A METHOD OF VISUALIZING CLUSTERS OF LARGE COLLECTIONS OF TEXT DOCUMENTS

Field of the invention

This invention relates to a method of visualizing clusters of large collections of text documents and refers particularly, though not exclusively, to a map facility for the user to not only browse a text collection in an intuitive and meaningful manner but also to navigate and discover useful trends from the document collection.

Background to invention

Text mining is the next wave of research after data mining. It is supported by a number of technologies such as categorization, clustering, summarization, visualization, text understanding, information extraction, and so forth. There are a small number of existing techniques to visualize clusters of documents. Most are traditional, and rather preliminary. They are facing usability problems due to the usability to meaningfully handle and display large number of clusters.

US Patent 5581797 discloses an interactive method and apparatus for displaying structure, statistics, and characteristics of a large software systems, i.e. of more than one million lines of code. This method and apparatus displays important structure and statistics in a manner where the entire software system can be visualized as an entity. The visualization technique is to represent similar subdivisions of the code in similar geometric shapes having a substantially equal reference frames, such as substantially constant row heights, so relative sizes of the subdivisions can be understood.

In addition to the sizes, different shading and colouring modes can be used to display changes, software errors and software corrections. Further, if data of the various releases of a large software system are available, the evolution of the system through its releases can be animated to provide a greater understanding of the history of the large software system.

This is used in the area of visualising the status of large software systems. The present invention is directed at text mining.

Furthermore, the number of rows displayed is at least three, and each row has about the same height. With the present invention the number of rows will be limited to a manageable number, and the row heights vary depending on the area occupied by the clusters. The method used to generate the interface is thus very different.

Also, the document discloses the use of different shading and coloring modes to display changes, software errors and software corrections. The present invention proposes to use meaningful symbols attached to each area to indicate changes in the underlying text collection. Colours are used for selection of clusters, zooming into sub-clusters, as well as personalization. Finally, the specification of the prior art discloses the use of histograms to represent information within each rectangle. The present invention makes use of meaningful symbols.

US Patent 5442778 is in relation to Scatter-Gather, which is the use of a computer based document browsing method which operates in time proportional to the number of documents in a target corpus. The Scatter-Gather method includes: preparing an initial ordering of the corpus using, for example, an off-line computational method; determining a summary of the initial ordering of the corpus for interactive utility; and providing a further ordering of the corpus

using, for example, an on-line non-deterministic method. The step of an off-line preparation of an initial ordering of corpus is non-time-dependent, thus an accurate initial ordering is prepared. The step of determining a summary includes determining a summary of presentation to a user without scrolling on a CRT. The step of providing a further ordering includes truncated group average agglomerate clustering, merging disjointed document sets, center finding, assign-to-nearest and other refinement methods.

Implementations of the above that appear in publications show a listbox interface to correspond to each cluster.

However, with the present invention the graphical display is different, and each cluster is represented by keywords and/or a summary, consisting of a list of summary points, of each cluster.

The disclosure of the specification of US Patent 5574837 reveals a cluster interface which is generated to represent similarities of semantics between segments of code with respect to both the physical construction of the code and the underlying operations performed by the code. The generated interface represents one or more code segments. Code segments to be analyzed are received by a computer system. Statistical internal information is extracted from each code segment. An external metric is generated which is based on the extracted statistical information. An interface display is created from the external metric which represents similarity relationship between the imputed code segments based on the extracted statistical information. As such, it makes use of a mixture of bars and trees to display the cluster interface, and is used in a different context from the present invention.

The Tree-Map technique, a space-filling approach, is applied to hierarchical databases. It is useful for displaying large hierarchical

structures in a limited space. The derivation of the visual interface is from tree diagrams to Venn diagrams, to a nested/un-nested tree map. With the present invention, tree diagrams/venn diagrams are not used as with them clusters of documents are not necessarily hierarchical. With the prior art the organization of the rectangles corresponds to the hierarchy of the tree, whereas with the present invention clusters with repeated keywords are preferably together.

The Kohonen's Self-Organizing Maps (SOM) is another known space-filling method of representing clusters of text documents. The regions of the two dimensional map vary in size and shape according to how frequently documents assigned to a corresponding theme occur within the collection. Regions are characterized by single words or phrases and similar clusters are placed adjacent to each other. A cursor moved over a document region causes the titles of the documents most strongly associated with the region to be highlighted. Documents can be associated with more than one region. Other SOM-based implementations include cMap, ETMap, Graphical Table of Contents.

The disadvantage of this technique is that there is no systematic way of proceeding through the clusters. Reading the labels in each other cluster is quite haphazard. The display becomes overwhelming, or cluttered, when the number of clusters is large, particularly with very large collections. This is because every cluster is displayed. Also, the size of the polygons does not accurately represent the actual size of the cluster, which can be misleading to users. Context is also not maintained when a user zooms into a more detail view of the display.

With the present invention, clusters may be represented in a systematic manner such, for example, left-to-right, top-to-bottom, to

facilitate reading and searching. Repeated keywords are preferably grouped together to reduce the users' mental workload. It provides a way of limiting the number of clusters displayed to reduce clutter due to many small clusters. The size of each cluster is represented accurately within the display. Context is maintained when more detail of a cluster is obtained.

Themescape is another method to display cluster information on a map-like interface. It uses a geographical map-like interface to display clusters. Mountain peaks denote clusters of interest but there is no accurately defined boundary for each cluster. Themescape also allows the user to vary the display of more than one keyword per cluster. This has the same problem as Kohonen's SOM. There is no systematic way of going from one peak (cluster) to the next. Repeated keywords can be quite far apart on the display. The display becomes overwhelming, or cluttered, when the number of clusters is large, particularly with very large collections, as every cluster is displayed. In addition, the boundaries for each cluster are not accurately defined which means that the size of each cluster is not obvious. Context is also not kept when a user zooms into a more detailed view of the display.

SmartMoney Map is a two-dimensional visualization algorithm for presenting detailed information on hundreds of items while emphasizing overall patterns in data. It is a modification of the Tree-Map display and makes use of both hierarchy and similarity information. It is implemented for stock data whereas the present invention is for document collections; and makes use of rectangle of varying sizes to group together similar stocks whereas with the present invention clusters are organized with similar keywords by rows/columns.

Wical 5, 918,236 describes a display in the form of a tree representation. That is, from the initial list of categories at the top level, the user can decide to show the next level of detail until the level of the individual document is reached. It is only at this level that it shows a summary of the document. Unlike the present invention, it is for pre-defined/supervised grouping categories rather than undefined/unsupervised grouping clusters.

Microsoft WO99/67727 describes the display in the form of a tree representation or circular format. It is also for categories and clusters.

Conrad 6,028,605 describes a system that extracts metadata for a set of documents. The results can be grouped in the form of categories. The display is mainly in the form of lists.

Jorna 6,029,172 describes a method of browsing a hierarchically classified database by interactively displaying a relevant portion of the classification scheme of the database as category names and sub-category names. It displays top level categories in bigger fonts, surrounded by its sub-categories in smaller forms.

Chen 6,009,442 describes a computer based electronic document and/or paper-based document management application program that automatically imports, indexes, categorizes, stores, searches, retrieves, manipulates and archives electronic documents. It makes use of conventional list boxes.

Martz 5,986,673 discloses cluster display for relational data, which is unlike the present invention which is for text documents. The cluster display is in form of hierarchical dendrograms.

Prager 5,943,670 describes a system and method to enable an object (e.g. a document) to be classified in more than one category.

Krellenstein 5,924,090 describes a method and apparatus to search a database of records and then assign the search result into a category. GUI is a list of folders.

Zhao 5,920,864 describes a system of a multi-level dynamic categorizer for user navigation.

Miyasaka 5,918,236 describes an information search and collection method. It is mainly for categories.

Kleinberg 5,884,305 generally relates to the field of data mining, not text mining. It relates to the determination of "categorical clusters" from databases.

Egger 5,832,494 describes a computer research tool for indexing, searching, and displaying data as applied to the legal domain. The GUI is a two or three dimension display showing clusters of searched documents along a time bar.

Duke-Moran 5,819,259 describes an expert system which is capable of searching media and text information such as newspaper items and placing them into predetermined categories. Output of categories is in the form of lists.

Vaithyanathan 5,819,258 describes a top-down clustering method, not visualization.

Shakib 5,752,025 describes a method for creating and displaying a categorization table. Display is in the form of collapsible/expandable hierarchical folders.

Homma 5,179,643 describes cluster analysis as applied to merchandise information. Displays of clusters are like scatter plots or two dimension graphical plots.

Summary of the invention

In this invention, we disclose a method to generate a display that aims to solve the above mentioned problems. The display is a space-filling visualization technique that has a simple interface, manages large number of clusters more effectively, displays meaningful information, and facilitates browsing. In addition, it extends the interface to provide users with the means to discover trends from the clusters through the use of user-defined meaningful symbols; and support for personalization features.

With the present invention these is provided an organized and systematic interface to look at the clustered output of a clustering algorithm. At a glance, it may provide an overview of the entire collection, as well as their relative cluster sizes. It preferably includes the ability to look at a higher level of information regarding a cluster, before proceeding with the re-clustering of clusters having a large number of documents. The context of view is preferably always kept, so the user should never get lost. It may also allow knowledge discovery of clusters of documents through use of special symbols, and enables personalization in the tracking of clusters, and defined user profiles.

Furthermore, the display of large number of clusters produced by clustering algorithms may be managed more effectively by grouping clusters with repeated keywords, or other characteristics, together so as to reduce the users mental workload of grouping it themselves.

Clusters may be arranged in a systematic manner to facilitate reading, and the handling of more clusters. This may be left-to-right, top-to-bottom, if arranged by row, and top-to-bottom, left to right if arranged by column.

The number of rows/columns may be fixed to display the clusters at a number that the user considers as being manageable conceptually.

By using such a display, discovery of a document collection may be enhanced by changes over time, through the use of meaningful symbols such as, for example, time-dependent symbols and cluster characteristic symbols. These symbols may be customisable by the user.

The display may be personalizable to track changes in any particular cluster or clusters of interest.

Description of drawings

In order that the invention may be fully understood and readily put into practical effect, there shall now be described by way of non-limitative example only a preferred embodiment of the present invention, the description being with reference to the accompanying illustrative drawings in which:

Figure 1 shows an example of the overall system architecture;

Figure 2 shows an initial view of clusters;

Figure 3 shows a "view summary" of clusters;

Figure 4 shows a view of a subcluster;

Figure 5 is an overview of the method used to generate the map of Figure 2;

Figure 6 shows a sample map file and how it corresponds to a typical prior art map;

Figure 7 is a flowchart for drawing a map display;

Figure 8 is a flowchart for displaying subclusters;

Figure 9 is a flowchart for displaying time–dependent symbols;

Figure 10 is a flowchart for displaying cluster characteristic symbols;

Figure 11 is a flowchart for tracking cluster; and

Figure 12 is a flowchart for defining user profiles.

Description of preferred embodiment

The present invention is built on the assumption that there is a clustering algorithm that clusters a large document collection into clusters and provides a summary of each cluster. The present invention makes use of the output of these algorithms.

Figure 1 shows an example of the overall system architecture. The present invention covers the "Display" box. The pre-processing stage takes care of all the computation needed to create the visual display.

The display step is an initial output 10 as shown in Figure 2, which shows a first polygonal area, in this case a rectangle, having a number of contained polygonal areas, again rectangles, in this case, arranged in a plurality of rows 12. The display area is subdivided such that each area 14 represents a cluster. Each area 14 may be rectangular, as shown, or other polygonal shape such as, for example, triangular, hexagonal, and so forth. Within each rectangle 14, it displays the top three keywords 16 of each cluster. The number of key words may be predetermined. This may be by the user, if desired. Furthermore, the number of key words may be varied. Clusters (e.g. 18, 20) with repeated top three keywords are grouped close together. The arrangement is from left-to-right and top-to-bottom. The size of each rectangle 14 corresponds to the relative size of each cluster with respect to the whole collection. The actual number of documents in each cluster is also displayed in

brackets within each rectangle 14. When the cursor is placed over a rectangle 14, a popup display shows the top number of keywords such as, for example, five represented in that cluster, as well as the document size. For rectangles 14 that are too small to display the necessary words, the user can move the cursor over the selected area to activate the popup display.

The display 10 is limited to a manageable number of rows 12 such as, for example, seven thus making the viewing of clusters manageable. The last row 22 gathers all those clusters with a predetermined, relatively small number of documents such as, for example, three, into a single cluster. The predetermined number may be expressed as a unit (e.g. three) or as a percentage of the number of documents in the collection (e.g. 1%). If the number of clusters in the last row is excessive, or the number of rectangles 14 excessive, the predetermined number can be changed and reclustering take place to give a result which is easily used.

At a glance, the user can see the relative sizes of each cluster in a document collection, that is, an overview of the text collection. With this information, the user may decide to or not to read those clusters with very small areas.

Clusters with repeated keywords are grouped together to reduce the users' mental workload of grouping it themselves. In the example display of Figure 2, documents related to different car companies are placed next to each other.

Clusters are displayed in a systematic left-to-right, top-to-bottom manner to facilitate reading. This enhances a user ability to read and thus handle more clusters. Existing clustering displays like SOM/Themescape become very hard to read when the number of clusters is large, as every cluster is displayed.

By limiting the number of rows 12 to be displayed, and grouping smaller clusters together into a cluster at the bottom 12, clutter on the display 10 is reduced.

The View Summary (Figure 3) of a cluster can be achieved by selecting the desired cluster or clusters. To activate, a pop-up menu or other form of list is displayed. When View Summary option is chosen, a window 24 opens to display the summaries related to the selected cluster, and the map 10 of Figure 2 is resized to another part of the screen, but is not overlapped with the new window 24 where it can be seen that the original screen 10 has been reduced in size to right half of the display. Each summary in window 24 is made up of a number of themes 26 so as to advise the user what the cluster is about. It is also possible to view the details of each of the themes by selecting the View Detailed Summary menu option.

This summary option gives the user more information about the cluster of documents before the user decides to look at the actual documents within the cluster. This is particularly useful if the collection is large.

By making use of a separate window 24 to display the summaries, the visual display 10 is not cluttered. This allows the user to make use of the visual display to 'jump' quickly to any cluster of interest, as opposed to scrolling down as in Scatter/Gather displays.

It is possible to view a subcluster if a cluster represented is too large, and it would be more practical recluster only that cluster to be able to see a finer granularity of subclusters within it. To do so, the user can select the desired cluster and select the View Subcluster menu option. The selected cluster is expanded to show subclusters 28 as shown in Figure 4. The rest of the clusters 30 are collapsed

and pushed to the edge of the display (in this case the right and bottom) and represented in a different or lighter colour thus giving a "fish-eye lens" effect. Up to three levels of viewing sub-clusters may be allowed so that there is enough display area to represent the subclusters. However, higher levels of viewing may be used, if desired.

The user may not want to see all the documents in a cluster if it has a large number of documents. By reclustering, it enables the user to decide which subcluster is more interesting to look into, thus reducing the time taken for a user to retrieve the necessary documents. Whenever the user zooms into a subcluster, they should not get lost as the higher level clusters 28 are collapsed to the edge of the display, thus maintaining the context of the display.

The display may be extended to allow for discovery of new information, especially if the underlying text collection is constantly changing. This is done by applying different types of symbols to each of the clusters.

For example, in Figure 3, there is displayed three legends: a seed 32 to represent a new cluster (day), a seedling 34 to show a cluster that has been appearing for a short time (week), and a tree 36 to represent a cluster that has been appearing for a long time (month). However, any form or number of symbols may be used, and any lengths of time may be used. Alternative intuitive symbols could be used or chosen by the user. The duration of day/week/month is user customisable as well. These time-dependent legends may only appear the second time the user views the collection. This is so that the initial display of Figure 2 remains uncluttered by symbols.

In a News Watch application, for example, the present invention display enables the user to spot new topics that are

appearing by looking at the seeds 32, or a topic of increasing attention by looking at those rectangles with seedlings 34. The user may not want to look at those that are marked as trees 36 as they are probably old topics.

Cluster characteristic symbols make use of information about the clusters in a document to help the user decide if a cluster is worth investigating.

For example, in the News Watch application, each rectangular area 14 could use a special symbol to show if a particular cluster is made up of documents that come from the one source or from a variety of sources. This may help a user to assess the reliability of the news mentioned in that cluster; or to decide if all the documents in that particular cluster need to be read, in which case the user can select one document to read. Also, if it is reported by many sources, it could indicate a hot topic, as opposed to being reported by only a single source.

The time dependent symbols enable a user to view document collections that change, e.g. news. To date, existing clustering visualization techniques offer only a snapshot of a document collection but no information of how it differs from a previous collection. Cluster characteristics symbols help a user to discover more information about a particular cluster, as well as enabling the user to decide if a cluster is worth investigating. A combination of time-dependent and cluster characteristics gives important information about a cluster. For example, a cluster with symbols 'seed' and 'variety of sources' indicates a new, hot topic.

Personalization features are options to customise the present invention to suit the personal needs of the end user.

To track a particular cluster, the user first selects a cluster. Next, the user can select a menu option (e.g. Option->Track). A dialog box pops up to allow the user to select a colour or other indicia to associate with this cluster. In future, whenever the particular cluster is changed (that is, documents within the cluster have increased/decreased), it will be highlighted in the chosen colour or indicia on the visual display.

The user can also define a profile of the keyword topics which they are interested in tracking, and it will allow system to highlight those clusters in a colour of the user's choice.

The flowcharts of Figures 5 to 7 show a possible implementation method of generating the maps discussed above.

Figures 5 to 7 shows how the map of Figure 2 is generated. Essentially, use is made of a standard Kohonen algorithm to generate a placing for each of the cluster on the display within a fixed number of rows. The input to the Kohonen algorithm is a matrix of clusters versus distinct words, where each row is a cluster vector, and each column corresponds to a distinct word. A cluster vector contains a "1" in a given column if the corresponding distinct word occurs in the keywords for that cluster, and a "0" otherwise. Following Kohonen's algorithm:

- each feature corresponds to a distinct word;
- each cluster is an input vector;
- after training, each cluster is mapped onto a grid node as shown in Figure 6 (right).

This output is converted to a map file as shown in Figure 6 (left) where the single numeral 6 indicates the number of rows. This is used to generate the display of Figure 2.

Figure 8 shows how the sub-cluster of Figure 4 is generated. Standard clustering techniques are used, with the exception that area - division is used to collapse the display of the clusters.

Figure 9 shows how time-dependent symbols are displayed. The main aspect to this is in the largest box, where the date difference is determined, after determining if the top predetermined number of keywords set appeared before in any of the clusters, and the relevant symbol is then assigned and the map displayed.

In Figure 10, there is shown the techniques used for displaying cluster characteristic symbols. Again, the largest box contains the description on what is done. As can be seen, single or multiple arrow symbols are used to indicate clusters with a corresponding number of unique sources of a document in the document list, prior to the map being displayed. In this way, a user can readily determine the relevance and/or importance of the document.

The tracking of a cluster is shown in Figure 11. After a new map of the current set of documents is generated, the previously stored cluster information, including keywords and marked clusters, is retrieved and the marked clusters compared with the new map. For each positive cluster, the new map is displayed with the positive clusters highlighted.

To define user profiles one follows the procedure of Figure 12. Like the tracking of a cluster, a new map is generated for the current set of documents. The previously stored user profile is retrieved. The user profile contains, for example, keywords, colours, and so forth. For each cluster in the new map, a check is conducted to determine if the user's preferred list of keywords appears in the cluster. For each positive response, the user's selected colour is assigned to that cluster, and the map displayed.

Whilst there has been described in the foregoing description preferred embodiments of the present invention, it will be understood by those skilled in the technology that many variations or modifications in technical details may be made without departing from the present invention.

The Claims

1. A graphical user interface to display clusters of a text collection on a display screen including a first two dimensional polygonal area of the display screen, and a plurality of contained two dimensional areas contained within the first two dimensional area, the plurality of contained two dimensional areas being arranged in a fixed number of user-defined rows, each row having varying heights depending on the number of said plurality of contained two-dimensional areas represented in each row.
2. A graphical user interface as claimed in claim 1, wherein the maximum number of the plurality of contained two dimensional areas corresponds to the number of clusters generated by a clustering algorithm.
3. A graphical user interface as claimed in claim 1, wherein each of the contained two dimensional areas represents a cluster and the size of the area of each of the contained two dimensional areas within each row corresponds to the relative size of its represented cluster with respect to the total number of documents in the text collection.
4. A graphical user interface as claimed in claim 1, wherein each said contained two dimensional areas has a description of the cluster; whereby each description contains a predetermined number of the most occurring keywords representative of that cluster, and the number of documents in that cluster.
5. A graphical user interface to display clusters of a text collection on a display screen including a first two dimensional polygonal area of the display screen, and a plurality of

contained two dimensional areas within the first two dimensional area, the number of the plurality of contained two dimensional areas corresponding to the number of clusters automatically generated by a clustering algorithm.

6. A graphical user interface to display clusters of a text collection on a display screen including a first two dimensional polygonal area of the display screen, and a plurality of contained two dimensional areas within the first two dimensional area, each of the contained two dimensional areas representing a cluster and the size of the area of each of the contained two dimensional areas corresponds to the relative size of its represented cluster with respect to the total number of documents in the collection.

7. A graphical user interface to display clusters of a text collection on a display screen including a first two dimensional polygonal area of the display screen, and a plurality of contained two dimensional areas within the first two dimensional area, each of the contained two dimensional areas having a description of the cluster, whereby each description contains a predetermined number of the most occurring keywords representative of that cluster and the number of documents in that cluster.

8. A graphical user interface as claimed in claim 6, wherein the total number of the plurality of contained two dimensional areas corresponds to the number of clusters automatically generated by a clustering algorithm.

9. A graphical user interface as claimed in claim 7, wherein the total number of the plurality of contained two dimensional

areas corresponds to the number of clusters automatically generated by a clustering algorithm.

10. A graphical user interface as claimed on claim 7, wherein each of the contained two dimensional areas represents a cluster and the size of the area of each of the contained two dimensional areas corresponds to the relative size of its represented cluster with respect to the total number of documents in the text collection.

11. A graphical user interface as claimed in claim 6, wherein the contained two dimensional areas are arranged in a plurality of rows.

12. A graphical user interface as claimed in claim 7, wherein the contained two dimensional areas are arranged in a plurality of rows.

13. A graphical user interface as claimed in claim 1, wherein the contained two dimensional areas are arranged such that those with the same overlapping keyword description will be placed adjacent to each other on the same row; or the next adjacent row if there are too many in the row.

14. A graphical user interface as claimed in claim 6, wherein the contained two dimensional areas are arranged such that those with the same overlapping keyword description will be placed adjacent to each other on the same row; or the next adjacent row if there are too many in the row.

15. A graphical user interface as claimed in claim 7, wherein the contained two dimensional areas are arranged such that those with the same overlapping keyword description will be placed adjacent row if there are too many in a row; or the next adjacent row if there are too many in the row.

16. A graphical user interface as claimed in claim 1, wherein the rows include a last row, the last row representing a group with less than a threshold number of documents.

17. A graphical user interface as claimed in claim 1, wherein the area of the first two dimensional polygonal area corresponds to the total number of documents in the text collection.

18. A graphical user interface as claimed in claim 1, wherein the first two dimensional polygonal area is responsive to a click-and drag operation of a mouse to give a new length and new width.

19. A graphical user interface as claimed in claim 18, wherein the plurality of contained two dimensional areas are adjusted in rows to the new length and new width of said first two dimensional area.

20. A graphical user interface as claimed in claim 17, wherein the threshold of number of documents is adjustable by user input.

21. A graphical user interface as claimed in claim 20, wherein the number of the plurality of rows is adjustable by user input.

22. A graphical user interface as claimed in claim 1, wherein the said contained two dimensional areas belong to the same class of polygon.

23. A graphical user interface as claimed in any one of claims 1 to 22, wherein the arrangement of the contained two-dimensional areas is by columns instead of by rows.

24. A graphical user interface as claimed in claim 1, including tracking means to track the changes in the size of each of the contained two dimensional areas.

25. A graphical user interface as claimed in claim 1, wherein define means are provided to enable keywords to be tracked to be defined.

26. A graphical user interface as claimed in claim 6, wherein define means are provided to enable keywords to be tracked to be defined.

27. A graphical user interface as claimed in claim 7, wherein define means are provided to enable keywords to be tracked to be defined.

28. A program storage device readable by a machine, including a program of executable instructions to display on a display screen a graphical user interface in accordance with any one or more of claims 1 to 27.

29. A method for obtaining textual information about at least one cluster of a plurality of clusters each represented in a plurality of contained two-dimensional areas within a first two dimensional polygonal area, including the steps of:

(a) selecting the at least cluster via selection of at least one of the contained two dimensional area;

(b) selecting a menu option to select desired details of the at least cluster;

(c) displaying a second two dimensional polygonal area in a first portion of the two dimensional polygonal area;

(d) reducing in area the contained two dimensional areas such that they occupy a second portion of the two dimensional polygonal area; and

(e) displaying requested details of the selected at least one cluster in the first portion.

30. A method as claimed in claim 29, wherein the requested details of the at least one cluster include one or more display items selected from the group including:

a summary of the set of documents represented by the at least one cluster,

a detailed summary of the summary, and

a list of titles of documents that appear in the at least one cluster,

31. A method as claimed in claim 29, wherein the contained two dimensional areas are within a region on at least one of the sides of the first two dimensional polygonal area.

32. A method as claimed in claim 31, wherein subdivisions of the cluster are displayed in the second portion.

33. A method as claimed in claim 32, wherein a row of unclustered documents does not appear in the second portion.

34. A method as claimed in claim 29, wherein at least one graphic symbol is displayed within at least one of the plurality of contained two dimensional areas to represent additional knowledge about a cluster.

35. A method as claimed in claim 34, wherein there are more than one of said graphic symbols within at least one of the contained dimensional areas.

36. A method as claimed in claim 34, wherein the graphic symbols are customised by a user.

37. A method as claimed in claim 29, including tracking the changes to the size of each of the contained two dimensional area.

38. A method as claimed in claim 29, including defining the keywords to be tracked.

39. A method of displaying clusters of a text collection on a display screen, the method including the steps of:

(a) generating a placing for each of the clusters within a number of rows of the display;

(b) creating a contained two dimensional area for each cluster by determining the relative size of its represented cluster with respect to the total number documents in the text collection;

(c) determining the height of all clusters in each of the rows;

(d) determining the number of clusters to be in each of the rows;

(e) determining the width of each cluster in each row; and

(f) creating the display.

40. A method as claimed in claim 39, where the number of contained two dimensional areas corresponds to the number of clusters generated by the clustering algorithm.

41. A method as claimed in claim 39, wherein the number of rows is fixed.

42. A method as claimed in claim 41, wherein the number of rows is fixed by a user.

43. A method as claimed in claim 39, including the step of including within each of the contained two dimensional areas a description of the cluster; each description containing a predetermined number of the most occurring keywords representative of that cluster.

44. A method as claimed in claim 43, wherein the description includes the number of documents in that cluster.

45. A method as claimed in claim 43, including the further steps of: determining those clusters with overlapping keyword descriptions; and arranging the resultant clusters in adjacent relationship.

46. A method as claimed in claim 39, including the additional steps of: determining all clusters with less than a threshold number of documents, and combining all such clusters in a single contained two dimensional area in a last of the number of rows.

47. A method as claimed in claim 39, wherein the display is contained within a first two dimensional polygonal area which corresponds to the total number of documents in the text collection.

48. A method as claimed in claim 46, wherein the threshold number is obtained by user input.

49. A method as claimed in claim 39, wherein the generating of the placing is by a clustering algorithm.

50. A method as claimed in claim 43, wherein the predetermined number is user selected.

51. A method as claimed in claim 43, wherein the predetermined number is variable.

52. A method as claimed in claim 39, wherein the generating of the placing is on a user-specified template.

53. A method as claimed in any one of claims 39 to 52, wherein the rows are vertical columns.