УДК 004.912

МЕТОД АВТОМАТИЗИРОВАННОЙ КЛАССИФИКАЦИИ НАУЧНЫХ СТАТЕЙ ПО ТИПУ РЕЗУЛЬТАТА В НАУЧНЫХ АНАЛИТИЧЕСКИХ СИСТЕМАХ

Кузнецов И.А.

ФГАОУ ВО «Национальный исследовательский ядерный университет «МИФИ», Москва, e-mail: iakuznetsov@mephi.ru

В данной работе представлен метод, который позволяет на основе аннотации научной статьи определить тип полученного в ней результата. Представленный метод основывается на выявлении связей между схожими статьями посредством тематического моделирования с применением латентного размещения Дирихле. В качестве основных признаков для формирования семантического ядра каждого типа научного результата используются словосочетания, состоящие из двух или трех слов. Полученный тип научного результата дополнит перечень характеристик, которыми уже обладает научная статья в явном виде. Каждая область знаний может быть описана своим специфическим набором словосочетаний. Подготовить в ручном режиме такой набор словосочетаний по каждому направлению без знания специфики предметной области практически невозможно, да и весьма трудозатратно. В данной работе был проведен эксперимент с применением представленного метода на научных статьях, относящихся к рубрике «Информатика». Дополненный перечень явных и неявных характеристик научных статей может быть использован в качестве входных параметров для рекомендательной системы, что позволит сформировать наиболее релевантный и пертинентный набор рекомендаций пользователю.

Ключевые слова: научный результат, текстовый анализ данных, нечеткая кластеризация данных, классификация данных, рекомендательные системы, латентное размещение Дирихле

METHOD OF AUTOMATED CLASSIFICATION OF SCIENTIFIC ARTICLES BY THE TYPE OF RESULT IN SCIENTIFIC ANALYTICAL SYSTEMS

Kuznetsov I.A.

National Research Nuclear University MEPhI, Moscow, e-mail: iakuznetsov@mephi.ru

In this paper we present a method that allows us to determine the type of the result obtained in it on the basis of the annotation of a scientific article. The method presented based on the identification of links between similar articles through thematic modeling using latent Dirichlet allocation. As the main signs for the formation of the semantic kernel of each type of scientific result, phrases consisting of two or three words are used. The obtained type of scientific result will complement the list of characteristics that the scientific article already has in its explicit form. Each area of knowledge can be described by its specific set of word combinations. Prepare in a manual mode such a set of phrases for each direction without knowledge of the specifics of the subject area is almost impossible, and in addition it is very labor-intensive. In this paper, an experiment was carried out using the presented method on scientific articles related to the «Informatics» rubric. The supplemented list of explicit and implicit characteristics of scientific articles can be used as input parameters for the recommender systems, which will allow to form the most relevant and pertinent set of recommendations to the user.

Keywords: scientific result, text data analysis, fuzzy data clustering, data classification, recommender systems, latent Dirichlet allocation

В современном мире уже давно наблюдается тенденция к распространению и доступности данных. В обиход входят такие концепции, как «open data», «open access» и «open source». Научная сфера деятельности также не стоит в стороне, а принимает самое активное участие в распространении и популяризации знаний для всех желающих. На текущий момент благодаря концепции «open access» в сети Интернет доступно огромное количество полнотекстовых научных работ для их последующего изучения без какого-либо ограничения для конечного пользователя. Как правило, доступ к таким данным обеспечивается через веб-сайт с удобным интерфейсом и наличием поисковой строки. Но несмотря на всё это, процесс поиска нужной информации среди огромного количества научных статей и публикаций становится весьма трудоемким, а классический поиск не позволяет решить эту проблему в полной мере.

Для решения таких проблем были разработаны специальные рекомендательные системы, которые с помощью определённых алгоритмов позволяют предоставить не только релевантный результат для поискового запроса, но и сделать результат выдачи пертинентным, то есть спрогнозировать потребность пользователя в определенной информации [1]. Реализация рекомендательных систем базируется на алгоритмах машинного обучения, где задачу по выдаче похожей информации можно свести к решению задачи классификации. В качестве входных данных для алгоритма классифи-

кации выступает определённый набор признаков, который описывает сам объект для классификации. Этап подготовки такого набора признаков является одним из самых важных. Когда речь заходит о работе с текстовыми данными, то возникают вопросы о подготовке и выделении таких признаков. Как правило, очень многое зависит от конкретной предметной области и самих объектов, для которых решается задача классификации [2].

В рамках научно-образовательного процесса выделяют два основных объекта: автор и научная статья. Для каждого объекта формируется свой профиль, в рамках которого выполняется дальнейшая аналитическая работа, связанная с определением приоритетных научных направлений, выявлением научных коллективов, проведением различных наукометрических исследований и прочее. Как правило, для построения научного профиля автора или профиля научной статьи используется явный набор признаков, указанный либо самим автором о себе, либо указанный в параметрах научной статьи [3]:

- для автора – указанные им личные данные, научные интересы, перечень его публикаций, научные достижения и производные значения.
- для научной статьи – ключевые слова, наукометрические показатели, язык, тематика, периодическое издание и производные значения.

Целью данной работы является дополнение и расширение профиля научной статьи путем извлечения из аннотации признака, отвечающего за тип научного результата. Данный признак будет использоваться наряду с другими признаками и будет подаваться на вход как перечень входных параметров для алгоритмов классификации. Дополненный набор признаков должен привести к повышению точности работы рекомендательной системы и сокращению заведомо нерелевантных данных в рекомендательном наборе.

Гипотеза о типе научного результата

Под типом научного результата понимаются представленные в научной статье выводы и заключения. Описываемый подход будет базироваться на принципах потребности пользователя в определенном типе научного результата, который ему необходимо получить исходя из его конкретных нужд и задач.

Основным предположением будет то, что каждый пользователь может нуждаться либо в теоретической информации, либо в практической. Два данных класса могут служить верхнеуровневым представлением типа научного результата, который был получен в рамках научной статьи. Но потребность пользователя может не ограничиться только разделением на теорию и практику. Возможно, ему потребуется некоторая конкретика результата в зависимости от того, какую цель он преследует. Можно сформулировать гипотезу о потребностях пользователя следующим образом:

- Пользователь хочет ознакомиться с предметной областью и ее основами.
- Пользователь хочет понять, в каком направлении развивается интересующая его область и какие есть проблемы.
- Пользователь хочет понять, каких результатов уже удалось добиться на текущий момент времени.
- Пользователь хочет понять, что уже опробовано и готово к использованию.

Данную гипотезу можно представить в виде следующей схемы по типам научного результата исходя из потребности пользователя (см. рис. 1).

Предлагаемый метод

В данной работе задача будет состоять в том, чтобы разделить научные статьи по двум типам: теоретический и практический. Решение этой задачи состоит из нескольких этапов.

1. Подготовка текстовых данных. Одним из главных шагов в работе с текстовыми данными является подготовка и очистка текста для последующего анализа, а также некоторые подходы по преобразованию текстов в иные структуры [4]. Обработка текста включает в себя удаление стоп-слов, удаление редких слов, лемматизацию, удаление специальных символов и ряд других действий. В данной работе применялась лемматизация, удалялись небуквенные символы и отбрасывались слова длиной менее трех символов.

2. Определение значимых словосочетаний. Данный шаг включает в себя два дополнительных действия [5]: генерацию словосочетаний и оценку полученных словосочетаний. Для того, чтобы словосочетания отражали научный результат, было введено ограничение, что одно из слов в словосочетании должно быть причастием. Соответственно, слова с такой частью речи не подвергались лемматизации, но приводились к единой форме с помощью расстояния Левенштейна. В работе использовались словосочетания с количеством слов равным два и три. При формировании словосочетаний учитывался тот факт, что два значащих слова могут находиться не только рядом, но и на некотором расстоянии друг от друга. При генерации словосочетаний было использовано «окно» равное трем.

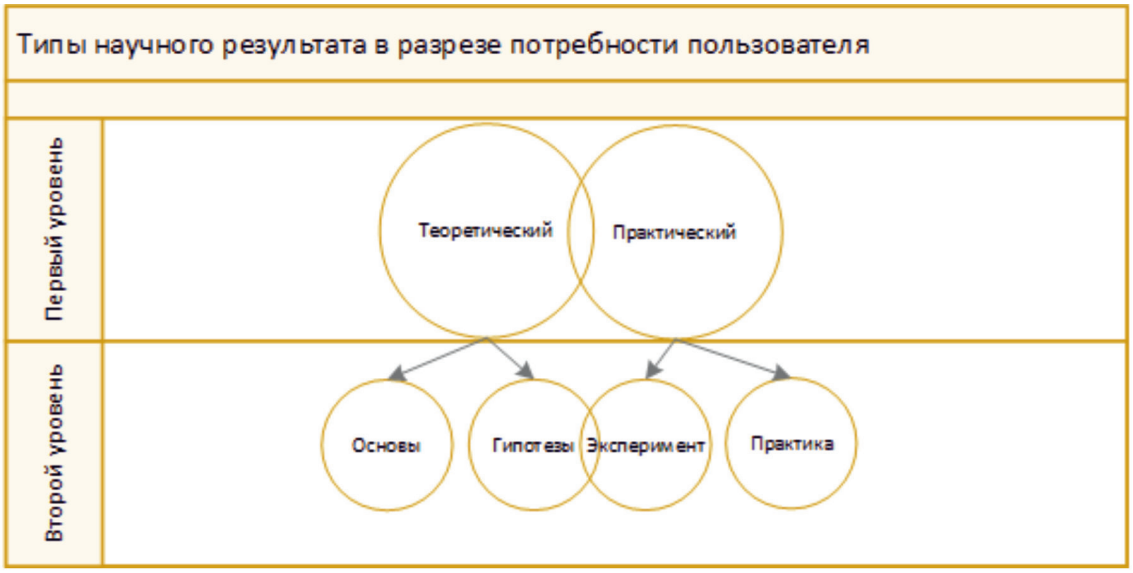

Рис. 1. Гипотеза о типах научного результата

После генерации словосочетаний необходимо определить, является ли данное словосочетание действительно значимым, а не случайным набором слов. Для решения этой задачи были использованы инструменты для проверки статистических гипотез: t-критерий Стьюдента и критерий Хи-квадрат.

t-критерий Стьюдента вычисляется следующим образом:

$$t = \frac{|\overline{x} - m|}{\sqrt{\frac{s^2}{N}}}, \qquad (1)$$

где $\overline{x}$ – выборочное среднее (вероятность встретить словосочетание), m – ожидаемое значение (произведение вероятностей для каждого слова), s^2 – дисперсия (в данном случае примерно соответствует $\overline{x}$), N – размер выборки.

Хи-квадрат вычисляется следующим образом:

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}, \qquad (2)$$

где O_{ij} – фактическое количество словосочетаний, E_{ij} – ожидаемое количество словосочетаний.

Выполнив таким образом расчёт представленных значений для каждого словосочетания в исходном наборе научных статей, можно выполнить их фильтрацию в соответствии с табличными значениями вероятности возникновения ошибки. В данном случае использовалась ошибка равная 0,005.

3. Нечеткое разделение данных на классы. Данный шаг подразумевает проведение кластеризации очищенного набора данных на два класса. В качестве алгоритма нечеткой кластеризации будет использоваться алгоритм латентного размещения Дирихле. С помощью данного алгоритма можно провести кластеризацию по словосочетаниям и построить обучающую модель, которая позволит определить вероятность отнесения научной статьи к каждому из двух типов научного результата.

Метод латентного размещения Дирихле основан на вероятностной модели [6]:

$$p(d, w) = \sum_{t \in T} p(d) * p(w|t) * p(t|d), \qquad (3)$$

где d – документ, w – слово, t – тема, T – множество тем, $p(d)$ – априорное распределение на множестве документов, $p(w|t)$ – условное распределение слова w в теме t, $p(t|d)$ – условное распределение темы t в документе d.

Проведенный эксперимент

Данный эксперимент был выполнен с использованием научных статей из базы данных ВИНИТИ, относящихся к рубрике «Информатика». При подготовке словосочетаний для последующего анализа использовались аннотации научных статей в количестве 5000. Для алгоритма латентного размещения Дирихле количество тем было задано равное двум, что соответствует верхнеуровневому разделению научных статей по типу результата на «Теоретический» и «Практический».

На графиках (см. рис. 2–3) приведены ключевые словосочетания, полученные с использованием представленного метода и отсортированные в порядке убывания по рассчитанной частоте словосочетания по каждому типу результата.

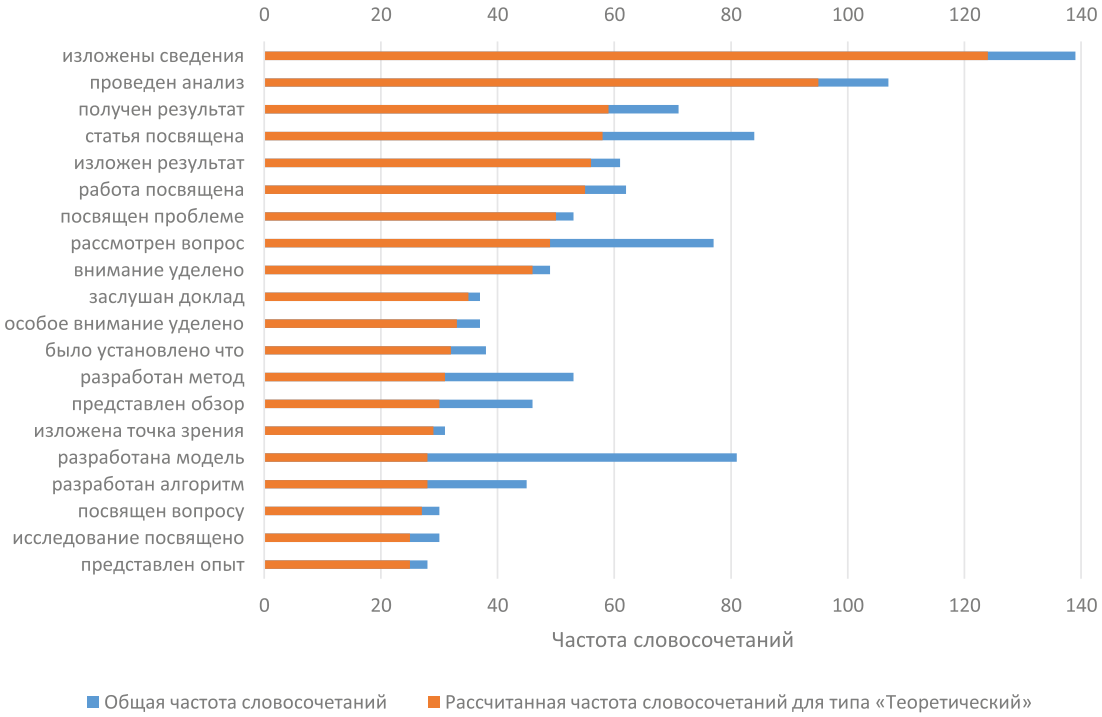

Рис. 2. Топ-20 наиболее релевантных словосочетаний для типа «Теоретический»

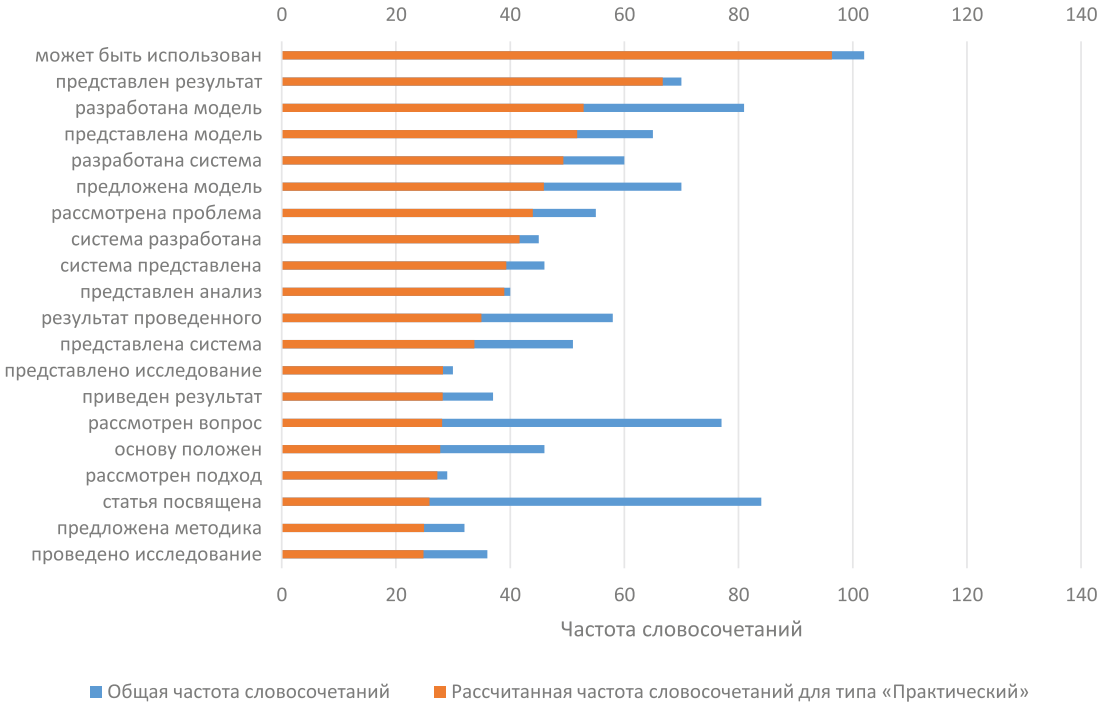

Рис. 3. Топ-20 наиболее релевантных словосочетаний для типа «Практический»

Из представленных графиков видно, что каждый тип результата описан своим набором словосочетаний. Для типа «Практический» отчетливо видны словосочетания, которые отражают именно практический результат, который был достигнут: разработана система, разработана модель, приведен результат и прочее. Для типа «Теоретический» также были выделены словосочетания, которые имеют в большей степени теоретический окрас: посвящено проблеме, рассмотрен вопрос, внимание уделено и прочее.

Однако можно наблюдать и некоторое пересечение похожих словосочетаний. Это связано с тем, что данное словосочетание с разной степенью относится к каждому из типов результата, следовательно, подход на основе нечеткой модели кластеризации в данном случае вполне оправдан. Но для таких словосочетаний необходим отдельный способ обработки и новый тип, который будет включать в себя «шаблонные» фразы и выражения.

Практическое преимущество представленного подхода заключается в том, что для различных направлений научной деятельности используется собственная терминология, индивидуальные обороты и словосочетания. Чтобы охватить каждое научное направление и составить для него свой словарь в ручном режиме – придется прибегнуть к огромному количеству экспертов по разным отраслям науки, что является весьма трудозатратным и малореализуемым.

Заключение

В данной статье был рассмотрен метод, который позволяет выделить новый признак из аннотации научных статей, характеризующий тип результата. На текущий момент данный подход позволяет выделить только два верхнеуровневых типа научного результата. Полученные типы предполагается использовать для уточнения и конкретизации рекомендаций пользователям на основе их поисковых запросов в научных аналитических системах.

При попытке использования данного метода для деления исходных научных статей сразу на вложенные уровни – возникает зашумленность отдельных типов «мусорными» словосочетаниями, и всё это приводит к невозможности однозначной идентификации каждого конкретного типа. В качестве решения представленной проблемы стоит уделить больше внимания специфике фильтрации и обработки текстовых данных, а также использовать другие части речи и учитывать их последовательность в предложении при определении новых типов научного результата [7].

Работа выполнена при поддержке гранта РФФИ № 15-07-08742.

Список литературы

1. Kireev V.S., Kuznetsov I.A. Development of algorithms ensemble in case of the solution of the task of statistical classification in recommender systems // International Journal of Applied Engineering Research. – 2016. – Vol. 11. – № 9. – P. 6613–6618.

2. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011. – 272 с.

3. Варенников Д.А., Шлей М.Д., Муромцев Д.И. Построение научных профилей участников научно-образовательного процесса в информационной системе университета // Современные проблемы науки и образования. – 2015. – № 2–2. URL: http://www.science-education.ru/ru/article/view?id=23109 (дата обращения: 15.02.2018).

4. Артамонова Е.В., Лештаев С.В. Преобразование естественно-языковых текстов в RDF-граф // Современные наукоемкие технологии. – 2016. – № 11–2. – С. 214–218.

5. Кузнецов И.А. Автоматизация процесса формирования онтологии на основе классов пользователей в научных рекомендательных системах // Новые информационные технологии в научных исследованиях: материалы XXII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов (Рязань, РГРТУ, 15–17 ноября). – Рязань, 2017. – С. 34–36.

6. Коляда А.С., Яковенко В.А, Гогунский В.Д. Применение латентного размещения Дирихле для анализа публикаций из наукометрических баз данных // Труды Одесского политехнического университета. – 2014. – № 1. – С. 186–191.

7. Shang J., Liu J., Jiang M., Ren X., Voss C.R., Han J. Automated Phrase Mining from Massive Text Corpora // CoRR. – 2017. – Vol. abs/1702.04457.