Проблемы Визуализации Многомерных Данных

Айтадж Агабейли
Институт Информационных Технологий НАН Азербайджана, Баку, Азербайджан
agabeyli96@gmail.com

Аннотация **– в этой статье описывается термин Big Data в аспектах представления и визуализации данных. Существуют определенные специфические проблемы визуализации больших данных, мы постарались определить эти проблемы и совокупность подходов, помогающих их избежать. Кроме того, сделан обзор существующих методов визуализации применительно к многомерным данным и с учетом описанных проблем. А также была представлена классификация методов визуализации для работы с большими данными.**

Ключевые слова – большие данные, 2D, 3D, визуализация, график, диаграмма, представление.

I. ВВЕДЕНИЕ

Обработка больших данных не тривиальная задача и требует специальных методов и подходов. Графическое мышление очень простой и естественный способ обработки данных для человека, изображение является эффективным методом представления, который позволяет объективно оценить результаты и помогает в принятии правильного решения задач. Но в случае с большими данными множество из классических методов представления становится менее эффективным или даже неприменимым для конкретных задач. Необходимо классифицировать существующие методы визуализации по критерию их применимости к тому или иному классу больших данных.

Для принятия решения и сортировки описанных классов многомерных данных необходимо проанализировать следующие характеристики: применимость для большого объема данных, возможность визуализации данных, представленных в разных форматах, скорость и производительность представления данных.

Визуализация информации – не такая новая область. Первым «визуализатором», по мнению автора [1], стал математик и астроном Урбен Леверье, наиболее известный своим «открытием Нептуна на кончике пера».

Визуальная передача информации известна человеку с тех пор, как мы начали рассказывать друг другу истории. Визуализацией является любая техника создания изображений, диаграмм, карт, таблиц или анимаций. Визуальные образы были эффективным способом общения с древнейших времен. Примеры из истории включают наскальные рисунки, египетские иероглифы, греческую геометрию, революционные методы технического рисования Леонардо да Винчи для инженерных и научных целей.

Автор книги Handbook of Data Visualization Michael Friendly отметил основные опорные точки истории визуализации. Это ранние карты и диаграммы, измерения и теории, новые графические формы, начало современной графики, золотой век статистической графики, смутные годы, возрождение визуализации информации, интерактивная и динамическая визуализации [1].

II. ТИПЫ ВИЗУАЛИЗАЦИИ БОЛЬШИХ ДАННЫХ

В литературе [2−6] были представлены следующие типы визуализации:

A. Линейные графики (line charts)

Линейный график, или line chart, показывает отношение одной переменной к другой. Такие графики наиболее часто используются для отслеживания изменений, происходящих в течение долгого времени. Линейные графики также помогают при сравнении нескольких объектов за один и тот же промежуток времени. Уложенные линии используются для сравнения значений тенденции изменения индивидуальных значений или нескольких переменных. Линейные графики используются, когда необходимо визуально изобразить изменение одной или нескольких переменных, и при изображении тренда или скорости изменения информации о значениях этих переменных [2].

Важно также отметить, что не обязательно выбирать линейный график только потому, что имеется определенное количество точек данных. Для этого можно использовать более простое отображение, например, просто перечислить их в определенном порядке с использованием таблиц. Линейная диаграмма используется тогда, когда необходимо не просто изобразить, а показать взаимосвязь между точками данных.

B. Бар графики (bar charts)

Гистограммы обычно используются для сравнения количественных показателей различных категорий или группы данных. Значения категории представлены в виде баров, они могут быть сконфигурированы либо вертикально, либо горизонтально, а высота или длина полос соответствует определенным значениям данных.

Если эти значения достаточно разнятся и отличия в барах могут быть обнаружены человеческим глазом, можно использовать простую диаграмму. Когда значения баров очень близки друг к другу, становится трудно сравнивать стержни между собой. Чтобы помочь обеспечить визуальный контраст, бары могут иметь различные цвета.

Другой формой гистограммы является прогрессивная диаграмма или waterfall (водопад) график. Водопад-график показывает, как начальное значение данных увеличивается или уменьшается в течение ряда операций или действий. Первый бар начинается в начальном значении, и каждый последующий начинается там, где заканчивается предыдущий бар. Длина и направление бара указывают величину и тип (положительный или отрицательный) действий. Каскад, образующийся в результате диаграммы, покажет, как данное действие или операция приведет к конечному результату [2].

C. Диаграмма рассеяния (scatter plot)

Диаграмма рассеяния (Scatter plot) - это тип математической диаграммы, использующий декартовые координаты для изображения набора данных обычно для двух переменных. Можно увеличить количество отображаемых переменных до трех в случаи, если точки имеют цветовую маркировку. С помощью этих диаграмм можно определить направление и линейность связи между переменными, представленными на диаграмме. С увеличением точек на диаграмме рассеяния увеличивается и степень корреляции.

Как отмечено в публикации [3], диаграммы рассеяния (scatter plots) обычно используются для визуализации многомерных данных. Однако 2D предоставления данных при большом количестве взаимодействий между ними имеют трудности в восприятии. Введенные диаграммы рассеяния, 3D расширений в визуализацию интерактивных данных называются регрессивными кубами (RC). Они значительно увеличивают 3D-диаграммы с тремя аспектами, на которых показаны корреляции между двумя переменными с помощью чувствительных линий и чувствительных потоковых линий (stream lines). Авторы статьи продемонстрировали свою систему двумя примерами и оценкой пользователей, а также показали, как регрессивные кубы дают возможность интерактивного визуального исследования многомерных наборов данных через различные классификации и задачи информационного поиска.

D. Пузырьковая диаграмма (bubble plots)

Пузырьковая диаграмма — это разновидность точечной диаграммы, в которых точки изображены в виде пузырьков, и их размер показывает дополнительные данные. Пузырьковая диаграмма так же, как и точечная, имеют оси значения, которыми являются горизонтальная и вертикальная оси. Значение Z является дополнением к значениям X и Y [2].

Часто пузырьковые диаграммы используют для визуализации финансовых данных. Отличие в размерах пузырьков помогает визуально выделить конкретные значения.

E. Круговые диаграммы (pie charts)

Этот тип диаграмм позволяет графически представить данные как сегменты круга или процентные доли от целого значения. С помощью круговых диаграмм возможно отображение только одного ряда данных, и они показывают отношение размера элементов, образующих ряд, к их сумме. Такие диаграммы используются для изображения составных частей одного целого в виде сегментов круга.

Существует много споров вокруг ценности круговых диаграмм. Могут возникнуть трудности, связанные с интерпретацией оценки результатов, т.к. человеческому глазу сложно сравнивать ломтики круга, близкие по размеру и расположенные не рядом друг с другом. Определенную сложность в оценке и сравнении создает различие углов обзора отдельных сегментов.

В статье [4] показано, что использование круговой диаграммы целесообразно в указанных ниже случаях:

- Нужно отобразить только один набор данных.
- Ни одно из значений, которое нужно отобразить, не является отрицательным.
- Ни одно из значений, которое нужно отобразить, не является нулевым (0).
- Число категорий небольшое.
- Категории представляют части целого круга.

F. Плоское дерево (tree map)

Этот метод показывает весь набор данных в виде элементов, которые являются составляющими иерархического дерева. Они изображаются в виде набора прямоугольников – ветвей дерева, внутри которых находятся дочерние ветви. Прямоугольники отличаются по размеру и цвету соответственно заданным параметрам. Tree map четко показывает отношения данных, но представленных только в определенный момент времени. Например, детальная структура бюджета компании, в котором цветом показан процент изменения каждого пункта по сравнению с предыдущим годом.

G. Секторные диаграммы (sunburst)

Секторные диаграммы (также известные как treepie) являются относительно новым способом представления древовидных структур, часто позиционируемым в качестве альтернативы «плоскому дереву» treemap [5].

Основным отличием между этими методами является то, что переменные параметры отмечаются не как ширина и высота, а как радиус и длина дуги. И это позволяет не

менять всю схему при изменении данных, а только один сектор, содержащий новые данные путем изменения его радиуса. И поэтому этот метод может быть использован для отображения динамики изменения данных, используя анимацию [6].

H. Карты (map)

Часто используемые виды карт это: географическая карта, фотографическая карта, дорожная карта и тематическая карта. Географические карты используются для схематичного изображения географических объектов. Фотографические карты являются фотографиями географических объектов со спутника. На дорожных картах показываются схемы магистралей, трасс, железных и других дорог. На тематических картах объекты изображаются в виде различных маркеров.

III. ПРОБЛЕМЫ ВИЗУАЛИЗАЦИИ БОЛЬШИХ ДАННЫХ

Учитывая свойства больших данных, в [7] были определены следующие проблемы их визуализации:

1. Visual noise / визуальный шум;
2. Large image perception / восприятие большого изображения;
3. Information loss / потеря информации;
4. High performance requirements / высокие требования производительности;
5. High rate of image change / высокая скорость изменения изображения.

1) Визуальный шум

Простая презентация целого ряда данных может создать полный беспорядок на экране, и мы увидим только одно большое пятно, состоящее из точек, представляющих каждую строку данных. Эта проблема состоит в том, что большинство объектов в наборе данных слишком связано друг с другом и на экране наблюдатель не может разделить их в виде отдельных объектов. Так, иногда, анализируя, сложно получить даже немного полезной информации от всей визуализации данных без какой-либо дополнительной обработки. Следует отметить, что в понятие «визуальный шум» не входит любое повреждение или искажение данных, его следует рассматривать как явление потери видимости.

2) Восприятие большого изображения

Следующей проблемой визуализации больших данных является ограничение восприятия слишком крупного изображения. Существует определенный уровень восприятия человеческим мозгом различных визуальных данных. Несмотря на то, что этот уровень для графической визуализации данных значительно выше, по сравнению с визуализацией данных таблицы он имеет свои ограничения. И после перехода этого уровня восприятия человек просто теряет способность приобретать любую дополнительную информацию из перегруженных визуально данных. Все методы визуализации ограничены разрешением технического устройства, которое отвечает за вывод этих данных. Конечно, мы можем заменить устройства на более современные или на группу устройств для частичной визуализации данных, что позволит нам представить более подробное изображение с большим количеством точек данных, но даже если бы мы могли повторить этот процесс бесконечное число раз, мы встретились бы с ограничением восприятия человека. С ростом объема данных, показанных одномоментно, человек сталкивается с трудностями в понимании и анализе этих данных. Таким образом, можно сказать, что методы визуализации данных ограничены не только соотношением и разрешением устройств, но и физическими пределами восприятия.

3) Потеря информации

В связи с вышеизложенным применяются подходы, которые в конечном счете приводят к уменьшению использованных видимых наборов данных. Но, несмотря на решение предыдущего препятствия, эти подходы приводят к появлению другой проблемы, которой является потеря определенного количества информации. Все методы уменьшения визуальной информации производят агрегацию и фильтрацию данных на основе родства объектов в конкретном наборе данных по одному или нескольким критериям. Использование этих подходов может ввести в заблуждение аналитика, который может не заметить некоторые интересные скрытые объекты, а сложный процесс агрегации может потребовать большего количества времени и ресурсов для того, чтобы получить точную и необходимую информацию.

4) Высокие требования производительности

Графический анализ не ограничивается только статической визуализацией изображения, а использует и динамическую визуализацию. Здесь может появиться еще одна проблема, не заметная при статической визуализации. При наличии определенной скорости визуализации появляются требования и к производительности процесса. Процесс анализа определенных данных может занимать много времени при непрерывном увеличении вычислительных ресурсов для фильтрации все большего и большего количества данных.

5) Высокая скорость изменения изображения

Последняя проблема связана с высокой скоростью изменения изображения. Она становится наиболее значимой в процессе мониторинга, когда человек, наблюдающий данные, просто не может реагировать на скорость изменения данных или их интенсивности на дисплее. Снижение скорости меняющихся данных не может обеспечить желаемую эффективность процесса, но скорость реакции человека накладывает определенные ограничения на этот процесс.

В [8] с помощью многих исследований были продемонстрированы преимущества использования прогрессивных иллюминационных моделей в объемных визуализациях. Интерактивная объемная визуализация, делающая этот процесс более эффективным, в качестве инструмента для исследования 3D-данных, включая прогрессивное освещение, была достигнута с ускорением GPU для регулярной решетки объемных данных. В статье [8] представлена интерактивная стратегия иллюминации, которая специально разработана и оптимизирована для объемной визуализации данных неструктурированной решетки. Основой конструкции является дифференциальное уравнение, основывающееся на модели освещения, для того чтобы имитировать распространение, поглощение и рассеяние света в объемной среде. В частности, двухуровневая схема вводится для решения проблем, возникающих в неструктурированных решетках. Результаты испытаний показывают, что добавленные иллюминационные эффекты, такие как глобальная слежка и многократное рассеивание, могут не только привести к визуально более привлекательному представлению, но и существенно повысить восприятие глубины информации и сложных пространственных отношений для людей. усовершенствование объемной визуализации вводится в то время, когда неструктурированные решетки используются в различных научно симулированных приложениях.

ЗАКЛЮЧЕНИЕ

В этой статье мы описали основные проблемы визуализации многомерных данных и проанализировали причины их возникновения. Будущие работы в этой области могут быть проведены по следующим направлениям: исследование методов визуализации и области их применения; принятие решений и рекомендации по выбору методов визуализации для конкретных классов больших данных; формализация требований и ограничений на методы визуализации, применяющиеся к одному или более классам больших данных.

ЛИТЕРАТУРА

[1] M. Friendly, A Brief History of Data Visualization, March 2006, York University.

[2] SAS Institute, Data Visualization Techniques,
http://smartest-it.com/sites/default/files/Data%20Visualization_SAS.pdf.

[3] Y.-H.Chan, C.D.Correa and K.-L.Ma, Regression Cube: A Technique for Multidimensional Visual Exploration and Interactive Pattern Finding, ACM Transactions on Interactive Intelligent Systems (TiiS), April 2014, vol. 4, no. 7.

[4] Cleveland W.S. and R.McGill, Theory, experimentation, and application to the development of graphical methods, Journal of the American Statistical Association, 1984, vol. 79, no. 387.

[5] M.Tennekes and de E.Jonge, Top-down data analysis with treemaps, Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP '11), March 2011, pp. 236–241.

[6] Treemap Visualizations for Analyzing Multi-Dimensional, Hierarchical Data Sets, Panopticon Software,
http://panopticon.com/images/stories/white_papers/wp_treemap_data_visualizations_for_multi-dimensional_data.pdf.

[7] E.Gorodov and V.Gubarev, Analytical Review of Data Visualization Methods in Application to Big Data, Journal of Electrical and Computer Engineering, 2013, pp.7
http://www.hindawi.com/journals/jece/2013/969458/

[8] M.Shih, Y.Zhang and K.-L.Ma, Advanced lighting for unstructured-grid data visualization, IEEE Pacific Visualization Symposium (PacificVis), April 2015, pp. 239-246.

[9] C.Johnson, Top Scientific Visualization Research Problems, IEEE Computer Graphics and Applications, July 2004, vol.24,
http://dl.acm.org/citation.cfm?id=1018051.

[10] N.Cawthon and A.V.Moere, The Effect of Aesthetic on the Usability of Data Visualization,
http://web.arch.usyd.edu.au/~andrew/publications/iv07b.pdf.

[11] J.Heer and B.Shneiderman, Interactive dynamics for visual analysis, Communications of the ACM, 2012, vol. 55, no. 4.

[12] J.Stasko, Visualization for Information Exploration and Analysis, Proceedings of the 4th ACM symposium on Software visualization, 2008, pp. 7-8.

[13] D.Selassie, B.Heller and J.Heer, Divided edge bundling for directional network data, IEEE Transactions on Visualization and Computer Graphics, 2011, vol. 17, no. 12, pp. 2354–2363.

[14] J.Tedesco, A.Sharma and R.Dudko, Theius: a streaming visualization suite for hadoop clusters, Proceedings of the IEEE International Conference on Cloud Engineering, 2013.

[15] T.-Y.Lee, C.Jones, B.-Y.Chen, and K.-L. Ma, Visualizing data trend and relation for exploring knowledge, Proceedings of the IEEE Pacific Visualization Poster, 2010.

[16] P.Zikopoulos, C.Eaton, D. deRoos, T.Deutsch and G.Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2012.

[17] C.Ahlberg and B.Shneiderman, Visual information seeking: tight coupling of dynamic query filters with starfield displays, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '94), April 1994, pp. 313–317.