ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION TECHNOLOGIES

УДК 519.711
https://doi.org/10.37661/1816-0301-2023-20-2-85-95

Оригинальная статья
Original Paper

Коррекция запросов в системе информационной поддержки принятия решений

С. Ф. Липницкий

Объединенный институт проблем информатики
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь
E-mail: lipn@newman.bas-net.by

Аннотация

Цели. Решается задача математического моделирования и алгоритмизации процессов коррекции запросов в системе информационной поддержки принятия решений. При этом преследуются три основные цели: построение обобщенной модели информационного поиска, разработка алгоритмов допоисковой коррекции запросов и разработка алгоритмов послепоисковой коррекции запросов.

Методы. Используются методы теории множеств и теории вероятностей.

Результаты. Разработана обобщенная модель информационного поиска. В рамках модели формализованы понятия поисковой функции, критерия выдачи, релевантности и пертинентности результатов поиска. Предложены алгоритмы допоисковой и послепоисковой коррекции запросов в системе информационной поддержки принятия решений.

Заключение. Разработанные алгоритмы коррекции запросов могут быть использованы при программной реализации информационных систем поддержки принятия решений. Универсальность алгоритмов обеспечивается благодаря применению тематических корпусов текстов в различных предметных областях. Они играют определяющую роль при планировании архитектуры информационных систем и их отдельных компонентов.

Ключевые слова: алгоритм, информационная поддержка, коррекция запроса, математическая модель, поисковая функция, принятие решений

Для цитирования. Липницкий, С. Ф. Коррекция запросов в системе информационной поддержки принятия решений / С. Ф. Липницкий // Информатика. – 2023. – Т. 20, № 2. – С. 85–95.
https://doi.org/10.37661/1816-0301-2023-20-2-85-95

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 14.02.2023
Подписана в печать | Accepted 28.02.2023
Опубликована | Published 29.06.2023

Correction of requests in the information system decision support

Stanislav F. Lipnitsky

The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus
E-mail: lipn@newman.bas-net.by

Abstract

Objectives. The problem of mathematical modeling and algorithmization of the processes of correction of requests in the system of information support for decision-making is being solved. At the same time, three main goals are pursued: building of a generalized information retrieval model, development of algorithms for pre-search query correction and development of algorithms for post-search query correction.

Methods. Methods of set theory and probability theory are used.

Results. A generalized information retrieval model has been developed. Within the framework of the model, the concepts of search function, issuance criterion, relevance and pertinence of search results are formalized. Algorithms for pre-search and post-search correction of queries in the information decision support system are proposed.

Conclusion. A mathematical model for correcting user requests in the information decision support system has been developed. Within the framework of the model, the efficiency of search processes in terms of the relevance and pertinence of the information found has been studied. Necessary and sufficient optimality of search functions are proved.

Keywords: algorithm, information support, query correction, mathematical model, search function, decision making

For citation. Lipnitsky S. F. *Correction of requests in the information system decision support*. Informatika [*Informatics*], 2023, vol. 20, no. 2, pp. 85−95 (In Russ.). https://doi.org/10.37661/1816-0301-2023-20-2-85-95

Conflict of interest. The author declares of no conflict of interest.

Введение. Эффективность информационного поиска определяется на основе оценок, принципиально различающихся между собой. В одних случаях оценивается релевантность, т. е. степень адекватности текста запросу, а в других – пертинентность, соответствующая смысловой близости текста и информационной потребности пользователя [1]. Различают два вида релевантности – смысловую и формальную. Соответствие текста содержанию запроса называют смысловой релевантностью, а соответствие поискового образа этого текста поисковому предписанию – формальной релевантностью. Факт смысловой релевантности устанавливается администратором информационной системы на основе сопоставления запросов и результатов поиска. Формальная релевантность определяется системой алгоритмически с учетом принятого в ней критерия выдачи. Запрос может значительно отличаться от информационной потребности, зафиксированной в сознании пользователя. Поэтому часто он оказывается неспособным точно выразить эту потребность. В подобных случаях необходима коррекция запросов в режиме итерационного поиска. В статье представлены алгоритмы их коррекции в системе информационной поддержки принятия решений. Алгоритмы основываются на использовании моделей представления знаний о предметной области, которые предложены автором в статьях [2, 3].

Модель поиска. Построим обобщенную модель информационного поиска с целью формализации основных его составляющих.

Критерий выдачи. Под критерием выдачи понимают правило, по которому вычисляется степень релевантности информации, найденной в процессе поиска, и принимается решение о выдаче (или невыдаче) соответствующего текста пользователю. Определим формально понятие критерия выдачи.

Пусть T и Q – некоторые множества соответственно входных и внутренних текстов системы информационной поддержки. Обозначим через Z ($Z \subseteq T$) непустое подмножество множества T, элементы которого будем называть запросами. Текст $q \in Q$ назовем поисковым образом произвольного входного текста $t \in T$, если существует такое инъективное отображение $\omega : T \to Q$, что текст q есть образ текста t при отображении ω. Если текст $z \in Z$ – запрос, то текст $q = \omega(z)$ будем называть поисковым предписанием, соответствующим запросу z. Отображение $\eta : \omega(T) \times \omega(Z) \to R$ декартова произведения множеств поисковых образов текстов и поисковых предписаний в множество R действительных чисел будем называть критерием выдачи.

Поисковая функция. Пусть $z \in Z$ – произвольный запрос. Один шаг поиска текстов промоделируем в виде частичного мультиотображения (по терминологии из монографии [4, с. 32]) $\pi : Z \to T$ множества запросов в множество текстов. Частичное мультиотображение π назовем поисковой функцией, если множество $\pi(z)$ включает те и только те тексты $t \in T$, для которых значение критерия выдачи не меньше некоторого η_0, т. е. $\eta(\omega(t), \omega(z)) \geq \eta_0$.

Релевантность и пертинентность. Определим биективное отображение $\theta : Z \to IP$, ставящее во взаимно однозначное соответствие множество запросов Z и множество IP соответствующих им информационных потребностей, т. е. выраженных вербально характеристик предметной области. Тогда понятия релевантных и пертинентных текстов можно формально ввести следующим образом. Произвольный текст $t \in T$ назовем релевантным запросу $z \in Z$, если существует сюръективное отображение $\mu : T \times Z \to \{0, 1\}$, такое, что $\mu(t, z) = 1$. Если же $\mu(t, z) = 0$, то будем считать, что текст t нерелевантен запросу z. Любой текст $t \in T$ будем называть пертинентным информационной потребности $\theta(z)$, если найдется сюръективное отображение $\nu : T \times \theta(Z) \to \{0, 1\}$, для которого справедливо соотношение $\nu(t, \theta(z)) = 1$, и непертинентным, когда $\nu(t, \theta(z)) = 0$.

Допоисковая коррекция запросов. При реализации поисковых функций будем различать два вида взаимодействия пользователей с информационной системой: допоисковое и послепоисковое. В процессе допоискового взаимодействия пользователю предъявляются сведения, которые используются при формулировании и коррекции запросов. Послепоисковое взаимодействие основано на оценке и использовании на последующих этапах поиска промежуточной выдачи найденной информации.

Пусть имеется запрос

$$z = \{a_1, a_2, \ldots \mid a_i \in W_T, i = 1, 2, \ldots\}, \quad (1)$$

где W_T – множество всех различных словоформ из множества текстов T.

Рассмотрим три варианта допоисковой коррекции запросов: путем дополнения ключевых слов словоизменениями и синонимами; путем применения корпусов текстов для вычисления информативности ключевых слов; путем разбиения исходного запроса на несколько.

Расширение запросов словоизменениями и синонимами ключевых слов. Словоизменения и синонимы ключевых слов содержат специальные лингвистические словари:

– словарь словоизменительных парадигм

$$Dic_{par} = \{(a, Par_a) \mid a \in W_T, a \in Par_a\},$$

состоящий из пар ⟨*словоформа*, *парадигма*⟩, где Par_a – совокупность всех словоизменений словоформы a;

– словарь синонимичных словоформ

$$Dic_{syn} = \{(a, Syn_a) \mid a \in W_T, a \in Syn_a\},$$

включающий в себя пары ⟨*словоформа*, *синонимичные словоформы*⟩, в которых каждой словоформе a соответствует множество ее синонимов Syn_a.

Представим запрос (1) в виде конъюнкции ключевых слов:

$$z^{(1)} = a_1 \wedge a_2 \wedge \ldots , a_i \in W_T, i = 1, 2, \ldots .$$

Дополним ключевые слова запроса $z^{(1)}$ найденными в словарях Dic_{par} и Dic_{syn} словоизменениями и синонимами. Запишем полученное выражение в виде конъюнктивной нормальной формы:

$$z^{(2)} = D_1 \wedge D_2 \wedge \ldots ,$$

где D_1, D_2 ... – ключевые слова или дизъюнкции ключевых слов.

Преобразуем запрос $z^{(2)}$ в дизъюнктивную нормальную форму:

$$z^{(3)} = K_1 \vee K_2 \vee \ldots ,$$

где K_1, K_2 ... – конъюнкции ключевых слов.

Каждый конъюнкт $K_i = b_1 \wedge b_2 \wedge \ldots$ (i = 1, 2, ...) запроса $z^{(3)}$ представим в виде совокупности $\{b_1, b_2, \ldots\}$ всех входящих в него ключевых слов. Таким образом, в результате коррекции исходного запроса (3) получим множество запросов $\{z_1, z_2, \ldots\}$.

Коррекция запросов на основе корпусов текстов. В корпусной лингвистике различают статические и динамические корпусы текстов. Примерами статических служат тематические корпусы и полный корпус, являющийся объединением всех тематических. Под динамическим понимают тематический корпус текстов, все документы которого релевантны некоторому тексту или запросу на поиск информации. Динамический корпус создается из релевантных текстов полного корпуса.

Пусть по-прежнему $z = \{a_1, a_2, \ldots | a_i \in W_T, i = 1, 2, \ldots\}$ – произвольный запрос пользователя информационной системы. Обозначим через Ct_1, Ct_2, ... тематические корпусы текстов, а через $Cf = Ct_1 \cup Ct_2 \cup \ldots$ – полный корпус. Коррекция запроса z на основе корпусов текстов сводится к соотнесению каждого ключевого слова a_i со значением его информативности $I_z^{a_i}$. Запросы являются, как правило, краткими сообщениями. Их объем не позволяет выявить статистические характеристики ключевых слов при вычислении их информативности. Поэтому в данном случае используются релевантные запросам тематические или динамические корпусы текстов.

Поиск релевантного тематического корпуса текстов. Исключим из всех поисковых образов текстов полного корпуса значения информативности ключевых слов, т. е. поисковый образ каждого текста $t \in T$ представим в виде

$$\text{ПО}_{Cf}^{t} = \{b_1, b_2, \ldots | t \in T, i = 1, 2, \ldots\}.$$

При поиске релевантного запросу текста будем использовать векторную модель описания данных [5], а в качестве критерия выдачи – косинус угла между векторами запроса и поискового образа текста [6]. Введем в рассмотрение n-мерное евклидово пространство E ($n = |W_T|$). Для этого лексикографически упорядочим все слова из множества W_T, т. е. представим его в виде кортежа $W_T = \langle c_1, c_2, \ldots, c_n \rangle$. Для каждого текста $t \in T$ построим вектор его поискового образа в пространстве E: $\mathbf{F}_t = (p_1, p_2, \ldots, p_n)$, где $p_i = 1$, если слово c_i входит в этот поисковый образ, в противном случае $p_i = 0$. Аналогично представим вектор, построенный для запроса z: $\mathbf{F}_z = (q_1, q_2, \ldots, q_n)$. Тогда для вычисления косинуса угла между векторами $\mathbf{F}_t$ и $\mathbf{F}_z$ воспользуемся формулой

$$\cos \varphi_{tz} = \frac{\mathbf{F}_t \mathbf{F}_z}{|\mathbf{F}_t||\mathbf{F}_z|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}} \sqrt{\sum_{i=1}^{n} q_i^{\,2}}}.$$

Обозначим через l количество совпавших ключевых слов поискового образа текста t и запроса z. Пусть также m_t – количество слов в поисковом образе текста t, а m_z – их количество в запросе z. Тогда данный критерий выдачи можно представить в виде

$$\cos \varphi_{tz} = \frac{l}{\sqrt{m_t m_z}}.$$

Приведем описание алгоритма поиска тематического корпуса текстов Ct, релевантного запросу z. На вход алгоритма поступает запрос z, по которому реализуется поиск поисковых образов тематических корпусов текстов в полном корпусе Cf. Результатом поиска считаем корпус Ct, которому соответствует наибольшее из значений критерия выдачи $\cos \varphi_{tz}$, такое, что $\cos \varphi_{tz} \geq \eta_0$. Если корпус Ct не найден, то для текста t необходимо сформировать динамический корпус текстов Dt.

Формирование релевантного динамического корпуса текстов. Для создания динамического корпуса текстов в множестве Cf нужно найти все тексты, релевантные запросу z. Пусть $d \in Cf$ – произвольный документ из полного корпуса текстов. Построим вектор $\mathbf{F}_d$ поискового образа документа d по аналогии с вектором $\mathbf{F}_t$: $\mathbf{F}_d = (r_1, r_2, \ldots, r_n)$. В качестве критерия выдачи будем использовать аналог критерия $\cos \varphi_{tz}$:

$$\cos \psi_{dz} = \frac{\mathbf{F}_d \mathbf{F}_z}{|\mathbf{F}_d||\mathbf{F}_z|} = \frac{\sum_{i=1}^{n} r_i q_i}{\sqrt{\sum_{i=1}^{n} r_i^{\,2}} \sqrt{\sum_{i=1}^{n} q_i^{\,2}}}.$$

Множество всех текстов из множества Cf, найденных в соответствии с критерием $\cos \psi_{dz}$, образует динамический корпус текстов Dt.

Вычисление информативности ключевых слов запроса. Информативность каждого ключевого слова a запроса z будем вычислять по формуле из статьи [2]

$$I_z^a = \frac{n_{Kt}^a + n_{Kt}^{Par_a} + n_{Kt}^{Syn_a}}{n_{Cf}^a + n_{Cf}^{Par_a} + n_{Cf}^{Syn_a}}, \qquad (2)$$

где Kt – релевантный запросу z тематический или динамический корпус текстов, n_{Kt}^a и n_{Cf}^a – частоты встречаемости слова a в корпусах текстов Kt и Cf соответственно;

$n_{Kt}^{Par_a}$ – число вхождений всех слов корпуса Kt, являющихся словоизменениями словоформы a:

$$n_{Kt}^{Par_a} = \sum_{b \in Par_a,\, b \neq a} n_{Kt}^b;$$

$n_{Kt}^{Syn_a}$ – количество синонимов словоформы a в корпусе Kt:

$$n_{Kt}^{Syn_a} = \sum_{c \in Syn_a,\, c \neq a} n_{Kt}^c.$$

Аналогичный смысл имеют параметры $n_{Cf}^{Par_a}$ и $n_{Cf}^{Syn_a}$.

В результате коррекции запроса z путем соотнесения его ключевых слов со значениями их информативности получим новый запрос

$$z' = \{(a_1,\ I_z^{a_1}), (a_2,\ I_z^{a_2}), \ldots \mid a_i \in W_{Cf},\ i = 1, 2, \ldots\}.$$

Разбиение исходных запросов. Данный вариант коррекции запросов будем использовать, если они состоят из нескольких предложений.

Классификация предложений запроса. Рассмотрим запрос $z = \langle \rho_1, \rho_2,...,\rho_l \rangle$, где $\langle \rho_1, \rho_2,...,\rho_l \rangle$ – кортеж предложений. Процессу разбиения кортежа предложений на классы предшествуют процедуры вычисления информативности вербальной ассоциации между словами предложений и между самими предложениями.

Информативность вербальной ассоциации между словами. Обозначим через W множество всех словоформ полного корпуса текстов Cf, а через $\prec_W$ – отношение строгого порядка на W (транзитивное и антирефлексивное бинарное отношение). Определим, кроме того, на множестве W антирефлексивное и антисимметричное бинарное отношение Θ, такое, что любая пара слов (a, b) из множества W является элементом отношения Θ тогда и только тогда, когда слова a и b из этой пары содержатся хотя бы в одном предложении корпуса Cf и выполняется соотношение $a \prec_W b$. Отношение Θ назовем отношением вербальной ассоциации слов в полном корпусе текстов Cf.

Информативность вербальной ассоциации между произвольными словами a и b некоторого предложения определим как вероятность его появления в корпусе Cf. При практической реализации информационной системы под указанной информативностью будем понимать дробь

$$I_{Cf}^{ab} = n_{Cf}^{ab} / N_{Cf} , \tag{3}$$

где n_{Cf}^{ab} – количество всех предложений в полном корпусе текстов Cf, в которых присутствуют слова a и b или их синонимы и словоизменения, а N_{Cf} – количество всех предложений в корпусе Cf. В развернутом виде формулу (3) можно переписать, используя информацию, которую содержат специальные лингвистические словари:

– частотный словарь словоформ

$$Dic_a = \{\langle a,\ n_{Cf}^a ,\ n_{Ct_1}^a ,\ n_{Ct_2}^a , \dots ,\ n_{Ct_n}^a \rangle |\ a \in W_{Cf}\},$$

в котором каждой словоформе приписаны частоты ее встречаемости n_{Cf}^a , $n_{Ct_1}^a$, $n_{Ct_2}^a$, … , $n_{Ct_n}^a$ во всех корпусах текстов;

– словарь вербально-ассоциативных пар слов

$$Dic_{ab} = \{\langle (a, b),\ I_{Cf}^{ab} \rangle \mid a, b \in \pi, \pi \in Cf\},$$

в котором каждой паре слов поставлена в соответствие информативность их вербальной ассоциации.

С учетом информации из лингвистических словарей формулу (3) представим в виде

$$I_{Cf}^{ab} = \frac{n_{Cf}^{ab} + n_{Cf}^{Par_{ab}} + n_{Cf}^{Syn_{ab}}}{N_{Cf}} . \tag{4}$$

Параметр $n_{Cf}^{Par_{ab}}$ в формуле (4) указывает на число вхождений всех пар словоформ, являющихся словоизменениями соответственно слов a и (или) b и встречающихся в одном и том же предложении корпуса текстов Cf:

$$n_{Cf}^{Par_{ab}} = \sum_{\substack{c \in Par_a ,\ d \in Par_b , \\ c \neq a \text{ и (или) } d \neq b, \\ (c,d) \in \Theta}} n_{Cf}^{cd} .$$

Аналогичное выражение справедливо для параметра $n_{Cf}^{Syn_{ab}}$:

$$n_{Cf}^{Syn_{ab}} = \sum_{\substack{d \in Syn_a,\ f \in Syn_b, \\ d \neq a \text{ и (или) } f \neq b, \\ (d,f) \in \Theta}} n_{Cf}^{df}.$$

Информативность вербальной ассоциации между предложениями и текстами. Пусть π и ρ – два предложения (или два текста) из корпуса *Cf*, а W_π и W_ρ – соответственно множества всех словоформ в этих предложениях, дополненные всеми синонимами и всеми словоизменениями из словарей Dic_{par} и Dic_{syn}. Построим вектор в пространстве *E*:

$$\mathbf{I}_{Cf}^{\pi\rho} = (I_{Cf}^{a_1 b_1}, I_{Cf}^{a_2 b_2}, \ldots, I_{Cf}^{a_n b_n}). \quad (5)$$

В формуле (5) значение информативности $I_{Cf}^{a_n b_n}$ для любого $i \in \{1, 2, \ldots, n\}$ определяется из словаря вербально-ассоциативных пар слов Dic_{ab}, если $(a_i, b_i) \in \Theta$ и выполняется хотя бы одно из двух условий: 1) $a_i \in W_\pi$, $b_i \in W_\rho$; 2) $b_i \in W_\pi$, $a_i \in W_\rho$. В противном случае $I_{Cf}^{a_i b_i} = 0$.

С учетом рассмотренных обозначений нормализованную информативность $I_{Cf}^{\pi\rho}$ вербальной ассоциации между предложениями (текстами) π и ρ можно интерпретировать как проекцию вектора $\mathbf{I}_{Cf}^{\pi\rho}$ на направление вектора $\mathbf{e} = (1, 1, \ldots, 1)$ размерности *n*, т. е. отношение скалярного произведения векторов $\mathbf{I}_{Cf}^{\pi\rho}$ и **e** к длине вектора **e**:

$$I_{Cf}^{\pi\rho} = \frac{\mathbf{I}_{Cf}^{\pi\rho} \cdot \mathbf{e}}{\sqrt{n}} = \frac{\sum_{i=1}^{n} I_{Cf}^{a_i b_i}}{\sqrt{n}}. \quad (6)$$

При программной реализации алгоритма вычисления информативности вербальной ассоциации между предложениями или текстами удобно пользоваться следующей формулой, полученной из выражения (6):

$$I_{Cf}^{\pi\rho} = \frac{I_1 + I_2 + \ldots + I_r}{\sqrt{n}}, \quad (7)$$

где $I_1, I_2, \ldots I_r$ – все отличные от нуля координаты вектора $\mathbf{I}_{Cf}^{\pi\rho}$.

Описание алгоритма классификации предложений. Алгоритм разбиения кортежа $z = \langle \rho_1, \rho_2, \ldots, \rho_l \rangle$ на классы работает следующим образом.

На начальном этапе в качестве единственного элемента первого класса S_1 будем рассматривать предложение ρ_1. Затем формируются множества словоформ предложений ρ_1 и ρ_2 и по формуле (7) вычисляется информативность вербальной ассоциации между ними. Если вычисленное значение не меньше некоторой пороговой величины ρ_0, то предложение ρ_2 помещается в класс S_1. Далее аналогичным образом вычисляется информативность вербальной ассоциации между предложениями из пар $(\rho_1, \rho_3), \ldots, (\rho_1, \rho_l)$. После завершения процесса формирования класса S_1 точно так же формируются и другие классы. В итоге будем иметь совокупность классов $\{S_1, S_2, \ldots, S_m\}$ $(m \leq l)$.

Индексирование информативных классов предложений. Среди сформированных классов предложений $S_1, S_2, \ldots, S_m$ могут быть неинформативные, использование которых в качестве запросов нецелесообразно. В связи с этим рассмотрим вопросы вычисления информативности классов предложений.

Информативность слов из полнотекстовых документов. Пусть T – полнотекстовый документ, объем которого обеспечивает вычисление статистических характеристик его словоформ и предложений. Информативность I_T^a слова a из текста T будем вычислять по формуле, аналогичной выражению (2):

$$I_T^a = \frac{n_T^a + n_T^{Par_a} + n_T^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \quad (8)$$

Информативность слов из полнотекстовых документов. Индексированию краткого сообщения предшествует процесс его расширения за счет включения релевантных предложений из полного корпуса текстов.

Рассмотрим краткое текстовое сообщение Q. Обозначим через W_Q множество всех его словоформ. Вычислим информативность $J_{Cf}^{Q\pi}$ вербальной ассоциации между текстом Q и некоторым предложением π из полного корпуса текстов Cf. Построим вектор $\mathbf{J}_{Cf}^{Q\pi} = (J_{Cf}^{c_1d_1}, J_{Cf}^{c_2d_2}, \ldots, J_{Cf}^{a_kb_k})$ в евклидовом пространстве. Для вычисления информативности $J_{Cf}^{Q\pi}$ воспользуемся аналогом формулы (7):

$$J_{Cf}^{Q\pi} = \frac{J_1 + J_2 + \ldots}{\sqrt{(J_1)^2 + (J_2)^2 + \ldots}}, \quad (9)$$

где J_1, J_2, ... – все отличные от нуля координаты вектора $\mathbf{J}_{Cf}^{Q\pi}$. Если информативность (9) не меньше некоторого критического значения, то предложение π занесем в текст Q. Аналогично поступим и с другими такими предложениями полного корпуса текстов. В результате получим расширенное множество предложений, которое снова будем считать текстом Q.

Информативность I_Q^a любого слова $a \in W_Q$ вычислим по формуле (8):

$$I_Q^a = \frac{n_Q^a + n_Q^{Par_a} + n_Q^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \quad (10)$$

Информативность предложений и текстов. При вычислении информативности предложений текста T будем также исходить из их векторного представления: $\mathbf{\Pi} = (I_\pi^{a_1}, I_\pi^{a_2}, \ldots, I_\pi^{a_l})$, где $I_\pi^{a_1}, I_\pi^{a_2}, \ldots, I_\pi^{a_l}$ – значения информативности слов произвольного предложения π (компонента вектора $\mathbf{\Pi}$ равна нулю, если соответствующего слова нет в предложении π). Тогда аналогично формуле (9) информативность I_T^π предложения π будем вычислять по формуле

$$I_T^\pi = \frac{I_1 + I_2 + \ldots}{\sqrt{(I_1)^2 + (I_2)^2 + \ldots}}, \quad (11)$$

где I_T^π, I_T^ρ, ... – значения информативности всех предложений документа T.

Описание алгоритма индексирования классов предложений. Алгоритм индексирования классов предложений функционирует в три этапа.

На первом этапе вычисляется информативность каждого из классов предложений S_1, S_2, ... , S_m по формуле (11). Класс будем считать информативным, если значение информативности не меньше некоторой пороговой величины. В результате выполнения первого этапа получим совокупность информативных классов предложений $\{U_1, U_2, \ldots, U_s\}$ ($s \le m$). Классы, имеющие недостаточный объем для вычисления статистических характеристик словоформ (т. е. являющиеся краткими сообщениями), дополняются релевантными предложениями из полного корпу-

са текстов *Cf* с использованием формулы (9). Полученные в результате такого расширения новые классы будем использовать в качестве запросов.

На втором этапе вычисляется информативность $I^a_{U_i}$ ($i = \overline{1, s}$) всех словоформ из предложений всех классов U_1, U_2, ... , U_s по формулам (8) и (10).

На третьем этапе формируются поисковые образы

$$\text{ПП}_i = \{(a, I^a_{U_i}); (b, I^b_{U_i}); ... \mid a, b, ... \in U_i\}, \; i = \overline{1, s},$$

всех классов U_1, U_2, ... , U_s предложений. Эти поисковые образы будут использованы в качестве поисковых предписаний, полученных после разбиения исходного запроса.

Послепоисковая коррекция запросов. Пусть по-прежнему $z \in Z$ – некоторый запрос пользователя, $\theta(z)$ – его информационная потребность, а $\pi(z)$ – совокупность найденных текстов по данному запросу.

Коррекция запросов на основе оценок пертинентности результатов поиска. Обозначим через t_z текст, полученный путем объединения предложений из всех текстов множества $\pi(z)$, которые пользователь оценил как пертинентные:

$$t_z = \{\rho \mid \rho \in t, t \in \pi(z), \nu(t, \theta(z)) = 1\}.$$

Вычислим информативность всех слов текста t_z по формуле, аналогичной выражению (2):

$$I^a_{t_z} = \frac{n^a_{t_z} + n^{Par_a}_{t_z} + n^{Syn_a}_{t_z}}{n^a_{Cf} + n^{Par_a}_{Cf} + n^{Syn_a}_{Cf}}. \qquad (12)$$

Исключим из запроса z все ключевые слова, не входящие в текст t_z, а также слова, информативность которых, вычисленная по формуле (8), меньше некоторого порогового значения. В итоге получим откорректированный запрос

$$z^+ = \{b_1, b_2, ... \mid b_i \in z \cup t_z, \; I^a_{t_z} \geq I^0, \; i = 1, 2, ...\}. \qquad (13)$$

Синтез поискового предписания из запроса (9) сводится к приписыванию каждому ключевому слову запроса z^+ его информативности:

$$z^{++} = \{(b_1, \; I^{b_1}_{t_z}), (b_2, \; I^{b_2}_{t_z}), ... \mid b_i \in z \cup t_z, \; I^a_{t_z} \geq I^0, \; i = 1, 2, ...\}.$$

Коррекция запросов на основе замены их пертинентными текстами-образцами. Пусть $z_1 = \{a_1, a_2, ... \mid a_i \in W_{Cf}, \; i = 1, 2, ...\}$ – запрос пользователя, а $Cf = Ct_1 \cup Ct_2 \cup ...$ – полный корпус текстов. Послепоисковую коррекцию запроса z_1 реализуем в три этапа.

На первом этапе проведем допоисковую коррекцию запроса z_1 на основе корпусов текстов. В результате получим запрос

$$z_2 = \{(a_1, \; I^{a_1}_{z_2}), (a_2, \; I^{a_2}_{z_2}), ... \mid a_i \in W_{Cf}, \; i = 1, 2, ...\},$$

где $I^{a_1}_{z_2}$, $I^{a_2}_{z_2}$, ... – значения информативности ключевых слов a_1, a_2, ... соответственно.

Рассмотрим реализацию второго этапа послепоисковой коррекции запроса пользователя. Пусть $Lex = \langle c_1, c_2, \ldots, c_n \rangle$ – множество всех различных слов полного корпуса текстов *Cf*, а E – n-мерное евклидово пространство. Построим в E вектор запроса z_2: $\mathbf{F}_{z_2} = (r_1, r_2, \ldots, r_n)$, где $r_i = I^{r_i}_{z_2}$, если слово r_i входит в запрос z_2, в противном случае $r_i = 0$. Аналогично для каждого текста $t \in T$ построим вектор его поискового образа в пространстве E: $\mathbf{F}_t = (s_1, s_2, \ldots, s_n)$.

На третьем этапе коррекции запроса пользователя осуществим поиск релевантных текстов в соответствии с критерием выдачи

$$\cos \varphi_{tz_2} = \frac{\mathbf{F}_t \mathbf{F}_{z_2}}{|\mathbf{F}_t||\mathbf{F}_{z_2}|} = \frac{\sum_{i=1}^{n} s_i r_i}{\sqrt{\sum_{i=1}^{n} s_i^{\,2}} \sqrt{\sum_{i=1}^{n} r_i^{\,2}}}.$$

Множество $\pi(z_2)$ текстов, найденных по запросу z_2, предъявляется пользователю, который выбирает наиболее пертинентный текст $t^+ \in \pi(z_2)$. Поисковый образ $\omega(t^+)$ текста t^+ приобретает статус откорректированного запроса:

$$z_3 = \{(c_1,\ I_{t^+}^{c_1}),\ (c_2,\ I_{t^+}^{c_2}),\ ...\}.$$

Заключение. Разработана математическая модель коррекции запросов пользователей в системе информационной поддержки принятия решений. В рамках модели предложены формулы для вычисления информативности слов, предложений и текстов, а также вербальных ассоциаций между ними. Рассмотрены два вида взаимодействия пользователей с информационной системой при коррекции запросов: допоисковое и послепоисковое. В процессе допоискового взаимодействия пользователю предъявляются сведения, которые используются при формулировании и коррекции запросов. Послепоисковое взаимодействие основано на оценке и использовании на последующих этапах поиска промежуточной выдачи найденной информации.

Разработаны алгоритмы допоисковой коррекции запросов путем дополнения ключевых слов запросов словоизменениями и синонимами, применения корпусов текстов для вычисления информативности ключевых слов и разбиения сложного исходного запроса на несколько простых.

Предложены алгоритмы послепоисковой коррекции запросов на основе оценок пертинентности результатов поиска и на основе замены исходных запросов пертинетными текстами-образцами.

Список использованных источников

1. Савотченко, С. Е. Современные аспекты повышения пертинентности результатов информационного поиска в глобальной сети [Электронный ресурс] / С. Е. Савотченко, Е. А. Проскурина. – Режим доступа: https://fundamental-research.ru/ru/article/view?id=34639. – Дата доступа: 08.02.2023.

2. Липницкий, С. Ф. Моделирование информационного поиска на основе динамических корпусов текстов / С. Ф. Липницкий, А. А. Мамчич // Вес. Нац. акад. навук Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 1. – С. 72–81.

3. Липницкий, С. Ф. Веб-поиск и адресное распространение информации на основе моделирования вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2019. – № 3. – С. 79–88.

4. Мальцев, А. И. Алгебраические системы / А. И. Мальцев. – М. : Наука, 1970. – 392 с.

5. Ландэ, Д. В. Поиск знаний в Internet. Профессиональная работа : пер. с англ. / Д. В. Ландэ. – М. : Диалектика-Вильямс, 2005. – 272 с.

6. Липницкий, С. Ф. Синтез запросов и поиск альтернатив в системе информационной поддержки принятия решений / С. Ф. Липницкий // Проблемы физики, математики и техники. – 2020. – № 2. – С. 91–95.

References

1. Savotchenko S. E., Proskurina E. A. Sovremennye aspekty povyshenija pertinentnosti rezul'tatov informacionnogo poiska v global'noj seti. *Modern Aspects of Increasing the Pertinence of Information Search Results in the Global Network* (In Russ.). Available at: https://fundamental-research.ru/ru/article/view?id=34639 (accessed at 02.08.2023).

2. Lipnitsky S. F., Mamchich A. A. *Modeling information retrieval based on dynamic text corpora*. Vestsi Natsyyanal'nai akademii navuk Belarusi. Seryya fizika-technichnych navuk [*Proceedings of the National Academy of Sciences of Belarus. Physical-technical Series*], 2011, no. 1, pp. 72–81 (In Russ.).

3. Lipnitsky S. F. *Web search and targeted dissemination of information based on the modeling of verbal associations.* Informatika [*Informatics*], 2019, no. 3, pp. 79–88 (In Russ.).

4. Maltsev A. I. Algebraicheskie sistemy. *Algebraic Systems*. Moscow, Nauka, 1970, 392 p. (In Russ.).

5. Lande D. V. Poisk znanij v Internet. Professional'naja rabota. *Knowledge Search in Internet. Professional Work*. Moscow, Dialektika-Viliams, 2005, 272 p.

6. Lipnitskiy S. F. *Synthesis of queries and search for alternatives in the system of information support for decision-making*. Problemy fiziki, matematiki i tehniki [*Problems of Physics, Mathematics and Technology*], 2020, no. 2, pp. 91–95 (In Russ.).

Информация об авторе

Липницкий Станислав Феликсович, доктор технических наук, главный научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: lipn@newman.bas-net.by

Information about the author

Stanislav F. Lipnitsky, D. Sc. (Eng.), Chief Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: lipn@newman.bas-net.by