УДК 004.912

А.А. Мамчич

АЛГОРИТМЫ ИНДЕКСИРОВАНИЯ И ПОИСКА ДОКУМЕНТОВ НА ОСНОВЕ ДИНАМИЧЕСКИХ КОРПУСОВ ТЕКСТОВ

Предлагаются тематические алгоритмы индексирования и поиска документов на основе динамических корпусов текстов. Данные алгоритмы могут применяться для индексирования и поиска как полнотекстовых документов, так и кратких сообщений в Интернете или в локальных базах данных. Предлагается метод вычисления весов ключевых слов на основе динамических корпусов текстов, который может использоваться в различных информационно-поисковых системах для повышения релевантности и полноты поиска текстовой информации.

Введение

Под классической задачей информационного поиска понимается процесс отыскания в некотором множестве текстов, которые удовлетворяют информационному запросу пользователей в рамках определенной коллекции документов [1, 2]. На современном этапе развития компьютерных сетей за счет экспоненциального увеличения количества текстовой информации круг задач информационного поиска существенно расширился и включает в себя такие базовые задачи, как тематическое индексирование, поиск с адаптацией к информационным потребностям пользователя, кластеризация и классификация документов и т. д. Поисковые системы, которые обеспечивают решение перечисленных выше задач, ориентированы на конкретную область, что позволяет осуществлять глубокий поиск по определенной тематике. В данной работе предлагается подход к организации процесса тематического индексирования и поиска документов на основе динамических корпусов текстов (наборов текстов, релевантных каждому конкретному тексту или запросу на поиск информации). Такие корпусы обеспечивают решение перечисленных выше задач и позволяют индексировать и искать не только полнотекстовые документы, но также и краткие сообщения.

Существующие методы решения задач тематического индексирования и поиска документов основаны главным образом на семантическом и статистическом анализе текстов [3, 4] и выборе специализированной стратегии обхода ресурсов, максимизирующей число обнаруженных документов по искомой тематике [5]. Главным недостатком таких подходов является то, что в процессе индексирования и поиска информации учитываются только статистические характеристики самого текста без привлечения знаний о предметной области, которые существенно повышают интеллектуальность данных процессов [6, 7].

В рамках рассматриваемой проблемы нерешенными являются следующие задачи:

– вычисление весов ключевых слов, определяющих их информативность, на основе динамических корпусов текстов;

– разработка алгоритма тематического индексирования документов на основе динамических корпусов текстов;

– разработка алгоритмов поиска текстовых документов по документу-образцу, поиска документов с учетом их тональности и стилистической окраски, алгоритмов кластеризации и классификации.

Решение этих задач является основной целью настоящей работы.

1. Определение динамического корпуса текстов

Моделирование знаний о предметной области позволяет повысить эффективность функционирования процессов индексирования и поиска за счет их интеллектуализации и служит теоретической базой для исследования этих процессов [6]. В качестве знаний информационно-поисковой системы будем использовать динамические корпусы текстов. Промоделируем и определим понятие динамического корпуса текстов.

Входная и выходная информация в процессе индексирования и поиска текстовой информации описывается на специализированных языках, посредством которых обеспечивается функционирование информационно-поисковой системы. Будем различать три информационных языка: входной $L(G_{вх})$, внутренний $L(G_{вн})$ и выходной $L(G_{вых})$.

Язык $L(G_{вх})$ предназначен для взаимодействия пользователя с информационно-поисковой системой, посредством которого он формулирует запросы на поиск информации. Алфавит данного языка включает символы естественного языка, цифры, разделительные символы, знаки арифметических и логических операций (И, НЕ, ИЛИ).

Язык $L(G_{вн})$ используется для описания основного содержания текстовых документов или темы документа (путем индексирования), поиска множества документов среди множества других и представляется в виде конечной последовательности пар «характеристика – числовое значение характеристики».

Выходной язык $L(G_{вых})$ часто совпадает с входным языком $L(G_{вх})$ и служит для отображения результатов поиска документов.

Перечисленные выше информационные языки можно определить и промоделировать следующим образом.

В работе [8] для определения информационных языков введена формальная порождающая грамматика в виде упорядоченной четверки вида

$$G = \langle V, N, I, R \rangle,$$

где V – непустое конечное множество терминальных элементов (слов), которое называют основным словарем грамматики G;

N – непустое конечное множество нетерминальных элементов, которое называют вспомогательным словарем грамматики G;

I – некоторый элемент множества V, который называют начальным символом грамматики G;

R – схема грамматики, т. е. конечное множество цепочек вида $\alpha \to \beta$ (α и β – различные непустые цепочки в словаре $V \cup N$ и $\to$ – символ, не входящий в $V \cup N$). Элементы множества R называются правилами грамматики G.

Определение информационного языка, заданного порождающей грамматикой G, связано с понятием выводимости.

Рассмотрим словарь V грамматики G. Произвольная конечная последовательность символов ρ называется цепочкой в словаре V. Пусть $\delta = \varphi \to \psi$ – правило грамматики G и $\xi_1\varphi\xi_2$ – вхождение φ в цепочку $\rho = \xi_1\varphi\xi_2$ в словаре $V \cup N$. В этом случае говорят, что цепочка $\chi = \xi_1\psi\xi_2$ получается из цепочки ρ путем применения правила $\delta = \varphi \to \psi$ к символу φ цепочки $\xi_1\varphi\xi_2$. Если цепочка χ получается из цепочки ρ путем применения какого-либо правила грамматики G, то полагают, что цепочка χ непосредственно выводима из цепочки ρ в грамматике G. Последовательность цепочек $D = (\rho_0, \rho_1, \ldots, \rho_n)$ ($n \geq 1$) называется выводом цепочки ρ_n из цепочки ρ_0 в грамматике G, если для каждого i ($1 \leq i \leq n$) цепочка ρ_{i-1} непосредственно выводима из ρ_i. Число n называется длиной вывода D. Если существует вывод цепочки χ из цепочки ρ в грамматике G, то говорят, что χ выводима из ρ. Тогда множество цепочек в основном словаре V грамматики G, выводимых из ее начального символа I, называется языком $L(G)$, порождаемым грамматикой G.

Рассмотрим порождающую формальную грамматику $G_1 = \langle V_N, N, I, R \rangle$, схему R которой определим следующим образом:

– для любого слова $a \in V_N$ существуют правила вывода $I \to a'$ и $a' \to a$;

– все остальные правила вывода имеют вид $a' \to a'b'$ или $a' \to b'a'$, где $a, b \in V_N$.

Пусть $V_{вх}$ – словарь некоторого естественного языка, который будем называть входным словарем, а его элементы – словами входного языка. По аналогии со схемой R грамматики G_1 построим совокупность правил вывода $R_{вх}$. Тогда язык $L(G_{вх})$, порождаемый грамматикой $G_{вх} = \langle V_{вх}, N, I, R_{вх} \rangle$, будем называть входным языком. Аналогичным образом определяются внутренний и выходной информационные языки $L(G_{вн})$ и $L(G_{вых})$, в качестве словарей которых вы-

ступают непустые множества терминальных элементов $V_{вн}$ и $V_{вых}$ (внутренний и выходной словари), а схемы грамматик аналогичны $R_{вх}$.

Любое непустое подмножество t языка $L(G)$, т. е. множество цепочек в основном словаре V грамматики G, выводимых из начального символа I, будем называть текстом, если на этом подмножестве определена редукция $\prec r = \prec \setminus \prec^2$ линейного порядка $\prec$ (транзитивного и антисимметричного бинарного отношения на множестве t, которое связанно на t, т. е. для любых $a, b \in t$ или $a \prec b$, или $b \prec a$, или $a = b$). Цепочки текста t назовем предложениями. Под полным корпусом текстов будем понимать объединение текстов t_i, т. е. множество $Cf = \bigcup_{i=1}^{n} t_i$.

Обозначим через $T_{вн}$ множество всех текстов внутреннего языка $L(G_{вн})$, а через $T_{вх}$ и $T_{вых}$ – множество всех текстов входного $L(G_{вх})$ и выходного $L(G_{вых})$ языков.

Пусть $Z_{вх}$ ($Z_{вх} \subseteq T_{вх}$) – непустое подмножество множества $T_{вх}$, элементы которого будем называть запросами. Текст $t_{вн} \in T_{вн}$ назовем поисковым образом произвольного текста $t_{вх} \in T_{вх}$, если существует такое инъективное отображение $\omega : T_{вх} \to T_{вн}$, что текст $t_{вн}$ является образом текста $t_{вх}$ при отображении ω. В этом случае будем говорить, что отображение ω моделирует процесс индексирования текстовых документов. В случае если текст $t_{вх}$ является запросом, текст $t_{вн} = \omega(t_{вх})$ будем называть поисковым предписанием, соответствующим запросу $t_{вх} \in Z_{вх}$. Отображение $\eta : \omega(T_{вх}) \times \omega(Z_{вх}) \to R$ декартова произведения множеств запросов и поисковых предписаний во множество R действительных чисел будем называть критерием выдачи, который является мерой смысловой близости между поисковым образом текста и поисковым предписанием.

Один шаг поиска текстов можно промоделировать в виде частичного мультиотображения $\pi : Z_{вх} \to T_{вх}$ множества запросов в множество текстов. Частичное мультиотображение π назовем поисковой функцией, если для любого запроса $z \in Z_{вх}$ множество $\pi(z)$ включает те и только те тексты $t \in T_{вх}$, для которых значение критерия выдачи не меньше некоторого η_0, т. е. $\eta(\omega(t), \varepsilon(z)) \geq \eta_0$.

Пусть $t \in Z_{вх}$ – некоторый текст (или, в частном случае, запрос) входного языка $L(G_{вх})$, $\omega(t)$ – поисковое предписание, полученное в результате перевода текста t на внутренний язык $L(G_{вн})$. Тогда множество текстов $\pi(t)$ полного корпуса текстов Cf, т. е. образ текста t при частичном мультиотображении $\pi : Z_{вх} \to Cf$, будем называть *динамическим корпусом текстов* для текста t ($\pi : Z_{вх} \to Cf$ – сужение поисковой функции π на множество Cf).

Модель базы знаний на основе динамических корпусов текстов можно представить в виде смешанного графа, вершинами которого являются основы слов, дугами – семантические связи между ними, а в качестве ребер выступают ситуативные отношения между словами. Под ситуативным отношением некоторой пары слов понимается выполнение условия того, что вероятность совместной встречаемости этой пары в корпусе текстов $\pi(t)$ (наличие их в одном и том же предложении корпуса) не меньше некоторого порогового значения (уровня ситуативной связи) [9].

На практике динамический корпус представляет собой набор документов, релевантных каждому конкретному тексту или запросу на поиск информации, а полный корпус Cf – совокупность всех документов по различным предметным областям. При программной реализации информационно-поисковой системы полный корпус текстов Cf хранится в памяти компьютера в виде комплекса словарей (например, комплекса, состоящего из словарей словоформ полного корпуса текстов, парадигм, синонимов и т. д.).

2. Определение весов ключевых слов документа на основе динамических корпусов текстов

Вес ключевого слова определяет информативность словоформы и учитывается при поиске информации по запросу пользователя.

В работе [10] авторами была получена формула для вычисления информативности словоформ на основе тематических корпусов текстов (совокупности текстов по конкретной тема-

тике) с учетом словоизменений и синонимии. Под словоизменением будем понимать образование словоформ той же лексемы (базовой единицы языка), имеющих разные грамматические значения. Обобщим формулу из работы [10] на случай динамического корпуса текстов.

Определим вес ключевого слова как условную вероятность того, что данная словоформа извлечена из динамического корпуса текстов $\pi(t)$ для индексируемого текста t (или из самого текстового документа) при условии, что она уже извлечена из полного корпуса текстов Cf:

$$P(S_{\pi(t)} / S_{Cf}) = \frac{P(S_{\pi(t)} \cdot S_{Cf})}{P(S_{Cf})} = \frac{P(S_{\pi(t)}) \cdot P(S_{Cf} / S_{\pi(t)})}{P(S_{Cf})}. \quad (1)$$

В формуле (1) задействованы следующие события:

$S_{\pi(t)}$ – словоформа извлечена случайным образом из динамического корпуса текстов $\pi(t)$ (или из самого текстового документа);

S_{Cf} – словоформа извлечена из полного корпуса текстов Cf.

Пусть $n_{\pi(t)}$, n_{Cf} – абсолютные частоты встречаемости словоформы a в корпусах текстов $\pi(t)$ и Cf соответственно. Тогда нетрудно установить, что при достаточно больших объемах этих корпусов формула для вычисления информативности словоформы примет вид

$$I_a = \frac{n_{\pi(t)}}{n_{Cf}}. \quad (2)$$

При определении весов ключевых слов документа t используется следующая методика:

1. Выбирается очередная словоформа a из индексируемого документа t.

2. Определяется n_1 – количество вхождений словоформы a в корпусе текстов $\pi(t)$ (частота словоформы в $\pi(t)$).

3. Определяется n_2 – количество словоизменений словоформы a в корпусе текстов $\pi(t)$ (частота словоизменений словоформы в $\pi(t)$).

4. Определяется n_3 – количество вхождений в $\pi(t)$ словоформ, которые являются синонимами данной словоформы (частота синонимов словоформы в $\pi(t)$).

5. Определяется N – суммарное число вхождений исходной словоформы, ее словоизменений и синонимов в полном корпусе текстов Cf (частота словоформы с учетом словоизменений и синонимии в Cf).

С учетом этого формула (2) для определения веса словоформы a примет вид

$$I_a = \frac{n_1 + n_2 + n_3}{N + (n_1 + n_2 + n_3)}. \quad (3)$$

3. Тематическое индексирование документов на основе динамических корпусов текстов

Процедура поиска текстовой информации сводится к сопоставлению поисковых образов документов, полученных на этапе индексирования, с информационными запросами пользователей. Таким образом, процесс индексирования текстового документа Td заключается в построении множества пар вида $D = \{(a, I_a) \mid a \in A, 0 \le I_a \le 1\}$, называемых поисковым образом документа, где A – множество словоформ в поисковом образе документа; I_a – веса соответствующих словоформ, вычисленные по формуле (3).

Под задачей тематического индексирования понимается автоматический процесс выявления документов, соответствующих заданной тематике, при котором отсеиваются все прочие документы. Эта задача актуальна при создании специализированных поисковых систем по различным предметным областям (например, технике, космосу, культуре и т. д.). За счет индекси-

рования ресурсов только определенной тематики увеличивается релевантность поиска, снижается объем индексной базы данных, что приводит к возрастанию скорости поиска и динамики обновления информации.

Перед началом процесса тематического индексирования эксперт формирует тематический фильтр, который представляет собой множество вида $F = \{(w_f) \mid w_f \in W_f\}$, где W_f – множество словоформ, определяющих тематику документов, предназначенных для индексирования. Помимо этого эксперт задает тематический порог K (нижняя граница оценки тематической близости документа), который устанавливается экспериментальным путем. На основании фильтра F из полного корпуса текстов Cf формируется динамический корпус текстов $\pi_F(t)$ посредством поиска документов в полном корпусе текстов, которые содержат все словоформы из множества W_f. Динамический корпус текстов $\pi_F(t)$ индексируется как один документ, при этом веса слов вычисляются по формуле (3). В итоге тематический фильтр будет представлять собой множество пар вида $T_{\pi_F(t)} = \{(w_{\pi_F(t)}, I_{\pi_F(t)}) \mid w_{\pi_F(t)} \in W_{\pi_F(t)}, 0 \le I_{\pi_F(t)} \le 1\}$, где $W_{\pi_F(t)}$ – множество словоформ в динамическом корпусе текстов $\pi_F(t)$; $I_{\pi_F(t)}$ – веса соответствующих словоформ, вычисленные по формуле (3).

Процедура тематического индексирования реализуется в два этапа. На первом этапе происходит предварительная обработка документа *Td*, которая включает в себя лексический анализ текста (удаление элементов форматирования, цифр, элементов пунктуации, математических формул и т. п.), исключение стоп-слов (слов малой информативности, которые не нужно учитывать при поиске документов: союзы, местоимения и т. д.). На втором этапе формируется поисковый образ документа $P = \{(b_i, J_i) \mid b_i \in B, 0 \le J_i \le 1, i = \overline{1,n}\}$, где B – множество ключевых слов, размерность которого равна количеству n словоформ в документе *Td*; J_i – веса соответствующих слов, вычисленные по формуле (3). Поскольку поисковый образ P документа *Td* и образ динамического корпуса текстов $\pi_F(t)$ можно представить в виде векторов в n-мерном евклидовом пространстве, в качестве координат которых выступают соответствующие веса словоформ, оценку тематической близости образа индексируемого документа *Td* и динамического корпуса текстов $\pi_F(t)$ можно представить как скалярное произведение

$$S_F = \sum_{i=1}^{n} J_i \cdot I^i_{\pi_F(t)}. \tag{4}$$

Тогда принадлежность документа *Td* тематике индексирования, заданной экспертом в виде тематического фильтра F, определяется исходя из следующих условий:

$$X = \begin{cases} S_F \ge K, \text{документ } Td \text{ удовлетворяет тематике индексирования;} \\ S_F \ge K, \text{документ } Td \text{ не удовлетворяет тематике индексирования.} \end{cases} \tag{5}$$

Опишем *алгоритм тематического индексирования текстовых документов*. На входе алгоритма – множество текстовых документов Td_j, прошедших предварительную обработку, тематический фильтр F и тематический порог K, на выходе – множество O_{Td} поисковых образов документов, удовлетворяющих тематике индексирования. Алгоритм индексирования включает в себя следующие шаги:

Шаг 1. $O_{Td} := \varnothing$.

Шаг 2. Получить очередной текстовый документ Td_j.

Шаг 3. Сформировать поисковый образ данного документа, вычисляя веса слов по формуле (3).

Шаг 4. Сформировать динамический корпус текстов $\pi_F(t)$ для фильтра F.

Шаг 5. Индексировать динамический корпус текстов $\pi_F(t)$ как конкатенацию текстовых документов и получить тематический фильтр $T_{\pi_F(t)}$.

Шаг 6. Вычислить тематическую близость S_F по формуле (4).

Шаг 7. На основании критерия (5) определить принадлежность документа Td_j тематике индексирования.

Шаг 8. Если Td_j принадлежит тематике индексирования, добавить его образ в множество O_{Td}.

Шаг 9. Если все текстовые документы Td_j обработаны, то КОНЕЦ (множество O_{Td} поисковых образов документов, удовлетворяющих тематике индексирования, сформировано), иначе перейти к шагу 2.

4. Тематический поиск документов на основе динамических корпусов текстов

Особенность задач тематического информационного поиска текстовых документов состоит в том, что в начале поиска пользователь четко не знает свою информационную потребность, а имеет лишь общее представление о ней – тему поиска. Поэтому он не может сформулировать точный первоначальный запрос, по которому будут найдены интересующие его документы, и решение задач тематического поиска сводится к уточнению первоначального запроса с целью уяснения информационной потребности пользователя. С помощью динамических корпусов текстов могут быть решены задачи тематического поиска документов за счет расширения запроса пользователя релевантным ему корпусом текстов и последующего динамического изменения пространства поиска.

Использование динамических корпусов текстов в системе поиска позволяет осуществить:

– поиск документов по документу-образцу, т. е. поиск с адаптацией к информационным потребностям пользователей;

– поиск документов с учетом их тональности и стилистической окраски;

– кластеризацию документов, выделение компактных подгрупп документов с близкими свойствами, т. е. разбиение их на непересекающиеся группы по тематике, без предварительного задания количества и характеристик этих групп;

– классификацию документов, т. е. отнесение документов к заранее определенным группам по тематике.

4.1. Алгоритм поиска документов по документу-образцу

Если в результате поиска было найдено избыточное количество текстовых документов и найден хотя бы один документ *Td*, удовлетворяющий информационной потребности пользователя (назовем его документом-образцом), то нахождение документов, сходных по содержанию с документом-образцом, приведет к получению более точных результатов поиска по теме первичного запроса *z*.

Рассмотрим *алгоритм поиска текстовых документов по документу-образцу*, который работает следующим образом: для документа, удовлетворяющего информационным потребностям пользователей, формируется динамический корпус текстов, образ которого в дальнейшем рассматривается как запрос на поиск информации.

На входе алгоритма – документ *Td*, пертинентный запросу пользователя *z*, и множество $T_{\text{вх.}}$ текстовых документов, на выходе – множество $\pi(Z)$ найденных документов. Алгоритм поиска включает в себя следующие шаги:

Шаг 1. $\pi(Z) := \varnothing$.

Шаг 2. Сформировать поисковый образ *P* документа *Td*, вычисляя веса слов по формуле (3).

Шаг 3. *z:=P*.

Шаг 4. По запросу *z* сформировать динамический корпус текстов $\pi_{Cf}(z)$ путем поиска в полном корпусе текстов *Cf* документов, релевантных *z*.

Шаг 5. Индексировать динамический корпус текстов $\pi_{Cf}(z)$ как конкатенацию документов и получить расширенный поисковый запрос $Z = \omega^{-1}(\pi_{Cf}(z))$.

Шаг 6. Найти во множестве текстовых документов $T_{\text{вх.}}$ множество документов $\pi(Z)$ по запросу Z на основании критерия выдачи принятого в системе поиска. КОНЕЦ.

4.2. Алгоритм поиска документов с учетом их тональности и стилистической окраски

Для реализации этого вида поиска в базе данных информационно-поисковой системы должен быть предварительно сформирован корпус текстов $Cf_{\text{тон.}} \subseteq Cf$., состоящий из документов, каждый из которых имеет необходимую тональность и стилистическую окраску. Такой вид поиска предназначен для нахождения документов, в которых используется, например, специальная (узкопрофильная) или своеобразная общеупотребительная лексика.

Рассмотрим *алгоритм поиска текстовых документов с учетом их тональности и стилистической окраски*, который сводится к формированию динамического корпуса текстов на основе корпуса $Cf_{\text{тон}}$ и проведению поиска, где в качестве запроса выступает образ соответствующего динамического корпуса.

На входе алгоритма – корпус текстов $Cf_{\text{тон}}$, множество $T_{\text{вх.}}$ текстовых документов и запрос пользователя z, на выходе – множество $\pi(Z)$ найденных документов с учетом их тональности и стилистической окраски.

Шаг 1. $\pi(Z) := \varnothing$.

Шаг 2. По запросу z сформировать динамический корпус текстов $\pi_{Cf}(z)$ путем поиска в корпусе текстов $Cf_{\text{тон}}$ документов, релевантных z.

Шаг 3. Индексировать динамический корпус текстов $\pi_{Cf}(z)$ как конкатенацию документов и получить расширенный поисковый запрос $Z = \omega^{-1}(\pi_{Cf}(z))$.

Шаг 4. Найти в множестве текстовых документов $T_{\text{вх.}}$ множество документов $\pi(Z)$ по запросу Z на основании критерия выдачи, принятого в системе поиска. КОНЕЦ.

4.3. Алгоритм кластеризации документов

Кластеризация текстовых документов удобна для уточнения первоначального запроса пользователя z путем автоматического определения классов для данного запроса.

Процесс кластеризации документов реализуется следующим образом: выбирается очередной документ, формируется его поисковый образ и осуществляется поиск по документу-образцу. Так формируется первый класс. Далее процедура циклически повторяется для остальных документов, подлежащих кластеризации. Алгоритм завершает работу, когда будут сгруппированы все документы.

На входе алгоритма – запрос пользователя z и множество $T_{\text{вх.}}$ текстовых документов, на выходе – множество W классов текстовых документов. Алгоритм кластеризации включает в себя следующие шаги:

Шаг 1. $W := \varnothing$.

Шаг 2. Найти множество $R \subseteq T_{\text{вх}}$ документов, релевантных запросу z.

Шаг 3. Выбрать очередной документ Td_j из множества R.

Шаг 4. Осуществить поиск текстовых документов по документу-образцу Td_j (см. алгоритм в п. 4.1).

Шаг 5. Сформировать из множества найденных документов $\pi(Z)$ по документу-образцу Td_j класс и добавить его в W.

Шаг 6. Если все документы Td_j обработаны, то КОНЕЦ (множество W классов текстовых документов по запросу пользователя z сформировано), иначе перейти к шагу 3.

4.4. Алгоритм классификации документов

Целью задачи классификации документов является упрощение процесса поиска путем определения каждого документа к одной из заранее заданных категорий (тематик).

Процесс классификации заключается в сравнении очередного текстового документа с поисковым образом класса. В соответствии с принятым в системе критерием выдачи документ включается в тот класс, для которого значение этого критерия наибольшее.

На входе алгоритма – множество текстовых документов $T_{вх.}$ и поисковых образов классов Q, на выходе – множество W_T документов, распределенных по классам Q. Алгоритм классификации включает в себя следующие шаги:

Шаг 1. $W_T := \varnothing$.

Шаг 2. Выбрать очередной текстовый документ Td_j из множества $T_{вх.}$.

Шаг 3. Сформировать поисковый образ P_j документа Td_j, вычисляя веса слов по формуле (3).

Шаг 4. $z_j := P_j$.

Шаг 5. По запросу z_j сформировать динамический корпус текстов $\pi_{Cf}(z_j)$ путем поиска в полном корпусе текстов Cf документов, релевантных z_j.

Шаг 6. Индексировать динамический корпус текстов $\pi_{Cf}(z_j)$ как конкатенацию документов и получить расширенный поисковый запрос $Z = \omega^{-1}(\pi_{Cf}(z_j))$.

Шаг 7. Найти в множестве Q поисковый образ класса, релевантный Z_j.

Шаг 8. Добавить документ Td_j в соответствующий класс множества W_T.

Шаг 9. Если все текстовые документы Td_j обработаны, то КОНЕЦ (множество W_T документов, распределенных по классам Q, сформировано), иначе перейти к шагу 2.

Заключение

Предложенные в статье алгоритмы могут применяться для индексирования и поиска документов определенной тематики как в Интернете, так и в локальных базах данных. Благодаря использованию динамических корпусов текстов тематические алгоритмы обладают универсальностью, т. е. независимостью от объема документов, за счет расширения первоначального образа документа (запроса пользователя) релевантным ему динамическим корпусом текстов и могут применяться для индексирования и поиска как полнотекстовых документов, так и кратких сообщений. Предложенный метод вычисления весов ключевых слов на основе динамических корпусов текстов может использоваться в различных информационно-поисковых системах для повышения релевантности и полноты поиска текстовой информации.

Список литературы

1. Черный, А.И. Введение в теорию информационного поиска / А.И. Черный. – М. : Наука, 1975. – 238 с.

2. Солтон, Дж. Динамические библиотечно-информационные системы / Дж. Солтон. – М. : Мир, 1979. – 560 с.

3. Chakrabarti, S. Mining the Web. Discovery knowledge from hypertext data / S. Chakrabarti. – Publisher: Morgan Kaufmann, 2002. – 344 p.

4. Ландэ, Д.В. Поисковые системы: поле боя – семантика / Д.В. Ландэ // Телеком. – 2004. – № 4. – С. 44–50.

5. Pinkerton, B. Finding What People Want: Experiences with the WebCrawler / B. Pinkerton [Электронный ресурс]. – Mode of access : http://thinkpink.com/bp/WebCrawler/ WWW94.html. – Date of access : 4.08.2009.

6. Тактаев, С. Поиск информации в компьютерных сетях: новые подходы / С. Тактаев [Электронный ресурс]. – Режим доступа : http://www.searchengines.ru/articles/004603.html. – Дата доступа : 4.08.2009.

7. Технологии извлечения знаний из текста / Н. Ильин [и др.] // Открытые системы [Электронный ресурс]. – 2006. – № 6. – Режим доступа : http://www.i-teco.ru/article104.html. – Дата доступа : 4.08.2009.

8. Automatic keyword extraction using domain knowledge / A. Hulth [et al.] // Lecture notes in computer science. – 2006. – Vol. 3930/2006. – P. 633–641.

9. Липницкий, С.Ф. Семантический анализ текста на основе ситуативно-синтагматической сети / С.Ф. Липницкий // Информатика. – 2005. – № 2 (6). – С. 102–110.

10. Липницкий, С.Ф. Веб-поиск и аннотирование научно-технической информации на основе тематических корпусов текстов / С.Ф. Липницкий, А.А. Мамчич, С.А. Сорудейкина // Информатика. – 2009. – № 2 (22). – С. 114–125.

Поступила 21.09.09

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: lexamam@newman.bas-net.by

A.A. Mamchich

ALGORITHMS OF INDEXING AND SEARCHING DOCUMENTS ON THE BASIS OF DYNAMIC CORPORA

Thematic algorithms of indexing and search of documents on the basis of dynamic corpora are suggested. The algorithms can be applied for indexing and searching both full-text documents and short messages on the Internet or in local databases. A method is proposed for calculation of the weights of keywords on the basis of dynamic corpora, which can be used in various information retrieval systems to increase relevance and completeness of searching textual information.