Proceedings of Two-Day National Seminar on
"ICT-Enabled User Driven Library Services:
Issues and Challenges"
Editor: Dr. Atashi Karpha

Published by Gokhale Memorial Girls' College
1/1 Harish Mukherjee Road
Kolkata 700 020, W.B.
on 23/12/2022

ISBN : 978-93-5768-438-5

Type Setting & Printing by Microprint Graphics
1A, Raja Lane, Kolkata 700 009, W.B.

Necessity of Caption Analysis for Providing Better Services to Scientific Community: A Case Study from Virology

Sri Debabrata Maity

Librarian, Khejuri College, Baratala, Khejuri, West Bengal-721431
Email: maitydebabrata6@gmail.com

Abstracts:

Designing of information retrieval system is a major technical work to provide relevant information to the information users. Though several databases use text mining technology or keyword-based techniques for abstract or text retrieval but some studies have shown the importance of figure's caption search and its parallel retrieval for better understanding of article's contents, especially in biological sciences. In the paper, using the Web of Science database, the efficiency of the processes of assigning the weightvalue to the technical terms is tested. Finally discussion has made about the importance of analyzing the captions of the objects presented in research articles for providing better services to scientific community.

Keywords: *Caption analysis; Caption analysis of figures; Caption analysis of objects; Keyword searching; Image retrieval; Weight assigning of keywords; Virology*

1. Introduction :

Development in the field of information and communication technology (ICT) has converted the library system from traditional phase to modern phase including electronic and digital library systems.Major parts of the library resources are converted from printed media to electronic and digital media including e-books, e-journals, e-notes, e-theses and dissertations, online databases, models, graphics, games, quizzes, animations, and so forth. As a result, to serve user community properly, findings of necessary documents from huge amount of resources, especially from electronic and digital resources, it is always necessary to design good information retrieval systems.

For experimental work or to stay update with the new knowledge,researchers, scientists or practitionersalways need micro documents. Ranganathan (as cited in Kumar, 1988) defined micro document as "an article in a periodical or the part of a book, not having an independent physical existence" (p. 18). Day by day number of micro document is increasing rapidly.As per specific need, retrieval of proper micro document especially in digital environment is a challenging task.

Many scientific documents include not only text but also non-textual elements like tables, photographs, graphical plots, charts, maps, drawings, figures, and so forth; here collectively called objects.Sanyal, Chattopadhyay, &Chatterjee(2019) mentioned that an object "along with its metadata - caption, figure text (i.e., text embedded on a figure), mentions (i.e., references to the figure from the full-text) - and the metadata of the original article contain a wealth of information that are otherwise buried deep with the article." They also wrote "browsing through the figures in an article or retrieving suitable figures from a journal database can quickly help a researcher or practitioner . . . " to form an idea of experiments and results in his/her domain of study. So object retrieval "is an invaluable aid to navigate and analyze the gargantuan volume of biomedical literature" (Sanyal et al., 2019).

2. Literature Review :

Yeh, Hirschman, and Morgan in the 2002 KDD competition found that figure captions can be helpful to the researchers for locating information about experimental results (2003).Divoli, Wooldridge, and Hearst (2010) observed that when researchers reading bioscience articles, it is a common trend "to start by looking at the title, abstract, figures, and captions" (p. 2).Scientific literature can be innovatively analyzed with the help of figure databases (Lee, West, & Howe, 2018).A recent study also noted that figures in an article quickly help researchers to find out whether the article is in their field of interest or not (Sanyal et al., 2019).

Information retrieval (IR) simply means recovery of necessary information embedded in documents. Two main approaches of IR technique are "matching words in the query against the database index (keyword searching) and traversing the database using hypertext or hypermedia links"(Gregersen, 2022).

Luhn (1957) proposed the idea to analyze the subject content of a document through automatic counting of terms occurrence within the document.The root concept of the theory was that "the more frequent the occurrence of a term in a given document, the more significant is that term in denoting the subject content of the document" (Chowdhury, 2010, p.120).So, within a given document, assigning weight to a particular technical term depends on the number of times of its occurrence. There are many search engines work in web-based environment, ranging from commonly used Google Scholar, includes scholarly articles from various disciplines, to domain-specific search engine like PubMed, provides search facility to biomedical literature only. Some of the search engines allow to search through title, abstract, and keywords of the paper; while some of them allow searching through full-text. For retrieval, keyword-based search techniques or semantic search techniques are used to compare the query strings against the text of the documents (Sanyal et al., 2019).PubMed provides full-text access to the articles only in the PubMed Central and using keyword-based search technique it not only shows research articles but also provides a thumbnail view of objects associated with each article.Except frequency of occurrence, assigning weight to a technical term also depends on its position of occurrence in the document.According to Shah, Perez-Iratxeta, Bork, and Andrade (2003), in case of bioscience literature, it is necessary to consider in which section of the article query terms found in. Here, the research question is that whether the above mentioned two process of assigning weight to the technical terms are sufficient or not?

3. Objectives of the Study :

The main objectives of the present study are:

- To verify the appropriateness of judging the importance of object keywordsas per the number of times they occur in the textual part of the article;
- To find out how far the object's caption analyzed keywords or object keywords match with the keywords derived from the titles, and abstracts, and with the author assigned keywords;
- To test whether the object keywords thatmatch with thekeywords derived from the titles, and abstracts or with the author assigned keywords have always high importance than comparing to unmatched keywords;and
- Making an interpretation about the necessity of caption analysis in the field of virology.

4. Methodology :

Major parts of the methodology, especially as written in the first two paragraphs are borrowed from the author's previous paper (Maity &Dutta, 2022). At first, 20 top-cited research articles (except review article, report, commentary, and etc.) in the field of virology were collected as sample size from the Web of Science database. During the search process the term Virology was putted within double inverted comma in the search box and the time span of searching was fixed for 1980-2014. Criterion was set to show the retrieved results as per relevance. Article must had at least one object (may include any of the non-textual elements like table, diagram, figure, chart, photograph, map and so forth) with proper caption and also had abstract and author assigned keywords was considered as sample. Then following steps were done:

Captions of the objects of the articles were analyzed to cull out keywords. Ifa single keyword occurredmore than one time within the same caption orin more than one object's caption within the same article, then it waslisted for single time only; the same was listedfor second time if occurred in any object's caption ofa separate article except the previous one;listed third time if occurred in the object's caption of an article except the first two,and the process continued.In all, 1566 keywords were derived from the captions of the objects of the articles. Secondly, titles and abstracts of the sampled articles were analyzed, which produced 83 keywords from the titles, and 402 keywords fromthe abstracts. A number of 97 author assigned keywords were also found in total.The phrase "object's caption analyzed keywords" was used for the keywords derived from the captions of the objects of the articles;whereas "title analyzed keywords" for keywords derived from titles; and "abstract analyzed keywords" for keywords derived from abstracts.Four important acronyms used in the study are OCAK for Object's Caption Analyzed Keyword; TAK for Title Analyzed Keyword; AAK for Abstract Analyzed Keyword; and AuAK for Author Assigned Keyword.

Usingthe same database and the same searching process as mentioned earlier, searching process was done again against each keyword to collect data about no. of result retrieved, total citation, availability of citation report. If the number of retrieved result for a particular OCAK was twenty or more than that, then the remarkwas fixed that the OCAK has high number of retrieved result.

Number of timeseach OCAK occurs within the textural part of the concerned article was counted.Verification process was also executed regarding the occurrence of each OCAK in the title, abstract, and AuAK sections of the concerned article. Finally tabulation and data analysis works were carried out.

5. Results :

Table 1 : below provides a clear picture about the types of keywords used in the present study.

Table 1

Total Number of Keywords used for study

Category of Keywords	TAK	AAK	AuAK	OCAK
Number of Keywords	83	402	97	1566

Distribution of retrieved result according to the frequency of occurrence of OCAKs only in the textual part of the concerned articles is represented in Table 2. From the table it is found that though 503 (32.12%) OCAKs don't occur in the textual part, yet more than half of them that is, 341 (21.78%) have high number of retrieved results. A total no. of 667 OCAKs has very low no. of occurrence (1-10), within it 506 (32.31%) carry high number of retrieved results. The OCAKs which individually occur more than 30 times always carry high number of retrieved results.

Table 2

Distribution of Retrieved Result According to the Frequency of Occurrence of OCAK

Frequency of Occurrence of OCAKsOnly in the Textual Part	Number of OCAKs				Total Number of OCAKs (%)
	For No. Result Retrieved	For Result Retrieved <20	For Result Retrieved =20	For Result Retrieved =20 and Citation Report Not Available	
0	89 (5.68)	73 (4.66)	281 (17.95)	60 (3.83)	503 (32.12)
1-10	97 (6.19)	64 (4.09)	328 (20.94)	178 (11.37)	667 (42.59)
11-20	5 (0.32)	20 (1.28)	57 (3.64)	84 (5.36)	166 (10.60)
21-30	0	2 (0.13)	10 (0.64)	24 (1.53)	36 (2.30)
31-40	0	0	21 (1.34)	50 (3.19)	71 (4.53)
41-50	0	0	23 (1.47)	13 (0.83)	36 (2.30)
>50	0	0	59 (3.77)	28 (1.79)	87 (5.56)
Total	191 (12.19)	159 (10.16)	779 (49.75)	437 (27.90)	1566 (100.00)

Table 3 shows the number of retrieved results against the matching of each OCAK, as it matches with the keywords derived from or used in the title, abstract, and author assigned keywords sections of the concerned article. It is found from the table that 31.16% of OCAKs match at least one of with TAK, AAK and AuAK; while 68.84% does not match with any. The amount of OCAKs match only with the TAKs is 0.89%; with the AAKs, 20.44%; and with the AuAKs, 0.64%.There always ?20results retrievedfor each OCAK that appears in any two or all three sections including the title, abstract, and author assigned keywords.Important to mention that 761 (48.60%) OCAKs that don't appear in anywhere of the title, abstract, or author assigned keywords section also carry high number (?20) of retrieved results.Within the 761 OCAKs, 256 (16.35%) have very high number of retrieved results and the citation report of total retrieved result for such amount of keywords is unavailable in the Web of Science database.

Table 3

Distribution of Retrieved Result According to the Matching of OCAKs

Matching of OCAKs with the TAKs, AAKs and AuAKs	Total Frequency (%)	For No Result Retrieved (%)	For Result Retrieved <20 (%)	For Result Retrieved =20 (%)	For Result Retrieved =20 and Citation Report Not Available (%)
Only with TAK	14 (0.89)	0	2 (0.13)	9 (0.57)	3 (0.19)
Only with AAK	320 (20.44)	14 (0.89)	15 (0.96)	177 (11.30)	114 (7.28)
Only with AuAK	10 (0.64)	1 (0.06)	1 (0.06)	6 (0.38)	2 (0.13)
With TAK and AAK	81 (5.17)	0	0	45 (2.87)	36 (2.30)
With AAK and AuAK	34 (2.17)	0	0	19 (1.21)	15 (0.96)
With TAK and AuAK	2 (0.13)	0	0	1 (0.06)	1 (0.06)
With TAK, AAK and AuAK	27 (1.72)	0	0	17 (1.09)	10 (0.64)
Total (Number of OCAK Match at Least One of with TAK, AAK and AuAK)	488 (31.16)	15 (0.96)	18 (1.15%)	274 (17.49)	181 (11.56)
Number of Unmatched Keywords	1078 (68.84)	176 (11.24)	141 (9.00)	505 (32.25)	256 (16.35)
Total	1566 (100.00)	191 (12.20)	159 (10.15)	779 (49.74)	437 (27.91)

6. Discussion :

Major findings of the present study are:

- The OCAKs that occur more than 30 times in the textual part always carry high number of retrieved results. High retrieved results also reflected for the OCAKs that don't occur in the textual part, 21.78%; occur with very less frequency (1-10), 32.31%;
- Only 31.16% of OCAKs matches at least one of with TAK, AAK and AuAK; while 68.84% does not match with any;

- The OCAKs that appear in any two or all three sections including the title, abstract, and author assigned keywords have always high number of retrieved result. Oppositely 48.60% of OCAKs that doesn't appear in anywhere of the title, abstract, or author assigned keywords section also has high number of retrieved results.

In information retrieval system, a document is indexed under some technical terms. These technical terms represent the subject(s) or topics embedded in the document. In automatic indexing system documents are retrieved according to the matching of query terms with the index terms of documents. Index terms occurs in the title, abstract, author assigned keyword sections or even in the full-text where searching possible. In case of web-based search environment or any IR environment "the objective should be to retrieved items in a ranked order, with those that best match at the top of the list. One way of achieving this might be to apply some sort of weighting to the index terms"(Chowdhury, 2010, p.121). According to Shah et al. (as cited in Divoli et al., 2010), for full-text journal articles, "there is evidence that bioscience literature ranking should consider which section of an article the query terms are found in, and assign different weights to different sections for different query types" (p. 1).

As mentioned earlier that in case of bioscience, browsing through the objects in a research article helps researchers to quickly identify about the necessity of the article. The same condition is applicable to virology also because it is not a separate field, but one of the important branches of modern day bioscience. In the field of bioscience including biomedical domain, in addition to full-text databases, there also exist specialized figure search engines. In the present study, usefulness of technical terms is judged by their capacity of retrieving research articles in the web environment. The more it retrieves, the high its usefulness. From the results of the study it is clear that the two process of assigning weight to the technical terms-by counting the frequency of occurrence, and by identifying the position of occurrence in a document are moderately true. Except the high number of occurrence, or position of occurrence in title, abstract, or author assigned keywords section a term may have same importance.In case of image/object searching from any full-text database or from other kinds of databases and repositories, if images/objects are retrieved according to the matching or occurrence of query terms in the captions, then there will be huge no. of retrieval, precision will be less. That is also a matter of time consuming and puzzling. If the terms occur in the caption of an object are weighted according to their magnitude of bearing the central concept of the object, then precision will be more. So when storing documentsespecially in scholarly databases,if there are some mechanisms to analyze the captions of the objects and to assign the weight value to each technical terms appearing in the caption of an object as per their magnitude of bearing the central concept of the object, then the indexing of objects will be better, which will be eventually helpful for providing better services to the scientific community.

7. Conclusion :

This study proves that the methods of assigning the weight value to the technical terms according to their frequency of occurrence and position of occurrence are always not appropriate. It suggests for assigningthe weight to the technical terms appearing in objects' captions as per theirmagnitude of bearing the central concept of the concerned objects. For keyword based searching, if query term(s) match with the terms of the captions of objects then during retrieval the rankingof the retrieved objects should also be as per the assigned weight value. Provision of hyperlink to the full-text from retrieved objects, where applicable, can easily navigate a researcher to the necessary documents.

References :

Chowdhury, G. G. (2010). Introduction to modern information retrieval (3rd ed.). London: Facet Publishing.

Divoli, A., Wooldridge, M. A., & Hearst, M. A. (2010). Full text and figure display improves bioscience literature search. PLoS ONE, 5(4), 1-15. doi: 10.1371/journal.pone.0009619

Gregersen, E. (2022). Information Retrieval.In Encyclopaedia Britannica. Retrieved from https://www.britannica.com/technology/information-retrieval

Kumar, K. (1988). Theory of classification (4th rev. ed.). New Delhi: Vikas Publishing House.

Lee, P., West, J. D., & Howe, B. (2018).Viziometrics: Analyzing visual information in the scientific literature. IEEE Transactions on Big Data, 4(1), 117-129.

Luhn, H. P. (1957). A statistical approach to the mechanical encoding and searching of literary information.IBM Journal of Research and Development, 1(4), 309-317.

Maity, D., &Dutta, B. (2022). In search of domain specific standard categories: A case study from cell biology. Journal of Indian Library Association, 58(1), 29-43.

Sanyal, D. K., Chattopadhyay, S., &Chatterjee, R. (2019). Figure retrieval from bioscience literature: An overview of techniques, tools and challenges. Article submitted for publication, School of Computer Engineering, Kalinga Institute of Industrial Technology.

Shah, P. K., Perez-Iratxeta, C., Bork, P., Andrade, M. A. (2003). Information extraction from full text scientific articles: Where are the keywords?.BMC Bioinformatics 4, doi: https://doi.org/10.1186/1471-2105-4-20

Yeh, A. S., Hirschman, L., & Morgan, A. A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup. Bioinformatrics, 19(Suppl 1).i331-i339.