PERGAMON

Computers & Industrial Engineering1 43 (2002) 841–862

computers & industrial engineering

www.elsevier.com/locate/dsw

DIVA: a visualization system for exploring document databases for technology forecasting

Steven Morris[a,*], Camille DeYong[b], Zheng Wu[a], Sinan Salman[b], Dagmawi Yemenu[b]

[a]*School of Electrical and Computer Engineering, Oklahoma State University, 202 Engineering South, Stillwater, OK 74078, USA*

[b]*School of Industrial Engineering and Management, Oklahoma State University, 322F Engineering North, Stillwater, OK 74078, USA*

Abstract

Database Information Visualization and Analysis system (DIVA) is a computer program that helps perform bibliometric analysis of collections of scientific literature and patents for technology forecasting. Documents, drawn from the technological field of interest, are visualized as clusters on a two dimensional map, permitting exploration of the relationships among the documents and document clusters and also permitting derivation of summary data about each document cluster. Such information, when provided to subject matter experts performing a technology forecast, can yield insight into trends in the technological field of interest. This paper discusses the document visualization and analysis process: acquisition of documents, mapping documents, clustering, exploration of relationships, and generation of summary and trend information. Detailed discussion of DIVA exploration functions is presented and followed by an example of visualization and analysis of a set of documents about chemical sensors. © 2002 Published by Elsevier Science Ltd.

Keywords: Technology forecasting; Information visualization; Knowledge discovery in databases; KDD; Data mining; Citation analysis; Document mapping; Bibliometrics; Scientometrics

1. Introduction

Bibliometric analysis, the analysis of large numbers of scientific documents for the purpose of technology forecasting (Kostoff, 1997), has been limited by the need to manually extract data from documents and further limited by the paucity of documents available in electronic form. The recent availability of Internet-based abstract services and patent databases, allowing easy access to documents in electronic form, has made the application of bibliometric techniques for technology forecasting quite practical.

* Corresponding author. Tel.: +1-405-744-1662; fax: +1-405-744-9198.
E-mail address: samorri@okstate.edu (S. Morris).

0360-8352/02/$ - see front matter © 2002 Published by Elsevier Science Ltd.
PII: S0360-8352(02)00143-2

Bibliometric analysis has usually been a process of summarizing document characteristics into tables for statistical analysis (Kostoff, 1997). Swanson and Smalheiser (1997) use an extension of such bibliometric techniques to find indirect links among concepts in documents from the medical literature. A useful addition to bibliometric analysis is to map documents onto a two dimensional space for visualization of relations among the documents (Small, 1997). Spasser (1997) and McCain (1998) are good examples of mapping and bibliometric analysis applied to the pharmaceutical literature and neural network literature, respectively. Bibliometric techniques, supplemented by visual mapping, have also been applied as an assessment and planning tool to collections of patents (Mogee, 1991). Such techniques, applied to patents, yield good results for competitive intelligence applications as well (Mogee, 1997).

Bibliometric analysis, when considered as a data mining problem, presents two challenges. The first challenge is to establish the magnitudes of relationships among documents based on their technological content. These inter-document relationships are expressed as 'similarity' values whose magnitudes are proportional to the strength of the relationships. Given a mathematical representation of those similarities, the second challenge is to classify (cluster) the documents by technological fields, such that measurable properties of the clusters and the links between clusters can be extracted to provide meaningful indicators of trends in technological progress.

For the first challenge above, inter-document similarity, we can apply *similarity functions*, methods that are used to mathematically express the strength of similarities among documents. Term co-occurrence similarity functions are based on finding commonalities in the text content among documents and are based on the work of Salton, whose book, *Automatic Text Processing* (Salton, 1989), contains a great number of valuable techniques for this purpose. Citation based similarity functions, based on analysis of inter-document citations, have been proposed and applied by Small (1997). For this work we apply both term co-occurrence similarity functions and citation based similarity functions.

For the second challenge, classification, we have chosen to use a technique that relies on document mapping, visualization, and interactive exploration. Documents are mapped on a rectangular surface such that closely related documents (as defined by the similarity function) are located close together on the map, while unrelated documents are far apart on the map. Such mapping tends to place the documents into clusters that are easily identified by a human using interactive exploration tools. The user identifies clusters and after exploring their properties, labels the clusters by technological field. The technical challenge of this technique is the formation of meaningful document maps from the inter-document similarities. We have successfully applied two mapping techniques to this application: force directed placement, and a self-organizing map (SOM) neural network technique.

This paper introduces a visualization tool, the 'Database Information Visualization and Analysis' system, DIVA, which is used for visualizing trends in the relationships among documents and clusters of documents in a database. Such clusters within collections of technical articles correspond to the different technologies covered by the documents. DIVA helps to identify emerging and declining technologies and how those technologies tend to sustain and borrow from each other. Additionally, DIVA's exploration tools allow identification of experts, centers of excellence, seminal articles and patents, and can also be used to show trends of technological change.

Section 2 of this paper describes the functioning of DIVA as a system, showing the process of building a database of documents, finding links among documents, mapping documents, exploring relations among documents and document clusters, generating summary information, and plotting trends. Section 3 discusses the exploration and report generation functions built into DIVA. Finally, Section 4 gives an

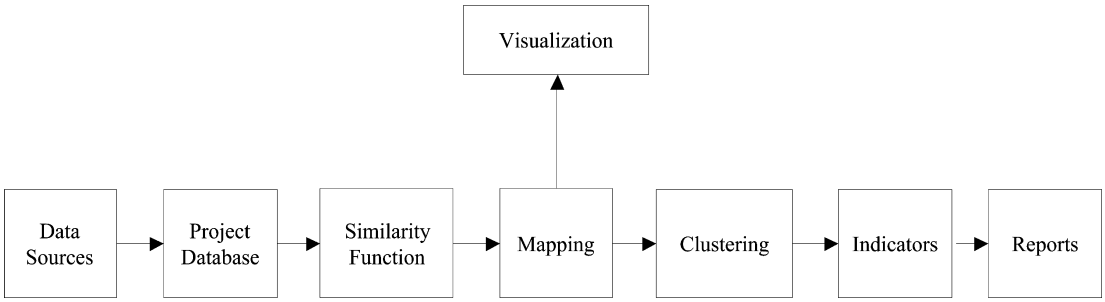

Fig. 1. Block diagram of document visualization process.

example, describing the application of DIVA to the analysis of collections of documents covering the subject of chemical sensors. To reduce the number of figures in this paper, figures from the chemical sensor example of Section 4 will be used when illustrating concepts in Sections 2 and 3, resulting in the figures being referenced out of sequence.

2. System description

Most of the DIVA exploration and analysis functions are written in MATLAB, but document storage is done using MS Access. Fig. 1 shows a block diagram of the flow of work in a DIVA project. Document data is loaded into the project database from a source file. A similarity function, based on citation data or co-occurrence of terms, is used to find the magnitude of the links between all pairs of documents in the database. Based on these links the documents are mapped onto a two dimensional space for visualization. Such mapping usually groups the documents into clusters as shown in Fig. 5. Clusters generally represent technologies and sub-technologies within the dataset. The clusters are manually identified and labeled by the user, who employs additional exploration functions to visualize relations among documents and document clusters. A report generator produces summary information and indicators for each cluster. Indicators are time plots of cluster properties, e.g. the number of documents in a cluster by year or the number of links to other clusters by year. A detailed description of the individual parts of the visualization system follows.

2.1. Document sources

DIVA uses three main sources of documents: patent abstracts, journal abstracts with citations, and journal abstracts with no citations. Patent abstracts, available from the US Patent and Trademark Office and from commercial services, contain patent citations which provide explicit links among the patents, are well refereed, are well classified, and many times contain information on technology not otherwise available in the scientific literature. Problematically, US patents are somewhat dated, because patent records in the US are publicly available only after a patent issues, which will be two years or more after the date of the patent application. Journal abstracts with citations are available from a citation service such as the Science Citation Index (Garfield, 1994), or from web based services such as Citeseer (Lawrence, Giles, & Bollacker, 1999). These services contain useful explicit links in the citation data. Author, institution, keywords, abstract text, and other data is available as well. Journal abstracts without

citations, available from abstract services such as Chemical Abstracts, are a third important source of documents. Since these services do not contain citation information it is necessary to find document links using similarity functions based on co-occurrence of keywords or terms.

2.2. Project database

The project database built using DIVA is formed from raw source files of documents. Source files can be generated from queries to the abstract service or patent database. It is often difficult to construct queries that return only documents of immediate relevance. Close interaction with subject matter experts is necessary to focus and evaluate queries for effectiveness. Datasets from sources that contain citation information, such as patent abstracts, can be built by finding all the documents that cite, or are cited by, a list of seed documents. The list of seed documents is usually constructed by asking subject matter experts for a list of seminal journal articles or patents in the technical field to be studied.

Data is generally downloaded to a file in tagged format, i.e. field labels precede each piece of data. DIVA project databases are usually not large, less than two to three thousand documents, and are stored in MS Access. Data is loaded from source files using procedures written in MS Access Visual Basic for that purpose. DIVA accesses data in the database using SQL queries through an ODBC driver (Mathworks, 1999).

2.3. Similarity functions

Similarity functions establish the magnitude of links between all pairs of documents in the database, storing the results in a similarity matrix. Constructing a similarity function that generates links that yield meaningful clusters after mapping is often difficult, especially for document sets that contain no citation information.

Useful similarity functions based on citations can be generated using the method of Small (1997) that uses the direct citations and three types of indirect relations to derive the inter-document similarity values. Starting with a function, designated *dc*, for direct citations:

$$dc_{ij} = \begin{cases} 1 & \text{if document } i \text{ cites document } j \\ 1 & \text{if document } j \text{ cites document } i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Additionally defining indirect citation functions *lc*, *cc*, and *bc*:

$$lc_{ij}(k) = \begin{cases} 1 & \text{if document } i \text{ cites document } k \text{ AND document } k \text{ cites document } j \\ 1 & \text{if document } j \text{ cites document } k \text{ AND document } k \text{ cites document } i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$cc_{ij}(k) = \begin{cases} 1 & \text{if document } k \text{ cites document } i \text{ AND document } k \text{ cites document } j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$bc_{ij}(k) = \begin{cases} 1 & \text{if document } i \text{ cites document } k \text{ AND document } j \text{ cites document } k \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Indirect citation functions *lc*, *cc*, and *bc* denote functions for longitudinal citations, cocitations and bibliographic coupling, respectively, as defined by Small (1997). The inter-document similarity values form the elements of the similarity matrix, **S**, and are found by weighting and summing the direct and indirect citation functions between each pair of documents:

$$s_{ij} = 2dc_{ij} + \sum_{k=1}^{N} [lc_{ij}(k) + cc_{ij}(k) + bc_{ij}(k)] \quad (5)$$

N is the number of documents in the database. The similarity matrix produced by this function is easily produced using matrix arithmetic. Regarding the documents and their citations as a directed graph, the elements of the adjacency matrix, **A**, of the citation graph can be found using the function:

$$a_{ij} = \begin{cases} 1 & \text{if document } j \text{ cites document } i \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The similarity matrix is computed using the matrix equation:

$$\mathbf{S} = 2(\mathbf{A} + \mathbf{A}^{\mathrm{T}}) + (\mathbf{AA} + (\mathbf{AA})^{\mathrm{T}}) + \mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{AA}^{\mathrm{T}} \quad (7)$$

The first term in the matrix equation above corresponds to the direct citations while the remaining three terms correspond to longitudinal citations, cocitations, and bibliographic coupling, respectively. Weighting the direct citations twice as much as the indirect citations places greater emphasis on direct citations and tends to produce a larger number of meaningful clusters after mapping (Small, 1997). As a final step in calculating the similarities it is necessary to scale the elements of the similarity matrix to be between zero and unity. The elements of the resulting normalized similarity matrix, $\mathbf{S}'$, are calculated using:

$$s'_{ij} = \frac{s_{ij}}{\sqrt{(N_i + 1)(N_j + 1)}} \quad (8)$$

N_i and N_j are the total number of indirect citations linked to document i and document j, respectively. Similarities calculated from citations generally produce meaningful document maps whose patterns expose clusters of documents and the relations among those clusters. Citation based similarity functions are used extensively when working with patents and journal abstracts from sources that provide citation data such as the Science Citation Index. For journal citation data, it is often difficult to identify any citation relations other than cocitations. Notwithstanding, similarity functions based solely on cocitations are very useful and are widely used in bibliometric studies (Garfield, 1994).

Similarity functions based on co-occurrence of terms among documents are also effective. Term co-occurrence similarity functions used here are essentially equivalent to the vector space models that can be found in Salton (1989). Certain data sources provide consistent index terms with each document that can be used to build a similarity function based on the number of common occurrences of index terms among documents. One simple function for generating a similarity matrix based on this technique

is

$$c_{ij}(k) = \begin{cases} 1 & \text{if term } k \text{ is an index term of document } i \text{ and also an index term of document } j \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The elements of the similarity matrix, **S**, are built using the function:

$$s_{ij} = \sum_{\substack{\text{all } k \\ i \neq j}} c_{ij}(k)$$

This function is easily implemented using a short chain of SQL queries. As an example of generating this type of similarity using SQL queries in MS Access, assume a normalized table 'Terms' with two fields: 'DocId' and 'Term'. Each record in the table contains a document identification number in its 'DocId' field and one of the index terms from that document in the 'Term' field. Each document will have as many records in table 'Terms' as there are index terms for the document. In MS Access, queries can be named and referenced by other queries as if they were tables. Construct a query, 'Term_pairs', that produces document number pairs for each co-occurrence of an index term between every pair of documents:

```
Term_pairs:  SELECT k1.DocId AS DocId1, k2.DocId AS DocId2
             FROM Terms AS k1, Terms AS k2
             WHERE k1.DocId > k2.DocId AND k1.Term = k2.Term;
```

Next construct a query that counts the number of pairs with the same document numbers from query 'Term_pairs':

```
Term_sim:  SELECT Term_pairs.DocId1, Term_pairs.DocId2, Count(*) AS Sim
           FROM Term_pairs
           GROUP BY Term_pairs.DocId1, Term_pairs.DocId2;
```

The resulting query, 'Term_sim', lists the pairs of documents in the database which have co-occurring index terms and for each document pair in that list, gives the number of co-occurring index terms for those two documents. The results of this query can be loaded into DIVA to produce the similarity matrix for the document set. In practice it is often necessary to limit the number of pairs generated by this similarity function. Some terms, especially those terms used in queries to originally gather the document set, produce large numbers of co-occurrence document pairs, resulting in a similarity matrix that is not sparse. Inclusion of such terms in the similarity function adds little information when mapping and visualizing the document set, yet dramatically increases computation time. Specific terms are easily excluded from the similarity function by adding terms in the WHERE clause of the SQL query used to generate pairs of documents with co-occurring terms. In the 'Term_pairs' query above, for example, the terms 'polymer' and 'cement' are excluded by modifying the query:

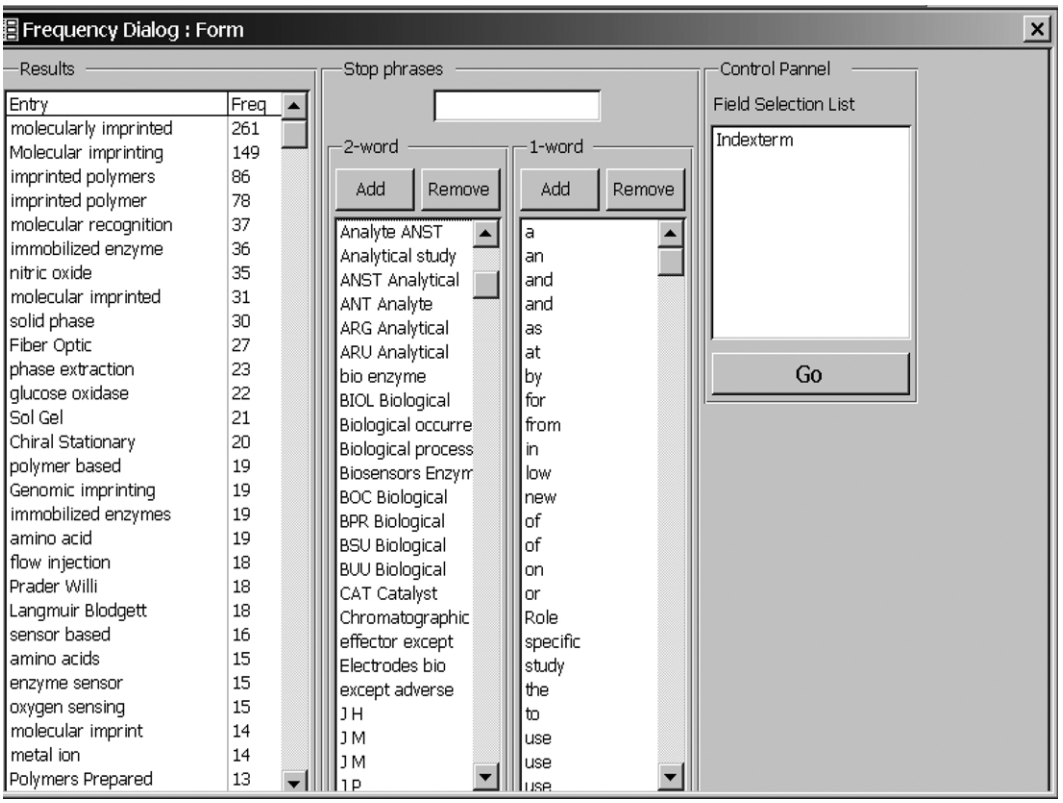

Fig. 2. A user interface for producing and editing tables of one-word, two-word and three-word terms from documents in the database.

Term_pairs:

```
SELECT k1.DocId AS DocId1, k2.DocId AS DocId2
FROM Terms AS k1, Terms AS k2
WHERE k1.DocId > k2.DocId AND k1.Term = k2.Term
AND k1.Term < > "polymer"
AND k1.Term < > "cement";
```

It is also possible to eliminate document pairs that have a low number of co-occurring terms. In the 'Term_sim' query above, for example, document pairs that have two or less co-occurring terms can be excluded by modifying the query:

Term_sim:

```
SELECT Term_pairs.DocId1, Term_pairs.DocId2, Count(*) AS Sim
FROM Term_pairs
GROUP BY Term_pairs.DocId1, Term_pairs.DocId2
HAVING Count(*) > 2;
```

Denote s_{ij} as the number of co-occurring index terms between document i and document j as provided by the queries above. DIVA constructs the scaled similarity matrix, $\mathbf{S}'$, by scaling the similarities to be

between zero and unity using the Dice coefficient similarity measure (Salton, 1989):

$$s'_{ij} = \frac{2s_{ij}}{N_i + N_j} \tag{10}$$

N_i and N_j are the total number of index term co-occurrences over all documents for document i and document j, respectively.

For document sources that do not have an index term field it is necessary to construct an index term table from a table of frequently occurring terms in the titles or abstracts of the documents. Terms in titles and abstracts can be one-word, two-word, or three-word terms. One-word terms are single words in the title or abstract, excluding user specified stop words. Two-word terms are any two consecutive words in the title or abstract, excluding user specified two-word stop terms or two-word terms with internal punctuation or containing user specified single stop words. Three-word terms are any three consecutive words in the title or abstract with exclusions similar to that for two-word terms.

For use by DIVA, tables of one-word, two-word, and three-word terms are constructed using Visual Basic Access procedures. When constructing such tables a good deal of user interaction is required to build lists of stop terms and to set a threshold for excluding less frequent terms. Fig. 2 shows an interactive MS Access form used for generating index term tables from titles or abstracts. The user selects the table and field from the database that will be the source of terms for the index term table. After clicking on the 'Go' command button the index term table is built and the terms are displayed in the 'Results' list box in descending order of frequency. Terms may be selected in the results box and added to the two-word stop term table by clicking on the 'Add' button provided for that purpose on the form. Other tools are provided on the form to edit the one-word and two-word stop term tables.

2.4. Mapping of documents

Documents are mapped to a rectangular planar surface for visualization of relations among documents. The document map also provides visualization of document clusters and relations among those clusters. Presently, DIVA uses two different methods to perform ordination for the purpose of mapping documents. The first method is force directed placement, while the other method is a SOM neural network technique.

Force directed placement (Davidson, Hendrickson, Johnson, Meyers, & Wylie, 1998) is an ordination method in which all the documents are initially placed in the center of the map. After this, a force rule, similar to Coulomb's Law, is iteratively applied. During each iteration the positions of the documents are moved in accordance with the vector sum of forces exerted by the other documents on the map. The attractive force between two documents i and j on the map is computed to be proportional to their similarity value, s_{ij} and inversely proportional to the distance between them on the map. If s_{ij} is zero then a repulsive force is computed between document i and document j which is inversely proportional to the distance between them on the map. The iterations proceed until a stable map results.

The SOM neural network technique (Morris, Wu, & Yen, 2001) is based on training a small rectangular neural network SOM using the rows of the similarity matrix as inputs. Assume N documents in the dataset, numbered from 1 to N. Each document is assumed to have N features, corresponding to the similarity of the document to each of the other documents in the dataset. For example, the value of the jth feature of document i will be s_{ij} from the similarity matrix. Thus, the N dimensional feature vector for

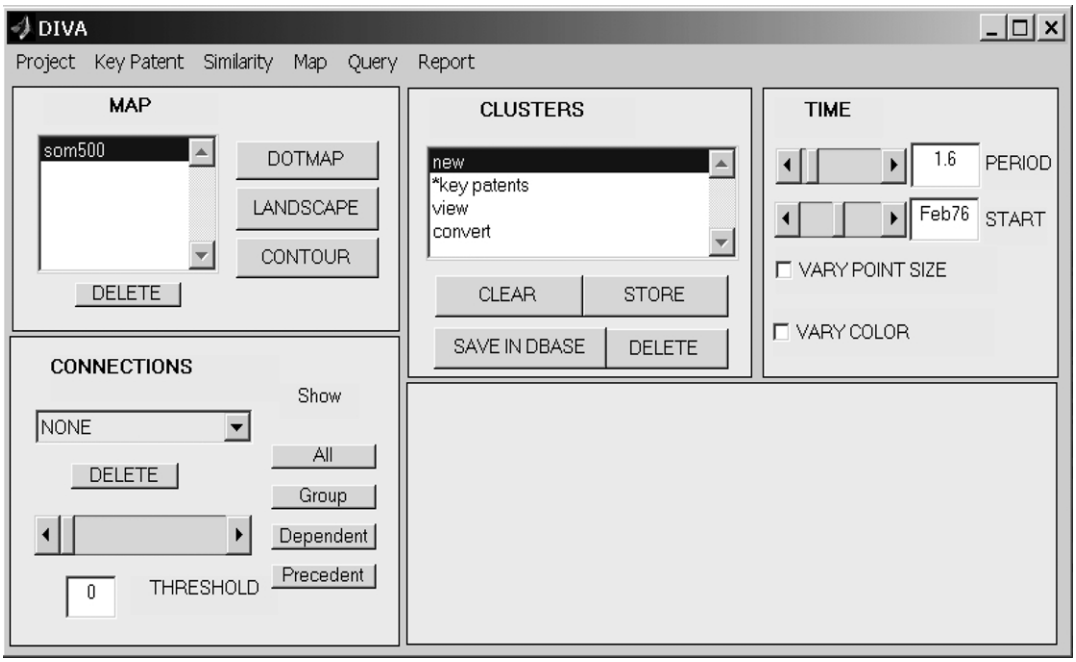

Fig. 3. The DIVA main graphical user interface (GUI).

each document is its corresponding row from the similarity matrix. Documents whose feature vectors are close together in the feature space are related to the same set of documents, and are therefore, related to each other.

The SOM generally used for this ordination technique is a 5 by 5 array of neurons on a rectangular grid. Each neuron has an N dimensional weight that is modified during training. After training, the density distribution of the neuron weights in N dimensional space will match the density distribution of the training vectors, in this case the rows of the similarity matrix. Neurons that are neighbors on the SOM rectangular grid will tend to have weights which are near each other in the N dimensional feature space. This allows a continuous mapping of points in the feature space back to the rectangular area of the SOM.

In practice, the method starts by training the SOM with the rows of the similarity matrix. Typically, for a project having 1000 documents and training for 20,000 epochs, training may take several minutes to a half-hour on a typical personal computer. Dimensionality reduction methods such as principal components analysis may be used to reduce training times. After training, the weights from the SOM neurons are retained. To map a document, the inner product of the document's row in the similarity matrix with each of the neuron weights is computed and negative values set to zero. The responses thus produced are assigned to their corresponding neuron positions on the SOM. Considering this set of responses as a spatial response across the map, the centroid of this spatial response is computed and the document is assigned the coordinates of the centroid.

2.5. Clustering of mapped documents

It is necessary to group the documents on the map into clusters for further exploration and analysis. In DIVA the user does all clustering manually by identifying interesting groups of documents on the

document map using a mouse and graphical user interface. Lists of documents in each cluster are stored in the project database as the user identifies them. Most clusters on the map correspond to specific technological fields or sub-fields in the document set. The user identifies cluster labels after analysis of the frequency of terms in the title and abstract fields of the documents in each cluster.

2.6. Visual exploration

Several types of maps can be generated by DIVA for visualization and exploration purposes. Two-dimensional maps, such as that shown in Fig. 5, are used to display the documents as dots on a rectangular area. This type of map is the most often used as it allows easy exploration of the document set. Functions are provided to manually group documents into clusters, display document links, display document dates, and highlight documents that satisfy specified query conditions. DIVA can also produce document timelines, where documents are clustered along the *y*-axis and plotted on the *x*-axis as a function of time, as in Fig. 7. This presentation format is quite useful for revealing time trends in document clusters, such as the birth and death of technologies. Timeline maps can also reveal patterns of borrowing among technologies.

2.7. Indicators and summary information

DIVA produces reports for each document cluster that summarize information about author frequency, institution frequency, and keyword frequency. Keyword frequency is used to establish cluster labels for technologies within the document set. Author and institution frequency data identifies experts and centers of excellence in technological fields within the dataset. The DIVA report generator additionally plots indicators, that is, time plots of cluster characteristics, such as document counts by year and links to other clusters by year. Document count plots are used to infer the growth or decline in technologies. Plots of links to other clusters by year show patterns of technological borrowing among technologies in the dataset.

3. Exploration functions

DIVA's exploration functions are used to gain insight into the relationships among documents and document clusters. These functions include means to produce different types of maps, cluster and identify documents on the map, visualize document links, visualize document dates, highlight documents which satisfy user queries, and to produce frequency tables of phrases occurring in document clusters.

3.1. DIVA main graphical user interface

Fig. 3 shows the graphical user interface (GUI) which is used to manage projects and control exploration of document maps. Along the menu bar at the top of the GUI are six pull down menus. The 'Project' menu is used for saving and recalling projects and for setting up the ODBC link to the project database. The 'Key Patent' menu is used to create patent databases based on lists of key patents. Patent data is extracted from a database of approximately two million US patent abstracts built from data

acquired from the US Patent and Trademark Office. The 'Similarity' menu is used to execute various similarity functions to build similarity matrices. The 'Map' menu is used to build document maps using either force directed placement or SOM document mapping. The 'Query' menu is used to execute queries. Documents matching such queries are highlighted on the document maps. The 'Report' menu is used to generate summary data for documents by cluster.

Below the menu bars are four groups of controls that are used for displaying maps and for map exploration. The 'Map' control group is used to display dotmaps, landscape maps, or contour maps of ordinations previously produced, named, and stored in memory. Names of previously produced ordinations are listed in the control groups list box for selection and display. Each map produced is displayed in a separate GUI. The 'Clusters' control group is used to identify, name and display clusters of documents on dotmaps. New clusters of documents are identified by dragging a rectangle over the documents on the map using a mouse. The 'Time' control group is used to display variations of document publication times on dotmaps, or to highlight documents within specified time windows. The 'Connections' control group is used to show relations between documents as lines between documents on dotmaps.

3.2. Maps

Two dimensional maps, most commonly used for exploration, show the documents mapped onto a rectangular area with closely related documents mapped near each other. Documents on the map generally fall into clusters according to their technology field. Documents are also commonly mapped as timelines, which are formed by mapping the documents onto the y-axis of a plot and using the document date as the x-axis position on the plot. In this representation the documents usually fall into groups horizontally stretched along the time axis as shown in Fig. 7.

The most convenient representation for two dimensional maps and timelines is the *dotmap*, where the documents are shown as small circles, or dots, on the map as shown in Fig. 5. This method allows easy selection of documents using mouse functions, e.g., pointing to individual documents or dragging a rectangle over a group of documents. Documents are easily highlighted on the map by coloring the interior of the dots. Document links are shown as lines connecting the document dots as shown in Fig. 11. Zooming and panning are easily done and cluster labels can usually be placed in the white space between clusters. Given the advantages of dotmaps, they are usually the map of choice when exploring two dimensional maps and timelines. However, when many documents are mapped on top of each other it is difficult to gauge the density of documents on the map. Additionally, in large document sets with many connections there may be so many lines plotted on the map that interpretation of link patterns is difficult.

Two alternate map formats, landscapes and contour maps, are also available, but are less often used than dotmaps. Landscapes are surface maps where the height of the surface above the plane is proportional to the density of documents at that position on the plane. This representation, shown in Fig. 6, allows better visualization of cluster document populations and document densities than the dotmap format. Clusters on this type of map show up as mountain peaks on the landscape. Labels can be manually attached to the peaks to identify the technologies they represent. Landscapes do not show connections well, because the surfaces obscure the links between clusters. Contour maps offer another method of visualizing document densities on the map. This format (not shown) is a modified dotmap

with contour lines of document density plotted on the map behind the dots and is often used for quick identification of high density clusters on corresponding dotmaps.

3.3. Clustering and marking documents

Functions for clustering and marking documents are available when exploring dotmaps. Clusters are normally created by dragging a rectangle over documents on the dotmap using a mouse. The software visually highlights all documents selected. The user enters the cluster label, after which the label and list of cluster documents is stored in the project database. Summary functions, such as author frequency tables, are available to operate on such lists to help explore them.

3.4. Visualization of document links

A pair of documents are linked if they have a non-zero similarity value. Document links are displayed by drawing lines between documents on the map as shown in Fig. 11. A threshold function is available which only displays links corresponding to similarity values above a threshold set by the user. It is also possible to display only links that connect to user selected documents and clusters. Additional functions are available which operate on citation lists to display dependent and precedent trees of selected documents. The dependent tree of a document is the collection of all documents that cite a document directly, or are linked to the document through a chain of citing documents. The precedent tree of a document is the list of documents that the document cites directly or to which it is linked to by a chain of citing documents.

3.5. Visualization of document dates

Most document sources supply document dates that correspond to the publication date of the document or the issue date of a patent. Dates are often limited to publication year only. Visualization of date information is very useful for investigating trends in document citation patterns and cluster sizes. The most useful method of displaying dates is to use timeline plots, as described in Section 3.2. However, DIVA provides additional functions to display date information on two dimensional dotmaps. A function is provided to vary the size and color of the dots by the date of the corresponding document. Older documents are displayed as small dark dots, while newer documents are displayed as large light dots as shown in Fig. 11. When displayed in this fashion, it is possible to find the younger technologies on the map by looking for clusters that show large concentrations of large, light colored dots. Another date function allows highlighting documents within specified time periods. This is useful to visualize activity in the document set over specific adjacent time periods, such as every five years.

3.6. Highlighting documents that satisfy user queries

One of the most useful functions provided by DIVA allows highlighting documents that satisfy user entered query terms. A query building GUI allows selection of the tables and fields within the database against which the query is executed. Query phrases may be combined using AND or OR functions and it is possible to specify whether phases should be matched exactly or using wildcards. Once executed, documents that match the query are highlighted on the map. Lists of documents that match a query can

be named and saved in the database as well. The query function is quite useful for finding regions of the map with documents dealing with subjects of interest to the user. Documents from specific authors or institutions, or which contain specified technical terms can be highlighted this way.

3.7. Word frequency functions

Word frequency functions are used to find the most common words and phrases in document text fields, such as the title and abstract fields. When exploring and labeling clusters on a map these functions are helpful for identifying the technologies associated with each cluster. Functions are provided for finding the frequency of one-word, two-word and three-word phrases in a selected document cluster.

4. Report generation functions

Reports are usually generated after the user initially explores the document maps. At this stage of the project the clusters have been identified and labeled and summary data can be generated. Reports are used to automatically produce summary information and indicators for each cluster in the dataset and for the dataset as a whole. The size of the reports generated by this function is greatly dependent on the number of clusters identified in the dataset. A dataset with 15–20 clusters will generally yield a report of about 100 pages. Because of this, reports are always initially produced in electronic format for browsing and editing by the user before printing.

4.1. Cluster reports

Cluster reports are produced for each list of cluster members identified by the user. Four frequency tables and three indicator plots are produced. The frequency tables are tables of frequency of authors, institutions, combined authors/institutions, and keywords. Figs. 8 and 9 show examples of these frequency tables. Author and institution frequency tables are usually used to identify experts and centers of excellence for the technological field associated with the document cluster. Keyword frequency tables are used to identify the technological field associated with the document cluster.

The indicator plots provided in cluster reports are plots of the number of documents in the cluster by date and plots of inter-cluster links by date. Fig. 9 shows an example of a plot of documents in a cluster by date. This type of plot gives a good indication of the trends in the technology that the cluster represents. These trends can be related to positions on the S shaped curve of technology growth. The upper plot of Fig. 10 shows a plot of total inter-cluster links by date between a cluster and all other clusters in an example dataset. The lower plot of Fig. 10 shows inter-cluster links by date broken out by individual cluster. Using such indicator plots helps to detect donor technologies and borrower technologies. A technology that is borrowing heavily may indicate that researchers in that technology are using new ideas from outside their field to solve their problems. Such activity may be a precursor to a technological breakthrough. A technology that is donating heavily may indicate the emergence of a new technology that can be broadly applied to many other technologies.

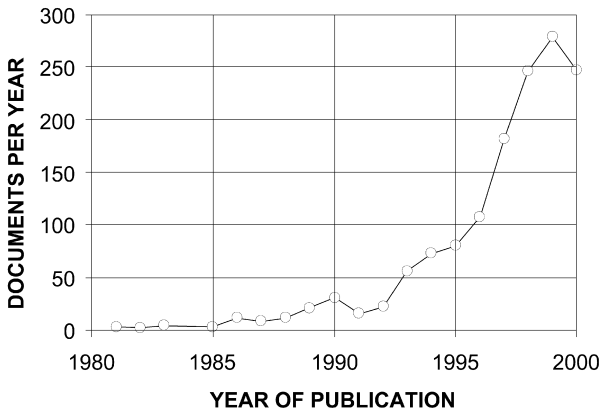

Fig. 4. Plot of chemical sensor articles by year of publication.

4.2. Overall report

The overall report is used to summarize the dataset as a whole. This report is generated by treating the whole dataset as a single cluster and generating a report. All the tables and plots described above for cluster reports are generated with the exception of the plots of inter-cluster links.

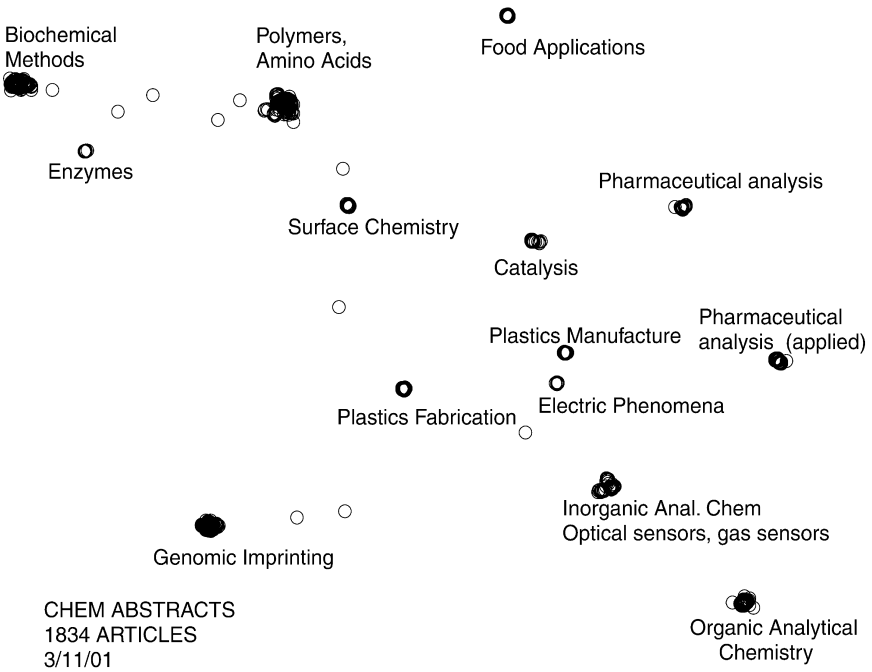

Fig. 5. Map of 1499 chemical sensor articles.

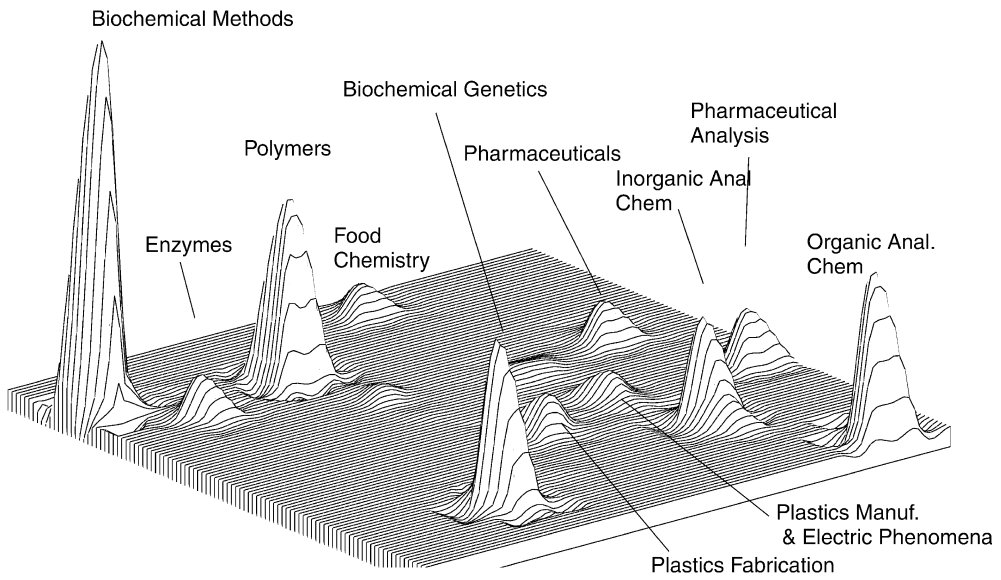

Fig. 6. Landscape map of 1499 chemical sensor articles.

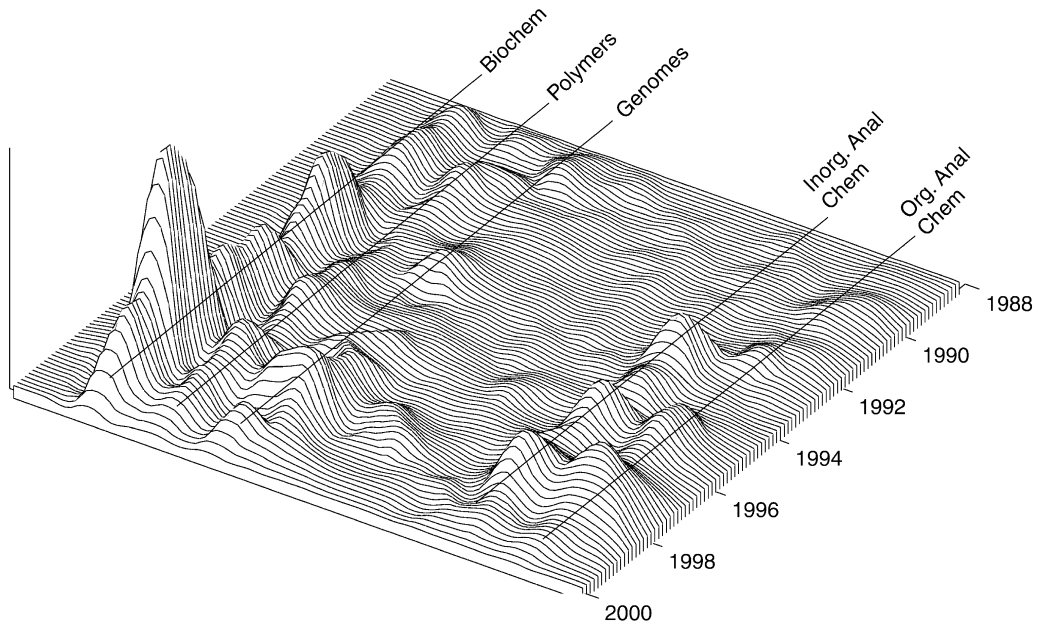

Fig. 7. Timeline of 1499 chemical sensor articles.

Cluster# 2 Report: Page 1 of 3

Most Frequent Authors

Mosbach, Klaus.	17
Karube, Isao.	9
Nicholls, Ian A.	8
Takeuchi, Toshifumi	6
Matsui, Jun	5
Yano, Kazuyoshi	5
Haupt, Karsten	5
Andersson, Lars I.	5
Andersson, Hakan S.	5
Ikebukuro, Kazunori	4

Most Frequent Index Terms

PREP Preparation	105
preparation PREP	72
mol imprinting	64
PROC Process	62
PRP Properties	58
SPN Synthetic	49
Synthetic preparation	47
molecularly imprinted	42
Amino acids	40
PEP Physical	39

Most Frequent Institutions

N/A	24
Chem. Cent.,Univ. Lund,Lund,Swed.	6
Department of Chemistry,University of Washington,Seattle,WA,USA.	5
Research Center for Advanced Science and Technology,University of Tokyo,Tokyo,Japan.	3
Bioorganic Chemistry Laboratory, Institute of Natural Sciences,University of Kalmar,Kalma	3
Department of Chemistry,Nagaoka University of Technology,Nagaoka,Japan.	2
Department of Chemistry,University of California,Irvine,CA,USA.	2
Department of Chemistry,University of Arizona,Tucson,AZ,USA.	2
Chemical Engineering,California Institute of Technology,Pasadena,CA,USA.	2

Fig. 8. Page 1 of cluster report showing frequency tables of authors, institutions, and index terms from the cluster labeled 'Polymers' in 1499 chemical sensor articles dataset.

5. Chemical sensor technology example

The technological field of chemical sensors is a field that has seen a recent explosion in activity starting in the late 1980s. There are several classes of problems to which chemical sensors can be

Cluster# 2 Report: Page 2 of 3

Most Frequent Institutions and Authors

Chem. Cent.,Univ. Lund,Lund,Swed.	Mosbach, Klaus.	4
Bioorganic Chemistry Laboratory, Institute of Natural Sciences,Univer	Nicholls, Ian A.	2
Research Center for Advanced Science and Technology,University of Tok	Karube, Isao.	2
Research Center for Advanced Science and Technology,University of Tok	Yano, Kazuyoshi	2
Research Center for Advanced Science and Technology,University of Tok	Cheong, Soo-Hwan	2
Institute Polymer Research Dresden,Dresden,Germany.	Karube, I.	1

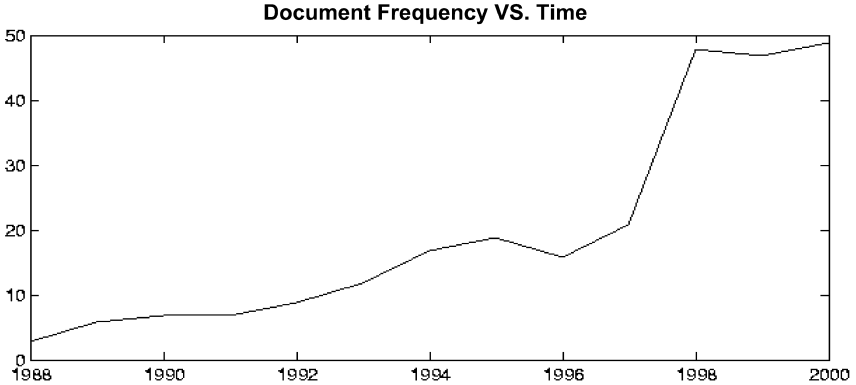

Fig. 9. Page 2 of a chemical sensor 'Polymers' cluster report showing a frequency table of institutions with authors and plot of articles by year of publication.

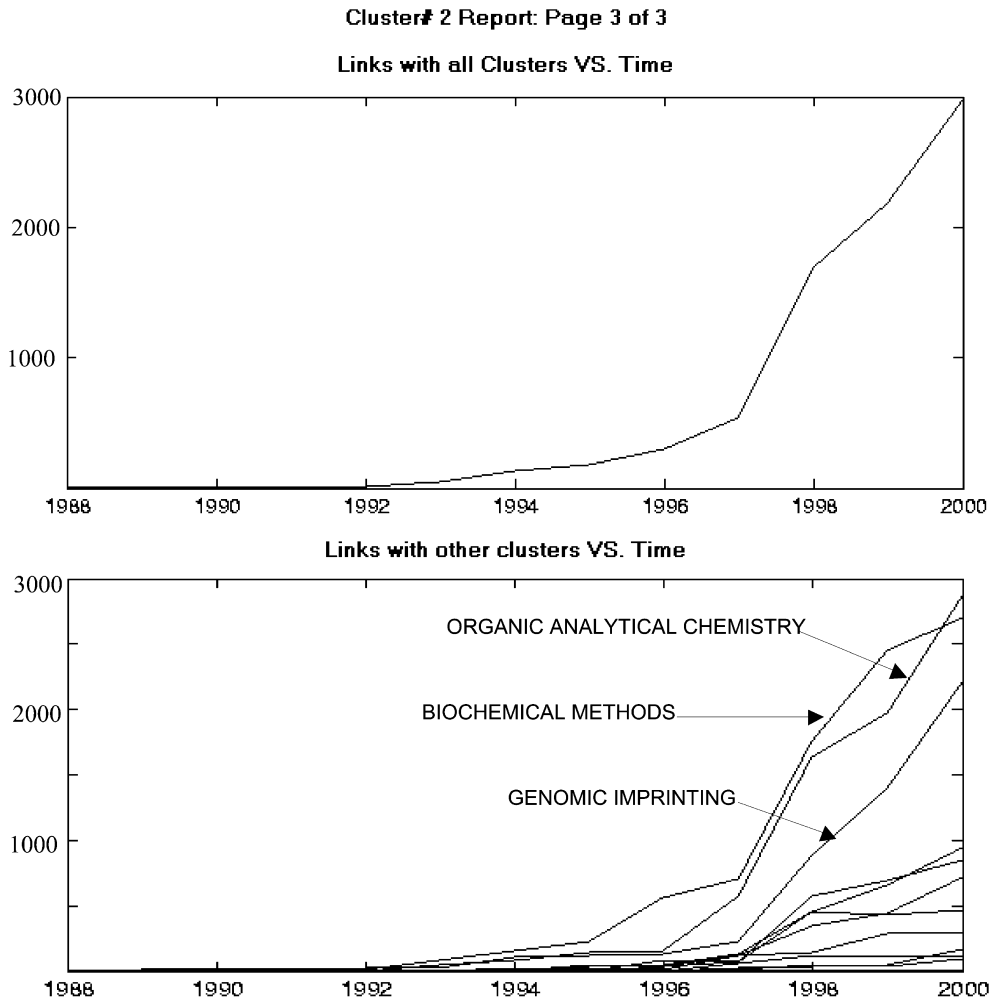

Fig. 10. Page 3 of 'Polymers' cluster report showing a plot of links to all other clusters by year and plot of links to each other cluster by year.

applied: detection of nerve gases and biological agents for military and antiterrorism applications, detection of food spoilage, and detection of drug abuse. A study of chemical sensor technology literature was conducted primarily to identify experts and centers of excellence within the field and to investigate trends within the field.

The project was conducted in the following sequence:

1. A subject matter expert was consulted to help define the scope of the study and to also define search terms for acquiring documents. It was decided to study three document sources: Chemical Abstracts, Science Citation Index, and US patent abstracts.
2. Documents were acquired from the three sources given above and loaded into three separate project databases.
3. For each database, similarity functions were constructed, maps were built, and clusters identified.

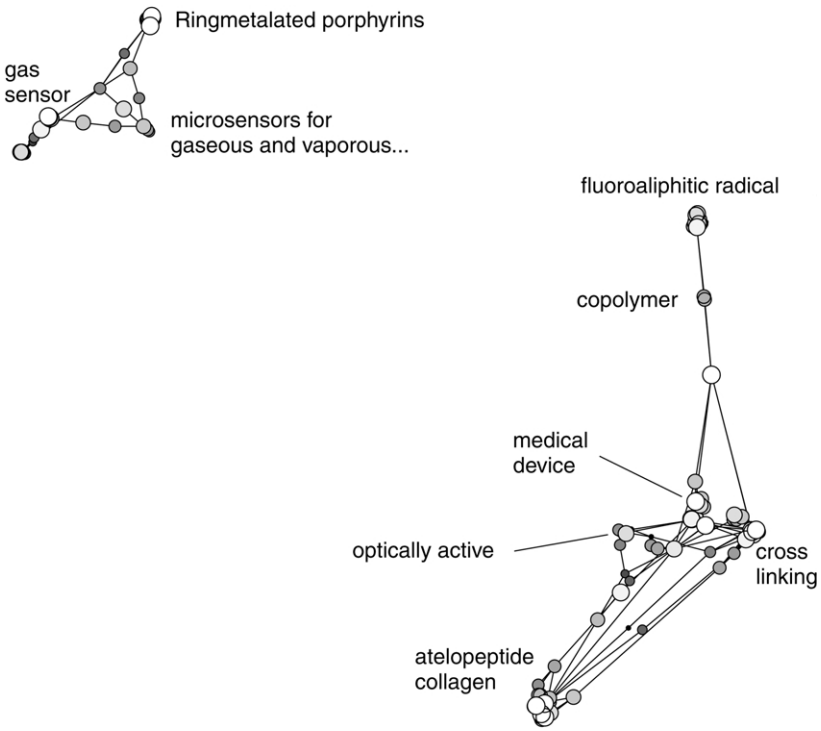

Fig. 11. Map of chemical sensor patents.

Working closely with the subject matter expert, the contents of the databases were checked to insure full subject matter coverage.

4. Final maps were constructed and cluster labels were generated by using phrase frequency tables generated by DIVA. Experts and centers of excellence for each cluster were identified using author and institution frequency tables.
5. Plots of indicators from each of the three databases were used to investigate trends in cluster sizes and inter-cluster links.

Sections 5.1–5.3 below describe the gathering and analysis of data for each of the three data sources used in this example.

5.1. Chemical Abstracts

Abstracts from Chemical Abstracts were gathered using various queries based on the root terms ‘molecular imprinting’, ‘porphyrins’, and ‘sensor’. Abstracts produced before 1988 were excluded, yielding a final count of 1834 abstracts for visualization and analysis. Fig. 4 shows a plot of the document publication rate for this dataset as a function of time. The exponential increase in documents indicates a technology in the early stages of growth.

The abstracts were processed to generate a table of two-word terms, after which a similarity function based on the sum of the co-occurrence of two-word terms between document pairs and co-membership in Chemical Abstracts sections and subsections was used to generate the

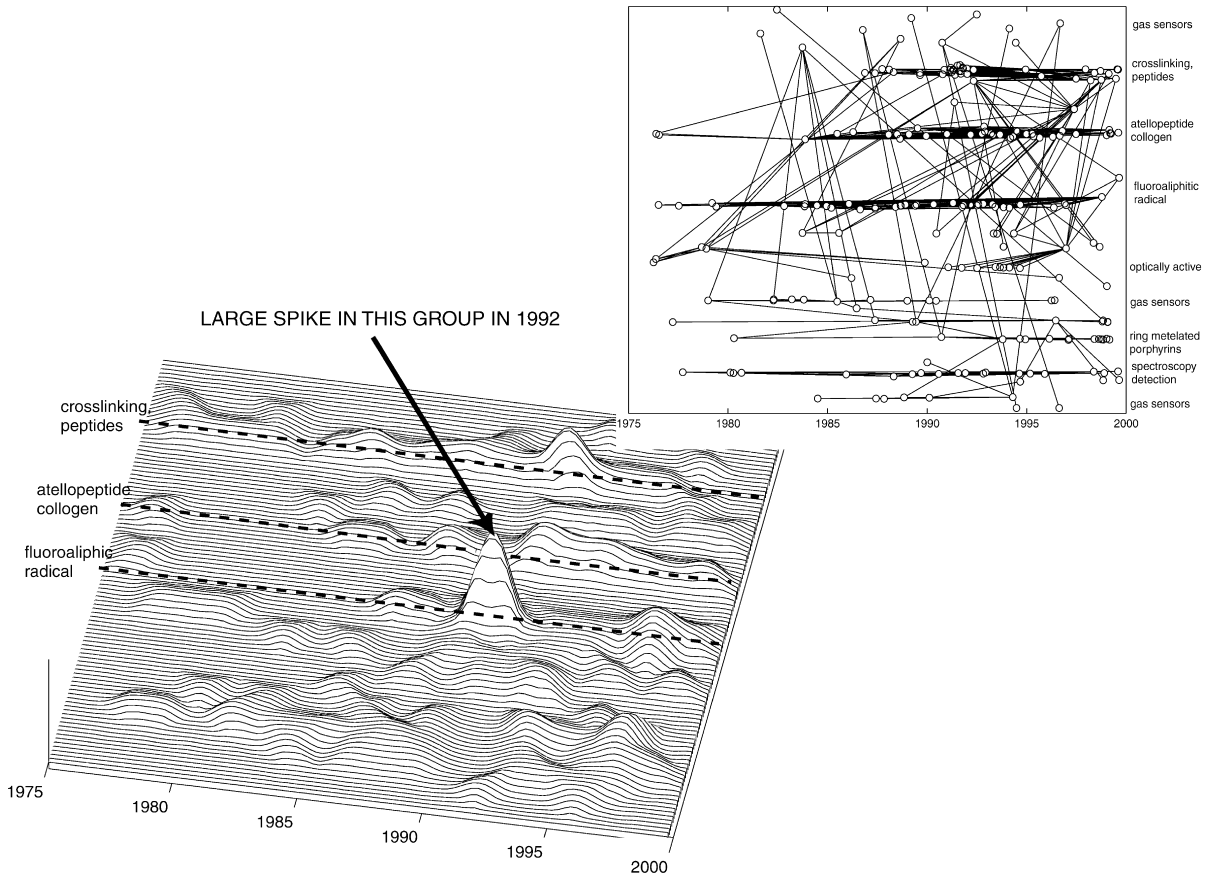

Fig. 12. Timeline of chemical sensor patents.

similarity matrix. Force directed placement was used to produce a map of the documents. Clusters were labeled by examining frequency tables of two-word terms occurring in the abstracts of documents within each cluster. Fig. 5 shows a labeled map of the 1834 chemical sensor abstracts. There are fourteen labeled clusters on the map of which 'biochemical methods' and 'polymers' contain the highest population of documents. Fig. 6 shows a landscape of the document map that helps to visualize the relative population of the document clusters. Fig. 7 shows a timeline of the documents with five of the largest clusters labeled. Note the great surges of publication activity in the 'biochemical methods' cluster in the years 1994 and 1998. This figure helps to establish the relative age of different technologies in the chemical sensor field.

Figs. 8–10 show a cluster report for the cluster identified as 'polymers'. Important authors and institutions in this field are readily found using the frequency tables provided. The plot in Fig. 9 shows a great surge in activity in this field in 1997. The lower plot in Fig. 10 shows that the clusters identified with 'organic analytical chemistry', 'biochemical methods', and 'genomic imprinting' are closely linked to the 'polymer' cluster.

5.2. Patents

A dataset of 223 patents was generated from a seed list of seventeen key patents identified by the

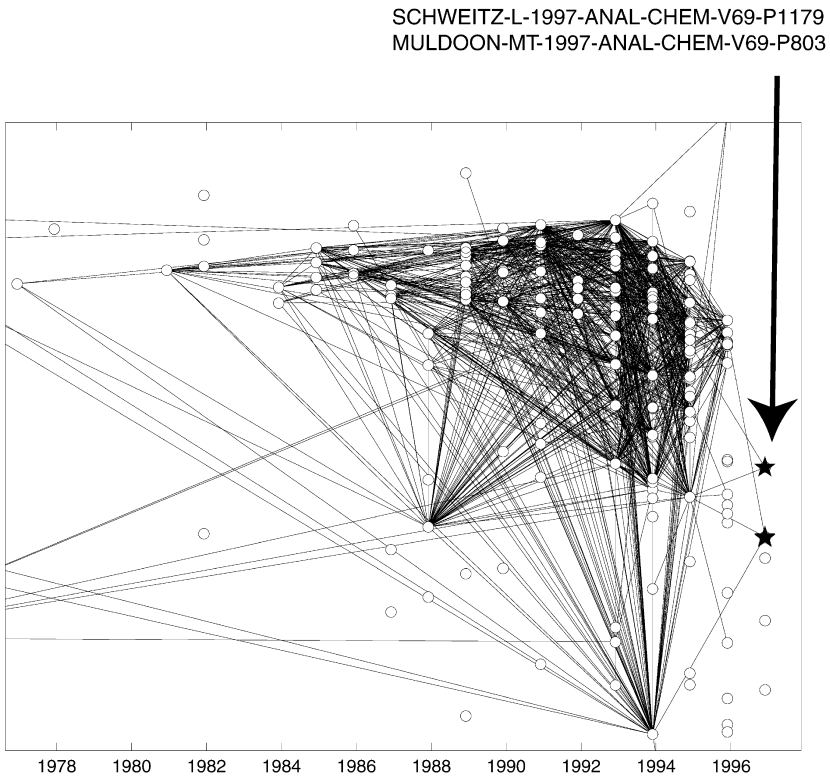

Fig. 13. A timeline of chemical sensor articles from the Science Citation Index showing cocitation links.

subject matter expert. A similarity function, based on patent citations, was applied and force directed placement was used to generate the map of Fig. 11. Labels for the patent clusters were derived from frequencies of two-word terms in titles and abstracts of the patents. In this map recent patents are displayed as large white circles, while older patents are displayed as smaller dark circles. Note the 'cross-linking, peptides' cluster that appears to have many recent patents. This is verified by noting that a great many patents in this cluster occur in 1998 and 1999 as shown in the timeline in the top right corner of Fig. 12.

Fig. 12 shows a timeline of the patents in dotmap format and a corresponding landscape. There does not appear to be a significant amount of inter-cluster citation activity in this set but a few individual patents cite heavily from several different clusters, inviting further investigation. Note a great surge of activity in the 'fluroaliphic radical' cluster in 1992 that invites further investigation.

5.3. Science Citation Index

Fig. 13 shows a timeline generated from documents extracted from the Science Citation Index CDROM product, for the years 1993–1998, for chemical sensors. Citations from the documents from this five year period extend from 1997 back into the 1800s, but most citations occur after 1980. An initial set of abstracts was generated by finding all documents that cite, or are cited by, a set of key articles supplied by the subject matter expert. Further key articles were found from the initial dataset by

MOST CITED DOCUMENTS

CITATION	freq
MOSBACH-K-1994-TRENDS-BIOCHEM-SCI-V19-P9	116
VLATAKIS-G-1993-NATURE-V361-P645	77
WULFF-G-1995-ANGEW-CHEM-INT-EDIT-V34-P1812	66
SHEA-KJ-1994-TRENDS-POLYM-SCI-V2-P166	60
FISCHER-L-1991-J-AM-CHEM-SOC-V113-P9358	44
SELLERGREN-B-1988-J-AM-CHEM-SOC-V110-P5853	41
KEMPE-M-1994-J-CHROMATOGR-A-V664-P276	40
ANDERSSON-LI-1995-P-NATL-ACAD-SCI-USA-V92-P4788	37
KRIZ-D-1995-ANAL-CHEM-V67-P2142	34
MATSUI-J-1993-ANAL-CHEM-V65-P2223	31
SELLERGREN-B-1993-J-CHROMATOGR-V635-P31	30
ANDERSSON-LI-1990-J-CHROMATOGR-V516-P313	29
RAMSTROM-O-1994-TETRAHEDRON-ASYMMETR-V5-P649	27
RAMSTROM-O-1993-J-ORG-CHEM-V58-P7562	26
BEACH-JV-1994-J-AM-CHEM-SOC-V116-P379	26
WHITCOMBE-MJ-1995-J-AM-CHEM-SOC-V117-P7105	26

MOST CITED AUTHORS

author	Count
WULFF-G	275
SELLERGREN-B	208
ANDERSSON-LI	157
KEMPE-M	140
MOSBACH-K	138
SHEA-KJ	113
MATSUI-J	89
VLATAKIS-G	77
KRIZ-D	74
RAMSTROM-O	64
FISCHER-L	44
NICHOLLS-IA	36
OSHANNESSY-DJ	35
MAYES-AG	34
PILETSKY-SA	31
HOSOYA-K	26
BEACH-JV	26
WHITCOMBE-MJ	26
DHAL-PK	24
MULDOON-MT	23
ROBINSON-DK	23
MORIHARA-K	23

Fig. 14. Frequency tables of cited authors and cited documents for chemical sensor articles taken from the Science Citation Index.

identifying a few other articles with very high citation counts. The final dataset consisted of 333 documents. The similarity function used on this dataset, based on cocitations, did not yield a useful map, but clustered most documents into one large 'clump' on a two dimensional map (not shown). Fig. 13 shows a timeline with pairs of documents that have more than six cocitations connected by lines. Several documents in the timeline show very large numbers of cocitations, indicating the importance of those documents. Additional information on important authors and documents in this dataset are revealed using frequency tables of cited documents and cited authors as shown in Fig. 14.

In Fig. 13 note the two 1997 documents in the set that have received eight citations in the period 1997–1998, the last two years in the dataset. This is a large number of citations for such a short period after publication. Such anomalously high citation counts can be generated by review articles and authors that cite their own papers frequently. It was determined that neither document falls into one of these categories so it is reasonable to assume that these two papers could contain important work in this field. A check of current citations reveals in the four years since publication these two documents have been cited 81 and 73 times, supporting the prediction that these two articles may be important work in the field.

6. Conclusion

DIVA provides a valuable tool for assessing technology for the purpose of technology forecasting or competitive intelligence. The ability to visualize relations among collections of documents from a technological field allows the user to make inferences and draw conclusions about the relations and

trends within the technology that would not otherwise be apparent. DIVA provides many useful tools for exploring links among documents, and for exploring the time trends of those links.

To date we have used DIVA to study scientific literature from many fields: chemical sensors, oilfield polymers, oilfield cements, water-in-oil emulsions, hydrophobically modified water soluble polymers, silicon-on-insulator integrated circuits, and even technological substitution curves. Based on experience with these applications we have plans for several extensions to the DIVA system for visualization, exploration and generation of technology indicators.

AcknowledgEments

The authors gratefully acknowledge support from Halliburton Energy Services under Contract Number AA559530. The authors would also like to thank Dr Kevin Boyack of Sandia National Laboratories for his helpful collaboration and cleverly constructed SQL queries. Finally, thanks to Dr James Harmon of Oklahoma State University, who acted as our subject matter expert for chemical sensors.

References

Davidson, G., Hendrickson, B., Johnson, D., Meyers, J., & Wylie, B. (1998). Knowledge mining with VxInsight: discovery through interaction. *Journal of Intelligent Information Systems*, *11*, 259–279.

Garfield, E. (1994). The concept of citation indexing: A unique and innovative tool for navigating the research literature. *Current Contents*, 1994 January 3.

Kostoff, R. N (1997). *The Handbook of Research Impact Assessment, 7th edition*, DTIC Report ADA 296021, Office of Naval Research, 800 N. Quincy St. Arlington VA 22217.

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, *32*(6), 67–71.

Mathworks (1999). *Database Toolbox*, The Mathworks, Inc. 3 Apple Hill Drive, Natick, MA.

McCain, K. W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics*, *41*(3), 389–410.

Mogee, M. (1991). Using patent data for technology analysis and planning. *Research Technology Management*, *34*(4), 43–49.

Mogee, M. E. (1997). Patents and technology intelligence. In W. B. Ashton, & R. A. Klavans (Eds.), *Keeping abreast of science and technology, technical intelligence for business*. Columbus, OH: Battelle Press.

Morris, S. A., Wu, Z., & Yen, G. (2001). A SOM mapping technique for visualizing documents in a database. *Proceedings of the IEEE International Joint Conference on Neural Networks, Washington DC, USA*, July 14–19.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA, USA: Addison-Wesley.

Small, H. (1997). Update on science mapping: creating large document spaces. *Scientometrics*, *38*(2), 275–293.

Spasser, M. A. (1997). Mapping the terrain of pharmacy: co-classification analysis of the international pharmaceutical abstracts database. *Scientometrics*, *39*(1), 77–97.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artifical Intelligence*, *91*, 183–203.