УДК 025.4.036ДИАНА

Б. А. Кузнецов, Е. К. Солнцева, М. В. Деревянкин, Д. В. Закамская

Обработка запросов на естественном языке — новое качество поиска в БД ВИНИТИ

Рассматривается применение принципиально новой поисковой системы — ДИАНА — для обработки запросов на естественном языке в БД ВИНИТИ. В качестве примера приводятся сравнительные результаты обработки запроса традиционными средствами диалогового поиска и автоматически с помощью ДИАНЫ. Отмечается, что автоматическая обработка запроса на естественном языке позволяет получить существенно более высокую полноту поиска. Поиск традиционными диалоговыми средствами давал в несколько раз меньше документов, при этом многие документы с высокой степенью релевантности оказывались пропущенными.

Анализируются основные принципы обработки запросов на естественном языке, распознавание слов в тексте запроса, ранжирование выдачи, терминологическая навигация по смысловым аспектам полученного результата, методы создания проблемных БД по "словесному портрету" проблемы. Система способна с хорошим временем реакции искать информацию по сложным запросам из 10–20 слов в представительных ретроспективных массивах за много лет.

ВВЕДЕНИЕ

Традиционные средства поиска в библиографических и реферативных базах данных в основном опираются на два метода.

Первый метод — использование классификации документальных источников по темам. В ВИНИТИ, например, имеется развитая система рубрикаторов, которая позволяет подробно делить информационный поток документов как в реферативном журнале, так и в БД по многочисленным широким и узким тематическим областям. Классификационное деление документов помогает сузить область поиска соответствующими разделами. Некоторые пользователи считают, что в ряде случаев поиск подходящей узкой рубрики может оказаться лучше, чем поиск по запросу из набора ключевых слов. Достаточно лишь знать, в какой рубрике и какого уровня помещаются соответствующие документы на заданную тему.

Второй метод ориентирован на прямой диалоговый поиск по ключевым словам или терминам, содержащимся в текстах документов. Пользователь формулирует булево выражение из слов запроса — ему выдаются документы, отвечающие формуле запроса. Этот метод используется для поиска во многих информационных системах, в том числе и для поиска в БД ВИНИТИ. Нередко на практике применяют оба метода — сначала выбирают БД по интересующей тематической области, а затем проводят более тонкий поиск в выбранной БД.

В настоящей статье рассмотрены возможности поиска в БД с помощью принципиально новой ИПС — ДИАНА [1–3], позволяющей довольно быстро искать в БД ВИНИТИ не по традиционным булевым формулировкам, а по текстовым запросам на обычном разговорном естественном языке, который используется на практике, когда пользователь поручает специалисту найти интересующие его документы. При этом никаких ограничений на формы представления слов, применяемый синтаксис и способы формулирования не накладывается. Система не имеет серьезных ограничений и на длину запроса — в нем может быть 10, 20 и более слов. Ни одно слово запроса не отвергается при проведении поиска. Широкие возможности системы позволяют применять на практике даже такие "экзотические" виды поиска, как поиск без всякой предварительной обработки по вырезанному "мышью" фрагменту любого документа, который показался пользователю интересным с точки зрения темы запроса. По каждому запросу система автоматически формирует множество смысловых аспектов в виде терминологических композиций из терминов запроса. Смысловые аспекты ранжированы по степени важности. Каждый смысловой аспект (ранг) может включать от одного до сотен документов в зависимости от сложности терминологической композиции. Все смысловые аспекты одновременно доступны для обозрения и выбора. Пользователь отмечает интересующий набор смысловых аспектов, формируя выдачу по множеству интересующих его аспектов запроса.

Новые свойства системы дают возможность пользователю кардинально изменить поведение при формировании и модификации запроса. Если в традиционной ИПС пользователь начинает сеанс "отладки" запроса с одного, двух, максимум трех слов, боясь получить нулевой результат или мало документов на излишне многословный запрос, то в ДИАНЕ все наоборот — лучше, не стесняясь, написать сразу подробный запрос, включая любой

набор терминов, которые хоть в какой-то степени отражают те или иные стороны информационной потребности пользователя. В результате пользователю нет нужды в течение длительного сеанса поиска многократно подбирать "подходящую" формулировку запроса — вместо этого достаточно просмотреть перечень смысловых аспектов, отвечающих исходному запросу на естественном языке, и отобрать то, что нужно в интегральный результат поиска.

Одновременно легко решается проблема выбора между полнотой и точностью. Если нужно немного документов, но в максимальной степени отвечающих содержанию запроса, достаточно выбрать интересные смысловые аспекты только высших рангов (попутно оценивается суммарное количество документов, которое пользователь согласен получить в качестве результата). Если необходима высокая полнота, продолжается просмотр смысловых аспектов средних и нижних рангов, где терминологические композиции содержат меньшее число терминов. Таким образом полнота увеличивается, но растет вероятность появления шумовых документов. А теперь, после краткого введения, остановимся на основных моментах более подробно.

ЧТО ДАЕТ ПОИСК ПО ЗАПРОСАМ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ?

Действительно, а дают ли новые средства поиска практический эффект? В поисковом сервисе находится большое число БД. Во всех информационных системах, в том числе и доступных через ИНТЕРНЕТ, используются испытанные механизмы диалогового поиска. Пользователи знают, как маскировать окончания слов, как применять операторы булевой алгебры или контекстуальные операторы, как расширить формулировку запроса, если получен нулевой ответ, как сузить запрос, если документов слишком много. Может быть никаких других средств на самом деле и не нужно? Попытаемся разобраться.

Проще всего понять проблемы обычного диалогового поиска, обработав конкретный запрос разными способами. Мы выбрали для эксперимента один из таких реальных запросов к БД ВИНИТИ: **"стоимость экологически чистой водопроводной питьевой воды с ее очисткой от загрязнений до ПДК"**. (ПДК — это предельно допустимые концентрации вредных веществ в окружающей среде по принятым нормам). Запрос показался интересным для иллюстрации процесса его итерационной отладки обычным последовательным набором пробных булевых формулировок.

Стартовый пробный поиск по формулировке **"очистка AND вода"** (это означает, что в документе — библиографической записи или реферате — должны быть обязательно оба слова) дал результат более 25 тыс. записей. Такое большое число записей по широкой формулировке создало у пользователя впечатление, что по любым узким формулировкам будет легко получить вполне представительное число записей.

Пользователю было предложено самостоятельно попытаться обработать исходный запрос в традиционной ИПС путем последовательных модификаций. Исходный запрос был преобразован в следующий набор слов: **"стоимость, экологически, чистая, питьевая, водопроводная, вода, очистка, загрязнения, ПДК"**.

Для упрощения эксперимента в качестве оператора пересечения слов был выбран наиболее общий — **AND** (слова, связанные этим оператором, могут встретиться в любом месте документа на любом расстоянии друг от друга — другие операторы вводят те или иные ограничения на взаимное расположение слов или расстояние между ними). Чтобы сделать результаты эксперимента сопоставимыми и сократить число ошибок из-за неверного распознавания слов в различном падеже и числе, вместо механизма усечения окончаний слов в традиционном поиске был использован механизм автоматического морфологического анализа и синтеза слов, принадлежащих общей основе. Постепенно расширяя запрос по своему сценарию (основное желание было при последовательном удалении тех или иных слов сохранить термин "стоимость"), пользователь получил следующую серию подзапросов, которые давали нулевой результат:

1. стоимость AND экологически AND чистая AND питьевая AND водопроводная AND вода AND очистка AND загрязнения — (удалено слово *ПДК*).

2. стоимость AND экологически AND чистая AND питьевая AND водопроводная AND вода AND очистка — (удалено слово *загрязнения*).

3. стоимость AND экологически AND чистая AND питьевая AND водопроводная AND вода — (удалено слово *очистка*).

4. стоимость AND экологически AND чистая AND водопроводная AND вода — (удалено слово *питьевая*).

Следующие итерации дали ненулевые результаты

5. стоимость AND чистая AND водопроводная AND вода — (удалено слово *экологически*), результат — 1 запись.

6. стоимость AND водопроводная AND вода — (удалено слово *чистая*), результат — 54 записи.

Другая серия итераций с сохранением термина "ПДК" дала следующий набор подзапросов с нулевыми результатами:

7. экологически AND чистая AND питьевая AND водопроводная AND вода AND очистка AND загрязнения AND ПДК — (удалено слово *стоимость*).

8. экологически AND чистая AND питьевая AND вода AND очистка AND загрязнения AND ПДК — (удалено слово *водопроводная*).

Следующие итерации дали ненулевые результаты:

9. экологически AND чистая AND вода AND очистка AND загрязнения AND ПДК — (удалено слово *питьевая*), результат — 1 запись.

10. экологически AND чистая AND вода AND очистка AND ПДК — (удалено слово *загрязнения*), результат — 3 записи.

11. чистая AND вода AND очистка AND ПДК — (удалено слово *экологически*), результат — 8 записей.

12. чистая AND вода AND ПДК — (удалено слово *очистка*), результат — 46 записей.

Эксперимент показывает, что даже, когда документов по теме в БД много, поиск по реальному многоаспектному запросу традиционными средствами превращается в трудоемкую операцию с результатами, которые трудно квалифицировать. Можно ли, например, считать по итогам сеанса отладки запроса результаты поиска по подзапросам 6 и 12 лучшими? Очевидно, что приведенные варианты отладки запроса далеко не единственные и при наличии у пользователя фантазии и времени он мог бы подобрать еще и не один десяток похожих формулировок. Вопрос заключается в том, сколько времени пользователь согласен потратить на такой диалоговый сеанс и что будет служить для него критерием его завершения?

После проведенного эксперимента по традиционному диалоговому поиску было решено обработать исходный запрос средствами ДИАНЫ без модификаций автоматически. В результате был получен набор, включающий 152 смысловых аспекта, ранжированных в порядке важности — самые релевантные, содержащие больше терминов запроса идут первыми, далее размещаются более широкие аспекты с меньшим числом терминов и с большей частотой встречаемости в массиве документов. Ниже приводится ранжированный список первых десяти из 152 смысловых аспектов.

1. стоимость экологически чистая питьевая вода очистка загрязнения [2/2]
2. стоимость экологически чистая вода очистка ПДК [1/1]
3. экологически чистая вода очистка загрязнения ПДК [1/1]
4. стоимость экологически питьевая вода очистка ПДК [1/1]
5. экологически чистая питьевая вода очистка загрязнения [1/3]
6. экологически чистая питьевая водопроводная вода загрязнения [1/1]
7. экологически чистая питьевая вода загрязнения ПДК [1/1]
8. стоимость питьевая водопроводная вода очистка загрязнения [1/1]
9. стоимость питьевая водопроводная вода загрязнения ПДК [1/1]
10. чистая питьевая водопроводная вода очистка загрязнения [2/2]

Слева указан номер ранга — смыслового аспекта, идентифицированного уникальной терминологической композицией из слов исходного запроса на естественном языке. Справа в скобках указаны два параметра: слева от косой черты указано число документов БД, которые включают указанную терминологическую композицию. Эти документы содержат перечисленный набор слов, но в них должны обязательно отсутствовать любые другие слова из запроса. Параметр справа от косой черты указывает общее число документов во всех рангах, которые включают указанную терминологическую композицию. Например, терминологическая композиция ранга 5 в "чистом виде" содержится в одном документе, но всего их три — еще два документа находятся в 1-м ранге, где содержится еще одно слово запроса — *стоимость*. Анализируя список смысловых аспектов, можно отобрать из них в результирующую выдачу те, которые представляют интерес. Ниже приводится перечень выбранных смысловых аспектов, связанных с термином *стоимость*.

1. стоимость экологически чистая питьевая вода очистка загрязнения [2/2]
2. стоимость экологически чистая вода очистка ПДК [1/1]
4. стоимость экологически питьевая вода очистка ПДК [1/1]
8. стоимость питьевая водопроводная вода очистка загрязнения [1/1]
9. стоимость питьевая водопроводная вода загрязнения ПДК [1/1]
15. стоимость экологически чистая питьевая вода [2/4]
16.* стоимость чистая питьевая водопроводная вода [1/1]
17. стоимость чистая питьевая вода загрязнения [1/3]
19. стоимость экологически питьевая вода очистка [4/7]
22. стоимость водопроводная вода очистка загрязнения [2/3]
26. стоимость питьевая водопроводная вода очистка [7/8]
27. стоимость питьевая вода очистка загрязнения [11/14]
29. стоимость питьевая вода очистка ПДК [2/3]
52. стоимость экологически вода ПДК [1/3]
53. стоимость чистая питьевая вода [3/9]
58. стоимость водопроводная вода очистка [9/19]
59. стоимость вода очистка ПДК [3/7]
60. стоимость водопроводная вода загрязнения [2/6]
65. стоимость питьевая вода очистка [28/56]
68. стоимость питьевая водопроводная вода [8/18]
69. стоимость питьевая вода загрязнения [25/41]
72. стоимость питьевая вода ПДК [2/6]
101. стоимость питьевая водопроводная [1/19]
102.* стоимость водопроводная вода [23/54]
103. стоимость питьевая загрязнения [1/42]
110. стоимость питьевая вода [114/218]

Выборка смысловых аспектов, связанных с термином *ПДК* выглядит следующим образом

1. стоимость экологически чистая вода очистка ПДК [2/2]
3.* экологически чистая вода очистка загрязнения ПДК [1/1]
4. стоимость экологически питьевая вода очистка ПДК [1/1]
7. экологически чистая питьевая вода загрязнения ПДК [1/1]
9. стоимость питьевая водопроводная вода загрязнения ПДК [1/1]
11. чистая питьевая водопроводная вода загрязнения ПДК [2/2]
13. экологически питьевая вода очистка загрязнения ПДК [4/4]
14. питьевая водопроводная вода очистка загрязнения ПДК [2/2]

20.* экологически чистая вода очистка ПДК [1/3] 1
21. экологически чистая вода загрязнения ПДК [9/11]
28. чистая вода очистка загрязнения ПДК [2/3]
29. стоимость питьевая вода очистка ПДК [2/3]
30. чистая водопроводная вода загрязнения ПДК [1/3]
33. экологически вода очистка загрязнения ПДК [8/13]
37. чистая питьевая водопроводная вода ПДК [3/5]
39. чистая питьевая водопроводная вода загрязнения ПДК [2/5]
42. экологически питьевая водопроводная вода ПДК [1/1]
43. экологически питьевая вода загрязнения ПДК [12/17]
44. водопроводная вода очистка загрязнения ПДК [2/4]
46. питьевая водопроводная вода очистка ПДК [2/4]
47. питьевая вода очистка загрязнения ПДК [17/23]
48. питьевая водопроводная вода загрязнения ПДК [11/16]
52. стоимость экологически вода ПДК [2/3]
57. экологически чистая вода ПДК [6/19]
59. стоимость вода очистка ПДК [3/7]
62. экологически очистка вода ПДК [6/19]
66.* чистая вода очистка ПДК [3/8]
71. чистая вода очистка загрязнения ПДК [7/25]
72. стоимость питьевая вода ПДК [2/6]
73. экологически вода очистка ПДК [13/29]
77. экологически вода загрязнения ПДК [56/93]
83. водопроводная вода очистка ПДК [1/7]
85. водопроводная вода загрязнения ПДК [4/25]
88. питьевая вода очистка ПДК [40/68]
90. питьевая водопроводная вода ПДК [11/33]
91. питьевая вода загрязнения ПДК [78/130]
111.* чистая вода ПДК [7/46]
114. экологически питьевая ПДК [1/28]
119. водопроводная загрязнения ПДК [1/26]
125. водопроводная вода ПДК [6/49]
130. питьевая вода ПДК [108/308]

Звездочкой указаны ранги, которые совпадают с найденными в процессе традиционного диалогового поиска. Включение в ранг **16** дополнительного термина *питьевая* не должно смущать — ДИАНА обнаружила этот термин в документе, в то время как пользователь не указал его в формуле запроса (в выдаче представлен один и тот же документ). По аспектам, связанных с термином *стоимость* оказалось выбрано 256 документов в 26 рангах. Заметим, что в традиционном диалоговом режиме было в конечном итоге найдено 54 документа — 21% от указанного выше числа. По аспектам, связанных с термином *ПДК* было выбрано 493 документа в 41 ранге; в традиционном диалоговом режиме было найдено 46 документов — 9% от этого значения. Более высокая полнота поиска средствами ДИАНЫ, как показывает сравнительное изучение терминологических композиций выбранных рангов и формулировок традиционных диалоговых подзапросов, вызвана не тем, что в смысловых аспектах меньше терминов, скорее наоборот. Можно, например, оставить в выдаче только те ранги, которые содержат не менее пяти слов запроса. Тогда для смысловых аспектов, связанных с термином *ПДК*, выборка ограничивается 48-м рангом и выдача составит 86 документов — много больше, чем выдача по полученной "вручную" пользовательской формулировке (ранг 111), где в запросе всего три слова.

ВОЗМОЖНОСТИ ТРАДИЦИОННЫХ СРЕДСТВ ДИАЛОГОВОГО ПОИСКА

Обычно утверждается, что разнообразие средств ИПС традиционного диалогового поиска предназначено для того, чтобы пользователь быстрее и эффективнее сумел бы отладить свой запрос в процессе уточнения формулировки. При этом то обстоятельство, что пользователь сам принимает решение на различных этапах диалогового поиска воспринимается, порой, как некая гарантия соответствия отлаженного запроса информационной потребности. В традиционном диалоговом поиске предполагается, что пользователь, ведя сеанс поиска, выбирает разумную стратегию отладки запроса, применяет те или иные операторы и термины и в конце концов выходит на ту формулировку запроса, которая отвечает требованиям полноты и точности поиска. Для этого пользователю предоставляется определенный набор средств обработки слов и составления формулировок запроса.

Рассмотрим эти средства. Чтобы ввести в запрос однокоренные слова, представленные в разных словоформах, используются средства так называемого усечения слов. Например, если необходимо, как это было в нашем примере, представить различные варианты слов *вода* и *водопроводная*, можно задать это в следующем виде: **"вод$"**, что означает — отыскиваются любые слова, имеющие начало *вод* с любым продолжением. Разумеется различные варианты слова *вода* (*водой, водами* и т. д.) и слова *водопроводная* (*водопроводного, водопроводных* и т. д.) будут найдены. Однако в сферу поиска, к сожалению, будут включены и другие слова — *водоросли, водитель, водка, водород* и т. д. В ИПС, правда, имеется средство, которое помогает в ряде случаев избежать таких неприятностей — указание так называемой глубины усечения. Например, можно написать для слова *вода* — **"вод$3"**, что означает — окончание не может быть больше трех символов. Однако указанное средство не позволяет все же исключить шумовое слово *водка*. К сожалению, в русском языке есть немало слов, где использование глубины усечения дает совсем мало эффекта — *корень, чай, белок* и т. д.

Чтобы учесть различные требования точности при составлении формулировки запроса предлагается множество операторов для связи слов запроса. Если написать в запросе — (**водопроводн$3**

ADJ вод$3), то указанные слова будут отыскиваться только тогда, когда они следуют контактно расположенными в указанном порядке. Но в этом случае будет отвергнуто словосочетание *водопроводная питьевая вода*. Тогда имеется возможность применить оператор **WITH**, допускающий расположение слов свободно в тексте, но с ограничением расстояния — в пределах одного предложения — (**водопроводн$3 WITH вод$3**). Однако и здесь не исключены потери. Например, такой вполне релевантный фрагмент (взят из реальной БД) будет пропущен: ... *водопроводная сеть расширяется. Вода к концу года начнет поступать и на эту бывшую окраину города* ... .

Оператор **OR** используется для объединения синонимичных или ассоциативно близких по значению терминов с целью увеличения полноты выдачи. Однако создание таких "синонимичных" групп в общем случае очень нетривиальное занятие даже для опытного пользователя. Проиллюстрируем это на примере нашего запроса: *стоимость, экологически, чистая, питьевая, водопроводная, вода, очистка, загрязнения, ПДК*. Терминологическая комбинация *экологически чистая, питьевая водопроводная вода*, очевидно, содержит ассоциативно близкие термины, но как и корректно представить? Можно, например, сделать это таким образом: **"(экологическ$3 OR чист$$3 OR питьев$3 OR водопроводн$3) AND вод$3"**. Это означает, что отыскивается слово *вода* в сочетании с любым из перечисленных определителей. Однако в этом случае в результат поиска попадает значительное число документов с такими фрагментами текста: *сточные воды завода нарушили экологический баланс региона, масло остается в камере, а чистая вода стекает в патрубок, ... питьевая сода растворена в воде... , водопроводный кран не выдержал напора воды*. Можно, конечно, использовать вместо **AND** другой оператор — **ADJ**, но тогда резко возрастут потери при поиске. Еще труднее придумать, как объединить имеющие между собой ассоциации термины: *очистка, загрязнения, ПДК*.

Приведенные примеры показывают, что разнообразие средств, используемых для обработки слов и формулировки запроса, требует большой аккуратности применения и пользователь далеко не всегда может предвидеть отрицательные последствия принятого решения при выборе тех или иных операторов информационно-поискового языка. Основная беда заключается в том, что пользователь может судить о качестве только тех документов, которые найдены по формулировке запроса, но он практически ничего не знает о тех документах, которые потеряны при неудачной стратегии поиска.

ПОДГОТОВКА И ОБРАБОТКА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ЗАПРОСОВ

Как уже указывалось, никаких особых требований на формулировку запроса на естественном языке не предъявляется. Такая свобода позволяет выходить из ситуаций, которые в традиционных системах кажутся непреодолимыми. Один из пользователей-медиков столкнулся с трудностями при обработке очень простого запроса — *веко* (найти документы, в которых говорится об этом). На этот запрос, естественно, выдавалось целое море документов, не имеющих никакого отношения к его теме. Обычные советы внести уточнения в запрос в данном случае ни к чему не привели. Внесенное уточнение скорее усилило парадоксальность ситуации — запрос получил формулировку *болезнь века*. При поиске в ДИАНЕ ситуация была разрешена тривиально простым способом. Совместно с пользователем запрос был искусственно дополнен набором терминов, определяющих предметную область — *воспаление, зрачок, глаз, кожа, набухание, офтальмолог, покраснение, зрение, ресницы*. Далее из списка смысловых аспектов были отобраны те, которые содержат термин *веко* в окружении нескольких из указанных терминов. В результирующей выдаче оказалось мало документов, не имеющих отношения к теме. Следует отметить, что доступные решения для традиционных диалоговых систем не дали желаемого результата. Например, когда был задан запрос в виде **веко AND (воспаление OR зрачок OR глаз OR кожа OR набухание OR офтальмолог OR покраснение OR зрение OR ресницы)**, в выдаче оказалось огромное число шумовых записей с такими, например, фрагментами ... *ведущие офтальмологи считают, что в 21-м веке...*; ... *с точки зрения специалистов проблема не решалась веками; производство кожи в стране снизилось к концу века* и т. п.

По сравнению с другими системами в ДИАНЕ очень просто реализуются так называемые функции отрицания (при этом никаких булевых операторов отрицания употреблять не надо). Приведем пример. Допустим вы задали запрос: **"Загрязнение воды в Волге"**, но вам хотелось бы исключить из результата такие уже известные вам виды загрязнителей, как, скажем, *фенол, пестициды, тяжелые металлы*. Ситуация разрешается очень просто — нужно добавить в любом порядке указанные термины в запрос, который в результате выглядит следующим образом: *Загрязнение воды в Волге, фенол, пестициды, тяжелые металлы*. Вы запускаете на исполнение этот запрос и получаете множество различных смысловых аспектов. Все, не интересующие вас ранги обязательно должны включать хотя бы один из указанных терминов *фенол, пестициды, тяжелые металлы*. Единственно, что от вас требуется, это пропустить (не смотреть) эти ранги. Система гарантирует, что в любых других рангах (независимо от их числа) нет ни одного из упомянутых терминов.

Одной из важных особенностей ДИАНЫ является то, что, во многих случаях нет никакой необходимости предварительного просмотра документов ранга для принятия решения о включении их в результат поиска. Дело в том, что, чем больше слов запроса содержит терминологическая композиция, тем ниже вероятность попадания в ранг нерелевантного документа. Это и понятно. Трудно ожидать, что четыре—пять слов запроса, находятся в таком текстовом окружении, что документ не имеет отношения к теме запроса. Это подтверждается реальной практикой поиска средствами ДИАНЫ. Поэтому при отборе смысловых аспектов с насыщенными терминологическими композициями можно ограничиться только проставлением метки об их включении в выдачу без просмотра собственно документов. Это дополнительный аргумент в пользу того, что при поиске средствами

ДИАНЫ следует использовать подробные формулировки запроса.

А теперь остановимся кратко на принципах поиска. В ДИАНЕ запросы обрабатываются полностью автоматически без привлечения пользователя. Когда пользователь предъявляет системе некоторый текст в качестве запроса, то на самом деле это всего лишь исходный материал, из которого система будет сама автоматически конструировать разные формулировки из имеющихся терминов. Эти термины автоматически распознаются в тексте запроса специальными средствами, оцениваются по значимости, а затем начинается невидимый пользователю супербыстрый автоматический сеанс диалогового поиска по сочетаниям различных терминов. При этом используются специальные терминологические базы знаний, полученные на основе обработки больших корпусов текстов с обнаружением терминологических композиций разной длины. Дело в том, что простой перебор вариантов возможных комбинаций слов при генерации подзапросов требует при достаточном числе слов запроса очень больших вычислительных ресурсов. А это создает препятствия для реализации системы на рядовых компьютерах с процессорами PENTIUM обычной производительности.

В системе использованы эвристические процедуры, позволяющие довольно быстро выходить на большинство наиболее ценных терминологических композиций, не занимаясь перебором комбинаций. Сначала система пытается выяснить, нет ли документов, содержащих все значимые термины запроса. Такие документы автоматически попадают в первый ранг. Но отсутствие таких документов не нарушает стратегии поиска наилучших терминологических композиций для первых рангов. Сценарий автоматического поиска построен так, что в ходе сеанса система самообучается и находит все новые и новые удачные формулировки. Поиск по каждой из таких формулировок по сути дела напоминает обычный пользовательский поиск традиционного диалогового сеанса, но не по сценарию пользователя, а по внутренним правилам интеллектуального поискового процессора системы. При этом есть существенная разница — скорость таких поисков в ДИАНЕ несоизмеримо выше. За несколько секунд выдаются результаты поисков по сотням формулировок, отражающих самые тонкие аспекты исходного запроса на естественном языке (каждая из формулировок имеет уникальный терминологический состав). Если сравнить результат с обычными диалоговыми системами, то там за те же секунды выдается ответ только на одну формулировку, да и то он может оказаться нулевым. Нетрудно себе представить абсолютную нереальность попробовать в традиционном диалоговом режиме не то, что сотню, но даже и пару десятков формулировок с разным терминологическим составом.

РАСПОЗНАВАНИЕ СЛОВ В ЗАПРОСЕ И ДОКУМЕНТАХ

То обстоятельство, что ДИАНА в полностью автоматическом режиме должна обрабатывать обычные текстовые запросы, вызывает необходимость использования средств автоматического распознавания и в запросе и в документах слов естественного языка. Для этого используются специально разработанные средства морфологического анализа и синтеза, ориентированные не только на представленные в словарях слова, но и неологизмы и слэнговую лексику. Такая необходимость связана с отсутствием каких-либо ограничений на естественный язык запросов. Известно, что ряд систем морфологического анализа, построенных на основе нормативных словарей (например, известного грамматического словаря русского языка А. А. Зализняка) хорошо справляются с обычной словарной лексикой, но, порой, допускают ошибки с новыми словами, встречающимися в реальных текстах.

Система морфологического анализа и синтеза, используемая в ДИАНЕ, построена на основе обработки больших корпусов реальных текстов. В зависимости от целей использования системы она может настраиваться по-разному. Основной режим настройки предусматривает раскрытие словоизменительной парадигмы для существительных, прилагательных и причастий. Местоимения, предлоги, частицы, союзы и другие подобные малозначащие части речи не учитываются при поиске. Части речи, принадлежащие глагольной парадигме: глаголы в инфинитиве, настоящем, прошедшем и будущем времени, причастия, деепричастия не расширяются другими формами (хотя соответствующая настройка на противоположный вариант может быть сделана). Это объясняется тем, что глаголы и деепричастия в подавляющем числе случаев не несут основной смысловой нагрузки в запросах и представление полной парадигмы могло бы привести к значительному числу избыточных шумовых смысловых аспектов при поиске по свободным фрагментам текста. Особое внимание уделено распознаванию омонимичных слов. Не секрет, что пропуск варианта существительного в словах *дорогой* и *пасть* или варианта фамилии в слое *волков* может привести к невосполнимым потерям при поиске. Система настроена так, что скорее она выдаст маловероятные варианты (для новых слов), чем пропустит какой-то возможный случай. Например, для слова *Гора* будут выданы не только *Горы, Горой* и т. д., но и *Гором*, что вполне допустимо, если речь пойдет о вице-президенте США.

При создании индексных файлов БД все слова оставляются в естественном виде (без нормализации). При поиске для каждого значащего слова запроса синтезируются все его возможные словоформы. Система ориентирована на автоматическое распознавание слов русского и английского языка.

СОЗДАНИЕ ПРОБЛЕМНЫХ БД ПО СЛОЖНЫМ ТЕКСТОВЫМ ЗАПРОСАМ

Под проблемными БД понимаются объемные наборы записей по определенной проблеме. Такие наборы могут содержать тысячи записей, отобранных по некоторым критериям из ретроспективных БД. Желание создавать проблемные БД понятно. Как правило, пользователи таких БД — это специалисты в определенных предметных областях, которые хотят в основном иметь дело не с универсальной политематической библиотекой, где есть все, что угодно — от физики до экологии, — а со своей библиотекой, где собрано наиболее важное в

своей профессиональной деятельности и своих коллег по работе. Проблемная база гораздо меньше политематической, удобнее, и она не требует для работы излишних вычислительных ресурсов. Одним из основных способов формирования проблемных БД считается использование классификационного деления документов на тематические разделы. В какой-то мере можно считать, что выпуски реферативных журналов отражают информацию по крупным проблемам, а деление массива документов по рубрикам внутри выпусков охватывает значительную часть более мелких проблем. К сожалению, взгляд на свои проблемы у большинства пользователей редко совпадает с традиционным классификационным делением на рубрики. Чаще всего для проблемной БД оказывается необходимым отбор записей из многих рубрик, при этом большинство документов каждой рубрики отвергается. Чтобы как-то компенсировать отмеченный выше недостаток обычно предлагается использовать все тот же механизм диалогового поиска по ключевым словам. Для формирования проблемной БД пользователю рекомендуют выбрать нужные рубрики, а затем, с помощью диалогового поиска отобрать наиболее важные аспекты для проблемы. Проблема, как правило, включает большое число аспектов, поэтому для отбора документов приходится формулировать не один, а множество разных запросов. Мы уже показывали выше, что отражение разных аспектов в запросах представляет собой трудоемкую задачу. А это означает на практике, что реальный отбор документов по всем интересующим аспектам в традиционном режиме требует очень больших затрат. К сожалению, поиски по разным запросам могут пересекаться и в результате в проблемную БД многократно включаются одни и те же документы.

С помощью ДИАНЫ создание проблемной БД требует существенно меньших усилий, а ее качество оказывается значительно выше. Во-первых, ДИАНА вполне справляется со сложными запросами в несколько десятков значащих слов. А это означает, что можно в рамках одного запроса сформулировать "словесный портрет" проблемы сразу в виде некоторой проблемной записки, где указываются термины, охватывающие суть проблемы. Во-вторых, система сразу генерирует сотни и тысячи непересекающихся терминологических композиций с абсолютно разными документами. А это значит, что нет повторяющихся документов и для формирования сложной проблемной БД пользователю достаточно лишь отметить интересные смысловые аспекты. Так как пользователь имеет дело с многословными терминологическими композициями, уровень шума гораздо ниже, чем в БД, получаемых традиционным путем. Наконец, весь отобранный перечень документов проблемной БД загружается в естественно-языковую поисковую оболочку, обеспечивая быстрый поиск по самым сложным запросам на естественном языке.

СПИСОК ЛИТЕРАТУРЫ

1. Кузнецов Б. А., Солнцева Е. К., Закамская Д. В., Леонтьев А. А., Деревянкин М. В., Быховский Д. В., Ашкинадзе Б. Л. Диана — система поиска в текстовых базах данных по запросам на естественном языке // Информационные продукты и технологии: Материалы конф. НТИ-96 — М.: ВИНИТИ, 1996 .— С. 156–158.

2. Кузнецов Б. А., Солнцева Е. К., Закамская Д. В., Леонтьев А. А., Деревянкин М. В., Быховский Д. В., Ашкинадзе Б. Л. Интеллектуальный поиск в текстовых БД с помощью системы ДИАНА // Информационные ресурсы, интеграция, технологии: Материалы конф. НТИ-97 — М.: ВИНИТИ, 1997 .— С. 131–135.

3. Кузнецов Б. А., Солнцева Е. К., Деревянкин М. В., Закамская Д. В., Быховский Д. В. Базы данных ВИНИТИ на CD-ROM в интеллектуальной поисковой оболочке "АРИАДНА": быстрый автоматический поиск текстовой информации по запросам на естественном языке // Материалы конф. НТИ-99 — М.: ВИНИТИ, 1999 .— С. 119–121.

Материал поступил в редакцию 21.08.01.