ИНФОРМАТИКА

УДК 004.912

СИНТЕЗ ЗАПРОСОВ И ПОИСК АЛЬТЕРНАТИВ В СИСТЕМЕ ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

С.Ф. Липницкий

Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск

QUERY SYNTHESIS AND SEARCH OF ALTERNATIVES IN THE SYSTEM OF INFORMATION SUPPORT FOR DECISION MAKING

S.F. Lipnitsky

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk

Предлагается математическая модель синтеза запросов и веб-поиска альтернатив в системе информационной поддержки процессов принятия решений. В рамках модели решены задачи вычисления информативности слов, предложений и текстов, а также информативности вербальной ассоциации между ними. Приведены описания алгоритмов синтеза запросов, поиска веб-страниц и фактографической информации.

***Ключевые слова**: индексирование, информативность, информационный поиск, математическая модель, принятие решений, синтез запросов.*

A mathematical model of query synthesis and web search of alternatives in the information support system for decision-making processes is proposed. Within the framework of the model, the problems of calculating the information content of words, sentences and texts, as well as the information content of the verbal association between them are solved. Descriptions of the algorithms for query synthesis, search for web pages and factual information are described.

Keywords: *indexing, informativeness, information retrieval, mathematical model, decision making, query synthesis.*

Введение

Процесс принятия решений в различных предметных областях включает, как правило, следующие основные этапы:

– описание проблемной ситуации и постановка 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 данной статье предлагается математическая модель синтеза запросов и веб-поиска альтернативных решений путем исследования вербальных ассоциаций между предложениями в описании проблемной ситуации [2]. Процесс синтеза запросов и поиска альтернатив осуществляется в три 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 на веб-поиск альтернативных вариантов решения поставленной задачи.

На третьем этапе осуществляется поиск таких вариантов.

1 Классификация предложений в описании проблемной ситуации

Пусть $T = \langle \rho_1, \rho_2, ..., \rho_l \rangle$ – описание проблемной ситуации, где $\langle \rho_1, \rho_2, ..., \rho_l \rangle$ – кортеж предложений текста T. Процессу разбиения кортежа предложений на классы предшествуют процедуры вычисления информативности вербальной ассоциации между словами предложений и между самими предложениями.

1.1 Информативность вербальной ассоциации между словами. 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 (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

$$I_{Cf}^{ab} = n_{Cf}^{ab} / N_{Cf}, \qquad (1.1)$$

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

– частотный словарь словоформ

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

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;

– 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 данной 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 информации из словарей формулу (1.1) представим в виде

$$I_{Cf}^{ab} = \frac{n_{Cf}^{ab} + n_{Cf}^{Par_{ab}} + n_{Cf}^{Syn_{ab}}}{N_{Cf}}. \qquad (1.2)$$

Параметр $n_{Cf}^{Par_{ab}}$ в формуле (1.2) указывает на число вхождений всех пар словоформ, являющихся словоизменениями соответственно слов *a* и (или) *b* и встречающимися 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 $n_{Cf}^{Syn_{ab}}$:

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

1.2 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

$$\Theta = \langle (a_1, b_1), (a_2, b_2), \ldots, (a_l, b_l) \rangle.$$

Согласно определению отношения Θ, для каждой пары (a_i, b_i) из данного кортежа существует хотя бы одно предложение в корпусе *Cf*, в котором содержатся обе словоформы a_i и b_i $(i = \overline{1, l})$ и для них справедливо соотношение $a_i \prec_W b_i$.

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*:

$$\mathbf{I}_{Cf}^{\pi\rho} = (I_{Cf}^{a_1 b_1}, I_{Cf}^{a_2 b_2}, \ldots, I_{Cf}^{a_l b_l}). \qquad (1.3)$$

В формуле (1.3) значение информативности $I_{Cf}^{a_i b_i}$ для любого $i \in \{1, 2, \ldots, l\}$ 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 $\mathbf{e} = (1, 1, \ldots, 1)$ 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 $\mathbf{I}_{Cf}^{\pi\rho}$:

$$I_{Cf}^{\pi\rho} = \frac{\mathbf{I}_{Cf}^{\pi\rho} \cdot \mathbf{e}}{|\mathbf{I}_{Cf}^{\pi\rho}|} = \frac{\sum_{i=1}^{l} I_{Cf}^{a_i b_i}}{\sqrt{\sum_{i=1}^{l} (I_{Cf}^{a_i b_i})^2}}. \qquad (1.4)$$

При реализации алгоритма вычисления информативности вербальной ассоциации между предложениями или текстами удобно пользоваться следующей формулой, полученной из выражения (1.4):

$$I_{Cf}^{\pi\rho} = \frac{I_1 + I_2 + \ldots + I_r}{\sqrt{(I_1)^2 + (I_2)^2 + \ldots + (I_r)^2}}, \qquad (1.5)$$

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

1.3 Описание алгоритма классификации предложений. Алгоритм разбиения кортежа *T* 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 (1.5) вычисляется информативнось 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 (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 $\{S_1, S_2, ..., S_m\}$ $(m \le l)$.

2 Индексирование информативных классов предложений

Среди сформированных классов предложений S_1, S_2, ..., S_m могут быть неинформативные, использование которых в качестве запросов на поиск альтернативных решений поставленной проблемы нецелесообразно. В связи с этим рассмотрим вопросы вычисления информативности классов предложений.

2.1 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 определена в статье [3] как вероятность того, что слово а имеется в данном текстовом документе при условии, что оно содержится в полном корпусе текстов. Рассмотрим совокупность событий (в вероятностном смысле):

– S_D – некоторая словоформа a извлечена случайным образом из текста T ($T \in Cf$);

– H_T – появление текста T;

– S_{Cf} – словоформа a содержится в полном корпусе текстов Cf.

Тогда $I_T^a = P(S_T \mid S_{Cf})$, где $P(S_T \mid S_{Cf})$ – условная вероятность того, что словоформа а извлечена из текста T при условии, что она уже извлечена из полного корпуса текстов Cf. Эта вероятность вычисляется следующим образом:

$$P(S_T / S_{Cf}) = \frac{P(S_T \cdot S_{Cf})}{P(S_{Cf})} = \frac{P(S_T) \cdot P(S_{Cf} / S_T)}{P(S_{Cf})}.$$

Учитывая, что $P(S_{Cf} / S_T) = 1$, воспользовавшись формулой полной вероятности, получим

$$P(S_T / S_{Cf}) = \frac{P(S_T / H_D)}{P(S_{Cf})} \cdot P(H_T).$$

При достаточно больших объемах полного корпуса текстов Ct и текстового документа T можно считать, что

$$P(S_T / H_T) \approx \frac{n_T}{N_T},\ P(S_{Cf}) \approx \frac{n_{Cf}}{N_{Cf}},\ P(H_T) \approx \frac{N_T}{N_{Cf}},$$

где n_T, n_{Cf} – частоты встречаемости (с учетом словоизменения и синонимии) словоформы a в тексте T и полном корпусе текстов Cf, а N_T, N_{Cf} – число вхождений всех словоформ в T и Cf соответственно. Тогда формула для вычисления информативности I_T^a словоформы a в тексте T имеет вид

$$I_T^a = \frac{n_T}{n_{Cf}}. \qquad (2.1)$$

Используя лингвистические словари Dic_{par} и Dic_{syn}, формулу (2.1) перепишем в виде

$$I_T^a = \frac{n_T^a + n_T^{Par_a} + n_T^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \qquad (2.2)$$

В формуле (2.2) $n_T^{Par_a}$ – это число вхождений всех словоформ текста T, являющихся словоизменениями словоформы a, т. е.

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

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 $N_{Cf}^{Syn_a}$:

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

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

2.2 Информативность слов из кратких текстовых сообщений. Под кратким сообщением будем понимать текстовый документ, объем которого не позволяет выявить статистические характеристики его словоформ. Поэтому 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. По аналогии с выражением (1.3) построим вектор $\mathbf{J}_{Cf}^{Q\pi} = (J_{Cf}^{c_1 d_1}, J_{Cf}^{c_2 d_2}, ..., J_{Cf}^{a_k b_k})$ в евклидовом пространстве. Для вычисления информативности $J_{Cf}^{Q\pi}$ воспользуемся аналогом формулы (1.5):

$$J_{Cf}^{Q\pi} = \frac{J_1 + J_2 + ...}{\sqrt{(J_1)^2 + (J_2)^2 + ...}}, \qquad (2.3)$$

где J_1, J_2, ... – все отличные от нуля координаты вектора $\mathbf{J}_{Cf}^{Q\pi}$. Если информативность (2.3) 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 (2.2):

$$I_Q^a = \frac{n_Q^a + n_Q^{Par_a} + n_Q^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \qquad (2.4)$$

2.3 Информативность предложений и текстов. При вычислении информативности предложений текста T будем также исходить из их векторного представления: $\mathbf{\Pi} = (I_\pi^{a_1}, I_\pi^{a_2}, ..., I_\pi^{a_l})$, где $I_\pi^{a_1}, I_\pi^{a_2}, ..., I_\pi^{a_l}$ – значения информативности слов произвольного предложения π (компонента вектора **Π** равна нулю, если соответствующего слова нет в предложении π). Тогда, аналогично формуле (2.3), нормализованную информативность I_T^π предложения π будем вычислять по формуле:

$$I_T^\pi = \frac{I_1 + I_2 + ...}{\sqrt{(I_1)^2 + (I_2)^2 + ...}}, \qquad (2.5)$$

где $I_1, I_2, ...$ – значения информативности всех слов предложения π [4].

Информативность произвольного текста *T* из полного корпуса текстов будем вычислять по формуле, аналогичной выражению (2.5):

$$I_{Cf}^T = \frac{I_T^\pi + I_T^\rho + ...}{\sqrt{(I_T^\pi)^2 + (I_T^\rho)^2 + ...}}, \qquad (2.6)$$

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

2.4 Описание алгоритма индексирования классов предложений. Алгоритм индексирования классов предложений из описания проблемной ситуации функционирует в три этапа.

На первом этапе вычисляется информативность каждого из классов предложений S_1, S_2, ..., S_m по формуле (2.6). Класс будем считать информативным, если значение информативности не меньше некоторой пороговой величины. В результате выполнения первого этапа имеем совокупность информативных классов предложений $\{U_1, U_2, ..., U_s\}$ $(s \le m)$. 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 (2.3). Полученные в результате такого расширения новые классы будем использовать в качестве запросов на поиск альтернатив.

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 (2.2) и (2.4).

На третьем этапе формируются поисковые образы

$$\text{ПП}_i = \{(a, I_{U_i}^a); (b, I_{U_i}^b)); ... \mid a, b, ... \in U_i\}, \quad i = \overline{1, s} \qquad (2.7)$$

всех классов U_1, U_2, ..., U_s предложений из описания проблемной ситуации. Эти поисковые образы будут использованы в качестве поисковых предписаний на поиск альтернативных вариантов решения поставленной задачи в рамках проблемной ситуации.

3 Поиск альтернативных вариантов решения проблемной ситуации

Выявление альтернатив при принятии решений связано с двумя видами информационного поиска – поиска веб-страниц, упорядоченных по убыванию их информативности, и фактографического поиска информативных фрагментов полнотекстовых документов на этих страницах.
При поиске альтернативных вариантов 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 альтернатив.

3.1 Критерии выдачи. Под критерием выдачи понимается правило, по которому вычисляется степень соответствия запросу веб-страниц или текстовых документов, найденных в процессе информационного поиска. В большинстве известных информационных систем критерии выдачи строятся на основе векторной модели описания данных [5] в виде косинуса угла между векторами поискового предписания и поискового образа документа. Рассмотрим эту меру близости, используя принятые выше обозначения. Пусть, по-прежнему, *W* – множество всех словоформ полного корпуса текстов *Cf*, а *E* – *m*-мерное евклидово пространство $(m = |W|)$. Для каждой веб-страницы *S* построим вектор ее поискового образа в пространстве *E*: $\mathbf{F}_S = (I_{a_1}, I_{a_2}, ..., I_{a_m})$. Аналогично запишем вектор поискового предписания (2.7):

$$\mathbf{F}_{\Pi\Pi_i} = (I_{b_1}, I_{b_2}, ..., I_{b_m}).$$

Тогда для поиска веб-страниц по поисковому предписанию $\mathbf{F}_{\Pi\Pi_i}$ в качестве критерия выдачи используем косинус угла φ между векторами $\mathbf{F}_S$ и $\mathbf{F}_{\Pi\Pi_i}$:

$$\cos\varphi = \frac{\mathbf{F}_S \cdot \mathbf{F}_{\Pi\Pi_i}}{|\mathbf{F}_S| \cdot |\mathbf{F}_{\Pi\Pi_i}|} = \frac{\sum_{j=1}^{m} I_{a_j} I_{b_j}}{\sqrt{\sum_{j=1}^{m} (I_{a_j})^2} \cdot \sqrt{\sum_{i=1}^{n} (I_{b_j})^2}}. \qquad (3.1)$$

3.2 Описание алгоритма поиска веб-страниц. Поиск альтернативных вариантов при принятии решений осуществляется в три этапа.

На первом этапе предложения текста описания проблемной ситуации разбиваются на классы в соответствии с алгоритмом, описанном в п. 1.3.

На втором этапе выполняется алгоритм из п. 2.4. Согласно этому алгоритму реализуется индексирование информативных классов предложений. В результате индексирования формируется совокупность поисковых предписаний (3.1) для интернет-поиска альтернатив при принятии решений.

На третьем этапе по каждому поисковому предписанию ПП_i $(i = \overline{1, s})$ проводится поиск веб-страниц, содержащих альтернативные варианты решения поставленной задачи в рамках проблемной ситуации. При поиске используется критерий выдачи (3.2). Все найденные страницы упорядочиваются по убыванию его значений.

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

На первом этапе вычисляется информативность I_T^a каждой словоформы a из найденного текста T по формуле, аналогичной выражению (2.2):

$$I_T^a = \frac{n_U^a + n_U^{Par_a} + n_U^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}.$$

Затем определяется информативность I_T^π, I_T^ρ, ... каждого предложения текста T по формуле (2.5).

На втором этапе фактографического поиска выявляется контекстное окружение всех информативных предложений текста T путем вычисления информативности вербальной ассоциации каждого из них с другими предложениями данного текста по формуле (1.5).

Сформированные таким образом фрагменты текстовых документов на найденных веб-страницах могут быть использованы при информационной поддержке процессов принятия решений.

Заключение

Разработана математическая модель синтеза запросов и веб-поиска альтернативных вариантов в системе информационной поддержки процесса принятия решений. Промоделированы три этапа данного процесса. 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 на веб-поиск альтернативных вариантов решения поставленной задачи. На третьем этапе осуществляется поиск таких вариантов.

ЛИТЕРАТУРА

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

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

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

4. *Липницкий*, *С.Ф.* Математическая модель синтеза текстов на основе слияния коммуникативных фрагментов / С.Ф. Липницкий // Проблемы физики, математики и техники. – 2018. – № 4 (37). – С. 106–110.

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

Поступила в редакцию 05.03.2020.