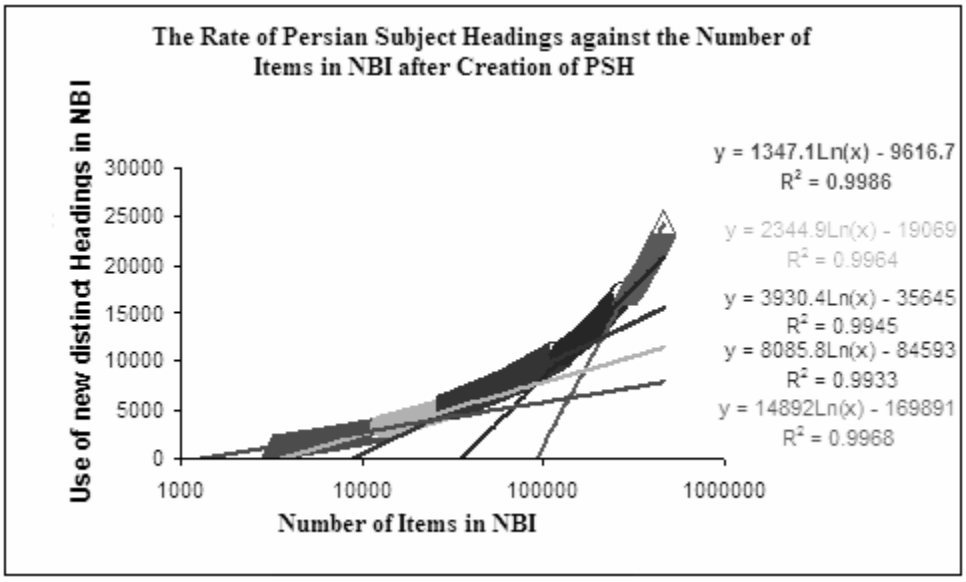

Figure 3: The rate of Persian subject headings against the number of items in NBI after creation of PSH

Figure 3 is a close look at the last five phases of the Figure 1. The fatter lines indicate the observed data and the thinners the trend lines. In addition, the x-axis is scaled logarithmically to illustrate the trend lines directly. The five associated functions in the figure show five similar logarithmic correlations with dissimilar coefficients. This above chunk can be called development period of Persian subject headings.

Use distribution of main subject headings

The frequency distribution of Persian subject headings shows that 24,396 distinct main headings have been used for subject cataloguing of 458,250 items in NBI. Furthermore, 19,715 (80.8%) headings were only assigned fewer than ten times among them 11,370 (46.6%) were used only once.

The correlation analysis was performed on the subject headings with rank number between 1 and 4,680. For performing the analysis, the rank number of each heading was plotted against its use frequency in NBI.

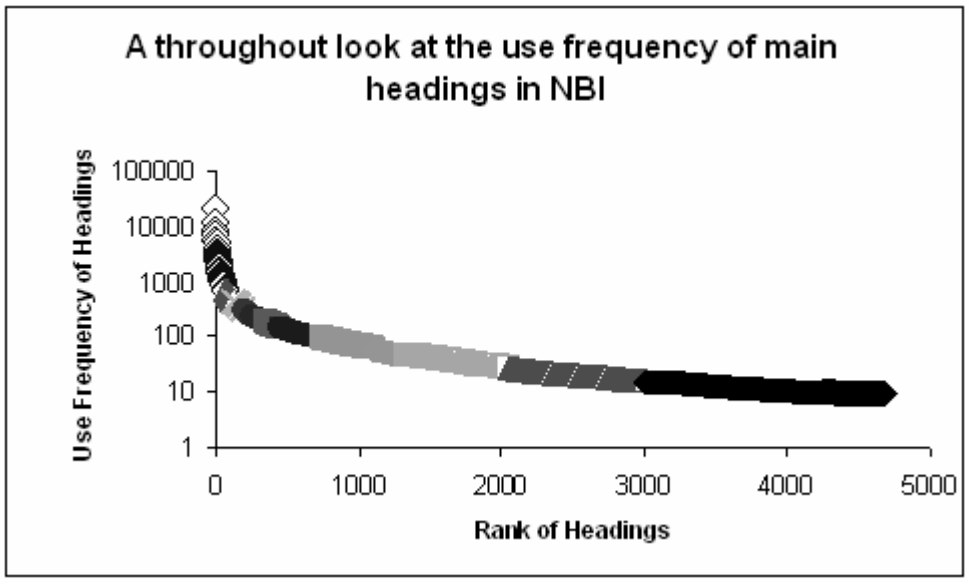

Figure 4: A throughout look at the use frequency of main headings in NBI

In Figure 4, the y-axis is scaled logarithmically to show a better shape of distribution. It includes twelve subclasses. A close look at this figure reveals three kinds of distributions: power-law (log-log) for mostly, linearly for frequently used, and exponentially (normal-log) for normal used headings. Based on these distributions, we can get a precise result by splitting the figure into three others:

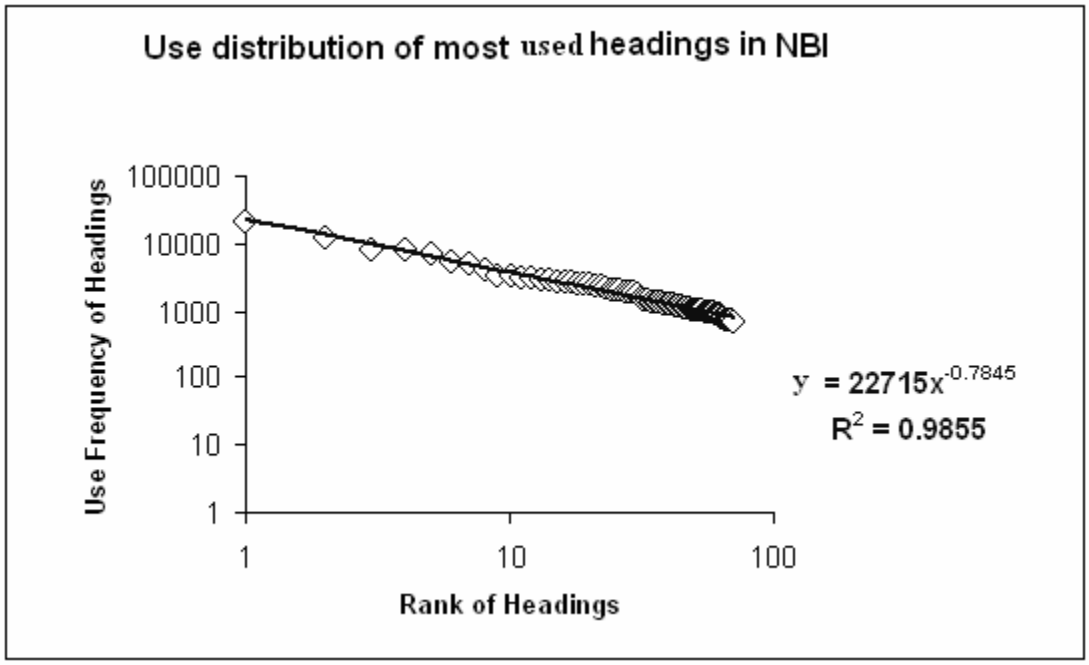

Figure 5: Use distribution of the most used Persian subject headings in NBI

Figure 5 shows the distribution of the most used subject headings. Regarding the power-law (log-log) correlation, both of the two axes scaled logarithmically to obtain a

direct line. It shows that by decreasing the rank number, the use frequency of headings diminishes by a power equal to -0.7845.

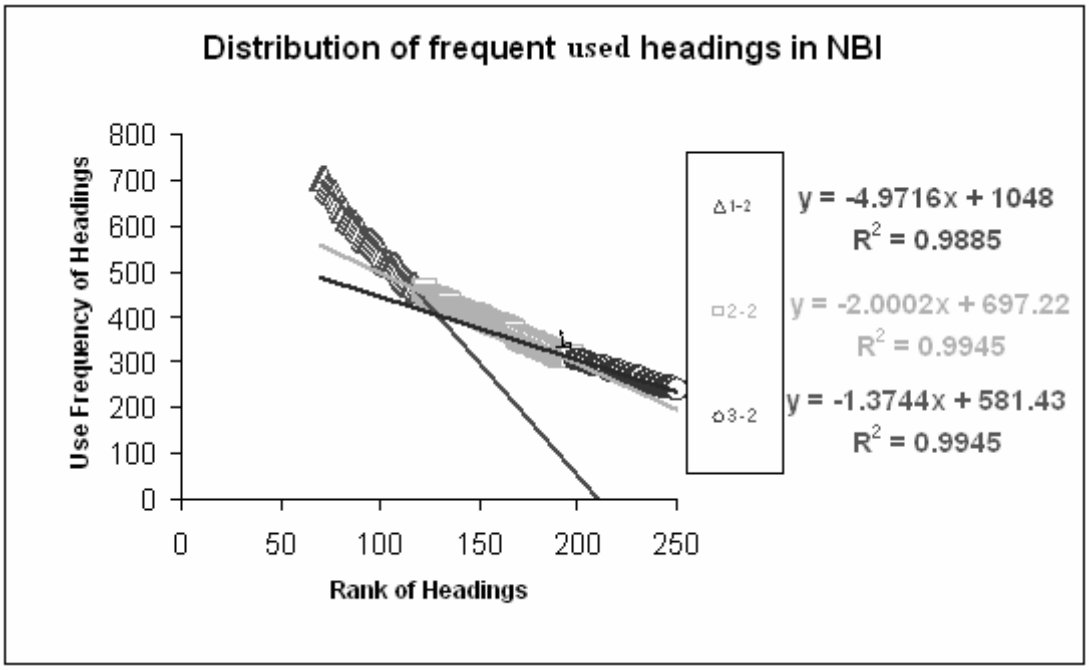

Figure 6: Distribution of frequent used Persian subject headings in NBI

Figure 6 is a cut of those headings with linear distribution. The fatter lines are observed data and the thinner ones represent the trend lines. We can distinguish three different coefficients for this part. The distribution began with a coefficient equals to -4.97. It changed to -2 and followed by -1.37. They present a category of the frequently used headings.

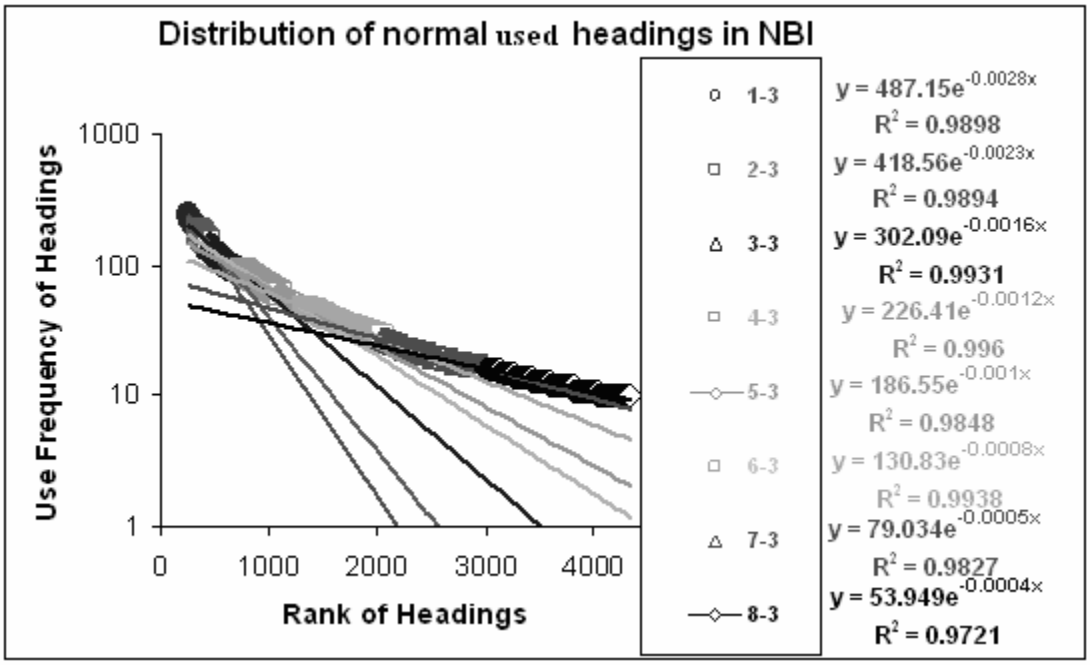

Figure 7: Distribution of normal used Persian subject headings in NBI

The last part follows an exponential (normal-log) distribution. In Figure 7, the y-axis is scaled logarithmically to obtain direct lines. Those fatter lines are based on the observed data and the thinner ones represent the trend lines. This part includes the majority of the Persian subject headings.

Discussion

The results related to the production of Persian subject headings reveal the answer to the first research question. Accordingly, the growth of Persian subject headings during the expansion of NBI resulted in two different kinds of distributions. These can be considered as creation and development periods of Persian subject headings. The logarithmic distribution of headings in development period matches with the Lancaster's prediction in 1986 as he states: "Of course, a vocabulary developed through an actual indexing operation will grow very fast at the first". On the other hand, it disagrees with Wurm's finding in 1964 where he reports: "it is not possible to establish with any accuracy what the relations between file size [number of documents] and term number might be".

The results disagree with the Lancaster as he predicts that the curve "will reach a plateau after X papers have been indexed". The findings show that the logarithmic growth of controlled vocabularies never reaches a plateau. They also agree with Wurm as he states "as the observations made [it] deviate too much from any ideal curve". It means that the resulted curve differs from an accurate one and should be broken into smaller parts.

The second part of results which addresses the use distribution of Persian subject headings ties to the second research question. Based on the correlation analyses, the Persian headings were divided into three classes. The findings showed that the broader terms were used more and were classified in toper categories. The seventy headings on the top, for example, have a role close to the main classes of a classification system. They make it possible to retrieve a certain class of publication materials like "Persian Poetry". Table 2 provides more examples for that category.

A preceding study on the Medical Subject Headings (MeSH) revealed two categories: frequently used with a power-law (log-log) and normal used with an exponential (normal-log) distribution (Tavakolizadeh-Ravari, 2007a: p. 58). Acquiring one more category in this research may be a consequence of some unknown occasions during the development of Persian headings. If we consider their count within the list of the two first categories, we will get a number close to 256 (70 + 183 = 253). Tavakolizadeh-Ravari (2007b) noticed an unexplained number equal to 256 through assembling the findings of his previous work with those done by Price (1963) and Umstätter (1986). He suggested that it is possible to initially divide the subjects into 256 classes. Browsing the category with linear distribution illuminates that the broadness level of its headings is close to those in the prior one. Therefore, unwelcome effects during the development of PSH possibly led to a change in

distribution type. Otherwise, the use distribution of its headings in NBI would actually produce two subject categories: one with a power-law (log-log) with ~256 headings and the other with an exponential (normal-log) distribution.

The correlation analyses on the use distribution of Persian subject headings also reveal that the vast majority of Persian subject headings have been rarely used (19,715 headings out of 24,396). Two main reasons might lead to this condition:

1. The occurred errors while typing data entry in electronic version of NBI (see research limitation).

2. Production of headings that were not actually used by cataloguers. This may be due to absence of policies in NLI to control the utility of headings.

One of the addressed policies is the problem of translating a number of Persian headings. Karimi (2008) studied the development of PSH to answer these questions: whether it has developed by support of new publications along with the consideration of the linguistic characteristics. She claims that these cases were not of the main interest at the beginning and possibly, most headings were translated from Library of Congress Subject Headings (LCSH).

The other problem was studied by Jafargholi-biglu (1998) and Fatahi and Arastoopour (2007). They concentrated on how identical are the Persian Subject headings with the users' and catalogers' minds. The work done by Fatahi and Arastoopour focused on Persian books to see how far the Persian subject headings match with the keywords of “titles” and “tables of contents” in human, social, applied, and pure sciences. The results showed that there is the most harmonious relationship between, the subject headings and the titles (40.5%) and the subject headings and the table of contents (39.7%) in human sciences. The matching level was more deficient for other fields. Creating headings with low level of agreement with catalogers’ mind likely led to a great number of rarely used subjects.

Conclusion

The main objectives of this research were: 1- to find the relationship between the inclusion of new items to NBI and the need for creation of further headings; and, 2- to determine the frequency distribution of Persian subject headings in NBI.

The first objective sought to determine how the list of Persian subject headings was expanded. The findings showed that the creation of this list in proportion to the items in NBI has followed a linear distribution. After creation, it has developed logarithmically. The findings related to this objective agree with Lancaster’s as he expresses that “a vocabulary developed through an actual indexing operation will grow very fast at the first” but are against his prediction as he denotes that “it will reach a plateau after X papers have been indexed”. The splitting of curves in this research matches with Wurm’s finding as he attained that the curves should be broken into smaller parts.

The second objective was concerned with the utility of Persian subject headings. Their frequency distribution in NBI showed that 19,715 Persian subject headings out of 24,396 were rarely used but the use of 4,681 subject headings was considerable. Spink, Amanda et al (2001) found the similar result for the terms that users use in their queries: "Of the 140,279 unique terms, some 57.1% were used only once, 14.5% twice, and 6.7% three times, i.e., some 78.3% of unique terms were used three times or less".

The correlation analysis between their use frequency and their rank revealed three major classes of subject headings: most, frequent, and normal used. This finding disagrees with a related work, which found two classes. Acquiring one more class may be a consequence of some unknown occasions during the development of Persian headings. The rare use of vast majority of subject headings is due to the occurred errors while typing data entry in electronic version of NBI and the absence of policies in NLI to control the utility of headings

Suggestions

In order to avoid the unnecessary or inadequate augmentation of controlled vocabularies, the policymakers should establish efficient mechanisms. The use of statistical analyses allows for objective decision-making for acceptance or rejection of terms. Subjective determination may lead to a list of terms that are not compatible with cataloguers' minds. So, descriptors and subject headings should be included to a controlled vocabulary after joining in a test phase. The length of this period depends on the proliferation pace of indexed documents in the corresponded indexing system. If a created term is assigned to a sufficient amount of documents during that time, it will be accepted as a relevant term for that indexing system. Otherwise, it should be revised. This operation provides a mechanism for controlling the utility of each unique term.

Similarly, the frequency distribution of headings and descriptors allows for controlling the level of their specificities and exhaustivities. As seen before, the number of exhaustive terms is fewer than those with more specificity. To keep this state, the relationship between the use frequency of terms and their ranks should be constantly kept under control. A power-law (log-log) distribution is expected for the most frequent terms and an exponential (normal-log) for the rest.

For keeping the size of controlled vocabularies in an optimum level, the relationship between the growth of indexed items and the expansion of its corresponded vocabulary should be persistently kept under control. The expectation is a logarithmic distribution during the development phase. However, the value of its growth coefficient should not exceed or fall over the time except some reasonable cautions like changes in the indexing policy.

References

Bennett, J.M. (1975). Storage design for information retrieval: Scarrott's conjecture and Zipf's law. International Computing Symposium Amsterdam, 233–237.

Booth, A. D. (1967). A "law" of occurrences for words of low frequency. *Information and Control, 10* (4), 386-393.

Burnett, J.E., et al. (1979). Document retrieval experiments using indexing vocabularies of varying size. I. variety generation symbols assigned to the fronts of index terms. *Journal of Documentation, 35* (3), 197–206.

Cleverdon, C. W. et al. (1966). Factors Determining the Performance of Indexing Systems 1: Design. Granfield, England: College of Aeronautics. Aslib Granfield Research Project.

Fattahi, R. & Arastoupoor, Sh. (2007). Study on matching Persian subject headings with keywords of titles and table of contents in human, social, applied, and pure sciences. *Library Faslnameh Ketabdari va Etela Resani*. *10* (3), 57 – 80.

Fedorowicz, J. (1982). A Zipfian model of an automatic bibliographic system: An application to Medline. *Journal of the American Society for Information Science, 33*, 223–232.

Hajizeynolabedini, M. et al. (2000). Evaluation of national bibliography database. *Faslnameh Etela Resani, 15* (3/4).

Houston, N. & Wall, E. (1964). The distribution of term usage in manipulative indexes. *American Documentation, 15* (2), 105-114.

Jafargholi-Biglu, M. (2003). Comparison on subjects in the mind of information seekers and the structure of Persian subject headings. *Oloome Etela Resani, 14* (3/4).

Karimi, M. (2008). Persian subject headings and library of congress subject headings: Composition or translation? *Majaleh Electronici Ertebate Elmi, 8* (2).

Lancaster, F. W. (1986). *Vocabulary control for information retrieval (Second Edition)*. Washington: Information Resources Press.

Nelson, M.J. (1989). Stochastic models for the distribution of index terms. *Journal of Documentation, 45* (3), 227–237.

Pratt, W. (1999). *Dynamic categorization: A method for decreasing information overload*. (Ph.D. Dissertation). Stanford: Medical Information Sciences, Stanford University.

Price, D. (1963). *Little science big science.* Colombia: Colombia University Press.

Spink, A. et al. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52* (3), 226–234.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*, 331–340.

Tavakolizadeh-Ravari, M. (2007a). *Analysis of the long term dynamics in thesaurus developments and its consequences.* (Ph.D. Dissertation). Germany: Institut für

Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin.

Tavakolizadeh-Ravari, M. (2007b). The growth of medical sciences subjects: A correlation analysis between development of mesh and Medline. *Modiriate Etelaate Salamat, 4* (2), 185-192.

Umstätter, W. (1986). Informetrische hilfen durch das intelligente terminal. *Deutscher Dokumentartag*, 556-564.

Wall, E. (1964). Further implications of the distribution of index term usage. *Proceedings of the American Documentation Institute. 1*, 457-466.

Wolfram, D. (1992). Applying informetric characteristics of databases to ir system file design, part I: Informetric models. *Information Processing & Management, 28*(1), 121–133.

Wurm, B. R. (1964). The relation between number of documents and number of terms and their discriminatory power in information retrieval for U.S. pharmaceutical patents. Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices ICIREPAT, 349-360.

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Cambridge: Addison-Wesley.