Study of Knowledge Evolution in Parallel Computing by Short Texts Analysis

Pavel Makagonov[1] and Alejandro Ruiz Figueroa[2]

[1] Postgraduate Division of Mixteca University of Technology, Huajuapan de León, Oaxaca,
69000, México
mpp@mixteco.utm.mx

[2] Institute of Electronic and Computation of Mixteca University of Technology
Huajuapan de León, Oaxaca, 69000, México
figueroa@nuyoo.utm.mx

Abstract. The problem of measuring and predicting the future of various branches of science is discussed. We propose an economical approach that is useful for the estimation of the stage of development for any branch of "normal" science with the help of abstract flow analysis. For this goal it is necessary to collect large amounts of abstracts uniformly distributed in years. As abstracts are poor knowledge objects, we use the procedure of aggregation in its annual sum of texts as an elemental unit for cluster analysis. For cluster analysis we use the tool kit «Visual Heuristic Cluster Analysis for Texts» developed earlier by one of the co-authors, with K. Sboychakov. To determine the topic of the cluster, we propose to use chapters of manuals and articles principal in the procedure of pattern recognition.

1 Introduction. Problems of Scientometric and Practical Demands

The problem of resource distribution between different branches of investigation call for evaluation and quantification of the scientific activity, its productivity and results. Public institutes involved in the process of sharing of restricted economical resources for investigation need tool kits for the analysis of effectiveness of their policy of investment and to help to improve it according to a plan.

In fact it is not possible to predict the time and the place of the appearance of new inventions, but it is possible to predict the development of "normal" science [1] that is gradually improving results and that these results warrant capital investment in investigation.

"Normal" science is less uncertain than the forefront of science. "Normal" science bases on antecedent results (the primaries of which as a rule are unexpected, unforeseen) and a study of predecessors gives us the possibility to predict the tendency of development of different (distinguishable) branches of science and gives a reason to correct financial planning.

Quantitative parameters of the system life cycle can be described as an S-curve [2]. When the system reaches the limits of its possible development it changes qualitatively or is substituted by another system. "Normal" science is growing in the stage of system development and declining about the moment of obsolescence.

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 439–445, 2004.

We propose a method that does not have the ability to predict a character of qualitative changes but can be used for fixing a time when interest in the subject of investigation falls precipitously or begins to grow up. To implement this approach it is necessary to have the criteria of the prospects for the branch of science or for the subject of investigation.

For problems of this type different methods of Scientometric and Bibliometric Mapping [3, 4, 5, and 6] are used. These methods are laborious, and demand a great quantity of articles that must be paid for in advance. Large scientific bodies, Government organizations and big companies can use these articles because of their financial ability. Our idea is to develop a simple tool kit that can be used by public institutes that are responsible for the financial support of scientific investigation and by investigators who are in the initial part of their activity, and who only have access to free abstracts of articles on the Internet, and need to investigate different branches of their science.

2 Preparing Samples for Revealing Models of Scientific Nowledge Flows

Our approach is based on an analysis of a corpus of articles' texts or at least abstracts' texts for a sufficient period of years in a special circle of problems or in a partial branch of science.

Every topic has a typical time period of substantial development. We chose a topic that has a rather short history of development (about 30 years) and has developed in the last 10 years very intensively. This topic is Parallel, Simultaneous, Concurrent, and Distributed Computing.

The characteristics almost coincide with the title of the book "Foundations of Multithreaded, Parallel, and Distributed Programming" [7]. The author of this book earlier, in 1991, issued another book with the title «Concurrent Programming Principles and Practices». So, even the names of these two books show us changes in the point of view of the same subject.

Information on this topic is accessible on the Internet for free. With the availability of this information we collected 710 abstracts on the mentioned topic for the years from 1990 until 2004 (about 50 per year) in the Digital library of IEEE [8].

The criteria for selection of abstracts into the corpus of texts was the presence of one of the key words of the topic (Parallel, Simultaneous, Concurrent, and Distributed) or those equivalents (multithreads, supercomputer, hypercube, cluster of computers etc.) of the same level of abstraction of the "ontology" of this topic.

Our task was to reveal clusters of words that give us a lower level of abstraction (ontology) for our topic.

3 Method of Analysis of Poor Knowledge Data

We used the toolkit Visual Heuristic Cluster Analysis for texts (VHCA for texts) [9] as an elemental step of methodology (algorithm).

We used this tool kit to obtain a Domain oriented dictionary (DOD) for the texts corpus and an image of every text as a vector of quantities of words from the DOD presented in the text [9].

These images were used to form three matrices:

1. matrix "text/word" with quantity of words of the DOD (column) in every text (row) as elements of it;
2. matrix "word/word" with quantity of texts which contain the pairs of words from column and row as elements of it;
3. matrix "text/text" with quantity of the same words of the DOD in every pair of texts as elements of it.

For construction of the DOD with VHCA for texts we select the words which satisfy the following criteria:

1. The relative frequency of a word in the corpus of texts must be K0 times greater than the relative frequency of the same word in the frequency list of common used lexis (in our case K0=400%).
2. Criterion K1 defines the upper boundary for the minimum of texts' quantity that contains a given word at least once.
3. Criterion K2 defines the lower boundary for the maximum of texts' quantity that contains a given word. (The goal is to exclude common scientific words for all subtopics.)

It is known that abstracts contain about 150-250 words and in very poor condition when K0 = 100% only about 2 - 8 words per one text (abstract) are new candidates to DOD. Real options of K0, K1, and K2 reduce this number 2- 5 times.

That is why abstracts are considered as poor knowledge objects (texts).

To obtain objects with enriched knowledge we join all abstracts of every year to one annual text and obtain about 100 words as new candidates to the DOD.

In figure 1 we present the matrix "text/word" where every text is the sum of abstracts of the same year. Every non-zero element of the matrix is substituted by a rectangle with corresponding darkness of grey color in according to scale located in the left part of the figure.

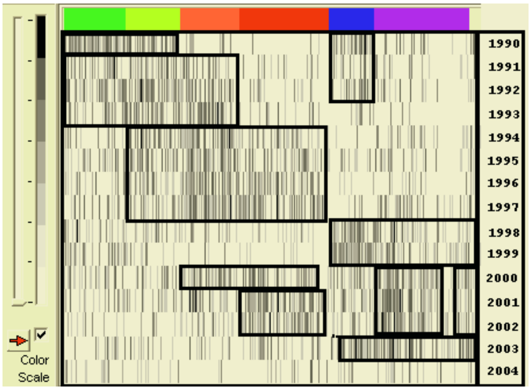

Fig. 1. Matrix text /word for annual sums of abstracts

This matrix is obtained in condition for DOD forming: K1 = 14%; K2= 86% of abstracts. The matrix is presented after clustering with the above mentioned tool kit VHCA. It is then possible to see the tight groups (clusters) of words for some groups of annual texts. If we construct the matrix of the “text/word” type directly for abstracts, we obtain a very rarefied matrix or a poor knowledge object.

The matrix of figure 1 was used for the calculation of the matrix “text/text” presented in figure 2 with outlined clusters of annual texts.

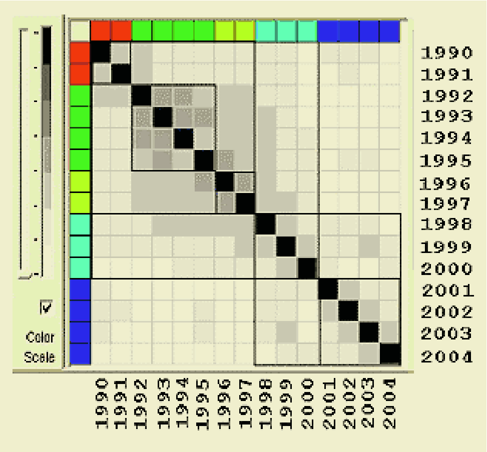

Fig. 2. Matrix “text/text” for annual sums of abstracts

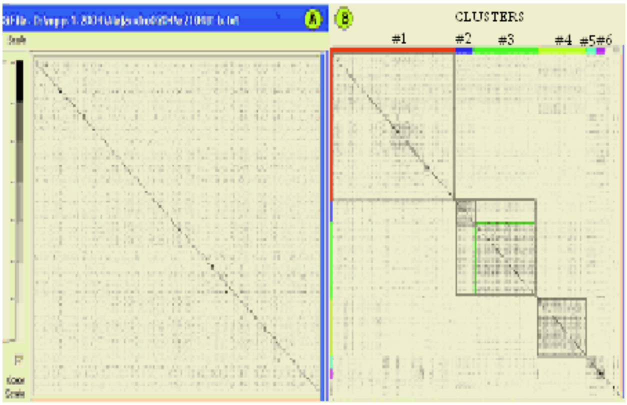

Fig. 3. Matrix “text/text” for 710 abstracts of different years (A) before and (B) after clustering

For a uniform aggregate of texts we can consider this part of our investigation almost complete, but our annual texts are a mixture of different topics (according to the method of preparation) and we can assume a greater diversity of topics for the same years.

For a more detailed investigation of an inhomogeneous aggregate of texts it is possible to obtain greater detail. For this goal we prepare the matrix “text/text” for every

710 abstracts. In figure 3a and 3b this matrix is presented before and after clustering. In figure 3b one can distinctly see 4 large clusters of abstracts that are evidently connected with more partial topics.

The quantity of abstracts in large clusters is represented in figure 4.

The deviation of the number of abstracts in clusters gives us information about the change of interest over time (during years) for topics that are connected with every cluster. The only problem is to get to know the contents of these topics.

For this goal we prepared the set of annual texts for every rich cluster. Every annual text contains a sum of abstracts from the same cluster for the same year. We can do it only for the clusters that contain a large amount of abstracts. Otherwise we would obtain poor knowledge texts. This new set of annual texts can be analyzed by the same method that we use for the initial corpus of texts.

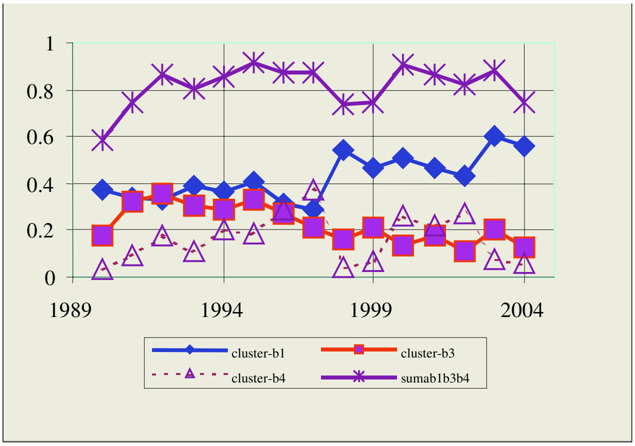

Fig. 4. The quantity of abstracts in large clusters

The difference is that new annual texts are more uniform (homogeneous) themselves. For this reason it is possible to reveal more definitively the topic of every new cluster of annual investigations. Indeed, with help of VHCA we can now obtain partial DODs of new clusters. And these DODs are dictionaries of subtopics that the experienced investigator can reveal by analyzing their contents.

Now the problem is what to do if we do not have an "experienced investigator" at our disposal? In the case of the absence of this specialist it is possible to use the same (or another) tool kit for pattern recognition. For this goal one can combine a new corpus of texts that contain:

1. a set of chapters of manuals or articles with different partial topics (and different sets of special words);
2. a set of annual sums of abstracts for different clusters, which are obtained in previous steps.

The results of this step for patterns and annual texts of two large clusters are presented in figure 5. The visible links (a) between annual texts of different clusters can be explained by using scientific words that are more common in articles than in textbooks. Texts of cluster #3 have fewer links with manuals than cluster #1.

The possible reason is that we do not have correspondent patterns in our aggregate or, alternatively, the topic corresponding to this cluster is a new one or is a discussion of very general problems without special words of partial subjects.

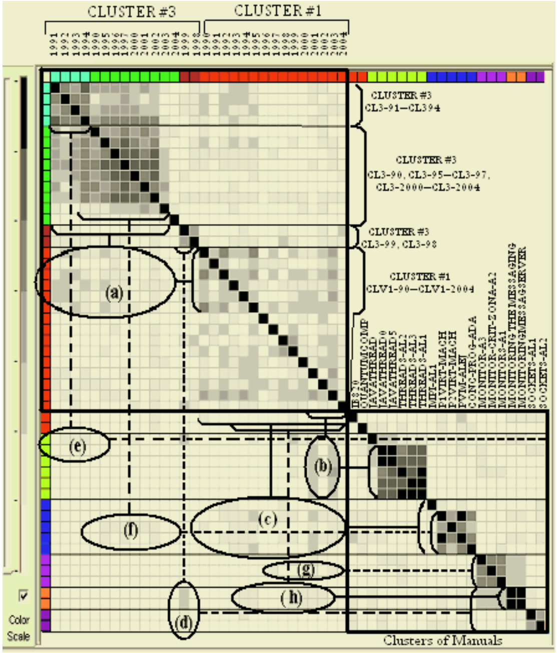

Fig. 5. Matrix "text/text" for patterns and annual texts of two large clusters

Cluster #1 has rather strong links (b) with topic Threads in a later years; with topics MPI (Message Passing Interface) and PVM (Parallel Virtual Machine) in almost all years (c) with the exception of 1998 and 1999, and with topic Monitor and Monitoring (g) and (h).

Cluster #3 has only poor links (e) with Java Threads, with the topic "Sockets" in 1999 (d), and with PVM (f).

Part of the words in manuals can be distributed between different clusters of annual texts. In this case it is possible to count the percentage of representation of every manual topic in every cluster of annual texts.

4 Conclusion

This article exposes only a short description of the proposed method of revealing prospective topics of "normal" science. We have postponed the problem of pattern quality. Those patterns could be selected by experts or with the tool kit used here (or with some other "standard instrument"). In any case, it is necessary to select texts enriched with words in a very narrow subject with minimal intersection of vocabulary with other patterns. This is the subject of future research.

The other problem is common scientific terminology. It can be an obstacle for more refined clustering in the case of poor knowledge texts, but there are some approaches for the solution of this problem.

The third problem is an eagerness of authors to synthesize new complex words that are ignored by our method. It is possible to include in the DOD such words as hypercube, multithreads, supercomputer, etc. in the same row with "cube, thread, computer". There are texts with exotic words that which could be dropped as nonexistent or as errors in writing, if the operator could not add those worlds to the DOD without confidence.

To obtain certain results with our method for poor knowledge data it is necessary to use only clusters with numerous abstracts (not articles). For small but important clusters it is still necessary to buy full texts of articles corresponding to abstracts.

This research is being partially supported by National Council for Science and Technology – CONACYT, México under project N-39011-A. We greatly appreciate the help given to us by Ms. Lorraine Parkin for reviewing the text.

References

1. La Estructura de las revoluciones científicas. Thomas S. Kuhn. Editorial Fondo de Cultura Económica de España. 2000(1962). Traducción: A. Contín. ISBN:84-375-0046-X, pp 320.
2. The Geography of Economic Development: Regional Changes, Global Challenges. Timothy J. Fik. A Division of the McGraw-Hill Companies, pp 260-265.
3. Bibliometric Mapping as a Science Policy and Research Management Tool. Ed C. M. Noyons. DSWO PRESS. Science Studies. 1999 Leiden University , The Netherlands, pp 225.
4. http://148.216.10.83/VIGILANCIA/capitulo_4.htm.
5. http://bvs.sld.cu/revistas/aci/vol10_4_02/aci040402.htm.
6. http://www.campus-oei.org/salactsi/elsa6.htm
7. Foundations of Multithreaded, Parallel, and Distributed Programming. Gregory R. Andrews, University of Arizona. Addison-Wesley 2000.
8. http://search2.computer.org/advanced/simplesearch.jsp
9. Makagonov P., Alexandrov, M., Sboychakov, K. A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: H.A. Kiers et al (Eds.), "Data Analysis, Classification, and Related Methods", Springer, 2000 (Studies in classification, data analysis, and knowledge organization), pp. 83-88.