The Research of Multi-Layer Topic Map Analysis using Co-word Analysis with Growing Hierarchical Self-organizing Map

[1] Yu-Hsiang Yang, [2] Rua-Huan Tsaih, [3]Huimin Bhikshu
*[1], *Dept. of Management Information Systems, National Chengchi University,*
yuxiang1001@gmail.com
[2], *Dept. of Management Information Systems, National Chengchi University,*
tsaih@mis.nccu.edu.tw
[3] *Dharma Drum Buddhist College, General Education, Taipei National University of the Arts,*
huimin@ddbc.edu.tw

doi:10.4156/jdcta.vol5.issue3.35

Abstract

The purpose of this study was to propose a multi-layer topic map analysis using co-word analysis of informetrics with Growing Hierarchical Self-Organizing Map (GHSOM). The topic map illustrated the delicate intertwining of subject areas and provided a more explicit illustration of the concepts within each subject area. We applied GHSOM, a text-mining Neural Networks tool, to obtain a hierarchical topic map. After taking up one example of altruism in evaluation, we suggest that topic map may disclose some important facts from a whole bunch of data.

Keywords*: Topic-map, Co-word, Growing Hierarchical Self-Organizing Map, GHSOM, Altruism*

1. Introduction

This study proposed a hierarchical mapping model using co-word analysis and Growing Hierarchical Self-Organizing Map (GHSOM) [1,2]. Since Price [3] first suggested the possibility of dynamic mapping using the scientific method, research in bibliometrics and scientometrics has developed techniques to analyze data sets from within publications [4]. Most early work in this field focused on identifying networks (or clusters) of authors, papers, or references. Based on the nature of words, which are important carriers of scientific concepts, ideas and knowledge [5], co-word analysis was also adopted to identify semantic themes [6]. Co-word analysis simplifies and projects data into specific visual representations while maintaining the essential information contained within it.

Noyons [7] suggested that bibliometric mapping of science appeared to have experienced a revival, due to increased interest in information technology, since the mid-1990s. Many studies, such as [7,8,9,10,11,12] have applied bibliometric maps using co-word analysis to visualize cognitive structures, based on scientific topics, as well as the relationships linking them. In particular, Noyons and van Raan [11] adopted the Self-Organizing Map (SOM) technique [13], to apply co-word approach to scientific mapping (i.e. the organization of science based topics). Furthermore, Shih et al [14] and Li and Chang [15] propose a layered knowledge-map using the clustering of keyterms through GHSOM [1,2]. This is an updated version of SOM, enabling the visualization of hierarchical topic maps.

The objectives of this study were to reveal the major topics or conceptual interrelations of research related to altruism as an example, in order to gain a better understanding of the quantitative aspects of recorded data and discover features of research relevant to altruism embedded in the SSCI database. Altruistic behavior is a selfless prosocial behavior for the welfare of others. It is also a traditional virtue in many cultures and a core aspect of various religions such as Buddhism, Islam, and Christianity and so on. Thus, we adopted GHSOM in co-word analysis to cluster the conceptual topics into a representation of dynamic 2-dimentional interrelated structures within the data.

2. Dataset and Method

The dataset used in this study was derived from the SSCI database of the Web of Science, created by the Institute for Scientific Information. It comprehensively indexes over 1,950 journals across 50 social sciences disciplines. It also indexes individually selected, relevant items from over 3,300 of the world's leading scientific and technical journals[1].

An empirical search command was used by “Topic=(altruism) OR Topic=("altruist* behavio*") OR Topic=("helping behavio*") OR Topic=("prosocial behavio*") refined by Document Type= (ARTICLE OR REVIEW) “to retrieve data related to altruism. The documents specifically included articles or reviews in the study. Book reviews, papers of proceeding, letters, notes, meeting abstracts were not taken into consideration. A total of 4,271 papers published between 1956 and 2009 were found.

The study applied co-word analysis with GHSOM to cluster the major topics of a large collection of documents based on research related to altruism, and provide a topical landscape of the field. As with co-citation analysis, co-word analysis has been used to determine the strength of relationships among textual containers, whether the containers are full-text documents, their surrogates, fields within documents (e.g. titles, descriptors), or queries submitted to information retrieval systems. Techniques for the analysis of word co-occurrence are generally similar to those used for co-citation analysis, consisting of cluster analyses, multidimensional scaling methods [16].

Co-occurrence analysis of document content is usually performed on substantive keywords appearing in a bibliographic database record field such as the title, descriptors, or abstract. These fields encapsulate the topicality of a document, although keywords from the body of text could be used as well [16]. The benefits of co-word analysis can he mixed depending on the application such as clustering major topics of a large collection of documents based on their content and providing a topical landscape of a field. Many studies, such as [7,8,9,10,11,12] had applied informetric maps using co-word analysis to visualize cognitive structures, based on scientific topics, as well as the relationships linking them.

Co-word analysis embraces a large number of different methods to determine the clusters of word co-occurrence. For the purposes of the present study, we choose GHSOM used successfully before in comparable studies to identify distinctive clusters of papers [14,15].

Self-Organizing Map was designed according to the concept of unsupervised artificial neural networks to process high-dimensional data and provided visual results [11,13,17,18]. However, SOM requires a predefined number of nodes (neural processing units) and implements a static architecture. These nodes result in a representation of hierarchical relations with limited capability. GHSOM approach was developed to overcome these limitations, and is often applied in field the information extraction [1,2,14,15,19]. GHSOM is based on the characteristic of SOM, but it can automatically grow its own multi-layer hierarchical structure, in which each layer encompasses a number of SOMs, as shown in Figure 1.

The process of applying GHSOM to topic analysis is illustrated in Figure 2. The three phases are: the data preprocessing phase; the clustering phase; and the interpreting phase.

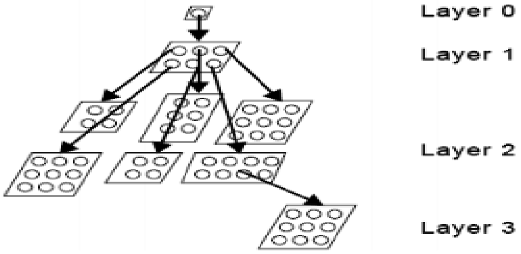

Fig 1. Structures of GHSOM [2]

[1] The SCIE information from http://images.isiknowledge.com/WOKRS49B3/help/WOS/h_database.html#ssci, retrieved at August, 19, 2010.

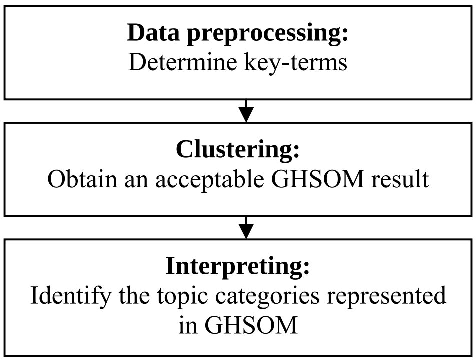
Data preprocessing:
Determine key-terms
Clustering:
Obtain an acceptable GHSOM result
Interpreting:
Identify the topic categories represented in GHSOM

Fig 2. The three phases of the topic analysis process

In the data preprocessing phase, key-terms such as titles, keywords, and subject categories are used to represent the contents of the documents. Meaningful key-terms describing the articles are extracted directly from the documents without any manual intervention. These key-terms are weighted according to a *tf* x *idf* the state-of-the-art weighting scheme shown in equation (1) [2,14,16,20].

$$w_i(d) = tf_i\,(d) \times \log\,(\,N / df_i\,) \qquad (1)$$

In equation (1), $w_i(d)$ represents the weight of the *i*th term in document (d), $tf_i(d)$ represents the number of times the ith term appears in document (d), N represents the total number of documents, and df_i represents how many documents contain the ith term. The weighted value for a term will always be greater than or equal to zero. This weighting scheme assigns high values to terms considered important for describing the contents of a document and discriminating between various documents. A high weight is earned by frequent appearances of a term in a given document, with infrequent appearance of terms within the entire collection of documents. In this manner, weight assignment tends to filter out common terms. Based up on weighting v alues, we selected the top order distinct key-terms for document representation [16,20]. The resulting key-term vectors were used for GHSOM training.

In the clustering phase, the GHSOM experiment[2] was conducted through the trial and error method, using various values for breadth and depth and different normalizations to gain an acceptable GHSOM model for the analysis. The results of GHSOM are shown as Figure 2.

In the interpreting phase, for each node of GHSOM of the first-layer and some nodes of the second-layer which will be re-grouped into the layer 3, we counted the df_i value of each key-term in all articles cluster them into a particular node and assigned a key-term with the highest df_i value (or several key-terms if their df_i values were very close) as the topic category. If there were more than five topics, we would denote it as multidisciplinary. For the remaining nodes, the utmost five important key-terms would be automatically assigned by the GHSOM using the *tf* x *idf* weighting scheme.

3. Results

3.1. Overview of Productivity

A total of 4,271 papers related to altruism were retrieved from the SSCI database. Figure 2 shows the number of published papers on the topic of altruism, between 1956 and 2009. According to numerical data, a large number of research papers published in recent years (2007-2009) have been catalogued in the SSCI database, with distribution rates of 326 (7.6 %), 416 (9.7 %), and 406 (9.5 %) against the total number of papers, respectively. It has also been observed that a trend in the growth of these numbers appears to have begun in 1991. Figure 3 shows the number of citations of published

[2] We used GHSOM toolbox in the Matlab R2007a® package to conduct the GHSOM experiment.

papers related to altruism made each year. The figures suggest that the number of these citations has also been increasing. Clearly, the topic of altruism has received a great deal of attention from researchers in the fields of social sciences.

The ten countries ranked as the top publishers of catalogues in the SSCI database are illustrated in Figure 4. The figure shows how the USA is the dominant country, followed by England, Canada and so on.

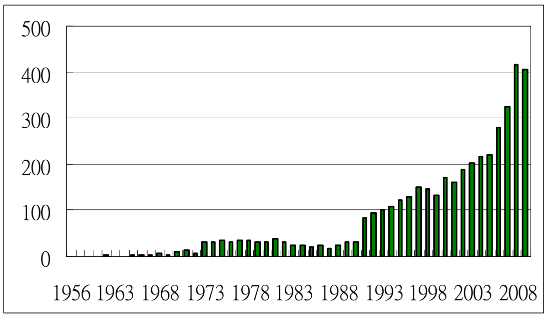

Fig 3. Number of published papers

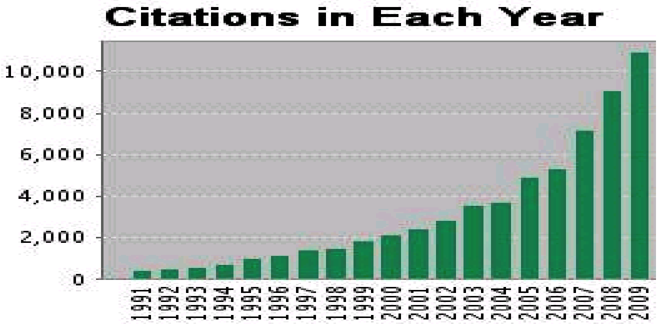

Fig 4. Number of citation (source: ISI Web of Science)

Figure 5 provides the top ten subject areas in which altruism was most widely studied, within the social sciences. The most highly ranked subject area was economics, followed by social psychology and developmental psychology related to altruism. It was also observed that the main part of studies was related to psychology, accounting for over 40 % of total.

Table 3 shows the 10 articles receiving the most citations. The results show how Trivers [21] was an icon in altruism; however, if we take into account the average number of citations per year, the work of Goodman [22] was more influential than that of Trivers [21]. Most of the articles were in the fields of psychology and biology. In addition, Robert Goodman and Ernst Fehr had the two most cited articles, shown in bold text.

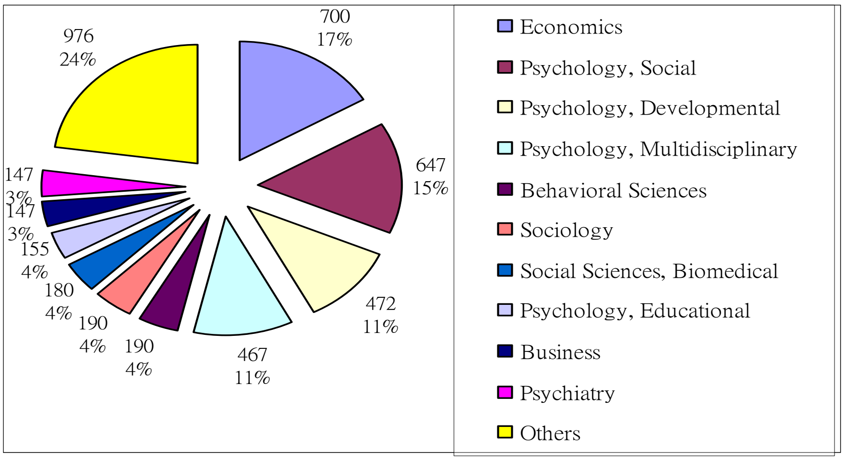

Fig 5. Top 10 subject areas for articles related to altruism

Table 3. The10 most cited articles (Data retrieved on August 23, 2010)

Authors	Article	Year	TC	ACPY
Trivers, R. L.	Evolution of reciprocal altruism	1971	**2,410**	**60**
Goodman, R.	**The strengths and difficulties questionnaire: A research note**	**1997**	**1,025**	**73**
Fehr, E. and Gachter, S.	**Altruistic punishment in humans**	**2002**	**591**	**65**
Pratto, F., Sidanius, J., Stallworth, L. M. and Malle, B. F.	Social-dominance orientation - a personality variable predicting social and political-attitudes	1994	**529**	**31**
Andreoni, J.	Impure altruism and donations to public-goods - a theory of warm-glow giving	1990	**455**	**21**
Goodman, R.	**Psychometric properties of the strengths and difficulties questionnaire**	**2001**	**434**	**43**
Conner, M. and Armitage, C. J.	Extending the theory of planned behavior: A review and avenues for further research	1998	**381**	**29**
Krebs, D. L.	Altruism - examination of concept and a review of research	1970	**370**	**9**
Fehr, E. and Fischbacher, U.	**The nature of human altruism**	**2003**	**361**	**45**
Colquitt, J. A.	On the dimensionality of organizational justice: A construct validation of a measure	2001	**359**	**36**

TC: times cited; ACPY: Average Citations per Year

3.2. GHSOM and Topic Analysis

Through the process of applying GHSOM to topic analysis as showed in Figure 2, we obtained the result as showed in Figure 7 in the clustering phase. The model comprised three layers and 56 nodes. All 4,271 articles were clustered into a SOM of 2 x 3 nodes in layer 1, where all articles that had been clustered into the six nodes were further re-grouped into a SOM of 2 x 2 (i.e. node 1, 2, 3, 5, and 6) or 2 x 3 (i.e. node 4) nodes in layer 2, respectively. The articles clustered into nodes 4.1, 4.3 and 6.2 were further re-grouped into a SOM of 2 x 2 nodes in layer 3. The articles clustered into node 4.1.4 were further re-grouped into a SOM of 2 x 2 nodes in layer 4.

In the interpreting phase, for each node of GHSOM, we count the df_i value of each key-term in all articles cluster them into a particular node and assigned a key-term with the highest df_i value (or several key-terms if their df_i values were very close), as the topic category. If there were more than five topics, we would denote it as multidisciplinary. The results are presented in Figures 8, 9, and 10 , in which the number in the parenthesis refers to the number of clustered articles. For instance, there were 283 articles clustered into node 1, and based upon the interpretation, it was named the "psychology applied in management and business category"; 434 articles in node 2 as "the economics category", 1308 articles in node 3 as the "psychology, multidiscipline category", 603 articles in node 4 as the "economics & multidiscipline category", 518 articles in node 5 as the "psychology, developmental category", 351 articles in node 6 as the "evolution category". Based on these dominant topical clusters in the collection of articles, further specific topics were obtained in layer 2, (Figure 9). For instance, articles in the "psychology applied in management and business category" were further re-grouped into sub-category topics including "organizational", "work predictors", "performance", and "applied psychology" in node 1.1; the sub-category topics including " management", " business, and applied psychology" in node 1.2; the sub-category topics including "management", "performance", "applied psychology", "work", and "job" in node 1.3; and sub-category topics including "donation", "attitudes", "biomedical social sciences", "PEOH (public, environmental, and o_ccupational health)", and "transplantation" in node 1.4. Articles in a number of nodes of layer 2 (that is, nodes 3.1, 3.3, 3.5, 4.1, 5.2, and 5.4) were further re-grouped into more specific subcategories in layer 3, as shown in Figure 10.

The interpretation results for the second- and third-layer of GHSOM shown in Figure 9 and 10 were more delicate than those in Figure 5 were. It was observed that the interpretation results for the second-layer were more specific than in the first-layer. For instance, articles in nodes 1.1 and 1.3 belonged to the category of "psychology applied in management and business" in node 1, but they both have further differentiations. Node 1.3 focuses on organizational, work, predictors, and performance, while node 1.1 focuses on the overall aspect of management and business. Another interesting observation shown in Figure 9 is that the two neighboring nodes are much more closely related than the remote nodes. For example, articles clustered in node 6.4 at the bottom-right corner of Figure 9 are obviously very different from those clustered in node 1.1 in the top-left corner of Figure 9, but they are more closely related to those in nodes 6.1, 6.2 and 6.3.

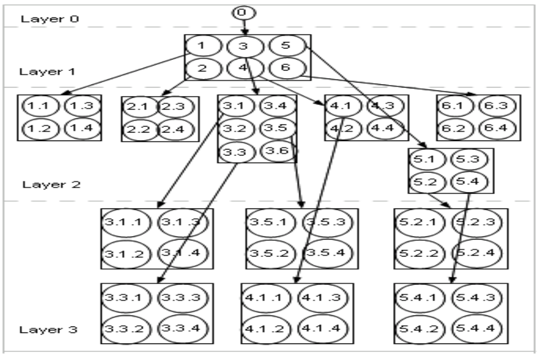

Fig 7. The GHSOM result

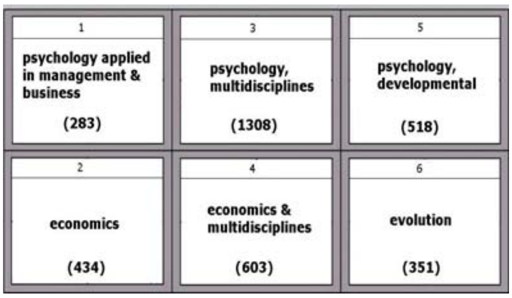
1
psychology applied in management & business
(283)
3
psychology, multidisciplines
(1308)
5
psychology, developmental
(518)
2
economics
(434)
4
economics & multidisciplines
(603)
6
evolution
(351)

Fig 8. First-layer interpretation results of GHSOM.

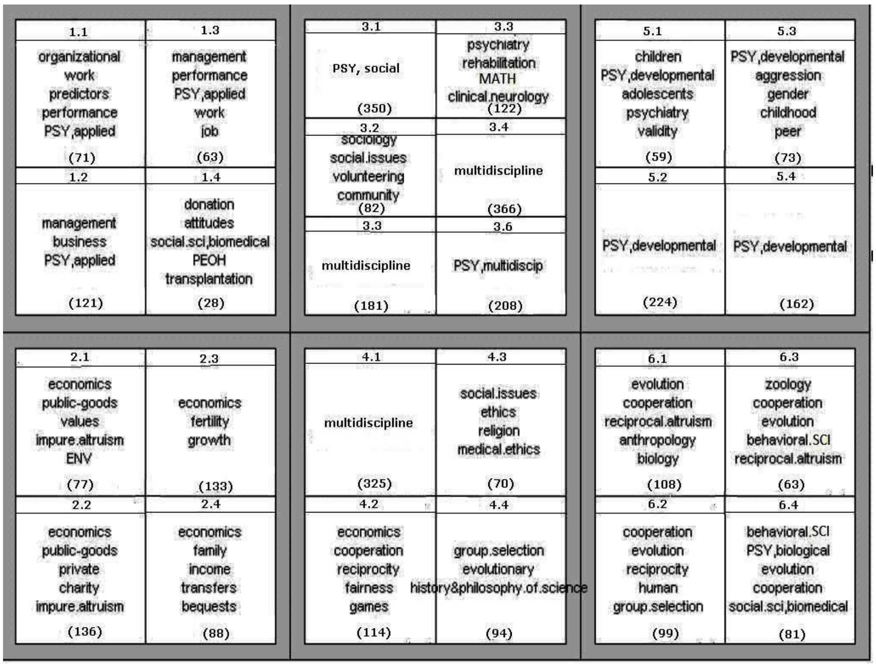

Fig 9. Second-layer interpretation result of GHSOM. PSY is the abbreviation for psychology; PEOH refers to public, environmental, and occupational health; SCI is science; MATH is mathematics; ENV is environment.

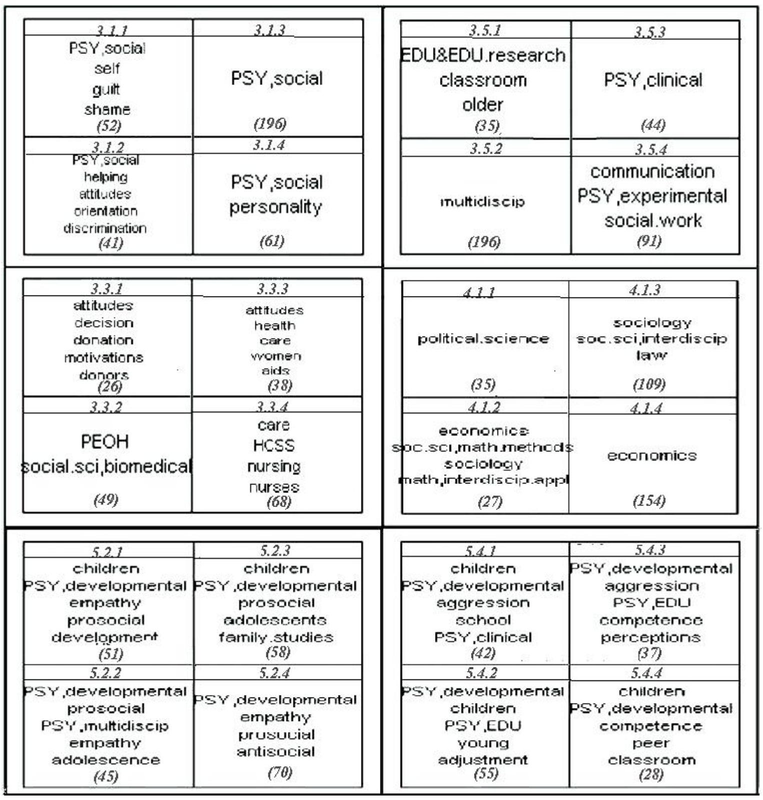

Fig. 10. Third-layer interpretation result of GHSOM. HCSS is the abbreviation for health care sciences and services; SOC refers to social; SCI is science; EDU is education.

4. Discussion

The results of the GHSOM complied with the subject area rankings in the first layer, and provided more explicit topics implying the interrelationship of the different subject areas in the second or third layers. For example, the sociology in Figure 5 is in the node 4.1.2 of Figure 9, indicating that research regarding altruism related to sociology was relevant to economics, social sciences with mathematical methods, and interdisciplinary applied mathematics. The first-layer interpretation results give the disciplinary map while the second- and third-layer interpretation results present topic maps indicating the relationship among different disciplines.

Furthermore, the evolution category in node 6 with 351 papers did not appear in Figure 5, where the top ten subject areas are listed. The category of evolution in node 6 co-exists with a number of disciplines such as anthropology and biology in node 6.1, zoology in 6.3, and biological psychology and biomedical social sciences in 6.4 , which implies that these studies were interdisciplinary and focused on evolution. To be more precise, the topics in nodes 6.1, 6.2, 6.3 and 6.4 explained why evolution had become one of the major clusters in Figure 8. For example, node 6.1 tells us that groups of research associated with anthropology were strongly related to cooperation and reciprocal altruism, which was based on the research of biological evolutionary findings. At the same time node 6.2 illustrates how the groups of cooperation and reciprocity were correlated with the research of group selection in biology. In addition, node 6.4 shows that the group of biomedical social science researchers targeted cooperation, which is based on the research of behavioral sciences and biological psychology. Node 6.3 gives us a hint that the behavioral sciences group applies the ideas of cooperation and reciprocal altruism in zoology. The above four groups of research are all based on evolutionary concepts which have had a long and sound history of since Darwin's great discovery. More specifically, the works such as Trivers [21], Fehr and Fischbacher [23], and Fehr and Gachter [24] in Table 3 could explain the above suggestion, because their articles were prominently cited in research related to altruism. This highlights the need for co-citation analysis in the future.

5. Conclusions

To sum up, we observed a steady growth in the number of papers related to altruism between the years of 1956 and 2009 in this study. The three most influential authors were Trivers, R., Goodman, R., and Fehr, E. with regard to the number of times cited. The Goodman's paper is a standard psychometric tool that has become widely used in psychology measuring prosocial or altruistic behaviours. Obviously, Trivers, R is a giant of the field, so it was reassuring he comes out on top!
The study also shows that the variety of research appeared to be scattered across a wide range of subject areas, and that the three main subject areas were primarily within the fields of economics, psychology, and sociology. However, the GHSOM tool had all of the benefit of SOM, in providing a map from a higher dimensional input space to a lower dimensional map space, as well as providing a global orientation of independently growing maps in the individual layers of the hierarchy, which facilitated navigation across branches. The topic map using GHSOM in co-word analysis illustrated the delicate intertwining of subject areas and provided a more explicit illustration of the concepts within each subject area. The result of the topic map may indicate that the concept of evolution played an importance role in multidiscipline within the research related to altruism. This suggests that topic map may disclose some important facts from a whole bunch of data.

6. References

1. Dittenbach M, Rauber A, Merkl D (2002) Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. Neurocomputing 48: 199-216.
2. Rauber A, Merkl D, Dittenbach M (2002) The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks 13: 1331.
3. Price DJS (1965) Networks of scientific papers. Science 149: 510-515.
4. Leydesdorff L (1987) Various methods for the mapping of science. Scientometrics 11: 295-324.

5. Van Raan AFJ, Tijssen RJW (1993) The neural net of neural network research. Scientometrics 26: 169-192.
6. Boyack K W, Klavans R, Borner K (2005) Mapping the backbone of science. Scientometrics 64: 351-374.
7. Noyons E (2001) Bibliometric mapping of science in a policy context. Scientometrics 50: 83-98.
8. Hassan E (2003) Simultaneous mapping of interactions between scientific and technological knowledge bases: the case of space communications. Journal of the American Society for Information Science and Technology 54: 462-468.
9. Chau M, Huang Z, Qin J, Zhou Y, Chen H (2006) Building a scientific knowledge web portal: The NanoPort experience. Decision Support Systems 42: 1216-1238.
10. Grupp H, Schmoch U, editors (1992) Perception of scientification as measured by referencing between patents and papers. Heidelberg: Spring Verlag. 73-128 p.
11. Noyons E, van Raan A (1998) Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. Journal of the American Society for Information Science 49: 68-81.
12. Ding Y, Chowdhury G, Foo S (2001) Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing & Management 37: 817-842.
13. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological cybernetics 43: 59-69.
14. Shih J, Chang Y, Chen W (2008) Using GHSOM to construct legal maps for Taiwan's securities and futures markets. Expert Systems with Applications 34: 850-858.
15. Li ST, Chang WC (2009) Design And Evaluation Of A Layered Thematic Knowledge Map System. Journal of Computer Information Systems 49.
16. Wolfram D (2003) Applied informetrics for information retrieval research. Westport, Connecticut: Greenwood Publishing Group.
17. Campanario J (1995) Using neural networks to study networks of scientific journals. Scientometrics 33: 23-40.
18. Kohonen T, Kaski S, Lagus K, Salojarvi J, Honkela J, et al. (2000) Self organization of a massive document collection. IEEE Transactions on Neural Networks 11: 574-585.
19. Tsaih R, Lin W, Huang S (2009) Exploring Fraudulent Financial Reporting with GHSOM. Intelligence and Security Informatics: 31-41.
20. Salton G (1989) Automatic text processing: the transformation, analysis, and retrieval of information by computer. Reading, MA: Addison-Wesley.
21. Trivers RL (1971) EVOLUTION OF RECIPROCAL ALTRUISM. Quarterly Review of Biology 46: 35-57.
22. Goodman R (1997) The strengths and difficulties questionnaire: A research note. Journal of Child Psychology and Psychiatry and Allied Disciplines 38: 581-586.
23. Fehr E, Fischbacher U (2003) The nature of human altruism. Nature 425: 785-791.
24. Fehr E, Gachter S (2002) Altruistic punishment in humans. Nature 415: 137-140.