УДК 004.421+519.178

МЕТОДЫ И СРЕДСТВА ВИЗУАЛИЗАЦИИ МАССИВОВ НАУЧНО-ТЕХНИЧЕСКИХ ПОКАЗАТЕЛЕЙ В ВИДЕ ГРАФОВ

Целых А.А.[1], Целых А.Н.[1], Матвеев Д.А.[2]

[1]ФГАОУ ВПО «Южный федеральный университет», Ростов-на-Дону, Россия (344006, г. Ростов-на-Дону, ул. Большая Садовая, 105/42), e-mail: tselykh@sfedu.ru
[2]ООО «Услуги инфо», Екатеринбург, Россия (620028, г. Екатеринбург, ул. Татищева, 49а, офис 415), e-mail: mda_job@rbcmail.ru

Статья содержит аналитический обзор и обоснование выбора методов, алгоритмов и программных средств с открытым исходным кодом для визуализации больших массивов научно-технических показателей в виде графов. Рассматриваются индикаторы научно-технического потенциала развития науки и техники, содержащиеся в базах данных федеральных целевых научно-технических программ. С позиции теории графов и гиперграфов дается классификация типов связей, которые могут быть использованы для представления и визуализации данных. Вводится понятие обобщенного гиперграфа специального вида, который позволяет отображать более сложные функциональные и смысловые связи между объектами. Для разрабатываемого программного комплекса предлагается клиент-серверная архитектура с компонентой визуализации на стороне клиента, реализуемой средствами HTML5 и Javascript, и компонентой анализа на стороне сервера с использованием внешней Java-библиотеки Gephi Toolkit.

Ключевые слова: визуализация данных, визуальная аналитика, граф, гиперграф.

METHODS AND TECHNIQUES FOR VISUALIZING ARRAYS OF SCIENTIFIC AND TECHNICAL INDICATORS USING GRAPHS

Tselykh A.A.[1], Tselykh A.N.[1], Matveev D.A.[2]

[1]Southern Federal University, Rostov-on-Don, Russia (344006, Rostov-on-Don, B. Sadovaya Street, 105/42), e-mail: tselykh@sfedu.ru
[2]«Uslugi info» LLC, Ekaterinburg, Russia (620028, Ekaterinburg, Tatischeva Street, 49 «a», office 415), e-mail: mda_job@rbcmail.ru

This paper is an analytical review and argumentation of methods, algorithms and open source software for visualizing arrays of scientific and technical indicators using graphs. We specifically consider indicators of scientific and technical development potential from the databases of federal scientific and technical programs. From the position of graph and hypergraph theories, we classify types of relationships used to represent and visualize data. We introduce a notion of a generalized hypergraph that allows representing complex functional and semantic relationships between the objects. For a software complex under development, we propose client-server architecture with a client-side visualization component based on HTML5 and Javascript and server-side analysis component using external Java library of Gephi Toolkit.

Key words: data visualization, visual analytics, graph, hypergraph.

Введение

Стремительному развитию средств бизнес-аналитики и интеллектуального анализа данных [3] сопутствует развитие таких методов поддержки аналитической работы с данными, как визуализация данных и визуальная аналитика, которые позволяют наглядно представить большие массивы числовой и другой информации в интуитивно понятной и информативной визуальной форме.

Объектом настоящего исследования является система управления большими массивами показателей в научно-технической сфере. Примерами таких показателей являются индикаторы научно-технического потенциала развития науки и техники, содержащиеся в

базах данных федеральных целевых научно-технических программ: базе заявок на формирование тематики и объемов финансирования работ и проектов, базе заявок на участие в конкурсе, базе государственных контрактов, заключенных государственным заказчиком, и базе результатов выполненных работ.

Предмет исследования – методы и средства визуализации структурированных данных, базовым формализмом для моделирования которых является граф как совокупность объектов и связей между ними.

Методы визуализации графов разрабатываются с начала 60-х годов XX века [6]. Мощный импульс развития направление исследований получило с появлением первых подходов на основе физических аналогий – «пружинных» алгоритмов Eades (1984), Kamada-Kawaii (1989) и Fruchterman-Reingold (1991). Среди алгоритмов, имеющих линейно-логарифмическую сложность, можно выделить многоуровневый алгоритм Yifan Hu, алгоритмы Force Atlas 2 и Open Ord (2011). В 1998 году в самостоятельное направление выделилась визуализация информации (в первую очередь текстовой), и всего несколько лет назад – визуальная аналитика. В фокусе современных исследований – методы визуализация больших графов, методы навигации при визуализации графов [1], методы визуализации социальных сетей. Основные результаты в этой области публикуются в международных изданиях Journal of Graph Algorithms and Applications, IEEE Transactions on Visualization and Computer Graphics, трудах Graph Drawing Symposia, IEEE Information Visualization Conference и других научно-технических мероприятий.

Цель исследования. Обоснование выбора методов, алгоритмов и программных средств с открытым исходным кодом для визуализации больших массивов научно-технических показателей в виде графов.

Методология исследования. Методы визуализации, теория графов и гиперграфов, теория баз данных.

Графы и гиперграфы специального вида для задач разметки и визуализации данных

Научно-технические показатели одной категории зачастую являются атрибутами различных сущностей и располагаются в разных таблицах в одной или в разных базах данных. Аналитикам же хочется видеть близкие по смыслу данные рядом. Процесс работы с исходными базами данных посредством группирования близких по смыслу атрибутов назовем выделением смысловых групп.

Для представления и визуализации связей как внутри одной смысловой группы, так и в нескольких группах обратимся к математическому аппарату теории графов и гиперграфов. Основное отличие гиперграфов [4] от графов заключается в том, что ребро гиперграфа может

соединять не две, а большее число вершин. И хотя любой гиперграф можно представить в виде двудольного графа, а затем применить широкий спектр алгоритмов на графах, гиперграфы являются иным уровнем представления и хорошо понятны для аналитиков. Для отображения более полного спектра связей между компонентами данных может потребоваться использование некоторых дополнительных уточнений графовой парадигмы. Это могут быть ориентированные гиперграфы, GH-гиперграфы специального вида [5] и нечеткие графовые модели [2].

Выделим типы связей, которые могут быть использованы для представления и визуализации данных.

Неориентированные связи между простыми компонентами данных. Две вершины связаны неориентированным ребром, все связи в графе имеют одинаковый смысл («принадлежать к», «быть частью» – «состоять из»). Связь типа «*gv-gv*» (*gv – graphvertex*), представляется при помощи матриц (таблиц) инцидентности (*Gv×Ge* определяет, какое ребро связывает какие две вершины, где *edge* – ребро, *vertex* – вершина, *Gv* – множество вершин графа, *Ge* – множество ребер графа) или смежности (*Gv×Gv* определяет, какая вершина связана с какой вершиной; в неориентированном графе матрица смежности симметрична относительно диагонали, т.е. существуют связи $gv_i \rightarrow gv_j$ *и* $gv_j \rightarrow gv_i$, элементы матрицы *Gv×Gv* $gv_{ij}=gv_{ji}=1$).

Ориентированные связи между простыми компонентами данных. Две вершины связаны ребром, все связи в графе имеют одинаковый смысл. Например, указание влияния, убывание/возрастание степени общности, обозначение последовательности процессов и т.д. Представляются также матрицами *Gv×Ge* и *Gv×Gv*, но матрица смежности не симметрична – при наличии связи $gv_i \rightarrow gv_j$, $gv_{ij}=1$, $gv_{ji}=0$.

Ориентированные и неориентированные связи между простыми компонентами данных. В графе присутствуют группы связей различных типов: «*gv-gv*» и «*gv→gv*». Например, на одном и том же множестве вершин одним типом связей представлены отношения персон в штатной структуре, другим типом – отношения во временном трудовом коллективе.

Каждая ориентированная или неориентированная связь графа имеет свое смысловое значение. В графе присутствуют группы связей различных типов: «gv_i-t_k-gv_j» и «gv_i-t_k→gv_j». Например, на одном и том же множестве вершин различными видами связей представлены отношения персон. На рис. 1 {*A, B, C, D*} – множество типов связей. Матрица *Gv×Gv* состоит из идентификаторов типов связей, 0 – отсутствие связи.

Неориентированные ребра гиперграфа представляют подмножества некоторого множества. Связь типа «*hv-he*» (*h – hypergraph*) представляется при помощи матриц инцидентности *Hv×He,* которые определяют, какое ребро какие вершины связывает.

Например, пересечение областей научных интересов, где предметная область – ребро, вершины – индивиды.

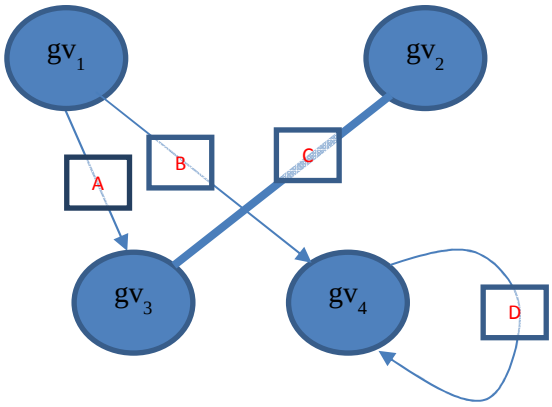

Рис. 1. Граф с группами связей различных типов

В ориентированных ребрах гиперграфа выделяется одна из вершин, которая имеет особый статус. Связь типа «*hv-he**». Связи внутри ребра типа «все к одному» (рис. 2). Например, руководитель научной школы.

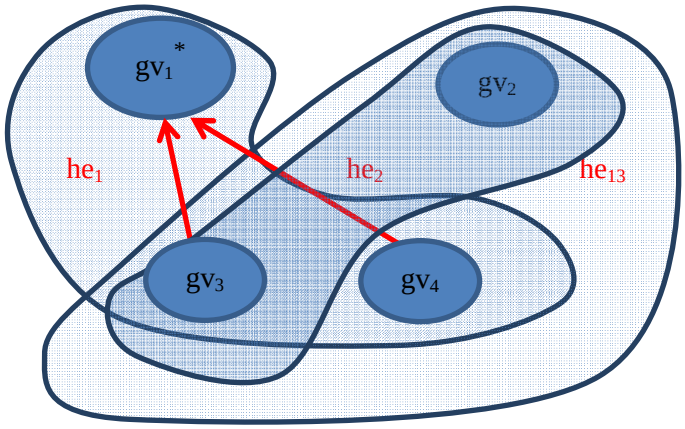

Рис. 2. Ориентированный гиперграф

Сложные связи внутри ребер гиперграфа. Графовые связи внутри ребра здесь не следует рассматривать как самостоятельные, они лишь описывают некоторые свойства конкретного ребра. Например, ребро гиперграфа – это конкурс, а вершины ребра – заявки, связанные внутри ребра гиперграфа ориентированными простыми (графовыми) ребрами в порядке возрастания значений некоторого параметра. Матрица инцидентности графа может содержать ссылки на матрицы смежности, описывающие отношения вершин внутри конкретного ребра.

«Графовые» связи между вершинами, входящими в разные ребра гиперграфа. Например, ребра гиперграфа – научные школы, вершины – индивиды, межреберные связи – связи цитирования.

Связи между ребрами гиперграфа. Например, ребра гиперграфа – стадии жизненного цикла заявки-вершины. Представляются несколькими матрицами инцидентности.

Развивая парадигму ориентированного гиперграфа [2; 4], предположим, что отношения вершин гиперграфа представлены неполным графом. Такая модель может отображать как отношения целых групп элементов, так и отношения между отдельными элементами.

Обобщенным гиперграфом назовем тройку следующего вида:

$H_X=(X,E , G_h (X))$, $X=\{x_1, x_2 ,\ldots,x_n\}$, $E =\{E_1, E_2 ,\ldots,E_m\}$, $E_1, E_2 ,\ldots,E_mX$,

$G_h (X)=(X,U)$, $U=\{u_1,u_2 ,\ldots,u_l,\ldots,u_s\}$, $u_l=(x_i, x_j)$, $x_i, x_j \in X$. (1)

Такое определение оставляет возможность представления отношений между вершинами разных ребер гиперграфа.

Частный случай обобщенного гиперграфа представлен определением (2). Он подразумевает графовые отношения только внутри ребер гиперграфа:

$H_X=(X,E_X)$,$X=\{x_1, x_2 ,\ldots,x_n\}$, $E_X = \{(E_1,G_{E1}), (E_2,G_{E2}),\ldots,(E_r,G_{El}),\ldots,$

$(E_m,G_{Em})\}$, $E_1, E_2 ,\ldots, E_r ,\ldots,E_mX$,$G_{Er}=(E_r,U_r)$,$U_r=\{u_1,u_2 ,\ldots,u_l,\ldots,u_k\}$,$u_l=(x_i, x_j)$, $x_i, x_j \in E_r$. (2)

Пример обобщенного гиперграфа показан на рис. 3.

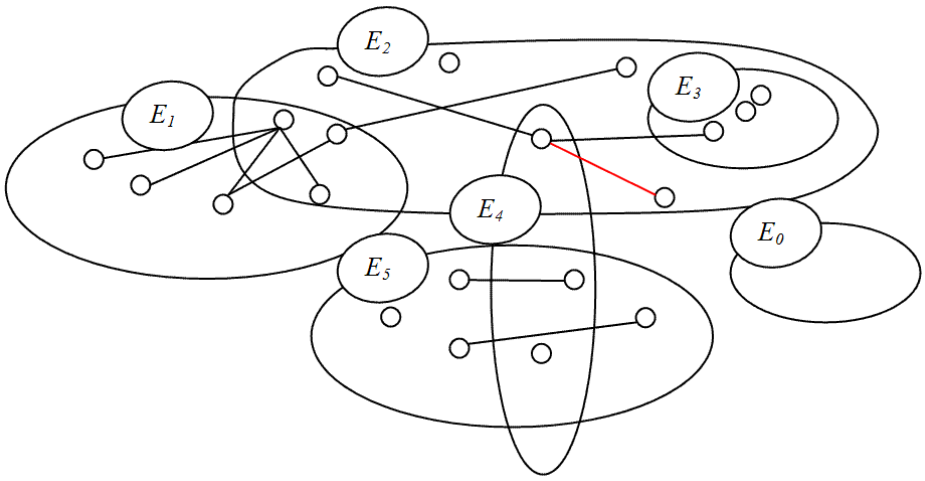

Рис. 3. Пример обобщенного гиперграфа

Нетрудно предположить, каким образом будет выглядеть таблица смежности нечеткого обобщенного гиперграфа. В ней должны присутствовать элементы следующих видов $<\mu_U(x_i, x_k)/(x_i, x_k)>$, $<\mu_X(x_i)/(x_i)>$, $<\mu_{Ei}(x_i)/(x_i)>$.

На рис. 4 предлагается пример представления обобщенного гиперграфа H_{X1}. На представление гиперграфа H_{X1} двудольным графом $G_K=(XUE_X ,V)$ наложены «графовые» связи вершин $G_S=G_KUG_{E1}UG_{E2},\ldots,G_{El}),\ldots, G_{Em}$.

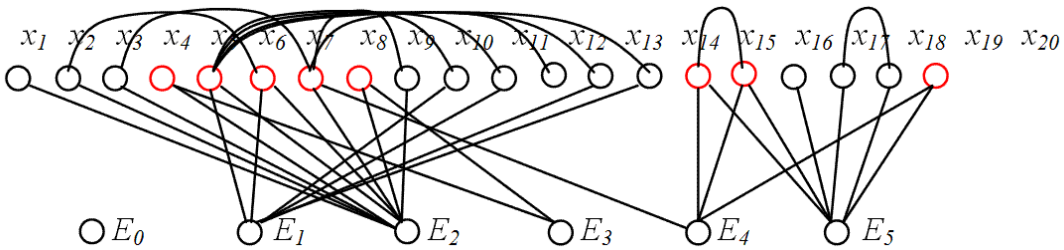

Рис. 4. Обобщенный гиперграф H_{X1}

Предлагаемая модель позволяет отображать более сложные функциональные и смысловые связи между объектами.

Программное решение для визуализации и анализа больших массивов научно-технических показателей в виде графов

Разрабатываемое программное решение имеет клиент-серверную архитектуру (рис. 5).

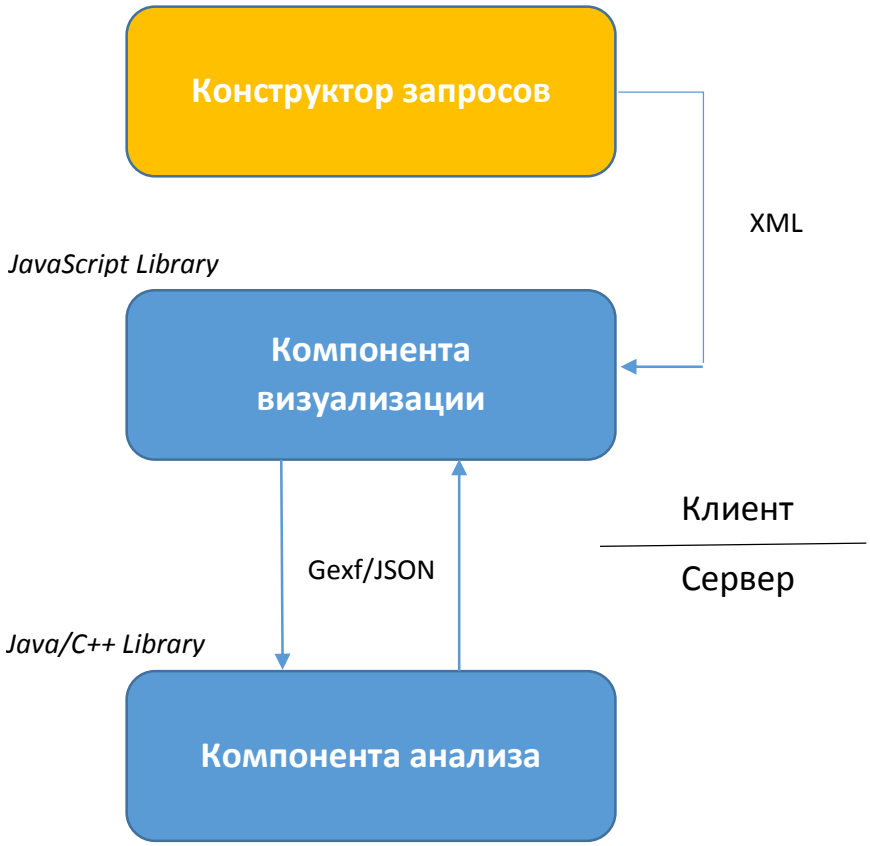

Рис. 5. Архитектура системы визуализации и анализа больших массивов научно-технических показателей в виде графов

Компонента визуализации работает на стороне клиента с использованием технологий HTML5 и JavaScript. Она предоставляет пользовательский интерфейс для формирования в режиме онлайн графового набора данных из реляционной таблицы, полученной по результатам работы конструктора запросов. Граф формируется последовательно с помощью несложных манипуляций на экране мобильного устройства. Поскольку даже небольшое

замедление при формировании графа будет сказываться на удобстве работы пользователя, компонента визуализация должна поддерживать базовые операции на графах (добавление/удаление вершин и связей) и быстрые алгоритмы укладки, обеспечивающие отрисовку графа в реальном времени без запросов к серверу.

Серверная *компонента анализа* должна поддерживать дополнительные наглядные виды укладки сложных графов на плоскости, а также раскраску и ранжирование размера вершин на основе метрик центральности, фильтрацию, построение эго-сети, кластеризацию вершин, визуализацию картографической информации. На вход компоненты анализа поступает исходный граф в формате GEXF/JSON, задающий множество вершин и связей между ними, а на выходе – результирующий граф в формате GEXF/JSON для отрисовки, в котором для каждой вершины уже определены все атрибуты визуализации, включая тип, цвет, размер и позиционирование вершины в пространстве.

Заключение

С целью представления и визуализации связей между научно-техническими показателями внутри одной смысловой группы или в нескольких группах введено понятие обобщенного гиперграфа специального вида, который позволяет отображать более сложные функциональные и смысловые связи между объектами. Дана классификация типов связей, которые могут быть использованы для представления и визуализации данных.

Разрабатываемый программный комплекс имеет клиент-серверную архитектуру с компонентой визуализации на стороне клиента, реализуемой средствами HTML5 и Javascript, и компонентой анализа на стороне сервера с использованием внешней Java-библиотеки Gephi Toolkit. В качестве формата хранения и обмена графовыми данными между компонентами рекомендуется использовать формат GEXF.

Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту от 10 августа 2012 г. № 14.514.11.4037 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

Список литературы

1. Апанович З.В. Методы навигации при визуализации графов // Вестник НГУ. Серия: Информационные технологии. – 2008. – Т. 6, вып. 3. – С. 35-47.

2. Берштейн Л.С., Боженюк А.В. Нечеткие графы и гиперграфы. – М. : Научный мир, 2005. – 256 с.

3. Боженюк А.В., Котов Э.М., Целых А.А. Интеллектуальные интернет-технологии (учебник) // Успехи современного естествознания. – 2010. – № 2. – С. 93-94.

4. Емеличев В.А., Мельников О.И., Сарванов В.И., Тышкевич Р.И. Лекции по теории графов. Глава XI: Гиперграфы. – М. : Наука, 1990. – С. 298-315.
5. Целых А.А. Графогиперграфовая модель семантической социальной сети // Известия Южного федерального университета. Технические науки. – 2012. – № 4 (129). – С. 225-229.
6. Batista G.Di, Eades P., Tamassia R., Tollis I.G. Algorithms for Drawing Graphs: an Annotated Bibliography // Computational Geometry: Theory and Applications. – 1994. – № 4. – Pp. 235-282.

Рецензенты:

Карелин Владимир Петрович, д.т.н., профессор, заведующий кафедрой прикладной математики и информационных технологий НОУ ВПО «Таганрогский институт управления и экономики», г. Таганрог.

Ромм Яков Евсеевич, д.т.н., профессор, заведующий кафедрой информатики ФГБОУ ВПО «Таганрогский государственный педагогический институт имени А.П. Чехова», г. Таганрог.