Title: METHOD AND APPARATUS FOR AUTOMATICALLY DETERMINING SALIENT FEATURES FOR OBJECT CLASSIFICATION

Abstract: A method and apparatus for automatically determining salient features (308) for object classification is provided. In accordance with one embodiment, one or more unique features are extracted from a first content group of objects to form a first feature list, and one or more unique features are extracted from a second anti-content group of objects to form a second feature list. A ranked list of features is then created by applying statistical differentiation between unique features of the first feature list and unique features of the second feature list. A set of salient features (308) is then identified from the resulting ranked list of features.
For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.
METHOD AND APPARATUS FOR AUTOMATICALLY DETERMINING
SALIENT FEATURES FOR OBJECT CLASSIFICATION

BACKGROUND OF THE INVENTION

1. Field of the Invention

The invention relates to the field of data processing. More specifically, the invention relates to the automatic selection of features of objects for use in classifying the objects into groups.

2. Background Information

The World Wide Web provides an important information resource, with estimates of billions of pages of information available for online viewing and downloading. In order to make efficient use of this information, however, a sensible method for navigating this huge expanse of data is necessary.

In the early days of Internet surfing, two basic methods were developed for assisting in Web searches. In the first approach, an indexed database is created based upon the contents of Web pages gathered by automated search engines which "crawl" the web looking for new and unique pages. This database can then be searched using various query techniques, and often ranked on the basis of similarity to the form of the query. In the second approach, Web pages are grouped into a
categorical hierarchy, typically presented in a tree form. The user then makes a series of selections while descending the hierarchy, with two or more choices at each level representing salient differences between the sub trees below the decision point, ultimately reaching leaf nodes which contain pages of text and/or multimedia content.

For example, Figure 1 illustrates an exemplary prior art subject hierarchy 102 in which multiple decision nodes (hereinafter "nodes") 130-136 are hierarchically arranged into multiple parent and/or child nodes, each of which are associated with a unique subject category. For example, node 130 is a parent node to nodes 131 and 132, while nodes 131 and 132 are child nodes to node 130. Because nodes 131 and 132 are both child nodes of the same node (e.g. node 130), nodes 131 and 132 are said to be siblings of one another. Additional sibling pairs in subject hierarchy 102 include nodes 133 and 134, as well as nodes 135 and 136. It can be seen from Figure 1 that node 130 forms a first level 137 of subject hierarchy 102, while nodes 131-132 form a second level 138 of subject hierarchy 102, and nodes 133-136 form a third level 139 of subject hierarchy 102. Additionally, node 130 is referred to as a root node of subject hierarchy 102 in that it is not a child of any other node.

The process of creating a hierarchical categorization for Web pages presents multiple challenges. First, the nature of the hierarchy must be defined. This is typically done manually by experts within a particular subject area, in a manner similar to the creating of categories in the Dewey Decimal System for libraries. These categories are then provided with descriptive labels so that users and
categorizers can make appropriate decisions while navigating the hierarchy. Content in the form of individual electronic documents for example are then placed into the categories by means of a manual search through the hierarchy.

In recent years attention has been directed towards automating the various stages of this process. Systems exist for the automatic categorization of documents from a corpus of documents. For example, some systems utilize key words associated with documents to automatically cluster or group similar documents. Such clusters can be iteratively grouped into super-clusters, thus creating a hierarchical structure, however, these systems require manual insertion of key words, and produce a hierarchy with no systematic structure. If the hierarchy is to be used for manual search, labels must be affixed to the nodes of the hierarchy by manual examination of the sub nodes or leaf documents to identify common feature(s).

Many classification systems utilize lists of salient words for classifying documents. Typically, salient words are either predefined or selected from the documents being processed to more accurately characterize the documents. Commonly these salient word lists are created by counting the frequency of occurrence of all words for each of a set of documents. Words are then removed from the word lists according to one or more criteria. Often, words that occur too few times within the corpus are eliminated, since such words are too rare to reliably distinguish among categories, whereas words that occur too frequently are eliminated, because such words are assumed to occur commonly in all documents across categories.
Further, "stop words" and word stems are often eliminated from feature lists to facilitate salient feature determination. Stop words comprise words which are common in the language such as "a", "the", "his", and "and", which are felt to carry no semantic content, whereas word stems represent suffixes such as "-ing", "-end", "-is", and "-able". Unfortunately, the creation of stop word and word stem lists is a language-specific task, requiring expert knowledge of syntax, grammar, and usage, which may change with time. Thus, a more flexible way of determining salient features is therefore desirable.

BRIEF DESCRIPTION OF DRAWINGS

The present invention will be described by way of exemplary embodiments, but not limitations, illustrated in the accompanying drawings in which like references denote similar elements, and in which:

Figure 1 illustrates an exemplary prior art subject hierarchy including multiple decision nodes;

Figures 2(A-C) illustrate an operational flow of a salient feature determination function, in accordance with one embodiment of the invention;

Figure 3 illustrates an example application of the salient feature determination facilities of the present invention, in accordance with one embodiment;

Figure 4 illustrates a functional block diagram of classifier training services of Figure 3, in accordance with one embodiment of the invention; and
Figure 5 illustrates an example computer system suitable for use in determining salient features, in accordance with one embodiment of the present invention.

**DETAILED DESCRIPTION OF THE INVENTION**

In the following description, various aspects of the present invention will be described. However, it will be apparent to those skilled in the art that the present invention may be practiced with only some or all aspects of the present invention. For purposes of explanation, specific numbers, materials and configurations are set forth in order to provide a thorough understanding of the present invention. However, it will also be apparent to one skilled in the art that the present invention may be practiced without the specific details. In other instances, well-known features are omitted or simplified in order not to obscure the present invention.

Parts of the description will be presented in terms of operations performed by a processor based device, using terms such as data, storing, selecting, determining, calculating, and the like, consistent with the manner commonly employed by those skilled in the art to convey the substance of their work to others skilled in the art. As well understood by those skilled in the art, the quantities take the form of electrical, magnetic, or optical signals capable of being stored, transferred, combined, and otherwise manipulated through mechanical and electrical components of the processor based device; and the term processor include microprocessors, micro-
controllers, digital signal processors, and the like, that are standalone, adjunct or embedded.

Various operations will be described as multiple discrete steps in turn, in a manner that is most helpful in understanding the present invention, however, the order of description should not be construed as to imply that these operations are necessarily order dependent. In particular, these operations need not be performed in the order of presentation. Further, the description repeatedly uses the phrase "in one embodiment", which ordinarily does not refer to the same embodiment, although it may.

In accordance with one embodiment of the invention, one or more unique features are extracted from a first group of objects to form a first feature set, and one or more unique features are extracted from a second group of objects to form a second feature set. A ranked list of features is then created by applying statistical differentiation between unique features of the first feature set and unique features of the second feature set. A set of salient features is then identified from the resulting ranked list of features.

In one embodiment, salient features are determined to facilitate efficient classification and categorization of data objects including but not limited to text files, image files, audio sequences and video sequences comprising both proprietary and non-proprietary formats within very-large-scale hierarchical classification trees as well as within non-hierarchical data structures such as flat files. In a text file for example, features may take the form of words where the term "word" is commonly understood to represent a group of letters within a given language, having some
semantic meaning. More generally, a feature could be an \(N\)-token gram, where a token is one atomic element of a language including \(N\)-letter grams and \(N\)-word grams in English, as well as \(N\)-ideogram grams in Asian languages for example. In audio sequences, for example, musical notes, intonation, tempo, duration, pitch, volume and the like may be utilized as features for classifying the audio, whereas in video sequences and still images, various pixel attributes such as chrominance and luminance levels may be utilized as features. In accordance with one embodiment of the present invention, once a group of features have been identified from a group of e.g. electronic documents, a subset of those features are then determined to be salient for the purposes of classifying a given group of data objects. The term "electronic document" is broadly used herein to describe a family of data objects such as those described above that include one or more constituent features. Although an electronic document may include text, it may similarly include audio and/or video content in place of, or in addition to text.

Once feature selection criteria have been determined (i.e. which of the various text/audio/video attributes will be utilized as determinative features within the set of data objects), the salient feature determination process of the present invention may be performed. To begin the salient feature determination process, the data objects in question are divided into two groups. An equation representing the "odds of relevance" is then applied to these groups of data objects (see e.g. equation 1), where \(O(d)\) represents the odds that a given data object is a member of a first group of data objects, \(P(R \mid d)\) represents the probability that the data object is
a member of the first group, and \( P(R' | d) \) represents that probability that the data
object is a member of the second group.

\[
O(d) = \frac{P(R | d)}{P(R' | d)} \quad \text{equation (1)}
\]

Because manual grouping of the data objects does not provide the probability
desired to compute the odds of relevance, equation (1) can be maximized to
approximate this value. Accordingly, the logarithm function in conjunction with
Baye's formula can be applied to both sides of equation (1), to yield equation (2):

\[
\log O(d) = \log P(d | R) - \log P(d | R') + \log P(R) - \log P(R'). \quad \text{equation (2)}
\]

If a data object is assumed to consist of a set of features \( \{F\} \), and \( X_i \) is either 1 or 0 if
a given feature \( f_i \) is present or absent in a data object, respectively, then:

\[
\log O(d) = \sum_i [\log P(X_i | R) - \log P(X_i | R')] + \log P(R) - \log P(R'). \quad \text{equation (3)}
\]

Since \( \log P(R) \) and \( \log P(R') \) are constant and independent of the features selected
as salient in the data object, a new quantity \( g(d) \) is defined such that:

\[
g(d) = \sum_i [\log P(X_i | R) - \log P(X_i | R')]. \quad \text{equation (4)}
\]

If it is assumed that \( p_i = P(X_i = 1 | R) \) represents the probability that the given feature
\( (f_i) \) occurs in a data object in the first group of data objects, and \( q_i = P(X_i = 1 | R') \)
represents the probability that the given feature \((f_i)\) occurs in a data object in the second group of data objects, then substitution and simplification yields equation (5):

\[
g(d) = \sum_{i} X_i \log \frac{p_i (1-q_i)}{q_i (1-p_i)} + \sum_{i} \left[ \log \frac{1-p_i}{1-q_i} \right] \quad \text{equation (5)}
\]

Because the second summation does not depend upon the feature occurrences in the data objects, it can be eliminated resulting in equation (6):

\[
\log \frac{p_i (1-q_i)}{q_i (1-p_i)} \quad \text{equation (6)}
\]

Because the log function is monotonic, maximizing the ratio of:

\[
\frac{p_i (1-q_i)}{q_i (1-p_i)} \quad \text{equation (7)}
\]

is sufficient to maximize the corresponding log value. In accordance with one embodiment of the invention, equation (7) is applied to each feature in the combined feature list for the two groups of data objects in order to facilitate identification of salient features. In order to do so, \(p_i\) is estimated to represent the number of data objects in the first group of data objects that contain feature \(f_i\) at least once, divided by the total number of data objects in the first group of data objects documents. Likewise, \(q_i\) is estimated to represent the number of data objects in the second group that contain feature \(f_i\) at least once, divided by the total number of data objects in the second group of data objects documents.

Figures 2(A-C) illustrate an operational flow of a salient feature determination function, in accordance with one embodiment of the invention. To begin, a first set
of data objects are examined to create a feature list consisting of unique features present within one or more data objects from at least the first set of data objects, block 210. For each unique feature identified, equation (7) is applied to generate a ranked list of features, block 220, and at least a subset of the ranked list of features are chosen as salient features, block 230. The salient features may comprise one or more contiguous or non-contiguous group(s) of elements selected from the ranked list of features. In one embodiment, the first $N$ elements of the ranked list of features are chosen as salient, where $N$ may vary depending upon the requirements of the system. In an alternative embodiment, the last $M$ elements of the ranked list of features are chosen as salient, where $M$ may also vary depending upon the requirements of the system.

In accordance with one embodiment of the invention, while creating the feature list, block 210, the total number of data objects contained within each group of data objects is determined, block 212, and for each unique feature identified within at least the first group of data objects, the total number of data objects containing the unique feature is also determined, block 214. Additionally, the list of unique features may be filtered based upon various criteria as desired, block 216. For example, the list of unique features may be pruned to remove those features that are not found in at least some minimum number of data objects, those features that are shorter than some established minimum length, and/or those features that occur a fewer number of times than an allotted amount.

In accordance with one embodiment of the invention, applying statistical differentiation to obtain a ranked list of features, as described with respect to block
220 of Figure 2A, further includes those processes illustrated in Figure 2C. That is to say, in applying statistical differentiation (i.e. as represented by equation (7)), a determination is made as to which of the unique features identified within the first set of data objects are also present within the second set of data objects, block 221, as well as a determination as to which of the unique features identified within the first set of data objects are not present within the second set of documents, block 222. In accordance with the illustrated embodiment, those features that are determined to be present in one set of data objects but not the other set are assigned a higher relative ranking within the ranked list of features, block 223, whereas those features that are determined to be present in both sets of data objects are assigned a lower relative ranking, as determined through statistical differentiation (i.e. equation (7)), block 224. Optionally, the features may further be ranked within the ranked feature list based upon the total number of data objects that contain each respective feature.

Example Application

Reference is now made to Figure 3, wherein a diagram illustrating an example application of the salient feature determination facilities of the present invention, in accordance with one embodiment, is shown. As illustrated, classifier 300 is provided to efficiently classify and categorize data objects such as electronic documents including but not limited to text files, image files, audio sequences and video sequences comprising both proprietary and non-proprietary formats, within a variety of data structures including very-large-scale hierarchical classification trees and flat file formats. Classifier 300 includes classifier training services 305, for training classifier 300 to categorize the new data objects based upon classification
rules extracted from a previously categorized data hierarchy, as well as classifier categorization services 315 for categorizing new data objects input into classifier 300.

Classifier training services 305 include aggregation function 306, salient feature determination function 308 of the present invention, and node characterization function 309. In accordance with the illustrated embodiment, content from the previously categorized data hierarchy is aggregated at each node in the hierarchy, through aggregation function 306 for example, to form both content and anti-content groups of data. Features from each of these groups of data are then extracted and a subset of those features are determined to be salient by way of salient feature determination function 308. Node characterization function 309 is utilized to characterize each node of the previously categorized data hierarchy based upon the salient features, and to store such hierarchical characterizations in data store 310 for example, for further use by classifier categorization services 315.

Additional information regarding classifier 300 including classifier training services 305 and classifier categorization services 315 are described in co-pending, US patent application number "<51026.P004>" entitled "Very-Large-Scale Automatic Categorizer For Web Content" filed contemporaneously herewith, and commonly assigned to the assignee of the present application, the disclosure of which is fully incorporated herein by reference.

Classifier Training Services

Figure 4 illustrates a functional block diagram of classifier training services 305 of Figure 3, in accordance with one embodiment of the invention. As illustrated in Figure 4, previously categorized data hierarchy 402 is provided for input into classifier training services 305 of classifier 300. Previously categorized data
hierarchy 402 represents a set of data objects such as audio, video and/or text objects, which have been previously classified and categorized into a subject hierarchy (typically through manual entry by individuals). Previously categorized data hierarchy 402 may represent one or more sets of electronic documents previously categorized by a Web portal or search engine for example.

In accordance with the illustrated example, aggregation function 406 aggregates content from previously categorized data hierarchy 402 into content and anti-content data groups so as to increase differentiation between sibling nodes at each level of the hierarchy. Salient feature determination function 408 operates to extract features from the content and anti-content groups of data and determine which of the extracted features (409) are to be considered salient (409').

Additionally, in accordance with the illustrated example, node characterization function 309 of Figure 3 operates to characterize the content and anti-content groups of data. In one embodiment, the content and anti-content groups of data are characterized based upon the determined salient features. In one embodiment, the characterizations are stored in data store 310, which can be implemented in the form of any number of data structures such as a database, a directory structure, or a simple lookup table. In one embodiment of the invention, the parameters for the classifiers for each node are stored in a hierarchical categorization tree having a file structure that mimics the previously categorized data hierarchy.

**Example Computer System**

Figure 5 illustrates an example computer system suitable for use in determining salient features, in accordance with one embodiment of the present invention. As shown, computer system 500 includes one or more processors 502 and system memory 504. Additionally, computer system 500 includes mass storage
devices 506 (such as diskette, hard drive, CDROM and so forth), input/output devices 508 (such as keyboard, cursor control and so forth) and communication interfaces 510 (such as network interface cards, modems and so forth). The elements are coupled to each other via system bus 512, which represents one or more buses. In the case where system bus 512 represents multiple buses, they are bridged by one or more bus bridges (not shown).

Each of these elements performs its conventional functions known in the art. In particular, system memory 504 and mass storage 506 are employed to store a working copy and a permanent copy of the programming instructions implementing the categorization system of the present invention. The permanent copy of the programming instructions may be loaded into mass storage 506 in the factory, or in the field, as described earlier, through a distribution medium (not shown) or through communication interface 510 (from a distribution server (not shown). The constitution of these elements 502-512 are known, and accordingly will not be further described.

Conclusion and Epilogue

Thus, it can be seen from the above descriptions, a novel method and apparatus for automatically determining salient features for object classification has been described. While the present invention has been described in terms of the above-illustrated embodiments, those skilled in the art will recognize that the invention is not limited to the embodiments described. The present invention can be practiced with modification and alteration within the spirit and scope of the appended claims. The description is thus to be regarded as illustrative instead of restrictive on the present invention.
CLAIMS

What is claimed is:

1. A method comprising:

   extracting one or more unique features from a first content group of data
   objects to form a first feature list;

   extracting one or more unique features from a second anti-content group of
   data objects documents to form a second feature list;

   creating a ranked list of features by applying statistical differentiation between
   unique features of said first feature list and unique features of said second feature
   list; and

   identifying a set of salient features from said ranked list of features.

2. The method of claim 1, wherein each of said first content group of data
   objects and said second anti-content group of data objects comprises one or more
   electronic documents.

3. The method of claim 1, further comprising:

   determining a first total number of data objects comprising said first content
   group of data objects; and

   determining a second total number of data objects comprising said second
   anti-content group of data objects.

4. The method of claim 3, further comprising:
determining, for each of said one or more unique features forming said first feature list, a first number of data objects of said first content group of data objects that contain at least one instance of each respective said one or more unique features of said first feature list; and

determining, for each of said one or more unique features forming said second feature list, a second number of data objects of said second anti-content group of data objects that contain at least one instance of each respective said one or more unique features of said second feature list.

5. The method of claim 4, wherein creating said ranked list comprises:

identifying as exclusive features those unique features of said first feature list that are not present in said second feature list;

identifying as common features those unique features of said first feature list that are also present in said second feature list; and

ordering said ranked list such that said exclusive features are ranked higher within said ranked list as compared to said common features.

6. The method of claim 5, further comprising:

applying a probabilistic function to each of said common features to obtain a result vector, wherein said probabilistic function comprises a ratio of the first number of data objects divided by said first total number of data objects, to said second number of data objects divided by said second total number of data objects; and
ordering said common features within said ranked list based at least in part upon the result vector of said probabilistic function.

7. The method of claim 5, wherein said exclusive features are further ranked based upon the first number of data objects.

8. The method of claim 1, wherein identifying said set of salient features from said ranked list of features comprises selecting a first $N$ contiguous features of said ranked list of features.

9. The method of claim 1, wherein identifying said set of salient features from said ranked list of features comprises selecting a last $M$ contiguous features of said ranked list of features.

10. The method of claim 1, wherein each of said unique features comprises a grouping of one or more alphanumeric characters.

11. The method of claim 1, further comprising:
    classifying a new data object as being most related to one of said first content group of data objects and said second anti-content group of data objects based at least in part upon said set of salient features.
12. The method of claim 1, wherein said first content group of data objects comprises those data objects corresponding to a selected node of a subject hierarchy having a plurality of nodes and any associated sub-nodes of the selected node; and

wherein said second anti-content group of data objects comprises those data objects corresponding to any associated sibling nodes of the selected node and any associated sub-nodes of the sibling nodes.

13. A method of identifying salient features, the method comprising:

identifying one or more unique features that are members of a first data class;

examining a second data class to identify those of said one or more unique features that are also members of said second data class, and those of said one or more unique features that are not members of said second data class;

generating a ranked list of unique features having an order based upon membership of each of said one or more unique features within said second data class; and

identifying as salient one or more of said ranked list of unique features.

14. The method of claim 13, further comprising:

determining, for each of said ranked list of unique features, a number of objects within said first data class that contain each respective unique feature.
15. The method of claim 14, wherein generating a ranked list further comprises ranking those of said unique features that are not members of said second data class higher in said ranked list than those of said unique features that are also members of said second data class.

16. The method of claim 15, wherein generating a ranked list further comprises ranking those of said unique features that belong to a greater number of object of said first data class higher in said ranked list than those of said unique features that belong to a lesser number of objects within said first data class.

17. The method of claim 13, wherein identifying as salient comprises selecting a first set of $N$ consecutive unique features from said ranked list of unique features.

18. The method of claim 13, wherein identifying as salient comprises selecting a last $M$ consecutive unique features from said ranked list of unique features.

19. An apparatus comprising:
   a storage medium having stored therein a plurality of programming instructions designed to implement a plurality of functions of a category name service for providing a category name to a data object, including first one or more functions to extract one or more unique features from a first content group of data objects to form a first feature list,
extract one or more unique features from a second anti-content group of data objects to form a second feature list,
create a ranked list of features by applying statistical differentiation between unique features of said first feature list and unique features of said second feature list, and identify a set of salient features from said ranked list of features; and a processor coupled to the storage medium to execute the programming instructions.

20. The apparatus of claim 19, wherein each of said first content group of data objects and said second anti-content group of data objects comprises one or more data objects.

21. The apparatus of claim 19, wherein said plurality of instructions further comprise instructions to determine a first total number of data objects comprising said first content group of data objects, and determine a second total number of data objects comprising said second anti-content group of data objects.

22. The apparatus of claim 19, wherein said plurality of instructions further comprise instructions to
determine, for each of said one or more unique features forming said first feature list, a first number of data objects of said first content group of data objects that contain at least one instance of each respective said one or more unique features of said first feature list, and
determine, for each of said one or more unique features forming said second feature list, a second number of data objects of said second anti-content group of data objects that contain at least one instance of each respective said one or more unique features of said second feature list.

23. The apparatus of claim 20, wherein said plurality of instructions to create said ranked list comprise instructions to
   identify as exclusive features those unique features of said first feature list that are not present in said second feature list,
   identify as common features those unique features of said first feature list that are also present in said second feature list, and
   order said ranked list such that said exclusive features are ranked higher within said ranked list as compared to said common features.

24. The apparatus of claim 23, wherein said plurality of instructions further comprise instructions to
   apply a probabilistic function to each of said common features to obtain a result vector, wherein said probabilistic function comprises a ratio of the first number
of data objects divided by said first total number of data objects, to said second number of documents divided by said second total number of data objects, and order said common features within said ranked list based at least in part upon the result vector of said probabilistic function.

25. The apparatus of claim 23, wherein said exclusive features are further ranked based upon the first number of data objects.

26. The apparatus of claim 19, wherein said plurality of instructions to identify said set of salient features from said ranked list of features further comprises instructions to select a first $N$ contiguous features of said ranked list of features.

27. The apparatus of claim 19, wherein said plurality of instructions to identify said set of salient features from said ranked list of features further comprises instructions to select a last $M$ contiguous features of said ranked list of features.

28. The apparatus of claim 19, wherein each of said unique features comprises a grouping of one or more alphanumeric characters.

29. The apparatus of claim 19, wherein said plurality of instructions further comprise instructions to
classify a new data object as being most related to one of said first content

group of data objects and said second anti-content group of data objects based at

least in part upon said set of salient features.

30. The apparatus of claim 19, wherein said first content group of data objects

comprises those data objects corresponding to a selected node of a subject

hierarchy having a plurality of nodes and any associated sub-nodes of the selected

node; and

wherein said second anti-content group of data objects comprises those data

objects corresponding to any associated sibling nodes of the selected node and any

associated sub-nodes of the sibling nodes.

31. An apparatus comprising:

a storage medium having stored therein a plurality of programming

instructions designed to implement a plurality of functions including first one or more

functions to

identify one or more unique features that are members of a first data

class,

examine a second data class to identify those of said one or more

unique features that are also members of said second data class, and those

of said one or more unique features that are not members of said second

data class,
generate a ranked list of unique features having an order based upon membership of each of said one or more unique features within said second data class, and identify as salient one or more of said ranked list of unique features; and a processor coupled to the storage medium to execute the programming instructions.

32. The apparatus of claim 31, wherein said plurality of instructions further comprise instructions to determine, for each of said ranked list of unique features, a number of objects within said first data class that contain each respective unique feature.

33. The apparatus of claim 32, wherein said plurality of instructions to generate a ranked list further comprises instructions to rank those of said unique features that are not members of said second data class higher in said ranked list than those of said unique features that are also members of said second data class.

34. The apparatus of claim 33, wherein said plurality of instructions to generate a ranked list further comprise instructions to rank those of said unique features that belong to a greater number of objects of said first data class higher in said ranked list than those of said unique features that belong to a lesser number of objects within said first data class.
35. The apparatus of claim 31, wherein said plurality of instructions to identify as salient further comprise instructions to select a first set of $N$ consecutive unique features from said ranked list of unique features.

36. The apparatus of claim 31, wherein said plurality of instructions to identify as salient further comprise instructions to select a last set of $M$ consecutive unique features from said ranked list of unique features.
CREATE A FEATURE LIST CONTAINING UNIQUE FEATURES PRESENT WITHIN AT LEAST 1ST SET OF DATA OBJECTS

FOR EACH IDENTIFIED UNIQUE FEATURE, APPLY STATISTICAL DIFFERENTIATION BETWEEN 1ST AND 2ND GROUPS OF DATA OBJECTS TO GENERATE A RANKED LIST OF FEATURES

IDENTIFY AT LEAST A SUBSET OF THE RANKED LIST OF FEATURES TO REPRESENT SALIENT FEATURES

FIGURE 2A
Determine the total number of data objects within each group of data objects

For each unique feature identified, determine the total number of data objects that contain the feature in the corresponding group of data objects

Filter unique features based on various criteria as desired

FIGURE 2B
DETERMINE WHICH OF THE UNIQUE FEATURES IDENTIFIED WITHIN THE FIRST SET OF DATA OBJECTS ARE ALSO PRESENT WITHIN THE SECOND SET OF DATA OBJECTS

DETERMINE WHICH OF THE UNIQUE FEATURES IDENTIFIED WITHIN THE FIRST SET OF DATA OBJECTS ARE NOT PRESENT WITHIN THE SECOND SET OF DATA OBJECTS

ASSIGN HIGHER RANKINGS TO THOSE FEATURES FOUND WITHIN ONE SET OF DATA OBJECTS AND NOT THE OTHER SET

ASSIGN LOWEST RANKINGS TO THOSE REMAINING BASED UPON OUTCOME OF STATISTICAL DIFFERENTIATION RATIO

FIGURE 2C
PREVIOUSLY CATEGORIZED DATA HIERARCHY

DATA OBJECT TO BE CATEGORIZED

AGGREGATION FUNCTION

SALIENT FEATURE SELECTION FUNCTION

NODE CHARACTERIZATION FUNCTION

CLASSIFIER TRAINING SERVICES

CLASSIFIER CATEGORIZATION SERVICES

CLASSIFIER

CATEGORIZED DATA OBJECT

FIGURE 3
FIGURE 5
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : G06F 17/00, 17/30
US CL : 715/3, 7
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
U.S. : 715/3, 7

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
IEEE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>US 6,353,825 B1 (PONTE) 5 March 2002 (05.03.2002), column 22, line 59 - column 23, line 32, lines 57-65, column 24, lines 34-53, column 30, line 48 - column 31, line 34.</td>
<td>1-36</td>
</tr>
</tbody>
</table>

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

*P* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

*X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

*Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

* #: document member of the same patent family

Date of the actual completion of the international search

Date of mailing of the international search report
09 JAN 2003

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703)305-3230

Authorized officer
Heather Hergens
Telephone No. (703) 305-4700

Form PCT/ISA/210 (second sheet) (July 1998)
Continuation of B. FIELDS SEARCHED Item 3:

EAST

search terms: unique, salient, ranking, classifying, common words or features