# ПРЕДСТАВЛЕНИЕ И ВИЗУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

А.В. Благов

Самарский государственный аэрокосмический университет им. академика С.П. Королёва
(национально исследовательский университет)

В настоящее время одним из самых перспективных направлений для исследований в различных областях является обработка и анализ данных сверхбольшого объема (Big Data). В данной статье рассматриваются вопросы представления данных социальных сетей, а также визуального представления систем, описанных с помощью графов. Предложены алгоритмы по уменьшению размерности графов.

**Введение**

В современном мире практически во всех областях деятельности человека генерируется большое количество данных, которые могут быть использованы для решения важнейших задач в области медицины, безопасности, финансов и т.д. Многие системы, такие как социальные, биологические, финансовые сети, а также сложные процессы, в том числе меняющиеся во времени, могут быть представлены в виде графов. Анализ подобных больших графов важен для описания и определения зависимостей, прогнозов и т.д. [1]. Анализ сложных систем бывает сопряжен и с их визуализацией, а зачастую эта задача является трудоемкой вследствие колоссального количества данных, нуждающихся в обработке и упрощении [1, 2].

**1. Данные социальных сетей**

Исследование данных социальных сетей является актуальной задачей в области социологии, маркетинга и безопасности. Так или иначе, эти данные отражают общественное мнение и определяют различного рода зависимости. В подобных исследованиях может быть получена информация о популярности той или иной тематики в различных географических локациях, а также её восприятие (положительное, отрицательное, нейтральное) населением. На рисунке 1 приведен пример популярности и восприятия тематики «big data» и «data mining» в мире.

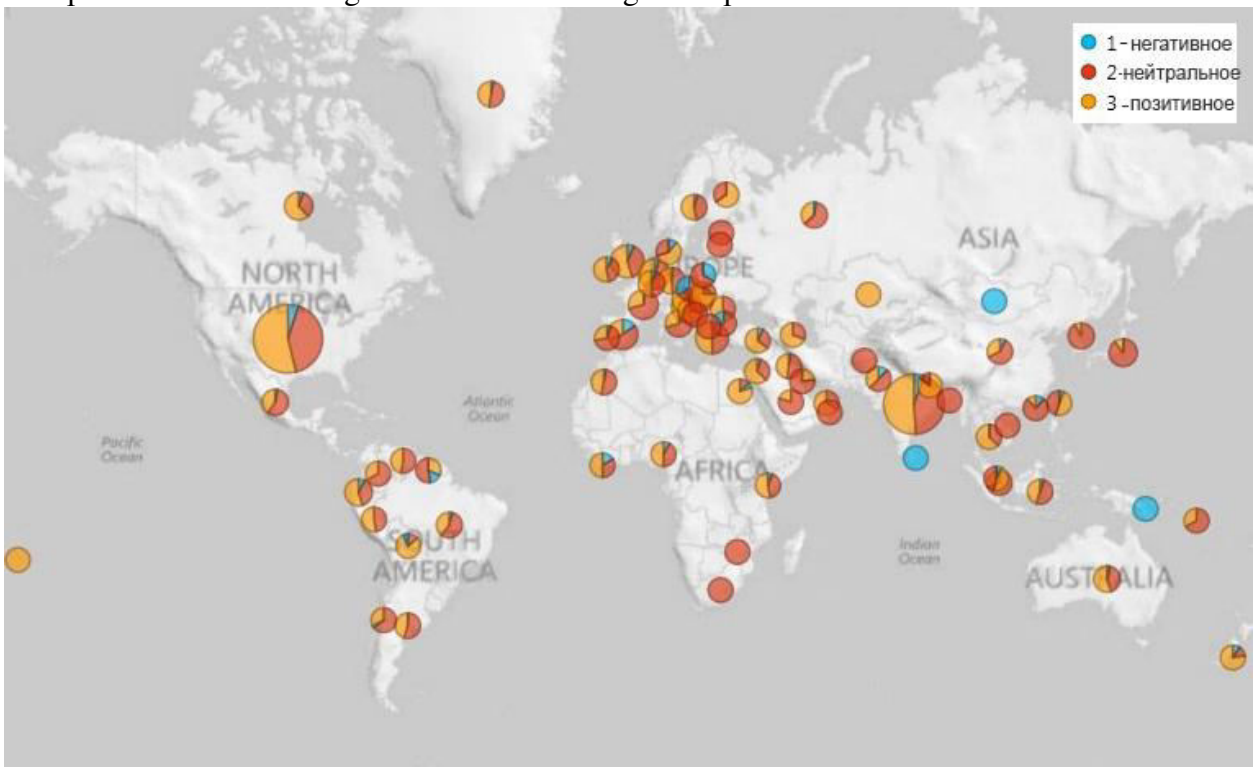


Рисунок 1 – Пример распределения сообщений сети twitter с эмоциональным окрасом по всему миру

Еще одной задачей является построение моделей, описывающих зависимости в социальных сетях. В одном из подходов, разобранных в работах [3, 4], рассматривается $G_{n,p}$ – случайная модель на графах, где $n$ – количество вершин в графе, $p$ – величина в диапазоне от 0 до 1, задающая для каждой пары ($i$, $j$), соединенной ребром, степень зависимости. Степень узла при этом задается следующим образом: $z = np$. Степенное распределение социальных сетей описывается распределением Пуассона: $p(k) = \frac{z^k}{k!}e^{-z}$. Таким образом, степенное распределение сконцентрировано вокруг своего среднего значения, а вероятность наличия узлов с высокой степенью экспоненциально мала.

Подобные модели могут применяться для описания отношений и взаимодействий в определенной большой группе лиц.

## 2. Визуальный анализ графов

Визуализация данных сверхбольшого объема, представленных в виде графов является нетривиальной задачей. Принято выделять три важнейших компонента визуального анализа данных [1]:

- визуальное представление графа;
- алгоритмический анализ графа;
- взаимодействие с пользователем.

Как правило, вовлечение пользователя в процесс анализа и чередование его решений с автоматическим процессом является необходимым для анализа данных сверхбольшого объема.

В работе [5] визуальный анализ данных представлен следующей схемой (рисунок 2).

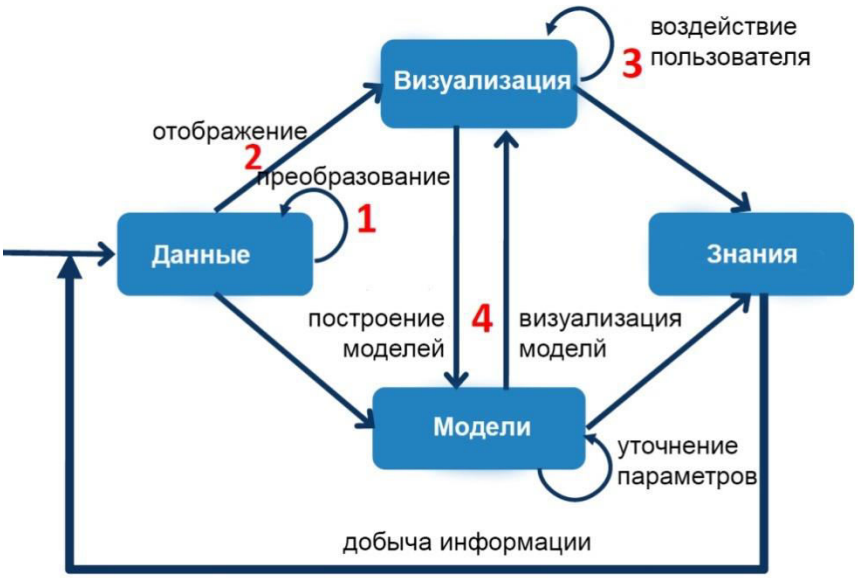


Рисунок 2 – Процесс визуального анализа

В классической теории графов граф $G = (V, E)$ характеризуется парой двух элементов вершин (узлов) и ребер. При описании потоков данных сверхбольших объемов возникает еще и временная компонента $T$, в этом случае имеют место графы изменяющиеся во времени [6]. Все большую актуальность приобретает задача упрощения графа путем уменьшения его размера без потери связности.

Выделяют два основных подхода для уменьшения размерности графа: фильтрация графа и агрегация графа.

Фильтрация графа бывает двух типов: стохастическая и детерминированная. Стохастическая фильтрация строится на основе случайного отбора узлов и ребер из исходного графа. Данные методы рассмотрены в работе [7]. Детерминированная фильтрация согласно своему названию использует детерминированный алгоритм для выбора узлов и ребер, пригодных к удалению. Как правило, данная фильтрация

основывается на атрибутах узла, а также на топологии графа. К примеру, фильтрация на основе данного метода может быть использована для удаления менее важных краев по различным метрикам, сохраняя при этом основную структуру (связность и другие функции) графа [8]. Одна из наиболее распространенных метрик "промежуточность-центральное," [9], которая показывает, что часто узел лежит на коротком (и, предположительно, наиболее часто используемом) пути между другими узлами:

$$BC(v) = \sum_{u \neq v \neq w \in V} \sigma_{u,w(v)} / \sigma_{u,w},$$

где $\sigma_{u,w}$ - количество кратчайших путей между $u$ и $w$, а $\sigma_{u,w(v)}$ – только те из них, которые содержат вершину $v$. При этом удаляются самые дальние по BC ребра графа, изолируя и кластеризуя подсети, однако не во всех задачах визуализации это бывает оправдано [10].

При агрегации графа узлы и ребра объединяются, при этом уменьшая размер графа и выявляя связи между группами узлов. Агрегация может быть применена несколько раз, образуя иерархический граф.

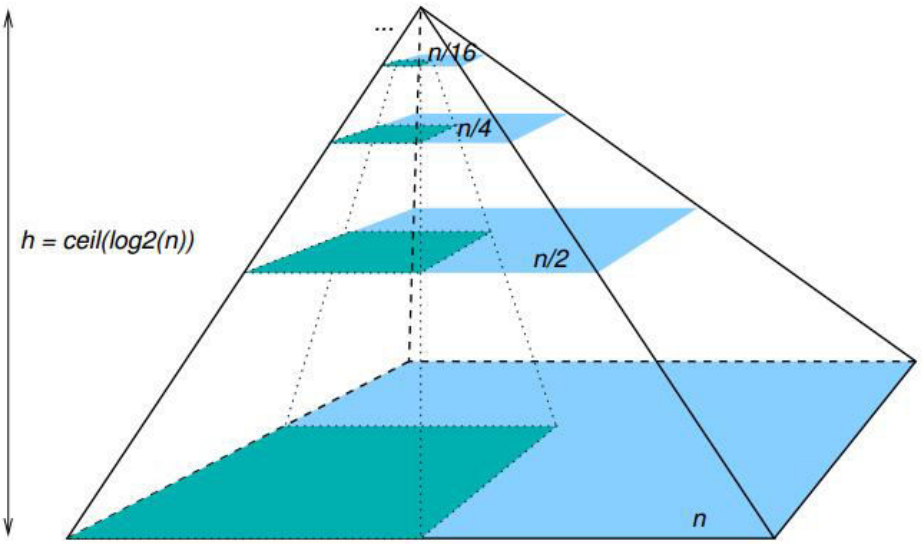


Рисунок 3 – Агрегация для многомасштабной визуализации графа

Существуют различные способы агрегирования графа, в том числе с использованием заданных иерархий узлов, или агрегации в соответствии с атрибутами узлов [11] (см. рисунок 4).

**Заключение**

Визуальный анализ данных сверхбольшого объема позволяет понимать сложные процессы и системы, извлекая из них полезную информацию и знания. Графы при этом служат удобным инструментом. Исследования в области визуального анализа данных находят широкое применение в различных направлениях, ускоряя время принятия решения там, где это необходимо.

**Литература**

1. Von Landesberger T. et al. Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges //Computer graphics forum. – Blackwell Publishing Ltd, 2011. – Т. 30. – №. 6. – С. 1719-1749.
2. Kazanskiy N.L., Protsenko V.I., Serafimovich P.G. Comparison of system performance for streaming data analysis in image processing tasks by sliding window // Computer Optics. – 2014. – Vol. 38(4). – P. 804-810.
3. Koloniari G., Pitoura E. Partial View Selection for Evolving Social Graphs. – 2013.
4. Margariti S. V., Dimakopoulos V. V. A study on the redundancy of flooding in unstructured p2p networks //International Journal of Parallel, Emergent and Distributed Systems. – 2013. – Т. 28. – №. 3. – С. 214-229.
5. Kein D., Andrienko G., Fekete J.-D., Görg C., Kohlhammer J., Melancon G.: Information Visualization, vol. 4950 of Lecture Notes in Computer Science. Springer, 2008, ch. Visual Analytics: Definition, Process, and Challenges, pp. 154–175.

6. Wehmuth, Klaus, Artur Ziviani, and Eric Fleury. "Model for Time-Varying Graphs." Workshop on Dynamic Networks. 2013.
7. Leskovec J., Faloutsos C.: Sampling from large graphs. In Proceedings of 12th ACM SIGKDD Int. Conference on Knowledge Discovery and data mining (2006), pp. 631–636.
8. Jia Y., Hoberock J., Garland M., Hart J.: On the visualization of social and other scale-free networks. IEEE Transactions on Visualization and Computer Graphics, 2008, 1285–1292.
9. L. C. Freeman. A set of measures of centrality based upon betweenness. Sociometry, 40(1):35–41, 1977.
10. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821–7826, 2002.
11. Elmqvist N., Do T.-N., Goodell H., Henry N., Fekete J.-D.: Zame: Interactive large-scale graph visualization. In Proceedings of IEEE Pacific Visualization Symposium (2008), pp. 215–222.