НАУЧНЫЙ ОТДЕЛ

ИНФОРМАТИКА

УДК 81.32

ЗАДАЧИ ИНТЕРАКТИВНОЙ ОБРАБОТКИ ПОИСКОВЫХ ЗАПРОСОВ В ТЕОРЕТИКО-МНОЖЕСТВЕННОЙ ПОСТАНОВКЕ

Д.А. Бодров, С.Л. Кожитов, В.Н. Поляков

Московский государственный институт стали и сплавов (Технологический университет),
Кафедра АСУ
E-mail: polyakovvn@misis.ru

В работе представлено теоретико-множественное описание основных интерактивных механизмов обработки запросов в рамках проекта «Интеллектуальная поисковая машина»: фокусировка и расширение полноты. Новизна подхода заключается в том, что впервые рассматривается систематизация интерактивных методов разрешения многозначности в человеко-машинном комплексе «пользователь –- поисковая система». Инженерная реализация методов выполнена в нотации языка *SQL*.
Результаты исследований найдут применение в поисковых модулях систем документооборота предприятия, в библиотечных системах, в сети Интернет.

The Tasks of Interactive Processing of Search Queries in the Set Theory Formalization

D.A. Bodrov, S.L. Kozhitov, V.N. Polykov

In the paper description in the set theory of main mechanisms of query processing is fullfiled. Focusing and widening of fullnes are described as query processing. The methods are used in the project of Intelligent Search Engine. Newness of the approach consists that there is the first systematization of interactive methods of word sense disambiguation in man-machine complex «user-search engine». Engineer realization of the methods was done in SQL. This results can be used in the search modules of documents workflow systems of ERP, in bibliotic systems, in network Internet.

ВВЕДЕНИЕ

Рост объемов информации в сети Интернет и возможности доступа к библиотечным ресурсам средствами коммуникаций поставили проблему качественного поиска чрезвычайно остро. Несмотря на обилие поисковых систем в Интернете, современные технологии не предоставляют достаточно средств для организации эффективного поиска. Поэтому результат поиска по-прежнему больше зависит от уровня подготовленности самого пользователя, нежели от поисковой системы, что достаточно ярко иллюстрируют исследования (табл. 1) [1].

Из результатов исследования видно, что пользователи поисковых систем Интернета и открытых библиотечных ресурсов гораздо менее точно формулируют свои информационные потребности (используют 1–2 коротких запроса) и практически не используют возможностей языка запроса поисковой системы. Это можно объяснить недостаточно высокой подготовкой и опытностью пользователей, в отличие от пользователей, работающих с традиционными поисковыми системами. Можно предположить, что релевантность поиска неопытными пользователями в сети Интернет будет невысокой, так как невысокими будут полнота и точность поиска. Однако для пользователей библиотечных ресурсов это не вполне верно, так как им для поиска доступны и специальные поля, такие как автор, название, ключевые слова. Что в результате может привести к достаточно высокой релевантности даже с использованием

коротких запросов и без использования расширенных возможностей поиска. Проблема улучшения качества работы поисковых систем тесно связана с фундаментальными задачами обработки естественного языка, которыми занимается компьютерная лингвистика. Так, например, хорошо известно, что неполнота поиска тесно связана с обработкой синонимии в различных ее аспектах, включая не только лексическую синонимию (*помидор — томат*), но и синонимические перифразы (*обучение — образовательный процесс*), аббревиатуры (*высшее учебное заведение — вуз*), использование терминов с расширительным толкованием (*гипоним — гипероним: студент — учащийся*). Другая актуальная проблема улучшения качества поисковых систем, проблема информационного шума, тесно связана с такими явлениями языка, как омонимия и полисемия[1]. Полисемия обычно трактуется как разновидность более широкого явления, многозначности в языке, на лексическом уровне. Проблеме лексической многозначности посвящено сравнительно много исследований в области компьютерной лингвистики. Наиболее полный обзор подходов и алгоритмов, сопровождаемый историческим экскурсом, представлен в работе [2]. В работе [3] приведена альтернативная классификация методов разрешения многозначности, основанная на видах используемой информации. В работе [4] представлен концептуальный проект информационной поисковой системы, построенный на систематических принципах разрешения многозначности, переформулирования запросов с целью расширения результатов поиска или снижения информационного шума. Традиционно разрешение лексической многозначности проводится на этапе индексирования текста, когда составляется индекс, содержащий ссылки не только на слово, но и его значение. Система значений в этом случае представляется в виде специализированной семантической сети по типу лексических онтологий [6,7]. В работе [5] впервые был поставлен вопрос систематизации на научной основе интерактивных методов переформулирования запросов в интерфейсном модуле поисковой системы. В настоящей работе представлена логико-математическая интерпретация базовых операций по фокусировке запросов, их расширению и переформулированию с целью улучшения поиска. Инженерная реализация методов выполнена в нотации языка *SQL*.

Сравнение типичных поисковых сессий в разных категориях

Характеристика	Поиск в Интернете	Традиционная ИПС	Библиотечные ресурсы
Длина сессии (число запросов пользователей в поисковой системе)	1–2	7–16	2–5
Длина запроса (число слов в запросе)	2	6–9	1–2
Число релевантных документов, просмотренных за сессию	не более 10	около 10	менее 50
Использование расширенных возможностей (с использованием дополнительных опций языка запросов)	9 %	9 %	8 %
Использование логических операций (число запросов с использованием логических операций)	8 %	37 %	1 %
Частота отказов (частота неправильно сформулированных запросов)	10 %	17 %	7–19 %

1. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ

Опишем исходные множества[2].

- множество документов $D = \{d_1, d_2, ..., d_n\}$,
- множество лексем $L = \{l_1, l_2, ..., l_m\}$,
- множество значений $I = \{i_{11}, i_{12}, ..., i_{m1}, i_{mk}\}$,
- поисковый запрос $Z = \{z_1, z_2, ..., z_v\} \subset L$, z_j — ключевая лексема.

Введем определения для исходных отношений:

- пословный индекс $R_1(D, L) : (l_j, d_f) \in R_1(D, L) \leftrightarrow$ лексема l_j содержится в документе d_f,
- толковый словарь $R_2(I, L) : (i_{jk}, l_j) \in R_2(I, L) \leftrightarrow$ значение i_{jk} относится к лексеме l_j,

[1]Омонимия — одинаковое написание различных морфологических форм одного и того же или различных слов (*стали (глагол) — стали (существительное)*). Полисемия — наличие нескольких значений одного и того же слова: (*ревун — обезьяна и маяк*).

[2]Лексема — слово в словаре. Значение —- смысловое содержание слова. Для многозначных слов одной лексеме соответствует несколько значений.

• индекс по значениям $R_3(D,I)$: $(d_f, l_j) \in R_3(D,I) \leftrightarrow$ лексема l_j содержится в документе d_f в значении i_{jk}.

«И-запрос» представляет собой поиск пересекающихся множеств документов по двум и более ключевым словам. При этом выполняются шаги 1–4.

Шаг 1. Строится сечение $R(z_1) = \{z_{1z}\} \subset I$ на отношении $R_2(I,L)$.

Шаг 2. Строится сечение каждого элемента $R(z_1)$ — множества $R(z_{1z}) = \{d_{zd}\} \subset D$ на отношении $R_3(D,I)$.

Шаг 3. Создается множество $M\{m_d\} : m_d = R(z_{1z}) \cap \bigcap_{u=2}^{v} R(z_u) \subset D$, где $R(z_u) \subset D$ — сечение отношения $R_1(D,L)$ по z_u.

Шаг 4. Осуществляется выбор пользователем одного из элементов множества M, обозначенного далее $M_p \subset D$, где p — номер значения из подмножества $I^{z1} = \{i_1, ...i_p, ...i_r\}, I^{z1} \subset I$.

Ниже приводятся соотношения, которые активируют сценарии

• фокусировки запроса

$$|M_p| > 10, \tag{1}$$

• расширения полноты поиска

$$|M_p| < 10. \tag{2}$$

При выполнении И-запроса разрешение многозначности выполняется для ядерной лексемы. Будем для простоты полагать, что ядерной лексемой является первое ключевое слово в запросе z_1. Сечение $R(z_1)$ представляет собой выборку — множество значений для ключевого слова z_1. Сечение $R(z_{1z})$ представляет собой выборку — множество документов в БД, которые включают ключевое слово z_1. Множество M описывает результаты обработки запроса поисковой системой с условием вхождения слова z_1 при одновременном вхождении остальных слов из Z, сгруппированные по значениям $1 \ldots r$ слова z_1. M_p — подмножество документов M, включающих лексему z_1 в значении р. В зависимости от мощности этого множества выполняются два сценария: фокусировки, т.е. сужения полноты поиска (при условии (1)) и расширения полноты (при условии (2)).

Фокусировка представляет собой операцию по формированию подмножества M путем использования следующих ассоциативных и семантических связей:

– между словом и другими терминами из данной предметной области (тематический кластер) (*сталь –> {домна, прокат}*);

– между словом и его словосочетаниями (*интеллект –> искусственный интеллект*);

– между словом и потенциальными вопросами, на которые может ответить данный текст (*МИСиС –> Как поступить в МИСиС?*);

– между словом и областью деятельности, сферой интересов, которую разработчики ИПМ назвали коммуникативным кластером (*ЕГЭ –> обучение в вузе*).

Операция фокусировки может быть сформулирована следующим образом.

Множества:

• множество тематических кластеров $K = \{k_k\}$,

• множество словосочетаний $E = \{e_p\}$,

• множество вопросов $Q = \{q_q\}$,

• множество коммуникативных кластеров $C = \{c_c\}$.

Отношения:

• индекс по тематическим кластерам $R_K(K,D)$ — кластер k_k содержится в документе d_i,

• индекс по словосочетаниям $R_E(E,D)$ — словосочетание e_p содержится в документе d_i,

• индекс по вопросам — вопрос q_j относится к документу d_i,

• индекс по коммуникативным кластерам $R_C(\ C\ , D)$ — кластер c_r (или его часть) содержится в документе d_i,

• онтологическая связь $R_O(I,I)$ — значение $i_{j1,l1}$ состоит в онтологической связи со значением $i_{j2,l2}$.

Расширение полноты поиска как операция переформулирования запроса (переход от M_1 к M_2) используется в том случае, если на запрос пользователя интеллектуальная поисковая система выдала незначительное количество документов ($|M| < 10$). К основным механизмам расширения полноты поиска относятся:

– переход от слова к словообразовательной парадигме (*борт –> бортпроводник*);

– переход от слова к синонимическому ряду (*бегемот –> гиппопотам*);

– переход от аббревиатуры к ее расшифровке, и наоборот (*МИСиС –> Московский институт стали и сплавов*)

Кроме того, существуют еще возможности переформулирования запросов по элементам онтологического дерева (Пример: *самолет –> когипонимы = виды транспорта: поезд, метро, трамвай, троллейбус, автобус*).

Операция расширения может быть сформулирована следующим образом.

Отношения:

• словообразовательная парадигма $R_W(L, I)$ — лексема l_j является морфологическим дериватом другой лексемы со значением i_{jl},

• синонимический ряд $R_S(L, L)$ — лексема l_i является синонимом лексемы l_j,

• аббревиатура $R_A(L, E)$ — лексема l_j является аббревиатурой словосочетания e_e,

• онтологическая связь $R_O(I, I)$ — значение $i_{j1,k\,1}$ состоит в онтологической связи со значением $i_{j2,k\,2}$.

Расширение в общем виде $R(x_x) : \ (x_x, y_y) \in R(X, Y), \quad R(x_x) \subset Y, \quad R(x_x) \cup R(z_1)$.

Далее представлена задача интерактивного поиска как задача принятия решения. Под пертинентностью документа мы понимаем субъективную оценку пользователем этого документа с точки зрения удовлетворения его информационной потребности.

Найти такой поисковый запрос S^*, который обеспечит выполнение следующих критериев:

$$\max_{s \subset S} P(s), \quad \max_{s \subset S}(-|M(s)|), \quad S = \{s : \ S \in \Omega, \ P(s) > 0, \ M(s) > 0\},$$

$P(s)$ — суммарная пертинентность множества найденных документов, $M(s)$ — множество результатов (найденных документов), S — множество результативных запросов, Ω — множество всех запросов.

Нечеткие лингвистические переменные — критерии

$$\langle \text{ПЕРТИНЕНТНОСТЬ}, T(L), [0, 1], G, H \rangle,$$

где $T(L) = \{$непертинентно, среднепертинентно, пертинентно$\}$ — терм-множество; G — процедура образования новых термов с помощью связок и модификаторов типа «очень», «слегка», «совсем», «не» и др. *Например: «малопертинентно»*; H — процедура задания на множестве $[0, 1]$ нечетких подмножеств, выполняемая пользователем в процессе работы с поисковой системой.

$$\langle \text{ЧИСЛО РЕЗУЛЬТАТОВ}, T(L), [0, |D|\,], G, H \rangle,$$

где $T(L) = \{$мало, много$\}$ — терм-множество; G — процедура образования новых термов с помощью связок и модификаторов типа «очень», «слегка», «совсем», «не» и др. *Например: «слишком много»*; $|D|$ — мощность множества всех документов; H — процедура задания на множестве нечетких подмножеств, выполняемая пользователем в процессе работы с поисковой системой.

На рис. 1–2 представлены алгоритм обработки запросов и обобщенный алгоритм сценария поиска.

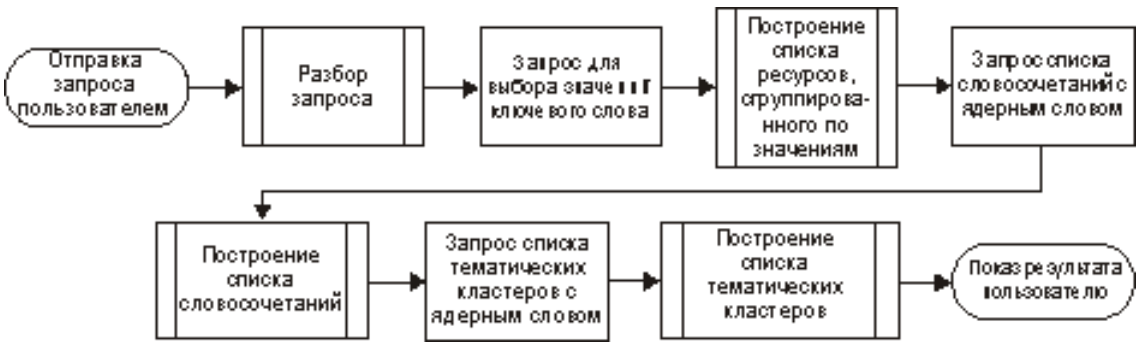

Рис. 1. Алгоритм обработки запросов в Интернете

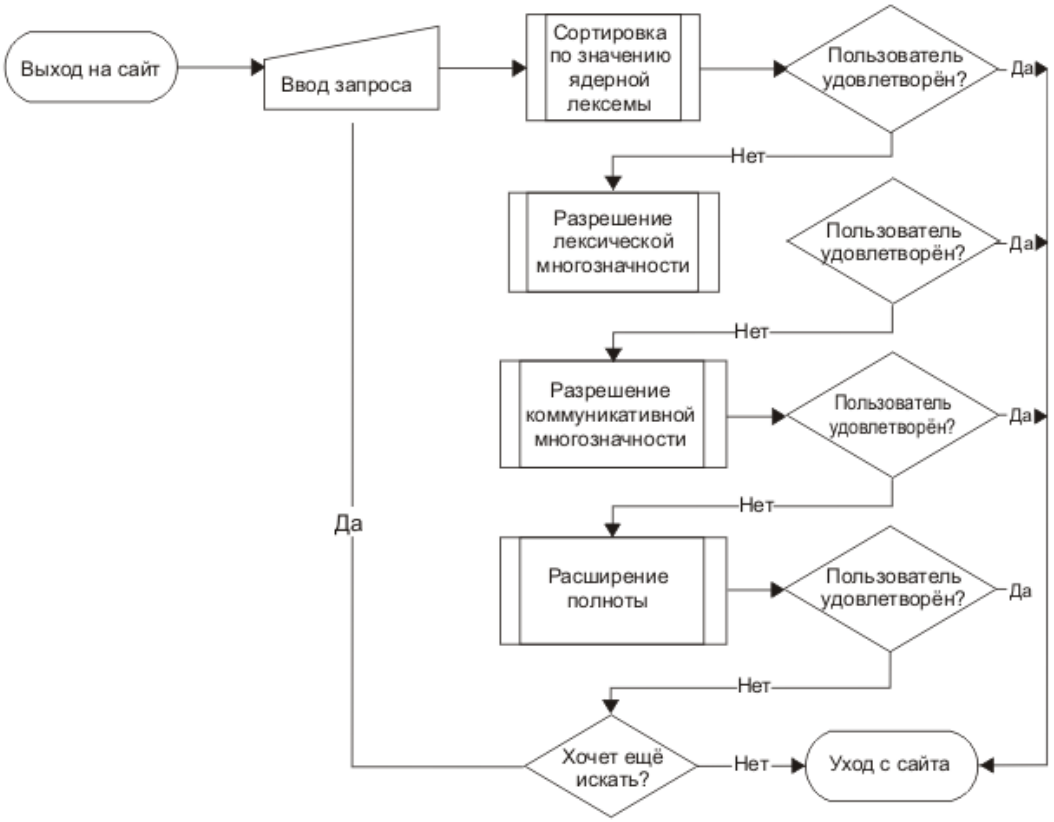

Рис. 2. Обобщенный алгоритм сценария поиска в Интернете

ЗАКЛЮЧЕНИЕ

В работе представлено теоретико-множественное описание основных интерактивных механизмов обработки запросов в рамках проекта «Интеллектуальная поисковая машина»: фокусировка, расширение полноты. Новизна подхода заключается в том, что впервые рассматривается систематизация интерактивных методов разрешения многозначности в человеко-машинном комплексе «пользователь – поисковая система». Инженерная реализация методов выполнена в нотации языка *SQL*. Результаты исследований найдут применение в поисковых модулях систем документооборота предприятия, в библиотечных системах, в сети Интернет.

Работа выполнена при частичной финансовой поддержке РФФИ (*проект* 05-07-90339).

Библиографический список

1. *Jansen B.J., Pooch U.* Web user studies: A review and framework for future work // J. of the Amer. Society of Information Science and Technology. 2000. V. 52(3). P. 235–246.
2. *Ide N., Veronis J.* Word Sense Disambiguation: The State of the Art. // Computational Linguistics. 1998. V. 24, № 1. P. 1–40.
3. *Поляков В.Н.* Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации // Проблемы прикладной лингвистики. Вып.2. Сборник статей /Отв. ред. Н.В. Васильева. М.: Азбуковник, 2004. С. 101–117
4. *Поляков В.Н.* Интеллектуальная поисковая машина. Концептуальный проект // Труды Казан. шк. по комп. и когнитивной лингвистике. Казань: Сэлэт, 2000. № 5.
5. *Бодров Д.А., Поляков В.Н., Точин А.В.* Интерактивные методы фокусировки и расширения поиска в интеллектуальной поисковой машине. М., 2002.
6. *Miller G.A.* WordNet: a lexical database for English // Communications of the ACM 38. 1995. № 11. http://www.acm.org/pubs/articles/journals/ cacm/1995-38-11/p39-miller/p39-miller.pdf.
7. *Поляков В.Н.* Проект WordNet и его влияние на технологии компьютерной и когнитивной лингвистики. М., 2003.
8. *Фор Р., Кофман А., Дени-Папен М.* Современная математика/ Пер. с фр.; Под ред. А.Н.Колмогорова. М.: Мир, 1966. 273 с.