SATC Tool for Analysis of the Evolution of Scientific Knowledge

Abstract— **In this work we present a software tool that allow identify tendencies that describe the evolution in a discipline of scientific knowledge, where information resources are classified. The tool search support the data mining as part of discovery knowledge process and the identification is supported by production analysis of information resources in science and technology and his visualization in graphs with respect to time.**

To illustrate the use of the tool we utilize the divulgation articles from two electronic magazines in computation of the digital library in ACM (Association for Computing Machinery), classified with multilevel thematic "The 1998 ACM Computing Classification System" that describe this discipline. The tendencies in categories or themes of this discipline are identified by the formulation of directed questions known as models and depending to the utilized model, the results are represented in graphs showing the tendencies, or by graphs considered as knowledge mappings that satisfied the model criterion. The knowledge analysis units are the themes in the first three levels localized in the ACM classification system.

The graphs can help to show the contribution in the themes production or categories of the discipline in a period time, showing production comparatives in his descriptors related. They permit identify in the themes if his popularity increase or in opposite case of others that are not interesting, or they are abandoned in his studied or investigation and for how many time.

Keywords: Database Management; Data Mining; Tendencies in the evolution of scientific knowledge; Software tool

I. INTRODUCTION. IMPORTANCE OF ANALYZING THE EVOLUTION OF KNOWLEDGE

The evolution of scientific knowledge in all themes or categories of investigation present changes and transformations in the time with significance, these changes are originated by economic, politic, social, culture, geographic and technologic changes presented in the world, in the history, factors that reflect the importance to do the analysis [2].

The utility of this analysis to identify the changes or tendencies in the different disciplines of knowledge fundament the decision making under the areas or topics that must be supported by the governments of countries or the responsible entities [1]. In addition, they permit review to the authors by his production affect the tendency and take them into account as knowledge source. Also permit to observe relations between themes in accord his production [6].

The organization of this document since point two commented about related works to this development. The point three describe the study motivation in the CIC-IPN by the knowledge evolution, the formulation of questions related to the knowledge evolution to reply, the selection of information resource, along with the design of a database structure that permit store data of information resource, well as the tool architecture. In the point four are showed results to reply the raised questions and finally in the point five it's presented the opinions and future works.

II. RELATED WORKS

There are diverse the works that over time are realized with the goal of determine the evolution of knowledge scientific in his diverse disciplines, by methodologies and work tools applied in other kind of analysis but that are applied and adapted by scholars of scientometrics and related areas. For example, in [7] Garfield shows by using the citation index and the clusters mapping how the biomedical engineering are developed in the global map of the science in 1984.

In [8], Makagonov exemplifies by using the development of a S model, the temporal tendency of hardware and software in the parallel and distributed computing between 1990 and 2004. In [9], Samoylenko build a map of scientific knowledge in 2 dimensions using minimum expansion trees and groupings between journals.

Finally, in [10] Guan compares the scientific production of China with respect to other Asian countries in the field of semiconductors using bibliometrics indicators and some techniques of the informetrics. There are some similar works in the analysis of scientific knowledge made by science historians as Karl Popper, Thomas Kuhn, Paul Feyerabend [6] that denote the importance of this kind of analysis. With the scientometrics appearance in 1963 as the evaluative science of quantitative aspects of scientific activity [1][2], it have developed techniques and methods that permit know the development and quality of scientific investigation using structured procedures and goals that can be applied across the board, as the work described in the present paper.

III. TOOL DESIGN

A. Motivation of the Study

The need of know which are the investigation centers where certain themes or categories in the computation field are cultivated to establish communication, I've been looking for ways to observe the behavior or evolution in the articles production and the identify the authors or the articles generation centers. Under this approach in this work is defined the evolution in a knowledge field as the production tendency of his information resources in a period of time in a theme or category of a knowledge field.

The questions to resolve are:

Q1 = what themes or categories of Computation have a strong production in the five last years?

Q2 = what themes or categories of Computation are in study or have a strong presence in the investigation into the same Computation?

Q3 = what themes or categories of Computation are into oblivion or no longer grown or have little presence in Computation?

Q4 = How are related the themes or categories of Computation in terms of production of articles?

These questions for his reply require of a data source, with essential data as the generation date and they are organized in terms of a classification into the Computation among other data.

B. Information Resources

For the analysis of the evolution in the field of Computation first it was think in perform an investigation with the thesis documents of master degree of CIC-IPN, but in general terms they lack of a classification and its necessary to do this work, work that it is outside the scope of present work. After to search information resources with a classification, it could to access the information resource of month journal articles into "Communications of the ACM" [3] in a digital format (www.acm.org). The available digital production is in the period 1958-2009 in pdf format but store in text or images.

The information resource in this case are classified, but an analysis under the resource show that only the production from 1981 to 2008 are classified under "The 1998 ACM Classification System" [4]. This period was used to respond the purpose of this investigation. The count of the classified production it can be observed in Fig. 1.

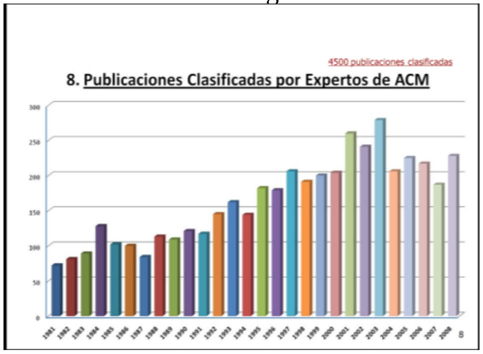


Fig. 1 Statistics of classification in the information resource

Data of the articles collected are:

- Publication year and month
- Title
- Authors
- Publication place
- Abstract
- Introduction
- Conclusions
- References
- The primary and secondary classification of each article.

In order to respond the questions Q1, Q2, Q3 and Q4 are defined a model group that is the next:

- Production and comparative statistics
- Temporal presence of themes in the field of study
- Thematic relation.

C. Models Formulation

The analysis of production and tendency it must be realized through time in different levels of granularity, in this case starting in the minimum unit of time that is the month, up to get to annual production. For the above is necessary to make aggregate to different levels. The dimensions of production and classification facilitate to obtain some interesting graphs through time; more if is raised as specific production models, as like these.

1) Production and Comparative Statistics Model

As a starting point we must to know the total production to revise in certain period of time, this is obtained with the next calculus:

$$SP_{cl} = SP\left(cl, r, u\right) = P_{cl}\left(F_i\right) + P_{cl}\left(F_i + 1\right) + \ldots + P_{cl}\left(F_f\right) \quad (1)$$

$$CP_{cl} = CP(cl, r, u) = \left(SP_c\right) / \left(TP_{tc}\right) \quad (2)$$

Where:

- SPcl = Production of articles in the time range of a selected classification.
- Pcl(Fi) = Production of a classification in the time unit of the analysis.
- CPcl = Porcentual contribution of a classification in the time range.
- TPtc = All production in all selected classification in this time range.
- r = time range, cl classification of interest, u = time unit of analysis, Fi = start year of analysis and Ff finish year of the range r of analysis, and tc is all selected classification.

To obtain graphs of production comparatives only are plotted the functions:

$$FP_{cl}(r,u) = P_{cl}(F_i) + P_{cl}(F_i+1) + \ldots + P_{cl}(F_f) \quad (3)$$

Where FPcl is the Production Function of a specific classification, in a range time r with time units u.

2) Tendency of Presence Model

This model leads to calcule in each point of interest (month, semester, year, etc.) the contribution of each one of the classifications in his three first levels. The trend chart of presence is equal to divide (3) between all production in the point of time to plot, so the values are in the interval [0,1] obtained by the next formula:

$$TP_{cl}(r,u) = P_{cl}(F_i)/TP_{tc}(F_i+1) + P_{cl}(F_i+1)/TP_{tc}(F_i+1) + \ldots + P_{cl}(F_f)/TP_{tc}(F_f) \quad (4)$$

3) Thematic Relation Model

This model requires a classification in order to search the other classifications mentioned in each article and account in a period of time. Our model can be as:

$$FR(c_1,c_2,r) = count(SP_{cl}) \geq n \quad (5)$$

Where FR = Relation frequency or mention count between two classifications c1 and c2 in a range r of time, c1 is the selected classification and c2 is the classification that are mentioned in the articles and to count the mentions is greater than the threshold n indicated.

4) Content Definition of Database

With the foregoing the database must contain three basic elements:

a) Generals of information resource to analyze (id, authors, name, publication date, generating source of the resource, classification in ACM, among other data).

b) Accumulated time of production according the interest of the analysis by unit time (monthly, bimonthly, quarterly, semiannual, and annual).

c) Accumulated of production classification according to each classification of "The 1998 ACM Classification System" by time unit (monthly, bimonthly, quarterly, semiannual, and annual).

In Fig. 2 we can observe the basic elements in the database that support to resolve the interest production models.

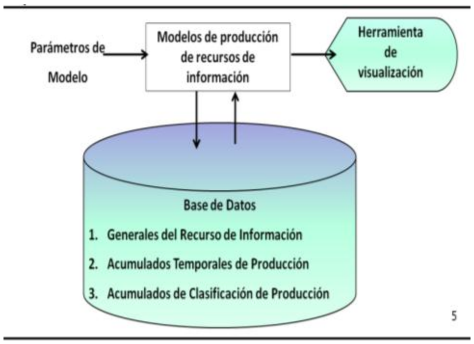

Parámetros de Modelo
Modelos de producción de recursos de información
Herramienta de visualización
Base de Datos
1. Generales del Recurso de Información
2. Acumulados Temporales de Producción
3. Acumulados de Clasificación de Producción
5

Fig. 2 Basic elements in the database

This brings us to design a starnet diagram or star [16] to store accumulates through time and the classification as can be observed in Fig. 3. With the above we have a data cube [17] with name <producción> and whose elements are formed by (x, y, z, p) where x€{artícle}, y€{time}, z€{classification} and p= production value in this point.

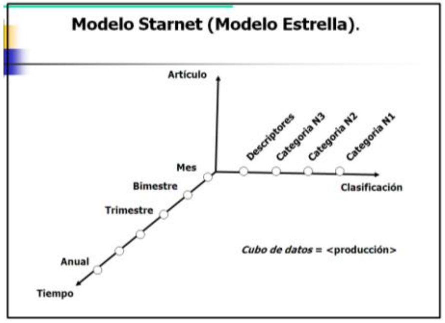


Fig. 3 Star model with dimensions and hierarchies

The data cube <producción> in the time dimension has the hierarchies:

Month ➔ two months ➔ quarter ➔ semester ➔ half ➔ year (6)

And the classification dimension has the hierarchies:

descriptor ➔ category 3 ➔ category 2 ➔ category 1 (7)

The relations between components of this system can be observed in the Fig. 4.

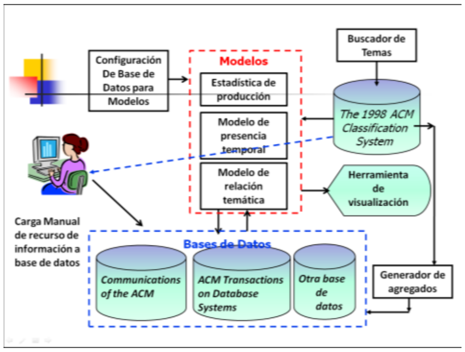

Configuración De Base de Datos para Modelos
Modelos
Estadística de producción
Modelo de presencia temporal
Modelo de relación temática
Buscador de Temas
The 1998 ACM Classification System
Herramienta de visualización
Carga Manual de recurso de información a base de datos
Bases de Datos
Communications of the ACM
ACM Transactions on Database Systems
Otra base de datos
Generador de agregados

Fig.4 Relation between elements in SATC tool

5) Tool Architecture

The build tool named as "Sistema de Analisis Tematico del Conocimiento" (SATC) contain the next elements:

a) Incorporation of an information resource. In this element is described the data organization of the information resource that wants to load to database for subsequent analysis. The load at the moment is manually.

b) Database configuration. In this element is presented the available databases and can be selected the database or information resource to analyze.

c) Production Statistics. In this component could be given the parameters in order to revise the production in a period of time as well as graphic display of production comparatives in these time period.

d) Temporal presence of themes. In this component could be given the parameters in order to revise interest classifications, the graphs are visualized in a time range in production percentages.

e) Thematic relation. In this component are searched for articles of a classification, and are accounted the citation number to others classifications (or thematic relations) in order to show only those that meet at least the value proportionate for a time range.

f) Thematic searcher. This element allows a word given unknown in the category or level is located into the ACM classification, indicating an approximate position.

g) Aggregate generator. This element allows generate the necessary aggregates to different time levels in the time and classification dimensions. The levels or hierarchies in the time dimension are indicated in the hierarchy (6) but in the classification dimension are just:

$$\text{category 3} \rightarrow \text{category 2} \rightarrow \text{category 1}$$

IV. Tool Test

In this point are described the results of respond some questions in the next three models:

a) Statistics and comparatives of production

b) Temporal presence of themes

c) Thematic relation.

A. *Statistics and Comparatives of Production*

For this model, we introduce the next values into the parameters:

- Classification in first level
- Selection of the 11 themes or categories of ACM classification
- Primary classification in the articles
- Graph by years
- Period 1981-2008 (28 years)

In Fig. 5 are visualized the graph obtained with the formula (1) result of this requirement. In this graph it could be appreciate the classification with category K – Computing Milieux have the best production, and there are some with less production.

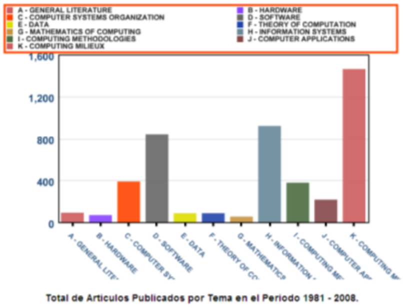


Fig. 5 Production statistics for first level of the classification

In the Fig. 6 now is showed the graph result of capture the next parameters with values:

- Classification in second level
- Selection of theme H (only 4 subthemes of H)
- Primary classification
- Graph by years
- Period 1981-2008 (28 years)

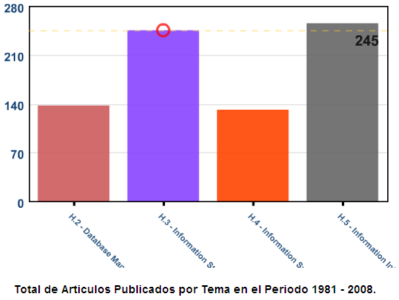


Fig. 6 Production statistics at the second level in some themes of classification H

In this graph also obtained with formula (1), it could be appreciate the themes or categories H.3 and H.5 Information Storage and Retrieval and Information Interfaces and Presentation respectively, have the best production in these time interval.

In Fig. 7 is showed the graph that result of capture the parameters in the comparative of production with values:

- Classification in second level
- Selection of theme H (only 4 subthemes of H)
- Primary classification
- Graph by years
- Period 1981-1990 (10 years)

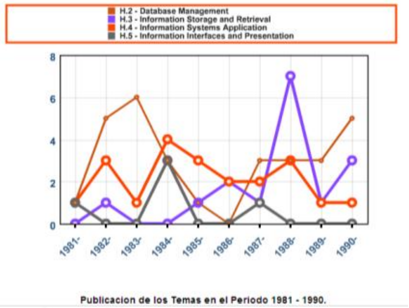


Fig. 7 Comparative at second level in some themes of classification H

In this graph obtained with the formula (3) it could be appreciate and comparable the productions of this four categories in the period of 10 years, showing the maximum and minimum productions in the same period.

B. *Temporal Presence of Themes*

In Fig. 8 are showed the results of capture the parameters for this model with the next values:

- Classification in first level
- Selection of 3 themes
- Primary classification

- Graph by months
- Period 1981-1981 (1 year)

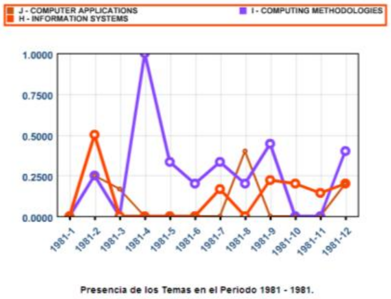


Fig. 8 Graph showing the presence of 3 themes in a year

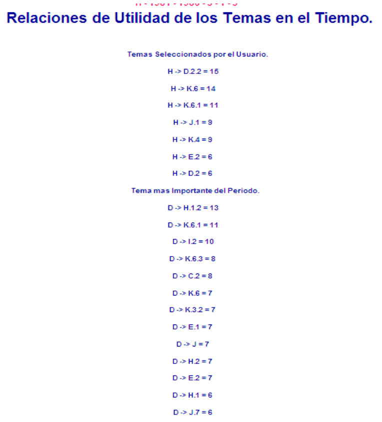
This graph obtained with the formula (4) show that of the three selected categories, the category I – Computing Methodologies has the best presence in this period of time.

C. Thematic Relation

For this model we capture the next values in parameters:

- Classification in first level
- Selection of theme H – Information Systems
- Period 1981 – 1986 (5 years)
- Time umbrae 1 year
- Related theme, it must be mentioned at least other five articles in the period as second classification.

In Fig. 9 and Fig. 10 are showed the graphs that result of this requirement in order to evaluate the formula (5). In Fig. 9 are listed the second categories with mention greater or equal than five give to the articles that have as principal theme H – Information System as category. Although they are visualized other categories by separate satisfied the umbrae of 5 as principal category and secondary category; in this case the category D – Software as principal and the second categories.



Relaciones de Utilidad de los Temas en el Tiempo.

Temas Seleccionados por el Usuario.

H -> D.2.2 = 15
H -> K.6 = 14
H -> K.6.1 = 11
H -> J.1 = 9
H -> K.4 = 9
H -> E.2 = 6
H -> D.2 = 6

Tema mas Importante del Periodo.

D -> H.1.2 = 13
D -> K.6.1 = 11
D -> I.2 = 10
D -> K.6.3 = 8
D -> C.2 = 8
D -> K.6 = 7
D -> K.3.2 = 7
D -> E.1 = 7
D -> J = 7
D -> H.2 = 7
D -> E.2 = 7
D -> H.1 = 6
D -> J.7 = 6

Fig. 9 Text indicating categories that satisfied umbrae with category H and other categories that satisfied the umbrae

In Fig. 10 are showed in other format the categories D and H that satisfied the umbrae of greater or equal of five mentions, format that permit to appreciate the principals as center categories.

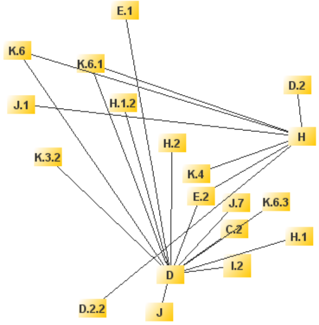


Fig. 10 Map showing categories that satisfied umbrae with category H and other categories that satisfied too

V. Results and future work

The analysis presented at the questions Q1, Q2, Q3 and Q4 have been resolved with the database organized, the tool architecture and the query models here described.

The work developed permit to conclude that in order to make analysis in other information resources is necessary that they are classified in some classification and it is important to model with data cubes when is necessary to work with aggregates and hierarchies.

The contribution of this work is to permit to people make the initial questions Q1, Q2, Q3 and Q4 to facilitate revise the source or articles (see Figure 18) that generate a tendency in order to act in search of his interests, although the tendencies showed can be object of study or interpretation.

This work has not have some functionalities taken as future works and that are the next:

a) Obtain a grade of automatization in the load component of a database for resource information. For example, if we wish to analyze the master degree thesis of CIC and Ph., first it is necessary to classify Degree under ACM classification (question to theists or thesis directors or classify using a classificatory system or a human classifier) and then make the manual load to the system.

b) The database it is growing of more article data in order to make analysis of the models specified in the introduction of this document, permitting have the database design showed in the Fig. 11.

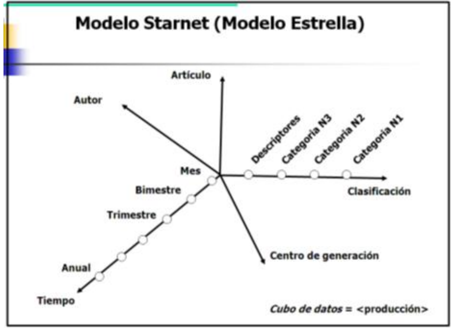


Fig. 11 Star model designs to obtain other aggregates of interest as authors and generate source

c) Verification of correlation between themes with similar comportments; in this part it is necessary fundament if there is a relation between two themes using correlation coefficients or similarity coefficients or equivalence as

$$eij = wij2 / wi*wj \qquad (7)$$

Where wi is the number of documents where the word wi is showed, wj is the number of documents where the wj word is showed and wij is the number of documents where both words are showed [5].

References

[1] E. Acevedo-Pineda "Lo que la Cienciometría no Alcanza a Medir", Organización de Estados Iberoamericanos, http://www.oei.es/salactsi/elsa6.htm. 05 de Febrero de 2009.

[2] H. Canales-Becerra and M. Mesa-Fleitas, "Bibliometría, Informetría, Cienciometría: Su Etimología y Alcance Conceptual", Biblioteca Virtual de las Ciencias en Cuba, http://www.bibliociencias.cu/gsdl/collect/eventos/index/assoc/HASH0160.dir/doc.pdf. 05 de Febrero de 2009.

[3] "ACM Portal", http://portal.acm.org/ 09 de Febrero de 2009.

[4] "The 1998 ACM Computing Classification System", http://www.acm.org/about/class/1998. 05 de Febrero de 2009.

[5] S. Sung and M. Jung, (2008), "Knowledge sources of innovation studies in Korea: A citation analysis", Scientometrics, 75 (1): 01–18

[6] E. Spinak, "Indicadores Cienciométricos", 1998, Ci. Inf., Brasília, v. 27, n. 2, p. 141-148.

[7] E. Garfield, "Mapping the world of biomedical engineering", Institute for Scientific Information. 1986.

[8] A. Ruiz Figueroa and P. Makagonov, "Modelos de desarrollo del hardware y software basados en el estudio de computación paralela", Interciencia Vol. 32 No. 3. 2007.

[9] I. Samoylenko, T.-C. Chao, W.-C. Liu and C.-M. Chen, "Visualizing the scientific world and its evolution", Wiley Interscience. 2006.

[10] J. Guan and N. Ma, "A bibliometric study of China's semiconductor literature compared with other major Asian countries", Scientometrics Vol. 70 No. 1. 2007.

[11] R. Cañedo-Andalia, "Los análisis de citas en la evaluación de los trabajos científicos y las publicaciones seriadas", ACIMED Vol. 7, No. 1. 1999.

[12] A F J. Van Raan, "Scientometrics: State of the Art", Scientometrics Vol. 38 No. 1 pp. 205-218. 1997.

[13] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence. 1996.

[14] "JGraph", http://www.jgraph.com/jgraph.html. 20 de Febrero de 2009.

[15] V. Herrero-Solana and J. Morales-del-Castillo, "Análisis geopolítico de los mapas de conocimiento", E-Prints in Library and Information Science. Vol. 179 No. 705 pp. 159-171. 2004.

[16] R. Kimball, "The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing and Deploying Data Warehouses", John Wiley & Sons, Inc. USA, 1998

[17] R. Agrawal, A. Gupta, and S. Sarawagi, "Modeling Multidimensional Databases", IBM Almaden Research Center, 650 Harry Road, San José, CA 95120; 1995.