

Представление агрегированной структуры текстов с помощью обобщенной контекстно-зависимой теоретико-графовой модели

Н.Д. Москин

Петрозаводский государственный университет

Аннотация: В работе рассматриваются вопросы, связанные с применением теоретико-графовых моделей при анализе текстов. Одной из задач является агрегирование подобных моделей для выявления более «простых» графов, вершины которых соответствуют подмножествам вершин первоначальной модели, а ребра отражают «сильные связи» между вершинами. На примере сюжета русской народной сказки показано, как можно построить агрегированную модель с заданным порогом значимости и представить ее для дальнейшего анализа. Для проведения экспериментов был построен набор теоретико-графовых моделей для сказочных сюжетов из сборника А. М. Афанасьева с помощью информационной системы «Фольклор», где был усовершенствован модуль агрегации графов.

Ключевые слова: анализ текстов, теоретико-графовая модель, агрегация, порог значимости, формат хранения, фольклорный текст, сказочный сюжет, информационная система «Фольклор».

Введение

Графы часто применяются в анализе текста благодаря своей способности эффективно представлять структуру взаимосвязей между словами, предложениями, абзацами и более объемными фрагментами текста [1, 2]. Рассмотрим некоторые методы анализа текста с использованием графовых моделей:

1. *Семантический граф.* Семантический граф позволяет визуализировать смысловые связи между различными элементами текста. Вершины графа представляют собой отдельные понятия (объекты), а ребра отражают отношения между ними. Это помогает выявить скрытые закономерности и взаимосвязи, что может быть полезно при работе с большими объемами информации, такими как научные статьи, юридические документы или литературные произведения.

2. *Задача выделения ключевых понятий.* Один из способов автоматического извлечения ключевых понятий заключается в построении

графа, вершины которого соответствуют словам или терминологическим единицам, а вес каждого ребра определяется частотой совместного появления соответствующих вершин в тексте. Анализируя этот граф, можно определить наиболее важные элементы текста, что в дальнейшем можно применить в задачах индексирования или классификации документов.

3. *Анализ структуры предложений.* Теоретико-графовая модель также применяется для анализа синтаксической структуры предложений. В этом случае вершинам сопоставляются слова определенной части речи, а ребрами обозначаются грамматические связи между ними. Такой подход способствует пониманию скрытых связей внутри текста и улучшает качество обработки естественного языка.

Важной проблемой является хранение слабоструктурированных данных, а также их извлечение с использованием графов, особенно с учетом все возрастающих объемов информации [3, 4]. В данной работе рассматривается задача представления агрегированной структуры текста с помощью обобщенной контекстно-зависимой теоретико-графовой модели (на примере фольклорных текстов).

Теоретико-графовые модели фольклорного текста

При исследовании устного народного творчества одной из важных задач является классификация и систематизация типов сказок. Для этого можно использовать агрегированные теоретико-графовые модели [5]. Рассмотрим построение такой модели на примере текста «Сказка о молодце-удальце, молодильных яблоках и живой воде» (№176 из сборника А. М. Афанасьева [6]), сюжет которой включает большое количество персонажей и связей между ними. Сначала выполним сегментацию текста. Общее количество слов $K = 2761$, множество слов текста обозначим $W = \{w_k\}_{k=1}^K$ (индекс k соответствует порядку появления слова w_k в тексте). Слова могут

объединяться в упорядоченные подмножества $W_l \subset W$, которые соответствуют элементам графа ($l=1, 2, \dots, L$). Подмножество W_l может состоять как из одного слова, так и из совокупности слов.

Вопросы построения теоретико-графовых моделей и филологические подходы к задаче агрегирования изложены в [5] (далее будем придерживаться обозначений, принятых в этой работе). Здесь возникает задача объединения схожих элементов графа с целью сокращения его сложности без потери информативности, а также хранения полученной структуры. Обозначим ее набором $G=(V, H, E, \alpha, \beta, \mu, \gamma)$ для текста T (рекурсивное определение обобщенной контекстно-зависимой теоретико-графовой модели для решения задачи анализа текстов изложено в [1]).

В сюжете упоминаются 39 объектов, которые разделяются на 9 групп (см. таблицу №1). Отмечу, что выделение объектов в сказке и деление их на группы выполнено с опорой на классические труды в области фольклористики (например, [7, 8]). Обозначим вершины, соответствующие объектам, с помощью $V = \{v_i\}_{i=1}^n$, где число объектов $n = 39$. В данном случае через A обозначим множество всех групп, а $\alpha: V \rightarrow A$ как функцию, которая каждой вершине ставит в соответствие идентификатор ее группы. Например, объекту $H1$ («Иван-царевич») соответствует вершина v_9 , которая принадлежит группе «герой», т. е. $\alpha(v_9) = \langle H \rangle$.

Между объектами существуют связи (всего в данном тексте их $s=92$), которые отражаются на графе ребрами e_j . Отметим, что по числу ребер, инцидентных той или иной вершине, можно судить о значимости соответствующего персонажа. Отображение γ задает соответствие между вершинами и подмножествами слов в тексте $W_l \subset W$. Например, персонаж «Иван-царевич» (вершина v_9) встречается в сказке 69 раз в совершенно разных частях текста. Аналогично можно задать слова, соответствующие

ребрам модели. Номера слов в тексте определяют темпоральность (упорядоченность) элементов графа.

Таблица № 1

Соответствие объектов текста и вершин графа

Объект текста	Вершина графа	Группа	Объект текста	Вершина графа	Группа
<i>T1</i>	<i>v₁</i>	родственник	<i>N7</i>	<i>v₂₁</i>	награда
<i>T2</i>	<i>v₂</i>	родственник	<i>N8</i>	<i>v₂₂</i>	награда
<i>T3</i>	<i>v₃</i>	родственник	<i>N9</i>	<i>v₂₃</i>	награда
<i>T4</i>	<i>v₄</i>	родственник	<i>N10</i>	<i>v₂₄</i>	награда
<i>T5</i>	<i>v₅</i>	родственник	<i>N11</i>	<i>v₂₅</i>	награда
<i>T6</i>	<i>v₆</i>	родственник	<i>N12</i>	<i>v₂₆</i>	награда
<i>G1</i>	<i>v₇</i>	глупец	<i>П1</i>	<i>v₂₇</i>	помощник
<i>G2</i>	<i>v₈</i>	глупец	<i>П2</i>	<i>v₂₈</i>	помощник
<i>H1</i>	<i>v₉</i>	герой	<i>П3</i>	<i>v₂₉</i>	помощник
<i>A1</i>	<i>v₁₀</i>	антигерой	<i>П4</i>	<i>v₃₀</i>	помощник
<i>A2</i>	<i>v₁₁</i>	антигерой	<i>П5</i>	<i>v₃₁</i>	помощник
<i>A3</i>	<i>v₁₂</i>	антигерой	<i>П6</i>	<i>v₃₂</i>	помощник
<i>A4</i>	<i>v₁₃</i>	антигерой	<i>П7</i>	<i>v₃₃</i>	помощник
<i>A5</i>	<i>v₁₄</i>	антигерой	<i>Д1</i>	<i>v₃₄</i>	даритель
<i>N1</i>	<i>v₁₅</i>	награда	<i>Д2</i>	<i>v₃₅</i>	даритель
<i>N2</i>	<i>v₁₆</i>	награда	<i>Д3</i>	<i>v₃₆</i>	даритель
<i>N3</i>	<i>v₁₇</i>	награда	<i>R1</i>	<i>v₃₇</i>	препятствие
<i>N4</i>	<i>v₁₈</i>	награда	<i>R2</i>	<i>v₃₈</i>	препятствие
<i>N5</i>	<i>v₁₉</i>	награда	<i>V1</i>	<i>v₃₉</i>	антипомощник
<i>N6</i>	<i>v₂₀</i>	награда			

В [1] изложено, как можно представить исходную структуру в терминах обобщенной контекстно-зависимой теоретико-графовой модели. Далее покажем, как определить агрегированную модель, которая храниться совместно с исходной и ее дополняет.

Представление агрегированной структуры текста

Задача агрегирования заключается в том, чтобы на основе исходной (более «сложной») теоретико-графовой модели построить более «простую»

модель, в которой отражаются «основные потоки связей». Обозначим ее G^* . При этом простой граф образуют m вершин, которым соответствуют подмножества вершин исходного графа V_j , и ребра, которые определяют «сильные» связи.

При агрегировании графа вершины объединяются в подмножества $V_j \subset V$. Подмножества не пересекаются и их объединение образует исходное множество V . Пусть задана матрица попарных показателей связи $A = \{a_{ij}\}_{i,j=1}^n$ между n вершинами из множества V . Определим элементы матрицы A следующим образом: $a_{ii} = \mu(v_i)$ для $\forall i = 1, 2, \dots, n$. Если существует ребро e_t , которое определяет связь от вершины v_i к вершине v_j для $\forall i, j = 1, 2, \dots, n$ и $i \neq j$, то $a_{ij} = \mu(e_t)$ при $\forall t = 1, 2, \dots, s$, иначе $a_{ij} = 0$. Функция μ принимает значения от 0 до 1 и задает нечеткость объектов теоретико-графовой модели (для данного примера нечеткие элементы, т. е. допускающие множественную интерпретацию, отсутствуют). Задача выявления структуры матрицы связей состоит в том, чтобы максимизировать функционал I :

$$I = \sum_{x=1}^m \sum_{y=1}^m \left| \sum_{v_i \in V_x} \sum_{v_j \in V_y} (a_{ij} - \alpha) \right|,$$

где α – порог значимости показателей связи, который делит все связи на положительные (значимые) и отрицательные (незначимые).

На рис.1 представлена теоретико-графовая модель фрагмента текста «Сказка о молодце-удальце, молодильных яблоках и живой воде», начиная с фразы «Вот старшие братья пошли к своему отцу домой...» и заканчивая словами «... тут сокол кашлянул и выкинул его икры и платье.», которая была агрегирована до пяти вершин (первоначальное число вершин фрагмента равно 15). В данном случае порог значимости был выбран $\alpha=0,1$ [5].

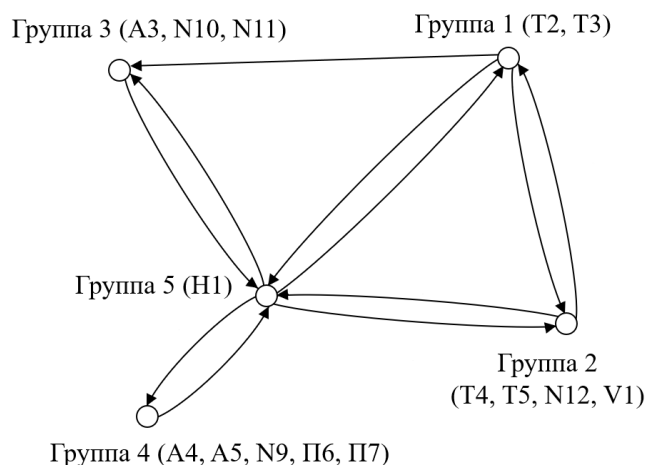


Рис. 1. – Агрегированный граф «сильных» связей (5 вершин)

Тогда вершины объединим в пять подмножеств: $V_1 = \{v_2, v_3\}$, $V_2 = \{v_4, v_5, v_{26}, v_{39}\}$, $V_3 = \{v_{12}, v_{24}, v_{25}\}$, $V_4 = \{v_{13}, v_{14}, v_{23}, v_{32}, v_{33}\}$, $V_5 = \{v_9\}$ которые соответствуют группам 1-5 на рис. 1. Несмотря на то, что пятая группа состоит из одного элемента, необходимо определить V_5 , так как $V = \bigcup_{j=1}^5 V_j$.

Пять подмножеств образуют четыре подструктуры G_1, G_2, G_3, G_4, G_5 .

- $G_1 = (V_1, H_1, E_1)$, $V_1 = \{v_2, v_3\}$, $E_1 = \emptyset$, $H_1 = \emptyset$;
- $G_2 = (V_2, H_2, E_2)$, $V_2 = \{v_4, v_5, v_{26}, v_{39}\}$, $E_2 = \{e_{51}, e_{52}, e_{60}, e_{61}, e_{62}, e_{69}, e_{71}, e_{72}, e_{80}\}$, $H_2 = \emptyset$;
- $G_3 = (V_3, H_3, E_3)$, $V_3 = \{v_{12}, v_{24}, v_{25}\}$, $E_3 = \emptyset$, $H_3 = \emptyset$;
- $G_4 = (V_4, H_4, E_4)$, $V_4 = \{v_{13}, v_{14}, v_{23}, v_{32}, v_{33}\}$, $E_4 = \{e_{43}\}$, $H_4 = \emptyset$;
- $G_5 = (V_5, H_5, E_5)$, $V_5 = \{v_9\}$, $E_5 = \emptyset$, $H_5 = \emptyset$.

В данном примере H_j определяются как пустые множества, поскольку нет иерархических (вложенных) структур. Однако если необходимо рассмотреть вторую агрегированную теоретико-графовую модель G^{**} с большим числом вершин (например, с десятью вершинами) и представить ее совместно с первой агрегированной моделью G^* , то некоторые элементы G^* могут включать в себя элементы G^{**} , т. е. образуется вложенность, которую можно задать с помощью H_j .

Далее введем множество связей в соответствии с весами «сильных» связей агрегированного графа (см. таблицу №2).

Таблица № 2

«Сильные» связи агрегированной модели

№	Связи	Вес связи	Функция принадлежности
1	$e_{93} = (V_1, V_2)$	$\beta(e_{93}) = 1,2$	$\mu(e_{93}) = 1$
2	$e_{94} = (V_1, V_3)$	$\beta(e_{94}) = 3,4$	$\mu(e_{94}) = 1$
3	$e_{95} = (V_1, V_5)$	$\beta(e_{95}) = 5,8$	$\mu(e_{95}) = 1$
4	$e_{96} = (V_2, V_1)$	$\beta(e_{96}) = 3,2$	$\mu(e_{96}) = 1$
5	$e_{97} = (V_2, V_5)$	$\beta(e_{97}) = 2,6$	$\mu(e_{97}) = 1$
6	$e_{98} = (V_3, V_5)$	$\beta(e_{98}) = 0,7$	$\mu(e_{98}) = 1$
7	$e_{99} = (V_4, V_5)$	$\beta(e_{99}) = 4,5$	$\mu(e_{99}) = 1$
8	$e_{100} = (V_5, V_1)$	$\beta(e_{100}) = 4,8$	$\mu(e_{100}) = 1$
9	$e_{101} = (V_5, V_2)$	$\beta(e_{101}) = 4,6$	$\mu(e_{101}) = 1$
10	$e_{102} = (V_5, V_3)$	$\beta(e_{102}) = 4,7$	$\mu(e_{102}) = 1$
11	$e_{103} = (V_5, V_4)$	$\beta(e_{103}) = 6,5$	$\mu(e_{103}) = 1$

В целом подобная структура напоминает вложенный метаграф. В [9] определяется это понятие как обобщение для графов, гиперграфов и метаграфов. Здесь каждый узел представляет собой отдельный подграф, состоящий из собственных вершин и ребер. Для хранения и последующего анализа таких моделей необходимо специализированное программное обеспечение.

Для этой цели в информационной системе «Фольклор» [1] был усовершенствован модуль агрегации графов с последующим хранением получившихся структур в формате обобщенной контекстно-зависимой теоретико-графовой модели. Данный модуль позволяет: загрузить начальные данные модели; выполнить расчеты в модуле «Агрегация», где строится начальное распределение вершин исходного графа на заданное число подмножеств, определяющих вершины «простого» графа; выполнить расчеты в модуле «Структура», который производит локальное улучшение

начального приближения; вывести параметры агрегации и полученные результаты. Апробация предложенных подходов была выполнена на выборке сказочных сюжетов (тексты из сборника А. М. Афанасьева [6]), для которых были построены теоретико-графовые модели и соответствующие им агрегированные структуры (при значении порога значимости показателей связи $\alpha = 0,1; 0,2; 0,3$). Отметим, что подобные структуры можно использовать не только для исследования народных, но также и авторских сказок.

Таким образом, рассмотренная модель позволяет сохранить не только первоначальный вариант разбора сказочного сюжета в виде графа, но и его агрегированную структуру, содержащую наиболее значимые группы объектов и связей. В дальнейшем ее можно исследовать с помощью современных моделей машинного обучения для решения актуальных задач в области компьютерной лингвистики и фольклористики. Например, в [10] отмечается большой потенциал в использовании графовых нейронных сетей для задач из области обработки естественного языка (англ. *Natural Language Processing* – NLP).

Литература

1. Москин Н.Д. Теоретико-графовые модели, методы и программные средства интеллектуального анализа текстовой информации на примере фольклорных и литературных произведений. Дис. ... докт. техн. Наук. 05.13.18. Петрозаводск. 2022. 370 с.
2. Castillo E., Cervantes O., Vilarino D. Text Analysis Using Different Graph-Based Representations // *Computación y Sistemas*. 2018. Vol. 21. № 10. Pp. 581-599.
3. Клименков С.В., Николаев В.В., Харитонов А.Е., Гаврилов А.В., Письмак А.Е., Покид А.В. Применение семантической сети для хранения

слабоструктурированных данных. Инженерный вестник Дона. 2020. №2. URL: ivdon.ru/magazine/archive/n2y2020/6339/.

4. Засядко Г.Е., Карпов А.В. Проблемы разработки графовых баз данных. Инженерный вестник Дона. 2017. №1. URL: ivdon.ru/magazine/archive/n1y2017/3994/.

5. Москин Н.Д., Лебедев А.А. Объяснительные возможности агрегирования теоретико-графовых моделей сказочных сюжетов в практике филологического анализа текста. Terra Linguistica. 2025. Т.16. №2. С. 85–100.

6. Афанасьев А.М. Народные русские сказки А. Н. Афанасьева. М.: Государственное Издательство Художественной литературы (Гослитиздат). 1957. Т. 1. 516 с.

7. Пропп В. Морфология сказки. Ленинград: Academia. 1928. 152 с.

8. Гаазе-Рапопорт М.Г. Поиск вариантов в сочинении сказок // Зарипов Р.Х. Машинный поиск вариантов при моделировании творческого процесса. М. Наука. 1983. С. 213–223.

9. Астанин С.В., Драгныш Н.В., Жуковская Н.К. Вложенные метаграфы как модели сложных объектов. Инженерный вестник Дона. 2012. №4 (часть 2) URL: ivdon.ru/magazine/archive/n4p2y2012/1434/.

10. Wu L., Chen Y., Shen K., Guo X., Gao H., Li S., Pei J., Long B. Graph Neural Networks for Natural Language Processing: A Survey // *ArXiv* abs/2106.06090. 2021. 127 p.

References

1. Moskin N.D. Teoretiko-grafovyye modeli, metody i programmnyye sredstva intellektual'nogo analiza tekstovoy informatsii na primere fol'klornykh i literaturnykh proizvedeniy [Graph-theoretical models, methods and software tools for the intellectual analysis of textual information on the example of folklore and literary works]. Dis. ... dokt. tekhn. Nauk. 05.13.18. Petrozavodsk. 2022. 370 p.

2. Castillo E., Cervantes O., Vilarino D. Text Analysis Using Different Graph-Based Representations. *Computación y Sistemas*. 2018. Vol.21. №10. Pp. 581-599.
3. Klimenkov S.V., Nikolayev V.V., Kharitonova A.E., Gavrilov A.V., Pis'mak A.E., Pokid A.V. *Inzhenernyy vestnik Dona*. 2020. №2. URL: ivdon.ru/magazine/archive/n2y2020/6339/.
4. Zasyadko G.E., Karpov A.V. *Inzhenernyy vestnik Dona*. 2017. №1. URL: ivdon.ru/magazine/archive/n1y2017/3994/.
5. Moskin N.D., Lebedev A.A. *Terra Linguistica*. 2025. Vol.16. №2. Pp. 85–100.
6. Afanas'yev A.M. *Narodnyye russkiye skazki* A. N. Afanas'yeva [Russian folk tales by A. N. Afanasyev]. Moskva: Gosudarstvennoye Izdatel'stvo Khudozhestvennoy literatury (Goslitizdat). 1957. Vol.1. 516 p.
7. Propp V. *Morfologiya skazki* [Morphology of a fairy tale]. Leningrad: Academia. 1928. 152 p.
8. Gaaze-Rapoport M.G. Zaripov R.H. *Mashinnyj poisk variantov pri modelirovanii tvorcheskogo protsessa*. Moskva. Nauka. 1983. Pp. 213–223.
9. Astanin S.V., Dragnysh N.V., Zhukovskaya N.K. *Inzhenernyy vestnik Dona*. 2012. №4 (CH. 2). URL: ivdon.ru/magazine/archive/n4p2y2012/1434/.
10. Wu L., Chen Y., Shen K., Guo X., Gao H., Li S., Pei J., Long B. Graph Neural Networks for Natural Language Processing: A Survey. *ArXiv abs/2106.06090*. 2021. 127 p.

Автор согласен на обработку и хранение персональных данных.

Дата поступления: 18.10.2025

Дата публикации: 27.11.2025