DOI: 10.15514/ISPRAS-2022-34(1)-6

Обобщенная контекстно-зависимая теоретико-графовая модель фольклорных и литературных текстов

Н.Д. Москин, ORCID: 0000-0001-5556-5349 <moskin@petrsu.ru>
А.А. Рогов, ORCID: 0000-0002-8815-7920 <rogov@petrsu.ru>
Р.В. Воронов, ORCID: 0000-0003-0104-6409 <rvoronov@petrsu.ru>
Петрозаводский государственный университет,
185910, Россия, г. Петрозаводск, пр. Ленина, д. 33

Аннотация. Одной из проблем при автоматической обработке текстов является их атрибуция. Под этим термином понимают установление атрибутов текстового произведения (определение авторства, времени создания, места записи и др.). В статье представлена обобщенная контекстно-зависимая теоретико-графовая модель, предназначенная для анализа фольклорных и литературных текстов. Минимальной структурной единицей модели (примитивом) является слово. Множества слов объединяются в вершины, причем одно и то же слово может иметь отношение к разным вершинам. Ребра и графовые подструктуры отражают лексические, синтаксические и семантические связи текста. Характеристиками модели являются ее нечеткость, иерархичность и темпоральность. В качестве примеров приводятся иерархическая теоретико-графовая модель составляющих (на примере литературных произведений А. С. Пушкина), темпоральная теоретико-графовая модель сказочного сюжета (на примере русских волшебных сказок А. М. Афанасьева) и нечеткая теоретико-графовая модель «сильных» связей грамматических классов (на примере анонимных статей из дореволюционных журналов «Время», «Эпоха» и еженедельника «Гражданин», которые редактировал Ф. М. Достоевский). Модель строится таким образом, чтобы в дальнейшем ее можно было исследовать с помощью методов искусственного интеллекта (например, деревьев решений или нейронных сетей). Для этой цели в информационной системе «Фольклор» был разработан формат для хранения подобных данных, а также реализованы процедуры для ввода, редактирования и анализа текстов и их теоретико-графовых моделей.

Ключевые слова: теоретико-графовая модель; атрибуция текстов; лексика; синтаксис; семантика; нечеткий граф; иерархический граф; темпоральный граф; информационная система «Фольклор»

Для цитирования: Москин Н.Д., Рогов А.А., Воронов Р.В. Обобщенная контекстно-зависимая теоретико-графовая модель фольклорных и литературных текстов. Труды ИСП РАН, том 34, вып. 1, 2022 г., стр. 73-86. DOI: 10.15514/ISPRAS–2022–34(1)–6

Generalized context-dependent graph-theoretic model of folklore and literary texts

N.D. Moskin, ORCID: 0000-0001-5556-5349 <moskin@petrsu.ru>
A.A. Rogov, ORCID: 0000-0002-8815-7920 <rogov@petrsu.ru>
R.V. Voronov, ORCID: 0000-0003-0104-6409 <rvoronov@petrsu.ru>
Petrozavodsk State University,
33, Lenin st., Petrozavodsk, 185910, Russia

Abstract. One of the problems of automatic text processing is their attribution. This term is understood as the establishment of the attributes of a text work (determination of authorship, time of creation, place of recording,

etc.). The article presents a generalized context-dependent graph-theoretic model designed for the analysis of folklore and literary texts. The minimal structural unit of the model (primitive) is a word. Sets of words are combined into vertices, and the same word can be related to different vertices. Edges and graph substructures reflect the lexical, syntactic and semantic links of the text. The characteristics of the model are its fuzziness, hierarchy and temporality. As examples, a hierarchical graph-theoretical model of components (on the example of literary works by A. S. Pushkin), a temporal graph-theoretic model of a fairy tale plot (on the example of Russian fairy tales by A. M. Afanasyev) and a fuzzy graph-theoretic model of «strong» connections of grammatical classes (on the example of anonymous articles from the pre-revolutionary magazines «Time», «Epoch» and the weekly «Citizen», edited by F. M. Dostoevsky). The model is built in such a way that it can be further explored using artificial intelligence methods (for example, decision trees or neural networks). For this purpose, a format for storing such data was implemented in the information system «Folklore», as well as procedures for entering, editing and analyzing texts and their graph-theoretic models.

Keywords: graph-theoretic model; text attribution; lexis; syntax; semantics; fuzzy graph; hierarchical graph; temporal graph; information system «Folklore»

For citation: Moskin N.D., Rogov A.A., Voronov R.V. Generalized context-dependent graph-theoretic model of folklore and literary texts. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 1, 2022, pp. 73-86 (in Russian). DOI: 10.15514/ISPRAS–2022–34(1)–6

1. Введение

Задачи анализа текстов, как одно из направлений искусственного интеллекта [12], все чаще решаются с помощью современных математических методов и компьютерных технологий. В научной литературе обозначены различные подходы и методы решения задач классификации и поиска текстов, атрибуции текстов, машинного перевода, автоматического реферирования, выявления плагиата, анализа тональности текстов, генерации текстов, реконструкции текстов и др. [7] Эти задачи объединяет необходимость поиска нестандартных, скрытых закономерностей, присущих текстам, которые можно обнаружить, например, с помощью методов искусственного интеллекта и машинного обучения. Также отметим, что результаты, полученные при решении одного класса задач обработки текстов можно применить и для другого класса.

В настоящее время активно развивается направление обработки естественного языка (машинный перевод, создание чат-ботов и т.д.), связанное с использованием нейросетевых технологий (Transformer, RNN, CNN) [11, 17]. Эти технологии позволяют выявлять скрытые закономерности языка, однако для их настройки и адекватной работы требуется большой объем данных. Другим недостатком этих технологий является закрытость получаемых языковых закономерностей (моделей) и отсутствие обоснования принимаемых решений. Это допустимо при их технологическом использовании, но часто не подходит для научных исследований по атрибуции текстов.

Основная идея математических подходов к решению задачи атрибуции текстов заключается в подсчете статистических параметров, которые, с одной стороны, идентифицируют стиль автора, а, с другой стороны, им слабо контролируются [15, 16]. Это направление известно под названием стилеметрия (stylometry). Ученые изучают такие характеристики как длина предложения, длина слова, богатство словарного запаса и т.п. Иногда исследования базируются на подсчете *n*-грамм - последовательностей текстовых элементов (букв, слов, идентификаторов частей речи и т. д.), взятых в порядке их появления в тексте [18]. Однако часто подобные эксперименты не давали убедительных результатов и являлись труднообъяснимыми для филологов.

Отметим, что по причине своей многоплановой и многоуровневой структуры текст является сложным объектом для изучения. С одной стороны, в нем можно выделить разные структурные единицы (например, на лексическом, синтаксическом и семантическом уровнях), а с другой стороны, установить разные виды связей, т.е. в результате одному и тому же тексту могут соответствовать несколько различных моделей. Такие модели можно

представить в виде графов, которые состоят из множества объектов (вершин) и связей между этими объектами (ребер). На наш взгляд, теоретико-графовые модели являются перспективным направлением в области атрибуции текстов. Например, использование технологии GNN (Graph Neural Network или в переводе с английского *графовые нейронные сети*) позволяет не преобразовывать структуру в числовой вектор, теряя при этом часть важной информации, а сохранить топологические отношения для последующего анализа [19].

Разработка обобщенной модели обусловлена несколькими факторами. Разные теоретико-графовые модели позволяют получить новую информацию об исследуемых текстах. Особенно это важно в случае фольклорных и литературных текстов, когда их число в коллекциях в силу исторических причин невозможно увеличить. При этом получаемые выборки могут быть несбалансированными (например, произведения одного автора сильно преобладают над произведениями другого автора). Интерес представляют и новые характеристики, получаемые в результате синтеза разных моделей, и возможности для дальнейшей разработки более совершенных гибридных структур. Реализованные в программе математические методы и алгоритмы без труда можно перенести с одного класса моделей на другой. В случае же, когда они имеют разное описание, подобный подход выглядит труднореализуемым и сложно интерпретируемым для филологов. Отметим также, что становится проще на одних и тех же данных провести сравнение результатов классификации, полученных с помощью разных методик, и выявить наиболее эффективные. Поэтому значимым являются не только единообразное структурное описание теоретико-графовых моделей для атрибуции текстов, но и разработка общего формата для их хранения и дальнейшего анализа. Данное исследование в целом представляется полезным для систематизации теоретико-графовых моделей текстов и методологии их построения.

В данной работе предлагается обобщенная контекстно-зависимая теоретико-графовая модель. Она была апробирована на различных коллекциях фольклорных и литературных текстов [10, 11, 14]:

- теоретико-графовая модель синтаксической структуры, рассмотренная на материале фольклорных песен (Лужские песни, бесёдные песни) и стилизованных под фольклор текстов (Н. А. Клюев, А. К. Толстой, С. А. Есенин и др.);
- иерархическая модель синтаксической структуры предложения (на материале текстов П. А. Вяземского, Э. По, И. А. Бродского, а также переводов С. Андреевского, Д. Мережковского, К. Брюсова, Г. Голохвастова, Н. Голя, В. Топорова и др.);
- нечеткая теоретико-графовая модель на основе деревьев зависимостей (на материале духовных стихов о Голубиной книге из сборника Кирши Данилова и «Собрания народных песен П. В. Киреевского», былин в записи П. Н. Рыбникова);
- теоретико-графовая модель семантической структуры фольклорных песен (на материале песен Заонежья XIX – начала XX века в записи Ф. Студитского, В. Дашкова, В. Лысанова и пр.);
- деревья решений, полученные на основе анализа анонимных статей из дореволюционных журналов «Время» (1861-1863), «Эпоха» (1864-1865) и еженедельника «Гражданин» (1873-1874), которые редактировал Ф. М. Достоевский.

2. *Обобщенная контекстно-зависимая теоретико-графовая модель фольклорных и литературных текстов*

2.1 Нечеткие, темпоральные и иерархические графы

Как отмечается в [10], важными характеристиками теоретико-графовых моделей текстов при решении задачи атрибуции являются нечеткость, иерархичность и темпоральность.

Одним из проявлений *нечеткости* на разных уровнях языковой структуры текста являются случаи омонимии. Омонимами (греч. *homos* – одинаковый, *onyma* – имя) называются слова, разные по значению, но одинаковые по звучанию и написанию. Различают лексическую омонимию, морфологическую омонимию, лексико-морфологическую омонимию (наиболее частый вид) и синтаксическую омонимию [5].

Понятие нечеткого графа (fuzzy graph) основано на определении функции принадлежности, которая ставит в соответствие вершине или ребру графа значение от 0 до 1. Более строго нечеткий граф второго рода $\tilde{G} = (\tilde{V}, \tilde{E})$ определяется следующим образом [3]. Пусть имеется некоторое универсальное множество X и задано нечеткое множество $\tilde{V}$ в X имеющее вид

$$\tilde{V} = \{\langle \mu_V(v) \backslash v \rangle\}, v \in X,$$

где $0 \le \mu_V(v) \le 1$ – значение функции принадлежности для вершины v (здесь V – носитель множества $\tilde{V}$ Зададим также нечеткое множество ребер

$$\tilde{E} = \{\langle \mu_E(v_i, v_j) \rangle\}, v_i, v_j \in V,$$

где $0 \le \mu_E(v_i, v_j) \le 1$ – значение функции принадлежности для ребра (v_i, v_j). Если множество вершин является четким в отличие от множества ребер, то такой граф называется графом первого рода.

Также в ряде работ Л.С. Берштейна (например, в [2]) вводится понятие *темпорального графа*, т.е. модели, где связи между элементами (вершинами графа) изменяются во времени (в случае моделирования текста под этим термином будем понимать упорядоченность слов в тексте и соответствующих им вершин). Автор отмечает, что понятие темпорального графа (temporal graph) в литературе трактуется в достаточно широком диапазоне – от временных графиков до ориентированных ациклических графов и сетей Петри.

В математических терминах назовем темпоральным графом тройку $G = (X, \{\Gamma_t\}, T)$, где X – множество вершин графа с числом вершин $n = |X|, T = \{1,2, \dots, N\}$ – множество натуральных чисел, определяющих (дискретное) время; $\{\Gamma_t\}$ – семейство соответствий, или отображений множества вершин X в себя в момент времени $t \in T$, т.е. $(\forall t \in T)\Gamma_t: X \to X$. Причем, для различных моментов времени эти отображения, в общем случае, различные:

$$(\forall x \in X)(\forall t_1, t_2 \in T | t_1 \ne t_2)\left[\Gamma_{t_1}(x) \ne \Gamma_{t_2}(x)\right].$$

Во многих случаях требуются более сложные графовые формализмы, обладающие иерархической структурой (hierarchical graph). Известны *иерархические графовые модели*, описание которых приводится, например, в [8]. Граф C называется фрагментом графа G (обозначим $C \subseteq G$), если C – это подмножество элементов графа G. Обозначим F – иерархию фрагментов G, если $G \in F$ и для любых двух фрагментов C_1 и C_2 из F либо фрагменты C_1 и C_2 не пересекаются, либо один из них является частью (подфрагментом) другого. Фрагмент G – основной (главный) фрагмент иерархии F. Фрагмент $C \in F$ – элементарный, если в F нет фрагментов G, являющихся подфрагментами фрагмента C.

Пусть задана некоторая иерархия фрагментов F графа G. Для любых $C_1, C_2 \in F$ фрагмент C_1 – прямой подфрагмент C_2, если C_1 – подфрагмент C_2 и не существует такого $C_3 \in F$ отличного от C_1 и C_2, что $C_1 \subseteq C_3 \subseteq C_2$. Иерархический граф $H = (G, T)$ состоит из графа G и корневого дерева T, вершины которого соответствуют элементам некоторой иерархии в G, а дуги отражают отношение их непосредственной вложенности. T называется деревом вложенности, а G – основным графом иерархического графа H.

Однако отметим, что подобные иерархические графы не подходят для описания структуры текста. Как показано в п. 2.3, связи могут существовать не только между вершинами, но и между какой-либо вершиной и фрагментом графа.

2.2 Рекурсивное определение обобщенной контекстно-зависимой теоретико-графовой модели

Дадим рекурсивное определение обобщенной контекстно-зависимой теоретико-графовой модели для решения задачи анализа текстов. Это набор $G = (V, H, E, \alpha, \beta, \mu, \gamma)$ для текста T, который определим в три этапа:

1) Сегментация текста T:

- Пусть текст T состоит из упорядоченной последовательности слов $W = \{w_k\}_{k=1}^{K}$, где $K > 0$ – общее количество слов (индекс k соответствует порядку появления слова w_k в тексте);
- $W_l \subset W$ – упорядоченные подмножества слов в тексте ($l = 1, 2, \dots, L$). Подмножество W_l может состоять как из одного слова, так и из совокупности слов (необязательно следующих подряд). Допускается, что подмножества могут пересекаться;

2) Определение элементов теоретико-графовой модели G:

- $V = \{v_i\}_{i=1}^{n}$ – непустое конечное множество вершин;
- $V_j \subset V, j = 0,1, \dots, m$ – подмножества вершин теоретико-графовой модели, таких что их объединение совпадает с $V = \bigcup_{j=1}^{m} V_j$. Допускается, что подмножества могут пересекаться;
- $H = \left\{\{v_i\}_{i=1}^{n} \cup \{G_j\}_{j=1}^{m}\right\}$ – множество, объединяющее вершины из V и совокупность вложенных теоретико-графовых структур $G_j = (V_j, H_j, E_j)$ уровня j, где множество H_j уровня $j = 2,3, \dots, m$ определяется либо как пустое множество, либо как подмножество вложенных теоретико-графовых структур уровней меньших j, т.е. $H_j \subset \{G_l\}_{l=1}^{j-1}$, а E_j представляет собой подмножество упорядоченных пар из $V_j \cup H_j$, т.е. ребер вложенной модели (иерархичность);
- $E \subset H \times H$ (подмножество упорядоченных пар элементов из H) – множество ребер G, которое состоит из s элементов. При этом подмножества ребер $E_j \subseteq E, j = 1,2, \dots, m$ и попарно не пересекаются;

3) Определение атрибутов элементов теоретико-графовой модели G:

- γ – отображение, задающее соответствие между объектами теоретико-графовой модели $x_i \in H \cup E$ и подмножествами слов в тексте $W_l \subset W$. Допускается, что некоторые вершины или подструктуры могут быть «фиктивные» (не связанные со словами в тексте), т. е. $\gamma(x_i) = \emptyset$. Таким образом, γ определяет упорядоченность (темпоральность) объектов теоретико-графовой модели.
- A – множество атрибутов (меток) вершин, которые определяются характеристиками текста. Элемент множества A может быть вектором, который определяет несколько атрибутов;
- $\alpha: V \to A$ – функция, задающая атрибуты (метки) вершинам;
- B – множество атрибутов (меток) ребер, которые определяются характеристиками текста. Элемент множества B может быть вектором, который определяет несколько атрибутов (в том числе, например, отсутствие направленности у ребра);
- $\beta: E \to B$ – функция, задающая атрибуты (метки) ребрам;
- $\mu: H \cup E \to [0,1]$ – функция, задающая нечеткость объектов теоретико-графовой модели.

Рассмотрим, как можно представить в терминах обобщенной модели три теоретико-графовые структуры, обладающие соответственно свойствами иерархичности (п. 2.3), темпоральности (п. 2.4) и нечеткости (п. 2.5).

2.3 Иерархическая теоретико-графовая модель составляющих

В литературе известны два вида деревьев, которые описывают синтаксическую структуру текста. *Деревья зависимостей* обычно используются в описаниях языков со свободным порядком слов (например, русского). Для описания языков с фиксированным порядком слов преимущественно используется второй тип графов – *деревья составляющих* [5]. При этом в предложении выделяются группы слов, функционирующие как отдельные синтаксические единицы – составляющие. Система составляющих – это множество отрезков предложения, которое обладает тем свойством, что каждые два входящих в него отрезка либо не пересекаются, либо один из них содержится в другом. Речь идет о так называемых *синтагмах*. Это совокупность нескольких слов, объединённых по принципу семантико-грамматической сочетаемости, единица синтагматики.

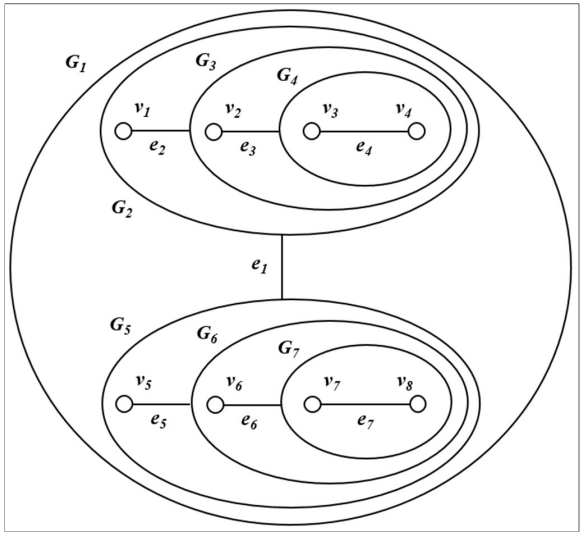

Рис. 1. Модель составляющих фрагмента «Онегин, добрый мой приятель, родился на брегах Невы»
Fig. 1. Model of the components of the fragment «Onegin, my good friend, was born on the banks of the Neva»

Табл. 1. Соответствие вершин и слов текста
Table 1. Correspondence of vertices and words of the text

№	Подмножества слов	Вершина или ребро	Группы	Функция принадлежности
1	$W_1 = \{w_1\}$ = {“Онегин”}	v_1	$\alpha(v_1) = N$	$\mu(v_1) = 1$
2	$W_2 = \{w_2\}$ = {“добрый”}	v_2	$\alpha(v_2) = A$	$\mu(v_2) = 1$
3	$W_3 = \{w_3\}$ = {“мой”}	v_3	$\alpha(v_3) = C$	$\mu(v_3) = 1$
4	$W_4 = \{w_4\}$ = {“приятель”}	v_4	$\alpha(v_4) = N$	$\mu(v_4) = 1$
5	$W_5 = \{w_5\}$ = {“родился”}	v_5	$\alpha(v_5) = V$	$\mu(v_5) = 1$
6	$W_6 = \{w_6\}$ = {“на”}	v_6	$\alpha(v_6) = P$	$\mu(v_6) = 1$
7	$W_7 = \{w_7\}$ = {“брегах”}	v_7	$\alpha(v_7) = N$	$\mu(v_7) = 1$
8	$W_8 = \{w_8\}$ = {“Невы”}	v_8	$\alpha(v_8) = N$	$\mu(v_8) = 1$

Рассмотрим модель составляющих (рис. 1) на примере фрагмента из романа в стихах А. С. Пушкина: «Онегин, добрый мой приятель, родился на брегах Невы» [5]. Отметим, что

существует множество текстов, которые приписываются Александру Сергеевичу и до сих пор в этом вопросе нет окончательного ответа [13]. Общее количество слов фрагмента $K = 8$. В данном случае подмножества W_l $(l = 1, \dots, L = 8)$ будут соответствовать не только словам w_i, но и вершинам графа v_i (табл. 1). Функция α может задавать, например, часть речи слова (т.е. быть атрибутом вершины). Описание множества A представлено в п. 3.2. Сразу отметим, функция μ для всех вершин и ребер принимает значение 1 (нечеткие связи отсутствуют), т.е. $\mu(v_i) = \mu(e_j) = 1$.

Данная теоретико-графовая модель содержит семь подструктур $G_j = (V_j, H_j, E_j), j = 1, 2, \dots, m = 7$. Первая подструктура G_1 содержит G_2 и G_5, которые соединяются ребром e_1, т.е. $V_1 = \emptyset$, $H_1 = \{G_2, G_5\}$, $E_1 = \{e_1 = (G_2, G_5)\}$. Аналогично опишем другие подструктуры (ребра из множества $E = \{e_t\}_{t=1}^{s=7}$ не являются ориентированными, атрибуты им не заданы, т.е. $B = \emptyset$, они могут соединять не только вершины, но и подструктуры):

- $V_2 = \{v_1\}, H_2 = \{G_3\}, E_2 = \{e_2 = (v_1, G_3)\}$;
- $V_3 = \{v_2\}, H_3 = \{G_4\}, E_3 = \{e_3 = (v_2, G_4)\}$;
- $V_4 = \{v_3, v_4\}, H_4 = \emptyset, E_4 = \{e_4 = (v_3, v_4)\}$;
- $V_5 = \{v_5\}, H_5 = \{G_6\}, E_5 = \{e_5 = (v_5, G_6)\}$;
- $V_6 = \{v_6\}, H_6 = \{G_7\}, E_6 = \{e_6 = (v_6, G_7)\}$;
- $V_7 = \{v_7, v_8\}, H_7 = \emptyset, E_7 = \{e_7 = (v_7, v_8)\}$.

Поскольку все вершины и ребра находятся «внутри» той или иной подструктуры, множества $V_0 = E_0 = \emptyset$.

2.4 Темпоральная теоретико-графовая модель сказочного сюжета

Вторая теоретико-графовая модель возникает при исследовании сказочных сюжетов (например, из волшебных сказок А. М. Афанасьева [1]). Основоположником подобного структурного направления является В. Я. Пропп и его последователи. В текстах выделяются инварианты – действующие лица сказки [4], которых можно объединить в десять групп:

- герой (*H*);
- антигерой (антагонист, вредитель) (*A*);
- прорицатель (*P*);
- даритель (снабдитель) (*Д*);
- помощник (*П*);
- антипомощник (*V*);
- глупец (*G*);
- антидаритель (*W*);
- награда (*N*);
- препятствие (*R*).

Тело сказки в самом общем виде есть конечная последовательность встреч действующих лиц, связанных соединительными фразами (например, «долго-ли, коротко-ли шел он и наконец увидел…»). Встречи непосредственно связаны с их поступками: например, «Даритель даст Герою совет о том, как действовать дальше». Возможные встречи действующих лиц сказки представлены в таблице 2 [4].

Табл. 2. Возможные встречи действующих лиц сказки
Table 2. Possible meetings of characters in the tale

Действующее лицо	**С кем может встретиться**							
H	*A*	*P*	*Д*	*П*	*V*	*N*	*G*	*W*
A		*P*		*П*	*V*	*N*	*G*	*W*
P				*П*		*N*	*G*	
Д				*П*		*N*		
П					*V*	*N*		*W*
V						*N*	*G*	
N							*G*	*W*
G								*W*

Построим теоретико-графовую модель, где вершинами являются действующие лица сказки, а ребра будут отражать их встречи, пронумерованные в соответствии с их появлением в теле сказки. Если встреч было несколько, то ребра будут кратными.

Табл. 3. Соответствие вершин/ребер и их групп
Table 3. Correspondence of vertices and their groups

№	**Подмно-жества слов**	**Вершина или ребро**	**Группы**	**Функция принадлежности**
1	W_1	v_1	$\alpha(v_1)$ = “Герой (*H*)”	$\mu(v_1) = 1$
2	W_2	v_2	$\alpha(v_2)$ = “Награда (*N*)”	$\mu(v_2) = 1$
3	W_3	v_3	$\alpha(v_3)$ = “Даритель (*Д*)”	$\mu(v_3) = 1$
4	W_4	v_4	$\alpha(v_4)$ = “Антигерой (*A*)”	$\mu(v_4) = 1$
5	W_5	v_5	$\alpha(v_5)$ = “Препятствие (R)”	$\mu(v_5) = 1$
6	W_6	e_1	$\beta(e_1)$ = “*H*-*Д*”	$\mu(e_1) = 1$
7	W_7	e_2	$\beta(e_2)$ = “*H*-*Д*”	$\mu(e_2) = 1$
8	W_8	e_3	$\beta(e_3)$ = “*H*-*Д*”	$\mu(e_3) = 1$
9	W_9	e_4	$\beta(e_4)$ = “*H*-*R*”	$\mu(e_4) = 1$
10	W_{10}	e_5	$\beta(e_5)$ = “*H*-*R*”	$\mu(e_5) = 1$
11	W_{11}	e_6	$\beta(e_6)$ = “*H*-*R*”	$\mu(e_6) = 1$
12	W_{12}	e_7	$\beta(e_7)$ = “*H*-*A*”	$\mu(e_7) = 1$
13	W_{13}	e_8	$\beta(e_8)$ = “*H*-*N*”	$\mu(e_8) = 1$

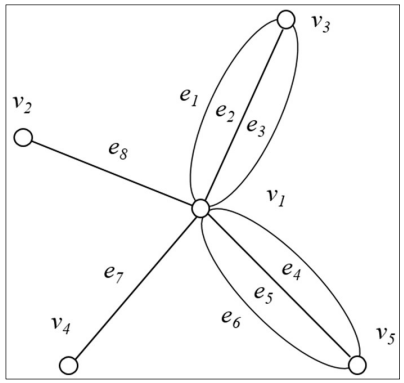

Рис. 2. Теоретико-графовая модель сказочного сюжета
Fig. 2. Graph-theoretic model of a fairy tale plot

Рассмотрим пример сказочного сюжета, в котором выделяются пять действующих лиц и восемь встреч [6]. Здесь не только каждому i–ому действующему лицу (вершине), но и j–й встрече (ребрам) будут соответствовать в тексте набор слов (словосочетаний) W_l, где $l = 1,2,\dots,L = n + s = 13$ (табл. 3). При этом подмножества W_l включают слова, которые могут повторяться и необязательно следовать друг за другом. Например, главный герой v_1 связан с набором слов, которые находятся в разных частях сказки (включая местоимения и синонимы). Ребра e_i, как правило, имеют отношение к цельным фрагментам, описывающим встречу действующих лиц. В данной модели отсутствует иерархия, поэтому $m = 0$ (рис. 2). Сравнивая между собой подобные графы можно посмотреть, как менялся сказочный сюжет с течением времени, какие были особенности у разных регионов и пр.

2.5 Нечеткая теоретико-графовая модель «сильных связей» грамматических связей

Третий вид теоретико-графовой модели основан на матрице частот парной встречаемости грамматических классов слов текста (биграмм). Подобные модели использовались, например, для атрибуции исторических текстов [9] и анонимных текстов из дореволюционных журналов «Время» (1861-1863), «Эпоха» (1864-1865) и еженедельника «Гражданин» (1873-1874), которые редактировал Ф. М. Достоевский [11]. Для получения такой матрицы необходимо:

- выбрать систему грамматических классов;
- перекодировать последовательность слов анализируемого текста в последовательность соответствующих обозначений грамматических классов;
- вычислить частоты парной встречаемости a_{ij} для каждой пары классов с учетом направления развертывания текста (слева направо).

Рассмотрим построение модели на примере текста Ф. М. Достоевского «Молодое перо. По поводу литературной подписи "Современник" № 1 и 2», опубликованной в журнале «Время» (1863 год. Разд. Современное обозрение. № 2. С. 221-226). Каждому i-му синтаксическому классу ($n = 8$ в примере на рис. 3) определяется вершина v_i (соответствующее подмножество слов W_i выбирается по принципу их принадлежности к этому классу, табл. 4).

Табл. 4. Соответствие вершин/ребер и их групп
Table 4. Correspondence of vertices and their groups

№	Подмно-жества слов	Вершина или ребро	Группы	Функция принадлежности
1	W_1	v_1	$\alpha(v_1)$ = “Существительное”	$\mu(v_1) = 1$
2	W_2	v_2	$\alpha(v_2)$ = “Прилагательное”	$\mu(v_2) = 1$
3	W_3	v_3	$\alpha(v_3)$ = “Местоимение”	$\mu(v_3) = 1$
4	W_4	v_4	$\alpha(v_4)$ = “Глагол”	$\mu(v_4) = 1$
5	W_5	v_5	$\alpha(v_5)$ = “Частица”	$\mu(v_5) = 1$
6	W_6	v_6	$\alpha(v_6)$ = “Предлог”	$\mu(v_6) = 1$
7	W_7	v_7	$\alpha(v_7)$ = “Союз”	$\mu(v_7) = 1$
8	W_8	v_8	$\alpha(v_8)$ = “Цитата”	$\mu(v_8) = 1$
9	-	e_1	$\beta(e_1)$ = “Существительное-Существительное”	$\mu(e_1) = 0{,}02041$
10	-	e_2	$\beta(e_2)$ = “Существительное-Местоимение”	$\mu(e_2) = 0{,}02099$

11	-	e_3	$\beta(e_3)$ = “Существительное-Глагол”	$\mu(e_3) = 0{,}02741$
12	-	e_4	$\beta(e_4)$ = “Существительное-Предлог”	$\mu(e_4) = 0{,}02041$
13	-	e_5	$\beta(e_5)$ = “Существительное-Союз”	$\mu(e_5) = 0{,}03032$
14	-	e_6	$\beta(e_6)$ = “Прилагательное-Существительное”	$\mu(e_6) = 0{,}05831$
15	-	e_7	$\beta(e_7)$ = “Местоимение-Существительное”	$\mu(e_7) = 0{,}02974$
16	-	e_8	$\beta(e_8)$ = “Местоимение-Глагол”	$\mu(e_8) = 0{,}02624$
17	-	e_9	$\beta(e_9)$ = “Глагол-Местоимение”	$\mu(e_9) = 0{,}02041$
18	-	e_{10}	$\beta(e_{10})$ = “Глагол-Союз”	$\mu(e_{10}) = 0{,}02332$
19	-	e_{11}	$\beta(e_{11})$ = “Частица-Глагол”	$\mu(e_{11}) = 0{,}02157$
20	-	e_{12}	$\beta(e_{12})$ = “Предлог-Существительное”	$\mu(e_{12}) = 0{,}03324$
21	-	e_{13}	$\beta(e_{13})$ = “Предлог-Местоимение”	$\mu(e_{13}) = 0{,}02507$
22	-	e_{14}	$\beta(e_{14})$ = “Союз-Местоимение”	$\mu(e_{14}) = 0{,}02741$
23	-	e_{15}	$\beta(e_{15})$ = “Цитата-Цитата”	$\mu(e_{15}) = 0{,}02857$

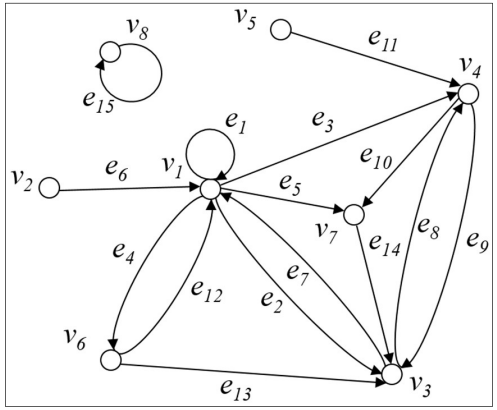

Рис. 3. Нечеткая модель «сильных» связей грамматических классов (текст Ф. М. Достоевского «Молодое перо», порог 0,02)
Fig. 3. Fuzzy model of «strong» connections of grammatical classes (text by F. M. Dostoevsky "Young Pen", threshold 0.02)

3. Формат представления теоретико-графовых моделей в информационной системе «Фольклор»

Приведенные примеры показывают, что в терминах обобщенной модели можно задавать структуры, обладающие разными свойствами. В [10] приводятся также другие виды

теоретико-графовых моделей, которые могут, например, обладать всеми тремя свойствами: иерархичность, нечеткость и темпоральность. Однако подобные формализации несомненно более сложны для описания и визуализации.

Возникает потребность в автоматической обработке текстов и их теоретико-графовых моделей, построенных на основе разных принципов. С этой целью в Петрозаводском государственном университете была разработана информационная система «Фольклор» [10]. Изначально она создавалась как проблемно-ориентированная система, предназначенная для сравнительного анализа одной коллекции бесёдных песен Заонежья конца XIX – начала XX века. Однако впоследствии программа была модифицирована таким образом, что позволила проводить исследование других коллекций на основе различных моделей.

Для хранения и последующего анализа текстов в информационной системе «Фольклор» был реализован формат SNG. Этот формат представляет собой текстовый файл, который можно легко редактировать. Рассмотрим на примере иерархической модели составляющих, как структурируются данные (табл. 5). Файл делится на пять частей: общие характеристики текста, слова, объекты, связи и матрица инцидентности (технически части разделены между собой одиночной строкой с комментариями, начинающиеся с символов //).

- Общие характеристики текста (1-11 строки). В первой строке указывается название текста, во второй строке – название группы, связанных между собой текстов (например, название теоретико-графовой модели), количество строк в тексте (если значение равно нулю, то граф строится без привязки к тексту), строки текста, количество характеристик текста, затем пары: название характеристики и ее значение.
- Слова текста (13-29 строки). В первой строке указывается количество слов, далее для каждого слова – номер строки, начало выделения, длина (параметр не является избыточным, т. к. в некоторых текстах возможно слияние слов), после дефиса – часть речи (*N* – существительное, *A* – прилагательное, *C* – местоимение, *O* – числительное, *V* – глагол, *E* – причастие, *G* – деепричастие, *D* – наречие, *S* - категория состояния, *P* – предлог, *L* – союз, *U* – частица, *I* – междометие). Теоретически в случае омонимии (например, слово «мой» может быть как местоимением, так и глаголом, но в данном случае из контекста понятно, что это местоимение) можно указать двойную часть речи (например, -*CV*). Если часть речи неизвестна или еще не определена, то ставится знак ‘?’.
- Объекты текста (31-62 строки). В первой строке указывается количество объектов, во второй строке – способ отображения вершин при визуализации, затем для каждого из них – название, уровень вложенности (если это обычная вершина, то указывается 0, если первый уровень, то – 1, если второй, то – 2, и т. д.), значение функции принадлежности, группа, количество слов объекта, номера слов, относящихся к объекту (если вершина фиктивная, то к ней не привязываются слова текста).
- Связи в тексте (64-75 строки). В первой строке указывается количество связей, затем - способ отображения ребер при визуализации, название связи, значение функции принадлежности, группа, количество слов связи, номера слов, относящихся к связи.
- Матрица связей (77-91 строки). Представляет собой матрицу инцидентности, где строки соответствуют объектам текста, а столбцы – связям. Если в столбце напротив вершины *i* указано -1, а напротив вершины *j* – 1, то это значит, что дуга идет из *i* в *j*.

Табл. 5. Представление иерархической теоретико-графовой модели составляющих фрагмента «Онегин, добрый мой приятель, родился на брегах Невы»
Table 5. Representation of a hierarchical graph-theoretic model of fragment components «Onegin, my good friend, was born on the banks of the Neva»

№	Строка	№	Строка
1	Онегин, добрый мой приятель, родился на брегах Невы	45	брегах
		46	0 1 1 1 6
2	иерархическая модель составляющих	47	Невы

	////////////////	48	0 1 1 1 7
3	1	49	мой приятель
4	Онегин, добрый мой приятель,	50	1 1 1 2 2 3
5	родился на брегах Невы	51	добрый мой приятель
	////////////////	52	2 1 1 3 1 2 3
6	2	53	Онегин, добрый мой приятель
7	Автор	54	3 1 1 4 0 1 2 3
8	А.С. Пушкин	55	брегах Невы
9	Источник	56	1 1 1 2 6 7
10	Роман в стихах "Евгений Онегин"	57	на брегах Невы
11	////////////////	58	2 1 1 3 5 6 7
	8	59	родился на брегах Невы
12	0 0 6	60	3 1 1 4 4 5 6 7
13	-N	61	Онегин, добрый мой приятель, родился на брегах Невы
14	0 8 6		
15	-A	62	4 1 1 8 0 1 2 3 4 5 6 7
16	0 15 3	63	////////////////
17	-C	64	7
18	0 19 8	65	group4
19	-N	66	ребро 1
20	0 29 7	67	1 0 0 0
21	-V	68	ребро 2
22	0 37 2	69	1 0 0 0
23	-P	70	ребро 3
24	0 40 6	71	1 0 0 0
25	-N	72	ребро 4
26	0 47 4	73	1 0 0 0
27	-N	74	ребро 5
28	////////////////	75	1 0 0 0
29	15	76	////////////////
30	group2	77	0 -1 0 0 0 0 0
31	Онегин	78	0 0 -1 0 0 0 0
32	0 1 1 1 0	79	0 0 0 -1 0 0 0
33	добрый	80	0 0 0 1 0 0 0
34	0 1 1 1 1	81	0 0 0 0 -1 0 0
35	мой	82	0 0 0 0 0 -1 0
36	0 1 1 1 2	83	0 0 0 0 0 0 -1
37	приятель	84	0 0 0 0 0 0 1
38	0 1 1 1 3	85	0 0 1 0 0 0 0
39	родился	86	0 1 0 0 0 0 0
40	0 1 1 1 4	87	-1 0 0 0 0 0 0
41	на	88	0 0 0 0 0 1 0
42	0 1 1 1 5	89	0 0 0 0 1 0 0
43		90	1 0 0 0 0 0 0
44		91	0 0 0 0 0 0 0

4. Заключение

В данной статье предложена обобщенная контекстно-зависимая теоретико-графовая модель для атрибуции текстов. Модель обладает свойствами иерархичности, нечеткости и темпоральности. Она была апробирована на материале фольклорных (народные песни, сказки, духовные стихи, былины) и литературных текстов (за авторством Н. А. Клюева, А. К. Толстого, С. А. Есенина, А. С. Пушкина, П. А. Вяземского, Э. По, И. А. Бродского и др.). Примеры моделей представлены в монографиях [10, 11] и на электронном ресурсе СМАЛТ (“Статистические методы анализа литературных текстов”), который расположен по адресу

http://smalt.karelia.ru/. В работе рассмотрен формат хранения теоретико-графовой модели в информационной системе «Фольклор» и системе СМАЛТ.

Описание единого подхода к построению различных теоретико-графовых моделей текстов позволяет находить расстояния между ними, т.е. решать задачи классификации и кластеризации текстов. Это можно использовать, например, при решении задачи атрибуции. В качестве такого расстояния можно использовать меры на основе операций редактирования, на основе максимального общего подграфа, минимального общего надграфа, структурных спектров и т. п. [10].

Список литературы / References

[1] Афанасьев А.М. Народные русские сказки А. Н. Афанасьева: в 3 т. М., Государственное Издательство Художественной литературы (Гослитиздат), 1957 г. / Afanasyev A.M. Folk Russian fairy tales by A. N. Afanasyev: in 3 volumes. Moscow, State Publishing House of Fiction (Goslitizdat), 1957 (in Russian).

[2] Берштейн Л.С., Боженюк А.В. Использование темпоральных графов как моделей сложных систем. Известия ЮФУ. Технические науки, № 4 (105), 2010 г., стр. 198-203 / Bershtein L.S., Bozhenyuk A.V. The use of temporal graphs as models of complex systems. Izvestiya SFedU. Engineering Sciences, vol. 4 (105), 2010, pp. 198-203 (in Russian).

[3] Берштейн Л.С., Боженюк А.В. Нечеткие графы и гиперграфы. М., Научный мир, 2005 г., 256 стр. / Bershtein L.S., Bozhenyuk A.V. Fuzzy graphs and hypergraphs. Moscow, Scientific world, 2005, 256 p. (in Russian).

[4] Гаазе-Рапопорт М.Г. Поиск вариантов в сочинении сказок. Дополнение в книге Зарипов Р.Х. Машинный поиск вариантов при моделировании творческого процесса. М.: Наука, 1983 г., стр. 213-223. / Gaaze-Rapoport M.G. Search for variants in the composition of fairy tales. Supplement in Zaripov R.H. Machine search for variants in modeling the creative process. Moscow, Nauka, 1983, pp. 213-223 (in Russian).

[5] Гладкий А.В. Синтаксические структуры естественного языка. М., ЛКИ, 2007 г., 152 с. / Gladky A.V. Syntactic structures of natural language. Moscow, LKI, 2007, 152 p. (in Russian).

[6] Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. М., Университетская книга, Логос, 2007 г., 320 стр. / Zubov A.V., Zubova I.I. Fundamentals of artificial intelligence for linguists. Moscow, University book, Logos, 2007, 320 p. (in Russian).

[7] Ильвовский Д.А., Черняк Е.Л. Системы автоматической обработки текстов. Открытые системы. СУБД, no. 1, 2014 г., стр. 51-53 / Ilvovsky D.A., Chernyak E.L. Systems of automatic processing of texts. Open systems. DBMS, no. 1, 2014, pp. 51–53 (in Russian).

[8] Касьянов В.Н., Евстигнеев В.А. Графы в программировании: обработка, визуализация и применение. СПб., БХВ-Петербург, 2003 г., 1104 стр. / Kasyanov V.N., Evstigneev V.A. Graphs in programming: processing, visualization and application. St. Petersburg, BHV-Petersburg, 2003, 1104 p. (in Russian).

[9] Милов Л.В., Бородкин Л.И. и др. От Нестора до Фонвизина: Новые методы определения авторства. М., Прогресс, 1994 г., 445 стр. / Milov L.V., Borodkin L.I. et al. From Nestor to Fonvizin: New methods for determining authorship. Moscow, Progress, 1994, 445 p. (in Russian)

[10] Москин Н.Д. Теоретико-графовые модели фольклорных текстов и методы их анализа. Петрозаводск, Изд-во ПетрГУ, 2013 г., 148 стр. / Moskin N.D. Graph-theoretic models of folklore texts and methods of their analysis. Petrozavodsk, PetrGU Publishing House, 2013, 148 p. (in Russian)

[11] Рогов А.А., Абрамов Р.В. и др. Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин». Петрозаводск, Изд-во «Острова», 2021 г., 391 с. / Rogov A.A., Abramov R.V. et al. The problem of attribution in the magazines «Time», «Epoch» and the weekly «Citizen». Petrozavodsk: Publishing house «Islands», 2021, 391 p. (in Russian)

[12] Соколов И.А. Теория и практика применения методов искусственного интеллекта. Вестник Российской академии наук, том 89, вып. 4, 2019, стр. 365-370. / Sokolov I.A. Theory and practice of application of artificial intelligence methods. Bulletin of the Russian Academy of Sciences, vol. 89, issue 4, 2019, pp. 365-370. (in Russian)

[13] Хозяинов С.А. Атрибуция публицистических произведений, приписываемых А. С. Пушкину: тексты 1830-1836 гг. Санкт-Петербург, 2008 г., 24 с. / Hozyainov S.A. Attribution of publicistic works attributed to A. S. Pushkin: texts of 1830-1836. St. Petersburg, 2008, 24 p. (in Russian)

[14] Щеголева Л.В., Лебедев А.А., Москин Н.Д. Методы анализа данных в задаче разграничения фольклорных и авторских текстов. Вопросы языкознания, 2020 г., no. 2, стр. 61-74. / Shchegoleva L.V., Lebedev A.A., Moskin N.D. Methods of data analysis in the problem of distinguishing between folklore and author's texts. Questions of linguistics, 2020, no. 2, pp. 61-74. (in Russian)

[15] Calle-Martin J., Miranda-Garcia A. Stylometry and Authorship Attribution: Introduction to the Special Issue. English Studies, vol. 93, no. 3, 2012, pp. 251-258.

[16] Stamatatos E. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, vol. 60, no. 3, 2009, pp. 538-556.

[17] Vaswani A., Shazeer N. et al. Attention is all you need. In Proc. of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000-6010.

[18] Zečević A. N-gram based text classification according to authorship. In Proc. of the Second Student Research Workshop associated with RANLP 2011, 2011, pp. 145-149.

[19] Zhou J., Cui G. et al. Graph neural networks: A review of methods and applications. AI Open, vol. 1, 2020, pp. 57-81.

Информация об авторах / Information about authors

Николай Дмитриевич МОСКИН – кандидат технических наук, доцент, доцент кафедры теории вероятностей и анализа данных. Сфера научных интересов: цифровые гуманитарные науки, теоретико-графовые модели, интеллектуальный анализ данных, компьютерная лингвистика, мультимедиа-технологии, компьютерная графика.

Nikolai Dmitrievich MOSKIN – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Probability Theory and Data Analysis. Research interests: digital humanities, graph-theoretic models, data mining, computational linguistics, multimedia technologies, computer graphics.

Александр Александрович РОГОВ – доктор технических наук, профессор, заведующий кафедрой теории вероятностей и анализа данных. Сфера научных интересов: математическое моделирование, прикладная статистика, математические методы распознавания образов, математические методы анализа литературных текстов.

Aleksandr Aleksandrovich ROGOV – Doctor of Technical Sciences, Professor, Head of the Department of Probability Theory and Data Analysis. Research interests: mathematical modeling, applied statistics, mathematical methods of pattern recognition, mathematical methods of analysis of literary texts.

Роман Владимирович ВОРОНОВ – доктор технических наук, профессор кафедры прикладной математики и кибернетики. Сфера научных интересов: математическое моделирование, задачи оптимизации, комбинаторные задачи на графах, математические методы и модели систем локального позиционирования мобильных объектов.

Roman Vladimirovich VORONOV – Doctor of Technical Sciences, Professor of the Department of Applied Mathematics and Cybernetics. Research interests: mathematical modeling, optimization problems, combinatorial problems on graphs, mathematical methods and models of mobile object local positioning systems.