Patinformatics: Tasks to tools

1 author:

Anthony J. Trippe
18 PUBLICATIONS   **725** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   Current Challenges in Patent Information Retrieval View project

Patinformatics: Tasks to tools

Anthony J. Trippe *

112 Grover Rd. Ext., Medford, MA 02155, USA

Abstract

This article starts with an overview of the field of patinformatics—the science of analyzing patent information to discover relationships and trends. This is followed by a survey of many common analysis tasks in this field, and many of the software tools available to tackle these tasks. The survey is set out under the tasks of list cleanup and grouping of concepts; list generation; co-occurrency matrices and circle graphs; clustering of structured data; clustering of unstructured data; mapping document clusters; adding temporal component to cluster map; citation analysis; subject/action/object functions. The author concludes that patinformatics has developed very rapidly over the last few years, and provides continuing challenges and opportunities in making optimal use of the resources available to achieve reliable and meaningful results. Useful tables summarizing aspects of this survey are included.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Patinformatics; Patent information analysis; Software analysis tools; Patent intelligence; List cleanup; Concept grouping; List generation; Co-occurrency matrices; Circle graphs; Data clustering; Mapping document clusters; Citation analysis

1. Introduction

To begin, a few definitions may be in order, starting with the term patinformatics [1]. This term is borrowed from the more common fields of bioinformatics or cheminformatics. By definition, bioinformatics [11] is the science of analyzing large amounts of biological data using computational methods. For example, researchers use genomic data to discover relationships or trends between different genes or biological pathways where looking at smaller datasets could mean missing a connection. In a similar fashion, the term patinformatics describes the science of analyzing patent information to discover relationships and trends, which would be difficult to see when working with patent documents on a one-on-one basis. In this way Patinformatics can be thought of as a macro-level science, that is analysis that deals with large quantities of patent information. There also exists the concept of micro-level analysis, which deals with relatively small numbers of patent documents but for which a high degree of precision and detail are required. Some of the tools described could potentially be used for micro-level tasks but often these tasks are best suited to a manual, intellectually based approach. Therefore, the term Patinformatics is meant to encompasses all macro-level forms of analyzing patent information including:

- patent intelligence (the use of patent information to identify the technical capabilities of an organization and the use of that intelligence to develop a strategy for strategic technical planning),
- patent mapping (sometimes described as white space mapping, which uses published patent data to create a graphical or physical representation of the relevant art pertaining to a particular subject area or novel invention),
- patent citation analysis (the study of patent citations for potentially determining a patent's value or, perhaps more reliably, the identification of potential licensing partners or leads based on the citation of an organization's patents by another company in the same or a completely different market space).

Patinformatics can also cover additional applications of patent information involving a subsequent analysis step. The key underlying property in each of these diverse areas is the analysis step. In the past few years there have been a number of articles published, which

* Fax: +1-617-444-6680.
E-mail addresses: tony@trippe.com, trippe@patinformatics.com (A.J. Trippe).

have dealt with various aspects of patent analysis. Ernst [2] has looked at patent applications and their effect on business performance. Mogee and Breitzman [3] have examined the different applications of patent analysis. Breitzman [4] discusses the use of patent citation analysis for targeting and valuing merger and acquisition candidates, Adams has looked at the patenting activity of Pharmacia prior to their merger with Pfizer [5] and Awaya [6] applies a twist to conventional patent citation analysis by combining it with a semantic overlap between the reference documents. In each of these instances the discussion has centered on describing an analysis technique and demonstrating the outcome towards a business objective.

At the same time there has been an increased use and availability of software tools for assisting patent analysts in performing patinformatics experiments. There has been an increase in articles that have looked specifically at how these tools can be applied to specific analysis tasks. In one instance [7] a comparison was made using Aurigin's ThemeScape tool and IBM/Synthema's Technology Watch, looking at clustering methods using intellectually assigned coding versus source titles and abstracts. A second tools based study [8] was conducted using tools from OmniViz to compare source titles and abstracts to the enhanced titles and abstracts produced by commercial patent abstracting services.

With the increased proliferation and variety of software tools for performing patinformatics a discussion aligning common patinformatics tasks to these tools seems appropriate. The remainder of this article will accordingly focus on examining a number of common patent analysis tasks and describing which software tools are most likely to be useful in performing these tasks. Hopefully this, along with other information they already have or decide to seek out, will help practitioners make smarter selections when it comes time to invest in patinformatics tools.

This paper will focus on nine tasks that are commonly conducted by practitioners of patinformatics including: List Clean-up and Grouping (Section 2), List Generation (Section 3), Co-occurrency Matrices and Circle Graphs (Section 4), Clustering of Structured (Fielded) Data (Section 5), Clustering of Unstructured (Text) Data (Section 6), Mapping Document Clusters (Section 7), Adding Temporal Components to Cluster Maps (Section 8), Citation Analysis (Section 9) and Subject/Action/Object (SAO) Functions (Section 10). A summary of these tasks along with their definition, utility and the tools, which can be used for conducting these tasks, can be found in Table 1. The tools covered include the Aureka Online System from Micropatent, the Citation Bridge from the Metrics Group, Citation Link, Text Clustering and PatLab II from Delphion/ Thomson, ClearResearch from ClearForest, DOORS from M-Cam, Knowledgist from Invention Machine

Corporation, OmniViz from OmniViz, SciFinder Panorama from Chemical Abstracts, SEMIO Map from SEMIO, Technology Watch from IBM/Synthema, VantagePoint from Search Technologies, Vivisimo from Vivisimo and the Analysis and Citation Modules from Wisdomain. A list of these tools and the Vendor's URL can be found in Table 2. A few additional tools that cannot be covered due to space and time constraints are also included in Table 2.

2. List cleanup and grouping of concepts

List cleanup can be defined as the manual or automatic standardization of terms within a data field. Grouping generally also includes standardization, such as when misspelled terms are cleaned up, but can also include the identification of synonyms and alternate terms for describing a similar concept. These tasks are required in order to produce statistically relevant results. Grouping allows synonymous terms to be combined together so that their true value in a data set can be accurately assessed. Without performing a list clean up or grouping exercise the numbers generated in a list could be misleading and not truly represent the relative importance of items or terms within a data field.

Tools that employ some form of list cleanup or grouping feature include: VantagePoint, Microsoft Excel, ClearResearch, Aureka ThemeScape and OmniViz.

One of the key advantages to the VantagePoint product is its list cleanup ability. Employing fuzzy logic algorithms the tool automatically examines each item in a data field and attempts to group those elements that look as if they belong together. The algorithm basically works by employing string matching, the assumption being that two or more items that share a percentage of letters in common are probably variations on the same theme. This is especially true with regard to misspellings and small changes such as the use of Co. versus Corp. or corporation. The algorithms do not work with abbreviations however (such as IBM for International Business Machines) so the system allows the algorithm to make a first pass at cleanup and then the user can manually finish the group by adding items to existing groupings or creating new ones.

VantagePoint also contains a grouping tool that is separate from the list cleanup tool. The feature is essentially used for grouping synonyms and alternate terms together. This can be very powerful when an analyst wants to take a collection of classification codes (Derwent Manual Codes for instance), where the code itself does not give an indication of its meaning, and group them together under a single, common heading with a meaningful name. There may be several codes that represent HIV Protease Inhibitors for instance but grouping them under a single heading allows all of them

Table 1
Summary of patinformatics tasks, definitions, utility and tools

Technique	Patinformatics definition	Patinformatics utility	Available Patinformatics tool
List Cleanup and Grouping of Concepts	Manual or automatic standardization of terms within a data field	List cleanup is required in order to produce statistically relevant results. Grouping allows synonymous terms to be combined together so that their true value in a data set can be accurately assessed	VantagePoint, MS Excel, Clear-Research, OmniViz, Aureka Theme-Scape
List Generation (Histograms)	Provides counts of various patent related metrics within individual data fields	Allows the statistical comparison of two or more entities in the same data field	VantagePoint, MS Excel, ClearResearch, Aureka Reporting, Knowledgist, Technology Watch, Wisdomain Analysis Module, Delphion PatLab II, SciFinder
Co-occurrency Matrices and Circle Graphs	Data fields are placed on an X and Y axes or on opposites sides of a circle. Number of overlapping occurrences of shared X and Y can be seen as numbers within the matrix or as lines of increasing width connecting items on the circle	Allows connections to be made between two or more fields of information and provides a representation of how strong the connection is	SciFinder Panorama, VantagePoint, ClearResearch, Aureka Reporting, Wisdomain Analysis Module, Delphion PatLab II
Clustering of Structured (Fielded) Data	Intellectually assigned classification systems produce a standardized code that can be used as a means of organizing documents that share a similar coding structure	Documents, which share a high percentage of codes in common, are likely to be similar. Allows a large number of documents to be organized	Technology Watch, ClearResearch, OmniViz, VantagePoint
Clustering of Unstructured (Text) Data	Raw text is processed to identify concepts and phrases contained within. As with the clustering of structured data, concepts instead of codes are used to group documents that share a high degree of overlap	Documents, which share a high percentage of concepts in common, are likely to be similar. Allows a large number of documents to be organized	Aureka ThemeScape, ClearResearch, OmniViz, Vivisimo, Delphion Text Clustering, VantagePoint
Mapping Document Clusters	Document clusters are arranged in two-dimensional space creating a map. Collections of documents, which share elements in common, are placed closer together geographically while collections with less similarity are placed further away	Allows relationships between clusters to be identified. Creates a visual representation of a document collection from a high-level view	Aureka ThemeScape, Technology Watch, ClearResearch, OmniViz, VantagePoint
Adding Temporal Component to Cluster Map	A time dimension can be called out on a map usually by means of alternate colors	User can follow the progression of a subject as it develops or evolves	Aureka ThemeScape, ClearResearch, OmniViz
Citation Analysis	When patent documents are examined relevant prior art is mentioned on the search report or on the front page of the documents. The number of citations can be counted or followed as they link documents together	Hyperbolic trees are used to show relationships between patents that cite one another. Citation counts are used to discover potentially pivotal documents	M-CAM, Aureka Citation Trees, Delphion Citation Link, Metrics Group Citation Bridge, Wisdomain Citation Module
SAO Functions	Parts of language that are used to describe the teachings that the author wants to portray. Key SAOs encapsulate the technical learnings contained in a document	SAOs can be described as problems and solutions. By identifying SAOs the teachings of a document can be isolated and examined from the rest of the document creating a knowledge base	Knowledgist

to be counted together when the subsequent analysis is done.

In an effort to also include tools that may already be readily available within an individual's organization it should be pointed out that Microsoft Excel contains a grouping tool within its Pivot Table function. It is beyond the scope of this article to explain the use of Pivot Tables but the function does allow items in a list to be grouped together manually by the analyst.

One of the features of the ClearResearch tool is the ability to create custom taxonomies. The tool generates many taxonomies automatically via its information extraction or smart tagging mechanisms but users may want to create their own customs one for specific analysis. New taxonomies can be created and items from existing taxonomies can be simply dragged and dropped onto the new one or an intelligence search can be initiated within the system for finding phrases that match or are related to the items of interest.

Table 2
Patinformatics tools and their vendors

Tool name	Vendor site URL
Aureka Citation Tree	http://www.aurigin.com/aureka_online.html#citation
Aureka Reporting Module	http://www.aurigin.com/aureka_online.html#reporting
Aureka ThemeScape	http://www.aurigin.com/aureka_online.html#themescape
Autonomy	http://www.autonomy.com
Citation Bridge—Metrics Group	http://www.patentcitations.com
Citation Link—Delphion	http://www.delphion.com/products/research/products-citelink
ClearLab Studio—ClearForest	http://www.clearforest.com
ClearResearch—ClearForest	http://www.clearforest.com/products/products.asp?id=1
Delphion PatLab II	http://www.delphion.com/products/research/products-patlab
Delphion Text Clustering	http://www.delphion.com/products/research/products-cluster
DOORS—M-Cam	http://www.m-cam.com/doors/
InXight Categorizer	http://www.inxight.com/products/categorizer/
Knowledgeist—Invention Machine Corp.	http://www.invention-machine.com/prodserv/knowledgist.cfm
OmniViz	http://www.omniviz.com
SciFinder Panorama—Chemical Abstracts	http://www.cas.org/SCIFINDER/panorama.html
SEMIO Map	http://www.entrieva.com/entrieva/products/semiomap.asp?Hdr=semiomap
Technology Watch—IBM Synthema	http://www.synthema.it
VantagePoint—Search Technologies	http://www.thevantagepoint.com
Vivisimo	http://www.vivisimo.com
Wisdomain Analysis Module	http://www.wisdomain.com/AnalysisModule.htm
Wisdomain Citation Module	http://www.wisdomain.com/CitationModule.htm

The Aurigin ThemeScape and OmniViz tools share a similar ancestry (both products were originally generated from Battelle) and hence have a number of features in common. They both cover grouping in a similar way for instance. Groups can be created by conducting a search for a specific item or collection of items or by selecting documents on the respective visualizations (ThemeScape map for ThemeScape and a Galaxy view for OmniViz). Both tools allow a form of list logic to be used on the existing groups. The list logic choices are union (analogous to an OR operator), intersection (an AND operator) and difference (a NOT operation). They also allow groups to be assigned different colors that allow the analyst to highlight how one group compares to another (or several) visually. A practical example of this will be described in Section 8 below (Adding Temporal Component to Cluster Map).

It could also be suggested that synonym lists are a form of grouping, and many of the tools in Table 2 have this feature. Since they are normally simple lists of synonymous terms and do not involve a great deal of interactive use (beyond the generation of the list) by the user they are not covered here.

3. List generation (Histograms)

List generation provides counts of various patent related metrics within individual data fields and allows the statistical comparison of two or more entities. This is perhaps the most common task conducted in patinformatics exercises. It may also be (with the possible ex-

ception of patent citation analysis) one of the more controversial tasks since it is easy to be misled by raw numbers without context. It is beyond the scope of this article to go into the debate revolving around "lies, damned lies and patent statistics" [1] but suffice it to say that practitioners should be careful when relying on lists of numbers alone.

Tools that can create lists include: VantagePoint, Microsoft Excel, ClearResearch, Aureka Reporting, Knowledgist, Technology Watch, Wisdomain Analysis Module, Delphion PatLab II and SciFinder.

In VantagePoint creating a list simply involves pressing the List button and selecting which field the list should be created from. For most lists it is strongly recommended that a cleanup routine be applied to it first to help increase its relevance. Once the list is generated it can be sorted and if the user selects the number next to the list item's name another list, this time containing the titles of the documents which contain the list element will be populated in the left most portion of the window. The titles can be clicked on to produce the document that contained the item from the list.

Technically Microsoft Excel can create lists but it normally requires typing them in manually or importing them from a delimited file. Again, the Pivot Table feature allows lists to be manipulated in a number of useful ways including reduction to a Top 10 (this could be any

[1] With apologies to Mark Twain (to whom the original quote, "lies, damned lies and statistics is attributed) and Edlyn Simmons (gave a presentation on this topic at the 2003 Patent Information Users Group Meeting in May of 2003).

number desired by the user, not just 10) and automatic generation of a chart incorporating the list and its values.

With ClearResearch the user selects a taxonomy they would like to create a list from and the system generates a histogram of the top 10 terms within it. Clicking on a bar takes the user to the documents, which contain the selected item. In past versions of this software the analyst was not able to expand the list beyond the top 10, which severely limited the usefulness of this function.

The Aureka Reporting module allows a user to generate a number of top 10 lists but they suffer from not being editable in the online version of the tool without first downloading the data to a local computer. When the data is received it can be opened with Excel and the Excel tools for list cleanup and reporting (Pivot Tables and Charts in its most powerful and flexible aspect) can be employed.

When Knowledgist generates its SAO functions (described in Section 10 below), the stock output incorporates a list of all the SAO functions found. This list can be sorted such that the SAO functions that were discovered most frequently within a collection of documents can be seen at the top of the list. Otherwise the list of SAOs can be sorted alphabetically by subject, action or object.

Technology Watch is mostly concerned with clustering documents—discussed in Section 6 below—but if a data field is designated as being available for analysis (but not necessarily used as a criteria for clustering) then the analyst can see a list of the top items within the field using this tool. Unfortunately, as with the ClearResearch and Aureka Reporting tools this list cannot be edited within the program and can often be misleading without some form of cleanup.

Since Delphion's PatLab II was created by Wisdomain it can be expected that the two tools would behave in a similar way. Both of them create a standard list of terms based on the field the user selected. Once again however, in order to clean and manipulate this information the data has to be exported.

The Analyze feature in SciFinder will create histograms from a collection of search results that are retrieved by this tool. A check box next to each item in the histogram allows the analyst to specify their interest in the documents containing that term and will narrow the answer set to just those. One of the biggest complaints patent information professionals have with lists of patent related data involves the lack of review and cleanup features. The SciFinder Analyze tool is also sparse in this area and cleanup has to be done once the data is exported.

Clearly the need to incorporate more advanced list cleanup features in products that feature lists of patent data will need to be an area of improvement in the future.

4. Co-occurrency Matrices and Circle graphs

This is a type of analysis where data fields are placed on an X and Y axes or on opposite sides of a circle. The number of overlapping occurrences of shared X and Y can be seen as numbers within the matrix or as lines of increasing width, connecting items on the circle. An analysis of this type allows connections to be made between two or more fields of information and provides a representation of how strong the connection is. This is often very useful for comparing data items such as Assignee and patent filing year or technology subject area (represented by a patent office classification code for instance). With Assignee on the Y-axis and filing year on the X patenting trends by year of a relatively large number of organizations can be quickly compared.

Co-occurrency Matrices and Circle graphs can be generated by: SciFinder Panorama, VantagePoint, ClearResearch, Aureka Reporting, Wisdomain Analysis Module and Delphion PatLab II.

Following the menu and button driven interface that SciFinder is known for the Panorama module also works by asking the user to choose from a number of different fields such as Chemical Abstracts Service (CAS) indexing terms and CAS Registry Numbers. Once the fields for the X and Y axes are selected the system produces a standard matrix with numbers representing the number of documents present that have an intersection of the two fields. The co-occurrency matrices can be exported to Excel for reporting the results to non-SciFinder users. The capacity to match CAS Registry Numbers to CAS indexing along with the ability to link back to the Registry record for compounds of interest provides powerful features for analysts interested in chemical compound information.

With VantagePoint, a matrix is created in almost the same way as a list was created. Once the matrix button is pushed the user is required to choose a field for each axis. If any grouping has been done, the user can decide if they want the raw results or the grouped results used for the matrix. Selecting a number within a matrix brings up the list of titles associated with the two attributes just as it did with list generation. The analyst can also choose to "flood" the list, which eliminates any number within the matrix below a threshold number. This allows the analyst to eliminate relationships that have a small number of documents supporting them and immediately draws the eye to only those portions of the matrix where major relationships are found.

In the Aureka Reporting Module co-occurrency matrices are constructed by once again taking advantage of Pivot Tables within Microsoft Excel. They share similar features to other types of matrices created by the other products but in addition can take advantage of "Round Trip Analytics", a term coined by the folks at Aurigin to characterize the systems ability to send data

back and forth between the major components seamlessly. An example of this feature might involve a search to create a group of patent records followed by the launch of the reporting module in order to produce a matrix that shows the various patent assignees by the priority filing year of their documents. Documents from a particular year and assignee can than be "round-tripped" back to the main Aureka online system where a citation analysis or ThemeScape map can be created from them.

Matrices created by PatLab II and the Wisdomain Analysis Module are constructed based on input from the user on which fields they would like to analyze. Once the matrix is created the user can export the data if they would like to manipulate it further with additional tools like Microsoft Excel.

Instead of creating a matrix, the ClearResearch tool relies on circle graphs [9]. Circle graphs share some characteristics with matrices but differ in a few key areas. Circle graphs do not use numbers to illustrate the number of supporting documents that define a relationship but instead use lines of increasing width and differing color. They also can incorporate more than two fields. Since the plot is on a circle and not a grid more than two types of data can be represented. Granted, if too many fields are represented the circle can grow cumbersome and the interrelationships between any two fields (since a line can only connect two items) can be difficult to pull out. Fortunately, ClearResearch also contains a "flooding" feature that eliminates lines drawn representing relationships supported by a small number of documents. A slide bar found on the right side of the graph easily controls the flooding. Double clicking a line between two items will produce the documents with those items in common while double clicking on one of the items listed around the circle will isolated that item and show the relationships it has with the other items around the circle. An additional nicety is the ability to identify the specific relationship between the items of interest within a document. A button allows the user to toggle between document proximity (where the two items are found anywhere together in the same document) and sentence proximity (where the two terms must be in the same sentence).

5. Clustering of structured (fielded) data

When data is tagged (such as what is found with XML) or contained within a field (an author field within a database for instance) it is referred to as being structured. The tagging of the data and the placement of it in a particular data field provides it with a structure not generated by the author when they were writing the document. Intellectually assigned classification systems (as described in the previous section), for instance,

produce a standardized code that can be used as a means of organizing documents that share a similar coding structure. Clustering is also sometimes referred to as bucketing where a relatively large number of documents are arranged in a systemic fashion. For instance, all the documents on quinolines may be grouped together while those on pyridines are collected in a separate area. The principal behind this type of analysis is based on the idea that documents, which share a high percentage of codes in common with one another, are likely to be similar.

The following tools can perform clustering of structured data: Technology Watch, VantagePoint, ClearResearch and OmniViz.

Data of this sort might be structured but it needs to go into a database before it can be manipulated further. All four tools start this process by employing parsing engines for extracting fielded data and placing the information in their own proprietary databases. File specific formatting describing what tags represent what data are created for each individual data source. For instance, in order to parse data from Derwent on STN the parsing engine needs to know which pieces of tagged data need to be extracted, which tags correspond to which type of data (TI for Titles for instance), if the data is multi-variant (more than one value in the field typically separated by a comma or semi-colon) and what if any spacers exist between the field code and the data. The software also needs to know how to distinguish one record in an online transcript from all the others since a single transcript may contain hundreds of records. Technology Watch uses a separate little program to pre-process data for incorporation into the main program while the other three tools have the parsing engine built into them.

To start a clustering experiment in Technology Watch the analyst selects which fields they want the clustering performed on. The user can also decide on the number of clusters they want the data divided into and the amount of time the system will take to perform the operation. The result is a collection of different sized bubbles some of which have lines of varying width and color between them. Without further processing the map at this point it can look quite confusing. The analyst needs to arrange the bubbles on the screen such that the lines representing shared relationships between clusters (this occurs when documents in two separate clusters share items in common but do not have a high enough overlap to push them all together into the same cluster) can be readily seen. This often requires looking for isolated bubbles and moving them off to the side and arranging the connecting bubbles. If a classification code was used as the means for clustering documents then the title of the bubble will contain the top three codes most common to all of the documents in the cluster. The documents within the cluster or the codes themselves

need to be examined and a translated title provided to identify the contents of the cluster. Finally, the analyst can move the outlier bubbles closer to other related ones and color the bubbles so relationships they want to call out but are not represented by connection lines can be observed.

Clustering in VantagePoint is not clustering per se (which normally involved the use of an algorithm such as K-Means or Hierarchical) but instead is referred to as Factor Mapping and involves principal components decomposition. Based on some reasonably daunting statistics, factor mapping in VantagePoint is easily accomplished by selecting the field of interest and the number of terms to be used in the analysis (the top 15 for instance). The software produces a bubble diagram similar to the one found in Technology Watch but without the ability to color and re-title the bubbles.

To produce clusters in ClearResearch the user selects a taxonomy and a number of terms (or an entire taxonomy) as context for the clustering. The result is a collection of circle graphs each representing a different cluster. An example would have the analyst selecting assignee as the taxonomy and certain technology terms for providing the context. Each circle graph would show assignees listed around the outside based on their sharing of the technology terms they share in common. An overview can be seen by looking at which assignees were associated with one another in context and specific details on the relationship can be seen by drilling into the lines connecting the companies within the circle.

Once data is imported into OmniViz the analyst is asked to select which field should represent the title of the document (this is straight forward when discussing written documents but in other types of analyses a chemical compound and its related property data might be the data which is being clustered in which case the compound name would be the title) and which fields should be clustered on. When dealing with numerical fields (the system automatically recognizes when fields are numeric, text, categorical, etc.) the tool allows the user to apply powerful statistical measures such as normalization to the data. There are also mechanisms for dealing with missing data, which can adversely effect clustering. The tool also allows the number of clusters and the type of clustering algorithm to be selected. Once the analysis is complete OmniViz will offer to open a visualization of the cluster. The Galaxy view is the default choice and will provide a two-dimensional image of the documents (or other data elements when dealing with compounds) clustered around a centroid (a small circle which represents the center of the cluster which at times can be quite spread out depending on the algorithm used). Clicking on a centroid and opening the document viewer will provide a list of the documents within the cluster.

6. Clustering of unstructured (text) data

Unstructured text is defined as text that has not been indexed or segmented into individual data fields. The only structure contained within the document is the structure that was implied by the author when they put words together into sentences, sentences into paragraphs, and so on. This type of data represents a special challenge for analysts since patent authors are their own lexicographers and can define concepts within the patent in nearly anyway they see fit as long as it is explained within the disclosure for the examiner looking at the document. Unstructured data can also be challenging since without tagging it is difficult to put the text into the proper context. Without further processing the software tools cannot tell which words are for instance, an individual's name or potentially an organization. The raw text needs to be processed to identify concepts and phrases contained within and put them into the proper context. As with the clustering of structured data, text concepts instead of codes can them be used to group documents that share a high degree of overlap.

Clustering of unstructured data can be accomplished using: Aureka ThemeScape, VantagePoint, ClearResearch, OmniViz, Vivisimo and Delphion Text Clustering.

Where the tools for clustering tagged or structured data start by parsing the fielded data into a database the systems for clustering unstructured text begin by identifying relevant terms within a document. This process is referred to as term extraction and it begins with a method called tokenization. A string of characters separated by a space, dash or other type of punctuation is called a token. Tokens are essentially words when an analyst is working with text. The software analyzes the document and identifies all of the tokens within it. With full-text documents the number of tokens is enormous so the next step in the process involves removing stop words. These are non-content bearing terms such as "the", "be", and "a" which do not provide any information on the content of the document. Once stop words are removed, combing words that differ only in their suffixes can further winnow the list of tokens, a process called stemming. Finally, most of the tools listed above employ an algorithm named term frequency inverted document frequency (TFIDF) [12] to produce a final list of terms/concepts/tokens for clustering. The TFIDF algorithm looks at the number of times a token is represented in a document and compares that to the number of documents it is contained in. Tokens at either extreme (which imply that the term is either used infrequently in a small number of documents or very frequently in a large number) are eliminated for consideration when the clustering begins. The theory is that these terms will not be useful for identifying differences

and similarities between documents allowing them to be placed accurately into clusters.

ThemeScape and OmniViz perform in similar ways with regards to term extraction and clustering of documents based on shared concepts within them. A useful feature that both tools employ is the ability for the analyst to select additional stop words and re-process their map after having had an initial look at the results. The term extraction process can create meaningful concepts that provide useful clusters but the process is improved dramatically when the domain expertise of the analyst can be incorporated by the selection of additional, subject and analysis specific non-content bearing terms. A comparison of maps created before and after additional stop words have been added show significant movement in documents within clusters.

When using VantagePoint a tag must be used to identify a body of text to be analyzed within a document but otherwise the system uses much of the term extraction process as described above. Once the text has been processed by the system it can be used to create lists, co-occurrency or factor maps just as it would with other forms of fielded data. Since most users do not use all of the text found within a document for doing analysis in VantagePoint the extensive use of stop words is not necessary.

Vivisimo does not create bubbles or a map but instead organizes documents by using a hierarchical folder structure. The algorithm used to cluster the documents is proprietary so interested users should speak to the vendor for additional information but in practice the system was able to classify documents favorably [10] compared to manual classification by company scientists. Vivisimo also differs from some of the other systems in that documents can reside in more than one folder. Most of the other tools try to find the single best cluster for a document to reside in. Vivisimo allows documents containing multiple topics to reside in multiple folders.

Delphion Document Clustering creates 25 clusters each labeled with terms from the documents contained within them. There are no visualization bells and whistles with this tool and the end result is just the list of clusters that can be clicked on to reveal their contents. The system is a little of a black box however since the user cannot control any of the options for clustering. What you see is what you get.

Of the tools listed only the ones from ClearForest extract unstructured text from documents and attempt to provide context by placing the terms in a taxonomy with other like concepts. This process goes by a few different names and is often called information extraction or smart tagging. It is beyond the scope of this article to explore the information extraction process in detail but it is an emerging area within the field of text mining. A number of new systems are being developed for creating automated taxonomies from documents and placing documents in clusters based on information in context. As for clustering, once the terms are extracted and organized into taxonomies they can be analyzed just like fielded data.

7. Mapping document clusters

Clustering documents can help organize like documents together into groups but it does not necessarily provide an overview of how the clusters relate to one another. Document clusters are arranged in two-dimensional space creating a map. Collections of documents, which share elements in common, are placed closer together geographically while collections with less similarity are placed further away. Mapping clustered documents together can create a powerful visualization that can be used to show, which subjects have more in common with one another, and show documents that may not fit well into a single cluster but have elements of several clusters. The map creates a visual representation of a document collection from a high-level view.

Tools for mapping document clusters include: Aureka ThemeScape, Technology Watch, ClearResearch, VantagePoint and OmniViz.

Generally speaking the user does not have to perform any additional steps beyond what they would normally do to produce documents clusters since the tools are configured to produce a two or three-dimensional (in the case of ThemeScape which uses document density to represent a third dimension on the Z-axis) map as the output of the clustering exercise. ThemeScape, OmniViz and VantagePoint automatically arrange the map so that clusters that have some inter-relationship to one another are placed closer together. The analyst, as discussed previously, conducts the arranging of clusters manually in Technology Watch. The ClearResearch tool presents a map of clustered data but inter-relationships are not implied by how the clusters are located next to one another.

8. Adding temporal component to cluster map

Once a cluster map is created additional intelligence can be gleaned from it by querying against it in different ways. One of the most useful queries is a temporal one that allows a time dimension to be called out on the map. The eye is drawn to this additional dimension usually by means of alternate colors. When different colors are used for varying years a user can follow the progression of a subject as it develops or evolves across the map with different colors.

The addition of a temporal component can be accomplished using: Aureka ThemeScape, ClearResearch and OmniViz.

ThemeScape and OmniViz (remember their shared ancestry with Battelle) each deal with temporal data by means of time slices. Time slices are simply groups that happen to represent a series of dates. Both tools allow the slices to be displayed in different colors over the top of the existing visualization (ThemeScape for Theme-Scape and Galaxy or Theme View for OmniViz) so color can be used to track the progression of an item over time. In ThemeScape for example if the titles of document clusters are represented by a subject heading and time slices are created for three year periods running from 1990 to the present than it would be possible to see which subjects enjoyed more rigorous activity (represented by the number of colored dots) as time progressed. A similar form of analysis can be conducted in OmniViz.

The temporal dimension is represented differently in ClearResearch. Instead of using groups and dots to call out dates a slide bar is added to the bottom of the window where a circle graph or clusters of circle graphs are represented. As the slide bar is moved along the temporal dimension elements in the circle graphs that no longer fit within the time range disappear from view. This allows the analyst to quickly see changes in time by sliding the date bar from side to side and watching the effect on the graphs.

9. Citation analysis

When patent documents are examined relevant prior art is mentioned on the search report or on the front page of the documents. These are referred to as patent citations. The number of citations can be counted or followed as they link documents together. By looking at how documents link together based on their citation history it is possible to see how early technology may have evolved over time as they spawned improvements, new technologies and patents. Hyperbolic trees are sometimes used to show relationships between patents that cite one another. Total citation counts can be used to discover potentially pivotal documents or ones that may make good licensing candidates based on the interest they have generated.

Tools for performing citation analysis include: M-CAM, Aureka Citation Trees, Delphion Citation Link, Metrics Group Citation Bridge and the Wisdomain Citation Module.

Instead of spending a lot of time discussing the different types of patent citation analysis (an exercise that has been done previously in some of the included references) this article will briefly examine the citation analysis features of each of the specified tools.

The citation analysis within the M-CAM tool takes place in the background and is combined with semantic analysis to identify relevant pieces of prior art or newly published documents that could potentially be infringing on the document of interest.

In the Aureka Online System users can prepare citation trees as well as a large variety of backward and forward citation reports. Within the citation tree the analyst can change the node titles, color the nodes by year or assignee and open several trees at once. An analyst can also specify how many generations backwards or forward they want the tree to occupy. Within the reporting module there are a number of pre-configured citation reports including total citation count, citation count by year and a citation data based Pivot Table.

Delphion's Citation Link also employs a tree structure but it does not expand back and forth as the Aureka hyperbolic tree does. Otherwise the two systems share a number of the same features and visually allows the analyst to call out assignees by color and provides instant access to the citing or cited documents within the tree.

The Metrics Group's citation bridge is a free service (after a no obligation registration at the patentcitations.com web site), which allows the user to get a text report of a single document going one level forward or backward in history. More full featured citation reports and analysis can be ordered at the patentcitations.com web site.

The Wisdomain Citation Module looks remarkably like Delphion's Citation Link raising the question as to whether, like PatLab II, Delphion has gone back to Wisdomain for a portion of their analysis capabilities. Wisdomain adds a unique feature to their citation tree by allowing a user to filter out assignees that are not of interest and narrow the scope of the tree by restricting it to only International Patent Classification codes of interest. The company also claims to allow collateral analysis where pending documents which would not normally be seen in a backward or forward analysis can be identified.

10. Subject/action/object functions

SAOs are parts of language that are used to describe the teachings that the author wants to share. Key SAOs encapsulate the technical learnings contained in a document. SAOs can also be described as problems and solutions. For example, the statement "soap cleans hands" contains a subject (soap), an action (cleans) and an object (hands) and while being simplistic does demonstrate how SAO functions generally present a teaching. By identifying SAOs the teachings of a document can be isolated and examined from the rest of the document. When these functions are accumulated a knowledge base is created of the corresponding collection of documents.

Table 3
Matrix of patinformatics tasks and tools

Tool	List Cleanup and Grouping of Concepts	List Generation	Co-occurrency Matrices and Circle Graphs	Clustering of Structured (Fielded) Data	Clustering of Unstructured (Text) Data	Mapping Document Clusters	Adding Temporal Component to Cluster Map	Citation Analysis	SAO Functions
Aureka ThemeScape	⊕				⊕	⊕	⊕		
ClearResearch	⊕	⊕	⊕	⊕	⊕	⊕	⊕		
OmniViz	⊕			⊕	⊕	⊕	⊕		
Vivisimo					⊕				
Delphion Text Clustering					⊕				
VantagePoint	⊕	⊕	⊕	⊕	⊕	⊕			
Technology Watch		⊕		⊕		⊕			
M-CAM								⊕	
Aureka Citation Trees								⊕	
Delphion Citation Link								⊕	
Metrics Group Citation Bridge								⊕	
Wisdomain Citation Module								⊕	
Knowledgist		⊕							⊕
MS Excel	⊕	⊕							
Aureka Reporting		⊕	⊕						
Wisdomain Analysis Module		⊕	⊕						
Delphion PatLab II		⊕	⊕						
SciFinder		⊕							
SciFinder Panorama			⊕						

This particular manifestation of building a knowledge base is the province of the Knowledgist tool. The user identifies a single or a collection of text documents and directs the software to begin the extraction process. When completed Knowledgist presents the user with a list of the extracted problems and solutions in a frame on the left hand side of the window. By clicking on the plus sign to the left of each problem and solution the program displays the ways that a particular problem is solved within the document or collection. Clicking on the desired solution method displays the portion of the document from which the solution came on the right hand side of the window. The SAO is highlighted and a link to the full document is located below the referenced portion. With long knowledge bases containing many different problems and solutions a search box is located towards the top of the screen so the user can narrow the collection of SAOs down to a more reasonably reviewed list.

11. Conclusions

The number, depth and breath of software tools for conducting patinformatics exercises have grown extensively over the past 5–10 years. This article has attempted to summarize some of the tasks associated with doing patent analysis and maps them to the tools that have been developed (please see Table 3 for a matrix of tools to tasks). It is hoped that this analysis will assist individuals in discovering the various software options available to them and, along with other information they already have or decide to seek out, pick the right tools for conducting each specific task. This is critical since as the old saying goes, "To a man with a hammer, everything looks like a nail." Truly valuable patinformatics work can be achieved when practitioners avoid the "one size, fits all" approach to selecting software and rely on having a clear understanding of the requirements for the job at hand and use the best available tool and data for conducting the research.

References

[1] Trippe AJ. Patinformatics: Identifying Haystacks from Space. Searcher 2002;10(9):28.

[2] Ernst H. Patent applications and subsequent changes of performance: evidence from time-series cross-section analysis on the firm level. Res Policy 2001;30:143.

[3] Mogee M, Breitzman A. The many applications of patent analysis. J Informat Sci 2002;28(3):187.

[4] Breitzman A, Thomas P. Using patent citation analysis to target/value M&A candidates. Res Technol Mgmt 2002;45(5): 28.

[5] Adams S. Pharmacia Corp.: Analysis of Patenting 1998–2002. Expert Opin Ther Patents 2002;13(2):223.

[6] Awaya, Kohei. Analysis method for patent documents utilizing references. In: Proceedings of the IASTED International Conference: Applied Informatics. 2002. p. 15.

[7] Trippe AJ. A comparison of ideologies: intellectually assigned co-coding clustering vs. themescape automatic thememaic mapping. In: Proceedings of the 2001 International Chemical Information Conference. 2001. p. 61.

[8] Trippe AJ. Visualization of chemical patents: source titles and abstracts vs. enhanced titles and abstracts. In: 223rd National American Chemical Society Meeting, Chemical Information Division, Orlando, FL, Spring 2002.

[9] Aumann Y, Feldman R, Yehuda YB, Landau D, Liphstat O, Schler Y. Circle graphs: new visualization tools for text-mining. In: Proceedings of the Principles of Data Mining and Knowledge Discovering Conference. 1999. p. 277.

[10] Trippe AJ, personal communication with consulting client.

[11] http://bioinformatics.org/faq/#definitions.

[12] http://classes.seattleu.edu/computer_science/csse470/Madani/ABCs. html.

Anthony Trippe currently holds the position of Senior Staff Investigator, Intellectual Property at Vertex Pharmaceuticals. He is responsible for designing and implementing patent intelligence and mapping activities at Vertex and for assisting with the leveraging of IP within and external to the company. Previously, Trippe was Practice Director, Intellectual Property Consulting for Aurigin Systems Inc. and was Technical Intelligence Manager for the Procter and Gamble Co.