Инженерный вестник Дона, №1 (2018)
ivdon.ru/ru/magazine/archive/n1y2018/4651

Нахождение подобия между неструктурированными объектами данных на основе метода сингулярного разложения спектра графа

Г.С. Мизюков

Ростовский государственный университет путей сообщения, Ростов-на-Дону

Аннотация: В статье рассматривается вопрос нахождения подобия между объектами, содержащими неструктурированную информацию на основании спектров двух объектов. Для нахождения спектра используется матрица смежности графа. Подобие между объектов сравнения определяется с использованием подхода основанного на сингулярном разложении матриц смежности графов. Также в статье рассматриваются существующие решения и приведены примеры сфер возможного применения описанного подхода.
Ключевые слова: спектр графа, сингулярное разложение, матрица смежности, неструктурированная информация, анализ больших массивов информации.

Большие данные сейчас являются новым трендом в области высоких технологий, в частности рынок аналитических инструментов, представлен широким спектром инструментов, позволяющим проводит различные манипуляции с информацией, хранящейся на серверах и представлять аналитические отчеты компании в различных разрезах. Росту информации, в свою очередь, способствует колоссальный «информации взрыв», произошедший в результате удешевления элементарной базы компонентов, необходимых для производства, как устройств хранения информации, так и устройств, которые генерируют информацию различной природы. Из-за этого многократного увеличения объема информации, а также многообразия хранимой информации различной природы, усложняется процесс управления данными. По данным аналитической компании *International Data Corporation* (*IDC)* до 60% информации, которая хранится на серверах в компаниях, не несёт в себе пользы. Эта информация представляет собой информационный шум, который усложняет процесс обработки и анализа информации. На рис. 1 представлен один из прогнозов представленных *IDC*, на котором можно увидеть динамику роста объемов информации. Так например, объем данных на планете вырастет до 40 зеттабайт к 2020 году,

т.е. каждый активный пользователь сети интернет будет генерировать по 5200 Гб. данных (рис. 1).

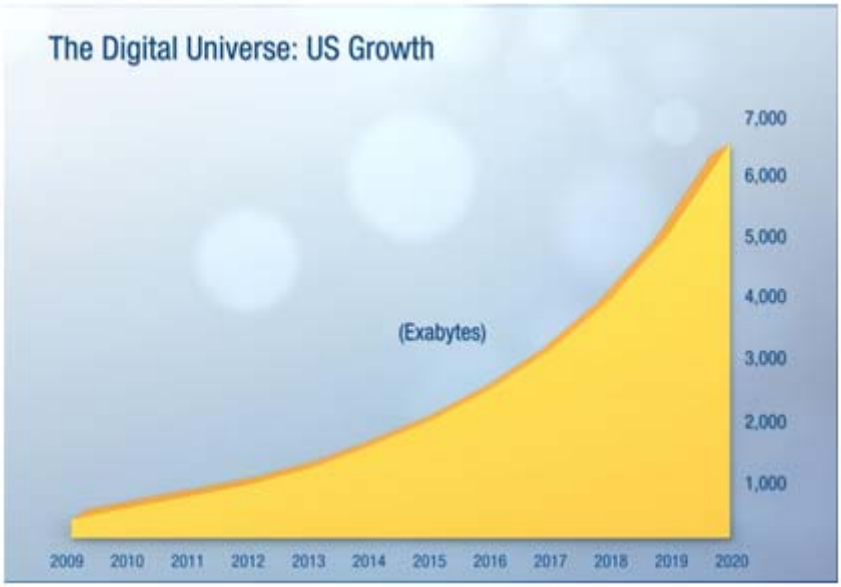


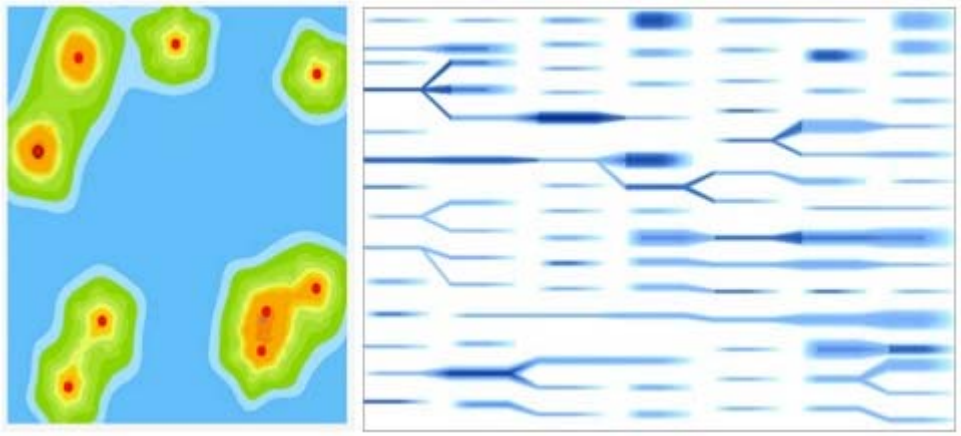
Рис. 1. – Прогноз компании *IDC* до 2020 года

Однако основной информационный поток будут формировать не люди, а устройства: сенсоры, смартфоны, интеллектуальные системы и т.д. Это в свою очередь приводит к потребности появления новых направлений в сфере информационных технологий, а также увеличению количество серверов, способных хранить и обрабатывать огромные массивы данных. На текущий момент все больше компаний заинтересовано в эффективном управлении информаций, так как неэффективное управление информацией, в условиях рыночных отношений, где преобладают информационные технологии, может оказать негативную динамику на прибыль компании, поэтому нельзя не отметить роль компаний в формировании мировоззрения по отношению к большим массивам информации и проблеме их анализа. Что в свою очередь обусловлено тем, что умение эффективно и качественно проводить анализ информации, а также оперативно реагировать на все изменения в структуре информации является одним из основных показателей зрелости компании в области информационной политики.

Тематике анализа данных посвящён один из документов представленных исследовательской и консалтинговой компанией, специализирующейся на рынках информационных технологий *Gartner* под названием «*Market Guide for File Analysis Software*». В документе приводится информация, касающаяся типовых сценариев использования аналитических инструментов, среди которых можно отметить следующие:

1. Оптимизация хранения;
2. Выявление ненужных данных и избавление от них при миграции ИТ-инфраструктуры;
3. Классификация;
4. Соблюдение нормативов и требований (*compliance*);
5. Управление уровнями доступа;
6. Автоматизация проведения расследований.

Для нас наиболее интересны первые три позиции, а именно: оптимизация хранения, выявление ненужных данных и классификация. На рынке существуют два инструмента, которые в равной степени позволяют качественно выполнять перечисленные выше сценарии при работе, в особенности с неструктурированной информацией. Это программные продукты компании *Hewlett-Packard* под названием *HP Storage Optimizer* и *HP Control Point*. Первый продукт специализируется на оптимизации хранения, второй продукт ориентирован на комплексный анализ с целью снижения бизнес-рисков, связанных с хранением данных. Оба инструмента обладают широким спектром функций, позволяющих качественно и эффективно управлять информацией. В частности хотелось бы отметить типы визуализации данных, которые могут представлять программные продукты – это карта кластеров информации рис. 2, *а* и спектограф рис. 2, *б* отображающий процесс изменения информации во времени внутри документа.



а) *б)*

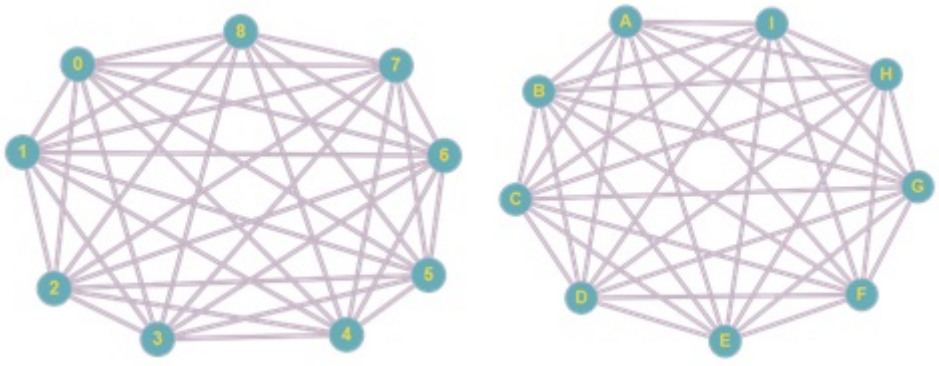
Рис. 2. – Типы визуализации данных программных продуктов Hewlett-Packard

а – карта кластеров; *б* – спектограф

Для определения подобия или схожести двух и более объектов информации программные продукты *Hewlett-Packard* использую мета данные содержащиеся в структуре документов. На основании метаданных осуществляется классификация и кластеризация данных со схожими наборами метаданных.

В качестве альтернативы программным продуктам *Hewlett-Packard* и методам, которые используются для нахождения подобия между объектами сравнения, мы будем использовать подход, основанный на спектральном разложении графа с использованием метода сингулярного разложения [1 – 3, 8]. Данный метод хорошо себя зарекомендовал в реконструкции и распознании 3D-объектов по спутниковым изображениям, с чем можно ознакомиться в статье с одноимённым названием [4]. Однако мы будем применять данный метод для нахождения подобия между неструктурированной информацией, на основе их спектра. В качестве исходных данных мы также будем использовать метаданные, содержащиеся в структуре документов.

На первом этапе нам необходимо построить два графа и описать их с помощью матрицы смежности (рис. 3). Первый граф будет выступать в качестве эталона, с которым необходимо будет производить математические операции, с целью определение подобия. Второй граф – это граф, полученный в результате выявления метаданных одного из множества документов хранящихся в *n* мерном массиве. Так как в данном решении предлагается взаимодействие с неструктурированными данными, построение графов и их структура может сильно отличаться, вплоть до использования вложенных метаграфом, которые в полной мере могут описать структуру документа и взаимосвязи внутри документа [5]. Для хранения подобных структур наиболее подходящими будут являться *NoSQL* базы данных, которые обладают широкими возможностями по описанию сложных структур данных с большим количеством связей [6, 7, 9, 10]. Однако же в нашем примере мы будем использовать простые полносвязные неориентированные графы. Матрица смежности неориентированного графа имеет вид: $G = (V, E)$ – квадратная симметричная матрица *A(G)* порядка *n*, элементы a_{ij} которой равны числу ребер, соединяющих вершины v_i и v_j.



а) *б)*

Рис. 3. – Нахождение подобия между вершинами графа

а – граф эталон; *б* – граф построенный на основе мета данных

В результате того, что оба графа изоморфны, матрица смежности обеих графов представленных на рис. 3 будут идентичны:

$$G_A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} G_B = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Рис. 4. – Матрицы смежности графа *A* и *B*

Нахождение подобия между сравниваемыми объектами на основе построенных графов состоит в поиске соответствий между структурами графов, поэтому предлагается использовать методы нахождения подобия на основе спектральной теории графов. Спектр графа представляет собой множества собственных значений $\{\lambda_1, \lambda_2 \ldots \lambda_{|V|}\}$ упорядоченных по убыванию или возрастанию. Спектральные методы основаны на следующем свойстве: собственные значения и собственные векторы матрицы смежности графа инвариантны относительно перестановок вершин в матрице. Следовательно, если два графа изоморфны, их матрицы смежности будут иметь одинаковые собственные значения и векторы, что собственно и показано на рис. 4. [4].

Из-за изоморфности двух графов мы будем производить разложение только матрицы G_A. Разложение матрицы G_A размерности $m \times n$ на собственные значения с использованием сингулярного разложения, можно представить в виде следующих формул:

$$G_A = U\Sigma V \quad (1)$$

или

$$G_A = U\Sigma V_R \quad (2)$$

где U и V – ортогонали матриц, если G_A – действительная или унитарная матрица; если G_A – комплексная матрица; V_H – сопряжённо-транспонированная матрица V с порядками m и n соответственно; Σ – диагонали матрица $m \times n$ с действительными элементами σ_i:

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\min(m,n)} \geq 0$$

где σ_i – сингулярные значения матрицы G_A, а первые $min(m,n)$ столбцы матриц U и V – левые и правые вектора матрицы G_A, которые должны удовлетворять следующему отношению:

$$G_A v_i = \sigma_i u_i$$

и

$$G_A^T u_i = \sigma_i v_i$$

где u_i и v_i- i-ые столбцы матриц U и V соответственно.

Следующим этапом выполним сингулярное разложение матрицы смежности графа G_A представленного на рис. 3 на основе формулы 1. В результате мы получим следующие результаты:

Сингулярные значения:

8,0000 1,0000 1,0000 1,0000 1,0000 1,0000 1,0000 1,0000 1,0000

Матрица собственных векторов:

$$U_A = \begin{bmatrix} -0,3333 & -0,0000 & -0,0000 & -0,0000 & -0,0000 & -0,0000 & 0,0000 & 0,0000 & 0,9428 \\ -0,3333 & -0,0000 & -0,0000 & -0,0000 & -0,0000 & -0,0000 & 0,0000 & 0,9354 & -0,1179 \\ -0,3333 & 0,3427 & 0,1349 & 0,7586 & 0,2182 & 0,2976 & -0,0994 & -0,1336 & -0,1179 \\ -0,3333 & 0,1449 & 0,0570 & -0,0406 & -0,5175 & 0,0602 & 0,7482 & -0,1336 & -0,1179 \\ -0,3333 & -0,8691 & 0,1301 & 0,0527 & 0,0599 & 0,2800 & -0,0108 & -0,1336 & -0,1179 \\ -0,3333 & -0,0356 & -0,8906 & 0,0527 & 0,0599 & -0,2313 & -0,0533 & -0,1336 & -0,1179 \\ -0,3333 & 0,2518 & 0,0991 & -0,5623 & 0,6372 & 0,2004 & 0,1468 & -0,1336 & -0,1179 \\ -0,3333 & -0,0356 & 0,3903 & 0,0527 & 0,0599 & -0,8286 & -0,1029 & -0,1336 & -0,1179 \\ -0,3333 & 0,2008 & 0,0791 & -0,3136 & -0,5175 & 0,2218 & -0,6286 & -0,1336 & -0,1179 \end{bmatrix}$$

Рис. 5. – Матрица собственных векторов графа G_A

В матрице собственных векторов U_A изменим значения отрицательных элементов, используя модуль числа, в результате получим матрицу $|\overline{U_A}|$.

$$|\overline{U_A}| = \begin{bmatrix}
0,3333 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,9428 \\
0,3333 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,9354 & 0,1179 \\
0,3333 & 0,3427 & 0,1349 & 0,7586 & 0,2182 & 0,2976 & 0,0994 & 0,1336 & 0,1179 \\
0,3333 & 0,1449 & 0,0570 & 0,0406 & 0,5175 & 0,0602 & 0,7482 & 0,1336 & 0,1179 \\
0,3333 & 0,8691 & 0,1301 & 0,0527 & 0,0599 & 0,2800 & 0,0108 & 0,1336 & 0,1179 \\
0,3333 & 0,0356 & 0,8906 & 0,0527 & 0,0599 & 0,2313 & 0,0533 & 0,1336 & 0,1179 \\
0,3333 & 0,2518 & 0,0991 & 0,5623 & 0,6372 & 0,2004 & 0,1468 & 0,1336 & 0,1179 \\
0,3333 & 0,0356 & 0,3903 & 0,0527 & 0,0599 & 0,8286 & 0,1029 & 0,1336 & 0,1179 \\
0,3333 & 0,2008 & 0,0791 & 0,3136 & 0,5175 & 0,2218 & 0,6286 & 0,1336 & 0,1179
\end{bmatrix}$$

Затем транспонируем полученную матрицу $|\overline{U_A}|$:

$$U_A^T = \begin{bmatrix}
0,3333 & 0,3333 & 0,3333 & 0,3333 & 0,3333 & 0,3333 & 0,3333 & 0,3333 & 0,3333 \\
0,0000 & 0,0000 & 0,3427 & 0,1449 & 0,8691 & 0,0356 & 0,2518 & 0,0356 & 0,2008 \\
0,0000 & 0,0000 & 0,1349 & 0,0570 & 0,1301 & 0,8906 & 0,0991 & 0,3903 & 0,0791 \\
0,0000 & 0,0000 & 0,7586 & 0,0406 & 0,0527 & 0,0527 & 0,5623 & 0,0527 & 0,3136 \\
0,0000 & 0,0000 & 0,2182 & 0,5175 & 0,0599 & 0,0599 & 0,6372 & 0,0599 & 0,5175 \\
0,0000 & 0,0000 & 0,2976 & 0,0602 & 0,2800 & 0,2313 & 0,2004 & 0,8286 & 0,2218 \\
0,0000 & 0,0000 & 0,0994 & 0,7482 & 0,0108 & 0,0533 & 0,1468 & 0,1029 & 0,6286 \\
0,0000 & 0,9354 & 0,1336 & 0,1336 & 0,1336 & 0,1336 & 0,1336 & 0,1336 & 0,1336 \\
0,9428 & 0,1179 & 0,1179 & 0,1179 & 0,1179 & 0,1179 & 0,1179 & 0,1179 & 0,1179
\end{bmatrix}$$

Для нахождения подобия необходимо перемножить матрицу U_A^T с матрицей собственных векторов сравниваемого объекта, т.е графа G_B. За счет того, что в нашем методе будут применяться только изоформные графы, матрицы собственных векторов, следовательно, будут идентичны с матрицей $|\overline{U_A}|$. В результате перемножения матриц U_A^T и $|\overline{U_A}|$, мы получим матрицу $U_A U_A^T$.

$$U_A U_A^T = \begin{bmatrix}
0,9998 & 0,6268 & 0,5936 & 0,6110 & 0,6900 & 0,7066 & 0,5966 & 0,6235 & 0,6286 \\
0,6268 & 1,0000 & 0,2540 & 0,5200 & 0,4704 & 0,4868 & 0,3206 & 0,2512 & 0,2217 \\
0,5936 & 0,2540 & 1,0000 & 0,2595 & 0,2475 & 0,6468 & 0,2094 & 0,2380 & 0,2161 \\
0,6110 & 0,5200 & 0,2595 & 1,0000 & 0,7166 & 0,4811 & 0,3943 & 0,2449 & 0,2161 \\
0,6900 & 0,4704 & 0,2475 & 0,7166 & 1,0000 & 0,4188 & 0,8377 & 0,2766 & 0,2441 \\
0,7066 & 0,4868 & 0,6468 & 0,4811 & 0,4188 & 1,0000 & 0,3441 & 0,2832 & 0,2499 \\
0,5966 & 0,3206 & 0,2094 & 0,3943 & 0,8377 & 0,3441 & 0,9999 & 0,2391 & 0,2110 \\
0,6235 & 0,2512 & 0,2380 & 0,2449 & 0,2766 & 0,2832 & 0,2391 & 0,9999 & 0,2205 \\
0,6286 & 0,2217 & 0,2100 & 0,2161 & 0,2441 & 0,2499 & 0,2110 & 0,2205 & 1,0000
\end{bmatrix}$$

В матрице $U_A U_A^T$ максимальные значения заменим на 1 остальные значения приравняем к 0, в итоге мы получим матрицу *P*.

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

В полученной матрице *P* номер столбца равен номеру узла в графе G_A, номер строки – номеру узла в графе с которым производилась операция сравнения G_B. Единица в матрице *P* указывает на соответствие между сравниваемыми объектами. Из всего выше перечисленного можно сделать вывод, что два объекта сравнения, описанные в виде графом подобны.

В заключении хотелось бы отметить, что в статье описан пример с идеально подобранными параметрами и, который является одним из немногих исключений из правил, но даже он нам демонстрирует, что на основании спектра графа, полученного путем сингулярного разложения матриц смежности графов, может использоваться для нахождения подобия между различными структурами. При этом в отличие от программных продуктов *Hewlett-Packard*, которые используют промежуточные состояния для хранения и анализа информации в структурированных БД, описанный метод позволяет работать напрямую с *n* мерным массивом данных и сохранять результаты как с использованием структурированных баз данных, так и базы данных основанные на *NoSql* подходе, что в свою очередь способствует не только сбалансированному распределению нагрузки, но и более эффективному и быстрому процессу анализа данных.

Литература

1. Chung F.R.K. Spectral graph theory. – AMS. – 1997. – 207 p.
2. Shokoufandeh A., Dickinson S.J., Siddiqi K., Zucker S.W. Indexing using a spectral encoding of topological structure // Int'l Conf. Computer Vision and Pattern Recognition. – 1999. – Vol. 2. – pp. 491 – 497.
3. Zakharov A., Zhiznyakov A. Synthesis of threedimensional models from drawings based on spectral graph theory // Applied Mechanics and Materials. – 2015. – Vol. 756. – pp. 598 – 603.
4. Тужилкин А.Ю. Распознавание и реконструкция 3D-объектов по спутниковым изображениям на основе сравнения спектров графов // Фундаментальные исследования. – 2015. – № 2-17. – С. 3727-3732; URL: fundamental-research.ru/ru/article/view?id=37846.
5. Г.Е. Засядко, А.В. Карпов, Проблемы разработки графовых баз данных // Инженерный вестник Дона, 2017, №1. URL: ivdon.ru/ru/magazine/archive/n1y2017/3994.
6. С.В. Астанин, Н.В. Драгныш, Н.К. Жуковская, Вложенные метаграфы как модели сложных объектов // Инженерный вестник Дона, 2012, №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1434.
7. Ian Robinson, Jim Webber, Graph Databases. O'Reilly, 2015. pp. 8 - 10.
8. Umeyama S. An eigendecomposition approach to weighted graph matching problems // IEEE transactions on pattern analysis and machine intelligence. – 1988. – Vol. 10, № 5. – pp. 695 – 703
9. Gavin Powell, Beginning Database Design. Wrox, 2006. p. 219.
10. Niklaus Wirth, Algorithms and Data Structures. Prentice-Hall, Inc, 1986. pp. 109 – 111.

References

1. Chung F.R.K. Spectral graph theory. AMS. 1997. 207 p.
2. Shokoufandeh A., Dickinson S.J., Siddiqi K., Zucker S.W. Int'l Conf. Computer Vision and Pattern Recognition. 1999. Vol. 2. pp. 491 – 497.
3. Zakharov A., Zhiznyakov A. Applied Mechanics and Materials. 2015. Vol. 756. pp. 598 – 603.
4. Tuzhilkin A.Ju. Fundamental'nye issledovanija. 2015. № 2-17. pp. 3727-3732; URL: fundamental-research.ru/ru/article/view?id=37846.
5. Zasjadko G.E., Karpov A.V. Inženernyj vestnik Dona (Rus). 2017, №1. URL: ivdon.ru/ru/magazine/archive/n1y2017/3994.

6. Astanin S.V., Dragnysh N.V., Zhukovskaja N.K., Inženernyj vestnik Dona (Rus). 2012, №4. URL: ivdon.ru/ru/magazine/archive/n4p2y2012/1434.

7. Ian Robinson, Jim Webber, Graph Databases. O'Reilly, 2015. pp. 8 – 10.

8. Umeyama S. IEEE transactions on pattern analysis and machine intelligence. 1988. Vol. 10, № 5. pp. 695 – 703

9. Gavin Powell, Beginning Database Design. Wrox, 2006. p. 219.

10. Niklaus Wirth, Algorithms and Data Structures. Prentice-Hall, Inc, 1986. pp. 109 – 111.