Методы и алгоритмы автоматического извлечения информации из научных текстов для создания тезауруса научной терминологии

Е. В. Вопилова[1] Е. Н. Крючкова[2]

В статье предлагается метод автоматического построения тезауруса научной терминологии, основанный на алгоритмах извлечения многословных терминов из специальных энциклопедий и научных публикаций.

Представлены результаты работы алгоритмов создания и пополнения тезауруса на примере обработки математических текстов.

Предложен алгоритм сравнительного семантического анализа научных публикаций, а также способы количественной оценки их семантического сходства.

Ключевые слова: аспектно-ориентированный анализ, научный лексикон, семантический граф, классификация научных текстов, автоматическая обработка неструктурированных текстов.

1. Введение

Разработка инструментов для автоматической обработки текстов научной тематики является одним из актуальных направлений исследований в области NLP и включает задачи реферирования и аннотирования [1], классификации и аспектного анализа [2], информационного поиска [3]. При проектировании алгоритмов обработки текстов в компьютерной лингвистике используются вероятностно-статистические методы [4], нейросетевые модели [5], в последнее время широкое распространение получили большие языковые модели (LLM) [6].

Наличие адекватной семантической модели предметной области является необходимым условием для эффективного решения задачи тематического анализа. В работе [7] предложена формальная модель лингвистической онтологии, содержащая универсальные для различных предметных

[1] *Вопилова Елена Владимировна* — аспирант каф. прикладной математики ф-та инф. технологий АлтГТУ, e-mail: vopilova.elena@gmail.com.

Vopilova Elena Vladimirovna — graduate student, Polzunov Altai State Technical University, Faculty of Information Technology, Chair of Applied Mathematics.

[2] *Крючкова Елена Николаевна* — к.ф.м.н., профессор каф. прикладной математики ф-та инф. технологий АлтГТУ, e-mail: kruchkova_elena@mail.ru.

Kryuchkova Elena Nikolaevna — PhD in Physics and Mathematics, Professor, Polzunov Altai State Technical University, Faculty of Information Technology, Chair of Applied Mathematics.

областей типы семантических отношений. Широко распространены графовые модели, аккумулирующие семантические связи между словами и статистику частотности слов [10].

В данной работе рассматриваются методы и алгоритмы автоматического построения модели тезауруса предметной терминологии, позволяющей выполнять семантический анализ научного текста на основе объединения семантики отношений между терминами со статистикой использования терминов в научной публикации. Предложен алгоритм пополнения предлагаемой модели представления знаний информацией из научных публикаций. Построенная модель может быть использована для аспектного анализа узкоспециализированных текстов, в том числе количественной оценки сходства тематики публикаций в текстовой коллекции. Адекватность предлагаемой модели продемонстрирована на примере сравнения математических публикаций. Автоматически созданный тезаурус математических терминов, источником данных которого является математическая энциклопедия [11], размещен в открытом доступе по адресу evvopilova.pythonanywhere.com. Сервисные функции просмотра терминологии позволяют переходить по семантическим связям построенного тезауруса.

2. Многословные термины как основа научного лексикона

Тексты научных публикаций отличаются от остальных особой морфологией и лексикой, а также определёнными синтаксическими и семантическими структурами. Научная терминология обладает рядом особенностей, из которых наибольшее влияние на построение доменной семантической модели оказывает многословность научных терминов. В данной работе используется модель научного лексикона, построенная в результате автоматической обработки математической энциклопедии и представляющая собой ориентированный семантический граф G_{domain}, связывающий специфические доменные термины семантическими отношениями [12]. Анализ алгоритма построения этого графа выходит за рамки данной статьи, в данной работе мы будем использовать уже построенный семантический граф G_{domain}, вершины которого соответствуют терминам домена, а взвешенные дуги представляют семантические отношения различного типа между терминами. Вес дуги соответствует значимости семантического отношения между терминами. В процессе разработки модели тезауруса основное внимание будет уделено ассоциативным связям между терминами, которые отражают наиболее существенные взаимосвязи между сущностями [7].

3. Сравнительный семантический анализ публикаций на основе научного лексикона

Из активно развивающихся ветвей автоматического контент-анализа в современной математической лингвистике можно выделить два направления: категоризация текстов или отдельных фрагментов – определение принадлежности текста к некоторому классу внутри заданной предметной области [8] и аспектно-ориентированный анализ [9], предполагающий работу с текстом на уровне отдельных аспектов/функций целевого объекта. Основная цель аспектного анализа состоит в извлечении аспектов – сущностей, позволяющих объединить однотипные по функциональности элементы. В научной терминологии такими сущностями являются термины, объединяющие в единую группу связанные с выделенным аспектным термином другие термины. В данной работе эта задача решается методом кластеризации терминологии, связанной с текстом научной публикации, в результате задача извлечения аспектных терминов публикации рассматривается как задача выделения центров построенных кластеров.

Структурно научный текст может быть представлен как последовательность общеупотребительных слов и специальных терминов, причем семантическую нагрузку несут только научные термины. Сравнение тематики двух публикаций в простейшем варианте можно провести на уровне сравнения частотностей используемых в текстах терминов, но такой подход не учитывает контекстное окружение использованной терминологии и, следовательно, непригоден для оценки семантического сходства текстов. Построим семантический граф $G_{text}(T)$ текста T, выделяя из G_{domain} не только вершины, соответствующие используемым в T терминам, но и связанные с ними вершины из некоторой окрестности, сохраняя при этом для каждой связывающей дуги не только ее вес, но и тип связи. В процессе построения по взвешенным дугам графа на термины из окрестности распространим частотность явно используемого термина с учетом пространственного затухания.

Пусть Y – выделенный из текста анализируемой публикации термин, для которого в базовом графе G_{domain} имеется одноименная вершина, Y_i – вершина G_{domain} из некоторой окрестности $\omega(Y)$ термина Y, $f_{init}(Y_i)$ – частотность термина Y_i в анализируемом тексте, γ – коэффициент пространственного затухания. Тогда в результате работы алгоритма распространения весов вершина графа Y_i, связанная с вершиной Y с частотностью $f(Y)$, будет иметь итоговую характеристику частотности

$$f(Y_i) = f_{init}(Y_i) + f(Y) \cdot m(Y, Y_i) \cdot \gamma^{l(Y,Y_i)} \tag{1}$$

где $l(Y, Y_i)$ – число дуг на этом пути,

$m(Y, Y_i)$ – произведение весов дуг на ориентированном пути от Y к Y_i с наименьшим количеством дуг.

При каждом последующем обновлении значения $f(Y_i)$ по формуле (1) в качестве $f_{init}(Y_i)$ будет использоваться текущее значение $f(Y_i)$. В результате контекст домена, представленный графом G_{domain}, будет объединен со статистикой использования терминологии в статье. Кластеризация построенного графа $G_{text}(T)$ позволит выделить тематические аспекты научной публикации, а центры кластеров можно рассматривать как главные аспектные термины статьи. Для кластеризации был использован алгоритм *k-medoids*, который пригоден для кластеризации вершин графа так как предназначен для решения тех задач, в которых центром кластера может быть только точка из набора обозначенных. Если для каждого кластера C_i вычислить его вес $p(C_i)$ как сумму весов принадлежащих кластеру вершин, то относительный вес кластера можно рассматривать как относительную значимость соответствующего кластеру аспектного термина. Здесь и далее в качестве относительного веса кластера будем использовать отношение веса кластера к сумме весов всех кластеров.

Рассмотрим сравнительный семантический анализ публикаций. Пусть T_1 и T_2 – сравниваемые тексты, T_{12} – конкатенация T_1 и T_2. Построим графы $G_{text}(T_1)$, $G_{text}(T_2)$ и $G_{text}(T_{12})$, представляющие статистику контекстного окружения в соответствующих текстах. Построим множество кластеров $C = \{C_1, C_2, \ldots, C_k\}$ для $G_{text}(T_{12})$, в каждом кластере C_i выделим медоид – центральный термин M_i. Граф семантического сходства $G_{cmpr}(T_1, T_2)$ определяет поток из T_1 в T_2 и содержит вершины источника и стока T_1 и T_2, а также промежуточные вершины C_1, C_2, $\ldots$, C_n. Вершины C_1, C_2, $\ldots$, C_n имеют пропускную способность, равную относительному весу кластера в $_{Gtext}(_{T12})$, ребра между $_1$ (или $_2$) и C_i имеют пропускную способность, равную относительному суммарному весу вершин $G_{text}(T_1)$ (соответственно $G_{text}(T_2)$), присутствующих в кластере C_i графа $G_{text}(T_{12})$. Коэффициент семантического сходства текстов T_1 и T_2 равен потоку из истока T_1 в сток T_2 в транспортной сети $G_{cmpr}(T_1, T_2)$.

При проведении экспериментов текст публикации T_1 сравнивался с текстом T_1': изначально $T_1 = T_1'$, затем в качестве «белого шума» в T_1' постепенно добавлялся текст из математической научной публикации T_2 другой тематики. На рисунке 1 представлены результаты сравнения текстов T_1 и T_1'. При непосредственном сравнении текстов T_1 и T_2 их семантическое сходство текстов равно 0.26.

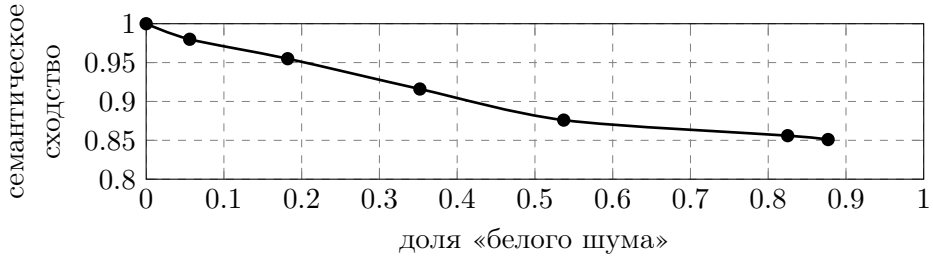

Рис. 1. Результаты сравнения текстов T_1 и T_1'

4. Алгоритмы обновления терминологии

В процессе эволюционного развития науки появляются новые разделы и научные направления, что приводит к появлению не только новой терминологии, но и к частичному изменению семантики существующей устоявшейся терминологии: возникают новые семантические связи между терминами, трансформируются некоторые существующие связи. Это приводит к необходимости пополнения базы знаний научной области, а, следовательно, и к выбору надежных источников обучающих данных. Источником надежной информации для систем автоматического обновления баз знаний могут служить публикации из рецензируемых изданий в ранжируемых научных издательствах.

Рассмотрим сначала задачу выявления новых предметных терминов. Несмотря на то, что с точки зрения семантической разметки текст научной статьи является неструктурированным, наличие в тексте статьи раздела «Ключевые слова» можно рассматривать как наличие некоторой слабой структурированности этого текста в целом. Определим специфику пополнения графа научной области G_{domain} при семантической обработке достаточно обширного текста научной публикации, некоторая часть которого связана с описанием нового понятия, в качестве которых мы будем рассматривать перечисленные в статье ключевые слова.

Пусть Y – абсолютно новый термин, тогда в семантический граф G_{domain} добавляется термин Y, а также термины, составленные из подмножества слов термина Y, связанные с Y семантическими связями типа «часть – целое» при условии, что новый термин имеет корректную синтаксическую структуру. Формирование связи типа «часть - целое» соответствует принципу формирования многословных терминов, составленных из уточняющих слов. Как правило, любое удаление хотя бы одного слова из термина приводит к определению более общего понятия. Например, «Маркова цепь сложная» является частью всех «цепей Маркова» или всех «сложных цепей». Для создания ассоциативных связей термина Y, во-первых, надо выделить в тексте статьи фрагменты, име-

ющих непосредственное отношение к новому термину, во-вторых, надо вычислить значимость связей нового термина Y с другими терминами, которые определяют главную тематику статьи. В силу того, что ключевое слово достаточно редко целиком используется в тексте статьи, попытка выделить физический фрагмент текста, имеющего непосредственное отношение к Y, в большинстве случаев заранее обречена на неудачу. Но остается возможность связать Y с тематикой статьи в целом. Семантика статьи T формируется в процессе построения графа $G_{text}(T)$. Таким образом, возникает задача выделения в построенных кластерах для графа $G_{text}(T)$ информации, контекстно связанной с новым термином. Вес ассоциативных связей между ключевым словом Y и термином M_i, который является центром кластера C_i, определяется как

$$Z(Y, M_i) = \frac{p(C_i)}{\sum_{i=1}^{n} p(C_i)} \tag{2}$$

где $p(C_i)$ – вес кластера C_i.

Исходный граф G_{domain} построен по исключительно авторитетному источнику информации – научной энциклопедии. Энциклопедия – совместный труд заслуженных авторов, аккумулирует устоявшиеся данные по большинству разделов науки. Поэтому вычисленный по формуле (2) на основании данных, извлеченных из единственного источника, следует умножить на некоторый коэффициент уровня доверия к источнику ξ ($\xi \leq 1$):

$$\tilde{Z}(Y, M_i) = \xi \cdot Z(Y, M_i) \tag{3}$$

В таблице 1 представлены результаты выделения новых терминов при обработке коллекции публикаций научных журналов [13, 14].

Таблица 1. Результаты пополнения базового семантического графа информацией из текстового корпуса

Число существующих вершин-кандидатов, переведенных в статус базового термина	36
Число новых вершин	107
Число новых связей – из них связей «часть-целое» – из них связей-ассоциаций	4345 2818 1527

5. Повышение весовых коэффициентов ассоциативных связей

Каждая новая статья может содержать информацию из новых разделов науки или новых научных направлений, данные о которых отсутствуют или плохо представлены в графеG_{domain}. Следовательно, обработка новой статьи может дополнить наши знания как новыми терминами, так и новыми семантическими связями между терминами – как новыми, так и уже имеющимися в тезаурусе. Семантика статьи может указывать на более сильные зависимости между терминами. Фактически это означает, что обработка статьи из доверенных источников может производиться в двух режимах:

1) в режиме пополнения новыми терминами и новыми связями,

2) в режиме модификации весов существующих семантических связей с поддержкой режима (1).

Выделенные из текста публикации ключевые слова могут быть как новыми терминами, так и уже существующими в графе G_{domain}. Новые термины добавляются вместе с новыми связями, веса которых рассчитаны по формуле (3). Если же семантическая связь уже существует, то возникает вопрос об изменении ее веса. Уже существующие связи построены на основе доверенных источников информации, и одна статья не может претендовать на кардинальное изменение весов, но может привести к некоторой их модификации.

Пусть $\tilde{Z}(Y_1, Y_2)$ – вес дуги в графе G_{domain}, $\tilde{Z}_1(Y_1, Y_2)$ – вычисленный по формуле (3) вес этой дуги. Если $\tilde{Z}(Y_1, Y_2) \geq \tilde{Z}_1(Y_1, Y_2)$, то новый текст не увеличивает ассоциативную связь между терминами, вес существующей связи не изменяется. В случае $\tilde{Z}(Y_1, Y_2) < \tilde{Z}_1(Y_1, Y_2)$ новый текст должен усилить эту ассоциативную связь и необходим перерасчет веса соответствующей дуги.

Вычислим относительное среднеквадратичное отклонение

$$d = \frac{|\tilde{Z}(Y_1, Y_2) - \tilde{Z}_1(Y_1, Y_2)|}{\tilde{Z}(Y_1, Y_2) + \tilde{Z}_1(Y_1, Y_2)}$$

Произведем корректировку веса дуги:

$$\tilde{Z}_{new}(Y_1, Y_2) = g(\tilde{Z}(Y_1, Y_2), \tilde{Z}_1(Y_1, Y_2))$$

где

$$g(\tilde{Z}(Y_1, Y_2), \tilde{Z}_1(Y_1, Y_2)) = \tilde{Z}(Y_1, Y_2) + \tilde{Z}_1(Y_1, Y_2) \cdot f(d) \qquad (4)$$

Выбор функции $f(d)$ и, следовательно, $g(\tilde{Z}(Y_1, Y_2), \tilde{Z}_1(Y_1, Y_2))$ определяется следующими критериями:

- при однократном появлении $\tilde{Z}_1(Y_1, Y_2)$, значительно превышающем $\tilde{Z}(Y_1, Y_2)$, не должен наблюдаться резкий рост $\tilde{Z}_{new}(Y_1, Y_2)$;
- при постоянном появлении большого значения $\tilde{Z}_1(Y_1, Y_2)$ вычисленное по формуле (4) значение $\tilde{Z}_{new}(Y_1, Y_2)$ должно за некоторое число итераций сходиться к $\tilde{Z}_1(Y_1, Y_2)$.

В данной работе используется

$$f(d) = \frac{d}{\sqrt{2}} = \frac{|\tilde{Z}(Y_1, Y_2) - \tilde{Z}_1(Y_1, Y_2)|}{\sqrt{2} \cdot (\tilde{Z}(Y_1, Y_2) + \tilde{Z}_1(Y_1, Y_2))}$$

Рассмотрим сходимость $\tilde{Z}_{new}(Y_1, Y_2)$ $\tilde{Z}_1(Y_1, Y_2)$. Вычислим количество итераций n, при котором $|\tilde{Z}(Y_1, Y_2) - \tilde{Z}_1(Y_1, Y_2)| < \varepsilon$. При $\varepsilon = 0.001$ потребуется от 10 до 14 шагов, а при $\varepsilon = 0.0001$ количество итераций лежит в диапазоне от 14 до 19. Фактически это означает, что даже при небольшом начальном весе дуги наличие 14-19 текстов с высоким уровнем ассоциативных отношений приводит к вычислению веса дуги, соответствующему данным из новых источников. С другой стороны, исходный вес дуги постепенно увеличивается только при наличии достаточного количества доверенных источников, что говорит о сбалансированном требовании к набору статей для устойчивого формирования нового веса $\tilde{Z}_{new}(Y_1, Y_2)$.

На рисунке 2 представлены результаты перерасчета веса существующей дуги при многократной обработке одной и той же научной публикации.

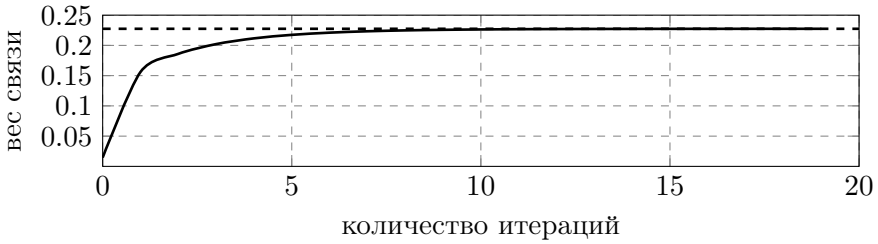

Рис. 2. Изменение веса дуги «дифференциальная игра» → «стратегия» графа G_{domain} при обработке текста научной публикации

6. Заключение

В статье рассмотрены проблемы автоматического формирования тезауруса научной области. Автоматически сформирован тезаурус математической терминологии, предложены алгоритмы пополнения построенного тезауруса. Рассматриваются вопросы выделения тематических аспектов

научной публикации, предложенные алгоритмы базируются на алгоритмах кластеризации графов.

Список литературы

[1] Altmami N., Menai M., "Automatic Summarization of Scientific Articles: A Survey", *Journal of King Saud University - Computer and Information Sciences*, **34** (2020), 1011–1026.

[2] Berna A, B., Ganiz M., "Semantic text classification: A survey of past and recent advances", *Information Processing & Management*, **54**:6 (2018), 1129–1163.

[3] Benites F., "Information Retrieval and Knowledge Extraction for Academic Writing", *Digital Writing Technologies in Higher Education*, 2023, 303–315.

[4] Hossari M., Dev S., Kelleher J. D., "TEST: A Terminology Extraction System for Technology Related Terms", *Proc. The 2019 11th International Conference on Computer and Automation Engineering*, 2019, 78-81.

[5] Danilov G., Ishankulov T., Kotik K., Orlov Yu. Shifrin M., Potapov A., "The Classification of Short Scientific Texts Using Pretrained BERT Model", *Public Health and Informatics*, **281** (2021), 83–87.

[6] Dunn A., Dagdelen J., Walker N., Lee S., Rosen A., Ceder G., Persson K., Jain A., "Structured information extraction from complex scientific text with fine-tuned large language models", 2022, 83–87.

[7] Лукашевич Н.В., Добров Б.В., "Проектирование лингвистических онтологий для информационных систем в широких предметных областях", *Онтология проектирования*, **5**:1(15) (2015), 47–69.

[8] Costa L.S., Oliveira I.A., Fileto R., "Text classification using embeddings: a survey", *Knowledge and Information Systems*, **65** (2023), 2761-2803.

[9] Marshalova A.,Bruches E., Batura T., "Automatic Aspect Extraction from Scientific Texts", *Proc. Recent Trends in Analysis of Images, Social Networks and Texts (AIST 2023), Communications in Computer and Information Science*, **1905** (2023), 67-80.

[10] Belwal R., Rai S., Gupta A., "A new graph-based extractive text summarization using keywords or topic modeling", *Journal of Ambient Intelligence and Humanized Computing*, **12** (2021), 8975–8990.

[11] *Математическая энциклопедия в 5 томах*, ред. Виноградов И.М., Советская энциклопедия, Москва, 1977.

[12] Вопилова Е.В., Крючкова Е.Н., "Методы автоматического анализа динамики изложения информации в текстах на основе адаптируемых словарей научных терминов", *Программная инженерия*, **15**:4 (2024), 206–215.

[13] *Вестник Южно-Уральского университета, серия «Математика. Механика. Физика»*, **14**:2–4 (2022).

[14] *Математический сборник*, **213**:9–12 (2022).

Methods and algorithms for automatic extraction of information from scientific texts for creating a scientific terminology thesaurus

Vopilova E.V., Kryuchkova E.N.

The paper proposes a method for automatic construction of a scientific terminology thesaurus based on algorithms for extraction of multi-word terms from special encyclopedias and scientific publications.

The results of the algorithms for thesaurus creation and replenishment are presented on the example of mathematical text processing.

We propose the algorithm for comparative semantic analysis of scientific publications and the ways of quantitative estimation of their semantic similarity.

Keywords: aspect-oriented analysis, scientific vocabulary, semantic graph, classification of scientific text, automatic processing of unstructured texts.

References

[1] Altmami N., Menai M., "Automatic Summarization of Scientific Articles: A Survey", *Journal of King Saud University - Computer and Information Sciences*, **34** (2020), 1011–1026.

[2] Berna A, B., Ganiz M., "Semantic text classification: A survey of past and recent advances", *Information Processing & Management*, **54**:6 (2018), 1129–1163.

[3] Benites F., "Information Retrieval and Knowledge Extraction for Academic Writing", *Digital Writing Technologies in Higher Education*, 2023, 303–315.

[4] Hossari M., Dev S., Kelleher J. D., "TEST: A Terminology Extraction System for Technology Related Terms", *Proc. The 2019 11th International Conference on Computer and Automation Engineering*, 2019, 78-81.

[5] Danilov G., Ishankulov T., Kotik K., Orlov Yu. Shifrin M., Potapov A., "The Classification of Short Scientific Texts Using Pretrained BERT Model", *Public Health and Informatics*, **281** (2021), 83–87.

[6] Dunn A., Dagdelen J., Walker N., Lee S., Rosen A., Ceder G., Persson K., Jain A., "Structured information extraction from complex scientific text with fine-tuned large language models", 2022, 83–87.

[7] Lukashevich N.V., Dobrov B.V., "Designing linguistic ontologies for information systems in broad subject areas", *Ontology of Designing*, **5**:1(15) (2015), 47–69 (In Russian).

[8] Costa L.S., Oliveira I.A., Fileto R., "Text classification using embeddings: a survey", *Knowledge and Information Systems*, **65** (2023), 2761-2803.

[9] Marshalova A.,Bruches E., Batura T., "Automatic Aspect Extraction from Scientific Texts", *Proc. Recent Trends in Analysis of Images, Social Networks and Texts (AIST 2023), Communications in Computer and Information Science*, **1905** (2023), 67-80.

[10] Belwal R., Rai S., Gupta A., "A new graph-based extractive text summarization using keywords or topic modeling", *Journal of Ambient Intelligence and Humanized Computing*, **12** (2021), 8975–8990.

[11] *Mathematical encyclopedia in 5 volumes*, ред. Vinogradov I.M., Soviet Encyclopedia, Moscow, 1977 (In Russian).

[12] Vopilova E. V., Kryuchkova E. N., "Automatic analysis methods of dynamics of information presentation in texts based on adaptable dictionaries of scientific terms", *Software Engineering*, **15**:4 (2024), 206–215 (In Russian).

[13] *Bulletin of the South Ural State University, Ser. Mathematics. Mechanics. Physics*, **14**:2–4 (2022) (In Russian).

[14] *Sbornik: Mathematics*, **213**:9–12 (2022) (In Russian).