ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION TECHNOLOGIES

УДК 004.912
https://doi.org/10.37661/1816-0301-2024-21-3-94-105

Оригинальная статья
Original Article

Обработка результатов веб-поиска в системе информационной поддержки процессов принятия решений

С. Ф. Липницкий, Л. В. Степура

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

Аннотация

Цели. Решается задача обработки результатов веб-поиска в системе информационной поддержки принятия решений с целью создания и коррекции содержательного описания проблемной ситуации. Предлагается подход к решению данной задачи на основе применения в качестве знаний о предметной области тематических корпусов текстов (совокупностей текстов по конкретной тематике), а также модели представления знаний на основе вербальных ассоциаций. При решении задач обработки результатов веб-поиска в системе информационной поддержки принятия решений преследуются пять основных целей: формирование расширенного описания проблемной ситуации, синтез поискового предписания, интернет-поиск информации о принятых решениях, синтез пересказа найденной информации, оценка качества найденных аналогов принятых решений.

Методы. Используются методы теории множеств, теории графов и математической лингвистики.

Результаты. Разработана математическая модель обработки результатов веб-поиска в системе информационной поддержки принятия решений. Формализованы понятия вербальной ассоциации слов и текстов, а также прагматически полной лексической структуры. Доказанные свойства таких структур обеспечивают алгоритмизацию информационных процессов в рассматриваемой модели.

Заключение. Подход к моделированию основывается на формализации понятий информативности слов, предложений, текстов и информативности вербальных ассоциаций между ними. В качестве реализации предложенной в статье модели разработаны алгоритмы создания словаря прагматически полных лексических структур, структурно-лексических шаблонов предложений, текстов и предметных областей, синтеза краткого пересказа найденной информации, оценки качества найденных аналогов принятых решений.

Ключевые слова: вербальные ассоциации, математическая модель, поисковое предписание, принятие решений, структурно-лексический шаблон

Для цитирования. Липницкий, С. Ф. Обработка результатов веб-поиска в системе информационной поддержки процессов принятия решений / С. Ф. Липницкий, Л. В. Степура // Информатика. – 2024. − Т. 21, № 3. – С. 94–105. https://doi.org/10.37661/1816-0301-2024-21-3-94-105

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 12.07.2024
Подписана в печать | Accepted 04.09.2024
Опубликована | Published 30.09.2024

© Липницкий С. Ф., Степура Л. В., 2024

Web search results processing in an information support system for decision-making processes

Stanislav F. Lipnitsky, Ludmila V. Stepura

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

Abstract

Objectives. The problem of the web search results processing in an information support system for decision-making is solved in order to create and correct a meaningful description of a problem situation. An approach to solving this problem is proposed based on the use of thematic text corpora (collections of texts on a specific topic) as knowledge about the subject area, as well as a knowledge representation model based on verbal associations. When solving problems of the web search results processing in a decision-making information support system, five main goals are pursued: the formation of an extended description of a problem situation, the synthesis of a search prescription, an Internet search for information about decisions made, the synthesis of a retelling of the information found, and an assessment of the quality of the found analogues of the decisions made.

Methods. Methods of set theory, graph theory and mathematical linguistics are used.

Results. A mathematical model has been developed for the web search results processing in an information support system for decision making. The concepts of verbal association of words and texts, as well as pragmatically complete lexical structure, are formalized. The proven properties of such structures provide algorithmization of information processes in the model under consideration.

Conclusion. The approach to modeling is based on the formalization of the concepts of the informativeness of words, sentences, texts and the informativeness of verbal associations between them. As an implementation of the model proposed in the article, algorithms have been developed for creating a dictionary of pragmatically complete lexical structures, creating structural-lexical templates for sentences, texts and subject areas, synthesizing a brief retelling of the information found, and assessing the quality of the found analogues of the decisions made.

Keywords: verbal associations, mathematical model, search prescription, decision making, structural-lexical template

For citation. Lipnitsky S. F., Stepura L. V. *Web search results processing in an information support system for decision-making processes*. Informatika [*Informatics*], 2024, vol. 21, no. 3, pp. 94–105 (In Russ.).
https://doi.org/10.37661/1816-0301-2024-21-3-94-105

Conflict of interest. The authors declare of no conflict of interest.

Введение. Процессы принятия решений в различных предметных областях имеют много общего и включают следующие основные этапы:

– содержательное описание проблемной ситуации. Под проблемной ситуацией понимают совокупность задач, требующих постановки и решения. Правильное ее описание является важнейшим условием принятия адекватного ситуации решения [1–5];

– сбор и анализ информации. Осуществляется поиск информации, которая находится в открытом доступе в Интернете. Проводится аналитическая обработка найденных сведений, в частности аналогов принятых решений по схожей тематике;

– коррекция описания проблемной ситуации. Корректирует описание пользователь (лицо, принимающее решение) на основе кратких пересказов найденных текстов и их оценок, генерируемых системой информационной поддержки;

– выявление вариантов (альтернатив) решения поставленной задачи. Альтернативы могут быть заданы на момент принятия решений, а могут быть неизвестны или частично известны;

– выбор критериев и оценка альтернатив. Критерии могут различаться по важности, определяемой соответствующим коэффициентом (весом);

– определение лучшей альтернативы;

– принятие решения с учетом результатов оценки альтернатив.

Реализация каждого этапа осуществляется, как правило, в итерационном режиме. В данной статье предлагается математическая модель обработки результатов веб-поиска с целью создания и коррекции содержательного описания проблемной ситуации. Формирование данного описания реализуется в режиме диалога с пользователем за пять шагов:

Шаг 1. Генерируется расширенное описание проблемной ситуации за счет пополнения исходного описания релевантными ему текстами.

Шаг 2. Формируется поисковое предписание путем индексирования исходного описания проблемной ситуации с привлечением статистических характеристик слов из расширенного описания.

Шаг 3. Осуществляется интернет-поиск информации по сформированному поисковому предписанию.

Шаг 4. Проводится аналитическая обработка результатов интернет-поиска и создаются краткие пересказы найденных текстов и их оценки.

Шаг 5. Пользователь корректирует описание проблемной ситуации.

Шаги 2–5 данного алгоритма могут повторяться по запросу пользователя.

При реализации процессов формирования и коррекции описания проблемной ситуации используются модели данных и знаний, предложенные в статье [6].

Процесс обработки результатов веб-поиска в системе информационной поддержки принятия решений состоит из пяти основных этапов, которые будут описаны в данной статье: формирования расширенного описания проблемной ситуации, синтеза поискового предписания, интернет-поиска информации о принятых решениях, синтеза пересказа найденной информации, оценки качества найденных аналогов принятых решений.

1. Формирование расширенного описания проблемной ситуации. Основой для расширения описания проблемной ситуации являются корпусы текстов. Под корпусом текстов в лингвистике понимают совокупность документов, накопленных и размеченных по определенным правилам в зависимости от назначения. В случае отсутствия разметки эти совокупности называют корпусами текстов первого порядка. Будем различать тематические и полные корпусы текстов. Тематический корпус – это множество неразмеченных текстов по некоторой конкретной тематике. Объединение всех тематических корпусов образует полный корпус текстов. Исходное описание проблемной ситуации будем дополнять текстами из полного корпуса, которые семантически связаны с этим описанием. Такие семантические связи называют вербальными ассоциациями.

Пусть Ct_i ($i=\overline{1,n}$; $n \geq 1$) – тематические корпусы текстов, а $Cf = \bigcup_{i=1}^{n} Ct_i$ – полный корпус, объединяющий все тематические. Обозначим через W множество всех словоформ корпуса Cf, а через $\prec_W$ – отношение строгого порядка на W_{Cf} (транзитивное и антирефлексивное бинарное отношение). Определим, кроме того, на множестве W антирефлексивное и антисимметричное бинарное отношение Θ, такое, что любая ориентированная 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.

1.1. Информативность вербальной ассоциации текстов. Обозначим через P описание проблемной ситуации, а через T – произвольный текст из полного корпуса текстов Cf. Рассмотрим l-мерное евклидово пространство E. Для его построения лексикографически упорядочим все пары словоформ из корпуса Cf, т. е. сформируем кортеж $\langle (a_1, b_1), (a_2, b_2), ..., (a_l, b_l)\rangle$. Пусть W_P и W_T – множества словоформ в текстах P и T соответственно, дополненные всеми синонимами и словоизменениями из следующих словарей Dic_{par} и Dic_{syn}:

$Dic_{par} = \{(a, Par_a) \mid a \in W_{Cf}, a \in Par_a\}$ – словарь словоизменительных парадигм, состоящий из пар $\langle$*словоформа, парадигма*$\rangle$. В позиции парадигмы Par_a представлены все словоизменения каждой словоформы a из множества всех словоформ W_{Cf} полного корпуса текстов;

$Dic_{syn} = \{(a, Syn_a) \mid a \in W_{Cf},\ a \in Syn_a\}$ – словарь синонимичных словоформ, включающий в себя пары ⟨*словоформа*, *синонимичные словоформы*⟩, в которых каждой словоформе a соответствует множество ее синонимов Syn_a.

Введем вектор в пространстве E:

$$\mathbf{I}^{PT} = (I^{a_1 b_1}, I^{a_2 b_2}, \ldots, I^{a_l b_l}). \quad (1)$$

Если словоформы a_i и b_i содержатся соответственно в множествах W_P и W_T, то значение информативности $I^{a_i b_i}$ согласно [7] в формуле (1) определяется из следующего словаря вербально-ассоциативных пар слов:

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

В противном случае $I^{a_i b_i} = 0$.

С учетом принятых соглашений естественно предположить, что информативность вербальной ассоциации между текстами P и T – это длина вектора (1). Для сравнения значений информативности вербальной ассоциации различных пар текстов необходима нормализация данного понятия. С целью нормализации рассмотрим вектор, все компоненты которого равны единице. Тогда нормализованную информативность I^{PT} можно интерпретировать как проекцию вектора $\mathbf{e} = (1, 1, \ldots, 1)$ размерности l на направление вектора $\mathbf{I}^{PT}$ [5], т. е. отношение скалярного произведения векторов $\mathbf{I}^{PT}$ и $\mathbf{e}$ к длине вектора $\mathbf{I}^{PT}$:

$$I^{PT} = \frac{\mathbf{I}^{PT} \cdot \mathbf{e}}{\left|\mathbf{I}^{PT}\right|} = \frac{I_1 + I_2 + \ldots + I_r}{\sqrt{(I_1)^2 + (I_2)^2 + \ldots + (I_r)^2}}, \quad (2)$$

где $I_1, I_2, \ldots I_r$ – все отличные от нуля координаты вектора I^{PT}.

Если значение информативности (2) превышает некоторое пороговое значение, то описание проблемной ситуации P расширяется путем присоединения к нему всех предложений текста T. При дальнейшем присоединении текстов описание расширяется до получения статистически значимого количества предложений в нем. Обозначим через Q расширенное описание проблемной ситуации.

1.2. Информативность вербальной ассоциации слов. Будем считать, что информативность I^{ab} вербальной ассоциации между произвольными словами a и b, так же как и в формуле (1) для словоформ, – это вероятность появления в полном корпусе текстов предложения, содержащего слова a и b. При практической реализации информационной системы под указанной информативностью будем понимать дробь [8]

$$I^{ab} = n_{Cf}^{ab} / n_{Cf}, \quad (3)$$

где n_{Cf}^{ab} – количество предложений в полном корпусе текстов Cf, в которых одновременно присутствуют слова a и b или синонимы и словоизменения хотя бы одного из этих слов, а n_{Cf} – количество всех предложений в корпусе Cf.

В развернутом виде формулу (3) перепишем, используя информацию, которую содержат словари Dic_{par} и Dic_{syn}:

$$I^{ab} = \frac{n_{Cf}^{ab} + n^{Par_{ab}} + n^{Syn_{ab}}}{n_{Cf}}. \quad (4)$$

Параметр $n^{Par_{ab}}$ 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*:

$$n^{Par_{ab}} = \sum_{c \in Par_a,\, d \in Par_b} n_{Cf}^{cd} .$$

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

$$n^{Syn_{ab}} = \sum_{d \in Syn_a,\, f \in Syn_b} n_{Cf}^{df} .$$

2. Синтез поискового предписания. Из исходного описания проблемной ситуации *P* формируется поисковое предписание в виде

$$\text{ПП}_P = \{(a, I_Q^a); (a, I_Q^b) \ldots \mid a, b \in P, I_Q^a \geq I^0, I_Q^b \geq I^0 \ldots\}, \quad (5)$$

где $I_Q^a, \ldots$ – значения информативности слов *a*, *b*… из описания *P*, I^0 – пороговое значение информативности слова в поисковом предписании, т. е. в поисковое предписание включаются все слова из текста *P*, информативность которых превышает пороговое значение I^0.

2.1. Информативность слов. Информативность I_Q^a каждого слова *a* из текста *Q* вычислим по формуле

$$I_Q^a = n_Q^a / n_{Cf}^a , \quad (6)$$

где n_Q^a и n_{Cf}^a – частоты встречаемости (с учетом словоизменения и синонимии) словоформы *a* в тексте *Q* и полном корпусе текстов *Cf* соответственно [5]. Таким образом, при вычислении информативности любого слова *a* из текста *P* в силу формулы (6) используется частота n_Q^a этого слова в расширенном описании проблемной ситуации, т. е. в тексте *Q*. При вычислении информативности I_Q^a будем использовать частотный словарь словоформ

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

в котором каждой словоформе из множества W_{Cf} всех словоформ корпуса *Cf* приписаны частоты ее встречаемости n_{Cf}^a , $n_{Ct_1}^a$, $n_{Ct_2}^a$, …, $n_{Ct_n}^a$ во всех тематических корпусах текстов Ct_i ($i = \overline{1, n}$; $n \geq 1$).

Учитывая словоизменения и синонимию, зафиксированные в лингвистических словарях Dic_{par} и Dic_{syn}, формулу (6) перепишем в виде

$$I_Q^a = \frac{n_Q^a + n_Q^{Par_a} + n_Q^{Syn_a}}{n_{Cf}^a + n_{Cf}^{Par_a} + n_{Cf}^{Syn_a}} . \quad (7)$$

Смысл параметров $n_Q^{Par_a}$, $n_{Cf}^{Par_a}$, $n_Q^{Syn_a}$ и $n_{Cf}^{Syn_a}$ аналогичен их смыслу в выражении (4).

2.2. Информативность текстов. При вычислении информативности любого текста *M*, являющегося фрагментом некоторого текста *L*, будем также исходить из его векторного представления $\mathbf{I}_L^M = (I_L^{a_1}, I_L^{a_2}, \ldots, I_L^{a_l})$, где $I_L^{a_1}, I_L^{a_2}, \ldots, I_L^{a_l}$ – значения информативности слов текста *M* (компонента вектора $\mathbf{I}_L^M$ равна нулю, если соответствующего слова нет в тексте *M*). Тогда нор-

мализованную информативность $\mathbf{I}_L^M$ текста M будем вычислять по формуле, аналогичной выражению (2):

$$I_L^M = \frac{I_1 + I_2 + ... + I_q}{\sqrt{(I_1)^2 + (I_2)^2 + ... + (I_q)^2}}, \quad (8)$$

где I_1, I_2, ... , I_q – значения информативности слов текста M.

3. Интернет-поиск информации о принятых решениях. Поиск проводится с целью выявления общих подходов к решению задачи и реализации синтеза кратких пересказов найденных текстов и их оценок.

Обозначим через R m-мерное евклидово пространство ($m = |W_{Cf}|$). Для каждой веб-страницы S построим вектор ее поискового образа в пространстве R: $\mathbf{F}_S = (I_S^{a_1}, I_S^{a_2}, ..., I_S^{a_m})$. Аналогично запишем вектор поискового предписания (5): $\mathbf{F}_{\text{ПП}_P} = (I_{\text{ПП}_P}^{b_1}, I_{\text{ПП}_P}^{b_2}, ..., I_{\text{ПП}_P}^{b_m})$. Тогда для поиска веб-страниц по поисковому предписанию $\mathbf{F}_{\text{ПП}_P}$ в качестве критерия выдачи, как показано в работе [9], используем косинус угла φ между векторами $\mathbf{F}_S$ и $\mathbf{F}_{\text{ПП}_P}$:

$$\cos\varphi = \frac{\mathbf{F}_S \cdot \mathbf{F}_{\text{ПП}_P}}{|\mathbf{F}_S| \cdot |\mathbf{F}_{\text{ПП}_P}|} = \frac{\sum_{j=1}^{m} I_S^{a_j} I_{\text{ПП}_P}^{b_j}}{\sqrt{\sum_{j=1}^{m} (I_S^{a_j})^2} \cdot \sqrt{\sum_{i=1}^{m} (I_{\text{ПП}_P}^{b_j})^2}}. \quad (9)$$

Все найденные веб-страницы упорядочиваются по убыванию значения (8).

4. Синтез пересказа найденной информации. Обычно синтез речи рассматривается как процесс последовательной генерации морфем, лексем, синтаксических фраз и, наконец, предложений. Возможен также подход к синтезу предложений, основанный на использовании языковой памяти человека [10]. Согласно этой концепции фразы при синтезе речи строятся из готовых хранящихся в памяти так называемых коммуникативных фрагментов. В случае синтеза текстовых сообщений будем использовать понятие прагматически полной лексической структуры (ПП-структуры).

4.1. Определение понятия прагматически полной лексической структуры. Пусть $\pi = a_1 a_2 ... a_n$ – произвольное предложение (или подцепочка предложения) некоторого текстового документа D. Определим формально понятие ПП-структуры:

1. Если $n = 1$, то слово a_1 цепочки π назовем ПП-структурой.

2. Если для всех m $(2 < m < n)$ справедливы неравенства $I_D^{(a_1 a_2 ... a_{m-1}) a_m} \geq I_D^{00}$ и $I_D^{(a_1 a_2 ... a_m) a_{m+1}} < I_D^{00}$, то цепочку $a_1 a_2 ... a_m$ назовем ПП-структурой. Значения информативности вычисляются по формуле (2).

3. Если для всех $n \geq 2$ выполняется неравенство $I_D^{(a_1 a_2 ... a_{n-1}) a_n} \geq I_D^{00}$, то цепочку $a_1 a_2 ... a_n$ назовем ПП-структурой.

В соответствии с данным определением работает алгоритм разбиения предложений на ПП-структуры и формирования соответствующего словаря:

Шаг 1. Проверяется условие 1. Если оно выполняется, то слово a_1 помещается в словарь ПП-структур и исключается из цепочки π. Оставшимся словам приписываются последовательно новые индексы, начиная с единицы.

Шаг 2. Проверяется условие 2. Если цепочка $a_1 a_2 ... a_m$ является ПП-структурой, то она заносится в словарь ПП-структур и исключается из цепочки π. Оставшимся словам приписываются последовательно новые индексы, начиная с единицы.

Шаг 3. Проверяется условие 3. Если цепочка $a_1 a_2 ... a_n$ является ПП-структурой, то она заносится в словарь ПП-структур.

Словарем ПП-структур будем называть множество

$$Dic_{str} = \{\langle str,\ I^{str}_{Ct_1},\ I^{str}_{Ct_2},\ \dots,\ I^{str}_{Ct_n}\rangle \mid str \in Str\},$$

где Str – совокупность ПП-структур. Значения информативности $I^{str}_{Ct_i}$ в словаре Dic_{str} вычисляются по формуле (8).

Будем различать базовые и связующие ПП-структуры. Пусть Ct – некоторый тематический корпус текстов. Рассмотрим предметную область, определяемую корпусом Ct. Обозначим через I^0_{Ct} пороговое значение информативности ПП-структуры. Тогда ПП-структуру str будем называть базовой, если значение ее информативности I^{str}_{Ct} удовлетворяет неравенству $I^{str}_{Ct} \geq I^0_{Ct}$. Если же $I^{str}_{Ct} < I^0_{Ct}$, то ПП-структуру str назовем связующей. Связующей, например, является ПП-структура «предлагается новый подход к решению проблемы», а базовой – ПП-структура «принятие решений в условиях неопределенности».

Обозначим через $Str_{\text{баз.}}$ множество всех базовых ПП-структур, а через $Str_{\text{св.}}$ – множество всех связующих. Тогда множество всех ПП-структур предметной области – это объединение множеств базовых и связующих ПП-структур, т. е. $Str_{Ct} = Str_{\text{баз.}} \cup Str_{\text{св.}}$.

4.2. Вербальная ассоциация прагматически полных лексических структур. Эффективность алгоритмов аналитической обработки текстовой информации существенным образом зависит от их интеллектуальности, т. е. способности работать не только с данными, но и знаниями об объектах и явлениях предметной области. При автоматизации процесса пересказа текста необходимые знания накапливаются в базе знаний о предметной области. Построим модель представления таких знаний.

Определим на множестве Str_{Ct} отношение толерантности Δ (рефлексивное и симметричное бинарное отношение), такое, что неупорядоченная пара (f, g) любых ПП-структур из множества Str_{Ct} является элементом отношения Δ, т. е. $(f, g) \in \Delta$ тогда и только тогда, когда ПП-структуры f и g из этой пары содержатся хотя бы в одном предложении корпуса Ct. Отношение Δ будем называть *вербально-ассоциативным отношением ПП-структур предметной области*, определяемой тематическим корпусом текстов Ct.

Вербально-ассоциативное отношение Δ – это отношение вербально-ассоциативной связи ПП-структур в тематическом корпусе текстов Ct.

Пусть f и g – произвольные ПП-структуры предметной области, определяемой тематическим корпусом текстов Ct. Для вычисления информативности вербальной ассоциации между ПП-структурами f и g будем использовать формулу (2).

4.3. Отношение конкатенации прагматически полных лексических структур. Для получения «хороших» предложений при их синтезе из ПП-структур будем использовать отношение их конкатенации («склейки», объединения). Понятие этого отношения введем следующим образом.

Определим на множестве Str_{Ct} всех ПП-структур в тематическом корпусе текстов Ct антирефлексивное бинарное отношение Λ, такое, что для любых фрагментов $f, g \in Str_{Ct}$ соотношение $(f, g) \in \Lambda$ выполняется тогда и только тогда, когда в некотором тексте $F_T \in Ct$ существует предложение π, в котором ПП-структура f непосредственно предшествует ПП-структуре g. Отношение Λ будем называть *отношением конкатенации ПП-структур* в тематическом корпусе текстов Ct.

Элементами конкатенации являются, например, следующие пары ПП-структур: «используется при» и «капитальном строительстве», «отложен ежегодный визит» и «председателя объединения».

Упорядоченные пары $(f, g) \in \Lambda$ будем хранить в специальном списке – словаре конкатенации ПП-структур:

$$Dic_{fg} = \{(f, g) \mid f, g \in Ft, (f, g) \in \Lambda\}. \qquad (10)$$

Основой для формирования словаря Dic_{fg} являются несовпадающие предложения полного корпуса текстов, представленные в виде цепочек, которые состоят из ПП-структур.

4.4. Вербально-ассоциативная сеть предметной области. Рассмотрим граф вербально-ассоциативного отношения Δ. Пометим каждую вершину f этого графа значением информативности I_{Ct}^{f} ПП-структуры (с учетом синонимии и словоизменения), а каждое ребро (f, g) – значением информативности I_{Ct}^{fg} вербальной ассоциации ПП-структур f и g (также учитывая синонимию и словоизменения). Пусть $(f, g) \in \Lambda$, то для всех таких пар (f, g) вершины f и g соединим дугой, направленной от f к g. Обозначим полученный смешанный граф через Net_{Ct}.

Граф Net_{Ct} назовем *вербально-ассоциативной сетью предметной области*, определяемой тематическим корпусом текстов Ct. Сеть Net_{Ct} является моделью поискового образа тематического корпуса текстов Ct. Построим этот поисковый образ в виде множества ПП-структур и вербально-ассоциативных пар таких структур, в котором им приписаны значения информативности, а парам – значения информативности вербальной ассоциации:

$$\text{ПО}_{Ct} = \{(f, I_{Ct}^{f}), \ldots, ((g, h), I_{Ct}^{gh}), \ldots, ((\overrightarrow{p, q}), I_{Ct}^{pq}), \ldots \mid I_{Ct}^{f} > I_{Ct}^{0}; I_{Ct}^{gh}, I_{Ct}^{pq} > I_{Ct}^{00}\}, \qquad (11)$$

где I_{Ct}^{0}, I_{Ct}^{00} – пороговые значения информативности ПП-структуры и вербальной ассоциации ПП-структур соответственно, стрелка над парой ПП-структур (p, q) означает, что $(p, q) \in \Lambda$.

4.5. Вербально-ассоциативная сеть пересказываемого текста. Пусть $Q \in Cf$ – некоторый текст, а Ct – релевантный ему тематический корпус текстов. Вычислим информативность I_{Q}^{π} каждого предложения π текста Q по формуле (9).

Обозначим через I_{Q}^{0} некоторое пороговое значение информативности предложений текста Q. Если $I_{Q}^{\pi} \geq I_{Q}^{0}$, то предложение π будем считать информативным. Исключив из текста все неинформативные предложения, т. е. такие, для которых $I_{Q}^{\pi} < I_{Q}^{0}$, получим кортеж информативных предложений $T = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$. Каждое предложение π_i ($i = \overline{1, m}$) представим в виде цепочки ПП-структур $\pi_i = f_1 f_2 \ldots$. Обозначим через Δ_T сужение вербально-ассоциативного отношения Δ на множестве F_T всех ПП-структур кортежа предложений T, т. е. $\Delta_T = \Delta \cap (F_T \times F_T)$, а через Λ_T – сужение отношения конкатенации ПП-структур Λ на этом же множестве. Тогда вербально-ассоциативную сеть пересказываемого текста Q построим следующим образом.

Рассмотрим вербально-ассоциативную сеть предметной области Net. Исключим из сети Net все ребра (f, g), такие, что $(f, g) \notin \Delta_T$, и все дуги (r, s), для которых $(r, s) \notin \Lambda_T$. Исключим также из сети Net инцидентные исключенным ребрам и дугам вершины. Полученный граф назовем *вербально-ассоциативной сетью пересказываемого текста Q*.

Для моделирования синтеза пересказа текстов введем понятия структурно-лексических шаблонов предложения, текста и предметной области.

4.6. Структурно-лексический шаблон предложения. Пусть имеется предложение $\pi = f_1 f_2 \ldots f_l$. Цепочку, полученную из предложения π заменой его базовых коммуникативных фрагментов слотами («пустыми» фрагментами), будем называть структурно-лексическим шаблоном предложения π.

Структурно-лексические шаблоны предложений создаются в автоматизированном режиме: сначала на основе специально подготовленных текстов программно формируется совокупность структурно-лексических шаблонов, а затем они корректируются экспертом-лингвистом информационной системы. При синтезе предложения слоты заменяются ПП-структурами.

4.7. Структурно-лексический шаблон текста. Пусть имеется текст в виде кортежа предложений $T = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$, такой, что каждому предложению π_i ($i = \overline{1, m}$) соответствует его структурно-лексический шаблон H_i. Рассмотрим кортеж структурно-лексических шаблонов

предложений $SH_T = \langle H_1, H_2, \dots, H_m \rangle$. В качестве характеристики связности структурно-лексических шаблонов предложений текста T определим на множестве SH_T антирефлексивное бинарное отношение Ω_T, элементами которого являются пары соседних структурно-лексических шаблонов предложений из множества SH_T, т. е. $\Omega_T = \{(H_i, H_{i+1}) | i = \overline{1, m-1}\}$. Отношение Ω_T назовем структурно-лексическим шаблоном текста T.

4.8. Структурно-лексический шаблон предметной области. Для формирования структурно-лексического шаблона предметной области необходимо предварительно подготовить множество $\{T_i | i = \overline{1, r}\}$ некоторых «хороших» текстов. Обозначим через Ω_{T_i} структурно-лексический шаблон текста T_i. Тогда объединение множеств $\Omega_{Ct} = \bigcup_{i=1}^{r} \Omega_{T_i}$ назовем структурно-лексическим шаблоном предметной области, определяемой тематическим корпусом текстов Ct.

4.9. Алгоритм синтеза краткого пересказа найденной информации. Пусть $T \in Cf$ – некоторый текст, Ct – релевантный ему тематический корпус текстов, а $Q = \langle \pi_1, \pi_2, \dots, \pi_m \rangle$ – кортеж информативных предложений этого текста. Алгоритм синтеза пересказа содержания текста T работает следующим образом:

Шаг 1. Из вербально-ассоциативной сети предметной области Net_{Ct} исключаются все вершины, которые соответствуют базовым ПП-структурам, отсутствующим в предложениях кортежа Q. Из сети Net_{Ct} удаляются инцидентные исключенным вершинам ребра и дуги. Полученный граф обозначим через Net_T^+.

Шаг 2. Из структурно-лексического шаблона предметной области Ω_{Ct} формируется множество начальных структурно-лексических шаблонов предложений $Form_1 = \{H_1, H_2, \dots\}$. Для каждого шаблона из множества $Form_1$ строится его вербально-ассоциативная сеть Net_{H_1}. Формируется множество всех орцепей $Req_{H_1}^{1,2}$ длиной 1 и 2 графа Net_{H_1}. Элементами множества $Req_{H_1}^{1,2}$ являются орцепи вида $f_1 s_1$, $s_2 g_2$, $f_3 s_3 g_3$, где $f_1 s_1$ – конечные фрагменты предложения, а $s_2 g_2$ – начальные; s_1, s_2 и s_3 – слоты; f_3 и g_3 – ПП-структуры, не являющиеся начальными и конечными в структурно-лексическом шаблоне H_1. Орцепи из множества $Req_{H_1}^{1,2}$ служат запросами на поиск релевантных орцепей в графе Net_T^+ с целью заполнения найденных слотов ПП-структурами. Орцепи графов Net_T^+ и Net_{H_1} считаются совпавшими, если совпали соответствующие ПП-структуры. Например, совпавшими являются орцепи fsg и frg, где r – ПП-структура, заполняющая слот s. Если все слоты структурно-лексического шаблона заполнены, то полученное в результате предложение включается в некоторое множество Sen. Далее описанная процедура повторяется для всех остальных шаблонов из множества $Form_1$.

Шаг 3. В сформированном множестве Sen проводится поиск релевантного графу Net_T^+ предложения π_l, полученного из шаблона H_l. Оно является началом формируемого пересказа текста.

Шаг 4. Создается множество $Form_2$, состоящее из всех структурно-лексических шаблонов предложений H, таких, что $(H_l, H) \in \Omega_T$. Процессы заполнения слотами шаблонов из множеств $Form_2$, $Form_3$ и т. д. аналогичны такой процедуре для шаблонов из множества $Form_1$.

Пример

Принимается решение о включении некоторой статьи в отраслевую энциклопедию. На этапе поиска информации о принятых решениях найден приведенный ниже текстовый фрагмент. Требуется построить краткий его пересказ.

В ходе интерпретации воссоздается мысленный мир, в котором по презумпции интерпретатора автор конструировал дискурс и в котором описывается реальное или нереальное положение дел. При этом анализ дискурса предполагает наличие языкового инструментария, при котором исследователь обращается не только к собственным лингвистическим знаниям, но и к общему фоновому знанию

о реальном мире, поскольку в процессах понимания и порождения речи взаимодействуют все базы данных, хранящиеся в когнитивном аппарате человека. В основном анализу подвергаются не отдельные слова, а более крупные объединения (предложения или даже целые тексты), так как известно, что трансляция смысла ведется с помощью именно текстов. Поэтому текст стал объектом исследования отдельного направления языкознания, лингвистики текста, которое стремится выйти за рамки предложения. Дискурс может разделиться на высказывания, в то время как существуют другие объединения, которые складываются из последовательных предложений, например текст.

На начальном этапе генерации пересказа данного текста формируется множество структурно-лексических шаблонов предложений $Form_1$: $Form_1$ = {⟨обсуждается проблема / ◊ ⟩, ⟨рассматриваются вопросы / ◊ / путем использования ◊ ⟩, ⟨речь идет о / ◊ ⟩, в работе / приведены результаты / ⟩,...}, где символом «/» обозначены разделители между коммуникативными фрагментами, а символом «◊» – слоты.

После построения для всех структурно-лексических шаблонов их вербально-ассоциативных сетей и поиска ПП-структур для заполнения слотов выбранного шаблона получим начальное предложение пересказа исходного текста в виде цепочки ПП-структур: *Рассматриваются вопросы / конструирования дискурса / путем использования / лингвистических знаний / о реальном мире.*

На последующих этапах синтеза выходного текста процесс поиска структурно-лексических шаблонов предложений и заполнения их слотов повторяется аналогичным образом. В результате получим следующие предложения сформированного краткого пересказа текста: *Для изучения дискурса / возникло направление / в языкознании / – / лингвистика текста. При конструировании / анализируются фрагменты / данного дискурса. Элементами дискурса / являются высказывания.*

5. Оценка качества найденных аналогов принятых решений. Для оценки качества принятых решений будем использовать совокупность оценочных тематических корпусов текстов. Каждому корпусу соответствует некоторая оценка качества. При n-балльной шкале оценок количество таких корпусов должно быть равно n. Всякий корпус включает текстовые документы одинакового оценочного качества. В простейшем случае формируются два корпуса текстов с соответствующей оценочной лексикой. Первый корпус создается для анализа положительного качества, а второй – для анализа отрицательного. Оценка качества принятых решений реализуется путем поиска наиболее релевантных им оценочных корпусов текстов.

Пусть Ct – тематический корпус текстов с оценочной лексикой, состоящий из n (по числу оценок в шкале оценивания) подкорпусов, т. е. $Ct = \{Cp_i \mid i = \overline{1, n}\}$. Каждый подкорпус Cp_i состоит из текстов с одинаковой оценкой качества и представляет собой пару $\langle Cp_i, Ev_i \rangle$ (Ev_i – оценка качества для всех текстов из множества Cp_i). Пусть также T – текстовое сообщение, полученное в результате формирования контекстного окружения к некоторому найденному на веб-странице $S_{\text{веб}}$ информативному предложению. Построим вектор $\mathbf{F}_T$ сообщения T и векторы $\mathbf{F}_{Cp_i}$ ($i = \overline{1, n}$). Для каждой пары ($\mathbf{F}_T$, $\mathbf{F}_{Cp_i}$) вычислим косинус угла между этими векторами по формуле (7). Тогда сообщению T будет соответствовать оценка качества Ev_i при таком значении i, при котором $\cos(\mathbf{F}_T, \mathbf{F}_{Cp_i})$ принимает наибольшее значение.

Заключение. Предложена математическая модель аналитической обработки результатов веб-поиска в системе информационной поддержки процессов принятия решений на этапе текстового описания проблемной ситуации. Подход к моделированию основывается на формализации понятий информативности слов, предложений и текстов и информативности вербальных ассоциаций между ними. Формализация этих понятий обеспечила реализацию следующих алгоритмов:

– создания словаря прагматически полных лексических структур;

– создания структурно-лексических шаблонов предложений, текстов и предметных областей;

– синтеза краткого пересказа найденной информации;

– оценки качества найденных аналогов принятых решений.

Словарь прагматически полных лексических структур создается в три шага в соответствии с их формальным определением. На каждом шаге последовательно проверяются условия из данного определения.

Фрагментно-слотовые шаблоны предложений формируются путем замены их базовых прагматически полных лексических структур слотами, а шаблоны текстов – как кортежи шаблонов их предложений.

Структурно-лексические шаблоны предметных областей создаются в виде реализации бинарных отношений на множествах шаблонов предложений из соответствующих тематических корпусов текстов. Каждый тематический корпус текстов определяет некоторую предметную область.

Вклад авторов. *С. Ф. Липницкий* предложил формальную модель обработки результатов вебпоиска в системе информационной поддержки принятия решений, *Л. В. Степура* разработала алгоритмы обработки текстовой информации и представила результаты исследования. Оба автора принимали участие в подготовке текста статьи.

Список использованных источников

1. Кравченко, Т. К. Системы поддержки принятия решений / Т. К. Кравченко // Информационные технологии для современного университета / под общ. ред.: А. Н. Тихонов, А. Д. Иванников. – М. : ГНИИ ИТТ «Информика», 2011. – С. 107–118.

2. Моисеенко, Е. В. Информационные технологии в экономике / Е. В. Моисеенко, Е. Г. Лаврушина ; ред.: М. А. Касаткина – М. : Софт, 2009. – С. 120–135.

3. Симанков, В. С. Методологическое обеспечение этапов поддержки принятия решений при синтезе сложных систем / В. С. Симанков, А. Н. Черкасов // Перспективы науки. – 2012. – № 12. – С. 85–89.

4. Липницкий, С. Ф. Описание проблемной ситуации в процессе информационной поддержки принятия решений / С. Ф. Липницкий, Л. В. Степура // Докл. XXI Междунар. конф. «Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2022)», Минск, 18 нояб. 2022 г. – Минск : ОИПИ НАН Беларуси, 2022. – С. 108–112.

5. Липницкий, С. Ф. Поиск и лексико-семантическая обработка научно-технической информации / С. Ф. Липницкий, Л. В. Степура // Докл. XXII Междунар. конф. «Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2023)», Минск, 16 нояб. 2023 г. – Минск : ОИПИ НАН Беларуси, 2023. – С. 181–185.

6. Липницкий, С. Ф. Модель представления знаний в информационных системах на основе вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2011. – № 4. – С. 21–28.

7. Буравкин, А. Г. Интернет-поиск альтернативных вариантов в процессе принятия решений / А. Г. Буравкин, С. Ф. Липницкий, Л. В. Степура // Докл. XX Междунар. конф. «Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2021)», Минск, 18 нояб. 2021 г. – Минск : ОИПИ НАН Беларуси, 2021. – С. 180–183.

8. 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

9. 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.

10. Гаспаров, Б. М. Язык, память, образ. Лингвистика языкового существования / Б. М. Гаспаров. – М. : Новое литературное обозрение, 1996. – 352 с.

References

1. Kravchenko T. K. *Decision support systems*. Informacionnye tehnologii dlja sovremennogo universiteta [*Information Technologies for a Modern University*]. In A. N. Tikhonov, A. D. Ivannikov (eds.). Moscow, GNII ITT "Informika", 2011, pp. 107–118 (In Russ.).

2. Moiseenko E. V., Lavrushina E. G. Informacionnye tehnologii v jekonomike. *Information Technologies in the Economy*. In M. A. Kasatkina (ed.). Moscow, Soft, 2009, pp. 120–135 (In Russ.).

3. Simankov V. S., Cherkasov A. N. Methodological support of decision support stages in the synthesis of complex systems. Perspektivy nauki [*Prospects of Science*], 2012, no. 12, pp. 85–89 (In Russ.).

4. Lipnitsky S. F., Stepura L. V. *Description of a problem situation in the process of information support for decision making*. Doklady XXI Mezhdunarodnoj konferencii «Razvitie informatizacii i gosudarstvennoj sistemy nauchno-tehnicheskoj informacii (RINTI-2022)», Minsk, 18 nojabrja 2022 g. [*Reports of the XXI International Conference "Development of Informatization and the State System of Scientific and Technical Information (RINTI-2022)", Minsk, 18 November 2022*]. Minsk, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 2022, pp. 108–112 (In Russ.).

5. Lipnitsky S. F., Stepura L. V. *Search and lexical-semantic processing of scientific and technical information*. Doklady XXII Mezhdunarodnoj konferencii «Razvitie informatizacii i gosudarstvennoj sistemy nauchno-tehnicheskoj informacii (RINTI-2023)», Minsk, 16 nojabrja 2023 g. [*Reports of the XXII International Conference "Development of Informatization and the State System of Scientific and Technical Information (RINTI-2023)", Minsk, 16 November 2023*]. Minsk, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 2023, pp. 181–185 (In Russ.).

6. Lipnitsky S. F. *Model of knowledge representation in information systems based on verbal associations*. Informatika [*Informatics*], 2011, no. 4, pp. 21–28 (In Russ.).

7. Buravkin A. G., Lipnitsky S. F., Stepura L. V. *Internet search for alternative options in the decision-making process*. Doklady XX Mezhdunarodnoj konferencii «Razvitie informatizacii i gosudarstvennoj sistemy nauchno-tehnicheskoj informacii (RINTI-2021)», Minsk, 18 nojabrja 2021 g. [*Reports of the XX International Conference "Development of Informatization and the State System of Scientific and Technical Information (RINTI-2021)", Minsk, 18 November 2021*]. Minsk, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 2021, pp. 180–183 (In Russ.).

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

9. Lipnitsky S. F., Mamchich A. A. *Modeling information retrieval based on dynamic text corpora*. Vesci Nacyjanal'naj akadjemii navuk Belarusi. Seryja fizika-tjehnichnyh navuk [*Proceedings of the National Academy of Sciences of Belarus. Physical-technical Series*], 2011, no. 1, pp. 72–81 (In Russ.).

10. Gasparov B. M. Jazyk, pamjat', obraz. Lingvistika jazykovogo sushhestvovanija. *Language, Memory, Image. Linguistics of Linguistic Existence*. Moscow, Novoe literaturnoe obozrenie, 1996, 352 p. (In Russ.).

Информация об авторах

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

Степура Людмила Васильевна, научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: stepura@newman.bas-net.by

Information about the authors

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

Ludmila V. Stepura, Researcher Scientist, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: stepura@newman.bas-net.by