УДК 004.9

СОВМЕСТНОЕ ИСПОЛЬЗОВАНИЕ ЧАСТОТНОГО АНАЛИЗАТОРА С КОНЦЕПТУАЛЬНОЙ СТРУКТУРОЙ ДЛЯ ПОВЫШЕНИЯ ПРОИЗВОДИТЕЛЬНОСТИ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Семенов Р.Э., Сорокин А.Б.

МИРЭА - Российский технологический университет, 119454, Россия, г. Москва, проспект Вернадского, 78, *e-mail: 9629790@gmail.com.*

Исследования в области концептуальной структуры текста позволяют более точно определить весовой критерий ключевых слов и исключить из списка объекты, которые в меньшей мере отражают достоверность информации, полученной из документа. На основании полученных данных система может быть модернизирована и получать постоянную поддержку в виде обновлений баз релевантности поисковых запросов и расчета критерий весовых характеристик. Проектируемая система поддержки принятия решений может быть внедрена в крупные системы интернет-ресурсов, использоваться как самостоятельный проект, поддерживающий автономную работоспособность, с помощью внедрения словарей слов и правил с возможностью обновления, или, для внедрения в интеллектуальные системы.

Ключевые слова: анализ текста, индексация терминов, концептуальная структура текста, вес термина, частотная модель, текстовое моделирование, вероятностная модель, сила термина.

SHARING A FREQUENCY ANALYZER WITH A CONCEPTUAL FRAMEWORK TO IMPROVE THE PERFORMANCE OF A DECISION SUPPORT SYSTEM

Semenov R.E., Sorokin A.B.

MIREA - Russian Technological University, 119454, Moscow, 78 Vernadskogo Avenue, Russia, *e-mail: 9629790@gmail.com.*

Research in the field of the conceptual structure of the text allows you to more accurately determine the weight criterion of keywords and exclude from the list objects that less reflect the reliability of the information obtained from the document. Based on the data obtained, the system can be upgraded and receive constant support in the form of updates to the databases of relevance of search queries and calculation of criteria for weight characteristics. The designed decision support system can be implemented in large systems of Internet resources, used as an independent project supporting autonomous operation, through the introduction of dictionaries of words and rules with the possibility of updating, or for implementation in intelligent systems.

Keywords: text analysis, indexing of terms, conceptual structure of the text, term weight, frequency model, text modeling, probabilistic model, term strength.

Введение

Ежедневно миллионы запросов, на электронных ресурсах, обрабатываются с помощью различных алгоритмов поиска, распознавания и обработки текста. Это позволяет находить нужную информацию среди похожих документов и тех, которые ее не содержат. Проблемой предметной области обработки и анализа текста выступает огромный поток информации, который постоянно увеличивается. Для решения проблемы зашумленности текстовых документов разрабатываются новые способы выявления главной мысли текстов и обработки поисковых запросов. Существуют общеизвестные алгоритмы, которые использовались с момента обнаружения данной проблемы и используются в настоящее время. Время обработки таких алгоритмов велико и не может быть использовано для быстрого получения ответа. Сейчас вычислительные возможности позволяют использовать одновременно некоторое количество алгоритмов, которые выполняются параллельно, что позволяет проводить многосторонний анализ за кротчайший промежуток времени. Необходимо выявить наиболее эффективные методы, которые будут взаимодействовать между собой для увеличения эффективности. Основная идея модели, на которой построен используемый алгоритм, была предложена Джерардом Солтоном в 1973 году. В векторном

представлении модели информации, смысл передаваемой информации выделяется списком терминов. С помощью векторного присвоения числа терминов каждому проанализированному документу, такой список позволяет найти запрашиваемую информацию. Образ смыслового представления информации документа для поискового запроса зависит от частоты использования выделенного термина и его относительного веса. Простая модель использует двоичный вектор, его значения зависят от того, принадлежит ли конкретный термин данному контексту или нет. Размерность вектора D увеличивается в зависимости от объема текстовой информации в анализируемом документе. При обработке запроса, который является текстом, происходит такой же разбор на термины поискового запроса и присвоение вектора q [5]. После проведенных исследований, необходимо получить структуру для сравнения образов поисковых запросов и образов документов, которые после представленных далее алгоритмов позволит выявить соответствия значимости терминов.

Модель представления анализируемой информации

Чтобы проанализировать текст большого объема, содержащийся в одном или нескольких документах, потребуется много вычислительной мощности и времени на обработку. Для упрощения и ускорения обработки необходимого материала используют информационный поток. Он представляет из себя массив, содержащий наборы документов. Матрица информационного массива изображена на Рисунке 1.

	Термин 1	Термин 2	...	Термин j	...	Термин D
Документ 1	W_{11}	W_{12}	...	W_{1j}	...	W_{1D}
Документ 2	W_{21}	W_{22}	...	W_{2j}	...	W_{2D}
...	...	...	...	...	...	...
Документ i	W_{i1}	W_{i2}	...	W_{ij}	...	W_{iD}
...	...	...	...	...	...	...
Документ N	W_{N1}	W_{N2}	...	W_{Nj}	...	W_{ND}

Рисунок 1. – Представление информационного потока в виде «документ-термин»

Где W_{ij}– вес термина t_j в документе d_i.

Для того, чтобы полностью определить векторную модель, необходимо выявить, каким образом будет определяться вес термина в документе для составления полного образа. Далее будут представлены стандартные способы задания функции взвешивания:

– Вес булевой алгебры. Присваивается значение «1», если термин содержится в анализируемом документе, если отсутствует, то присваивается значение «0».

– Частота термина (TF – term frequency). При определении частоты термина, определяется его вес. Для этого по функции определяется переменная, которая зависит от количества появления термина в анализируемом документе.

– Обратная частота документа (TF-IDF – term frequency – inverse document frequency). Значение высчитывается по произведению функции, в которой определяется количество использования термина в документе, и функции, в которой определяется обратное значения количества документов, в которых присутствует используемый термин [7].

Для составления векторного образа смыслового содержания документа, и дальнейшего представления концепта анализируемого текста, необходимо правильно, быстро и однозначно индексировать поступающую информацию. Первым этапом индексирования является отбор круга терминов, которые наилучшим образом характеризуют конкретно выбранный документ. Это обусловлено проблемой захламленности информационного потока. В 2022 году, большое количество документов, содержащих крупный объем текстовой информации, находится в свободном и общем доступе. Каждый поисковый запрос обрабатывает документы, которые содержат введенные слова и словосочетания, но на самом деле, найденный материал не в полной мере соответствует тому, который имелся ввиду. Поэтому, важно правильно произвести отбор терминов каждого анализируемого документа. Это позволит приблизится к решению проблемы зашумленности информационного поля в Интернете и позволит более быстро представить текст для моделирования текстового ответа на запрос.

Определение частотных параметров терминов

Частота появления в тексте определенных слов, называемая абсолютной частотой, является эффективной при, сравнительно, небольшом объеме текста, так как при использовании поискового запроса по одному, термину будет найдено огромное количество документов, содержащих такое же слово или словосочетание. Другие документы также содержали бы термины абсолютной частоты и одинаковые стоп-словари, что в итоге привело

бы к крайне низким результатам текстового анализа. Поэтому используются методы приведения относительной частоты, которые выделяют термины с высокой частотой употребления в данном документе с использованием этих же терминов во всем информационном массиве сети Интернет. Модели с использованием относительной частоты выделяют термины, которые больше встречаются в отдельных документах, а вне документов употребляются значительно реже. Это действие позволяет добиться отобразить содержание конкретного документа и различать разные, проанализированные, документы

Для расчета документной частоты, в которой термин t_i встречается, относительно числа документов всего массива используется взвешивающая функция, представленная выражением (1).

$$(IDF)_i = \log \frac{N}{(DF)_i} \qquad (1)$$

Где $(DF)_i$ – частота появления термина в документе, относительно появления данного термина в массиве документов.

N – число документов, которые находятся в рассматриваемом информационном потоке. Используя приведенную функцию, наибольшие значения присваиваются тем терминам, которые появляются только в нескольких документах. Чем больше раз термин встречается в документах массива, тем меньше будет значение обратной документной частоты.

Для конкретизации выбора термина при анализе и присвоении индекса, используется оценка силы термина. Под силой имеется ввиду индивидуальность термина, который характеризует документы максимально непохожими друг на друга. Наименьшей силой обладают ключевые слова, которые только приближенно описывают документы, отличить которые друг от друга трудно. Чем сильнее различия по выделенным терминам отдельных документов, тем легче находить определенные документы. Если сравниваемые документы представлены похожими векторами терминов, то пространство индексирования сжимается, поэтому достижение разграничения релевантных и нерелевантных документов будет затруднительным.

Значимость силы термина t_i определяется его силой в документе $(DF)_i$. Она определяется разностью средними значениями показателей документов, которые попарно подобны и когда термин t_i отсутствует в векторах документов массива. Также характеризуется средним значением документов, которые попарно подобны, когда термин t_i присутствует [2]. Если рассматриваемый термин представляет ценность для индексирования, то наличие его в векторе делает документы менее похожими друг на друга. Это значит, что среднее попарное подобие документов уменьшается, а коэффициент различия положителен. В ситуации, когда подобие увеличивается, коэффициент различия будет отрицательным.

При индексировании терминов для каждого документа нужно рассматривать все проанализированные правила и алгоритмы. Одним из главных способов, по которому можно выделить определенные термины в документе, является распределение их по документной частоте $(DF)_i$ и частоте встречаемости F_i. Встречаемость термина t_i в массиве документов определяется выражением (2).

$$F_i = \sum_{k=1}^{N} (f_i)_k \qquad (2)$$

Описанные ниже правила будут определять этапы проведения анализа автоматической системы поддержки принятия решений текстового моделирования при приведении текстовой информации документа к концепту. При индексации терминов, следует уделять большее внимание таким словам, которые имеют наивысшую различительную силу [4]. Наибольшей силой будут обладать термины со средним значением частоты встречаемости, обозначающейся как F_i и документной частотой, при которой суммарная частота в информационном массиве составляет менее половины частоты встречаемости термина. Менее эффективными являются термины с низкой документной и суммарной частотой, у которых сила будет близка к нулю. Неэффективными являются термины с отрицательным значением силы, которые определяются высокой частотой употребления в информационном массиве.

Концептуальная структура распределения параметров

Система поддержки принятия решений текстового моделирования для расчета критериев веса использует концепт, по структуре которого, выделяются наиболее ценные термины, используемые в тексте. Концептом называется логическая структура, позволяющая провести оценку информации с помощью морфологического анализа. Концептуальная структура, представленная на Рисунке 2, демонстрирует взаимосвязи элементов для определения наибольшего веса каждого проанализированного участка текста.

Каждый элемент структуры соотносится с набором слов конкретных частей речи. Наборы слов объединяют части речи морфологическими признаками, синтаксической ролью в предложении и общим грамматическим значением [2], [6]. Пара характеристик <название_параметра, значение_параметра> называется

морфологическим параметром. Для параметра характерны род, число, время, склонение и другие признаки слов, принятые в используемом языке [9]. Значением параметра является конкретное значение, которое может принимать данный признак. Так, например, род может быть средним, мужским, женским, Несколько параметров может быть сопоставлено в одной словоформе [11]. После получения нормальных форм слов с набором выделенных параметров, проводится анализ слов в тексте, которые, после частотного анализа, оказались наиболее релевантными [10]. Данный алгоритм позволяет с наименьшим количеством ошибок и неточностей проиндексировать термины в документе и подготовить составляющие текста для более глубокого анализа для системы текстового моделирования.

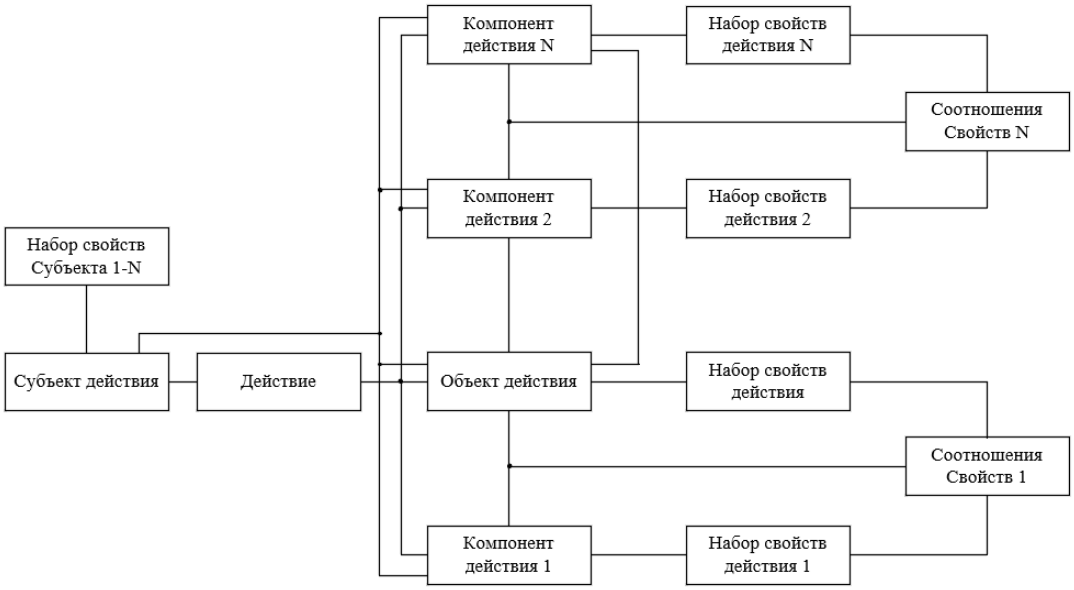

Рисунок 2. – Концептуальная структура текстового моделирования

Для составления частотной модели терминов отдельно взятого текста необходимо выявить и произвести расчет весов. Частотная модель взвешивания терминов применяется в частотном методе индексирования [8].

Приведенная выражением (3) весовая функция позволит привести расчет весовой характеристики терминов.

$$\mathrm{TF}-\mathrm{IDF}=W_i=(TF)_i\times(IDF)_i \qquad (3)$$

Где W_i– вес, приписываемый термину.

$(TF)_i$ – частота термина в документе.

$(IDF)_i$– обратная документная частота.

После проведения расчета весов терминов, необходимо сравнить значения и выявить ключевые слова с наибольшим значением. Самыми релевантными будут являться некоторые термины, количество которых зависит от объема анализируемого документа, они будут иметь значения, намного превышающие остальные. Более точную информацию возможно получить руководствуясь алгоритмом расчета вероятностной модели оценки весов терминов. Данный алгоритм основан на оценке вероятности того, что документ соответствует поисковому образу и является релевантным. В сравнении с частотным методом взвешивания, с помощью которого можно лишь формально произвести абстрактный расчет, без учета информационных потребностей [1], вероятностная модель сопоставляет термины с поисковыми образами документов и представлена выражением (4), которое использует теорему Байеса.

$$P(w_1|d)=\frac{P(d|w_1)P(w_1)}{P(d)} \qquad (4)$$

Где $P(w_1|d)$ – вероятность события, при котором определенный документ является релевантным поисковому запросу.

w_1– случайно выбранный документ для соответствия поисковому запросу.

d – анализируемый документ.

$P(w_1)$ – вероятность случайно выбранного документа оказаться релевантным поисковому запросу.

$P(d)$ – вероятность того, что из всего множества документов для рассмотрения выбран анализируемый документ.

$P(d|w_1)$ – вероятность того, что анализируемый документ выбран из множества релевантных документов.

По результатам определенной вероятности, необходимо выбрать термины с наибольшим значением. Значение обозначает вероятность события, при котором документ, проиндексированный конкретным термином, будет релевантным по соответствующему поисковому запросу.

Заключение

В ходе проведенных исследований были выявлены наиболее эффективные методы анализа текстового документа. Продемонстрированы наиболее эффективные алгоритмы, которые по поставленной задачи имеют возможность одновременной обработки данных. Этот метод позволит создать автономную систему, которая способна одновременно проводить анализ, поиск информации и моделировать текстовые конструкции. На основании полученных результатов, будет разработано программное обеспечение, содержащее алгоритмы анализа, представленные в статье, модули текстового моделирования, на основании модели концептуальной структуры, и базы данных для составления логически выверенных словарей. Задача текстового моделирования для системы поддержки принятия решения заключается в выборе метода индексации терминов и сопоставлении таких слов с концептуальной структурой.

Каждый анализируемый документ индексируется терминами, которые выбираются после проведения определенных алгоритмов. Для документов малого объема система проведет общий частотный анализ слов, затем данные, полученные после приведения текстовой информации к концепту, будут соотнесены с полученными терминами и некоторые будут исключены из списка индексирования анализируемого документа, так как не будут являться релевантными для дальнейшего анализа.

В случаях, когда текст содержит много сложных предложений, система текстового моделирования проводит дополнительные алгоритмы для поиска релевантных ключевых слов. Необходима постоянная поддержка системы, которая включает обновление баз правил языка, на котором проводится анализ, обновление алгоритмов вычисления весов терминов и релевантности поисковых запросов.

Список литературы

1. Амаева Л.А. Сравнительный анализ методов интеллектуального анализа данных. Инновационная наука. 2018;2(1):27-29.

2. Андрейчикова О.Н., Андрейчиков А.В., Интеллектуальные цифровые технологии концептуального проектирования инженерных решений: учебник. М.: ИНФРА-М; 2019. 511 с. ISBN 978-5-16-014884-7.

3. Анфёров М.А. Генетический алгоритм кластеризации. Russian Technological Journal, 2019. – № 6 (7), с. 134 – 150.

4. Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. – М.: МИЭМ, 2011. – 272 с.

5. Кронгауз М.А. Семантика: Учебник – М., издательский центр «Академия», 2005. – 352 с.

6. Сорокин А.Б., Лобанов Д.А. Концептуальное проектирование интеллектуальных систем. Информационные технологии, 2018. – №1 (24). – с. 3-10.

7. Сорокин А.Б., Смольянинова В.А. Концептуальное проектирование экспертных систем поддержки принятия решений. Информационные технологии, 2017. – №9 (23). – с. 634 – 641.

8. Соснина Е. П. Введение в прикладную лингвистику: учебное пособие. –Ульяновск: УлГТУ, 2012. – 110 с

9. Feldman S.E. The answer machine. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2012. – Vol. 4. – Pp. 1–137.

10. Rodrigues da Silva, A. Model-driven engineering: A survey supported by the unified conceptual model. Computer Languages, Systems & Structures 43, 139-155 (2015).

11. Zhang, X., Zhao, H., Chen, D.-Y., Semantic Mapping Methods Between Expert View and Ontology View. Journal of Software 31(9), 2855-2882 (2020).

References

1. Amaeva L.A. Comparative analysis of data mining methods. Innovative science., 2(1), pp. 27-29, 2018. Journal article.

2. Andreychikov A.V., Andreychikova O.N. Intelligent digital technologies of conceptual design of engineering solutions: textbook, INFRA-M, p. 511, 2019.

3. Anferov M.A. Genetic clustering algorithm. Russian Technological Journal, No. 6, pp. 134 – 150, 2019.

4. Bol'shakova E.I. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika, MIEM, p. 272, 2011.

5. Krongauz M. A. Semantika: Uchebnik, Izdatel'skij centr «Akademiya», p. 352, 2005.

6. Sorokin A. B., Lobanov D. A. Konceptual'noe proektirovanie intellektual'nyh sistem, Informacionnye tekhnologii,

№1 (24), pp. 3 – 10, 2018.
7. Sorokin A. B. Smol'yaninova V. A. Konceptual'noe proektirovanie ekspertnyh sistem In-formacionnye tekhnologii., № 9 (23). pp. 634 – 641, 2017.
8. Sosnina E. P. Vvedenie v prikladnuyu lingvistiku: uchebnoe posobie, Ulyanovsk, UlGTU, p. 110, 2012.
9. Feldman S. E. The answer machine. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, Vol. 4, pp. 1 – 137, 2012.
10. Rodrigues da Silva, A. Model-driven engineering: A survey supported by the unified conceptual model. Computer Languages, Systems & Structures 43, 139-155 (2015).
11. Zhang, X., Zhao, H., Chen, D.-Y., Semantic Mapping Methods Between Expert View and Ontology View. Journal of Software 31(9), 2855-2882 (2020).