An Objective Keyword Selection of Papers Using Phrase Frequency

Sivamani M[1], Syamala R[2]
[1]The Librarian (SS)
Vellalar College for Women (Autonomous)
Erode-12. India
sivavcw68@gmail.com

[2]The Librarian
Sri Ramakrishna College of Arts and Science for Women
Coimbatore. India

ABSTRACT: *In this work, we investigated the use of naturally occurring words and phrases and compared them with the keywords assigned by the authors. Normally, authors generate keywords of papers with a subjective view of the essential words and concepts treated in the paper. However, how they are correlated with the highly used words and phrases in the text is a question to be investigated. This work found no correlation between the author-assigned keywords and highly occurring words in the texts.*

Keywords: Phases and Words, Phrase Frequency, Keyword Selection

Received: 30 August 2023, Revised 17 September 2023, Accepted October 2023

DOI: http://doi.org/10.6025/stm/2023/4/93-97

1. Preliminaries

Words are the formal descriptions of the language used to express the content of scientific papers. These words are used not only to represent concepts but also to access and retrieve content. Among the words, the content-bearing words signal the key concepts associated with a document, and they are called key terms. These keywords are part of the papers and are normally generated by authors and supplemented by database producers. The critical question is how these words effectively reflect the content of documents. The better way is to analyse the contents and match them with the selected keywords.

The text words are subjected in this work to frequency count to identify the important core words or subject-bearing words. We counted not just words but also phrases and their frequency. Phrases normally consist of two words and a maximum of five to six words. These sets of words normally indicate the context of the words used. Thus, phrases offer more meaning than words. The phrase is a key element in the text analysis.

We counted the phrase frequency in natural language text. Highly occurring phrases are identified. These identified key phrases are matched with the keywords assigned by the authors.

Naturally existing words and author-assigned words
Author-assigned keywords are based on subjective decisions, whereas the highly used words of texts are naturally available.

2. Early work

One of the most commonly used unstructured pieces of data is text. As a rule of thumb, the primary purpose of text mining is transforming text into structured data. Phrase frequency effects show that the language processing system is sensitive to multi-word language units. (Jacobs) Liu et al investigated one promising paradigm for representing unstructured text, that is, through automatically identifying high-quality phrases from innumerable documents (Jialu Liu, Jingbo Shang, Jiawei Han). Wu et al proposed a novel phrase-based text representation method that considers the integrity of semantic units and utilizes vectors to represent the similarity relationship between texts. However, no study compared words and phrases and the keywords versus phrases.

3. Methods and Materials

We took the full text of selected papers, including title, abstract, keywords and text. The authors assign the keywords. These keywords are expected to express the content of papers. We keep the keywords as the test set one. We then subject the full text to phrase analysis. The natural language parsers identify the frequently occurring phrases and order them with a maximum of eight-word phrases to two-word phrases. These eight-word phrases are then ranked based on their frequency. In the phrases, we eliminate the stop words, and the resulting phrase is the content-bearing words. After the list of eight-word phrases, we proceed to the two-word phrases. In all the stages, we remove stop words, and the result is the content expression keywords.

4. Case Study

Below, we present the case study results we took for this work.

Source Paper

- **Title:**

Successes and Challenges: Inhaled Treatment Approaches Using Magnetic Nanoparticles in Cystic Fibrosis

- **Abstract:**

Magnetic nanoparticles have been largely applied to increase the efficacy of antibiotics due to passive accumulation provided by enhancing permeability and retention, which is essential for the treatment of lung infections. Recurring lung infections such as in the life-shortening genetic disease cystic fibrosis (CF) are a major problem. The recent advent of the CF modulator drug ivacaftor, alone or in combination with lumacaftor or tezacaftor, has enabled systemic treatment of the majority of patients. Magnetic nanoparticles (MNPs) show unique properties such as biocompatibility and biodegradability as well as magnetic and heat-medicated characteristics. These properties make them suitable to be used as drug carriers and hyperthermia-based agents. Hyperthermia is a promising approach for the thermal activation therapy of several diseases, including pulmonary diseases. The benefits of delivering CF drugs via inhalation using MNPs as drug carriers afford application of sufficient therapeutic dosages directly to the primary target site, while avoiding potential suboptimal pharmacokinetics/ pharmacodynamics and minimizing the risks of systemic toxicity. This review explores the multidisciplinary approach of using MNPs as vehicles of drug delivery. Additionally, we highlight advantages such as increased drug concentration at disease site, minimized drug loss and the possibility of specific cell targeting, while addressing major challenges for this emerging field.

- **Keywords:**

Cystic Fibrosis, Magnetic Nano Particles, Ivacaftor, CFTR Modulator, Gene Therapy, Pulmonary, Non-Viral Gene Delivery

5. Analysis

Some top phrases containing 8 words (without punctuation marks)	Occurrences
challenges inhaled treatment approaches using magnetic nanoparticles in	8
treatment approaches using magnetic nanoparticles in cystic fibrosis	8
and challenges inhaled treatment approaches using magnetic nanoparticles	8
inhaled treatment approaches using magnetic nanoparticles in cystic	8
successes and challenges inhaled treatment approaches using magnetic	8

Some top phrases containing 7 words (without punctuation marks)	Occurrences
and challenges inhaled treatment approaches using magnetic	8
approaches using magnetic nanoparticles in cystic fibrosis	8
challenges inhaled treatment approaches using magnetic nanoparticles	8
inhaled treatment approaches using magnetic nanoparticles in	8
treatment approaches using magnetic nanoparticles in cystic	8
successes and challenges inhaled treatment approaches using	8

Some top phrases containing 6 words (without punctuation marks)	Occurrences
sing magnetic nanoparticles in cystic fibrosis	8
challenges inhaled treatment approaches using magnetic	8
treatment approaches using magnetic nanoparticles in	8
inhaled treatment approaches using magnetic nanoparticles	8
successes and challenges inhaled treatment approaches	8
approaches using magnetic nanoparticles in cystic	8
and challenges inhaled treatment approaches using	8

Some top phrases containing 5 words (without punctuation marks)	Occurrences
approaches using magnetic nanoparticles in	8
using magnetic nanoparticles in cystic	8
inhaled treatment approaches using magnetic	8
and challenges inhaled treatment approaches	8
magnetic nanoparticles in cystic fibrosis	8
successes and challenges inhaled treatment	8
challenges inhaled treatment approaches using	8
treatment approaches using magnetic nanoparticles	8
cystic fibrosis transmembrane conductance regulator	6
for the treatment of lung	4

Some top phrases containing 4 words (without punctuation marks)	Occurrences
patients with cystic fibrosis	10
in patients with cystic	10
nanoparticles in cystic fibrosis	9
approaches using magnetic nanoparticles	8
challenges inhaled treatment approaches	8
and challenges inhaled treatment	8
successes and challenges inhaled	8
treatment approaches using magnetic	8
magnetic nanoparticles in cystic	8
inhaled treatment approaches using	8
using magnetic nanoparticles in	8

6. Results

The keywords given by the authors and available in the high frequent words if are similar, then they are marked in bold letters under the keywords list.

1. High-occurrence Keyphrases

- The more frequently used key phrases are presented below.
- Inhaled treatment approaches using magnetic
- Magnetic Nanoparticles in cystic fibrosis
- Cystic fibrosis
- Cystic fibrosis transmembrane conductance regulator
- Treatment of lung
- Patients with cystic fibrosis

2. Key phrases identified (but not used by authors as keywords)

- Authors do not use the following high-occurring keywords in the text as keywords. This is a significant issue we would like to emphasise here.
- Patients with cystic fibrosis
- Treatment of lung
- Transmembrane conductance regulator

7. Conclusion

We found that authors do not list highly used phrases in the text as keywords. Thus, the high-occurrence key phrases differ from the author's keywords. The Keyword generation is subjective. We conclude that the naturally occurring key phrases may express the content well.

References

[1] Cassandra L. Jacobs., Gary S. Dell. Colin Bannard. Phrase frequency effects in free recall: Evidence for re-integration,December 2017. *Journal of Memory and Language* 97:1-16. DOI. 10.1016/j.jml.2017.07.003

[2] Jialu Liu, Jingbo Shang, Jiawei Han Phrase Mining from Massive Text and Its Applications Springer International Publishing. June 2022. 97 p. 9783031019104, 3031019105

[3] Yongliang Wu, Shuliang Zhao, Wenbin Li. Phrase2Vec: Phrase embedding based on parsing, *Information Sciences*, Volume 517, May 2020, Pages 100-127