Визуелизација на мултидимензионални податоци

Снежана Савоска[1], Виолета Маневска[2]

[1)] snezana.savovska@uklo.edu.mk, [2)] violeta.manevska@uklo.edu.mk

Апстракт

Мултидимензионалните податоци се секојдневна реалност во ерата на информации. Ваквите видови податоци секојдневно се собираат од сензорите, читачите на картици, автоматски или со рачно внесување. Се складираат во различни формати, во релациони табели и складишта на податоци. Овие податоци се чуваат и анализираат бидејќи луѓето знаат дека се корисни за предвидување на иднината. За да се анализираат, потребно е да се користат полуавтоматски методи кои вклучуваат софтвер за визуелизација и експлорација и човечкото око како алатка за препознавање на модели. Целта на овој труд е кратко претставување на методите за визуелизација на мултидимензионални податоци, како и запознавање со развојот и најцитираната таксиономија на овие методи за визуелизација на мултидимензионални податоци.

Клучни зборови: Визуелизација, Визуелна експлорација, Мултидимензионлни податоци

ВОВЕД

Појавата и екстремно брзиот развој на персоналните компјутери овозможија огромни бенефиции за корисниците. Најважна корист е можноста за постојан пристап до информации кои ги има како никогаш порано и тоа 24 часа на ден, 7 дена во неделата преку интернет. Но, во природата на човекот е секогаш да бара повеќе. Во случајов, ги имаме податоците но да се најдат информации во податоците често е екстремно тешка задача. Во овој труд ги разгледуваме само нумеричките податоци, кои се можеби оние во кои најчесто бараме информации. Тие се често повеќе-димензионални (пример висина, тежина, години, состојба во

банките и т.н.) и секое од овие полиња чува една димензија и сите димензии заедно формираат повеќе-димензионален вектор кој го опишува секој поединечен податок. Задачата може да се дефинира како идентификување на групи на единки кои делат некои заеднички својства. За да се добие решение на оваа задача, мора да најдеме начин да ја идентификуваме структурата и границите во кои се движат димензиите. Ова може да се постигне со истражување со методите за полуавтоматска идентификација на групи во податоците, кои методи се состојат во креирање на симбиоза на човечкиот ум и можностите на софтверот за истражување на податоци и визуелизација.

Компјутерите и луѓето заедно можат многу подобро да ја идентификуваат структурата на податоците ако тие се мултидимензионални. Софтверот одлично се справува со големи количества податоци и може да манипулира со нив, но луѓето се многу добри во идентификување на модели. Софтверот кој манипулира со мултидимензионалните податоци им ги претставува на луѓето податоците во визуелна форма и на начин кој го олеснува препознавањето на модели, т.е. моделите полесно се идентификуваат од човечкото око. Пример за претходно наведеното се Андреовите криви во кои секоја податочна точка е претставена со крива. Човекот може да идентификува поединечни области на кривите кои се оптимални за идентификување на кластерите низ множестата на такви податоци. Подвижните 3D слики може да се креираат за да се видат облаци од податочни точки кои се движат како што се движиме низ кривите. Многу е реално, податоците кои се движат заедно, да се членови на ист кластер.

Некои методи кои ги користиме за наоѓање на структури (т.е. делот софтвер од партнерството човек-софтвер) се базирани на невронски мрежи, кои емулираат човечки неврони. Сепак ова не значи дека софтверот е способен да ги преземи човечките задачи. Иако оваа претпоставка може да има некој смисол, заедничките искуства покажуваат дека тоа е невозможно. Целта е да се најде интелигентен софтвер кој ќе помогне да се најдат структури во мултидимензионалните множества на податоци кои на човечкото око ќе им го олеснат процесот на пронаоѓање на модели.

Историја на мултидимензионалната визуелизација на информации

Корените на визуелизацијата на информации, како практично поле, може да се бараат во работата на Tukey, Bertin и Tufte[1] кои се фокусирале на 2D и 3D визуелизација. Тие ги дале основните правила за визуелизација во рамнина, композициите на бои кои се користат, ги дефинирале означувањата на атрибутите како и други детали. Но, основата на концептот на мултидимензионалните техники за визуелизација лежи во користењето на релационите табели од базите на податоци и атрибутите кои тие ги чуваат во форма на колони, а кои се всушност димензии на мултидимензионални податоци. Сепак, потребно е да нагласиме дека проучувањето на мултидимензионалната визуелизација започнува неколку века пред откривањето на релационите техники на бази на податоци [3]. Според Wong и Bergeron [2,9], еволуцијата на оваа дисциплина може да се подели во 4 периоди [3]:

1. Период на состојба на пребарување (од 1782 до 1976) која се карактеризира со релативно мали големини на податоците и алатки за визуелизација кои се состојат од колор молив и графичка хартија.

2. Период на состојба на будење (од 1977 до 1985) каде дво или тро-димензионите просторни податоци се проучуваат како најважни типови на податоци, а мултидимензионалните податоци веќе почнуваат да го добиваат потребното внимание.

3. Период на состојба на откривање (од 1986 до 1991) која се карактеризира со ограничена достапност до графичкиот хардвер и софтвер (брзи и скапи графички станици поседуваат само мал број компании со mainframe компјутери). Повеќето од методите за визуелизација кои денес се познати, развиени се токму во овој период.

4. Период на состојба на истражување, развивање и проценка на софтвер (од 1992 до денес) кој период се карактеризира со брз и интензивен развој на нови техники за визуелизација и визуелна експлорација на податоци.

Таксиономија (основи за класификација) на методите на визуелизација

Не постои единствен договор и консензус за најдобра таксиономија помеѓу научниците кои се занимаваат со проблемот на визуелизација. Сепак, најелаборирани и најцитирани се следните поделби на техниките за мултидимензионална визуелизација:

1.Таксиономија на техниките според Shneiderman

2.Таксиономија на техниките според Keim

3.Алтернативна таксиономија

I. Класификација според Shneiderman

Како поткрепа на визуелната информациона мантра „Прво преглед, зум и филтер и потоа детали по барање“[4], Shneiderman, предлага дефинирање на седум задачи при визуелизацијата на информации[3]. Тие задачи се:

1.Преглед: Добивање на глобален преглед на целото множество податоци.

2.Зум: Зумирање на податочните членови кои се од интерес, со што се овозможува подетален поглед на посакуваните податоци.

3.Филтер: Филтрирање на интересните групи на податоци и исфрлање на неинтересните групи или членови при што се прави редукција на големината на пребарувањето.

4.Детали по барање: Избор на податочни членови или групи и добивање на детали кога се потребни.

5.Поврзаност: Се осознаваат релациите помеѓу членовите.

6.Историја: Се чува историјата на сите дејствија за да се поддржат операциите undo, replay и прогресивно прочистување на податоците со кое се овозможува undo на грешките или пак некоја серија на чекори е заменета со друга.

7.Екстракција: Овозможува екстракција на групи и подгупи на податоци и нивно снимање, печатење или префрлање во други задачи или делови од задачи.

Седумте разгледувани типови на податоци кои се визуелизираат може да бидат:

а) Еднодимениионални податоци, кои може да бидат линеарни типови на податоци што вклучуваат текст документи, изворни програмски кодови, листи на имиња по алфабетски ред и др.

б) Дводимензионални податоци, како рамнини или дводимензионални цртежи кои вклучуваат географски карти, скици и планови или планирање на весници.

в) Тродимензионални податоци кои се реални објекти како молекули, човечко тело, згради, предмети. При оваа претстава, кога се гледаат објектите, покрај перцепцијата за самите објекти, корисниците треба да ја разберат нивната позиција и ориентација во просторот.

г) Мултидимензионални податоци се податочни членови со n-атрибути кои стануваат точки во n-димензионален простор.

д) Темпорални податоци одразуваат промени во време. Овие податоци се разликуваат од еднодимензионалните податоци по тоа што имаат време на почеток и крај, а и членовите – темпоралните податоци може да се преклопуваат.

ѓ) Податочни структури во форма на дрвја се збирки на членови од кои секој има врска со родителскиот член. Членовите и врските помеѓу родителот и детето може да имаат повеќе атрибути.

е) Податоци во мрежи каде членовите се поврзани со произволен број на други членови и врските помеѓу членовите не може да се спознаат со дрво структура.

II. Класификација според Keim

Keim и Kriesel [5], ги категоризирале техниките за визуелизација на следниот начин:

а) Геометриски техники чија цел е да се најдат „интересни“ трансформации на мултидимензионалното множество на податоци. Оваа класа техники вклучува матрици на точкасти цртежи, анализа на основните елементи, анализа на фактори, мултидимензионално скалирање, техника наречена Projection Pursuit и паралелни координати.

б)Техниките базирани на икони претпоставуваат дека секој мултидимензионален член е означен со икона или глифа. Количината на

информации која е возможно да се визуелизира со оваа техника во исто време е ограничена и зависи од карактеристиките на иконата. Примери на вакви техники се Чероф (Chernoff Herman) ликови, звездести икони, колор икони и икони со стапчести фигури.

в) Пиксел-ориентирани техники се техники каде атрибутите на податоците се означени со пиксели. Бојата на секој пиксел зависи од вредноста на атрибутот. Важен аспект е просторната дистрибуција на пикселите. Примери за овие техники се рекурзивните модели, техниките на кружни сегменти и спиралните техники.

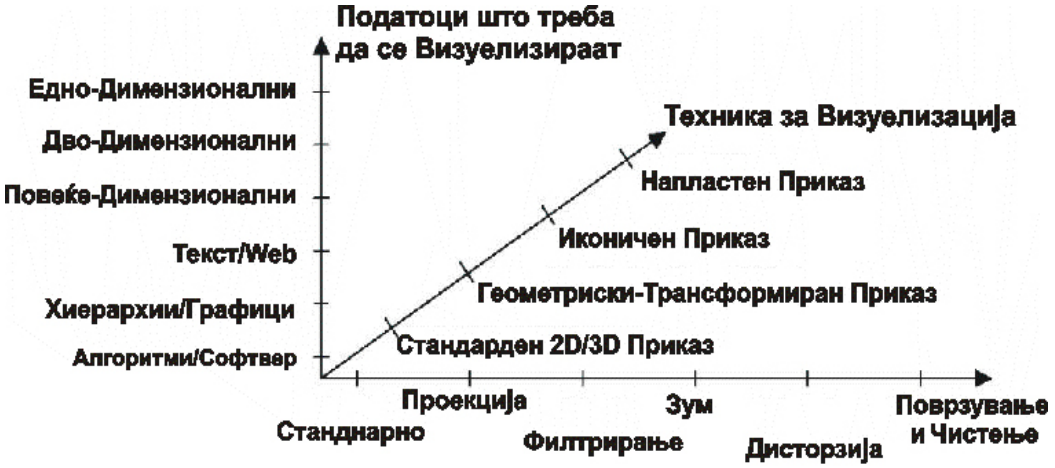
г) Хиерархиски техники претпоставуваат делење на мултидимензионалниот простор на хиерархиски подпростори. Примери за вакви техники се димензионално пластење и светови-во-светови.

д) Техники базирани на графови се специјализирани техники за претставување на големи графици со користење на специфични алгоритми за скицирање, query јазици и апстрактни техники. Пример за ваква техника е Treemap.

Оваа класификација според Keim е проширена со два ортогонални критериуми за поделба. Првиот критериум е избор на техника за дисторзија (изобличување), а вториот е избор на техника за интеракција. Првиот избор претпоставува приказ на дел од податоците со различно ниво на детали. Вториот избор овозможува корисниците да можат директно да вршат интеракција со податоците во самата визуелизација. Сите три критериуми на класификација може да се прикажат како ортогонални оски на системот на класификација (Слика 1).

Слика 1: Класификација на техниките за визуелизиција на информации според Keim [3],[7]

Како техниките за визуелизација на информации кореспондираат на податоците кои се визуелизираат и техниките за интеракција и дисторзија, е прикажано на Слика 2.



Слика 2: Класификација на техниките за визуелизација на информации според видот на податоци и техниките за интеракција и дисторзија [7]

III. Алтернативна таксиономија

Во практиката и во алатките кои се достапни за визуелизација на информации од повеќето производители на софтвер за визуелизација[2,3], методите за визуелизација ги групираме според следната класификација:

Методи на линеарна проекција каде се употребува пониска димензионална репрезентација на податоците со користење на линеарна проекција на повеќе-димензионалниот простор. Некои примери се анализа на основните компоненти (PCA), истражувачки Projection Pursuit (EPP) и матрица на точкасти цртежи.

Методи на тополошко зачувување се методи на нелинеарно поврзување на повеќе-димензионалниот простор со ниско-димензионалната претстава на податоците при канонична анализа на компонентите (CCA), самоорганизирачки пресликувања (мапи - SOM) и спирални (Spring) методи.

Методи на високо-димениионални претстави се методи каде нема намалување на димензиите и се користат сите димензии за да се произведе графичка претстава. Пример за вакви техники се Черноф ликови, паралелни координати, Андриеви криви и мултидимензионално пластење.

Grand Tour се методи каде место статичка претстава на податоците се прикажуваат секвенци од проекции за да се проучи структурата на податоците. Алгоритмите кои се користат за да се добие ваква секвенца се Asimov-Buja алгоритмот за „навивање“, алгоритмот за случајни криви, алгоритмот на поделени криви и привидна Grand Tour метода [1,3,7]. Карактеристично е што во сите алгоритми е искористена човечката способност за визуелно препознавање на модели.

Заклучок

Напредувањето на истражувањата на техниките за визуелизација на мултидимензионални податоци се незапирливи. Тие се поместуваат надвор од дизајнирањето на визуелните прикази на податоците и одат во правец на визуелна експлорација на податоци. Техниките кои се имплементирани во повеќето алатки за експлорација и визуелизација на мултидимезнионални податоци одат кон зголемување на точноста на резултатите од експлорацијата, зголемена продуктивност и подобро разбирање на информациите во податоците.

Научниците треба да се справат со огромни количества податоци кои се многу илјадници пати поголеми од бројот на пиксели кои може да се прикажат на дисплејот. Поради тоа техниките за визуелизација постојано ја менуваат визуелна перцепција и начинот на визуелна анализа. Научните бази, мултимедијалните системи, виртуелната реалност и мултидименионалната анализа бараат се помоќни и помоќни алатки кои имаат тенденција да ги сменат начините на визуелно размислување [9].

Користена литература

[1] Series editors, (2005, *Morgan Kaufman,), Information Visualization*
[2] Ward M., Yang J., (2003, *EUROGRAPHICS, Volume 22 Number 3), Interaction Spaces in Data and Information Visualization*
[3] Series editors, (2009) *Multidimensional Visualization methods,* http://pisuerga.inf.ubu.es/cgosorio/Visualization/?Multidimensional_visualization_methods ; http://www.cmsimple.com/

[4] Ware C., (2009, *Morgan Kaufman), Visual Thinking for Design*
[5] Wad CV, Rundensteiner EA, Ward MO, *(ACM 9781595936493/ 07/09), Quality Driven Data Abstraction Generation for Large Databases*
[6] Fry B., (2008, *O'Reilly Media), Visualizing Data*
[7]Berthold M., Hand DJ., (2007, *Springer), Intelligent data analysis*, Second edition
[8] Seo J, Shneiderman B., (2004, grant N01 NS-1-2339 from the NIH), *Understanding Clusters in Multidimensional Spaces: Making Meaning by Combining Insights from Coordinated Views of Domain Knowledge*
[9] Wong PC, Bergeron RD, (2003, *University of New Hampshire), 30 Years of Multidimensional Multivariate Visualization*

Multidimensional Data Visualization

Savoska S.

Abstract

Multidimensional data is the reality of the information era. The data is acquired with sensors, card readers in every day's operations, automated or by manual data input. It's collected in different formats, in relational tables and data warehouses. People collect and save data because they believe that it is very useful for gaining information about some future predictions. To analyze data, it is useful to have some semi-automated methods which use visualization and exploration software and also the human eye as a tool for pattern recognition. The aim of this paper is a short presentation of methods of visualization of multidimensional data history and the most citation taxonomy of these methods.