УДК 004.021 DOI: 10.18698/2541-8009-2023-01-856

МЕТОД ОПРЕДЕЛЕНИЯ ОБЪЕКТА ИЗ ОГРАНИЧЕННОЙ ВЫБОРКИ ПО НЕЧЕТКОМУ ОПИСАНИЮ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Е.В. Брянская bryanskayakatya@yandex.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Статья посвящена решению проблемы определения объекта из ограниченной выборки по нечеткому описанию на русском языке. Разработанный метод заключается в совмещении двух основных подходов к решению типичных задач в этой области, один из которых базируется на статистическом алгоритме, а второй основан на использовании семантической сети. Для каждого из них требуется своя онтология. Для формирования базы знаний первого этапа используется адаптированный метод TF-IDF, для второго за основу берется совокупность синтаксических графов. Для поиска нечетких дубликатов между запросом пользователя и базой знаний, созданной заранее, применяется косинусное сходство. В работе исследовано влияние размера выборки на меру сходства и точность определения объекта. Проведена оценка доли обращений ко второму шагу предложенного метода, в том числе с целью определить, какая доля этих обращений приходится на неверное предположение, сделанное на первом этапе.

Ключевые слова

Естественный язык, обработка текстов на естественном языке, онтология, «мешок слов», векторизация, TF-IDF, нечеткие дубликаты, косинусное сходство, семантическая сеть, синтаксический граф

Поступила в редакцию 21.12.2022

Введение. Согласно ежегодному прогнозу, который опубликовала Международная корпорация данных (IDC, International Data Corporation), занимающаяся мониторингом количества информации, объем данных в 2020 г. резко вырос. Установлено, что этот показатель достиг 64,2 Збайт (1 Збайт = 10^{21} байт) данных, что примерно в 2 раза больше, чем в 2018 г., когда отметка достигла 33 Збайт. Также предполагается, что в период с 2020 по 2025 г. объем будет продолжать активно расти. В связи с этим проблема поиска необходимой информации в больших массивах данных приобретает все бóльшую актуальность. Задача идентификации объекта по его словесному описанию достаточно сложна, поскольку, с одной стороны, один и тот же объект может быть по-разному описан в различных источниках. С другой стороны, возможна ситуация, когда объект имеет разную смысловую нагрузку в зависимости от контекста использования.

Определение объекта по его описанию может быть полезно в таких сферах, как медицина, например, в системах, ориентированных на определение вида

заболевания с последующим подбором лечения по описанию симптомов, результатов обследований и т. п. Также это может быть применено в системах контроля и оценки знаний учащихся, в частности, в заданиях с развернутым ответом. Определение объекта по его описанию может быть внедрено в системы поиска документов по формальному описанию их содержимого.

Среди распространенных подходов к решению подобных задач можно выделить статистический и синтаксический. В силу особенностей русского языка (многообразие синонимичных выражений, синтаксических особенностей построения предложения и т. п.) каждый из этих методов, применяемый по отдельности, демонстрирует не очень высокую результативность. Поэтому в данном проекте была сделана попытка объединить их в целях повышения качества конечного результата распознавания.

Формальная постановка задачи. Цель разработки метода — определение объекта из ограниченной выборки терминов определенной предметной области по его нечеткому описанию на русском языке.

Во избежание сложностей, возникающих при анализе текста на естественном языке, предлагается скомбинировать существующие решения типичных задач NLP (*Natural Language Processing* — обработка естественного языка), такие как статистический и синтаксический алгоритмы. В последнем используется семантическая сеть, основанная на синтаксических графах.

Для каждого подхода необходима своя онтология. Для формирования базы знаний первого этапа алгоритма требуется наличие качественной выборки. Что касается второго, то для построения синтаксических графов, которые в совокупности составляют семантическую сеть, используется словарь конкретной предметной области.

Второй шаг предлагаемого алгоритма является опциональным, т. е. он используется в том случае, если на первом шаге были получены неточные результаты: расхождение между запросом и подобранным описанием составило более 50 %.

Разработанный алгоритм в качестве входных данных принимает словесное описание объекта на русском языке из ограниченного набора терминов. Полученный запрос подвергается предобработке, затем начинается сам поиск. Результатом является название термина, эталонное описание которого наиболее совпадает с введенным. Термин представляет собой одно слово или словосочетание.

Описание основного алгоритма. Общая схема алгоритма представлена на рис. 1. На этапах вычисления косинусных расстояний и поиска по сети привлекаются заранее сформированные онтологии на базе статистических данных и синтаксических графов.

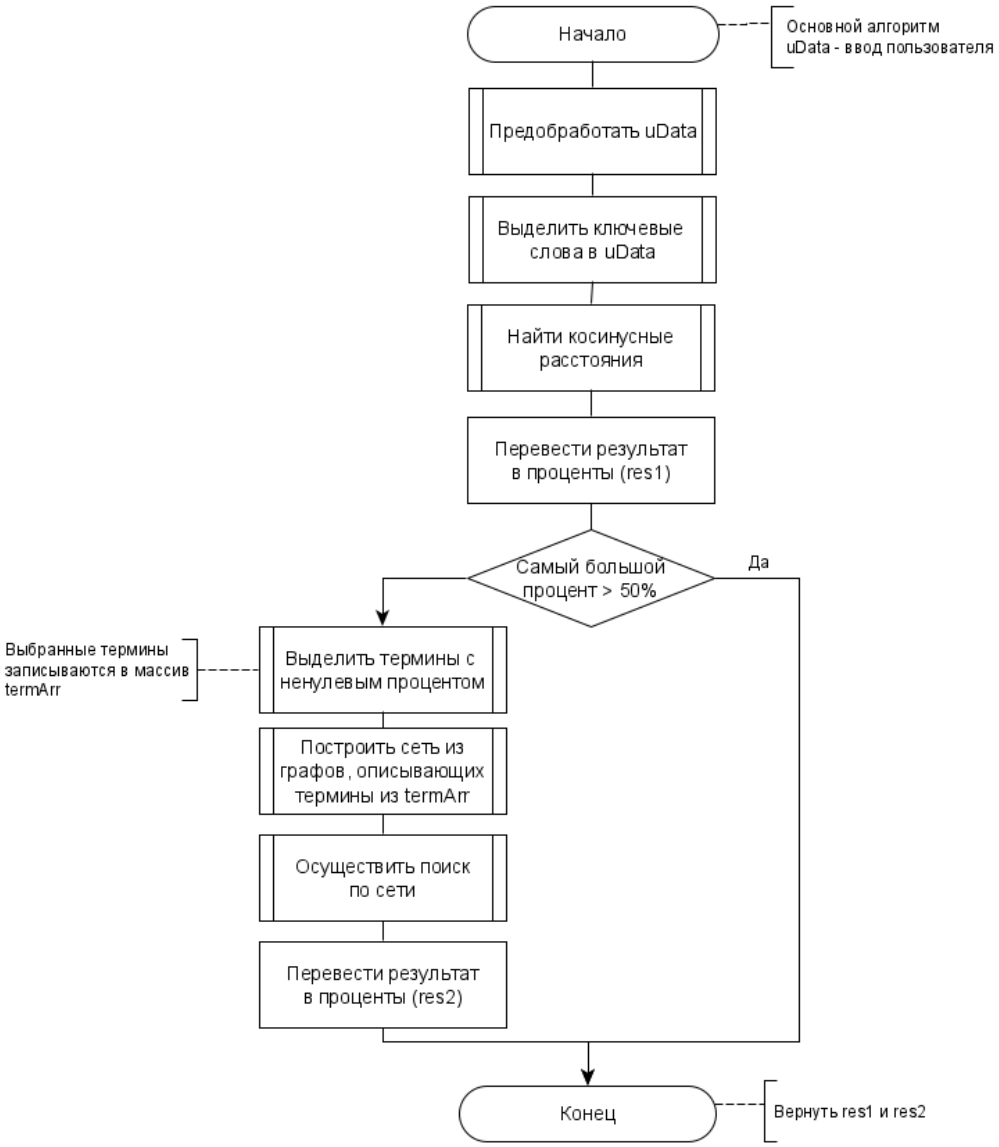

Рис. 1. Схема основного алгоритма

Онтология для статистического подхода. Полученные в результате опроса данные составляют единую систему — онтологию. В сфере информационных технологий этот термин рассматривается как удобная абстракция для отображения накопленных знаний в некоторой предметной области [1, с. 22]. Как правило, онтология включает не только общие, но и специфические предметные термины.

Формальной моделью онтологии предметной области будем называть пару

$$O = (\sigma, A),$$

где σ — сигнатура онтологии предметной области, т. е. множество ее ключевых понятий; *A* — множество предложений, описывающих их смысл [2, с. 287].

Предобработка исходных текстов. Для достижения наилучшего результата в определении объекта сначала осуществляется предобработка [3, с. 49–50] собранных данных, включающая процесс векторизации, необходимый для представления текста в удобном формате. Наиболее простой способ для этого — «мешок слов» (от англ. *bag-of-words*, или набор ключевых слов или терминов. При таком подходе игнорируется порядок единиц, входящих в состав рассматриваемого текста.

Под терминами коллекции документов *D* будем понимать все одиночные слова, которые встретились в тексте хотя бы в одном из документов. В итоге получается множество всех терминов коллекции:

$$\tau = \{t - \text{термин}\}.$$

Каждый документ в пространстве терминов представляется в виде вектора

$$\mathbf{d} = (t_1, ..., t_{|\tau|})^{\mathrm{T}}, \quad \mathbf{d} \in D,$$

где каждое число — координата вектора, соответствует конкретному термину и равняется его весу в данном документе.

Способ определения ключевых слов для статистического подхода. Для определения веса каждого слова используется метод TF-IDF (от англ. *Term Frequency — Inverse Document Frequency* — частота слова — обратная частота документа). Предполагается, что значимость термина прямо пропорциональна частоте его появления в документе и обратно пропорциональна доле документов в наборе, в которых он употреблен.

При таком подходе наибольший вес получает тот термин, который часто встречается в одном или небольшой группе документов, но не встречается в остальных, т. е. является отличительной особенностью на фоне часто употребляемых слов [4, с. 45–65].

Таким образом, с учетом коллекции документов *D*, термина *t* и текущего документа $\mathbf{d} \in D$, вес вычисляется по формуле

$$t(\mathbf{d}) = f(t, \mathbf{d}) \ln\left(\frac{|D|}{f(t, D)}\right),$$

где $f(t, \mathbf{d})$ — количество появлений *t* в **d**; $|D|$ — число документов в коллекции; $f(t, D)$ — количество документов, в которых встречается рассматриваемый термин

$$f(t, \mathbf{d}) = \frac{n_t}{\sum_k n_k},$$

где $n_{t,\,k}$ — количество появлений в рассматриваемом документе **d** соответствующего термина *t*, *k*:

$$f(t, D) = \left|\left\{\mathbf{d}_i \in D \mid t_i \in \mathbf{d}_i\right\}\right|.$$

Адаптация алгоритма TF-IDF к рассматриваемой задаче. Учитывая, что разрабатываемый метод планируется для использования с целью распознавания терминов в конкретных предметных областях, было сделано предположение о том, что вне зависимости от квалификации экспертов, привлекаемых к формулированию терминов, и их опыта работы в данной сфере большинство из данных ими определений, несмотря на их стилистические различия, будут содержать текстовые совпадения, описывающие ключевые признаки, присущие определяемому объекту, поэтому такие слова должны иметь наибольший вес.

В отличие от классического метода TF-IDF, в котором часто встречающиеся слова рассматриваются как слишком общие, не несущие информационной нагрузки, в основе разрабатываемого метода лежит противоположное предположение. Исходя из этого, была модифицирована формула расчета значимости слова, которая была приведена к виду

$$t(\mathbf{d}) = f(t, \mathbf{d}) \ln\left(\frac{|D|}{|D| - f(t, D)}\right).$$

С целью отсеять общие слова с минимальной дискриминационной силой дополнительно были введены следующие пороговые значения: если слово употребляется меньше, чем в 10 % определений, считается, что оно никак не характеризует термин. Аналогично, если слово встречается более чем в 90 % определений, оно также не считается ключевым.

Поиск нечетких дубликатов с помощью векторной модели. Два объекта считаются дубликатами, если они полностью совпадают. Если же один из них представляет собой видоизмененную копию другого, то в таком случае они являются нечеткими дубликатами [5]. В этом случае для двух векторов, сформированных на основе описанного выше подхода, определяется мера сходства, которая называется косинусной [6, с. 75–78]. То есть при наличии вектора запроса *q* и вектора *i*-го документа $\mathbf{d}_i$, косинусное сходство вычисляется по формуле

$$k_i = \frac{(q, \mathbf{d}_i)}{|q||\mathbf{d}_i|}.$$

Соответственно, чем значение ближе к единице, тем угол между векторами ближе к нулю градусов, и тем более схожи два рассматриваемых вектора.

Таким образом, формируется множество $K = \{k_1, ..., k_{|\tau|}\}$. Наиболее подходящим под исходное описание считается тот термин, косинусное сходство которого является наибольшим в полученном множестве *K*.

Семантические сети. Кроме подхода к представлению знаний в виде «мешка слов», также используются семантические сети, основанные на графах. Семантическая сеть — ориентированный граф, в котором вершины соответствуют конкретным фактам, общим понятиям, объектам, а дуги — отношениям или ассоциациям между ними [7, с. 7–9]. При этом объект может быть представлен как совокупность, состоящая из множества объектов, характеристик и свойств, набора состояний, действий, так или иначе связанных с ним.

Пример простейшей сети представлен на рис. 2.

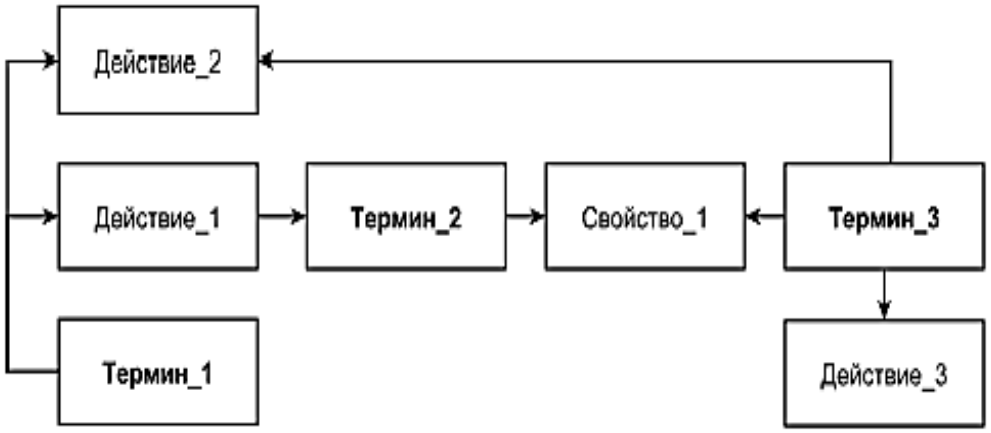

Рис. 2. Простейшая семантическая сеть

В предлагаемом методе сеть строится из синтаксических графов [8, с. 24–27], и в случае совпадения узлов двух или более графов, они объединяются и обновляются в соответствии с информацией, которая содержалась в узлах до их соединения. Следует учитывать, что второй этап метода реализуется только в том случае, если результаты применения статистического алгоритма дают степень уверенности меньше 50 %, что недостаточно для принятия решения.

Верификация разработанного метода. Для верификации разработанного метода использовали словарь терминов по тематике двигателей внутреннего сгорания [9], предоставленный кафедрой «Комбинированные двигатели и альтернативные энергоустановки» (Э-2) факультета «Энергомашиностроение» МГТУ им. Н.Э. Баумана.

Был подготовлен датасет, который формировался на основе устного опроса студентов третьего и четвертого курса бакалавриата, а также аспирантов кафедры Э-2. Всего было собрано 800 описаний 50 терминов из словаря. Собранные в ходе опроса аудио файлы были автоматически переведены в текстовый формат посредством использования специально разработанного бота для кроссплатформенной системы Discord.

Исследование влияния размера выборки на меру косинусного сходства. Для проведения эксперимента из собранного ранее датасета формировались онтологии следующих размеров: 100, 250, 300, 450 и 600 определений 50 терминов из словаря. Размер выборки основан на предположении, что малый размер датасета не позволяет в достаточной степени определить сходство двух определений. Кроме того, размер выборки напрямую будет влиять на принятие решения о необходимости применения второго этапа метода, основанного на синтаксических графах.

Результаты исследования сведены в таблицу.

Влияние размера выборки на меру косинусного сходства

Термин	Размер выборки, число определений				
	100	250	300	450	600
Аккумуляторный элемент	0,066	0,092	0,149	0,267	0,358
Гистерезис регулятора	0,113	0,105	0,110	0,121	0,152
Двигатель внутреннего сгорания	0,107	0,051	0,191	0,227	0,268
Изохорный процесс	0,671	0,811	0,949	0,975	0,984
Рабочее тело	0	0	0,221	0,282	0,330

Приведенные в таблице данные позволяют сделать следующие выводы.

1. С увеличением размера выборки увеличивается и мера сходства между определением из онтологии и соответствующим запросом, что, впрочем, было абсолютно предсказуемым.

2. Имели место случаи, когда накопление данных приводило к уменьшению меры сходства, что можно объяснить человеческим фактором. Например, определяя двигатель внутреннего сгорания, каждый из опрошенных использовал формулировку, которая казалась ему наиболее понятной и правильной. Таким образом, полученные описания термина значительно различались между собой, что сильно затруднило выявление ключевых слов.

3. В онтологии присутствовали термины, которые вызвали у опрашиваемых большое затруднение, например, «гистерезис регулятора». Многие студенты не знали, что это такое, либо ошибались в своих предположениях, поэтому косинусная мера сходства с увеличением выборки росла незначительно.

Таким образом, качество и размер выборки играют ключевую роль в повышении результативности работы рассматриваемого метода и качества результата. От них также зависит вероятность проведения дополнительного анализа на основе семантической сети.

Влияние размера выборки на процент обращения к вспомогательному методу. Так же, как и в предыдущем исследовании, все датасеты формировались из определений, которые давали студенты в ходе опроса. Анализ проводили на

наборах данных в 100, 250, 400, 550 и 700 описаний 50 терминов из словаря. Также вычисляли ключевые слова и их вес в рамках рассматриваемой онтологии. В качестве входных данных использовали определения из предметного словаря на каждой созданной онтологии. Каждый раз, когда результат применения статистического метода оценки не проходил по критерию принятия решения, фиксировалось обращение к дополняющему методу. Для получения статистики по неверным результатам ответ сравнивали с правильным определением термина, приведенным в словаре. В итоге по полученным значениям была построена диаграмма, представленная на рис. 3, которая позволяет сделать следующие выводы.

1. На маленьком датасете (100 определений) доля обращений очень высока (более 70 %), из них 10 % приходится на неверный результат.

2. В целом с увеличением размера данных корректность работы статистического метода увеличивается, о чем свидетельствует уменьшающийся процент обращений ко второму этапу алгоритма. Так, доля запросов, которые обрабатываются с привлечением вспомогательного метода при объеме данных в 700 определений, меньше примерно на 26 %, чем при 100 элементах, что потенциально сокращает время обработки запроса пользователя.

3. Аналогичная ситуация наблюдается с процентом неправильных ответов, которые даются при обработке вспомогательным методом, он также снижается. Это обусловливается прежде всего тем, что алгоритм косинусного сходства с увеличением выборки лучше идентифицирует потенциально верные термины, из которых на втором этапе алгоритма строится сеть.

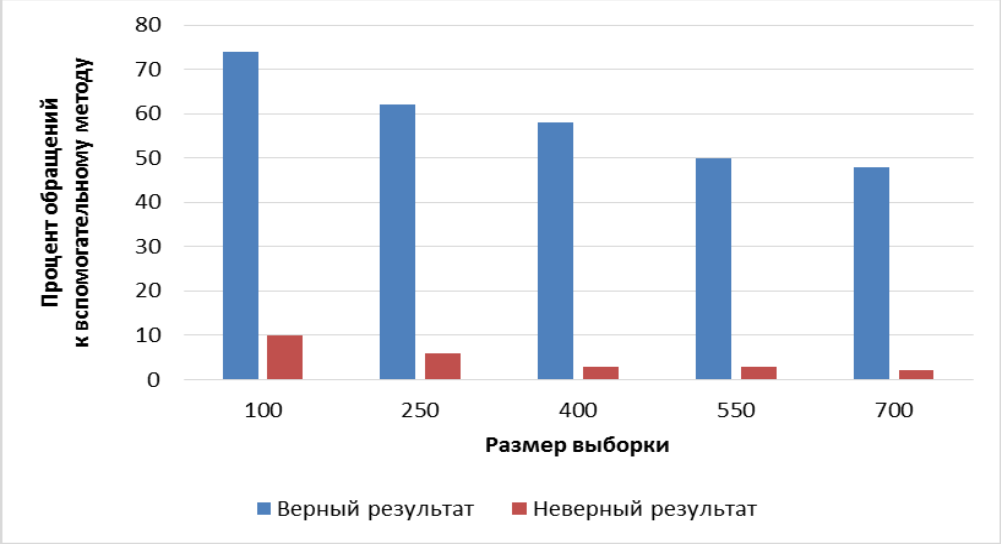

Рис. 3. Влияние размера выборки на процент обращения к вспомогательному методу

Заключение. Таким образом, разработан метод определения объекта из ограниченной выборки терминов определенной предметной области по их нечеткому описанию на русском языке. Предложен способ решения данной задачи, основанный на комбинировании метода TF-IDF для формирования статистической онтологии и сети синтаксических графов в качестве вспомогательного метода, который привлекается только в случае, если результаты обработки первым методом не удовлетворяют критерию для принятия решения. Для поиска нечетких дубликатов между запросом пользователя и собранными заранее данными предложено использовать косинусное сходство.

Проанализировано влияние размера выборки на меру сходства и на долю обращения к вспомогательному методу, который подтвердил, что наличие качественного и репрезентативного набора входных данных существенно влияет на результат работы метода и уменьшает количество обращений к вспомогательному методу на основе семантической сети, что позволяет сократить время обработки запроса. Кроме того, сокращается число случаев, когда сеть определяет термин неверно. В дальнейшем метод может быть усовершенствован путем расширения используемых онтологий. Также для уменьшения времени обработки запроса следует привлечь параллельный поиск по базам знаний.

Литература

[1] Большакова Е.И., Воронцов К.В., Ефремова Н.Э. и др. Автоматическая обработка текстов на естественном языке и анализ данных. М., Изд-во НИУ ВШЭ, 2017.

[2] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб., Питер, 2000.

[3] Srividhya V., Anitha R. Evaluating preprocessing techniques in text categorization. *Int. J. Comput. Sci. Appl.,* 2010, vol. 47, no. 11, pp. 49–51.

[4] Aizawa A. An information-theoretic perspective of TF–IDF measures. *Inf. Process. Manag.,* 2003, vol. 39, no. 1, pp. 45–65. DOI: https://doi.org/10.1016/S0306-4573(02)00021-3

[5] Зиберт А.О., Хрусталев В.И. Разработка системы определения наличия заимствований в работах студентов высших учебных заведений. Алгоритмы поиска нечетких дубликатов. *Universum: Технические науки*, 2014, № 3.
URL: https://7universum.com/ru/tech/archive/item/1139

[6] Преображенский Ю.П., Коновалов В.М. О методах создания рекомендательных систем. *Вестник ВИВТ*, 2019, № 4, с. 75–79.

[7] Бабкин Э.А., Козырев О.Р., Куркина И.В. Принципы и алгоритмы искусственного интеллекта. Нижний Новгород, НГТУ, 2006.

[8] Теньер Л. Основы структурного синтаксиса. Прогресс, 1988.

[9] Еникеев Р.Д., Рудой Б.П. Двигатели внутреннего сгорания. Основные термины и русско-английские соответствия. М., Машиностроение, 2004.

Брянская Екатерина Вадимовна — студентка кафедры «Программное обеспечение ЭВМ и информационные технологии», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Барышникова Марина Юрьевна, кандидат технических наук, доцент кафедры «Программное обеспечение ЭВМ и информационные технологии», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Брянская Е.В. Метод определения объекта из ограниченной выборки по нечеткому описанию на естественном языке. *Политехнический молодежный журнал*, 2023, № 01(78). http://dx.doi.org/10.18698/2541-8009-2023-01-856

A METHOD FOR DETERMINING AN OBJECT FROM A LIMITED SAMPLE BASED ON A FUZZY DESCRIPTION IN NATURAL LANGUAGE

E.V. Bryanskaya bryanskayakatya@yandex.ru

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

The article is devoted to solving the problem of determining an object from a limited sample by a fuzzy description in Russian. The developed method consists in combining two main approaches to solving typical problems in this area, one of which is based on a statistical algorithm, and the second is based on the use of a semantic network. Each of them requires its own ontology. To form the knowledge base of the first stage, the adapted TF-IDF method is used, for the second, a set of syntactic graphs is taken as a basis. Cosine similarity is used to find fuzzy duplicates between the user's query and the knowledge base created in advance. The paper investigates the influence of the sample size on the similarity measure and the accuracy of the object definition. The proportion of requests to the second step of the proposed method is also evaluated, including in order to determine what percentage of these requests falls on an incorrect assumption made at the first stage.

Keywords

Natural language, natural language text processing, ontology, "bag-of-words", vectorization, TF-IDF, fuzzy duplicates, cosine similarity, semantic network, syntactic graph

Received 21.12.2022

References

[1] Bolshakova E.I., Vorontsov K.V., Efremova N.E. et al. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh [Automatic natural language text processing and data analysis]. Moscow, Izd-vo NIU VShE Publ., 2017 (in Russ.).

[2] Gavrilova T.A., Khoroshevskiy V.F. Bazy znaniy intellektualnykh system [Knowledge bases of intelligent systems]. Sankt-Petersburg, Piter Publ., 2000 (in Russ.).

[3] Srividhya V., Anitha R. Evaluating preprocessing techniques in text categorization. *Int. J. Comput. Sci. Appl*., 2010, vol. 47, no. 11, pp. 49–51.

[4] Akiko Aizawa An information-theoretic perspective of TF–IDF measures. *Inf. Process. Manag*., 2003, vol. 39, no. 1, pp. 45–65.
DOI: https://doi.org/10.1016/S0306-4573(02)00021-3

[5] Zibert A.O., Khrustalev V.I. Development of a system for determining the existence of adoption in the works of the students. The search algorithms of indistinct duplicates. *Universum: Tekhnicheskie nauki*, 2014, no. 3.
URL: https://7universum.com/ru/tech/archive/item/1139 (in Russ.).

[6] Preobrazhenskiy Yu.P., Konovalov V.M. About methods for creating recommendation systems. *Vestnik VIVT*, 2019, no. 4, pp. 75–79 (in Russ.).

[7] Babkin E.A., Kozyrev O.R., Kurkina I.V. Printsipy i algoritmy iskusstvennogo intellekta [Principles and algorithms of artificial intelligence]. Nizhniy Novgorod, NGTU Publ., 2006 (in Russ.).
[8] Tesniere L. Elements de syntaxe structurale. Librairie C. Klinckseick, 1959. (Russ. ed.: Osnovy strukturnogo sintaksisa. Progress Publ., 1988.)
[9] Enikeev R.D., Rudoy B.P. Dvigateli vnutrennego sgoraniya. Osnovnye terminy i russko-angliyskie sootvetstviya [Internal combustion engines. Basic terms and Russian-English correspondences]. Moscow, Mashinostroenie Publ., 2004 (in Russ.).

Bryanskaya E.V. — Student, Department of Software for Computers and Information Technologies, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Baryshnikova M.Yu., Cand. Sc. (Eng.). Assoc. Professor, Department of Software for Computers and Information Technologies, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Bryanskaya E.V. A method for determining an object from a limited sample based on a fuzzy description in natural language. *Politekhnicheskiy molodezhnyy zhurnal* [Politechnical student journal], 2023, no. 01(78). http://dx.doi.org/10.18698/2541-8009-2023-01-856.html (in Russ.).