

SciMAT

Version 1.0

User guide

M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma, F. Herrera

University of Granada. Spain

1) Introduction

SciMAT (Science Mapping Analysis software Tool) is a new open source science mapping software tool developed at University of Granada. It integrates the advantages of the science mapping software tools available while reducing dependence on third party software. It can be freely downloaded, modified and redistributed according to the terms of GPLv3 license.

It is based on the science mapping analysis approach presented in Cobo et al. (2011) which allows us to carry out science mapping studies under a longitudinal framework (Garfield, 1994; Price & Gürsey, 1975).

The main characteristics of SciMAT are:

- It incorporates all modules necessary to carry out all the steps of the science mapping workflow, which can be configured ad-hoc. It helps the analyst to carry out the different steps of the science mapping workflow, from data acquisition and preprocessing to the visualization and interpretation of the results.
- It incorporates methods to build the majority of the bibliometric networks, different similarity measures to normalize them and build the maps using clustering algorithms, and different visualization techniques useful for interpreting the output.
- It implements a wide range of preprocessing tools such as detecting duplicate and misspelled items, time slicing, data reduction and network preprocessing.
- According to Cobo et al. (2011), SciMAT allows the analyst to perform a science mapping analysis in a longitudinal framework in order to analyze and track the conceptual, intellectual or social evolution of a research field through the course of consecutive time periods.
- Similarly, according to Cobo et al. (2011), SciMAT builds science maps enriched with bibliometric measures based on citations such as: h-index (Alonso et al., 2009; Hirsch, 2005), g-index (Egghe, 2006), hg-index (Alonso et al., 2010), q^2 -index (Cabrerizo et al., 2010), etc.

SciMAT is divided into three different modules: i) a module dedicated to the management of the knowledge base and its entities, ii) a module responsible for carrying out the science mapping analysis, and iii) a module to visualize the generated results and maps. These modules allow the analyst to carry out the different steps of the science mapping workflow.

In the following sections the structure of the knowledge base used by SciMAT is described and each of its modules is shown.

2) Knowledge base

SciMAT generates a knowledge base from a set of scientific documents, where the relations of the different entities related with each document (authors, keywords, journal, references, etc.) are stored. This structure helps the analyst to edit and preprocess the knowledge base in order to improve the quality of the data and consequently, obtain better results in the science mapping analysis.

The knowledge base is composed of sixteen entities. The principal one is the *Document* which represents a scientific document (usually, articles, letters, reviews or proceedings papers). It contains information such as, the title, abstract, doi, citations, etc. The Document has a variety of information associated with it, such as the authors, affiliations, keywords, cited references, the journal (or conference), and the publication year. Each one is considered an entity in the knowledge base.

The *Author* is the entity that represents the person who has been involved in the development of a Document. An Author can be associated with a set of Documents, and in a similar way, a Document has a set of Authors. Furthermore, an Author has an associated position in his/her Documents.

The *Affiliation* represents the author's affiliations. Due to the fact that the authors may work in different places (universities, institutes, etc.) during their research, an Author has a set of associated Affiliations.

Usually, the scientific documents have a set of keywords associated with them. Furthermore, depending on the bibliometric database used to retrieve the data, the documents can contain descriptive words provided by the database. For example, the ISIWoS adds a set of keywords called ISI Keywords PLUS to each document. In this sense, the entity *Word* represents a descriptive term of a document. A set of Words can appear in different Documents and each Document can have a set of Words. Each Word can have different roles in the Documents in which it appears. In this way, a document can have words provided by the authors (author's words), provided by the database (source's words), or added in the preprocessing step (extracted words).

The entity *Reference* represents the intellectual base of a scientific document. Similarly to the Word, a Document has a set of References associated with it, and each Reference can be present in different Documents. The references can often be divided into small pieces of information. Depending on the database used to retrieve the data, these pieces may be different, but some information appears more often, such as the author, journal and the year. For this reason, there are two entities related to the Reference: the *Author-Reference* and the *Source-Reference*.

Other entities associated with a Document are the *Journal* and the *Publish Date*. Logically, a Document can only have one Journal (or conference) and one Publish Date associated with it, whereas both entities can have one set of Documents associated. Moreover, these entities have an associated *Subject Category* which represents a global category, often given by the bibliometric database, which classifies the journal in main knowledge categories. The Journal can be associated with many Subject Categories, and this relation can change throughout the years.

The entity *Period* represents a set of (not necessarily disjointed) years. Usually, a set of Periods are defined to perform a longitudinal science mapping analysis.

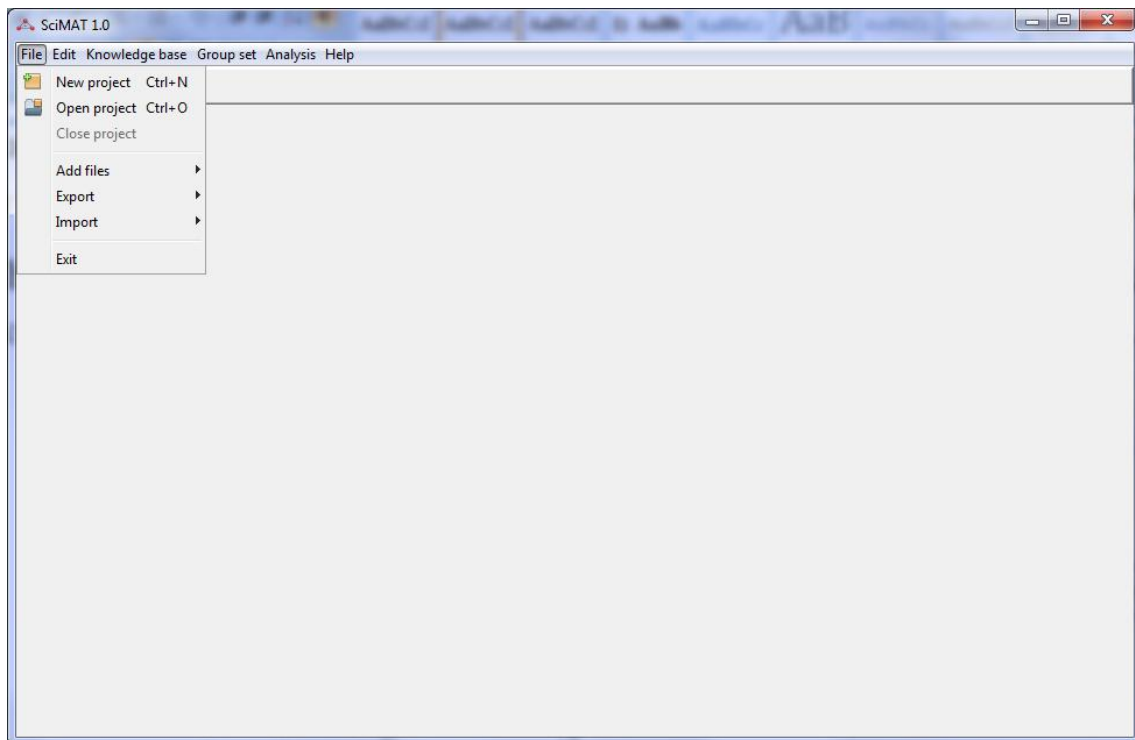
We should point out that five of the above described entities can be used as a unit of analysis in the science mapping analysis carried out by SciMAT: Author, Word, Reference, Author of Reference and Source of Reference. These entities should be

carefully preprocessed, paying special attention to the misspelling and de-duplicating process. Usually, the de-duplicating process joins the similar items in only one way. For example, two items stored in the knowledge base: Garfield, E. and Eugene Garfield. Both items represent the same author, and therefore, they should be joined (joining its association with the other entities). But, when two items are joined, only one of them is kept in the knowledge base (obviously, this item contains the association of the second item), and it is impossible to know the initial items joined. For this reason, our knowledge base provides the concept of *group* for each unit of analysis. A group is a set of items that represents the same entity. In this way, the knowledge base contains five kinds of groups: *Author Group*, *Author-Reference Group*, *Source-Reference Group*, *Reference Group* and finally *Word Group*. A group can be marked as *stop group*, and it will not be taken into account in the science mapping analysis.

3) Knowledge base manager

The module to manage the knowledge base is responsible for building it, importing the data from different bibliographical sources, and cleaning and fixing the possible errors in the entities. It can be considered as a first stage in the preprocessing step.

The first step in this module is to build a new project or load an existing one. It can be done through the menu *File* or using the buttons of the toolbar.

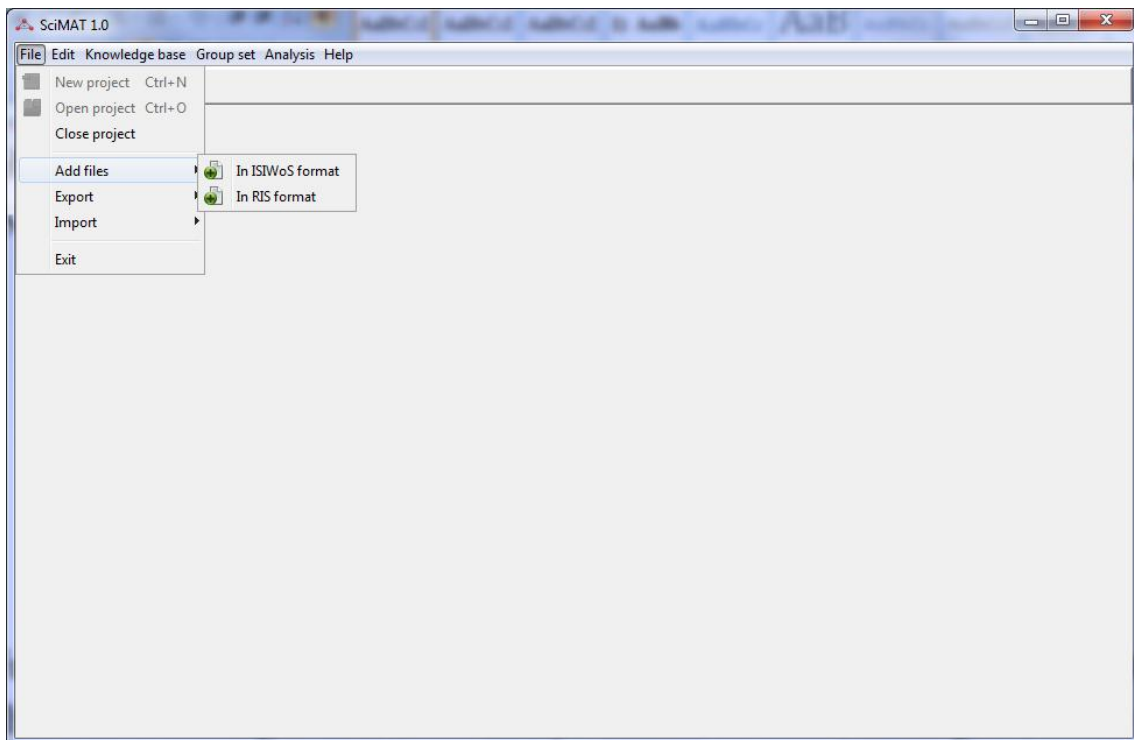


If a new project is selected, a new window will appear asking for the path where the knowledge base file will be stored and the name of the file. We can give any extension for the file.

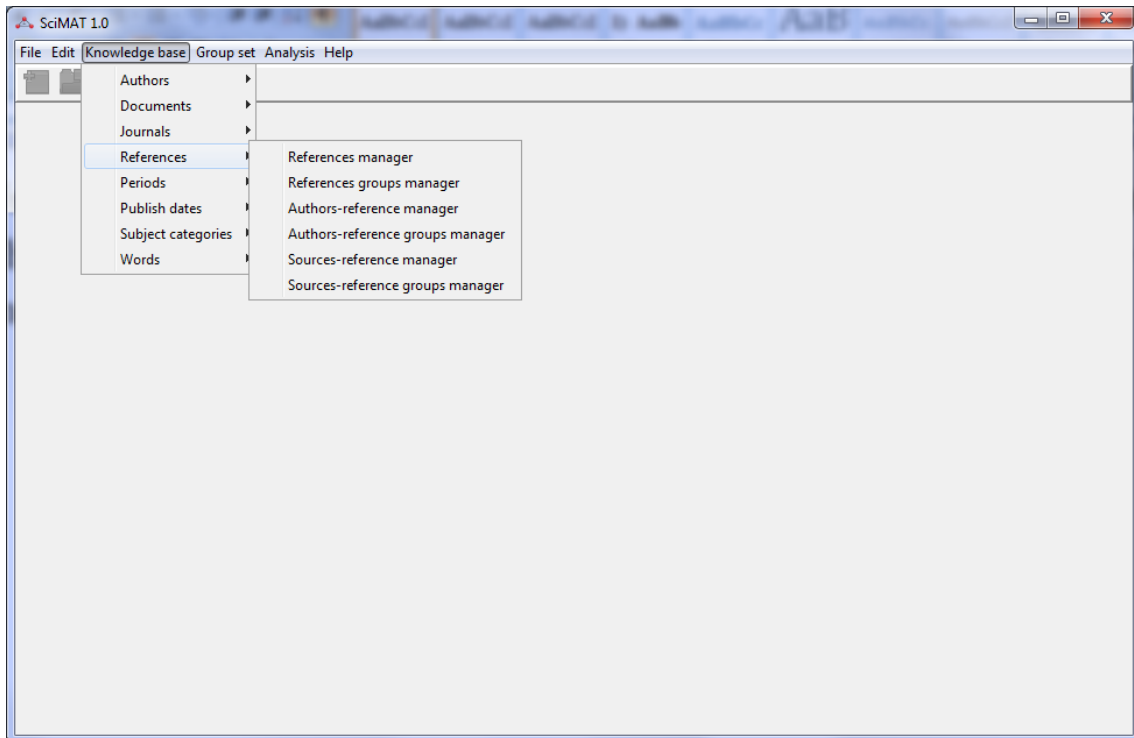
SciMAT uses the SQLite database engine in order to store the knowledge base built in the previous step. Thanks to these capabilities, the knowledge base can be opened with any database browser that reads SQLite files.

Once we have a project loaded (new or existing), the options under the menus *Knowledge base* and *Group set* will be enabled. Furthermore, the import, export and add options under the menu *File* will be enabled too.

The add files option allows the user to add bibliographical information, exported from bibliographical databases, to the knowledge base. Particularly, SciMAT is able to read bibliographical information exported in ISI Web of Knowledge format (ISI-CE) or RIS (Scopus) format.

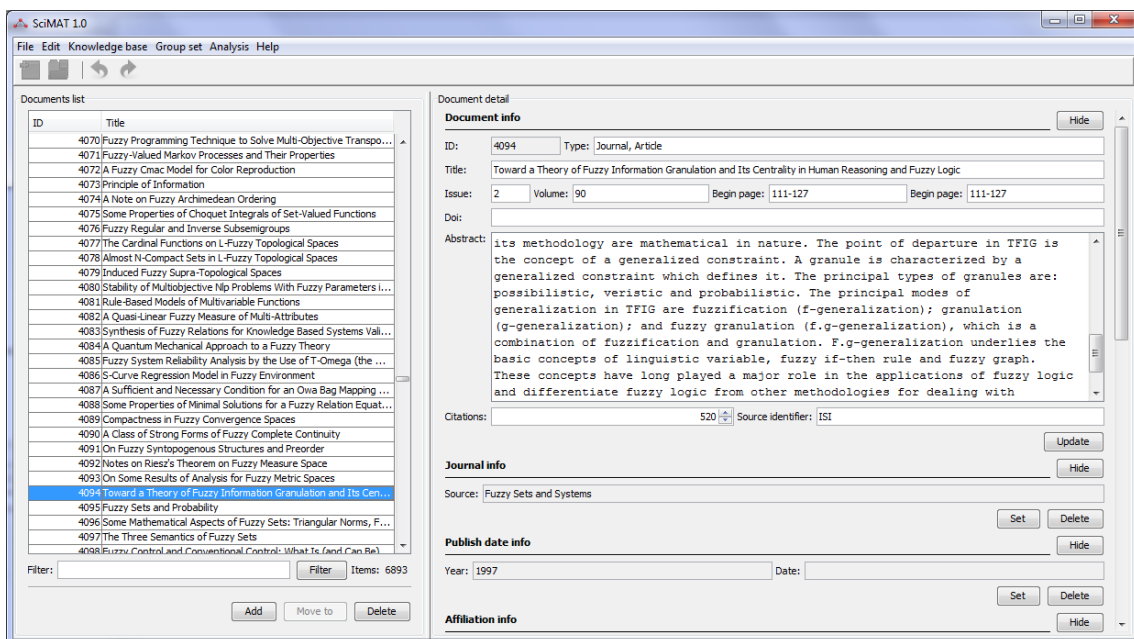


Under the menu *Knowledge base* the manager for the sixteen entities can be found. There is a manager for each entity. Thank to these managers, each entity can be edited and its attributes and associations can be modified.

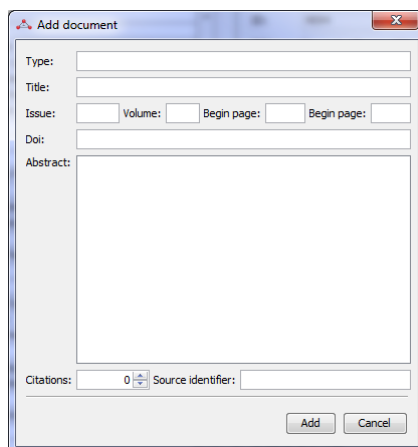


We should point out that all the managers have the same structure, on the left-side a list of entities is shown, and on the right-side the fields of the selected entity and its relations with other entities are shown.

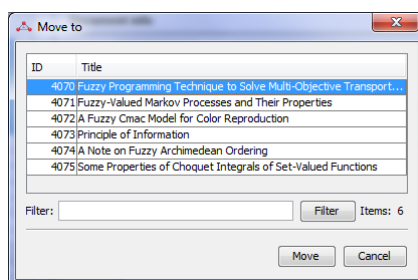
As an example the Document's manager is shown below. In the list of documents (left-side), one of the most cited articles in the knowledge base is selected, and on the right-side its associated information (title, abstract, publication data, citations, etc.) and associations are shown.



The manager allows us to add a new Document (filling manually each attribute), delete a set of Documents and join (*move to* button) a set of Document. Furthermore, the user can use the filter box to introduce a regular expression and find the wanted entities.



The *move to* or *join* capability allows us to join a set of entities under other. It is especially useful when we are working with groups. Once we have selected the set of entities that we want to join, a new dialog will appear. In this dialog, the user has to select the entity under the remaining entities will be joined. The main or target entity will maintain its associations with others entities and the associations of the joined entities. As an example, the six selected documents below will be joined under the document with ID 4070. So, the target document will be associated with the words, references, affiliations, etc of the remaining five documents.



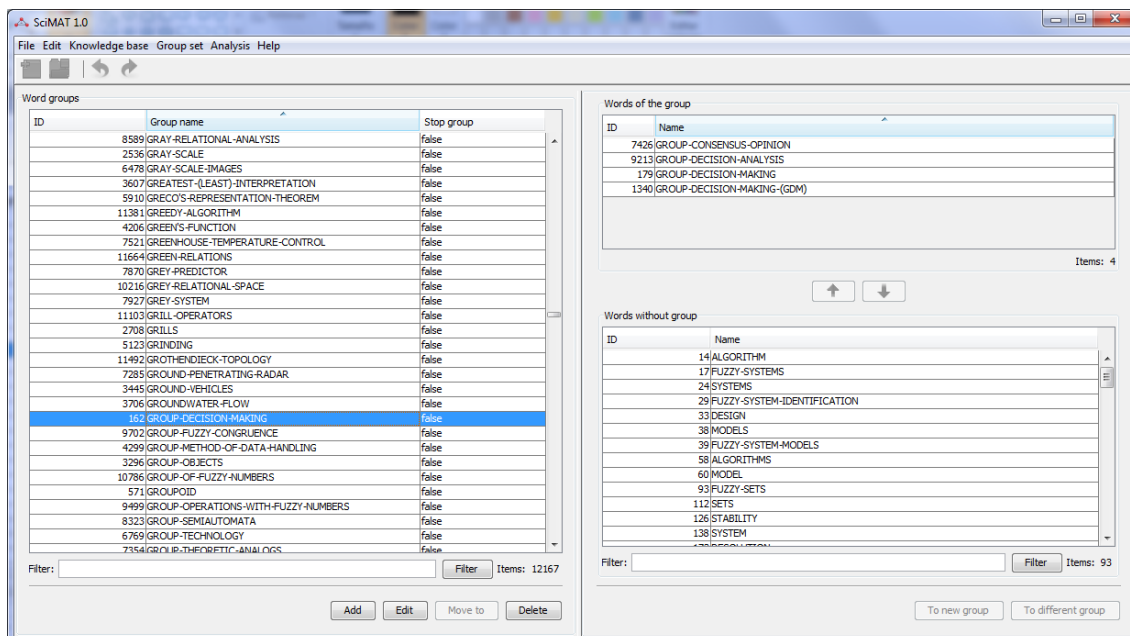
The right-side of the manager allows us to edit the field of the selected entity or its associations with other entities.

SciMAT incorporate powerful capabilities to perform a de-duplicating step over the unit of analysis by means of groups. To do that, there are five especial managers to perform this task (they can be found under the menu *Group set*). Similarly to the entity manager, the manual groups set manager have a common structure: the left-side shows a list of defined groups, and the right-side shows the entities associated with the selected entity (header-table) and the entities without groups (foot-table).

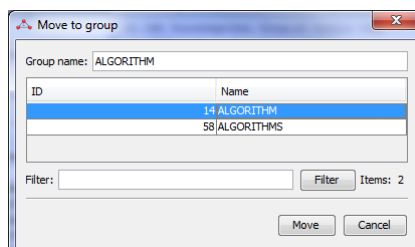
As an example, the manager to perform the manual set of the *Words Groups* is shown below. It can be seen that a particular word group with the name *GROUP-DECISION-MAKING* has been defined (left-side). It can also be observed that this word group collects four different word names or variants (top right-side) for the concept. The lower right-side allows the user to add more variants of the concept *GROUP-DECISION-MAKING*.

The manual set group manager allows us to add a new group, delete a set of groups, join a set of groups under other, and finally edit them. Furthermore, this manager allows us

to add a set of entities to a selected group, or delete a set of entities from a group. This can be done using the up-row and down-row from the middle of the right-panel.



The groups can also be added from a set of entities without group through the buttons “to new group” and “to different group”. The former build a new group adding the selected entities to it. The name of the group can be chosen from the entities or can be given by the user. In the latter, each entity will be associated with a group and the group’s name will be the main field of the entity.



Additionally, this module incorporates methods to help the analyst in the de-duplicating process, such as, finding similar items by plural or by Levenshtein distance, or importing the groups and theirs associated items from a file (in XML format).

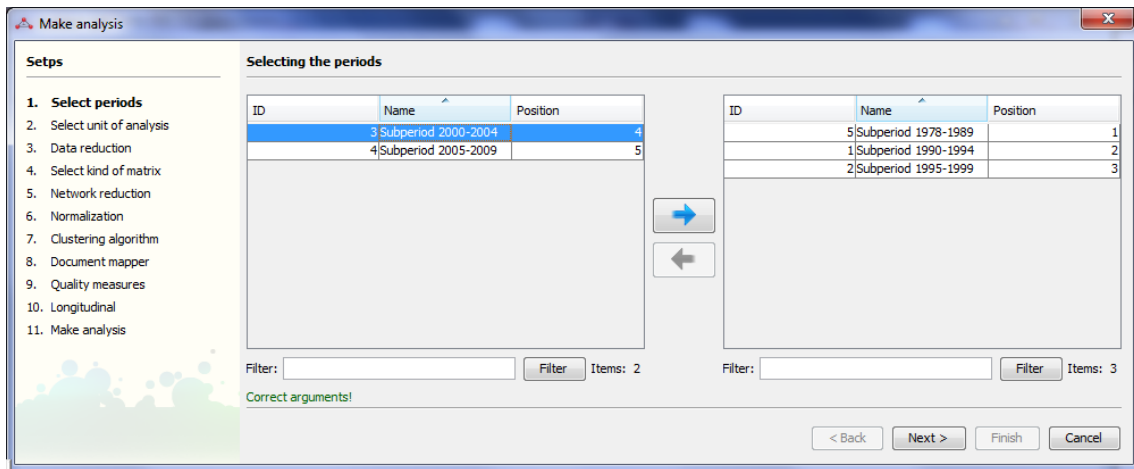
4) Science mapping analysis wizard

Once the knowledge base is ready for the science mapping analysis, this module helps the user to configure the process, choosing the methods and algorithms that have to be used by SciMAT to build the maps.

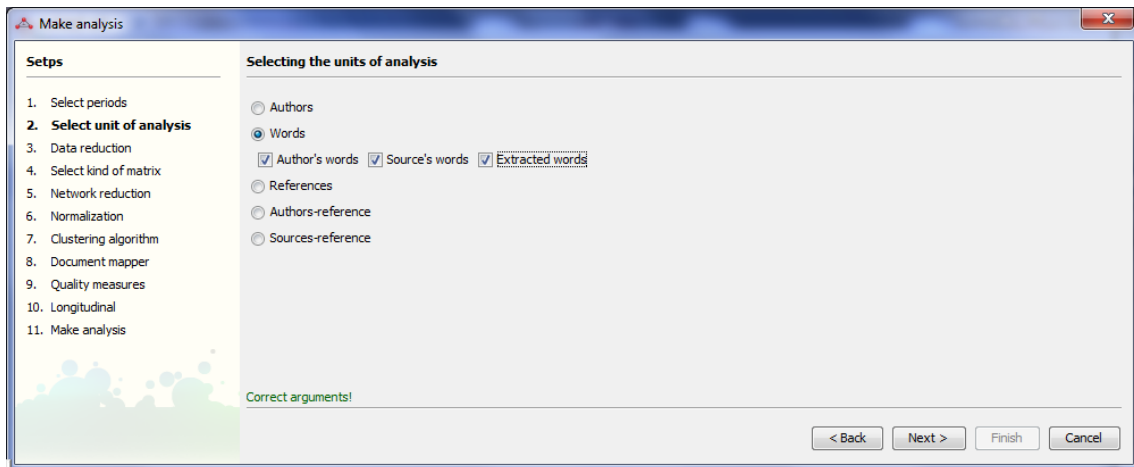
This module is implemented through a wizard and it is composed of eleven consecutive steps:

In the **first step** the user has to select the periods that he/she wants to analyze. Each period will produce a map. These periods will be used in the longitudinal or temporal analysis in order to study the structural evolution of the field. The position of the period

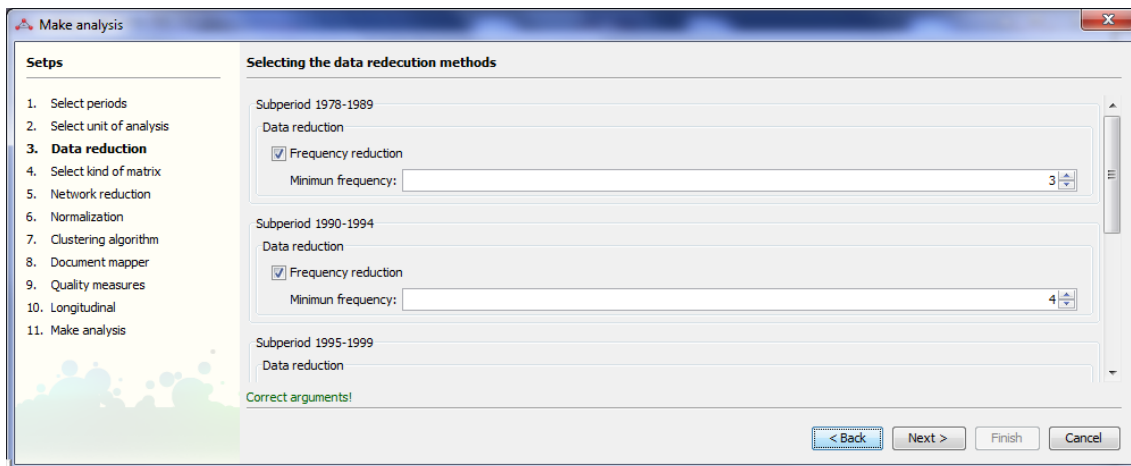
will indicate the order in which the period will be used in the process. So, the period with lowest position will be processed first and will be the first in the longitudinal results.



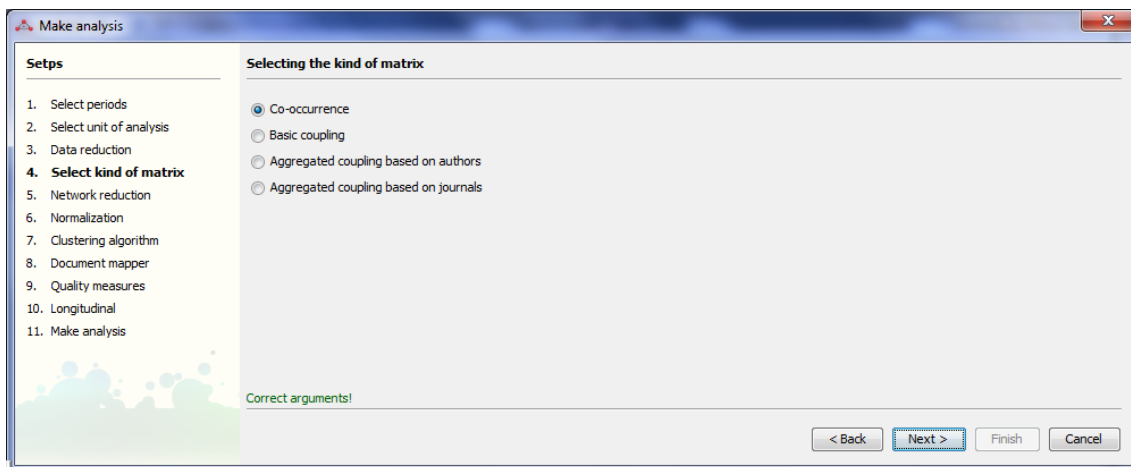
The **second step** is the selection of the unit of analysis. As the unit of analysis the user can select any of the five groups existing in the knowledge base: *Author Group*, *Author-Reference Group*, *Source-Reference Group*, *Reference Group*, or *Word Group*. Only one of them can be selected. If the *Word Group* has been selected, the role of the word with which the user wants to perform the analysis has to be chosen. In this case, the user has to select the author's word, source's word or extracted word, or indeed, any combination of them.



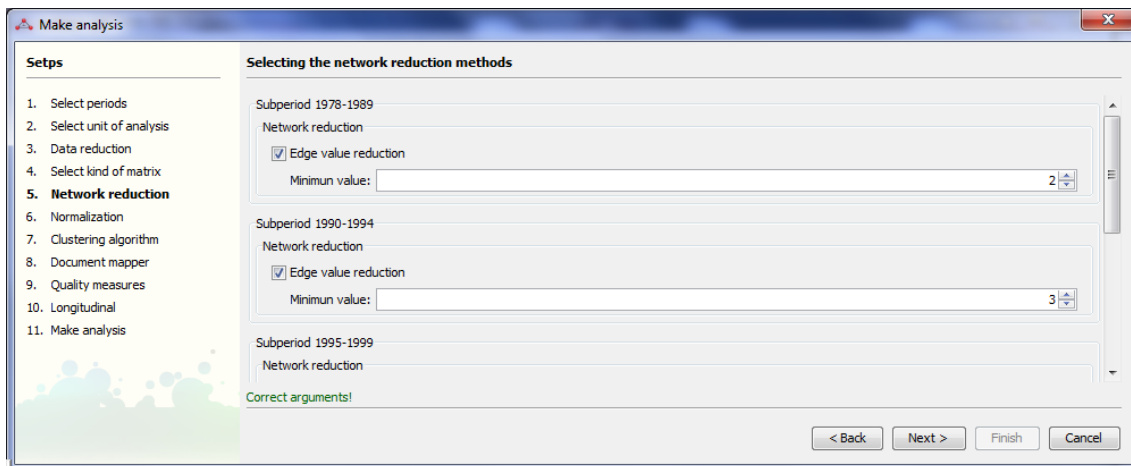
The **third step** is the data reduction. SciMAT allows the data to be filtered using a minimum frequency threshold. For each selected period, a threshold must also be selected. That is, only the item that appears in almost n documents in a given period will be taken into account.



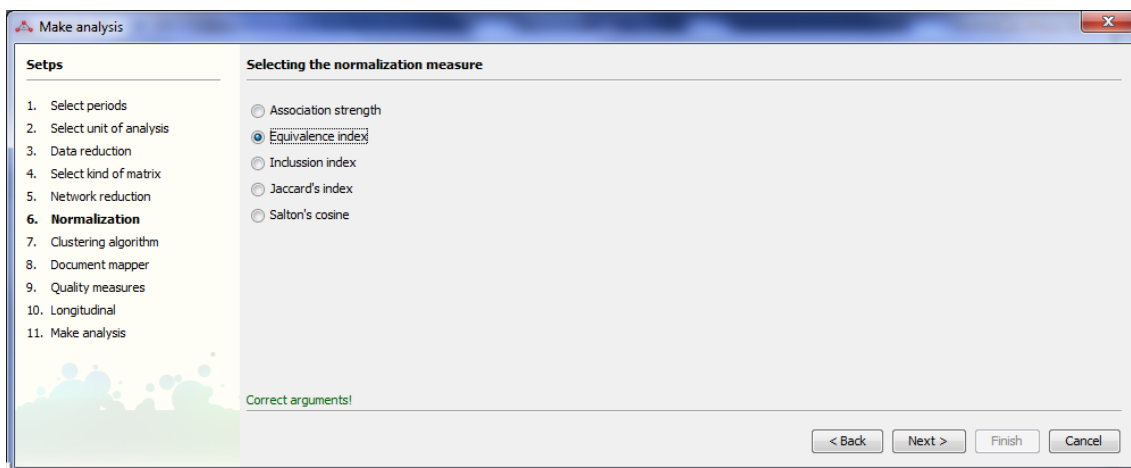
The **fourth step** is the selection of the way in which the network will be built: co-occurrence or coupling. Using co-occurrence, co-author, co-word, co-citation (using the references), author co-citation (using the authors-reference), and journal co-citation (using the sources-reference) network can be built. Otherwise, the coupling can be used in a basic or aggregated way. In the former, a document coupling network can be built using the selected unit of analysis as coupled items. That is, if the Reference has been chosen, a document bibliographic coupling network will be built. In the latter, an author or journal coupling network can be made, and again, the coupled item will be the selected unit of analysis.



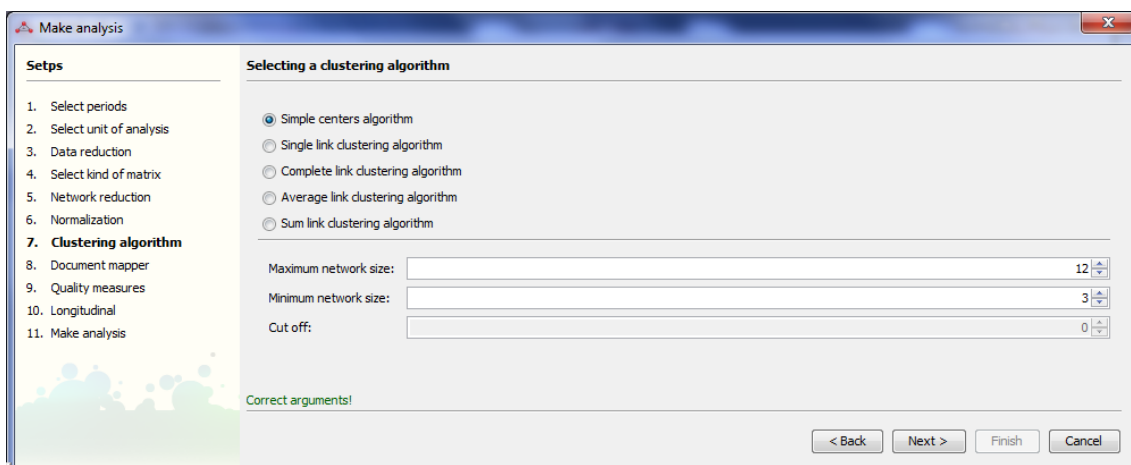
The **fifth step** is the network reduction. SciMAT allows the network to be filtered using a minimum edge value threshold. For each selected period, a threshold value must be set. That is, only the edges with a value greater or equal to n in a given period will be taken into account.



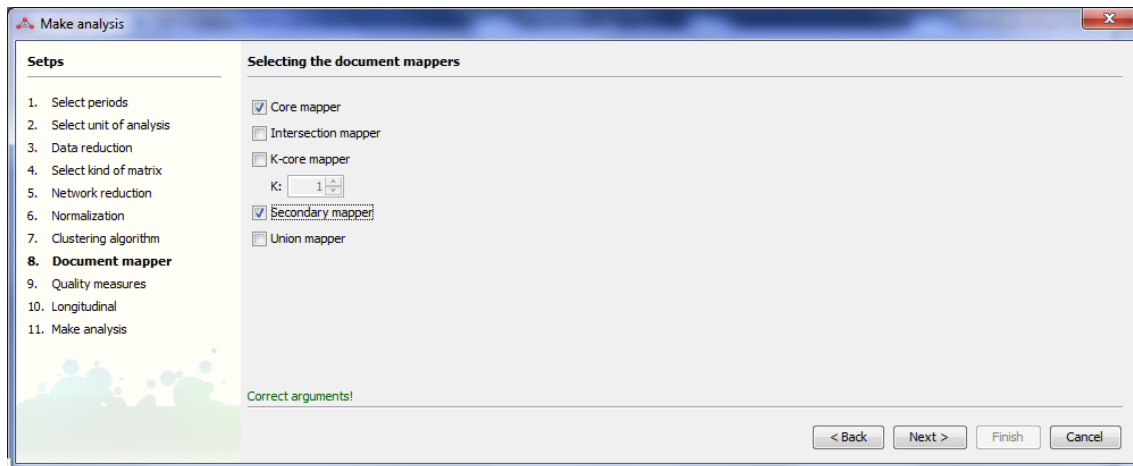
The **sixth step** is the selection of the similarity measure used to normalize the network. SciMAT allows the user to choose the similarity measures commonly used in the literature to normalize networks: Association Strength, Equivalence Index, Inclusion Index, Jaccard's Index and Salton's Cosine.



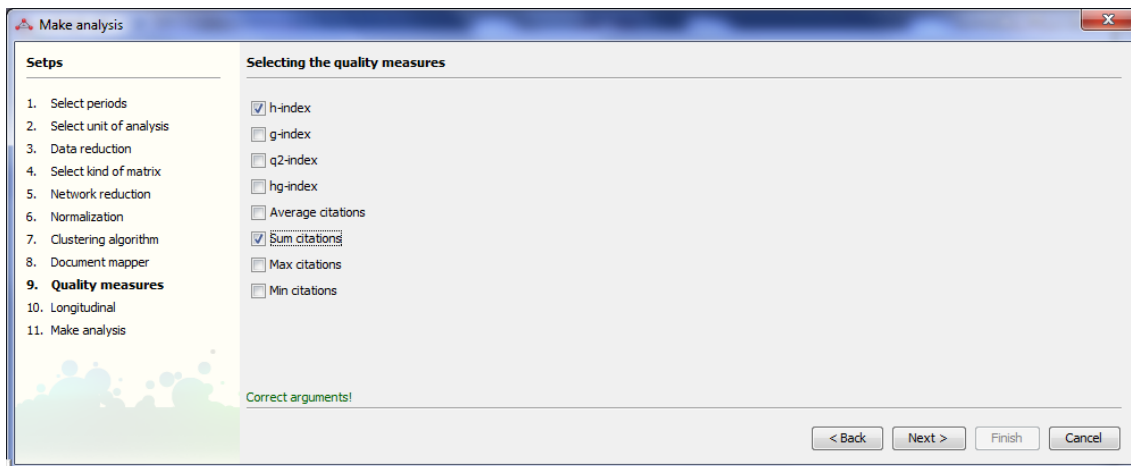
The seventh step is the selection of the clustering algorithm used to get the map and its associated clusters or subnetworks. Different clustering methods are available in SciMAT, such as, the Simple Centers Algorithm (Coulter et al., 1998; Cobo et al., 2011), Single-linkage (Small & Sweeney, 1985) and variants such as Complete-linkage, Average-linkage and Sum-linkage.



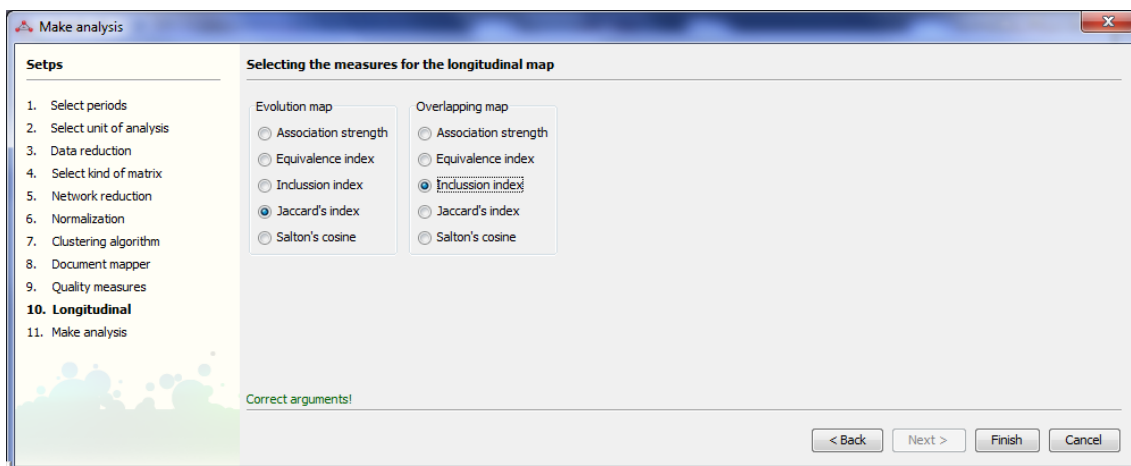
The **eighth step** is the selection of the documents mapper used in the performance analysis. SciMAT incorporates five different document mappers for co-occurrence networks: i) core mapper (Cobo et al., 2011), ii) intersection mapper which adds the documents that have all the items of the cluster, iii) k-core mapper which adds the documents that have at least k items in common with the cluster, iv) secondary mapper, and v) union mapper which adds documents that have at least one item in common with the cluster (this is the union of the documents associated with the core and secondary mappers). For coupling networks, SciMAT has two kinds of document mappers depending on the kind of coupling used. That is, if a basic coupling has been selected (each item of the cluster will be a document), the basic coupling document mapper is the only one available, which adds the items of the cluster as documents. If an aggregated coupling is selected, the aggregated coupling document mapper can be selected, which adds the documents associated with its items to each cluster (author's or journal's oeuvres). We should point out that each item or node of the cluster also has a set of associated documents. These documents correspond to the set of items associated with the item in the corresponding dataset.



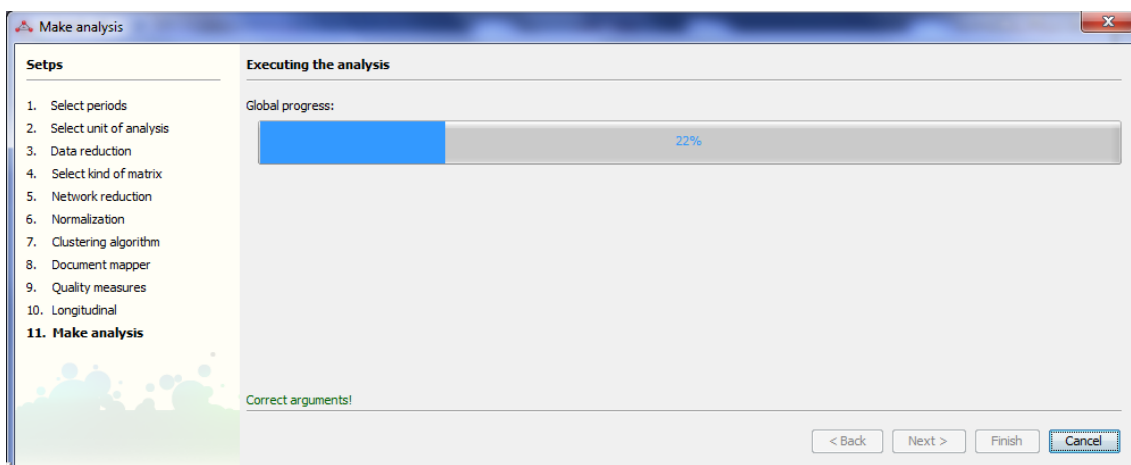
The ninth step is the selection of the performance and quality bibliometric measures. SciMAT adds by default the number of documents as performance measure. Moreover, the citations of a set of documents are used in order to assess the quality and impact of the clusters. In this sense, basic measures such as the sum, minimum, maximum and average citations, or complex measures such as the h-index (Alonso et al., 2009; Hirsch, 2005), g-index (Egghe, 2006), hg-index (Alonso et al., 2010) or q^2 -index (Cabrerizo et al., 2010) can be selected.



The tenth step is the selection of the similarity measure used to build the evolution map and the overlapping map. SciMAT allows us to choose between: Association Strength, Equivalence Index, Inclusion Index, Jaccard's Index and Salton's Cosine.



Finally, the eleventh step is responsible to perform the science mapping analysis. This process can be cancelled at any time.



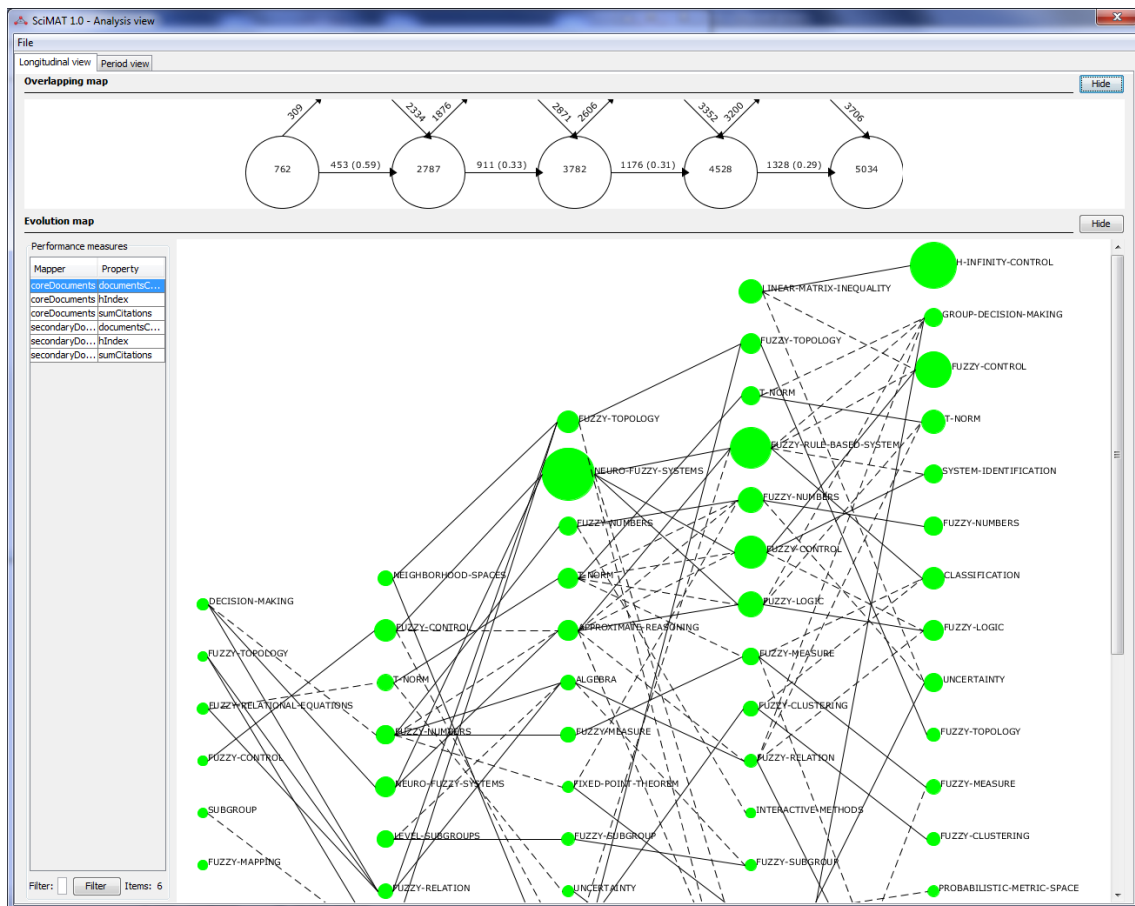
At the end, the analysis has to be saved (a new save window will be open when the process end), and then the results are visualized in the visualization module.

5) Visualization module

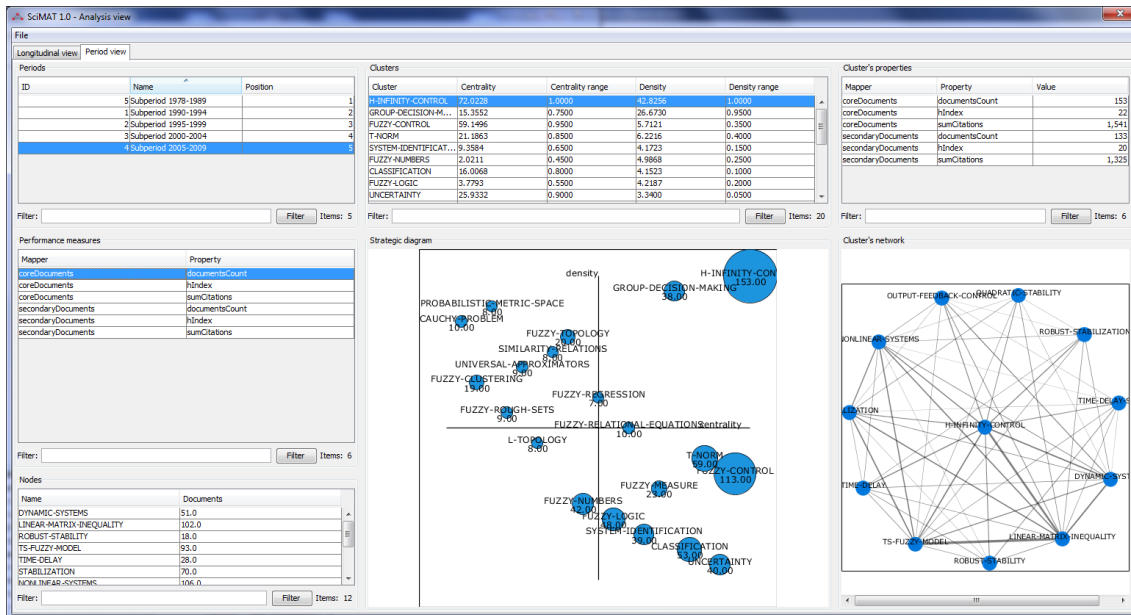
Once the science mapping analysis has been done, the results are visualized through this module. Furthermore, we can load an analysis previously done (to do that it is not necessary that a knowledge base is loaded).

The visualization module has two views: longitudinal view and period view.

In the Longitudinal view the overlapping map (top) and evolution map (down) are shown. This view helps us to detect the evolution of the clusters throughout the different periods, and to study the transient and new items of each period and the items shared by two consecutive periods. The right-table allows us to choose the measure used to draw the nodes in the evolution map.



Finally, the Period view shows detailed information for each period, its strategic diagram, and for each cluster, the bibliometric measures, the network and their nodes.



The visualization module is able to build a report in HTML or LaTeX format. The images (strategic diagrams, overlapping items map, etc.) are exported in PNG and SVG format, so the user can edit them. Furthermore, the cluster networks and evolution maps are exported in Pajek format.

References

- Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2009). h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3 , 273–289.
- Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2010). hg-index: A new index to characterize the scientific output of researchers based on the h- and g- indices. *Scientometrics*, 82 , 391–400.
- Cabrerizo, F. J., Alonso, S., Herrera-Viedma, E., & Herrera, F. (2010). q²-index: Quantitative and qualitative evaluation based on the number and impact of papers in the hirsch core. *Journal of Informetrics*, 4 , 23–28.
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, 5 , 146–166.
- Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49 , 1206–1223.
- Garfield, E. (1994). Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioural Sciences*, 7 , 5–10.
- Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69 ,131–152.
- Hirsch, J. (2005). An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences*, 102 , 16569–16572.
- Price, D., & Gürsey, S. (1975). Studies in scientometrics I: Transience and continuance in scientific authorship. *Ci. Informatics Rio de Janeiro*, 4 , 27–40.
- Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations. *Scientometrics*, 7 , 391–409.