VISUAL DATA MINING WITH PIXEL ORIENTED AND PARALLEL CO-ORDINATE TECHNIQUES ON MEDIACL DATA

[1]HARISHBABU.KALIDASU, [2]GUDIPATI MURALI

[1]Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
[2]Professor, Dept. of CSE, KKR & KSR Institute of Technology & Sciences, Guntur, Andhra Pradesh, India.

Abstract— Visual data mining as a sculpture and discipline of testing significant discernments out of enormous capacities of data that are inexplicable in another way requires consistent visual data representations. In contrast, traditional data mining techniques are not suitable to process these enormous volumes of data. On the other hand Visual data mining procedures have proven to be of more significance in investigative data analysis and they also have a high potential for mining large datasets. Particularly, biological data have been increasing ecologist are stepping up their determinations in appreciative the biological processes that trigger sickness pathways in the medical contexts. In this paper we evaluate the cancer patient data under visual data mining two popular techniques which are pixel oriented and parallel co-ordinate techniques and the performance is also presents as a graphs below. And also make comparison between two techniques in handling of data and accuracy of results. Data mining is the procedure of finding possibly important examples, affiliations, patterns, arrangements and conditions in information. Key business samples incorporate site access examination for enhancements in e-trade promoting, misrepresentation recognition, screening and examination, retail site or item investigation, and client division. Information mining systems can find data that numerous conventional business investigation and measurable strategies neglect to convey. Furthermore, the use of information mining methods further endeavors the estimation of information distribution center by changing over costly volumes of information into important resources for future strategic and vital business improvement. Administration data frameworks ought to give propelled abilities that give the client the ability to ask more modern and applicable inquiries. It engages the right individuals by giving the particular data they require.
Numerous learning revelation applications, for example, on-line administrations and internet applications, require precise mining data from information that progressions all the time. In such a situation, successive or intermittent upgrades may change the status of a few standards found before. More data ought to be gathered amid.

I. INTRODUCTION

Over the past few decades data have been generated great number of volumes. For examining and exploring these gigantic volumes of data has becoming progressively difficult. In contrast, traditional data mining techniques are not suitable to process these enormous volumes of data. On the other hand Visual data mining procedures have proven to be of more significance in investigative data analysis and they also have a high potential for mining large datasets. Especially, biological information have been expanding scholars are venturing up their endeavors in comprehension the biological procedures that underlie sickness pathways in the clinical settings. This has brought about a surge of biological and clinical information from genomic and protein successions, DNA microarrays, protein communications, biomedical pictures, to infection pathways and electronic wellbeing records. To abuse these information for finding new learning that can be deciphered into clinical applications, there are basic information investigation troubles that must be succeed. Practical issues such as handling noisy and incomplete data, processing compute-intensive tasks, and integrating various data sources, are new challenges faced by biologists in the post-genome era. Data representation methods are turning out to be progressively vital for investigation of vast multidimensional information sets. A noteworthy point of preference over programmed information investigation and examination strategies is that representation permits an immediate connection with the client and gives a quick input and additionally guiding which is hard to accomplish in most non-visual methodologies. The handy significance of visual information mining methods is along these lines consistently expanding and essentially all business information mining frameworks to fuse perception systems of one kind or other. In this article we mine biological data set using pixel orientation method and parallel co-ordinate method. Mainly medical data is very complex and important data which need to explore and mine exactly. The remaining article is organized as follows part-2 gives existing work regarding visual data mining, part-3 describes the method of visual analysis, part-4 presents the experimental analysis and finally part-5 concludes the article.
Geographic information mining can be characterized as an arrangement of exploratory computational and measurable methodologies for examining huge spatial and spatiotemporal information sets. Information mining procedures are frequently assembled into classes that incorporate grouping, arrangement, rundown, guideline mining, and highlight extraction. These sorts of methods are for the most part situated towards distinguishing spatial or spatio-transient examples in geographic perceptions or estimations. The recognizable proof of these examples is planned to goad theory era with regards to the geographic procedures from which the examples are produced.

Information mining can be viewed as one stage in the bigger procedure of geographic learning disclosure (Fayyad et al. 1996). This procedure is both intuitive and iterative and incorporates steps, for example, information choice, information cleaning, and the understanding of information mining results. While there is an assortment of scholastic and business information mining programming accessible, there are couple of complete learning revelation programming situations. The procedure of learning revelation is bolstered completely by the investigator, who is in charge of keeping track of the consequences of various information mining "keeps running," for occurrence depictions of tenets or removed components. These outcomes are in perfect world components of learning – synopses of examples implanted inside the observational information. In any case, these outcomes can likewise be viewed as a type of information themselves that regularly need further examination to yield helpful elucidation. By and large, the information sets coming about because of information mining are extensive and complex, yet there are couple of computational systems for dealing with these information mining results to completely bolster the learning revelation process.

The examination of the aftereffects of information mining is called meta-mining (Abraham and Roddick 1999). I contend here that geographic learning revelation programming requests support for information digging as well as for the capacity to outwardly and algorithmically meta-mine the consequences of information mining in an intelligent and iterative way. Computational backing for meta-mining is subject to the perception and database representation of the principles, bunches, and components that are the consequences of information mining. In this way, geographic information revelation programming situations must join perception and semantic database demonstrating systems for the representation of, and communication with, these learning components.

Database Support for Geographic Meta-Mining

There has been considerable examination in PC and data science, manmade brainpower, and related fields in computational learning representation and semantic information models.

This exploration has been stretched out by geographers and others for geographic databases. As of late, quite a bit of this examination has gone under the heading of philosophy – the improvement and formal encoding of the theoretical components and connections creating specific application spaces. Just a modest bunch of these examination endeavors have been coordinated towards geographic meta-mining, nonetheless.

There has been some exploration in meta-mining affiliation rules. Affiliation guideline mining distinguishes rules in value-based databases in view of the rate of co-event of specific qualities inside an arrangement of exchanges. Spatial affiliation standard mining is characterized as when a guideline contains a spatial relationship. An information set with only a modest bunch of traits, each with an only a couple of potential qualities, can create countless principles. Not at all like in a formal factual examination, there is no measure of noteworthiness in affiliation decide mining that one may use to winnow these outcomes to a reasonable (and interpretable) size. Or maybe, results might be assessed by relegating a measure of "interestingness" to every tenet, for example, the certainty and bolster measurements. Interestingness measures are genuinely subjective gadgets, in any case, and in my own particular experience I have found that decides that are of real hobby may score as generally "uninteresting," putting in uncertainty the utility of such measurements.

Representation Support for Geographic Meta-Mining

There are numerous built up methodologies for picturing geographic information: maps, scatterplot lattices, and parallel direction plots, just to give some examples. Systems have additionally been created for envisioning spatio-worldly information, for example, activity and little products. Other visual information mining strategies have been created for extensive, multidimensional information sets, conceivably spatial, and are regularly arranged as geometric, symbol based, various leveled, and pixel-based methodologies. These procedures regularly are utilized for spatial information as a part of which various properties are recorded for every area, or every area at different times. A sign of exploratory perception is the capacity to interface with various visual representations through brushing and connected showcases.

Representation systems to bolster geographic meta-mining may now and again be effortlessly exchanged from those utilized for envisioning geographic information. In the event that the aftereffects of geographic information mining take the type of an arrangement of individual georeferenced perceptions with numerous traits joined, then the "customary" representation systems depicted above can be effectively connected. For instance, if highlight extraction is connected to a spatial information set and yields an arrangement of geographic elements, each with its own particular arrangement of properties, these components can be mapped or investigated utilizing a scatterplot lattice.

Assuming, in any case, the aftereffects of information mining take a structure that is entirely unique in relation to the standard area/quality set organization of most geographic information, for example, an arrangement of affiliation guidelines, new sorts of particular perception systems must be created. A straightforward methodology for picturing the aftereffects of affiliation tenet mining would be to just guide where certain principles happen. A more

novel methodology may utilize spatialization (the utilization of a spatial allegory to "guide" non-spatial information qualities) to show the diverse standards in a trait space where the measurements comprise of various characteristics and interestingness measurements.

A key part of visual meta-mining is intelligence and the association with database learning representation. For instance, consider a speculative circumstance in which an element extraction calculation has been utilized to distinguish an arrangement of individual elements installed in the information. These components might be mapped and plotted on a scatterplot framework. The expert may recognize one specific component that is of hobby and need to discover different elements like that one in the information. A meta-mining framework ought to bolster the capacity of the examiner to choose a specific component from a representation, encode that element in the database as a learning component, and utilize that information to discover comparable elements in the database.

Adaptability comes in no less than two assortments, cardinality versatility (an excessive number of occurrences) and dimensionality adaptability (an excessive number of qualities). The adaptability issues can be thrown as far as plain phrasing as excessively numerous lines and an excessive number of segments. Obviously, an excessive number of tables can likewise still be an issue, however one could say that that issue has been unraveled to an attractive degree by the database research group in the course of recent years, confirm by the way that most database analysts are presently doing information mining. The two issues will be alluded to as the scourge of cardinality and the scourge of dimensionality. The essential answer for date has been testing.

The essential answer for the scourge of cardinality has been to choose (arbitrarily?) an agent subset of records (occurrences or lines), then to investigate or mine that subset. The inferred supposition is that the data (connections, designs, outlines, and so forth.) found in the subset applies to the full information set also. While, that implied (measurable) presumption can (and ought to dependably) be legitimized much of the time (especially when the answers looked for are of a synopsis nature), it is exceptionally hard to legitimize in others, e.g., in exemption mining. An irregular subset will quite often miss special cases, since exemptions are, in some sense, of measure zero, and little arbitrary sub-tests cross measure zero sets with measure zero. Put another way, if the likelihood is high that sub-testing will incorporate a special case, then it might be wrong to call it an exemption in any case.

There is an unequivocal requirement for a class of full-example answers for the scourge of cardinality. It is proposed in this publication, that such arrangements ought to structure the information vertically rather than the omnipresent level (record-based) organizing. Why? Roughly, compacted vertical structures don't develop in number as more records are created. Each of them will develop in size, however with legitimate pressure, they will develop exceptionally sub-directly. Another conceivably critical normal for a decent packed vertical innovation is that handling of the structures can be refined on the compacted form (and not require decompression first).

Two perceptions should be made instantly. In the first place, lists to flat information sets are vertical, so vertical organizing is not new. In any case, files are assistant vertical information structures which are made (and kept up) notwithstanding the flat information sets they record. One approach to see the vertical arrangement suggested here is that it replaces the flat information set with one all-inclusive record, maybe. Second, completely vertical databases have been proposed before (e.g., the Bubba venture of the 1980's). Perhaps, the reason that these vertical database innovations have been over-shadowed by the pervasive flat social advances, is that for database preparing, the craved result as a rule has level structure (the yield of a social inquiry is normally a connection itself). In this way, to structure the information vertically, prepare vertically, then need to change over the outcome to an even structure, may have been excessively wasteful.

The essential answer for the scourge of dimensionality is additionally to choose (non-arbitrarily) an appropriate subset of components (segments or properties). This procedure is frequently alluded to as highlight choice (e.g., primary segment examination). It can likewise include custom pivot first and after that element choice. Indeed, this arrangement is not in the way of a work-around (which the sub-testing of examples is for the scourge of cardinality). Given there IS a diminished subset of components which ARE the relevant ones for the examination embraced (i.e., the subset holds about all the data required), those ARE the elements that ought to be engaged upon. Be that as it may, infrequently that sub-gathering of components is still substantial (and here and there all elements are appropriate – i.e., hold imperative data). In these later cases the, supposed, condemnation of dimensionality might be all the more fittingly termed the certainty of high appropriate dimensionality, which is to say, there may not exist an adaptable answer for it.

II. EXISTING WORK

Ankerst [2] arranges current visual information mining methodologies into three classifications. Systems for the first gathering apply representation methods free of information mining calculations. The second gathering uses representation keeping in mind the end goal to speak to examples and results from mining calculations graphically. The third

classification firmly incorporates both mining and representation calculations in a manner that halfway strides of the mining calculations can be pictured.

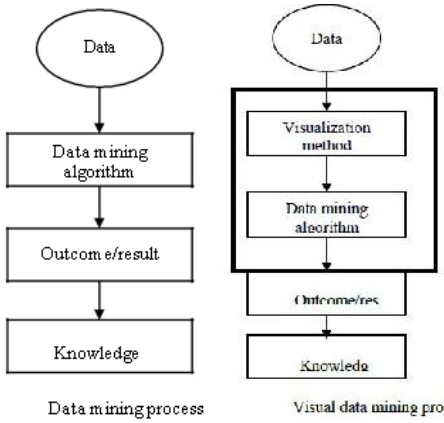

Figure-1: flow diagram of datamining and visual datamining

Moreover, this tight reconciliation permits clients to control and guide the mining process specifically taking into account the given visual criticism. An assortment of representation strategies which have been created in distinctive spaces can be arranged into the first gathering alluding to the order given above. Among these are systems for envisioning multidimensional data. These strategies attempt to guide relationships of articles in high-dimensional data space to spatial connections in a 2D or 3D presentation space. Among these are methodologies like IVORY, VR-VIBE, and Narcissus, which endeavor spring models to place articles as per their similitudes, whereby comparative items are put spatially near one another. Different frameworks, as Lyberworld [3] and SPIRE [3], use diverse visual allegories like Relevance Spaces, Information Galaxies, or Themescapes [5] with a specific end goal to picture report accumulations or results from information base recovery. Center [2] is an intuitive table viewer which underpins the investigation of complex item characteristic tables by a blend of a focus plus context method, a progressive outliner for expansive quality sets and a general simple to-utilize element inquiry instrument. Other visual interfaces have been created for picturing and connecting with chains of command, similar to Cone Trees [8] or Disk Trees [8], which utilize flat and vertical cones or plates to design progressive systems. FSN [6] and Information Pyramids [6] abuse the allegory of 3D data scenes to delineate extensive various leveled data spaces. Different methodologies, for example, Treemaps [8] and CHEOPS [8], are surely understood 2D systems which utilize accessible screen space viably. The representation of mining models (class 2 of the characterization of visual information mining methodologies) can be found in [6], where progressive group structures are found and imagined in view of verifiable surfaces. Different illustrations are WebSOM [7], which applies shading coded planes to envision after effects of a Self-Organizing Map calculation,

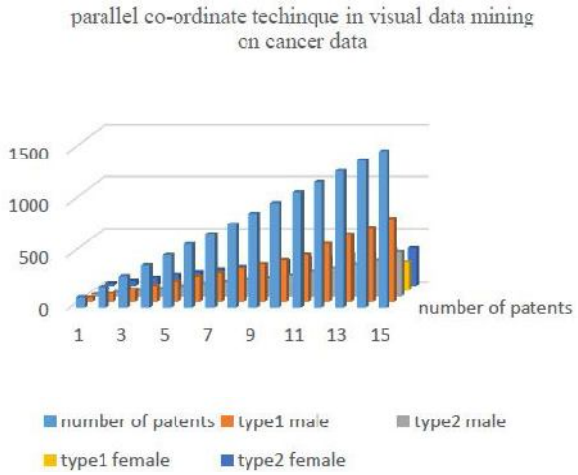

Figure-2: pixel oriented method

or OPTICS [8], which shows progressive clusterings. Frameworks like Descartes [7] or Devise [6] give answers for imagining geologically related data. Distinctive sorts of symbols, graphs, shaded confronts, and maps are utilized for delineating information inside of their spatial casing of reference. These frameworks, be that as it may, don't bolster the representation of rather complex data structures, as, for example, conceptual hub connection charts or chains of importance. A large portion of the frameworks said above explain, each in its own particular way, a portion of the single issues presented before in this segment. Up to now, there are still open inquiries of how to give an adaptable system to taking care of those issues in a more broad manner.

Datamining

Huge or Vast amount of data is produced and available daily. Data is sometimes noisy and incomplete; some important patterns are also missing. These types of data is not understand easily. The reliability of data is very low. So we have to correct the data. Data analyst will decide what kinds of data mining procedures stay used to correct the data.

Data mining is the process of discovering patterns and knowledge from large databases. Knowledge discovery process or simply KDP is synonym for data mining. The knowledge discovery process has iterative sequence of steps.

III. VISUAL DATA MINING

Visual data mining as a sculpture and discipline of testing significant discernments out of enormous capacities of facts that are bizarre in alternative way necessitates consistent visual data illustrations. The frequently used manifestation “the art of information visualization” appropriately describes the situation. Though substantial work has been done in the area of information visualization. It is tranquil a stimulating

activity to find out the methods, techniques and analogous tools that support visual data mining of a particular type of information. Visual data mining can assistance in dealing with the tidal wave of information. The benefit of visual data investigation is that the user is openly involved in the data mining process, through investigation the outcomes of the information visualization, user can

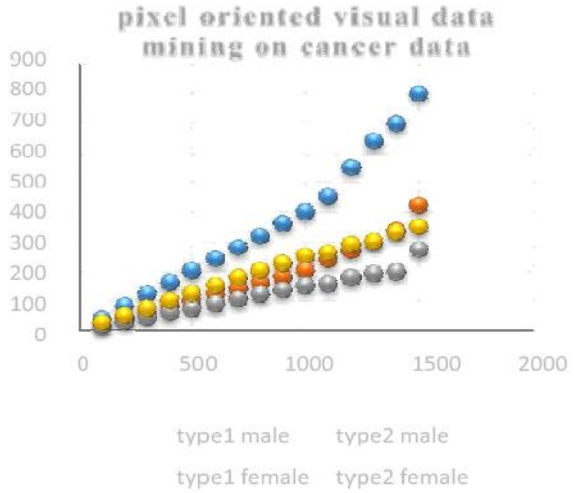

Figure-3: parallel coordinate method

Assimilate the professional acquaintance with the data mining algorithm

Visual data mining techniques:

Pixel-oriented techniques have been initiated by Keim for visual data mining. Which representing large amounts of high dimensional data with veneration to a specified query. As a result the user of the system is able to refine his query based on the knowledge gathered from the visual representation of the data. Pixel-oriented technique is characterized with different attributes and each and every attributed recognized as a distinct colored pixel. The range of possible attribute values are grouped and mapped to a standard color map. Diverse attributes are displayed with dissimilar sub colors. The main aim of this pixel oriented visualization method is to exploit the amount of information signified at one time without any intersection. It maintains comprehensive view by effectively preserve the perception of each small region of interest. These possessions makes pixel oriented visualization technique will handle large data sets effectively. Particularly, this technique widely useful to handle vast amount of heterogeneity and diverse dimensionality issues

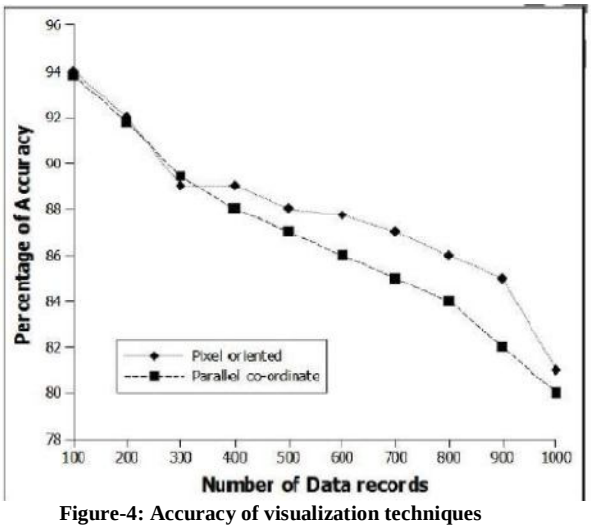

Figure-4: Accuracy of visualization techniques

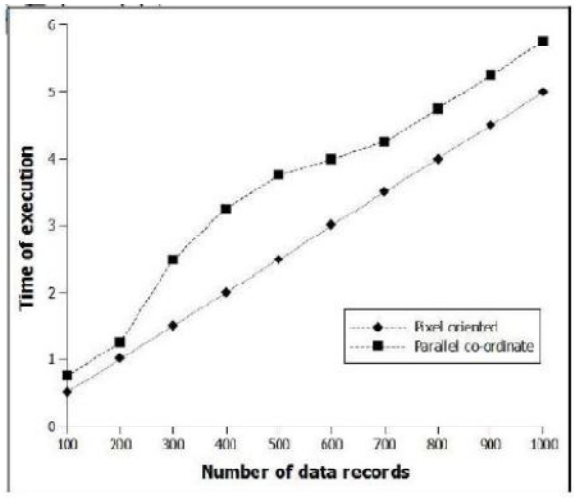

Figure-5: execution time of visualization techniques

Algorithm to pixel-oriented technique:

Algorithm to pixel-oriented ()

{

Scan the high volumes of data HD;

Data processing ()

{

Data cleaning;

Integration;

}

Query to analyze data; // user can give a query according to his view of mining or generating the patterns

Map the data as colored pixels;

}

Geometric Projection Techniques:

Geometric projection technique is one of the visualization technique which is used to find the fascinating predictions of multi-dimensional informatics. Basically, geometric projection

procedures embraces practices of empirical statics. Projection pursuit is a term which includes principal component analysis, factor analysis and multi-dimensional scaling. Subsequently for to project multi-dimensional data into the 2D data there are infinite number of possibilities but projection pursuit system intention at spontaneously finds the interested patterns or helps users to identify. Parallel co-ordinate technique is the one of the geometric visualization technique. The principle idea of the parallel coordinate visualization technique is relatively modest, it is powerful in enlightening a wide range of data individualities such as different data allocations and functional enslavements. The parallel coordinate technique maps the n-dimensional space into the 2D by using n-equidistant axes which are parallel to one of the spectacle 2D axes. The axes match to the dimensions and are linearly mounted from the minimum value to the maximum value of the equivalent dimension.

Algorithm to parallel coordinate ()

{

Scan the high volumes of data HD;

Data processing ()

{

Data cleaning;

Integration;

}

Map the data parallel to any one of the axes;

}

IV. EXPERIMENTAL MODEL

This section presents the experimental results and analysis done for this work. For the experiments, two different visualization techniques have been applied on the 1000 patents cancer dataset taken from UCI repository.

Here we test the cancer patents data contains patent-id, name of the patent, age, sex, occupation, alcohol, smoke, BP and weight. Based on these attributes we process the data under two methodologies those are pixel oriented and parallel co-ordinate techniques. The nature of cancer is still ambiguous in nature. Both the genetics and environmental factors such as patents habits, age sex are subsidizing factors. Research on this complex issue has been done by many scientists using data mining techniques. Visual data mining techniques helps the scientists to know the exact cause. By the analysis the doctors and scientist move ahead to improve the treatment and search for new way to get drugs also.

In this paper we evaluate the two popular visual data mining techniques to mine the medical data of 1000 patents. And we present the performance of the two techniques below and also we compare the two techniques under the time complexity.

The experiment is done with well popular tool weka. Here we test the cancer patents data under pixel oriented and parallel coordinate techniques with the same attributes. And both are analyzed the medical data and give results accuracy. The rate accuracy is also presented below. Figure-4 shows the accuracy of visualization techniques pixel oriented and parallel co-ordinate techniques. Pixel oriented accuracy is slight more than parallel co-ordinate technique. And in both the cases the percentage of accuracy is up to 80%. Figure-5 shows the time to process a data set using visualization techniques pixel oriented and parallel co-ordinate techniques. Pixel oriented processed rapid than parallel co-ordinate technique.

CONCLUSION

Visual data mining procedures have proven to be more significance in investigative data analysis and they also have a high potential for mining large datasets. Particularly, biological data have been increasing biologists are stepping up their efforts in understanding the biological processes that underlie disease pathways in the clinical contexts. In this article we comparing two popular visual mining techniques pixel oriented and parallel co-ordinate techniques and also analyze the biological data. And also compare the both the techniques under accuracy results and also execution time. Pixel oriented technique perform slight better than parallel co-ordinate technique.

A brief inquiry on the web yields numerous individual information mining programming apparatuses. Some of these apparatuses bolster, or might be adjusted for, spatial and spatiotemporal information. Be that as it may, the computational backing for the procedure of geographic information revelation is for the most part inadequate. This is because of the way that the administration of the information revelation process (i.e. monitoring, information mining decisions and parameterizations, results, and translations) must be kept up specially appointed by the investigator. Progresses in creating information disclosure programming situations ought to look for not just to fuse an assortment of information mining apparatuses, additionally the way to bolster geographic meta-mining and, thus,the intuitive and iterative nature of the learning revelation process.

Challenges:

1. Development of information structures to bolster

the capacity of both information and learning (i.e. information mining results and in addition investigator characterized area learning), and the mapping from one to the next

2. Development of methods of cooperation for database representations of information
3. Development of perception methods for the consequences of information mining
4. Development of methods of cooperation for the representation of information mining results
5. Development of the linkage among visual meta-mining systems and information structures for learning representation

REFERENCES

[1] Alfredo Cuzzocrea, Davood Zall "Parallel Coordinates Technique in Visual Data Mining: Advantages, Disadvantages, and Combinations" 2013 17th International Conference on Information Visualization.

[2] David Otasek, Chiara Pastrello, Andreas Holzinger, and Igor Jurisica "Visual Data Mining: Effective Exploration of the Biological Universe "A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 19–33, 2014. © Springer-Verlag Berlin Heidelberg 2014.

[3] Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: DATA 2012, Rome, Italy, 9–20. INSTICC (2012)

[4] Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)

[5] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Rhee, S.Y.: Big data: The future of biocuration. Nature 455(7209), 47–50 (2008)

[6] Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)

[7] Kreuzthaler, M., Bloice, M.D., Faulstich, L., Simonic, K.M., Holzinger, A.: A Comparison of Different Retrieval Strategies Working on Medical Free Texts. J. Univers. Comput. Sci. 17(7), 1109–1133 (2011)

[8] Wu, X.D., Zhu, X.Q., Wu, G.Q., Ding, W.: Data Mining with Big Data. IEEE Transactions

[9] Holzinger, A.: Weakly Structured Data in Health-Informatics: The Challenge for Human- Computer Interaction. In: Proceedings of INTERACT 2011 Workshop: Promoting and Supporting Healthy Living by Design. IFIP, pp. 5–7 (2011)

[10] Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R.: Combining HCI, Natural Language Processing, and Knowledge Discovery – Potential of IBM Content Analytics as an assistive technology in the biomedical domain. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 13–24. Springer, Heidelberg (2013)

★★★