ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

УДК 004.912
https://doi.org/10.37661/1816-0301-2020-17-4-73-82

Поступила в редакцию 28.10.2020
Received 28.10.2020

Принята к публикации 20.11.2020
Accepted 20.11.2020

Интернет-поиск и лексико-семантическая обработка аналогов принятых решений в различных предметных областях

С. Ф. Липницкий

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

Аннотация. Предлагается математическая модель интернет-поиска и лексико-семантической обработки аналогов принятых решений в различных предметных областях. Поиск проводится по запросам, синтезированным из описаний проблемных ситуаций. Процесс синтеза осуществляется в два этапа. На первом этапе предложения из текста, содержащего описание проблемной ситуации, классифицируются с учетом информативности вербальной ассоциации между ними. На втором этапе вычисляется информативность каждого класса. Наиболее информативные из сформированных классов после их индексирования используются в качестве запросов. При лексико-семантической обработке найденные аналоги решений исследуются на тональность. Оценка тональности реализуется путем использования лингвистических словарей тонально окрашенной лексики, которые формируются на основе специальных тонально окрашенных тематических корпусов текстов. В предельном случае создаются два типа словарей. Первый тип предназначен для анализа положительной тональности в описаниях принятых решений, а второй – для анализа отрицательной тональности, т. е. в процессе лексико-семантического анализа рассматриваются главным образом положительные и отрицательные аспекты принятых решений. Результаты анализа предъявляются пользователю (лицу, принимающему решения).

Ключевые слова: интернет-поиск, лингвистические словари, математическая модель, проблемная ситуация, релевантность, синтез запросов, тональная окрашенность

Для цитирования. Липницкий, С. Ф. Интернет-поиск и лексико-семантическая обработка аналогов принятых решений в различных предметных областях / С. Ф. Липницкий // Информатика. – 2020. − Т. 17, № 4. – С. 73–82. https://doi.org/ 10.37661/1816-0301-2020-17-4-73-82

Internet search and lexical-semantic processing of analogs when making decisions in various subject areas

Stanislav F. Lipnitsky

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

Abstract. A mathematical model of Internet search and lexical-semantic processing of analogs of the decisions made in various subject areas is proposed. The search is carried out on queries synthesized from descriptions of problem situations. The synthesis process is carried out in two stages. At the first stage, sentences from the text containing a description of the problem situation are classified taking into account the informativeness of the verbal association between them. At the second stage, the information content of each class is calculated. The most informative of the generated classes, after their indexing, are used as queries. In lexical-semantic processing, the found analogs of solutions are examined for sentiment. When assessing sentiment, linguistic dictionaries of tone-colored vocabulary are used, which are formed on the basis of special tone-colored thematic corpora of texts. In the extreme case, two types of dictionaries are created. The first type is intended for the analysis of the positive sentiment in the descriptions of the decisions made, and the second is intended for the analysis of

the negative sentiment, that is, in the process of lexical-semantic analysis, mainly positive and negative aspects of the decisions made are considered. The results of the analysis are presented to the user (decision-maker).

Keywords: internet search, linguistic dictionaries, mathematical model, problem situation, relevance, query synthesis, tonal coloration

For citation. Lipnitsky S. F. Internet search and lexical-semantic processing of analogs when making decisions in various subject areas. *Informatics*, 2020, vol. 17, no. 4, pp. 73–82 (in Russian). https://doi.org/10.37661/1816-0301-2020-17-4-73-82

Введение. Процессы принятия решений в различных предметных областях имеют много общего. Как правило, они включают в себя следующие основные этапы:

– описание проблемной ситуации и постановку задачи принятия решения;
– поиск вариантов (альтернатив) решения поставленной задачи;
– выбор критериев оценки альтернатив для описания вариантов решения;
– выявление ограничений на критерии;
– принятие решения с учетом результатов оценки альтернатив.

При вербальном анализе решений используется качественная (нечисловая) информация на всех его этапах [1].

В статье [2] автором предложена математическая модель информационной поддержки процесса принятия решения в части веб-поиска его альтернативных вариантов. Поиск альтернатив реализуется по запросам, синтезированным из описания проблемной ситуации.

В настоящей статье предлагаются модель и алгоритмы лексико-семантической обработки аналогов уже принятых решений, найденных в Интернете по запросам, которые были синтезированы из описаний проблемных ситуаций в соответствии с алгоритмами из работы автора [2]. В процессе лексико-семантического анализа рассматриваются главным образом положительные и отрицательные аспекты принятых решений, т. е. оценивается тональность при их обсуждении. Результаты анализа предъявляются пользователю (лицу, принимающему решения).

Архитектура информационной системы. Функциональными компонентами системы интернет-поиска и лексико-семантической обработки аналогов принятых решений являются четыре подсистемы (рис. 1): синтеза запросов, индексирования запросов, поиска аналогов решений в Интернете, лексико-семантической обработки аналогов.

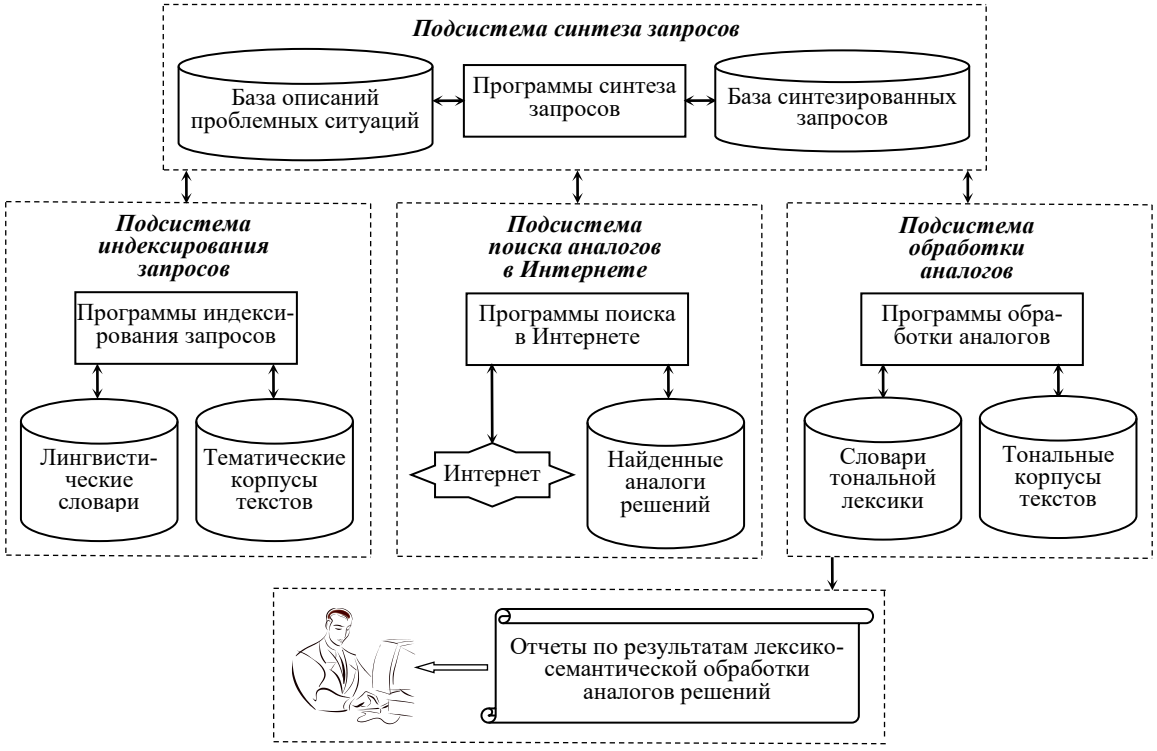

Рис. 1. Структурная схема системы интернет-поиска и лексико-семантической обработки аналогов принятых решений

Подсистема синтеза запросов включает базу описаний проблемных ситуаций при принятии решений, базу синтезированных запросов из этих описаний, а также соответствующие программные средства. Подсистема обеспечивает формирование запросов на основе исследования вербальных ассоциаций между предложениями в описании проблемной ситуации [3, 4]. Предложения из текста описания классифицируются с учетом информативности ассоциации между ними. Далее вычисляется информативность каждого класса. Информативные классы после их индексирования используются в качестве запросов на интернет-поиск аналогов принятых решений.

В состав подсистемы индексирования входит совокупность лингвистических словарей для вычисления информативности слов и вербальных ассоциаций между ними. Словари формируются из специальных наборов публикаций по каждой предметной области – тематических корпусов текстов. Поисковые образы классов предложений представляются в виде множеств слов и вербально-ассоциативных пар слов с соответствующими значениями информативности.

Подсистема поиска аналогов решений в Интернете состоит из специализированных программных агентов, основная задача которых заключается в систематическом получении и накоплении новых данных из Интернета. Поиск аналогов решений реализуется в порядке, определяемом упорядочивающим отношением, которое задается на множестве веб-страниц каждого сканируемого веб-сайта.

Подсистема лексико-семантической обработки аналогов принятых решений имеет в своем составе лингвистические словари тонально окрашенной лексики, а также тонально окрашенные тематические корпусы текстов. Они используются при формировании оценок тональности сообщений. В предельном случае создаются два типа словарей. Первый тип предназначен для анализа положительной тональности в описаниях принятых решений, а второй – для анализа отрицательной тональности.

Синтез запросов. Процесс синтеза запросов в системе интернет-поиска и лексико-семантической обработки аналогов принятых решений реализуется в два этапа. На первом этапе предложения из текста, содержащего описание проблемной ситуации, классифицируются с учетом информативности вербальной ассоциации между ними. На втором этапе вычисляется информативность каждого класса. Информативные классы после их индексирования используются в качестве запросов на интернет-поиск принятых решений.

Обозначим через $T = \langle \rho_1, \rho_2, ..., \rho_l \rangle$ описание проблемной ситуации, где $\langle \rho_1, \rho_2, ..., \rho_l \rangle$ – кортеж предложений текста T. При разбиении данного кортежа предложений на классы будем использовать формулы вычисления информативности вербальной ассоциации между словами предложений и между самими предложениями [2].

Информативность вербальной ассоциации между словами. Под вербальными ассоциациями в компьютерной лингвистике понимают семантические связи между словами в языке, соответствующие ассоциативным отношениям между обозначаемыми ими сущностями в реальном мире. Различают два типа таких ассоциаций: парадигматические и синтагматические [5, 6]. Парадигматические ассоциации существуют между словами языка независимо от контекста и объединяют понятия, обозначающие предметы или явления, между которыми имеется постоянная связь (например, пары слов *книга – знание*, *человек – дом*). В противоположность парадигматическим синтагматические ассоциации возникают в тексте, т. е. между словами и словосочетаниями каждого конкретного его предложения (например, в парах слов *технология – информационная*, *текст – шрифт*). Считаем, что информативность вербальной ассоциации между произвольными словами a и b некоторого предложения – это вероятность его появления в полном корпусе текстов. При практической реализации информационной системы под указанной информативностью будем понимать дробь

$$I_{Cf}^{ab} = n_{Cf}^{ab} / N_{Cf}, \qquad (1)$$

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 (1) можно переписать, используя информацию, которую содержат специальные лингвистические словари:

– 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;

– словарь словоизменительных парадигм $Dic_{par} = \{(a, Par_a) \mid a \in W_{Cf},\ a \in Par_a\}$, состоящий из пар ⟨*словоформа*, *парадигма*⟩. В позиции парадигмы Par_a представлены все словоизменения словоформы *a*;

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

С учетом словоизменений и синонимов, зафиксированных в словарях Dic_{par} и Dic_{syn}, формулу (1) перепишем в виде

$$I_{Cf}^{ab} = \frac{n_{Cf}^{ab} + n_{Cf}^{Par_{ab}} + n_{Cf}^{Syn_{ab}}}{N_{Cf}}. \tag{2}$$

Параметр $n_{Cf}^{Par_{ab}}$ в формуле (2) 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. Рассмотрим *l*-мерное евклидово пространство *E*. Для его построения лексикографически упорядочим все пары словоформ из полного корпуса текстов *Cf*, т. е. сформируем кортеж $\Theta = \langle (a_1, b_1), (a_2, b_2), ..., (a_l, b_l) \rangle$.

Пусть π и ρ – два предложения (текста) из корпуса *Cf*, а W_π и W_ρ – соответственно множества словоформ в этих предложениях, дополненные всеми синонимами и словоизменениями из словарей Dic_{par} и Dic_{syn}.

Построим вектор в пространстве *E*:

$$\mathbf{I}_{Cf}^{\pi\rho} = (I_{Cf}^{a_1 b_1}, I_{Cf}^{a_2 b_2}, \ldots, I_{Cf}^{a_l b_l}). \tag{3}$$

Если словоформы a_i и b_i содержатся соответственно в множествах W_π и W_ρ (или W_ρ и W_π), то значение информативности $I_{Cf}^{a_i b_i}$ в формуле (3) определяется из словаря вербально-ассоциативных пар слов Dic_{ab}. В противном случае $I_{Cf}^{a_i b_i} = 0$.

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

$$I_{Cf}^{\pi\rho} = \frac{\mathbf{I}_{Cf}^{\pi\rho} \cdot \mathbf{e}}{|\mathbf{I}_{Cf}^{\pi\rho}|} = \frac{\sum_{i=1}^{l} I_{Cf}^{a_i b_i}}{\sqrt{\sum_{i=1}^{l} (I_{Cf}^{a_i b_i})^2}}. \tag{4}$$

При реализации алгоритма вычисления информативности вербальной ассоциации между предложениями или текстами удобно пользоваться следующей формулой, полученной из выражения (4):

$$I_{Cf}^{\pi\rho} = \frac{I_1 + I_2 + ... + I_r}{\sqrt{(I_1)^2 + (I_2)^2 + ... + (I_r)^2}}, \quad (5)$$

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

Описание алгоритма классификации предложений из текста описания проблемной ситуации. Алгоритм классификации предложений текста T работает следующим образом.

Обозначим через S_1 первый класс предложений. В качестве единственного элемента первого класса S_1 будем рассматривать предложение ρ_1. Затем вычислим информативнось вербальной ассоциации между предложениями ρ_1 и ρ_2 по формуле (5). Если вычисленное значение не меньше некоторой пороговой величины ρ_0, то предложение ρ_2 поместим в класс S_1. Далее аналогичным образом вычислим информативность вербальной ассоциации между предложениями из пар (ρ_1, ρ_3), ..., (ρ_1, ρ_l). После завершения процесса формирования класса S_1 точно так же формируются и другие классы. В итоге будем иметь совокупность классов $\{S_1, S_2, ..., S_m\}$ $(m \le l)$. Информативные классы будем использовать в качестве запросов на поиск аналогов принятых решений.

Индексирование запросов. Запросы являются, как правило, краткими сообщениями. Их объем не позволяет выявить статистические характеристики словоформ. Поэтому индексированию запроса предшествует процесс его расширения за счет включения релевантных предложений из полного корпуса текстов.

Информативность слов. Пусть T – текстовый документ, объем которого обеспечивает вычисление статистических характеристик его словоформ и предложений. Информативность I_T^a слова a из текста T вычислим по формуле

$$I_T^a = n_T^a / n_{Cf}^a, \quad (6)$$

где n_T^a и n_{Cf}^a – частоты встречаемости (с учетом словоизменения и синонимии) словоформы a в тексте T и полном корпусе текстов Cf соответственно [7]. При вычислении будем использовать 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 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 ($i = \overline{1, n}$; $N \ge 1$).

Используя лингвистические словари Dic_{par} и Dic_{syn}, формулу (6) перепишем в виде

$$I_T^a = \frac{n_T^a + n_T^{Par_a} + n_T^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \quad (7)$$

В выражении (7) $n_T^{Par_a}$ – это число вхождений всех словоформ текста T, являющихся словоизменениями словоформы a, т. е. верно равенство

$$n_T^{Par_a} = \sum_{b \in Par_a,\ b \ne a} n_T^b.$$

Параметр $n_T^{Syn_a}$ означает количество синонимов словоформы a в тексте T:

$$n_T^{Syn_a} = \sum_{c \in Syn_a,\ c \ne a} n_T^c.$$

Аналогичный смысл имеют параметры $N_{Cf}^{Par_a}$ и $N_{Cf}^{Syn_a}$:

$$N_{Cf}^{Par_a} = \sum_{b \in Par_a,\ b \ne a} n_{Cf}^b, \quad N_{Cf}^{Syn_a} = \sum_{c \in Syn_a,\ c \ne a} n_{Cf}^c.$$

Пусть теперь S – краткое текстовое сообщение. Обозначим через W_S множество всех его словоформ. Вычислим информативность $J_{Cf}^{S\pi}$ вербальной ассоциации между текстом S и некоторым предложением π из полного корпуса текстов Cf. По аналогии с выражением (3) построим вектор $\mathbf{J}_{Cf}^{S\pi} = (J_{Cf}^{c_1 d_1}, J_{Cf}^{c_2 d_2}, \ldots, J_{Cf}^{a_k b_k})$ в евклидовом пространстве. Для вычисления информативности $J_{Cf}^{S\pi}$ воспользуемся аналогом формулы (5):

$$J_{Cf}^{S\pi} = \frac{J_1 + J_2 + \ldots}{\sqrt{(J_1)^2 + (J_2)^2 + \ldots}}, \tag{8}$$

где J_1, J_2, ... – все отличные от нуля координаты вектора $\mathbf{J}_{Cf}^{S\pi}$. Если информативность выражения (8) не меньше некоторого критического значения, то предложение π занесем в текст S. Аналогично поступим и с другими такими предложениями полного корпуса текстов. В результате получим расширенное множество предложений, которое снова будем считать текстом (запросом) S.

Информативность I_S^a любого слова $a \in W_S$ вычислим по формуле (7):

$$I_S^a = \frac{n_S^a + n_S^{Par_a} + n_S^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \tag{9}$$

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 (8) нормализованную информативность I_T^π предложения π будем вычислять по формуле

$$I_T^\pi = \frac{I_1 + I_2 + \ldots}{\sqrt{(I_1)^2 + (I_2)^2 + \ldots}}, \tag{10}$$

где I_1, I_2, ... – значения информативности всех слов предложения π.

Информативность произвольного текста T из полного корпуса текстов будем вычислять по формуле, аналогичной выражению (10):

$$I_{Cf}^T = \frac{I_T^\pi + I_T^\rho + \ldots}{\sqrt{(I_T^\pi)^2 + (I_T^\rho)^2 + \ldots}}, \tag{11}$$

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

Описание алгоритма индексирования запросов. Алгоритм индексирования запросов (классов предложений из описания проблемной ситуации) функционирует в три этапа.

На первом этапе вычисляется информативность каждого из множеств S_1, S_2, ... , S_m по формуле (11). Класс предложений будем считать информативным, если значение информативности не меньше некоторой пороговой величины. В результате выполнения первого этапа получаем 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 корпуса текстов Cf с использованием формулы (8). Полученные в результате такого расширения новые классы будем применять в качестве запросов на поиск аналогов принятых решений.

На втором этапе вычисляется информативность $I_{U_i}^a$ ($i = \overline{1, s}$) всех словоформ из предложений всех классов U_1, U_2, ... , U_s по формулам (7) и (9).

На третьем этапе формируются поисковые образы

$$\text{ПП}_i = \{(a, I^a_{U_i}); (b, I^b_{U_i}))\,; \ldots \mid a, b, \ldots \in U_i\}, \quad i = \overline{1, s}, \tag{12}$$

всех классов U_1, U_2, ... , U_s предложений из описания проблемной ситуации. Эти поисковые образы будут использованы в качестве поисковых предписаний на поиск аналогов решений.

Интернет-поиск аналогов принятых решений. Выявление аналогов принятых решений связано с двумя видами информационного поиска: поиска веб-страниц, упорядоченных по убыванию их информативности, и фактографического поиска информативных фрагментов текстовых документов на этих страницах.

При поиске аналогов решения поставленной задачи нужно учитывать тот факт, что в Интернете индексируются не сами документы, а веб-страницы, на которых они расположены. Это обстоятельство существенным образом влияет на выбор критериев выдачи и построение алгоритмов поиска аналогов.

Критерии выдачи. Критерий выдачи – это правило, согласно которому вычисляется степень соответствия запросу веб-страниц или текстовых документов, найденных в процессе информационного поиска. В большинстве информационных систем критерии выдачи строятся на основе векторной модели описания данных в виде косинуса угла между векторами поискового предписания и поискового образа документа [8].

Пусть по-прежнему W_{Cf} – множество всех словоформ полного корпуса текстов *Cf*, а R – m-мерное евклидово пространство ($m = |W_{Cf}|$). Для каждой веб-страницы S построим вектор ее поискового образа в пространстве R: $\mathbf{F}_S = (I_S^{a_1}, I_S^{a_2}, \ldots, I_S^{a_m})$. Аналогично запишем вектор поискового предписания (12): $\mathbf{F}_{\text{ПП}_i} = (I^{b_1}_{\text{ПП}_i}, I^{b_2}_{\text{ПП}_i}, \ldots, I^{b_m}_{\text{ПП}_i})$. Тогда для поиска веб-страниц по поисковому предписанию $\mathbf{F}_{\text{ПП}_i}$ в качестве критерия выдачи используем косинус угла φ между векторами $\mathbf{F}_S$ и $\mathbf{F}_{\text{ПП}_i}$:

$$\cos\varphi = \frac{\mathbf{F}_S \cdot \mathbf{F}_{\text{ПП}_i}}{|\mathbf{F}_S| \cdot |\mathbf{F}_{\text{ПП}_i}|} = \frac{\sum\limits_{j=1}^{m} I_S^{a_j} I^{b_j}_{\text{ПП}_i}}{\sqrt{\sum\limits_{j=1}^{m} (I_S^{a_j})^2} \cdot \sqrt{\sum\limits_{i=1}^{n} (I^{b_j}_{\text{ПП}_i})^2}}. \tag{13}$$

Описание алгоритма поиска веб-страниц. Поиск аналогов принятых решений осуществляется в три этапа.

На первом этапе предложения текста описания проблемной ситуации разбиваются на классы.

На втором этапе реализуется индексирование информативных классов предложений. В результате индексирования формируется совокупность поисковых предписаний (12) для интернет-поиска аналогов принятых решений.

На третьем этапе по каждому поисковому предписанию ПП_i ($i = \overline{1, s}$) проводится поиск веб-страниц, содержащих аналоги принятых решений. При поиске используется критерий выдачи (13), все найденные страницы упорядочиваются по убыванию его значений.

Описание алгоритма фактографического поиска. Поиск сводится к выделению в найденных текстах информативных фрагментов, релевантных каждому классу предложений из множества классов $\{U_1, U_2, \ldots, U_s\}$ ($s \le m$). Процедура включает два этапа.

На первом этапе вычисляется информативность I_Q^a каждой словоформы a из найденного текста Q по формуле, аналогичной выражению (7):

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

Затем определяется информативность I_Q^π, I_Q^ρ, ... каждого предложения текста Q по формуле (10).

На втором этапе фактографического поиска выявляется контекстное окружение всех информативных предложений текста Q путем вычисления информативности вербальной ассоциации каждого из них с другими предложениями этого текста по формуле (5).

Сформированные таким образом фрагменты текстовых документов на найденных веб-страницах могут быть использованы как аналоги принятых решений.

Лексико-семантическая обработка аналогов принятых решений. Данная обработка заключается в выявлении тонально окрашенной информации в описаниях принятых решений, найденных в Интернете. Для оценки тональности информации будем использовать совокупность специальных тематических корпусов текстов.

Тонально окрашенные тематические корпусы текстов. Для оценки тональности сообщений в Интернете будем использовать совокупность тонально окрашенных тематических корпусов текстов. Каждому корпусу соответствует некоторая оценка тональности. При n-балльной шкале оценок количество таких корпусов должно быть равно n. Всякий i-й корпус Ct_i включает текстовые документы одинаковой тональности, т. е. корпус Ct_i – эта пара $\langle Ct_i, Et_i\rangle$, где Et_i – оценка тональности каждого документа из множества Ct_i. В простейшем случае формируются два корпуса текстов с тонально окрашенной лексикой. Первый корпус создается для анализа положительной тональности в описаниях принятых решений, а второй – для анализа отрицательной тональности. На основе тонально окрашенных тематических корпусов текстов формируется частотный словарь тонально окрашенной лексики $Lex_a = \{\langle a, n^a_{Cf}, n^a_{\langle Ct_1, Et_1\rangle}, \ n^a_{\langle Ct_2, Et_2\rangle}, \ \ldots, \ n^a_{\langle Ct_n, Et_n\rangle}\rangle | \ a \in W_{Cf}\}$. Здесь a – словоформа, n^a_{Cf} и $n^a_{\langle Ct_i, Et_i\rangle}$ $(i = \overline{1, n})$ – абсолютные частоты ее появления соответственно в полном корпусе текстов Cf и в i-м тонально окрашенном корпусе, W_{Cf} – множество всех словоформ полного корпуса.

Оценка тональности аналогов принятых решений. Пусть по-прежнему Ct_i $(i = \overline{1, n})$ – тонально окрашенные тематические корпусы текстов. Каждый корпус Ct_i состоит из текстов одинаковой тональности и представляет собой пару $\langle Ct_i, Et_i\rangle$ (Et_i – оценка тональности для всех текстов из множества Ct_i). Пусть также Q – текстовое сообщение, полученное в результате формирования контекстного окружения к некоторому найденному на веб-странице информативному предложению. Построим вектор $\mathbf{F}_Q$ сообщения Q и векторы $\mathbf{F}_{Ct_i}$ $(i = \overline{1, n})$ по аналогии с построением векторов $\mathbf{F}_s$ и $\mathbf{F}_{\text{ПП}'_2}$. Для каждой пары $(\mathbf{F}_Q, \mathbf{F}_{Ct_i})$ вычислим косинус угла между этими векторами по формуле

$$\cos(\mathbf{F}_Q, \mathbf{F}_{Ct_i}) = \frac{\mathbf{F}_Q \mathbf{F}_{Ct_i}}{|\mathbf{F}_Q||\mathbf{F}_{Ct_i}|}, \ i = \overline{1, n}.$$

Тогда сообщению Q будет соответствовать оценка тональности Et_i при таком значении i, при котором $\cos(\mathbf{F}_Q, \mathbf{F}_{Ct_i})$ принимает наибольшее значение.

По результатам интернет-поиска аналогов принятых решений и анализа тонально окрашенной информации формируется отчет (таблица).

Отчет по результатам поиска аналогов принятых решений и анализа тонально окрашенной информации на примере приобретения участка или дачи

Тонально окрашенное сообщение	Адрес веб-страницы	Информативность, %	Оценка тональности
Купить участок. Вариант для терпеливых, трудолюбивых и богатых. Дачу нужно будет построить. Участок – облагородить. Сад – посадить. При больших вложениях, с наймом специалистов можно построиться примерно за два года. Но чаще всего нужно лет пять. И еще лет 15–20 на сад	https://myfin.by/stati/view/13498-za-i-protiv-stoit-li-pokupat-dachu	43	3 (из 10)
Купить готовую дачу. Выйдет где-то на 1/3 дешевле, чем было потрачено на ее строительство. И это только денежные расходы, компенсировать прошлому хозяину его труды и время не нужно		61	7 (из 10)

В таблице каждому сообщению соответствует адрес веб-страницы, на которой оно расположено, а также приведены информативность и числовое значение оценки тональности в принятой шкале.

Заключение. Разработана математическая модель интернет-поиска и лексико-семантической обработки аналогов принятых решений, найденных в Интернете по запросам, которые были синтезированы из описаний проблемных ситуаций в соответствии с алгоритмами из работы автора [2]. Промоделированы четыре этапа данного процесса. На первом этапе синтезируются запросы на основе исследования вербальных ассоциаций между предложениями в описании проблемной ситуации. На втором этапе полученные запросы индексируются. Поисковые предписания представляются в виде множеств слов с соответствующими значениями информативности. На третьем этапе реализуется поиск аналогов принятых решений в порядке, определяемом специальным упорядочивающим отношением, которое задается на множестве веб-страниц каждого сканируемого веб-сайта. На четвертом этапе проводится лексико-семантическая обработка информационных сообщений, в процессе которой найденные аналоги решений исследуются на тональность. При оценке тональности используются лингвистические словари тонально окрашенной лексики, которые формируются на основе специальных тонально окрашенных тематических корпусов текстов. В предельном случае создаются два типа словарей. Первый тип предназначен для анализа положительной тональности в описаниях принятых решений, а второй – для анализа отрицательной тональности.

Предложенная модель интернет-поиска и лексико-семантической обработки аналогов принятых решений может быть использована в тех предметных областях, где необходимо работать с крупными объемами текстовой информации. После получения результатов анализа тональности сообщений пользователю информационной системы будет проще принять окончательное решение.

Список использованных источников

1. Ларичев, О. И. Вербальный анализ решений / О. И. Ларичев. – М. : Наука, 2006. – 181 с.

2. 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.

3. Мартинович, Г. А. Вербальные ассоциации и организация лексикона человека / Г. А. Мартинович // Филологические науки. – 1989. – № 3. – С. 39–45.

4. Еленевская, М. Н. Хранение и описание вербальных ассоциаций: словари и тезаурусы [Электронный ресурс] / М. Н. Еленевская, И. Г. Овчинникова. – Режим доступа: https://cyberleninka.ru/article/n/hranenie-i-opisanie-verbalnyh-assotsiatsiy-slovari-i-tezaurusy/viewer. – Дата доступа: 05.11.2020.

5. Морковкин, В. В. Идеографические словари [Электронный ресурс] / В. В. Морковкин. – Режим доступа: http://rifmovnik.ru/ideog_book.htm. – Дата доступа: 05.11.2020.

6. Мартинович, Г. А. Вербальные ассоциации в ассоциативном эксперименте / Г. А. Мартинович. – СПб. : Изд-во СПбГУ, 1997. – 72 с.

7. Липницкий, С. Ф. Модель представления знаний в информационных системах на основе вербальных ассоциаций / С. Ф Липницкий // Информатика. – 2011. – № 4(32). – С. 21–28.

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

References

1. Larichev O. I. Verbal'nyj analiz reshenij. *Verbal Analysis of Decisions.* Moscow, Nauka, 2006, 181 p. (in Russian).

2. Lipnitsky S. F. Sintez zaprosov i poisk al'ternativ v sisteme informacionnoj podderzhki prinjatija reshenij [Synthesis of queries and search for alternatives in the decision support information system]. Problemy fiziki, matematiki i tehniki [*Problems of Physics, Mathematics and Technology*], 2020, no. 2, p. 91–95 (in Russian).

3. Martinovich G. A. Verbal'nye associacii i organizacija leksikona cheloveka [Verbal associations and organization of the human lexicon]. Filologicheskie nauki [*Philological Sciences*], 1989, no. 3, p 39–45 (in Russian).

4. Yelenevskaya M. N., Ovchinnikova I. G. Hranenie i opisanie verbal'nyh associacij: slovari i tezaurusy. *Storage and Description of Verbal Associations: Dictionaries and Thesauri* (in Russian). Available at: https://cyberleninka.ru/article/n/hranenie-i-opisanie-verbalnyh-assotsiatsiy-slovari-i-tezaurusy/viewer (accessed 05.11.2020).

5. Morkovkin V. V. Ideograficheskie slovari. *Ideographic Dictionaries* (in Russian). Available at: http://rifmovnik.ru/ideog_ book.htm (accessed 05.11.2020).

6. Martinovich G. A. Verbal'nye associacii v associativnom jeksperimente. *Verbal Associations in an Associative Experiment.* Saint Petersburg, Izdatel'stvo Sankt-Peterburgskogo gosudarstvennogo universiteta, 1997, 72 p. (in Russian).

7. Lipnitsky S. F. Model' predstavlenija znanij v informacionnyh sistemah na osnove verbal'nyh associacij [Model of knowledge representation in information systems based on verbal associations]. Informatika [*Informatics*], 2011, no. 4(32), p. 21–28 (in Russian).

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

Информация об авторе

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

Information about the author

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