doi: 10.3724/SP.J.1085.2012.00041	March 2012 Vol. 23 No. 1: 41-46

Content analysis of documents using neural networks: A study of Antarctic science research articles published in international journals

DASTIDAR, Prabir G[1*] & JHA, Deepak Kumar[2]

[1]Ministry of Earth Science, Prithvi Bhavan, Lodi Road, New Delhi 110003, India;
[2]Tata Consultancy Services (TCS), 4th & 5th Floor, PTI building, Parliament street, New Delhi 110001, India

Received 1 September 2011; accepted 11 January 2012

Abstract Content analysis of scientific papers emanating from Antarctic science research during the 25 years period (1980—2004) has been carried out using neural network based algorithm–CATPAC. A total of 10 942 research articles published in Science Citation Indexed (SCI) journals were used for the study. Normalized co-word matrix from 35 most-used significant words was used to study the semantic association between the words. Structural Equivalence blocks were constructed from these 35 most-used words. Four-block model solution was found to be optimum. The density table was dichotomized using the mean density of the table to derive the binary matrix, which was used to construct the network map. Network maps represent the thematic character of the blocks. The blocks showed preferred connection in establishing semantic relationship with the blocks, characterizing thematic composition of Antarctic science research. The analysis has provided an analytical framework for carrying out studies on the content of scientific articles. The paper has shown the utility of co-word analysis in highlighting the important areas of research in Antarctic science.

Keywords **Antarctica, content analysis, thematic analysis, scientometrics, neural network, co-occurrence, co-word, social network analysis**

Citation: Dastidar P G, Jha D K. Content analysis of documents using neural networks: A study of Antarctic science research articles published in international journals. Adv Polar Sci, 2012, 23: 41-46, doi: 10.3724/SP.J.1085.2012.00041

0 Introduction

The research articles published in scientific journals provide a macro level view of the main field, subfields and their linkages in the research domain. Since the outcome of any research is expressed through words, the analysis of these significant words can provide the micro level information about the content of the article vis-à-vis area of research. A particular knowledge domain has to be further organized into several sub-domains and to understand these knowledge sub-domains, analysis of the micro level indicators in the form of representative words and their association patterns leading to the concept formation is required. Word usage is more codified, and it seems possible to distinguish between words with a major theoretical, methodological, or observational meaning within the context of a given specialty. According to Leydesdorff[1], it provides an analytical framework for carrying out dynamic analysis of the contents of articles. The keywords are often used to identify sub-domains of research specialties all over the world.

The analysis of knowledge sub-domain in a research specialty is carried out through the study of dominant words in titles or abstracts of published scientific articles. The network of co-occurrences between different words, collected on a specific set of publications, allows the quantitative study of the structure of publication contents, in terms of the nature and strength of linkages. For the present study, the sub-domains were identified using the structural equivalence techniques by grouping keywords at different

* Corresponding author (email: Prabirgd11@gmail.com)

levels[2].

A scientific field is characterized by a group of "words", which signify its concepts, operations and processes or methodologies. The structure depicted by the frequency of co-occurrences of conceptual words reveals the important and informative linkages across them and provides a further insight into research field. These contextual analyses of co-occurrences of words enable the investigators to grasp the static and dynamic aspects of the manner in which scientists relate and place their work in a hierarchy of scientific research concepts.

This method labeled as "co-word analysis", provides a direct quantitative way of linking the conceptual contents of scientific publications, by comparing and classifying these publications based on the occurrences of similar word-pairs. Hence, such a co-word structure can represent the research activities in a scientific field. In the present study, the co-word analysis was applied to identify topics/research themes in the area of Antarctic science based on the articles published in international journals during a period of 25 years viz. 1980 to 2004. This is a valuable supplementation to the studies on intellectual structure of the field of Antarctic science during 1980 to 2004[3-4].

1 Materials & methods

1.1 Title words as indicator of research activity

The titles of a scientific article is an important indicator of its contents, and provides a clue to the importance of the research work. Numerous surveys have shown that bibliographies appearing in the papers are one of the most valuable sources of information in literature searching[5]. Garfield[6] showed that the "title words" provide a special perspective on scientific and scholarly activity and help in identifying research fronts. Search terms extracted from the titles of research articles are useful search terms for retrieval of information from databases and for augmenting retrieval efficiency[7].

1.2 Content analysis using neural network

To carry out content analysis, a self-organizing neural network based algorithm (software)–CATPAC, was used to derive the normalized matrix of word associations. Each word that CATPAC finds is associated with an artificial "neuron" in CATPAC's simulated brain. As a result of the learning and forgetting rules, CATPAC produces a 'brain" consisting of a network of interconnected neurons, each of which represents a word in the text. Some of these neurons are tightly and positively connected, indicating that they are closely associated. Whenever one of them is activated, the likelihood is high that the other will also be called to mind. Other neurons will be strongly negatively connected, indicating that another is very unlikely to be active when the one is active[8].

The neural network based algorithm makes it possible to retrieve episodic memories of the text document. Remembering episodic memories is generally more complex than recalling semantic memories, involving the evaluation of cued memories based upon the current goal[9].

The algorithm works by passing a moving window of the size n (in the present analysis 3-word window was used) through the text. In our study, the text was a collection of all the titles of the papers. Each title was separated by delimiter "-1" to single out contributions from individual publications. Any time the window encounters a word, the neuron representing the word becomes active, connections among active neurons are strengthened, so the words that occur close to each other in the text tend to have higher level of connections. In a subsequent scanning, if that word is encountered again, its value will go up, while in the absence of it, the activation level of words (neurons) goes down.

A word association normalized matrix was constructed by taking into account the connection strengths among the neurons that represent the top 35 most-used words. It is not a simple co-occurrence matrix. It represents not only the direct co-occurrences among the words, but also their indirect connections. For example, if word 1 and word 2 co-occur, and word 2 and word 3 co-occur, but word 1 and word 3 never co-occur, nevertheless, algorithm links the words 1 and 3 also because of their indirect connection through the word 2. The resultant matrix is a generalized scalar product matrix normalized to approximately plus or minus 1.1. This may be treated as a generic similarities matrix. The resultant matrix gives a better expression than the results obtained from simple co-occurrence of words. Like "Pacific" and "Ocean" do not convey much meaning independently but if the word "wave" comes with this group, it conveys that "wave research on Pacific Ocean".

1.3 Structural equivalence blocks as specialty areas

Lorrain and White[10] have proposed that if nodes are people, then social positions may be conceived as equivalence classes or "blocks" of people who relate in a similar way to other such blocks. A concrete network can be transformed into a simplified model of itself where the nodes are combined into blocks and the relation(s) between nodes are transformed into the relations between blocks. Ideally, if two nodes (words) have exactly the same pattern of giving and receiving ties, they are structurally equivalent to each other. A set of such nodes jointly occupy a common position in the network.

This is not the same for the groups that are formed through cluster analysis. In a cluster grouping, only strong cohesive linkages among members result in their being in a particular group. In structural equivalence, the main criteria of a member being present in a block is that it has same relationship with all other nodes. Thus, this provides a new method of looking at the relationships[11]. The model proposed by Breiger et al.[12] relies on the iterated correlations. Burt[13] used structural equivalence in studying social contagion and innovation. Doreian and Fararo[14] have used these techniques to study published literature. Words with strong structural connections were observed to be coming in a structurally equivalent block. The connections are mainly associated with properities, types, effects or methods used for the investigations. The blocks are categorized into possible research areas. This assigning is done based on observing the strength of linkages among the words inside the blocks. Further, the context of these words is seen from the titles, i.e., words which are embedded in the titles. Bhattacharya and Basu[15] have used the empirical or operational methods of reducing a concrete social network to a simpler image of itself which is referred to as "block modeling". The method of Structural Equivalence which looks at the relationships among words as well as structural equivalent blocks is more appropriate for mapping the research specialties at the microlevels, as it considers indirect linkages also. As proposed by Doreian and Fararo[14], in the present study the mean densities of the matrix were used as cut-off points to generate image matrices from the density of the blocks. These structures were viewed as reduced images of initial cognitive networks. The image matrices were used to draw network maps.

UCINET software[2] was used to study the structural equivalent blocks and for calculating the Freeman's centrality values of the most-frequently used words.

2 Data cleaning

SCI Database search with "Antarc*" in title, from the year 1980 through 2004 (25 years), retrieved a total of 10 942 records. Following synonyms and word variants were clubbed to bring similar words together. It ensured that the words with similar meaning were placed together and were not listed under variant entries.

All "Antarctica" words were replaced by "Antarctic".

All "Island" were replaced by the word "Islands".

All "Waters" were replaced by the word "Water".

The Words—"Art", "Sp", "Superba", "Land", "Late", "Polar", "Sub", "Study", etc. were kept excluded from the analysis.

3 Results and discussion

The rank-ordered list of 13 672 words was prepared from which top 35 most-used words were selected to produce the co-occurrence (co-word) matrix. The descending frequency list and alphabetically sorted list are given in Table 1. A perusal of Table 1 revealed that the word "Ice" depicted the maximum frequency of 1 681, which indicates that ice-related research dominates in the area Antarctic science. It is followed by the words "Sea" (frequency of 1 040) and "Islands" (frequency of 921) related research. The presence of words like "Peninsula" (word frequency of 463) and "Weddle" in the frequency list signified the growing importance of research on geographical locations. It is also observed that a considerable amount of research is underway to uncover the composition of various attributes as the word "Composition" has also recorded its presence in the frequency list. The neural network parameters used for the analysis are given in Table 2.

Table 1 Frequency statistics of the most-used words in Antarctic Science subject specialty

Descending Frequency List					Alphabetically Sorted List				
Word	Word Frequency		Case		Word	Word Frequency		Case	
	Frequency	Percentage	Frequency	Percentage		Frequency	Percentage	Frequency	Percentage
Ice	1 681	12.3	5 318	38.9	Bay	263	1.9	1 011	7.4
Sea	1 040	7.6	3 683	26.9	Changes	226	1.7	881	6.4
Islands	921	6.7	3 052	22.3	Composition	228	1.7	886	6.5
Water	628	4.6	2 292	16.8	Distribution	376	2.8	1 446	10.6
East	621	4.5	2 206	16.1	East	621	4.5	2 206	16.1
Peninsula	463	3.4	1 698	12.4	Euphausia	251	1.8	929	6.8
Southern	444	3.2	1 679	12.3	Evidence	284	2.1	1 072	7.8
Species	396	2.9	1 439	10.5	Fish	353	2.6	1 246	9.1
Krill	393	2.9	1 358	9.9	Ice	1 681	12.3	5 318	38.9
Distribution	376	2.8	1 446	10.6	Implications	264	1.9	1 030	7.5
Ocean	360	2.6	1 312	9.6	Islands	921	6.7	3 052	22.3
Ross	359	2.6	1 363	10.0	Krill	393	2.9	1 358	9.9
Fish	353	2.6	1 246	9.1	Lake	281	2.1	947	6.9
Marine	320	2.3	1 178	8.6	Marine	320	2.3	1 178	8.6
Ozone	294	2.2	903	6.6	McMurdo	227	1.7	865	6.3
West	291	2.1	1 066	7.8	Measurements	229	1.7	862	6.3
Evidence	284	2.1	1 072	7.8	Observations	238	1.7	883	6.5
Surface	282	2.1	1 058	7.7	Ocean	360	2.6	1 312	9.6
Lake	281	2.1	947	6.9	Ozone	294	2.2	903	6.6

(To be continued on the next page)

(Continued)

Temperature	275	2.0	1 041	7.6	Peninsula	463	3.4	1 698	12.4
Implications	264	1.9	1 030	7.5	Polar	242	1.8	888	6.5
Bay	263	1.9	1 011	7.4	Ross	359	2.6	1 363	10.0
Weddell	256	1.9	981	7.2	Sea	1 040	7.6	3 683	26.9
Shelf	254	1.9	968	7.1	Sheet	229	1.7	879	6.4
Euphausia	251	1.8	929	6.8	Shelf	254	1.9	968	7.1
Snow	246	1.8	896	6.6	Snow	246	1.8	896	6.6
Polar	242	1.8	888	6.5	Southern	444	3.2	1 679	12.3
Observations	238	1.7	883	6.5	Species	396	2.9	1 439	10.5
Station	237	1.7	900	6.6	Station	237	1.7	900	6.6
Measurements	229	1.7	862	6.3	Study	220	1.6	852	6.2
Sheet	229	1.7	879	6.4	Surface	282	2.1	1 058	7.7
Composition	228	1.7	886	6.5	Temperature	275	2.0	1 041	7.6
McMurdo	227	1.7	865	6.3	Water	628	4.6	2 292	16.8
Changes	226	1.7	881	6.4	Weddell	256	1.9	981	7.2
Study	220	1.6	852	6.2	West	291	2.1	1 066	7.8

Note: The case frequency indicates the total number of windows in which a word was used.

Table 2 Neural Network Parameters used for the analysis

The used euralnetwork parameters	Number
Total words (The total number of words in the text)	13 672
Total unique words (The number of words used in the analysis)	35
Total episodes (The total number of windows used in the analysis)	13 669
Cycles-Periodically updated network	1

3.1 Thematic blocks

Since the concepts and ideas are generated through the co-occurrences of words, structurally equivalent blocks were constructed to visualize the neural connections among the significant words.The blocks signify areas of research. Four blocks model solution was found to be optimum at R^2=0.998. 35 most frequently used words into four blocks is depicted in Table 3. Block 1 and Block 2 have 3 words each, while Block 3 is the biggest congener of 18 words, followed by Block 4 with 9 words.

Table 3 Block assignment (thematic blocks) of 35 most-used words

Thematic block	Most-used word
Block 1	Ice, Island, Sea Water
Block 2	Euphausia (superba), Krill, Measurement
Block 3	Bay, Distribution, East, Implication, Lake, Marine, Ocean, Peninsula, Polar, Ross, Sheet, Shelf, Snow, South, Species, Study, Surf, Weddle, Sea, West,
Block 4	Changes, Composition, Evidence, Fish, McMurdo, Observation, Ozone, Station, Temperature

The density matrix of these four blocks is given in Table 4. The density table was dichotomized using the mean density of -0.33 using the following rule:

$$y(I, j) = 1 \text{ if } x(I, j) > -0.33, \text{ and 0 otherwise.}$$

Table 4 Density metrics of four blocks of word association

Thematic block	Block 1	Block 2	Block 3	Block 4
Block 1	0.81	-0.93	0.71	-0.87
Block 2	-0.93	0.78	-0.83	0.72
Block 3	0.71	-0.83	0.62	-0.78
Block 4	0.87	0.72	-0.78	0.67

The binary matrix derived from the value matrix, is given in Table 5. The binary matrix was used to draw the network map.

Table 5 Binary matrix derived from the value matrix

Thematic block	Block 1	Block 2	Block 3	Block 4
Block 1	1	0	1	0
Block 2	0	1	0	1
Block 3	1	0	1	0
Block 4	0	1	0	1

3.2 Network map

Block to block network map was generated to find the linkages of the words in one block with those of the other blocks (Figure 1). Four distinct blocks have emerged, as can be visualized from Figure 1. The network map has generated two distinct clusters, one between Block 1 and Block 3, and the other between Block 2 and Block 4. Block 1 contains words like “Ice”, “Island”, “Sea” and “Water”, while Block 3 mostly identifies geographical locations, indicating prevalence of research on this subject in the stated locations like Peninsular regions, Ross islands, etc. “Changing” scenarios have been the focus of a substantial amount of research in this area. This may be due to the worldwide concerns about “global warming” and its relation with Antarctic ice-shelf or changing fish population in the Antarctic water. Substantial research has been done in and around the McMurdo station of the USA, which is Antarctic largest community. USA sends maximum number of expedition members to Antarctica. They maintain a huge research base in the icy continent, and USA is the largest producer of scientific information, as is evident from the published papers on Antarctic Continent[3].

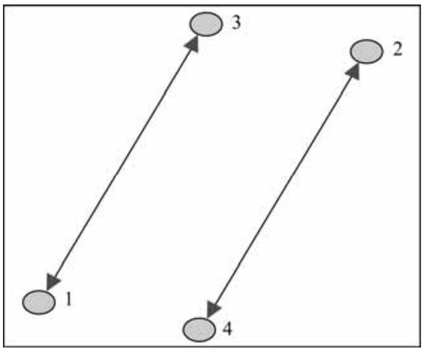

Figure 1 Network map of thematic blocks in Antarctic science research.

Block 2 is consisted of words like “Krill” and its scientific name “*Euphausia*” and “Measurement” and is linked with the Block 4 consisting of words like “Changes”, “Composition”, “Fish”, etc. It is evident from this block modeling that there is prevalence of research on the biological resources like krill, fish, etc., in the Antarctic water.

3.3 Degree centrality

The Freeman’s degree centrality, normalized degree centrality and share of centrality of words in the field of Antarctic science were calculated and are recorded in Table 6. A perusal of Table 6 revealed that among 35 most-used words, the word “Composition” has emerged as the topmost-connected word with a centrality value of 10.56. The next words with high centrality values are “Sea” (5.59), “Ice” (5.57), “Water” (4.96), “Island” (4.95), “East” (4.94), and “Southern” (4.70). The higher values of centrality depict their more use with other words in the area of Antarctic science. It also shows that a considerable amount of research is underway to uncover the “Composition” of various attributes. The relatively higher values of words “Peninsula”, “Weddle” signify the importance being accorded to the research on geographical locations.

4 Conclusions

The study has concluded that the “co-word analysis” can provide a direct quantitative way of linking the conceptual contents of scientific publications by classifying them on the basis of the occurrence of similar word-pairs. Such a co-word structure can highlight the on-going research in a scientific field. In the present study conducted on 10 942 articles published in SCI journals during the period 1980—2004, a total of 13 672 unique words have been identified. The word “Ice” had the maximum frequency of 1 681, indicating that ice-related research dominated the area of Antarctic science during the study period. It is followed by the words “Sea” and “Islands”, depicting their high association with Antarctic science research. The presence of words like “Peninsula” and “Weddle” in the frequency list is a pointer to the growing importance of geographical locations in the Antarctic science.

Table 6 Freeman’s degree centrality, normalized degree centrality and share of centrality of words in Antarctic science subject specialty

Sl. No.	Words	Freeman’s degree	Freeman’s Normalized degree	Share of Centrality
1	Sea	5.586	16.428	-0.216
2	Ice	5.57	16.382	-0.215
3	Water	4.964	14.6	-0.192
4	Island	4.948	14.554	-0.191
5	East	4.935	14.515	-0.191
6	Southern	4.697	13.814	-0.182
7	Ross	4.653	13.686	-0.18
8	Implication	4.613	13.567	-0.178
9	Marine	4.607	13.549	-0.178
10	Species	4.573	13.449	-0.177
11	Peninsula	4.562	13.418	-0.176
12	Ocean	4.558	13.405	-0.176
13	Distribution	4.545	13.369	-0.176
14	Snow	4.543	13.36	-0.176
15	Sheet	4.479	13.172	-0.173
16	Surface	4.446	13.077	-0.172
17	Lake	4.429	13.028	-0.171
18	Weddle	4.407	12.963	-0.17
19	Shelf	4.354	12.805	-0.168
20	West	4.336	12.752	-0.168
21	Bay	4.224	12.423	-0.163
22	Study	4.148	12.2	-0.16
23	Polar	4.058	11.937	-0.157
24	Composition	10.563	-31.068	0.408
25	Evidence	-10.58	-31.108	0.409
26	Temperature	-10.63	-31.259	0.411
27	Change	-10.83	-31.84	0.419
28	McMurdo	-10.86	-31.934	0.42
29	Station	-10.86	-31.947	0.42
30	Observation	-10.97	-32.252	0.424
31	Fish	-11.08	-32.595	0.428
32	Ozone	-11.28	-33.168	0.436
33	Measurement	-11.31	-33.265	0.437
34	Euphausia	-11.46	-33.715	0.443
35	Krill	-11.69	-34.372	0.452

Another significant observation is the emergence of the word "Composition" in the frequency list. It has recorded the highest centrality value of 10.56 among the 35 most-used words, which depicts its maximum use with other words in the area of Antarctic science. This means that a considerable amount of research is underway to uncover the "Composition" of various attributes in this scientific specialty. The higher centrality values of words "Peninsula" and "Weddle" also have pointed towards growing interest in research on geographical locations.

The network map generated in the study has two distinct clusters; one between Block 1 (having words like "Ice", "Island", "Sea" and "Water") and Block 2 (having words identifying geographical locations). This shows the prevalence of research in the locations like Peninsula regions, Ross islands, etc. The study has also recorded the focus of research on the "Changing" scenarios which could be due to the worldwide concerns about "global warming", and its relation with Antarctica ice-shelf or changing fish population in the Antarctic waters. The block modeling has also depicted prevalence of research on the biological resources like krill, fish, etc., in Antarctic water.

To sum up, the study is a valuable supplementation to the existing knowledge in the field of Antarctic science in terms of identification of present research areas and emerging areas in this subject specialty. In understanding research dynamics in Antarctic science, this study will help to draw a bigger picture of the field by complementing the previously done macro level studies[3-4].

Acknowledgements The authors express their deep appreciation and gratitude to Prof. Joseph Woelfel, Department of Communication, State University of New York at Buffalo for his guidance and support for the study. The authors are also thankful to Prof. Robert Hanneman, University of California, Riverside, USA for providing suggestions in doing the analysis.

References

1 Leydesdorff L. Full text analysis of scientific articles in: The challenges of scientometrics: The development, measurement, and self-Organization of scientific communication. Universal Publishers, USA, 2001.

2 Borgatti S P, Everett M G, Freeman L C. UCINET for Windows: Software for social network analysis. Harvard: Analytic Technologies, 2002.

3 Prabir G D. National and institutional productivity and collaboration in Antarctic science: An analysis of 25 years of journal publications (1980–2004). Polar Research, 2007, 26(2): 175-180.

4 Prabir G D. Intellectual structure of Antarctic science: A 25-year analysis. Scientometrics, 2008, 77(3): 389-414.

5 Garfield E. The value of articles in bibliographic citations. Current Contents, 1968, 5: 25-26.

6 Garfield E. ISIs master list of title words provides a special perspective on science and scholarly activity. Part 2. Comparative etymology of neologisms and research fronts. Current Contents, 1986, 28: 214-221.

7 Garfield E. Keyword Plus: ISI's breakthrough retrieval method. Part 1. Expanding your searching power on Current Contents on Diskette. Current Contents, 1990, 32: 295-299.

8 Woelfel J K. CATPAC II User's Manual. Rah Press Place. New York. 1998.

9 Raye K J H, Johnson C L, Mitchell M K, et al. FMRI investigations of left and right PFC contributions to episodic remembering. Psychobiology, 2000, 28: 197-206.

10 Lorrain F, White H D. Structural equivalence of individuals in networks. Journal of Mathematical Sociology, 1971, 1: 49-80.

11 Hanneman R. Introduction to social network methods. http://www.faculty.ucr.edu/~hanneman/, 2011.

12 Breiger R L, Boorman S A, Arabie P. An algorithm for clustering rational data with applications to social networks analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology, 1975, 12: 328-383.

13 Burt R S. Social contagion and innovation: Cohesion versus structuralequivalence. American Journal of Sociology, 1987, 92: 1287-1335.

14 Doreian P, Fararo T J. Structural equivalence in a journal network. Journal of the American Society for Information Science, 1985, 36: 28-37.

15 Bhattacharya S, Basu P. Mapping a research area at the micro level using co-word analysis. Scientometrics, 1998, 43(3): 359-372.