УДК 004.02:004.62

РАЗРАБОТКА МАТЕМАТИЧЕСКОЙ МОДЕЛИ ПРОЦЕССА ПОИСКА И ОБРАБОТКИ НЕСТРУКТУРИРОВАННОЙ ДОКУМЕНТИРОВАННОЙ ИНФОРМАЦИИ

Гагарина Л.Г., Слюсарь В.В., Федоров А.Р., Федоров П.А.

ФГАОУ ВО «Национальный исследовательский университет «Московский институт электронной техники», Москва (Зеленоград), e-mail: vslyusar@mail.ru

В статье рассмотрен процесс построения формализованной математической модели поиска и обработки неструктурированной документированной информации в рамках реализации концепции промышленного Интернета вещей для автоматизации управления распределенными производственными структурами. Дан краткий обзор наиболее распространенных способов математических методов извлечения контента из массивов неструктурированной документированной информации (метод TF-IDF, применение скрытых марковских моделей). Разработано формализованное представление для дальнейшего построения алгоритмов извлечения данных из массивов неструктурированной информации с использованием моделей ссылочного ранжирования (ранжирование документов массива представлено в виде может быть представлено в виде системы итерационных линейных алгебраических уравнений), индексации документов (определена зависимость веса термина в данной документе от частоты термина и количества документов, в которых данный термин встречается) и формализации обработки неструктурированных данных веб-документов (определено конечное множество выделенных из неструктурированного массива релевантных данных, очищенных от шумов, над которыми возможны дальнейшие действия по структуризации). На основе проведенных исследований будут разработаны алгоритмы и методы обработки неструктурированной информации в распределенных АСУП, ориентированных на работу в условиях предприятий, реализующих концепцию промышленного Интернета вещей.

Ключевые слова: промышленный Интернет вещей, поиск и обработка информации, математическая модель

DEVELOPMENT OF MATHEMATICAL MODEL OF UNSTRUCTURED DOCUMENTED INFORMATION SEARCHING AND PROCESSING

Gagarina L.G., Slyusar V.V., Fedorov A.R., Fedorov P.A.

National Research University of Electronic Technology, Moscow (Zelenograd), e-mail: vslyusar@mail.ru

The article deals with the process of constructing a formalized mathematical model for searching and processing unstructured documented information within the framework of the implementation of the industrial Internet concept. Things for automating the management of distributed production structures. A brief overview of the most common methods of mathematical methods for extracting content from arrays of unstructured documented information (TF-IDF method, application of hidden Markov models) is given. A formalized representation for the further construction of algorithms for extracting data from arrays of unstructured information using reference ranking models is developed (the ranking of the array documents can be represented in the form of a system of iterative linear algebraic equations), indexing of documents (the weight of the term in this document is determined from the frequency of the term and the number of documents in which the term occurs) and the formalization of processing is unstructured s Web document data (defined by a finite set of isolated unstructured array of relevant data, cleared of noise over which possible further actions to structure). Based on the conducted research, algorithms and methods for processing unstructured information in distributed automated control systems will be developed, oriented to work in the conditions of enterprises that implement the concept of the industrial Internet of things.

Keywords: Industrial Internet of Things, information retrieval and processing, mathematical model

В настоящее время в промышленном производстве всё большее распространение получает концепция так называемого индустриального (или «промышленного») Интернета вещей (Industrial Internet of Things, IIoT), под которым понимается интернет вещей для корпоративного/ отраслевого применения. Реализация данной концепции предполагает развертывание в рамках предприятия (корпорации) системы объединенных компьютерных сетей и подключенных промышленных (производственных) объектов со встроенными датчиками и ПО для сбора и обмена данными, с возможностью удаленного контроля и управления в автоматизированном режиме, без участия человека [1].

Одной из особенностей функционирования подобных систем является необходимость обработки значительных объемов неструктурированных данных, их фильтрации и адекватной интерпретации, что является приоритетной задачей для предприятий.

Для выполнения функций обработки неструктурированных данных в рамках реализации концепции промышленного Интернета вещей в рамках распределенной автоматизированной системы управления предприятием (АСУП) реализуется технология Web Mining – использование методов Data Mining для исследования и извлечения информации из Web-документов и сервисов [2].

Цель исследования: при создании эффективных подсистем поиска и обработки неструктурированной информации в рамках АСУП промышленного Интернета вещей, реализующих технологию WebMining разработчики могут столкнуться с проблемами, аналогичными проблемам, характерным для работы с информацией в сети Интернет, а именно:

1. Поиск значимой информации. Зачастую множество ссылок, предоставляемых поисковыми системами, только незначительный процент предоставляет релевантную информацию; кроме того, поиск неиндексированной информации затрудняется за счет низкой повторяемости вызовов.

2. Создание новых знаний вне полученной из поисковой системы информации.

3. Персонализация информации.

4. Изучение потребителя или индивидуального пользователя (необходимости персонализации поисковых систем) [3].

Проведем краткий обзор наиболее известных математических методов извлечения контента из массивов неструктурированной документированной информации.

Метод TFIDF

Данный метод предполагает использование двух понятий (метрик):

1. Частота термина TF (*term frequency*), определяемая как отношение количества выявлений некоторого термина к общему количеству слов документа. Данная метрика позволяет оценить вестермина t_i в пределах отдельного документа:

$$tf\left(t,d\right)=\frac{n_i}{\sum_k n_k}, \qquad (1)$$

где n_i – количество вхождений *i*-го термина в документ, $\sum_k n_k$ – общее количество слов в данном документе, включающее общеупотребительные и связующие слова [4].

2. «Инвертированная частота документа» IDF (*inverse document frequency)* инверсия частотности встречаемости заданного термина встречается в документах массива, позволяющая снизить веса связующих и общеупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF (2):

$$idf\left(t,D\right)=log\frac{\left|D\right|}{\left|\left\{t\in d_i|d_i\in D\right\}\right|}, \qquad (2)$$

где |D| – количество документов массива, $\left|\left\{t\in d_i|d_i\in D\right\}\right|$ – количество документов в массиве D, в которых встречается *t*.

Таким образом, мера TF-IDF определяется по следующей формуле:

$$tf-idf\left(t,d,D\right)=tf\left(t,d\right)\times idf\left(t,D\right). \qquad (3)$$

При использовании данной меры максимальный вес получают термины с высокой частотой в пределах одного документа и с низкой частотой употреблений в других документах [4].

Известны также различные модификации модели tf-idf. Одним из примеров является мера Okapi BM25 [5]. Наиболее существенным ограничением использования tf-idf является необходимость неизменности набора данных в течение всего времени расчета, что значительно усложняет вычисления, если требуется их проведение в режиме реального времени.

Скрытые марковские модели

Рассмотрим формальное описание скрытой марковской модели (СММ).

Каждая модель определяется следующими параметрами:

1. Множество $S = \{s1, s2, ..., s_N\}$ из N состояний.

2. Начальное распределение вероятностей $P = \{p_i\}$.

3. Матрица вероятностей переходов между состояниями $A = \{ai\}$.

4. Матрица вероятности генерации наблюдений $B = \{b_j(Ot)\}$, где $bj(Ot)$ – вероятность генерации наблюдения Ot в момент времени t в состоянии $qt = sj$, $bj(Ot) = P(Ot|qt = sj)$.

Представленная таким образом модель является одномерной, но для поиска и обработки документированной информации интерес представляют псевдомногомерные СММ, состоящие из конечного количества элементов, называемых суперпозициями, каждый из который в свою очередь представляет собой отдельную СММ [6].

В рамках функционирования автоматизированной поисковой системы, использующей СММ, документ может быть представлен в форме случайного n-мерного дискретного сигнала, обладающего n признаками, образующими индекс. В свою очередь индекс (вектор-документ) может извлекаться из документа различными способами, такими как преобразование Карунена – Лоэва [7].

Полученные вектор-документы распределяются по состояниям модели, определяющим некоторые ключевые признаки поиска. Например, СММ, реализующая поиск журнальных статей, может состоять из шести суперсостояний, соответствующих параметрам статьи (Название, автор(ы), наименование журнала, год, номер, ISBN), каждое из которых делится на отдельные состояния.

В данном разрезе поиск документов основывается на выявлении схожести в многомерном пространстве. Мерой схожести является функция, определяющая значение схожести между двумя или более объектами на основе некоторых предопределенных критериев или метрик. Под метрикой понимается функция расстояния r, определенная на метрическом множестве, для любых точек a, b, c которого верна следующая система условий:

$$\begin{cases} r(a,b) = r(b,a) \\ r(a,b) = 0 \leftrightarrow a = b \\ r(a,b) + r(b,c) \geq r(a,c) \end{cases}. \quad (4)$$

Рассмотрим процесс поиска информации следующим образом.

Пусть O – объект массива, L – объект запроса, K – ключевые документы, r – метрики. На основе (1) можно утверждать истинность следующих неравенств:

$$\begin{cases} r(O,K) \leq r(O,L) + r(L,K) \\ r(L,K) \leq r(O,L) + r(L,K) \end{cases}. \quad (5)$$

Соответственно, расстояние от объекта запроса до объекта массива $r(O,L) \geq r(O,K) - r(L,K)$.

Таким образом, путем последовательного сравнения объектов массива и запроса с ключевым объектом, может быть выделена нижняя граница метрик схожести между документом и запросом, позволяющая отсечь ту часть массива документов, которая не удовлетворяет запросу.

Первое представление математической модели, для формализованного представления задачи поиска и обработки информации в рамках концепции промышленного интернета вещей – математическая модель ссылочного ранжирования [8]. Рассмотрим web-документы и ссылки на них в виде графа, в котором документы являются вершинами, а ссылки – дугами графа. Пусть G(V, E) – ориентированный граф, где V – множество вершин, а E – множество дуг. ОСИ можно представить в виде разреженной матрицы S – матрицы смежности графа G, состоящей из элементов (2.1)

$$S_{ij} = \begin{cases} 0, (i,j) \in E \\ 1, \text{ иначе} \end{cases}, i,j \in V. \quad (5)$$

В матрице могут одновременно присутствовать ссылки как вида (i, j), так и (j, i). Пусть Θ – некое множество тематик, $T(i, j, t)$ – функция веса ссылки (j, i) в тематике $t \in \Theta$ (так называемого тематического веса), $\deg(i, t)$ $\deg(i,t) = \sum_{j=0}^{N} T(i,j,t) \cdot S_{ij}$ – сумма весов всех исходящих ссылок i-й вершины в тематике t, $P = T(i,j,t)/\deg(i,t)$ – вероятность перехода пользователя по ссылке (i,j). Матрица тематических весов для ссылок (i,j) может быть записана в следующем виде:

$$\forall e \in \Theta : S(i,j) = S_{ij} \cdot P(i,j,t).$$

Таким образом, тематическое ранжирование i-й вершины может быть представлено в виде системы итерационных линейных алгебраических уравнений:

$$X_i^{k+1} = (1-d) + d \cdot \sum_{j=1}^{N} X_j^k \cdot S(i,j),$$

где d – коэффициент затухания, k – номер итерации, X_j^k – значение ранга j-й вершины в ходе k-й итерации k.

Второе математическое представление задачи поиска и обработки неструктурированной документированной информации – модель индексации документов

Математическая модель индексации может быть сформирована в виде

$$S = \{D, T, Q, R\},$$

где $D = \{d_i\}_n$ – множество документов в массиве,

$T = \{t_j\}_m$ – множество терминов, индексирующих смысловую нагрузку документов,

$Q = \{q_i\}_m$ – множество запросов пользователей либо операторов систем,

$R = \{r_i\}_n$ – множество выданных в результате поиска ссылок на документы,

n, m – соответственно количество документов в массиве и используемых терминов.

Если $W = \{w_{ij}\}_{n \times m}$ – матрица, определяющая отношения терминов к документам, где w_{ij} – вес j-го термина в i-м документе с учетом всех документов, для которых $w_{ij} \in [0;1]$, то $w_{ij} = 0$, если j-й термин не встречается в i-м документе; наоборот $w_{ij} = 1$, когда j-й термин 100 % соответствует i-му документу.

Работа информационно-поисковой системы (ИПС) может быть определена как вычисление вектора ответа $R = W \cdot Q$ путем преобразования вектора запроса Q в соответствии с матрицей W.

Для реализации преобразования необходимо произвести определение весовых коэффициентов терминов. Данное определение может быть представлено в виде следующего алгоритма:

1. Производится определение веса j-го термина для i-го документа. Обозначим её как $f_{ij} = u_{ij}^t / u_j^w$, где u_{ij}^t – общее число выявленных появлений j-го термина в i-м документе, u_i^w – общее количество слов в i-м документе. Тогда f_{ij} – коэффициент веса j-го

термина в *i*-м документе без учета остального массива (в случае, когда данный коэффициент максимален для $j \in [1, N]$, *j*-й термин определяется как отражающий содержание *i*-го документа.

2. Производится определение общего веса *j*-го термина в рамках всего массива документов. Сравнивая соотношение общего количества документов в массиве *n* и n_j^d – количество документов, в которых встречается *j*-й термин (так называемая документная частота), можно определить, является ли слово значимым термином для данного документа (чем меньше значение n_j^d, тем большим весом может обладать *j*-й термин в документе). Для нормализации возможно проведение операции натурального логарифмирования *n* и n_j^d. Таким образом, обратная документная частота определяется как $f_i^d = \ln(n) - \ln(n_j^d) = \ln\left(\frac{n}{n_j^d}\right)$ [9].

Общий вес *j*-го термина в *i*-м документе вычисляется как

$$w_{ij} = f_{ij} \cdot f_j^d = \left(u_{ij}^t / u_i^w\right) \cdot \ln\left(n / n_j^d\right).$$

Таким образом, можно говорить о сосредоточении *j*-го термина в *i*-м документе при повышении его частоты в данном документе и снижения количества документов, содержащих *j*-й термин.

На основе приведенного метода возможно построение матрицы отношений *W*, которая является основой базы индексов информационно-поисковой системы.

Третье представление – формализация обработки неструктурированных данных веб-документов.

Данная задача может быть сформулирована следующим образом: Имеется множество веб-ориентированных документов $P = \{p_1, p_2, \ldots, p_n\}$, каждый из которых определяется набором признаков (переменных) $p_j = \{x_1, x_2, \ldots, x_m, y\}$, где x_m – блок информации, содержащейся в документе, определяющей значение переменной *y* – значение интересующей пользователя информации, очищенной от шума [10].

В свою очередь каждая переменная x_m может принимать значения из некоторого множества $Z = \{z_1, z_2, \ldots\}$. Таким образом, конечное множество выделенных из неструктурированного массива релевантных данных, очищенных от шумов, над которыми возможны дальнейшие действия по структуризации, определяется как $D = \{Z \to y\}$.

Заключение

Представленные формализованные представления являются основой для дальнейшей разработки методов и алгоритмов обработки неструктурированной информации в распределенных АСУП, ориентированных на работу в условиях предприятий, реализующих концепцию промышленного Интернета вещей. На основе разработанных представлений предполагается создание и программная реализация модуля интегрированной корпоративной информационной системы для управления цифровым предприятием с распределенными информационными массивами.

По результатам предварительных прогнозов можно предполагать повышение эффективности работы программных модулей, использующих реализацию разработанных формализованных представлений, в среднем на 7–10 % по сравнению с существующими системами поиска и обработки неструктурированной информации.

Кроме того, существуют практические перспективы применения результатов разработки в области Web Mining для создания программных средств извлечения неструктурированной информации с веб-страниц.

Список литературы

1. Алан Рот. Внедрение и развитие Индустрии 4.0. Основы, моделирование и примеры из практики. М., Editorial URSS, 2017. 294 с.

2. Alalouf C. Hybrid OLAP. St. Laurent, Canada: Speedware Corporation Inc., 1997.

3. Shiryaev A.P., Dorofeev A.V., Fedorov A.R., Gagarina L.G., Zaycev V.V. LDA models for finding trends in technical knowledge domain. Proceedings of the 2017 IEEE Russia section Young researchers in electrical and electronic engineering conference (ElConRus 2017). Saint-Petersburg, 2017. P. 551–554.

4. Jones K.S. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. MCB University Press. 2004. Т. 60, № 5. P. 493–502.

5. Okapi BM25. URL: https://ru.wikipedia.org/wiki/Okapi_BM25 (дата обращения: 19.06.2018).

6. Гультяева Т.А., Попов А.А. Классификация последовательностей с использованием скрытых марковских моделей в условиях неточного задания их структуры // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 3. С. 57–63.

7. Hidden Markov Model Toolkit Book. Cambridge University Engineering Department, 2001–2009. URL: http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml (дата обращения: 20.06.2018).

8. Kapitanov A.I., Kapitanova I.I., Troyanovskiy V.M., Slyusar V.V., Fedotova E.L. Investigation of the influence of outliers on text documents probabilistic classifier quality. Proceedings of the 2017 IEEE Russia section Young researchers in electrical and electronic engineering conference (ElConRus 2017). Saint-Petersburg, 2017. P. 438–439.

9. Касумов В.А. Поисковые механизмы библиотечно-информационных систем Internet. URL: http://www.gpntb.ru/win/inter-events/crimea2000/doc/tom1/444/Doc15.HTML (дата обращения: 10.08.2018).

10. Христий С.П., Слюсарь В.В. Разработки модели решения взаимодействия информационных систем различного назначения // Инновационные подходы к решению технико-экономических проблем-2017: сборник трудов Международной конференции. М.: МИЭТ, 2017. С. 213–217.