IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

A Review on Knowledge Map Visualization Using Co-Word Analysis

Utkarsh Malkoti[1], Vidhi Jain[2]

Student, Computer Science and Engineering, Inderprastha Engineering College, Delhi, India[1]

Student, Computer Science and Engineering, Inderprastha Engineering College, Delhi, India[2]

Abstract: Sociologie de 1'Innovation of the Ecole Nationale Superieure des Mines of Paris and the CNRS (Centre National de la Recherche Scientifique) of France and previously it was called "LEXIMAPPE". For decades, scientists struggled to map the interrelations in a subject to realize an effective and efficient plan structure of research. Hence, to resolve this problem some quantitative methods were developed like co-citation analysis, co-word analysis, and co-nomination analysis. Co-word analysis is a bibliometric technique that measures the co-occurrences of keywords to examine the content in the textual data. This technique has proved to be apowerful tool among the scientists to quantify and visualize the relationships between the various subject areas within the corpus. This paper aims to outline a timeline review on development of co-word analysis and discuss the issues of the researches. It summarizes the current state of knowledge of the topic to create an understanding of the topic for the researchers by discussing the findings presented in recent research papers.

Keywords: Co-word analysis, Social Network Analysis, Co-occurrence Matrix, Strategic Diagram, Clustering. Multi-dimensional Scaling.

I.INTRODUCTION

In the era of computer science and technology, according to research from the University of Ottawa, we have already crossed 50 million marks in terms of the total number of scientific papers published since 1665, and approximately 2.5 million new scientific papers are published each year and the growth is exponential. Hence, it is almost impossible for an expert to read even half of the published research papers in the respective field. Therefore, there is a need to effectively plan the structure of the research and map the interrelations of subfields to solve this problem multiple bibliometric techniques were developed, and now the most popular technique among them was co-word analysis. The modern development of Co-word analysis originated in the 1980s of the 20th centuries by Michel Callon at the Centre de Sociologie de I'Innovation at the Ecole Des Moines de Paris. [1] The initial motivation to develop a method was to help evaluate the state of research which would have a broader scope, be more objective, and provide a supplement to, panels of experts [1]. While a panel of experts could provide an acceptable view of the trends and interrelationships within a narrowly-defined research area, identification of the connectivity of a broad range of areas is well beyond the expertise of any one panel of experts, and perhaps beyond a group of panels. Co-word analysis is based on the principle that: two keywords that express the theme of the corpus co-occurring in an article or a document are considered to be correlated and the more the frequency of their co-occurrence (will be discussed later in the paper), the more will be the strength of their co-relation. Co-word analysis is a content analysis technique that uses patterns of co-occurrences of the pair of keywords to analyze the interrelationships within the topic and the evolution of the topic.

II.METHODS

A. Keywords Extraction

Keywords are the important words that can be used to define the theme of the corpus. Since it is not possible to read every document to extract the keywords, hence, the easiest way to extract keywords is to extract the nouns with the most frequency from the corpus. But many improved and efficient algorithms have been developed by scientists such as Gensim, Yake, KeyBert, TF-IDF (term frequency–inverse document frequency), etc. Different algorithms use different logic and hence produce different outputs. There are also many keyword extraction tools such as Keyword extract is necessary because the examination of only 100 items yields 4,950 possible pairings and for a large textual dataset, the number of items would be much higher, hence reducing computational stress significantly keyword extraction is a requisite. After the keywords are extracted, there is a need of standardizing the keywords by vocabulary tool as some

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

related concept is presented by different words, this process is called Data Standardization. Standardization of keywords refers to the removal of all synonyms, ambiguity, and different variants form of the word. This process can be achieved with the help of LCSH, SLSH, and Bibliometric Dictionary. Keywords representing the same concepts should be clubbed into standardized form and words having low frequency should be merged into broader terms.

B. Co-occurrence Analysis

Co-occurrence analysis is simply the counting of paired data within a collection unit. For example, buying shampoo and a brush at a drug store is an example of co-occurrence. Here, the data is the brush and the shampoo, and the collection unit is the particular transaction. In this example, the paired data is {shampoo, brush} and it occurs once. Of course, more items can be purchased at a time, so the pairings become more numbers as each item is paired with each other item. For example, if in addition to the two items, a third item is purchased, say, goo, then there are three pairings ({shampoo, brush}, {shampoo, goo}, {brush, goo}), again each with a count of one.

Another example can be considered like, suppose there are two sentences "Apple is sweet" and "Apple is red and sweet", in this example "Apple" is related to "sweet" and "red" but the co-relation of "Apple" with "sweet" is stronger as they co-occur two times. Although, Co-relation strength of {"Apple","red"} and {"red", "sweet"} can be considered equal as they co-occur only one time each in the corpus.

When two items co-occur, there is an association between the two entities as indicated by the grouping agent. The agent can be consumer purchasing items, that is another instance that strengthens that association.

The collection of all the co-occurring elements forms a framework from which to mine associations, be them in the form of clusters, association rules, or transitive associations, and can be found within systems that highlight for the analyst those unknown facts [2].

The data of pairs of items extracted from the corpus in the form of a matrix is called the co-occurrence matrix. In this matrix, the main diagonal elements represent the actual frequency of the words, and the rest of the elements, each represents the co-occurrence frequency of a pair. The actual size of the co-occurrence matrix is N*N, where N is the number of items. However, many of the methods assume it to be a symmetric matrix and also exclude main diagonal elements as they do not add any significant value to the further analysis.

This is the most crucial step in the co-word analysis as this matrix serves as the basis for all further analysis techniques.

C. Data Visualization

Data visualization is the graphical representation of the data in the form of networks, graphs, and maps. The development of Data Visualization for data interpretation has been adopted as a well-established practice for the analysis of large amounts of data by scientists and domain experts. There are several data visualization techniques and tools which make this step a lot easier for the analysts. Multiple techniques for data visualization include Hierarchical cluster analysis, Strategic diagrams, MDS (Multidimensional scaling), and Social Network analysis.

i. *Hierarchical Cluster Analysis:* Clustering analysis is based on a simple principle of grouping the objects into similar groups. It is a widely used analysis method used to classify data into structures that are easier to understand and interpret. This process can be achieved by multiple different algorithms and methods, one of the most popular methods is hierarchical cluster analysis. There is the common belief that in terms of clustering quality, partitional algorithms are actually inferior and less effective than their agglomerative counterparts. This belief is based both on experiments with low dimensional datasets as well as a limited number of studies in which agglomerative approaches in general outperformed partitional K -means based approaches. For this reason, existing reviews of hierarchical document clustering methods focused mainly on agglomerative methods and entirely ignored partitional methods. Hierarchical cluster analysis starts by treating each entity as an individual and then starts identifying the closest and most similar entities and merging them into a parent cluster, these steps are repeated until all the entities are merged into a parent cluster. This particular method is also called as Agglomerative method. The output of the hierarchical clustering method is a dendrogram, which shows the hierarchical relationships of the clusters. This technique has been applied in many types of research, for example, [3] [4] and [5]

ii. *Strategic Diagram:* A strategic diagram is a two-dimensional space built by slitting themes according to their centrality and density, where the abscissa axis represents the centrality, the ordinate axis represents the density, and the origin is denoted by the median or mean value of the two, centrality and density [6]. Centrality is the measure of the strength of correlation of one theme with other themes, whereas density is the measure of the importance of the theme. Strategic diagrams are used to visualize the internal relations within a cluster, as well as interactions among other clusters. Several studies have adopted this method of analysis because of its easy-to-understand representation, for example, [4].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

A strategic diagram is divided into 4 quadrants with the x-axis being centrality and the y-axis being density. In quadrant 1, containing clusters with high density and high centrality, themes or clusters are more important and more mature. In quadrant 2, themes or clusters are not that important but are well developed. In quadrant 3, the themes or clusters are neither central nor developed. In quadrant 4, themes or clusters are important and central but are underdeveloped

iii. *Multidimensional Scaling:* Multidimensional scaling is a technique that creates a perceptual map displaying the relative positioning of a number of objects, i.e, it visualizes the level of similarity of each individual object in a dataset. In this technique, a proximity matrix is inputted into multidimensional scaling software directly to generate a map that describes pairwise distinctions between N objects. As Kruskal and wish (1978, p.7) formulated, "A proximity is a number which indicates how similar or how different two objects are or are perceived to be, or any measure of this kind." Proximity matrices can be either similarity or dissimilarity matrices. The dissimilarity matrix can be calculated using Minskowski distance,

$$(\sum_{i=1}^{n} |X_i - Y_i|^p)^{1/p}$$

Where X and Y are the two data points and the case where p = 1 is equivalent to the Manhattan distance and the case where p = 2 is equivalent to the Euclidean distance.

According to our research, the most widely used method to calculate proximity matrix is Pearson' Correlation which measures how well the two items are related to each other. In other words, it indicates the similarity or dissimilarity of a pair of items.

$r = (N\sum xy - (\sum x)(\sum y))/\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}$

This formula returns a value between -1 to 1. The value near or equal to -1 or 1 denotes a strong relationship whereas value near zero denotes low or no relationship. The positive value denotes the direction of correlation as positive and negative value denotes the direction of correlation as negative.

iv. *Social Network Analysis:* *Social network analysis is one of the most popular methods of data analysis which is used to assess the latent content of the subject and inter-relationships within a subject using network and graph theory. Social network analysis has been extensively used to evaluate the unique structure of interrelationships in multiple disciplines such as psychology, social science, scientometrics, etc. The final result of this method is a sociogram which is developed to provide information about the relationship between the members of the network. At its simplest, each member in a network is called a "vertex" or a "node". Each vertex represents a member and the size of the vertex represents the frequency of the word. The larger the size of a node is, the higher the frequency is and vice-versa. The line between two nodes illustrates the connection relationship between two words, which directly means that the two words co-occurred in the same dissertation, and the thickness of the line depicts the strength of correlation between two keywords. The thicker line between two keywords, the stronger the correlation is.*

This method has attracted much attention in recent years. It is also referred to as "structural analysis". In recent years, several studies have used this method to obtain more intuitive and comprehensive results and to understand concealed meaning in the corpus or dataset.

D. Data analysis tools

In the last couple of decades, data analysis has been one of the most trending fields of study all over the world. Hence to simplify the main tasks of analysis which are used widely for different purposes, tools are developed and some of the most popular tools used in the co-word analysis are listed below:

Ucinet: UCINET is a comprehensive social network analysis software program. It can be used to visually analyze the co-occurrence matrix of the keywords. This software is currently developed for only windows but it has extremely rich metrics and it is the software most often used in academic circles. Also, Pajek, Mage, and NetDraw are bundled in Ucinet. The program itself does not include a graphics program for network visualization, but it can output data and process results to software such as NetDraw, Pajek, Mage, and KrackPlot for plotting. The Ucinet includes a large number of network analysis programs including the detection of cliques, clans, plexes, components, cores, centrality, personal network analysis, and structural hole analysis. It also contains numerous process-based analysis programs such as cluster analysis, multidimensional scaling, two-mode scaling (singular value decomposition, factor analysis, and correspondence analysis), role and status analysis (structure, role, and regular equivalence), and the center-edge fitting model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

VOSViewer: This is the most popular tool for constructing and visualizing bibliometric networks.Networks can be constructed based on bibliographic coupling, co-citation, co-authorship, or co-word relations. VOSviewer also offers zoom and scroll functionality, which helps in acquiring more detailed information about a map. Network maps and density maps can be constructed using this software.
CiteSpace: Citespace is a freely available Java-based application used for visualizing and analyzing trends and patterns in scientific literature. It helps in finding critical and pivotal intellectual structures. Citespace also facilitates in understanding and interpretation of network patterns and historical patterns, including the fast-growth topical areas, finding citation hotspots in the land of publications, decomposing a network into clusters, automatic labeling clusters with terms from citing articles, geospatial patterns of collaboration, and unique areas of international collaboration. [7]
Bibliographic Item Co-occurrence Mining System (BIOCOMS): This software was adopted as a way to manage a large amount of data to analyze the trends, identify the core journals and calculate the keywords co-occurrence matrix.
Statistical Package for The Social Sciences (SPSS): This is a statistical data analysis tool owned by IBM for complex data analysis and large data processing. It facilitates a user-friendly interface that makes it easy to use software and advanced statistical procedures which make the output with high accuracy and quality. It is feature-rich software that offers advanced statistical analysis, a huge library of machine learning algorithms, text analysis, open-source extensibility, and integration with big data.

III.LITERATURE REVIEW

[6] The role of scientific research in the innovation process is one of the major issues raised by the economy of technical change. For impact, a network of interactions (techno-economic networks) between science (S), technology (T), and the market (M) was attempted. The main goal of them is to explain the importance of co-word analysis which can be useful in the task.
Narin, for example, has studied the relationships that exist between science and technology by using patent citations to scientific articles. [8] Her study of S/T/M interactions sought to explain the time lags she observed between these three variables.

Co-word analysis had been used to describe the interactions that exist between different stages of the innovation process as well as to determine whether basic research or applied research is the driving force. There were two limitations of their research i.e, the first was that no in-depth examination of the domain would be conducted. Rather, they presented the technique, demonstrate how to use it, and provide examples chosen to demonstrate how different types of results can be interpreted. The second was that, in order to simplify the paper, they had focused their attention on a very small portion of the overall techno-economic network picture.

[9] The research motivation was totally dependent upon how various types of interventions by sponsors and policymakers could influence the evolution and impact of research. Modern quantitative techniques make extensive use of computer technology, which is usually supplemented by network analytic approaches, in order to integrate disparate fields of research.

The origins of co-word phenomena could be traced back at least five decades to pioneering work in 1) lexicography of Hornby (1942) to account for co-occurrence knowledge, and 2) linguistics of De Saussure (1949) to describe how the affinity of two languages units correlates with their appearance in the language. Early co-word studies classified words based on their co-occurrence with other words as well as their meanings [10] [11]. Chomsky added that the reasons for two words co-occurring in the same context are not always relevant to a general linguistic description of a given language. [12]

One of the studies used collocations as part of a linguistic model whose goal was to connect any given meaning and the texts that express it [13]. A combinatory dictionary containing only general English collocations was recently created. [14]

Information retrieval research focused on designing more efficient indexing tools using pairwise lexical affinities instead of keywords. [15] [16] [17] [18]

Applications based on stochastic language models built on previous work in speech recognition and text compression, and treated collocations as statistical entities [19] [20]

The Turner study's goal was to conduct a co-word analysis on industrial ceramics patents in order to determine priority areas for Ireland in this field. The LEXINET system, a computer-assisted indexing method, was used to reduce the

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

'indexer effect.' The high-frequency significant words extracted from the titles and summaries of the 16,000-patent database were clustered into clusters of strongly related words. Each cluster received two quantitative figures of merit: density and centrality.

It presented a new approach to co-word analysis which requires no index or keywords but deals with the text directly and useful tool for rapidly scanning large bodies of text to identify pervasive thrust areas and their connections. The current approach necessitates the physical proximity of the words in addition to co-occurrence.

The main limitations were that the methods for data generation and interpretation were insufficiently understood and required development and maturation, computers were not large or fast enough to provide technical resolution at a sufficient level of detail in a reasonable period of time to be useful, and the appropriate combination of expertise (analyst, technical expert, decisionmaker) was not brought to bear when the analyses were performed.

[21] A large number of publications from 1982 to 1994 were analyzed in this study to determine themes and trends in software engineering. Co-word analysis was used to analyze the publications. This methodology identifies patterns of association among publication descriptors (indexing terms) from the Computing Classification System and generates networks of terms that reveal those patterns.

The co-word methodology was applied to textual data that has been indexed. This chapter discusses the study's two components. The index terms used in this case are directly taken from a standard taxonomy.
At the SEI, the software is used to generate index terms directly from the studied corpora in other similar applications. There were 16,691 documents in total. The number of documents was limited until 1986. A total of 57,727 descriptors were used to index the 16,691 items (a mean of 3.46 per item). Metrics for co-word analysis had been extensively researched. [22]

Then Strength S of association between descriptors was given by the expression.

The primary patterns of emphasis in a period were revealed by network analysis. Representing node and link interactions within larger contexts allows for the identification of research and commentary trends in software engineering publications. The constraint of the algorithm used in the research was, without some minimum constraints, descriptors that appear infrequently but almost always together could dominate networks; thus, a minimum co-occurrence value is required to generate a link.

Because this particular study was based solely on refined publications, it represented topics that were more developed than others. There was certainly a lot of activity in cost/time estimation, programming team management, and other important but still relatively immature areas. The time lag between the invention of software technology and its acceptance into common practice was estimated to be 15-20 years, so this gap was not surprising.
Co-word analysis applied to author-defined descriptors, abstracts, or the text of a document may reveal observations that are complementary to the ones we noted.

[23] The main objective of the study was to map the intellectual structure of the field of Information Retrieval (IR) from 1987 to 1997. By measuring the association strengths of terms representative of relevant publications or other texts produced in the IR field, co-word analysis was used to reveal patterns and trends in the IR field. For the years 1987 to 1997, data were gathered from the Science Citation Index (SCI) and the Social Science Citation Index (SSCI). Other important keywords were manually extracted from titles and abstracts in addition to the keywords added by the SCI and SSCI databases. These keywords were standardized further using vocabulary control tools.

Traditional bibliometric techniques, such as author and journal co-citation analyses, were based on citation analysis in scientific papers.

Data were collected by using two ways. One method was to extract keywords from keyword lists, titles, abstracts, and, in some cases, classification codes. And another one was to download keywords from the online database. The number of times two keywords appear together in the same publication was calculated using custom-built Foxpro programs. As a result, we created a 240 x 240 keyword co-occurrence matrix. We put the co-occurrence frequency of XY and Y in the cell of keywords XY and Y. The diagonal matrix values were treated as missing data. Using Pearson's correlation coefficient, the matrix was transformed into a correlation matrix, indicating the similarity and dissimilarity of each keyword pair. MDS maps were generated for the Data mapping technique for further clustering.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

This study demonstrated the viability of co-word analysis as a viable approach for extracting patterns and identifying trends in large corpora where the texts collected are from the same domain or sub-domain and are divided into roughly equivalent amounts for different time periods.

[24] Co-occurrence matrices, such as co-citation, co-word, and co-link matrices, had been widely used in the information sciences. However, confusion and controversy had been hampered proper statistical analysis of these data.

The underlying issue, in their opinion, was understanding the nature of various types of matrices.
This article discussed the distinction between a symmetrical co-citation matrix and an asymmetrical citation matrix, as well as the statistical techniques that can be applied to each of these matrices. They suggested that similarity measures (such as the Pearson correlation coefficient or the cosine) should not be applied to the symmetrical co-citation matrix, but could be used to derive the proximity matrix from the asymmetrical citation matrix.

A set of data gathered using the Google Scholar search engine was analyzed using both traditional multivariate analysis methods and the new visualization software Pajek, which was based on social network analysis and graph theory. PROXSCAL had been taken into account for the measurement scale, it may be more appropriate for visualizing co-citation data than Pajek. However, as previously stated, Pajek allows the user to indicate the strength of the relationship through the thickness of the lines. Users of PROXSCAL and other MDS programs must draw the relevant lines and groupings themselves.

They concluded that while the Pearson correlation coefficient should not be applied to a symmetrical co-citation matrix, it could be applied to an asymmetrical citation matrix to derive the proximity matrix, which is required for analysis such as multidimensional scaling. The article also made a clear distinction between similarity and dissimilarity matrices, and they demonstrated how to define them when using statistical software like SPSS.

[3] The purpose of this article was to fill a knowledge gap in co-word analysis. They compared three visualization methods: cluster tree, strategy diagram, and social network maps, and integrate different results into one result using medical informatics co-word analysis.

They also mentioned that cluster trees depict the subject structure, strategic diagrams highlight the significance of topic themes in the structure, and social network maps interpret the internal relationship between themes. Among these techniques, co-classification analysis [25], co-citation analysis [26], and co-word analysis [23] were representative methods.

The principle of the technique used was when two professional terms expressing a specific research topic appear in the same article, there was an inherent relationship between the two words. And the more frequently these two words appear together, the closer their relationship was. Many applications than used mature visualization skills of co-word analysis. Nanotechnology [27], knowledge management, other subjects and disciplines, international scientific studies bioinformatics, and others. Using Bibliographic Item Cooccurrence Mining System (BIOCOMS), they retrieved major Mesh words and counted their frequencies from 2004-to 2008.

Adhesion strength is the average value of a word's co-occurrence frequency with others in the same category, which is used to assess the role of words in the clustering process.

The most significant advantage of co-word analysis is that the strategic diagram allows you to see the overall structure of the specific technology domain [28]

Mapping techniques in bibliometrics have a long tradition and go beyond text analysis. Co-word analysis condenses a large space of related descriptors into multiple related smaller spaces that are not only easier to understand but also suggest actual partitions of interrelated concepts in the literature under consideration [23].

Limitations: Had been limited by clustering and being unable to identify the relationship between all of the topic words. Visualization is not the primary goal of social network analysis. It is a quantitative sociology method that allows for the analysis of information about a social group in terms of mutual relations between group members.

[4] The purpose of this research was to map the intellectual structure of the digital library (DL) field in China from 2002 to 2011. Co-word analysis was used to uncover patterns in the DL field in China by measuring the association

DOI: 10.17148/IJARCCE.2022.11576

strength of keywords in relevant journals. Data was gathered from the Chinese Journal Full-Text Database between 2002 and 2011. The co-occurrence matrix of keywords was then analyzed using multivariate statistical analysis and social network analysis methods.

From 1998 to 2007, Li collected 1,948 articles on DL published in 17 core journals in Library & Information Science. The aspects of time and space distribution, content distribution, and author distribution were highlighted in this analysis. Using co-word analysis, this study attempted to map the intellectual structure of the DL field in China, including keyword relationships, research structure, and situation.

In the particular research, they depicted that any two keywords that occur in the same article were relevant to the topics to which they refer [29]. The presence of many co-occurrences of a pair of keywords within articles suggests that they might be part of the same research theme [23]. The correlation between keywords is determined by the number of articles that contain these two keywords.

Pearson's correlation indicates the similarity and dissimilarity of each keyword pair, as well as the degree of correlation among research themes.
In this article, each cluster represented a large research theme or research direction in the Chinese DL field.

Conclusion: The conclusion of the research was that SPSS19.0 and Ucinet6.0 tools were used for analysis results of the DL field in China.

[30] The purpose of this paper was to broaden recent reflection on the evolution of strategic management by examining the field's object of study: strategy. They had chosen content analysis for this study, combining consensus and co-word analysis with social network analysis techniques.

One of three general approaches in information science for demonstrating the evolution of socio-cognitive structures from a set of documents is co-word analysis. Co-word analysis, like co-citation analysis [31] and co-author analysis [32] uses co-occurrence and co-absence patterns of pairs of objects (e.g., words, nouns) in a corpus of texts to identify the relationship between ideas presented in such texts. They coded the matrix using the values obtained from calculating the inclusion index [33] of each pair of key terms. After constructing the co-occurrence/co-absence matrix, they used social network analysis techniques to determine the degree of centrality (closeness) of the key terms and then traced the evolution of the structure of the definition of the strategy concept in the three stages studied.

The centrality degree depicts the evolution of the key term's influence in its location in each of the stages investigated.

Findings: They concluded that the essence of the strategy concept is the dynamics of the firm's relationship with its environment, for which the necessary actions are taken to achieve its goals and/or to improve performance through the rational use of resources. The study reveals that the evolution of the internal cohesion of the key terms in the definition's structure resulted in the formation of new research subfields, which favored the field's rapid propagation and enrichment of its theoretical corpus.

[34] This article described a keyword analysis conducted to reveal publication patterns in the field of renewable energy, including the temporal evolution of its various research lines over the last two decades.

Researchers developed the technique of co-word analysis using keywords, in which a network is generated with the various keywords connected by links weighted by the number of papers in which the keywords at each end of the link co-occur [6]. In this analysis, the traditional method is to use a clustering procedure based on the two characteristics of cohesion and centrality.
Their research questions were: What is the main topics that structure renewable energy research? How do they relate to one another? Which themes are more central, and which are more specialized? What is the internal cohesion of each topic?

The data used for the study were taken from Elsevier's Scopus database [35] [36] one of the bibliographic databases.

The final results revealed a clear increase in scientific production related to alternative energies, as well as a structure corresponding to five major clusters, which were decomposed into 22 at a finer level of resolution. To address the issues represented by synonyms, singular/plural, hyphenated or unhyphenated descriptors, and so on, the multiple variants of any given term were unified into a single version.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

It was based on revealing communities in networks by gradually removing linkages with the highest betweenness, and on using modularity as a measure of the strength of participation in those communities. (Modified by Van Eck and Waltman) [37]
Van Eck and Waltman's study proposes an optimization algorithm with weighable and parameterizable modularity that allows one to adjust the size of the communities by changing a resolution parameter [37]. The cohesiveness defined by Callon, Courtial, and Penan (1995) has been employed as a measure of cluster coherence. [38]

Limitations: The author has only examined publications for this task since they were higher-ranking scientific contributions. Although reviews earn more citations, their scientific value was less, and they might bring significant noise when dealing with often huge issues.

Conclusion: The network map consisting of the clusters was formed at a finer level of resolution. There was some mixing in the central part of the map.

[5] The primary concepts addressed in this research field were identified using co-word analysis. To cluster the terms, hierarchical cluster analysis was utilized, and a strategy diagram was produced to assess trends.

Co-word analysis was employed in the study to characterize the evolution and current state of the literature on gender inequalities in science, with an emphasis on factors that influence gender inequality in higher education and science. Callon et al. proposed this bibliometric technique to help them visualize the division of the field (in this case, the explanatory factors for gender differences in science) into several subfields and show the relationships between them, providing insights into the evolution of the main topics discussed in the field over time. [39]

Their sole idea of a co-word analysis was to map the dynamics of a subject and identify the essential study subjects based on the pattern of co-occurrence of pairs of keywords that represent the various themes in a body of literature. [40]

SPSS v20 was used to generate the word-document occurrence matrix automatically. The Jaccard similarity index was also used to calculate the similarities between keywords. Ward's approach was used for hierarchical clustering analysis and squared Euclidean distance was used as the distance metric in SPSS v20.

In the last result, 16 clusters of keywords (themes) were discovered based on the hierarchical clustering of 106 keywords.

[41] Cluster analysis and social network analysis identified 12 theme clusters, cluster network features (centrality and density), the strategic diagram, and the correlation network. According to the study findings, there are numerous major topics with a strong correlation in Chinese RecSys research, which is deemed to be generally concentrated, mature, and well-developed overall.

According to a survey of the literature on RecSys research in China, most studies employed qualitative methodologies and were done based on the personal opinion of a small group of specialists.

The primary purpose of this work was to fill the gaps and restrictions by offering a detailed examination and analysis of RecSys' research developments in China during the last ten years. This strategy was used in this study to provide light on the general research structure, the association between topics, and the overall evolutionary tendencies in RecSys studies in China. Using co-word data, cluster analysis had been performed to study topics. Researchers created novel co-word analysis methods and tools, such as co-word clustering [6], multidimensional scaling, social network [23], and the strategic diagram. The strategic diagram, in particular, considers both centrality and density, and may therefore depict the dynamics of research areas.

SATI was used to generate a co-occurrence matrix from a set of keywords. Values in diagonal cells were considered as missing data in this matrix, whereas values in non-diagonal cells were co-word frequencies. The co-word matrix was then converted into a Pearson's correlation coefficient matrix, with each cell representing the degree of similarity between the row element (a keyword) and the column element. For the following data analysis, the co-word correlation matrix was employed.

Hierarchical clustering was used in cluster analysis, with Ward's technique as the cluster method and Squared Euclidean distance as the distance measure. Several social network analysis indicators were adopted for the study, including

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11576

network centrality, density, core-periphery structure, and strategic diagram.

Result: According to the power-law distribution, the research structure in the field of RecSys in China was unevenly distributed.
Conclusion: Co-word analysis was employed in this study, together with clustering and social network analysis methodologies, to show research patterns and trends in the RecSys area in China from 2004 to 2013.

[42] This paper suggests that articles were retrieved from the journal Scientometrics using SpringerLink (full-text database), and keywords were extracted non-parametrically from the LISA database and the articles themselves.

Co-word analysis is a technique that uses the co-occurrence pattern of words and phrases in a corpus. It creates a relationship between an idea and a notion within the topic area as described in the corpus. The presence of two keywords in the same document implies a connection between the issues to which they allude [29]. The presence of several co-occurrences with a keyword or phrase shows a primary point that has numerous links with other terms in a corpus that may be related to a study issue. It determines the strength of word co-occurrence and generates a set of lexical graphs that effectively demonstrate the strongest relationship between diverse phrases.

MDS was used to plot and detect the closeness of keywords. The Salton index was used to create a network and comprehend the link between terms.

Focusing on their data, non-parametric and parametric measures were taken into account. Non-parametric approaches deal with the manual efforts of gathering keywords provided by the author, journal databases, abstract databases, and citation databases.

[43] The goal of their work was to give a full bibliometric analysis of OR/MS in Mainland China, including statistical analysis of the quantity and kind of papers, as well as co-author analysis at the author, institution, and national levels. Finally, study ideas and citations were explored by them. They have used the JRAP index as a metric for each institution.

They also mentioned that Shang et al. measure researcher cooperation by the percentage of papers with three or more authors and the average number of authors per published article [44]. Therefore, they have used co-author analysis to assess the extent of collaboration in the OR/MS area. The two most commonly utilized variables in the research of academic collaboration were cooperation rate and cooperation degree [45]. These are their definitions:

Cooperation rate = total number of papers that cooperated / total number of papers 100

Total number of co-authors (cooperated institutions or countries)/Total number of authors = Cooperation degree (institutions, or countries)

They have used the bibliometric method to conduct a statistical analysis of the state of OR/MS research in Mainland China from 2001 to 2013, looking at the number of papers published, the quality of papers, the most productive institutions, co-author analysis, citation analysis, and research topic analysis.

IV.CONCLUSION

Co-word analysis is an important content analysis technique that is being used in multiple different disciplines to discover new insights such as studying academic trends of a topic, studying interrelations of subfields within a field, identifying hotspot keywords and themes inside a field, and so on. The biggest benefit of co-word analysis is that we can see the overall structure of the specific technology domain through the strategic diagram.
According to a study [3], the Clustering tree shows the research focus of MI using high-frequency words by quantitatively reflecting the content structure of these words. But it ignores the semantic relations and logical connections between words in the clustering process, so there are some drawbacks: (1) It is difficult to determine which keywords play a major role in the process. To solve this problem, this study introduced the evaluation of the type of contribution MeSH indicators to determine the adhesion of the center of each category word. (2) No specific details on the relationship and interaction of inter-cluster so we can't determine which cluster is core or mature. (3) Cluster result is often affected by subjective factors.
As well known, a strategic diagram reflects evolutionary trends of every cluster and the relationship between all clusters clearly highlighting the importance or feature of each research topic in the structure of some field by quantifying their

DOI: 10.17148/IJARCCE.2022.11576

maturity and core degree. However, it has some shortcomings, such as, it is limited by the result of clustering and it cannot identify the relationship among all the topic words.

According to this synthesis of co-word analysis, the majority of papers use this methodology to gain insights into a particular field but there is only a handful of papers that takes this methodology one step further.

This paper attempts to review this methodology and provide a detailed view of the most popular techniques, procedures, and tools that are well-accepted among researchers and scientists. This study will guide upcoming researchers researching in this field of bibliometrics analysis.

REFERENCES

[1] C. J. P. a. T. W. A. Callon M., "PROXAN: A Visual Disp1ay Technique for Scientific and Technical Problem Networks," *Second Workshop on the Measurement of R&D Output, Paris, France*, pp. 5-6, 1979.

[2] J. W. Buzydlowski, "Co-occurrence analysis as a framework for data mining," *Journal of Technology Research 6*, pp. 1-19, 2015.

[3] Y. M. W. a. L. C. Yang, "Integration of three visualization methods based on co-word analysis," *Scientometrics*, vol. 90, no. 2, pp. 659-673, 2012.

[4] G.-Y. J.-M. H. a. H.-L. W. Liu, "A co-word analysis of digital library field in China," *Scientometrics*, vol. 91, no. 1, pp. 203-217, 2012.

[5] T. A. V. a. M. B. Dehdarirad, "Research trends in gender differences in higher education and science: a co-word analysis," *Scientometrics* , vol. 101, no. 1, pp. 273-290, 2014.

[6] M. J.-P. C. a. F. L. Callon, "CO-WORD ANALYSIS AS A TOOL FOR DESCRIBING THE NETWORK OF INTERACTIONS BETWEEN BASIC AND TECHNOLOGICAL RESEARCH: THE CASE OF POLYMER CHEMISTRY," *Scientometrics* , vol. 22, no. 1, pp. 155-205, 1991.

[7] C. F. I. a. J. H. Chen, "The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis," *Journal of the American Society for information Science and Technology*, vol. 61, no. 7, pp. 1386-1409, 2010.

[8] F. N. E. &. P. R. Narin, "Callon M., Courtial, J. P., and Tumer, W. A., "PROXAN: A Visual Disp1ay," *Research policy*, vol. 16, no. 2-4, pp. 143-155, 1987.

[9] R. N. Kostoff, "Co-word analysis," *Springer, Boston, MA*, pp. 63-78.

[10] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, 1957.

[11] Z. S. Harris, "Mathematical structures of language," 2020.

[12] N. Chomsky, "Aspects of the Theory of Syntax. MIT Press, Cambridge, Mass, 1965," 1965.

[13] I. A. Mel'čuk, "Meaning-Text Models: A recent trend in Soviet linguistics," *Annual review of Anthropology* , vol. 10, no. 1, pp. 27-62, 1981.

[14] M. B. E. a. I. R. Benson, "The BBI Combinatory Dictionary of English: A Guide to Word Combinations," *John Benjamins, Amsterdam and Philadelphia*, 1986.

[15] K. S. Jones, "Automatic keyword classification for information retrieval," 1971.

[16] C. J. Van Rijsbergen, "Information retrieval 2nd edition butterworths," *London available on internet* , 1979.

[17] G. a. M. J. M. Salton, "Introduction to modern information retrieval," *mcgraw-hill*, 1983.

[18] Y. S. a. F. Z. S. Maarek, "Full text indexing based on lexical relations an application: Software libraries," *ACM SIGIR Forum*, vol. 23, no. SI, pp. 198-206, 1989.

[19] L. R. F. J. a. R. L. M. Bahl, "A maximum likelihood approach to continuous speech recognition," *IEEE transactions on pattern analysis and machine intelligence* , pp. 179-190, 1983.

[20] W. A. a. K. C. Gale, "Poor estimates of context are worse than none," *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June*, pp. 24-27, 1990.

[21] N. M. I. &. K. S. Coulter, " Software engineering as seen through its research literature: A study in co-word analysis," *Journal of the American Society for Information Science*, vol. 49, no. 13, pp. 1206-1223., 1998.

[22] J. Whittaker, "Creativity and conformity in science: Titles, keywords and co-word analysis," *Social Studies of Science*, vol. 19, no. 3, pp. 473-496, 1989.

[23] Y. C. G. G. &. F. S. Ding, "Bibliometric cartography of information retrieval research by using co-word analysis.," *Information processing & management*, vol. 37, no. 6, pp. 817-842, 2001.

[24] L. a. L. V. Leydesdorff, "Co-occurrence matrices and their applications in information science: Extending ACA to

DOI: 10.17148/IJARCCE.2022.11576

the Web environment," *Journal of the American Society for Information Science and technology,* vol. 57, no. 12, pp. 1616-1628, 2006.

[25] A. a. G. 2010, "A co-classification approach to learning from multilingual corpora," *Machine Learning* , vol. 79, no. 1-2, p. 105–121, 2010.

[26] C. A. R. E. M. &. M. T. Cottrill, "Co-citation analysis of the scientific literature of innovation research traditions diffusion of innovations and technology transfer," *Journal of Information,* vol. 36, no. 3, p. 383–400, 2010.

[27] R. N. S. J. A. J. D. M. J. S. L. C. G. Y. &. T. W. M. Kostoff, "The structure and infrastructure of the global nanotechnology literature," *Journal of Nanoparticle Research,* vol. 8, no. 3, pp. 301-321, 2006.

[28] B. &. J. Y. -I. Lee, "Mapping Korea's national R&D domain of robot technology by using the co-word analysis," *Scientometrics,* vol. 77, no. 1, p. 17, 2008.

[29] A. L. C. C. J. P. &. L. F. Cambrosio, "Historical scientometrics? Mapping over 70 years of biological safety research with co-word analysis," *Scientometrics,* vol. 27, no. 2, p. 119–143, 1993.

[30] G. A. a. L. Á. G. Ronda-Pupo, "Dynamics of the evolution of the strategy concept 1962–2008: a co-word analysis," *Strategic management journal* , vol. 33, no. 2, pp. 162-188, 2012.

[31] I.-S. F. H. J. Chen C, "The structure and dynamics of co-citation clusters: a multiple perspective co-citation analysis," *Journal of the American Society for Information Science and Technology,* vol. 61, no. 7, p. 1386–1409, 2010.

[32] S. A. Zhao D, "Information science during the first decade of the Web: an enriched author co-citation analysis," *Journal of the American Society for Information Science and Technology,* vol. 59, no. 6, p. 916–937, 2008b.

[33] L. J. R. A. Callon M, "Mapping the Dynamics of Science and Technology," *Sociology of Science in a Real World. Macmillan: London, UK,* 1986.

[34] L. M. V. P. G.-B. a. F. M.-A. Romo-Fernández, "Co-word based thematic analysis of renewable energy," *Scientometrics,* vol. 97, no. 3, pp. 743-765, 2013.

[35] P. Hane, "Elsevier announces Scopus service," *Information today,* 2004.

[36] B. Pickering, "Elsevier prepares Scopus to rival ISI Web of science," *Information world review. Amsterdam,Elsevier* , 2004.

[37] N. J. &. W. L. Van Eck, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics,* vol. 84, no. 2, p. 523–538, 2010a.

[38] M. C. J. P. &. P. H. Callon, "Cienciometrı´a. El estudio cuantitativo de la actividad cientı´fica: de la bibliometrı´a a la vigencia tecnolo´gica," *Gijo´n: Ediciones TREA,* 1995.

[39] M. C. J. P. T. W. A. &. B. S. Callon, "From translations to problematic networks:An introduction to co-word analysis," *Social Science Information,* vol. 22, no. 2, p. 191–235, 1983.

[40] Q. He, "Knowledge Discovery through Co-Word Analysis," *Library Trends,* vol. 48, no. 1, p. 133–159, 1999.

[41] J. a. Y. Z. Hu, "Research patterns and trends of Recommendation System in China using co-word analysis," *Information processing & management,* vol. 51, no. 4, pp. 329-339, 2015.

[42] S. A. A. a. S. N. S. Ravikumar, "Mapping the intellectual structure of scientometrics: A co-word analysis of the journal Scientometrics(2005–2010)," *Scientometrics,* vol. 102, no. 1, pp. 929-955, 2015.

[43] D. e. a. Wu, "A systematic overview of operations research/management science research in Mainland China: Bibliometric analysis of the period 2001–2013.," *Asia-Pacific Journal of Operational Research* , vol. 33, no. 06, p. 1650044, 2016.

[44] G. B. S. T. F. e. a. Shang, "Twenty-six years of operations management research (1985–2010): Authorship patterns and research constituents in eleven top rated journals," *International Journal of Production Research,* vol. 53, no. 20, p. 6161–6197, 2015.

[45] G. a. L. G.-M. Ronda-Pupo, "Dynamics of the scientific community network within the strategic management field through the Strategic Management Journal 1980–2009: The role of cooperation.," *Scientometrics,* vol. 85, no. 3, p. 821–848, 2010.

[46] Y. C. G. G. &. F. S. Ding, "Bibliometric cartography of information retrieval research by using co-word analysis.," *Information Processing & Management,* vol. 37, no. 6, p. 817–842, 2001.