Распределение научных направлений развития научно-технического центра в нефтегазовой отрасли на основании графа соавторства

Ф.В.Краснов, М.М.Хасанов

***Аннотация*— Планирование научно-технического развития исследовательской организации не должно идти в отрыве от реального положения дел. Такие явления как организационная инертность, диверсификация исследований и увлечение созданием ИТ-продуктов могут существенно исказить любые стратегии и планы развития. Тем не менее исполнимость планов является важной характеристикой развития, существенно подкрепляющей мотивацию персонала на результат. Поэтому постановка выполнимых задач имеет существенной значение.**

Количественных инструментов оценки выполнения научно-исследовательских работ не может быть достаточно. Формально-бумажная отчетность по НИР не способна отразить увлеченность и вовлеченность исследователей в работу. В то время как, малые формы исследовательских работ такие как презентация на научно-технической конференции или научная статья в рецензируемом периодическом издании требуют существенно более неформального отношения со стороны исследователей. Анализ развития научно-технической организации на основе публикационной активности является распространенной практикой. Многие исследования анализируют корпусы текстов научных статей и делают заключения о трендах развития. Текстовые данные обладают высоким уровнем шума и даже современные методы анализа LDA, Word2Vec выдают точные прогнозы только на основании огромных объемов текстов, которые не всегда имеются у небольших организаций. А именно небольшие научно-исследовательские организации больше страдают от неточности планирования научной деятельности.

Авторы данного исследования предлагают воспользоваться преимуществами анализа статей (презентаций) на основе двудольного графа соавторства для выделения научных направлений с целью дальнейшей их оценки и планирования.

***Ключевые слова*— кластеризация, граф соавторства, характеристики научно-исследовательской деятельности, наукометрия, организационные гипотезы, модель.**

Статья получена 15 января 2018.
Ф.В.Краснов, к.т.н., эксперт, ООО «Газпромнефть НТЦ», 190000 г. Санкт-Петербург, набережная реки Мойки д.75-79,. email:krasnov.fv@gazprom-neft.ru, orcid.org/0000-0002-9881-7371, РИНЦ 8650-1127
М.М.Хасанов, д.т.н., профессор, Директор дирекции по технологиям ПАО «Газпром нефть», 190000, Россия, Санкт-Петербург, ул. Почтамтская, д.3-5, (email:Khasanov.MM@gazprom-neft.ru)

I. Введение

Современный фокус применения научных подходов к управленческим решениям приобретает все большую актуальность. С ростом объемов данных традиционные аналитические средства руководителей организаций становятся все менее эффективными. С другой стороны необходимого объема данных для устойчивой работы современных алгоритмов часто бывает недостаточно. В авангарде этой тенденции возникает задача адаптации и создания новых эвристик для таких классических задач как кластеризация для применения в организационной среде.

Кластеризация данных на основании статической модели получила развитие с открытием таких алгоритмов как PAM [1], CLARANS [2], DBSCAN [3], CURE [4] и ROCK [5]. Тем не менее в последнее время особое внимание привлечено к алгоритмам кластеризации на основании динамической модели, например, CHAMELEON [6]. Основная идея алгоритма CHAMELEON [6] в использовании метрик близости графа, построенного на основании набора кластеризуемых данных с помощью метода «k наиболее близких соседей» (KNN). Метрики графа оказываются более эффективными для разбиения данных «сверху в низ» в случае сложных объектов (Рисунок 1).

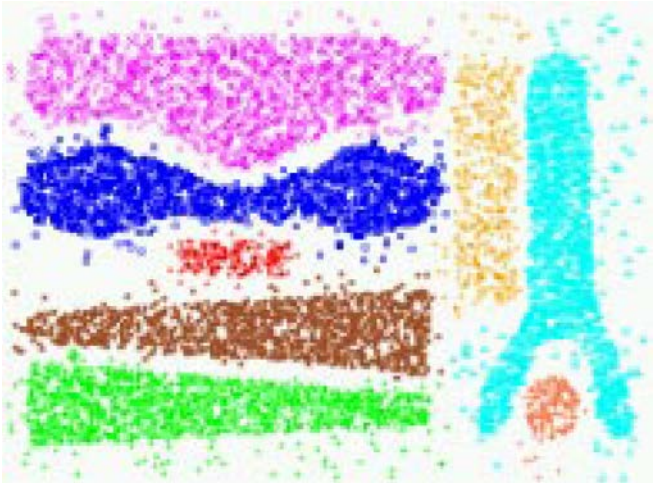

Рис. 1 Пример использования алгоритма CHAMELEON [6] для кластеризации сложных объектов.

Разнообразие алгоритмов кластеризации не умаляет задачу оценки их качества. Но в условиях

ограниченного количества данных и для обеспечения управленческих решений качество кластеризации должно иметь не только математически обоснованную, но и уверенную наглядную составляющую. Другими словами, чтобы «с одного взгляда было понятно» и не нужно было вникать в формулы. Таковы требования современного бизнеса.

II. Методология исследования

С формальной точки зрения нам необходимо решить задачу обучения без учителя (unsupervised machine learning) для графа соавторств, отнести кластеры к определенным тематикам и выявить изменения в кластерах со временем.

Кластеризация графа соавторства может быть осуществлена на основании различных метрик вершин:

- Degree centrality
- Betweenness centrality
- Closeness centrality
- Harmonic centrality
- Clustering

Рассмотрим содержательный смысл метрики *Betweenness centrality* применительно к задаче кластеризации графа соавторств научно-технической организации. Метрика *Betweenness centrality* характеризует насколько данный узел важен для связанности графа. Связи в графе соавторств отражают совместную исследовательскую работу. Графы соавторств не всегда являются связанными, обычно это несколько связанных компонент разного размера.

Связанные компоненты являются естественными кластерами. Небольшие связанные компоненты отражают начальные инициативы – это первые статьи сотрудников. Но главная связанная компонента может содержать 90% вершин графа соавторства и нуждается в отдельном подходе к кластеризации.

Для выделения кластеров из главной связанной компоненты графа соавторств возможно использовать методику искусственного удаления вершин с наибольшей метрикой *Betweenness centrality*. При каждом таком удалении вершины граф может распадаться на несколько несвязанных компонент. На рисунке 2 приведена модель такого разделения.

Каждую из получившихся при таком распаде связанных компонент можно анализировать на однородность тематики на основании текстов статей, которыми она образована. В результате нескольких итераций мы получим набор кластеров.

Предложенный авторами метод является эвристическим и нуждается в проверке по определенному формальному критерию. Для задач кластеризации таким критерием принято считать метрики близости объектов в кластере и расстояния между объектами в разных кластерах.

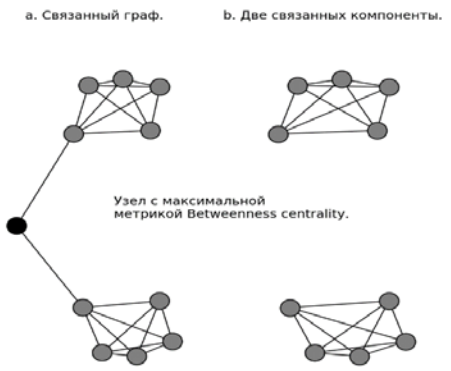

Рис. 2 Модель разделения графа. а.Связанный изначальный граф. б.Тот же граф, но после удаления вершины с наибольшей метрикой Betweenness centrality уже представляет две связанных компоненты.

Сходимость предложенного авторами метода обеспечивается путем поиска минимума функционала ошибок определения кластеров:

$$|WSS - BSS| \to min, \quad (1)$$

где WSS – это функция связанности кластера C_i:

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2, \quad (2)$$

а BSS – это функция разделения кластеров C_i:

$$BSS = \sum_i |C_i|(m - m_i)^2, \quad (3)$$

где $|C_i|$- это размер кластера C_i.

Междисциплинарные исследования приводят к тому, что статьи будут относится к нескольким тематикам, так что полученные кластера будут пересекающимися – не эксклюзивными.

III. Результаты

В качестве объекта исследования была выбрана публикационная активность НТЦ «Газпромнефть». Данные были получены из открытой электронной библиотеки OnePetro международного сообщества нефтегазовых инженеров (SPE). После очистки было получено 172 статьи.

Построим прогноз на основании графа соавторства. Для этого построим двудольный граф соавторства с вершинами: автор (479) и статья (171). Авторы обладают техническими компетенциями, статьи характеризуются названием, годом издания и ключевыми словами.

Полученный граф соавторства имеет 26 связанных компонент наибольшая из которых содержит 556 вершин, а остальные не более 8. Связанные компоненты с количеством узлов до 8 являются считать первыми статьями сотрудников.

Рассмотрим наибольшую связанную компоненту (556 вершин). Выделим подграф из основного графа соавторства на основании узлов, относящихся к

наибольшей связанной компоненте. Получившийся подграф отображен на рисунке (Рисунок 3).

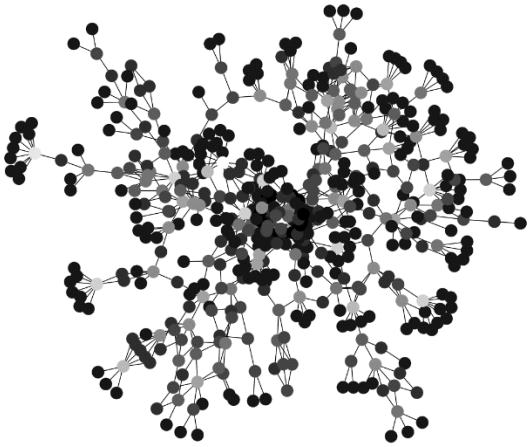

Рис. 3 Подграф наибольшей связанной компоненты графа соавторства.

Рассчитаем для полученного подграфа метрику *Betweenness centrality*. Полученные значения *Betweenness centrality* отображены на рисунке (Рисунок 4). Нулевые значения *Betweenness centrality* не отображены.

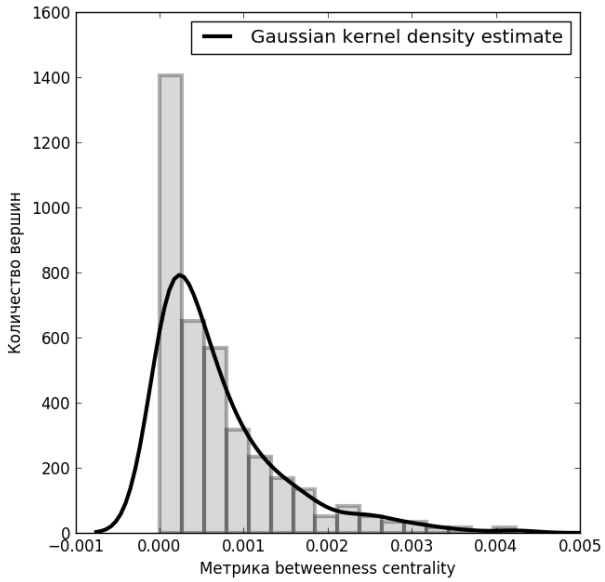

Рис. 4 Гистограмма значений Betweenness centrality для подграфа наибольшей связанной компоненты графа соавторств.

Как мы видим из рисунка 4, значения метрики *Betweenness centrality* в третьем квартиле принадлежат всего 23 вершинам, что составляет менее 5 % от всего количества вершин.

Применим алгоритм искусственного удаления вершин с наибольшим значением метрики *Betweenness centrality*. На рисунке 5 отображена зависимость количества связанных компонент от количества искусственно удаленных вершин.

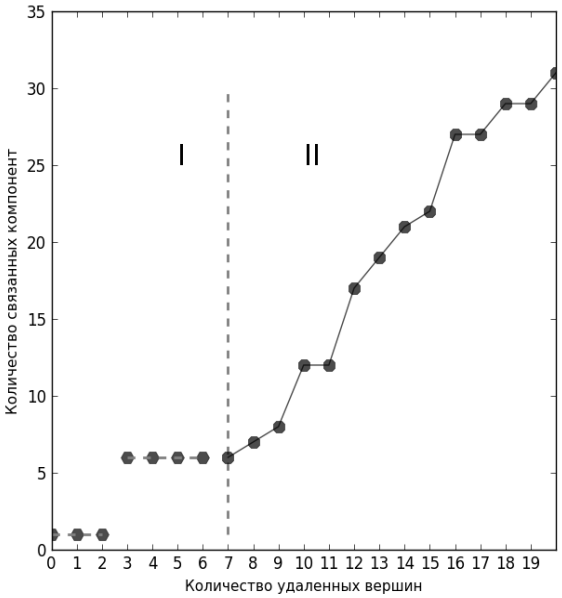

Рис. 5 Зависимость количества связанных компонент от количества искусственно удаленных вершин.

При удалении вершин поведение графа происходит в двух режимах:

I. Удержание связанности
II. Экспоненциальный распад

Отличительной чертой режима I является то, что граф остается связанным при удалении вершин с высокими значениями метрики *Betweenness centrality*. Это означает, что удаляемые вершины не являются единственными связующими между кластерами.

Отличительной особенностью режима II является следование степенной модели распада графа, когда удаление каждого узла вызывает степенной рост появления новых связанных компонент.

Рассмотрим более подробно вторую половину режима I алгоритма, когда граф разделился на 6 связанных компонент. Размеры этих компонент составляют 511, 34, 1, 1, 1, 1. И среди них ярко выраженное направление исследований по Теме 1 представлено именно компонентой с 34 узлами, представленной на рисунке (Рисунок 6).

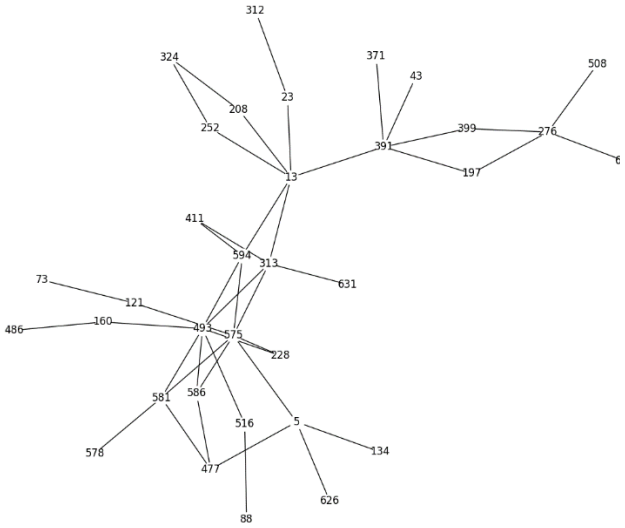

Рис. 6 Кластер исследователей по Теме 1, выделенный в результате применения метода удаления вершин в наибольшим значением метрики Betweenness centrality.

Мы рассмотрели выделение одного кластера подробно. Полный алгоритм выделения кластеров будет состоять из следующих шагов:

1. Построение двудольного графа соавторств: G
2. Определение метрики *Betweenness centrality* для графа G
3. Определение вершины с максимальной метрикой $N_{max(Betweenness\ centrality)}$
4. Удаление вершины $Nmax_{(Betweenness\ centrality)}$ из графа G
5. Получение списка связанных компонент графа G
6. Вычисление метрики качества полученных кластеров *BSS* и W*SS*
7. Далее алгоритм повторяется для каждой связанной компоненты
8. Алгоритм завершается, когда все связанные компоненты представляют кластеры удовлетворительного качества.

Для выбранного графа соавторства были выделены 16 кластеров.

Для вычисления значений *BSS* и W*SS* на основании текстов статей было использовано векторное представление текста статьи (VSM). Каждая статья представлена в виде вектора со значениями метрики BM25 [7] для каждого слова. Статьи рассматривались как «мешок слов» (bag of words). Для измерения дистанции между векторными представлениями статей была использована косинусная мера.

На рисунке (Рисунок 7) изображена матрица раздельности кластеров *BSS*.

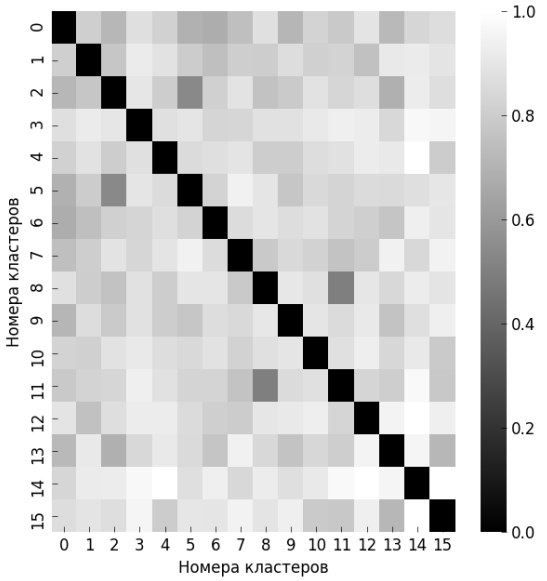

Рис. 7 Матрица раздельности кластеров. По осям отображены номера кластеров. В ячейках значения функции BSS.

Для сравнения полученной кластеризации статей была проведена кластеризация с помощью алгоритма *KMeans*, показавшая схожие результаты (Рисунок 8).

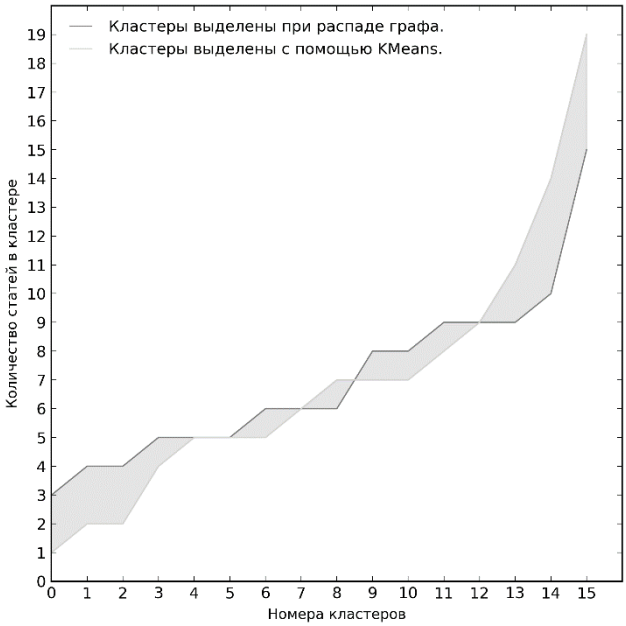

Рис. 8 Сравнение предлагаемого в статье алгоритма кластеризации и алгоритма KMeans.

С помощью *KMeans* была произведена кластеризация статьей, а затем из графа соавторства были определены кластеры авторов на основании полученных кластеров статей.

IV. Заключение

Авторами предложен метод выделения направлений научных исследований на основе графа соавторства. Содержательно предложенный метод относится к top-down алгоритмам кластеризации. В качестве критерия выделения кластеров выбрана метрика *Betweenness centrality*. В качестве критерия проверки качества

кластеров выбрана метрика близости членов кластера и метрика удаленности различных кластеров на основе тематик научных статей, входящих в граф соавторства. Результатом применения предложенного метода является укрупненное виденье научных направления развития организации, сделанное на основе публичных данных о публикационной активности сотрудников.
Разработанный авторами метод выделения направлений научных исследований на основе графа соавторства опробован на Научно-техническом центра ГазпромНефть. В результате выделены 16 кластеров, характеризующих деятельность организации.

Важными особенностями разработанного авторами метода выделения направлений научных исследований на основе графа соавторства являются следующие:

1. Рекурсивность алгоритма позволяет работать с графами различных порядков.
2. «Жадный» алгоритм определения качества кластеров позволяет корректировать оптимизацию на каждом шаге.
3. Применение двудольного построения графа соавторства позволяет анализировать различные проекции.
4. Работа на основании публичных данных дает широкие возможности для применения в бизнес разведке.

Новизна предложенного авторами метода выделения направлений научных исследований на основе графа соавторства состоит в использовании двудольного построения графа соавторства и в динамической модели кластеризации, использующей структурные метрики графа соавторства и метрики близости текстов научных статей.

Библиография

1. Kaufman L., Rousseeuw P. J. Finding groups in data: an introduction to cluster analysis. – John Wiley & Sons, 2009. – Т. 344.
2. Ng R. T., Han J. CLARANS: A method for clustering objects for spatial data mining //IEEE transactions on knowledge and data engineering. – 2002. – Т. 14. – №. 5. – С. 1003-1016.
3. Ester M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise //Kdd. – 1996. – Т. 96. – №. 34. – С. 226-231.
4. Guha S., Rastogi R., Shim K. Cure: an efficient clustering algorithm for large databases //Information Systems. – 2001. – Т. 26. – №. 1. – С. 35-58.
5. Guha S., Rastogi R., Shim K. ROCK: A robust clustering algorithm for categorical attributes //Information systems. – 2000. – Т. 25. – №. 5. – С. 345-366.
6. Karypis G., Han E. H., Kumar V. Chameleon: Hierarchical clustering using dynamic modeling //Computer. – 1999. – Т. 32. – №. 8. – С. 68-75.
7. Lv Y., Zhai C. X. Adaptive term frequency normalization for BM25 //Proceedings of the 20th ACM international conference on Information and knowledge management. – ACM, 2011. – С. 1985-1988.

Allocation of the scientific directions of development of science and technologies center in oil and gas industry based on the co-authorship network

Fedor Krasnov, Mars Khasanov

***Abstract* — Planning of scientific and technological development research organizations should not be divorced from the real situation. Phenomena such as organizational inertia, diversification of research and the hobby of creating IT products can significantly distort any strategies and development plans. However, the feasibility of the plans is an important characteristic of development, significantly reinforcing the motivation for the result. So setting realistic goals is essential.**

Quantitative tools for assessing the implementation of scientific research may not be enough. The formal paper reporting the research is not able to reflect the passion and engagement of researchers in the work. While, small-scale research such as presentations at scientific conference or a scientific paper in a peer-reviewed periodical journal require significantly more casual attitude on the part of researchers.

Analysis of the development of scientific and technological organizations based on publication activity is a common practice. Many studies analyse a corpus of texts of scientific articles and draw conclusions about the development trends. Textual data have a high noise level and even modern methods of analysis such as LDA and Word2Vec give accurate predictions only based on huge volumes of texts that were not always available to small organizations. Namely, a small research organization suffer more from inaccuracies of the planning of scientific activities.

The authors of this study propose to take advantage of the analysis of the articles (presentations) based on a bipartite graph of co-authorship to highlight research directions for further evaluation and planning.

***Keywords*— clustering, co-authorship network, the characteristics of research activities, scientometrics, organizational hypotheses, model.**