ВИЗУЕЛИЗАЦИЈА КАКО АЛАТКА ЗА АНАЛИЗА НА ПОДАТОЦИ

Снежана Савоска[1], Сузана Лошковска[2]

[1] Факултет за Администрација и Менаџмент на Информациски системи – Партизанска бб, 7000 Битола, Македонија, savoskasnezana@yahoo.com

[2] Факултет за Електротехничка и информациони технологии – Скопје, Карпош II бб, 1000 Скопје, Македонија, suze@feit.ukim.edu.mk

***Апстракт* – Улогата на човечкиот фактор во процесот на визуелното податочно истражување е посебно важна, бидејќи на тој начин се комбинира флексибилноста, креативноста и знаењето на луѓето со огромните податочни простори и моќта на пресметувањето на компјутерите. Поради тоа, визуелизацијата е една од најзначајните техники за анализа на податоци. Овој труд претставува преглед на визуелизацијата на информации, визуелното истражување, класификација на видовите податоци што се визуелизираат, техниките за визуелизација и техниките за интеракција.**

***Клучни зборови* – Визуелизација, Визуелна податочна анализа, Визуелно истражување на податоци**

1. ВОВЕД

Прогресот што го носи денешната технологија и можноста за зачувување големи количества на податоци и нивно обработување, придонесоа за создавање огромни складишта на податоци во дигитална форма. Податоците се собираат бидејќи луѓето веруваат дека се потенцијални извори на корисни информации кои овозможуваат проникнување во проблемите или стекнување на конкурентска предност. Денешните статистики покажуваат дека 1,5 Exabytes податоци се генерираат во светот секоја година. Следните години, се претпоставува дека ќе се генерираат повеќе податоци отколку што се генерирале во целата човечка историја [1]. Податоците најчесто се снимаат со сензори и системи за мониторинг, а се снимаат дури и при секојдневните трансакции како плаќањата со кредитни картички или користење на телефоните. Податоците што се снимаат, најчесто се мултидимензионални односно секој податок содржи одреден број атрибути. Ако податоците се прикажуваат на корисниците текстуално, бидејќи количеството податоци содржи илјадници податочни членови, наоѓањето на саканата информација во податоците станува исклучително тешко. Во таков случај не е можно податоците да бидат истражувани или анализирани бидејќи стануваат некорисни а базите не ги даваат очекуваните информации. Визуелизацијата на информациите може да помогне да се справиме со поплавата на податоци. Постојат голем број на техники за визуелизација на информации кои се развиени во последните декади а главна цел им е да ги поддржат големите множества на податоци. Улогата на човечкиот фактор во процесот на визуелното податочно истражување е посебно важна, бидејќи на тој начин се комбинира флексибилноста, креативноста и знаењето на луѓето со огромните податочни простори и моќта на пресметувањето на компјутерите. Основната идеја на Visual Data Mining (VDM) е да се претстават податоците во визуелна форма, овозможувајќи корисникот да проникне во податоците, да извлече заклучоци и да врши директна интеракција со податоците. Техниките за визуелна анализа на податоци имаат голема вредност во истражувачката анализа на податоци. Особено е корисна кога е познато многу малку за податоците но, кога целите на истражувањето се познати и треба да се постават хипотези. И верификацијата на хипотезите може исто така да се направи низ визуелизација на податоци. Главни предности на визуелната анализа на податоци е нивната способност да се справат со нехомогени податоци и податоци со шум, нивната интуитивност и тоа што не мора да се разберат комплексностите. Исто така, техниките за визуелно истражување на податоци (VDE) овозможуваат повисок степен на доверба во истражувањето поради вклученоста на човекот.

Визуелното истражување на податоци претставува процес што се одвива во три чекори: преглед, зумирање и филтрирање, и добивање на

бараните детали. Ben Shneiderman, го нарекол процесот информациска пребарувачка мантра [3]. Во анализирањето на големи множества на податоци, корисникот прво треба да добие целосен преглед на податоците. Во прегледот, тој идентификува интересни модели на групи на податоци и се фокусира на еден или повеќе. Тоа се овозможува со доделување на голем процент на дисплејот на множеството од интерес, додека се намалува користењето на екранот за податоците што не се од интерес. Технологијата на визуелизација овозможува основни техники за визуелизација во сите три чекори, но во исто време го премостува јазот меѓу трите чекори [1]. Во првиот дел на трудот се опишани видовите податоци, а во вториот техниките кои се користат во процесот на визуелизација. Во третиот дел се опишани техниките кои се користат за интеракција со податоците со цел да се направи подетална анализа.

2. КЛАСИФИКАЦИЈА НА ТЕХНИКИТЕ ЗА ВИЗУЕЛНА АНАЛИЗА НА ПОДАТОЦИ

Техниките за визуелизација на податоци како x-y графикони, линиски графикони и хистограми се многу корисни за анализа и истражување на податоците, но се ограничени на релативно мали и ниско-димензионални множества на податоци. Во последната декада развиени се голем број нови техники за визуелизација, кои овозможуваат визуелизација на мултидимензионални множества на податоци. Техниките може да се класифицираат на основа на три критериуми: податочните видови кои се визуелизираат, техниките за визуелизација и техниките за интеракција.

Видовите на податоци може да бидат:

- Еднодимензионални податоци
- Дводимензионални податоци
- Мултиудимензионални
- Текст
- Хиерархиски структури
- Алгоритми и софтвер.

Техниките за визуелизација може да се класифицираат како:

- Стандардни 2D/3D прикази
- Геометриски трансформирани прикази
- Иконично-базирани прикази
- Пиксел базирани техники
- Напластени (stacked) прикази

Третата димензија на класификацијата се техниките за интеракција. Техниките за интеракција овозможуваат корисниците директно да навигираат и ја модифицираат визуелната претстава. Овие техники вклучуваат:

- Динамички проекции
- Интерактивно филтрирање
- Зумирање на регионот од интерес
- Дисторзија или нерамномерно зголемување на регионот од интерес
- Селектирање и поврзување на податоци прикажани во повеќе проекции

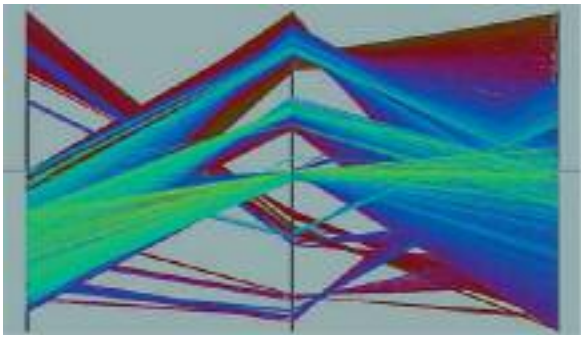
Сите три димензии на оваа класификација се ортогонални, т.е. која било од техниките за визуелизација може да се употреби заедно со која било техника за интеракција за кој било вид на податоци.

3. ВИДОВИ ПОДАТОЦИ

Во визуелизацијата на информации, податоците обично се состојат од голем број записи. Секој запис содржи различен број на променливи и димензии и одговара на едно набљудување, мерење или трансакција. Бројот на атрибути може да се разликува од едно множество на податоци до друго, во зависност од тоа со колку атрибути е опишан ентитетот. Бројот на променливите се нарекува димензионалност на множеството. Според димензионалноста, податочните множества се делат на еднодимензионални, дводимензионални или мултидимензионални. Податоците се нарекуваат и универијанти, биваријантни и мултиваријантни. Покрај оваа поделба денес од голем интерес е визуелизацијата на податоци кои се хиерархиски организирани, како хипертекстовите.

Многу множества на податоци во информациските системи вообичаено претставуваат повеќедимензионални податоци за кои не е можна едноставна визуелизација во 2D или 3D простор. Такви мултидимензионални (или мултиваријантни) податоци претставуваат табелите од релациските бази кои најчесто имаат десетина или стотици колони (или диманзии).

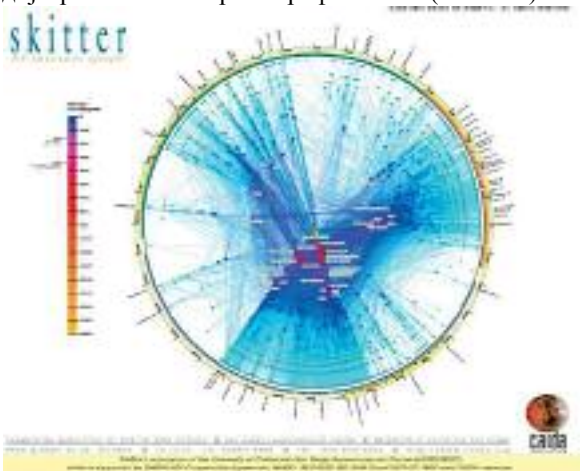
Кога не е можно да се реализира едноставно пресликување на податочните димензии во двете димензии на екран, потребни се пософистицирани техники за визуелизација. Една таква техника што овозможува визуелизација на мултидимензионални податоци е техниката на паралелни координати која ги прикажува податочните членови како множество на линиски сегменти кои ја сечат секоја паралелна оска на позиција која соодветствува на вредноста на атрибутот репрезентиран со оката. На слика 1 е прикажана визуелизација добиена со паралелни координати,



Сл. 1 – Техника на паралелни координати

Во ерата на WWW важни видови на податоци се хипертекстот и мултимедијалните Web содржини. Овие видови податоци се разликуваат бидејќи не можат лесно да се опишат со броеви и на нив не може да се применат стандардните техники за визуелизација. Во овие случаи, потребна е трансформација на податоците во описни вектори пред да се примени некоја техниката за визуелизација. Пример на едноставна трансформација е броењето на зборови [ThemeRiver], кое често се комбинира со техниките за статистички анализи како анализа на главните компоненти (PCA) или мултидимензионалното скалирање за да се намали димензионалноста на две или три димензии.

Кога податоците имаат врски со други делови на податоците или информации, станува збор за хиерархиски или мрежни модели на врски и во овој случај се користат специфични техники да се изразат ваквите меѓузависности. Овие техники подразбираат рендерирање на граф што се состои од множество на објекти, наречени јазли, а врските меѓу објектите се наречени линкови или врски. Примери за ваков вид на податоци се e-mail врските, човечките потрошувачки навики, структурата на датотеките на хард-дискот и хиперлинковите во www. Од многуте специфични техники за визуелизација кои се справуваат со хиерархиите и графичките податоци, може да ги споменеме интернет хиерархиите, дрво-дијаграмите и интернет графиконите (Слика 2).



Сл. 2 – Skitter-graph на интенет мрежа која покажува глобална структура на интернет („Skitter Graph“). Јазлите кореспондираат со ISP.

Други класи на податоци се алгоритмите и софтверот. Целта на визуелизацијата на софтверот е да го поддржи развојот на софтверот со помагање да се разберат алгоритмите, со тоа што ќе се прикаже протокот на информации во програмите. Притоа треба да се претпостави структурата на стотици изворни линии и да се претстават како графици. Цел на овој вид на визуелизација е и поддршката на програмерите во дебагирањето т.е. визуелизација на грешките [4].

4. ТЕХНИКИ ЗА ВИЗУЕЛИЗАЦИЈА НА ПОДАТОЦИ

Голем број техники за визуелизација може да се употребат за визуелизација на податоци, пред се, стандардните 2D/3D техники како x-y (x-y-z) графикони, бар графици, линиски графици, карти [2]. Но, постојат многу софистицирани техники за визуелизација, кои одговараат на основните принципи на визуелизација, а кои може да се комбинираат да се имплементира специфичен систем за визуелизација.

4.1. Геометриски трансформиран приказ

Овие техники се користат да помагнат во наоѓање на „интересни“ трансформации на мултидимензионални множества на податоци. Класата методи на геометриски прикази вклучуваат техники од експлораторна статистика, како матрици со точкасти графици (scatterplot matrices) и техники кои се означуваат како „projection pursuit”, итн. Други техники на геометриски проекции вклучуваат проекциски погледи (prosection view) и паралелни координати. Техниката на паралелни координати пресликува k-димензионален простор во две димензии со користење на k-паралелни оски. Оските одговараат на димензиите и се линеарно скалирани од минималната до максималната вредност на соодветната димензија. Секој податочен член е претставен со низа поврзани линиски сегменти, пресекувајќи ја секоја од оските на локацијата која соодветствува на вредноста на димензијата што ја претставува оската (Слика 1).

4.2. Прикази со икони

Класични техники за визуелно податочно истражување се и приказите со икони. Идејата е да се означат (мапираат) вредностите на атрибутите на мултидимензионалните податоци во својствата на иконата. Иконите може да се дефинираат произволно – на пример, да претставуваат мали ликови, иглички, sвезди стапчести фигури, итн. Слика 3 претставува пример на оваа класа на техники. Секоја податочна точка е претставена со sвезда, каде секоја податочна димензија ја контролира должината на зракот кој произлегува од центарот на иконата.

4.3. Пиксел базирани прикази

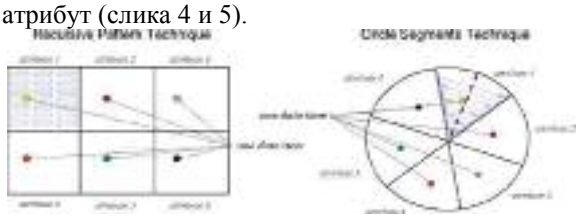
Идејата на пиксел базираните техники е да се означи секоја вредност на димензијата со обоен пиксел и потоа да се изврши групирање на пикселите. Овој приказ користи пиксел за секоја вредност на податокот. Ако секоја вредност на податокот е претставена со еден пиксел, основното прашање е како пикселите се поставени на екранот. Техниката користи

различни подредувања да се овозможат детални информации на локални корелации, зависности и клучни места.

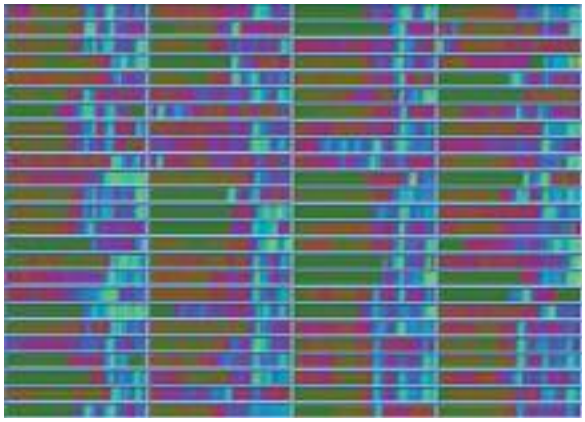
Добро познати подредувања се рекурзивните модели и техниките на кружни сегменти кои се базирани на генетички рекурзивни „напред-назад“ подредувања на пикселите на еден атрибут .

Слика 3 Техника на ѕвездести икони

Корисникот може да ги специфицира параметрите за секое рекурзивно ниво и да ги контролира поставувањата на пикселите за да се формира семантички значајна подструктура. Основен елемент на секое рекурзивно ново е шаблонот (правоаголник) со висина h_i и широчина w_i, (слика 4). Кај техника на кружни сегменти, податоците се поставуваат во круг, од центарот кон надвор. Кругот е поделен на сегменти, по еден за секој атрибут (слика 4 и 5).



Слика 4 Рекурзивни модели и техники на кружни сегменти



Сл. 5 – Техника на резкурзивни модели на густ дисплеј со податоци на Франфуртската берза за 100 производи

4.4. Мапи на хиерархии

Мапите на хиерархии се направени за да прикажат партиционирање на хиерархиски начин. Кога податоците се мултидимензионални, димензиите треба соодветно да се изберат и да се прикажат димензиите. Основната идеја е да се вгнезди еден координатен систем во друг, два атрибути да се вгнездат во друг координатен систем и т.н.

Приказот се генерира со делење на координатниот систем во правоаголни келии. Внатре во келијата, двата следни атрибути се искористени за да се скрати второто ниво на координатен систем. Овој процес може да се повторува повеќепати. Прво треба да се извлече најважната димензија. Визуелизацијата на хиерархиска мапа за истражување на податоци е покажана на слика 6. Други примери на техники на хиерархиски мапи вклучуваат светови-во-светови (worlds-within-worlds), мапи на дрва (treemap), конусни дрва (cone trees) и други техники.

Слика 6 Димензионално пластење на drill-mining

5. ТЕХНИКИ ЗА ИНТЕРАКЦИЈА

Техниките за интеракција се користат како додаток на техниките за визуелизација, за ефективно истражување на податоците. Тие овозможуваат аналитичарите на податоци да вршат директна интеракција со визуелизацијата и динамички да ја менуваат во согласност со објектите на истражување. Техниките за интеракција може да се поврзат и комбинираат со која и да е техника за визуелизација.

Техниките за интеракција може да се категоризираат врз основа на ефектите што ги предизвикуваат на дисплејот. Навигациските техники се фокусираат на модификација на проекцијата на податоците. Методите за контрола на погледите што овозможуваат порамнување на нивото на детали на целата визуелизација или на дел од податоците. Техниките за селекција овозможуваат изолирање на подмножество прикажани податоци за операции како нагласување, филтрирање и квантитативна анализа. Изборот може да се направи директно на визуелизацијата (директна манипулација) или низ дијалог прозорци и други механизми (индиректна манипулација).

4.1. Динамичка проекција

Основната идеја на динамичката проекција е автоматски да се менуваат проекциите, со цел да се истражат мултидимензионални множества на податоци како добро одредени кластери. Добро познат пример е системот GrandTour [10], во кој множества на мултидимензионални податоци се прикажуваат како дводимезнионални проекции на

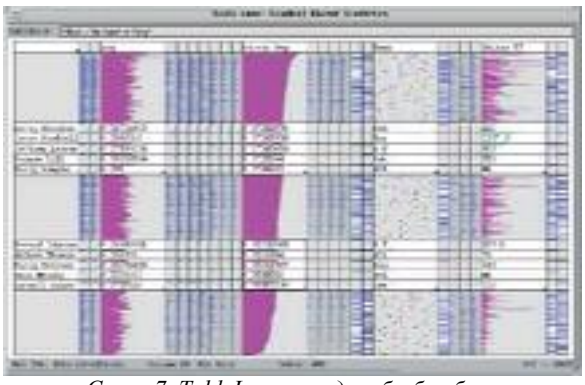
точкасти графици. Со софтверот се подджани се и динамички проекции кои вклучуваат техники како XGobi, XLispStat и ExplorN [10,6,7].

4.2. Интерактивно филтрирање

Интерактивното филтрирање е комбинација на избор на погледите при истражувањето на големи множества на податоци и техники за интеракција, каде интерактивно се избира и партиционира множеството на податоци во сегменти и се фокусира на интересни подмножества. Изборот на поглед може да се реализира со директен избор на подмножеството податоци (browsing) или со спецификација на својствата на посакуваното подмножество (query). Алатка која може да се користи за интерктивно филтрирање е MagicLens. Алатката се користи како лупа за да се филтрираат податоците директно во визуелизацијата. Податоците под лупата се прикажуваат поинаку од останатото множество на податоци, односно MagicLens покажува модифициран поглед на селектираниот регион, додека остатокот на визуелизацијата останува непроменет. Може да се користат неколку различни филтри, или нивна комбинација. Пример на техники на интерактивно филтрирање и алатки се InfoCrystal, DinamicQuery и Polaris [8].

4.3. Зумирање

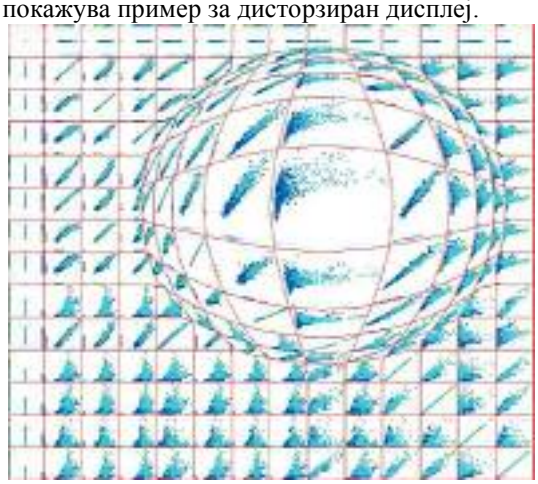
Зумирањето е добро позната техника за модифицирање на погледи која се справува со големи количества на податоци. Кај оваа техника е важно да се претстават податоците во високо компресирана форма и во различни резолуции. Зумирањето не значи зголемување на објектите, туку нивно прикажување со повеќе детали. Објектите може да бидат претставени како пиксели на ниско зум ниво или икони на средно ниво на зумирање но и како означени објекти со висока резолуција на пример, TableLens методот [9]. Добивањето на прегледи на вакво множество на податоци е тешко ако податоците с претставени во текстуална форма. Иницијалниот поглед им овозможува на корисниците да откријат модели, корелации и екстреми во множеството на податоци, а корисникот може да зумира подмножество што ќе се прикаже со повеќе детали. На слика 7 е прикажан пример за TableLens приод. Други примери на техники и системи кои користат интерактивно зумирање се PAD++, IVEE, SpotFire, DataSpace [4,6].



Слика 7 TableLens приод на безбол база

4.4. Дисторзија

Дисторзијата е техника за модификација на погледите која го поддржува процесот на истражување на податоците со зачувување на погледот на податоците за време на drill-down операциите. Основната идеја е да се покажат дел од податоците со многу детали, сите останати податоци со помалку детали. Популарна дисторзиона техника се хиперболичките и сферните дисторзии. Тие често се користат при визуелизација на хиерархии, но може да се употребат, исто така, во која било друга техника за визуелизација. Примери за дисторзиони техники вклучуваат бифокален приказ, перспективен ѕид, рибина перспектива, хиперболична визуелизација и хиперкоцки. Слика 8 покажува пример за дисторзиран дисплеј.



Слика 8 Дисторзиран приказ со рибина перспектива

4.5. Селекција и поврзување

Техниката најчесто се применува кога множеството податоци се визуелизира со користење на повеќе дводимензионални погледи. Селекцијата е процес на интерактивна селекција на податок во некој од погледите на визуелизацијата, што потоа се прикажува во сите останати погледи на визуелизацијата на множеството податоци. Идејата на поврзување и селектирање е да се комбинираат различни методи на визуелизација за да се надминат недостатоците на индивидуалните техники. Пример се матриците од точкати графици што

може да се комбинираат со боење на селектираниот податок во сите можни проекции на визуелизацијата. Како резултат, избраните точки се потенцирани, откривајќи ги зависностите и корелациите меѓу податоците, односно овозможувајќи повеќе информации отколку ако техниките се користат независно. Типичен пример на вакви техники за визуелизација се матрици со точкасти графици, бар графици, паралелни координати, пиксел базирани прикази и мапи. Голем број интерактивни системи за истражување на податоци го овозможуваат овој метод, како на пример алатките SPlus, XGobi, XmdvTool, DataDesc [5,6].

6. ЗАКЛУЧОК

Истражувањето на големи множества на податоци е важен но тежок проблем. Визуелното истражување на податоците, и многу примени како откривање на грешки и DM може да користи техники за визуелизација на информации за подобрување на анализата на податоците.

Постапката вклучува интеракција на техники за визуелизација со традиционалните техники какви што се статистиката, машинското учење, операционите истражувања и симулациите. Интеграцијата на техниките за визуелизација и овие методи овозможува развој на автоматизирани алгоритми за анализа на податоци притоа подобрувајќи го квалитетот, брзината и процесот на анализа на податоците. Техниките за визуелна анализа на податоци треба да бидат интегрирани со системите кои се користат за управување со големи количества на врски и полуструктуирани информации, кои вклучуваат управување со базите на податоци и со data warehouse системите. Нивната цел е да се донесе моќта на техниките за визуелизација на секој десктоп и да се овозможи подобро, побрзо и поинтуитивно истражување на податоците на многу големи податочни ресурси. Ова не е важно само во економска смисла туку и за да се постигне задоволство на корисникот

7. ЛИТЕРАТУРА

[1] Berthold M, Hand D.J.,: *Intelligent Data Analysis, Chapter 11,* Daniel Keim, Matthew Ward, Springer, 2007.

[2] Tagarden D.P., *Business Information Visualization,* Tutorial, CAIS, 1999

[3] B. Shneiderman. *The eye have it: A task by data type taxonomy for information visualizations. In Proc. Visual Languages, 1996*

[4] B. Price, R. Baecker, and I. Small. *A principled taxonomy of sofware visualization.* Journal of Visual Languages and Computing, 4(3):211-266, 1993.

[5] W. J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective.*Number 3 in Oxford Statistical Science Series. Oxford University Press, Oxford, 1988.

[6] B. Shneiderman. *Tree visualization with treemaps: A 2D space-filling approach,* ACM Transactions on Graphics, 92-99, 1992

[7] M. Ankerst, M. Breunig, H. Kriegel, and J. Sander. OPTICS: *Ordering Points To Identify the Clustering Structure.* In Proc. ACM SIGMOD'99, Int. Conf on Management of Data, Philadelphia, PA, pages 49-60, 1999.

[8] http://www.polaris.co.in/ http://www.scils.rutgers.edu/~aspoerri/InfoCrystal/InfoCrystal.htm

[9] http://www.avizsoft.com/contents/table-lens-demo.htm

[10] http://wareseeker.com/free-grandtour/

Summary

VISUALIZATION, TOOL FOR INTELLIGENT DATA ANALYSIS

Snezana Savoska[1], Suzana Loskovska[2]

[1] Faculty of Administration and Management Information systsems, Ss. Climent Ohridski University – Bitola, Macedonia, savoskasnezana @yahoo.com

[2] Faculty of Electrical Engineering and Information Technologies – Skopje, Karpoš II bb, 1000 Skopje, Macedonia, suze@feit.ukim.edu.mk

***Abstract* – The role of human being in the data visualization process is important because in this way we can combine the flexibility, creativity and general knowledge of the human with the enormous storage capacity and the computational power of today's computers. For these reasons, data visualization is one of the most popular techniques of data analysis. This paper presents an overview of information visualization, visual data exploration, classification of data that has to be visualized, the visualization techniques and the interaction techniques.**

***Keywords* – Visualization, Visual data analysis, Visual data exploration**