A Study on Evaluating Documents Regarding Their Lengths

Mohammad Tavakolizadeh-Ravari*

29 September 2008

Abstract

Due to the proliferation of information, today's users should be able to receive the most contents in a short time. Because of this, the art of writing the scientific documents is to write the most contents with the least words. Despite this fact, this factor has been not considered in evaluating published documents. The number of index terms for a document can be counted as an indicator of the amount of contents that a document bears. The amount of index terms versus the length of documents will prepare the possibility to evaluate the appropriateness of the documents lengths.

The problem in this way is the lack of indicators or standards to evaluate the appropriateness of the documents lengths. To find a norm for the evaluating the length of a document versus its contents, the relatedness between the length of documents and their index terms has been studied.

About one million medical journal articles in English that were indexed in MEDLINE between the years 1965 and 2005 were selected. The articles longer than thirty pages were eliminated. Two parameters of remained articles were noted by a program written in Delphi: 1. number of pages of articles and, 2. the number of MeSH headings of the articles.

The efforts were to find out the average number of MeSH headings for articles considering their lengths. We wanted to see how many MeSH headings had the articles with one page on average, how many with two pages, and so on. The other effort was determining the length of articles with the number of their types and tokens. In this case, nine full text documents with different lengths in page were selected. The number of their types and tokens were counted. The results showed that the relationship between the number of pages and Average number of MeSH headings is always logarithmic. The other result revealed that this function is not applicable for the articles longer than ten pages. It means a journal article longer than ten pages has not necessarily more contents. Here the text of the abstract of your paper has to be filled in. Please note, that the figures, references, and data used below are only dummies which serve as illustrations of formats.

1 Introduction

The exponential growth of information offers a huge number of documents to the users. It makes them difficult finding the appropriate ones to their information needs. Information specialists and systems try to ease this task. One function of abstracts or index-terms, as example, is to provide the users a fast way to scan the texts contents and to evaluate the relatedness between the corresponded texts and their needs. Consequently, the users select those documents that seem to fulfil their needs.

The other problem is the high number of appropriate documents that one retrieves during a search process or the plenty of documents around a subject. It causes that the readers not be able to study all of existing documents. Thus, the art of writing scientific documents is to offer the users the most content with fewer words. It provides the users the opportunity to be able to read more texts in less time.

* Assistant Professor, Yazd University, Department of Library and Information Sciences, Iran (tavakoli at yazduni dot ac dot ir)

There are not common standards for determining the documents length. There are only limits on the number of papers submitted to a conference or to a journal. Other limitation is the number of words used in writing an abstract. The limits don't take other factors into consideration. They show only the maximum number of pages. In the case of abstracts, we see some systems let the certain kinds of abstracts exceed the limits of words to a maximum. NLM, for example, permits the maximum length of abstracts in MEDLINE for records created after the year 2000 up to 10,000 characters. The original policy on inclusion of abstracts set a limit of 250 words for acceptance by NLM. In 1984, two changes were made in the policy: 1) the limit of words was raised to 400 words for articles of more than ten pages in the core journals identified by National Cancer Institute, and 2) abstracts exceeding the 250- or 400-word limit were to be included in truncated form at the end of the sentence closest to the word limit.[1]

However, we can assume that the document length can be affected by the amount of subjects that it contains. It means, by increasing the number of subjects within a document, its length increases as well. If it is so, the number of subjects describing a document determines its length. To test this hypothesis we should study the relatedness between the document length and the number of descriptor that they have received. Descriptors can be counted as the representative of documents' subjects.

2 Method

To test the hypothesis we need to describe the existing relationship between the length of documents and the number of descriptors they have received. The randomly selected records from MEDLINE prepared the possibility to find the mentioned relationship. In the following we will see the steps of selecting records and processing them:

PubMed makes it possible to search in MEDLINE. Using it, two keywords were queried by

---

1   See "MEDLINE®/PubMed® Data Element (Field) Descriptions" in References section.

"AND" operator. They were "HUMAN" and "MEDICINE". The search returned about 1,000,000 records. Besides querying the keywords mentioned above, the search was limited to the "Entrez Date" between the years 1965 and 2005, "Journal Articles" AND "English". Following to save the retrieved records in text format, they were processed by a computer program written in Delphi. The concentration was on two fields of records: "MeSH Headings" (MH) and "Number of Pages". The process was as follows:

1. Check tags were excluded from them. It means they were not considered as real descriptors.
2. The qualifiers (Subheadings) were also excluded.
3. Articles longer than thirty pages were eliminated from the research.
4. The rest were divided into thirty groups based on their number of pages. It means, documents with one page were inserted in group one, with two pages in group two, …, and with thirty pages in group thirty.
5. The total MeSH Headings received the documents in every group was determined and divided by the number of documents of that group. This process helped to determine the average number of descriptors that documents with different lengths received.

As the length can be determined by the number of words of a text, other concentration was on their tokens and types. In this case, nine full-text articles of different lengths were downloaded and their tokens and types were determined.

3 Results

The first effort was to determine the relationship between the number of pages of documents and number of their descriptors. It will show us if there is any relationship between these two variables. Figure 1 reveals the relationship between them.

The figure shows there is a logarithmic relationship between the two variables only if the lengths of journal articles don't exceed from 10 pages. The number of descriptors received by

articles longer than ten pages remains around nine. The cutting point states that the least descriptors received by articles in an indexing system with a level of in-depth indexing like MEDLINE are four. It means articles should cover at least four subjects if the specificity of terms within corresponded controlled vocabularies is like MeSH.
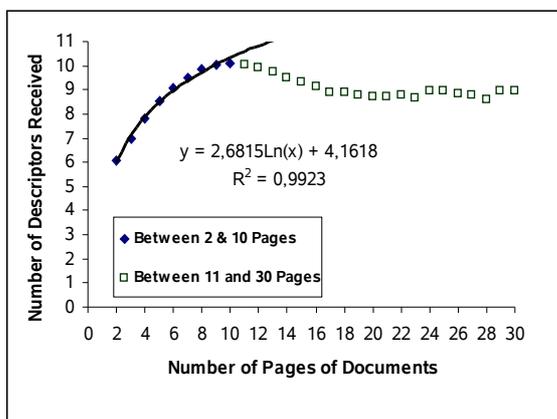

Figure 1: Relationship between number of pages of documents and the number of index terms they have received.

Table 1: Reduction of the types per page rate in relation to larger articles.

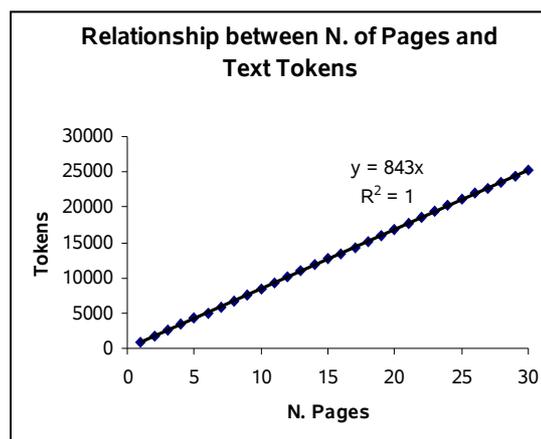| N. Pages	Tokens	Types	Types per Page
1	843	426	426
2	1686	635	318
3	2529	803	268
4	3372	948	237
5	4215	1078	216
6	5058	1197	200
7	5901	1309	187
8	6744	1413	177
9	7587	1513	168
10	8430	1608	161
11	9273	1698	154
12	10116	1786	149
13	10959	1870	144
14	11802	1952	139
15	12645	2031	135
16	13488	2108	132
17	14331	2183	128
18	15174	2256	125
19	16017	2328	123
20	16860	2398	120
21	17703	2466	117
22	18546	2533	115
23	19389	2599	113
24	20232	2663	111
25	21075	2727	109
26	21918	2789	107
27	22761	2851	106
28	23604	2911	104
29	24447	2971	102
30	25290	3029	101

Figure 2: Relationship between the number of pages and text tokens.
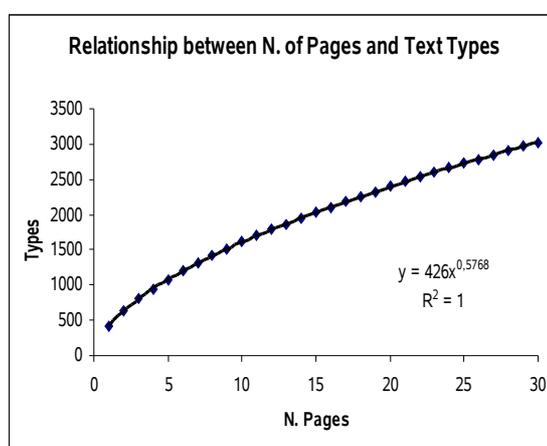

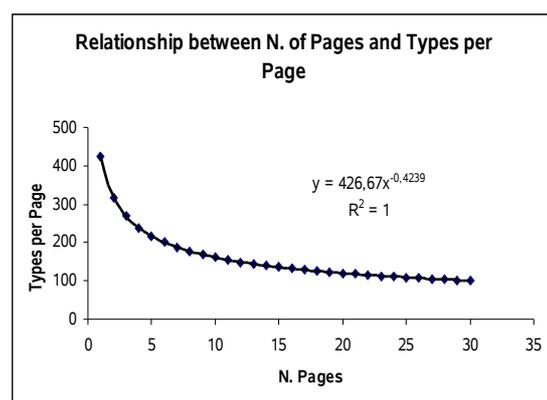Figure 3: Relationship between the number of pages and the text types.

Figure 4: Relationship between the number of pages and types per page.

To find out why such a relationship exists between length of articles and their received descriptors we can refer to the number of text words and vocabularies used for creating that text. We call the number of words within a text "tokens" and vocabularies used within it as its "types". Besides we should take the relationship between tokens and types with the number of pages. The data in Table 1 is the result of analysis nine full text articles in the field of medicine.

Determining the correlation between the data in the table above give us a best sight to know why there is such a relationship in Figure 1.

As the Figure 2 shows, it is a linear relationship between the number of pages of articles and their tokens. It means by enlarging a text one page, 843 words will add to it on average.

On other hand, the relationship between the number of pages of an article and its types is power law (Figure 3). It means, in an article with one page one half of words are types on average. But in a one with thirty pages, the number of types is only one tenth of tokens. This is due to exponent of function that is ~0.58. Figure 4 shows this fact better.

4 Discussion

The effect of the text length on the several variables has been studied by researchers like Anderson, M. D. (1971) and Wellisch, H. H. (1991) on the size of book index, Abt, H. A. and Garfield, E. (2002) on the number of references of an article, Tavakolizadeh-Ravari, M (2007) on the depth of indexing. But their emphasis is not on how evaluate the appropriateness of the documents lengths.

 Is there any way to see if the length of a document is appropriate to the amount of contents and information that a text is going to present? Is its length appropriate or it could be fewer? Results of the current research show how the amount of subjects within a text can relate to the appropriateness of its length. The data and numbers yielded in this research can't be generated to evaluate every kind of documents. Thus, how many subjects an article with a certain length should have relates to the indexing method and the specificity of the terms used for its indexing.

Thus, the current results yield a norm for evaluating the appropriateness of the lengths of articles indexed in an indexing system such MEDLINE with a controlled vocabulary such MeSH in special and for documents indexed in other indexing systems in general. They tell us if a document receives fewer subjects than a norm, they can't be counted as a good one at all. In our example, the least expected subjects for a medical journal article is four (except for check tags). Enlarging documents without any increase in the amount of their subjects makes them poor ones from the current point of view. From this point of view, adding one more page to a text should increase the number of subjects logarithmic.

Relationship of tokens and types can answer why such a relationship exists between the document length and its index terms. Being more types in a text increase the chance of emerging new subjects. For example a text with 1000 tokens having 500 types has more probability of having content bearing words than a text with the same amount of tokens but fewer types. In other words, increasing the lengths of documents causes an increase in text tokens, and therefore an increase of tokens brings an increase in types. On the other hand, the types per page decrease gradually through the enlargement of articles. As a result of this gradual reduction, when the lengths of articles grow, the function of the relationship between number of pages of articles and average number of their index terms becomes logarithmic.

We saw that the index terms of articles longer than ten pages remained around nine. It means such articles have not necessarily more subject. They may only have more figures or tables or other explainable features.

References

Berberich, K., S. Bedathur, M. Vazirgiannis, and G. Weikum (2006). Buzzrank … and the trend is your friend. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, pp. 937–938. ACM Press.

Abt, Helmut A. and Garfield, Eugene (2002). Is the Relationship between Numbers of References and Paper lengths the Same for all Sciences? *Journal of the American Society for Information Science and Technology*. 53(13): 1106– 1112.

Anderson, M. D. (1971). *Cambridge Authors' and Printers' Guides: Book indexing.* Cambridge, UK: At the University Press.

MEDLINE®/PubMed® Data Element (Field) Descriptions, National Institutes of Health, National Library of Medicine. [WWW Document] http://www.nlm.nih.gov/bsd/mms/medlineelements.html. 28.05.2008.

Tavakolizadeh-Ravari, M. (2007). *Analysis of the Long Term Dynamics in Thesaurus Developments and its Consequences.* Ph.D. Thesis, Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft, Berlin, Germany.

Wellisch, Hans H. (1991). *Indexing from A to Z.* New York: The H. W. Wilson Company.