

УДК 81'33

UDC 81'33

**АВТОДОПОЛНЕНИЕ ПОИСКОВОГО
ЗАПРОСА НА ОСНОВЕ
АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ
ПОДСКАЗОК ИЗ
ПРОИНДЕКСИРОВАННЫХ
ДОКУМЕНТОВ ПРЕДМЕТНОЙ
ОБЛАСТИ****SEARCH QUERY AUTOCOMPLETION
BASED ON AUTOMATIC
RETRIEVAL OF SUGGESTIONS
FROM PREINDEXED
DOCUMENTS OF
PARTICULAR DOMAIN**

Ю. Д. Голяк,
*соискатель кафедры прикладной
лингвистики филологического
факультета БГУ*

J. Haliak,
*Applicant of the Department of Applied
Linguistics, Philological Faculty, BSU*

Поступила в редакцию 11.07.18.

Received on 11.07.18.

В статье рассматривается один из подходов к решению задачи автодополнения (предиктивного ввода) поискового запроса пользователя на основании предварительного извлечения потенциальных подсказок из проиндексированных документов предметной области. Описываемый метод в общем виде не зависит от языка запроса и документов поискового пространства, обеспечивает как автоматическое завершение введенной части запроса, так и погружение его в контекст. Подход позволяет избежать зависимости системы автодополнения от истории запросов различных пользователей, что является важным в условиях ее отсутствия либо невозможности использования. Особое внимание в работе уделено лингвистическому аспекту задачи извлечения из текста соответствующих грамматических структур с наполнением и формированию базы подсказок на примере русского языка.

Ключевые слова: поисковый запрос, автодополнение, релевантность подсказки, поисковая история, контекст запроса.

The paper deals with an approach to solving the Query Type-Ahead problem (search query predictive autocompletion) basing on the corpus of potential suggestions previously extracted from the texts that have been indexed for the following search purposes. The described method generally does not depend on language of the query and the documents to be searched in, provides not just an auto-completion of the entered query prefix, but its immersion into a wider or, conversely, more specified context, as well as gives an opportunity to avoid dependence on the user's search history in conditions of the impossibility to reuse one user's search history for query autocompletion of another's. Particular attention has been paid to the linguistic aspect of pre-extraction of relevant grammatical structures with their concrete content from the texts and the suggestion corpus development for Russian language.

Keywords: Query Type-Ahead, search query autocompletion, suggestion relevance, search history, query context.

Введение
Автодополнение (предиктивный ввод) пользовательского поискового запроса (Query Type-Ahead, QTA) является крайне востребованной функциональностью для информационных систем, в том числе с интерактивным пользовательским ЕЯ-интерфейсом. Основная цель решения задачи QTA, иногда рассматриваемой в совокупности с задачей автокоррекции запроса, заключается в сокращении общего времени, затрачиваемого пользователем на ввод поискового запроса. Это достигается за счет того, что система, анализируя введенную пользователем часть запро-

са и опираясь на имеющуюся в ее распоряжении информацию, предлагает ему в виде подсказок варианты автоматического завершения запроса, выбор которых выполняется быстрее его ввода вручную. Существует целый ряд решений для реализации такого рода функциональности, в частности, в [1] описано решение задачи автодополнения методом поиска вхождений набранной части запроса в базе сохраненной предшествующей истории пользовательских запросов и ранжирования предложенных вариантов по релевантности, в [2] автодополнение формируется на основании личных данных пользователя и общей стати-

стики запросов, в [3] оно осуществляется параллельно с предоставлением поисковых результатов на введенный и автодополненный запрос (search-as-you-type), а предъявление подсказок и найденных документов обновляется динамически с каждым вводом одного или нескольких символов.

Отметим, что требования к системе QTA могут ограничиваться не только последовательным завершением строки вводимого запроса, рассматриваемого в качестве его префикса. Может предполагаться более сложная процедура, а именно погружение уже набранной пользователем части запроса в контекст, то есть дополнение его в начале, в конце или внутри набранной строки. Решение задачи в такой общей постановке не может не учитывать лингвистическую составляющую, а также историю поиска, причем не только уже проведенного, но и предполагаемого. При этом большинство современных систем, реализующих автодополнение запроса, как уже упоминалось выше, опираются в этом смысле только на историю поиска – персональную или общую статистику поисковых запросов в системе. Такое решение не всегда является оптимальным, а для некоторых коммерческих систем в целом может быть неприемлемо, так как история пользовательских запросов в некоторых профессиональных сферах является конфиденциальной информацией, представляющей собой коммерческую тайну. При этом объема персональной истории лишь одного пользователя или же истории запросов внутренней сети лишь одной организации, как правило, оказывается недостаточно для формирования качественной базы подсказок. В таком случае эффективным решением представляется предварительное извлечение потенциальных подсказок из индексируемых текстов, представляющих собой потенциальное поисковое пространство. Именно в таком аспекте в настоящей работе рассматривается задача QTA, решение которой осуществлено в рамках многоязычной поисковой платформы IHS Goldfire [4].

1. Общая постановка задачи QTA и принципиальная схема ее решения

Будем предполагать, что вводимый пользователем запрос в определенный момент времени представляет собой цепочку

$$W_1 W_2 \dots W_n, \quad (1)$$

где каждое $W_i, i = 1, n - 1, n > 1$ является словом естественного языка, а $W_n, n \geq 1$ – словом либо префиксом слова. Задача автодополнения запроса вида (1) заключается в автоматическом формировании списка таких грамматически и семантически корректных цепочек слов естественного языка, которые включают в себя члены цепочки (1). При этом предлагаемые подсказки должны быть отранжированы по релевантности предложенных цепочек относительно исходной. Как уже упоминалось выше, основной ресурс предиктивных знаний о пользовательских запросах должен определяться с учетом требований к конфиденциальности истории запросов. Таковым ресурсом была определена сама база текстов, индексируемых системой для последующего поиска, из которых предварительно с помощью лингвистического процессора автоматически извлекаются наиболее релевантные с точки зрения соответствия потенциальному запросу последовательности слов, а именно именные, глагольные и другие группы, хранящиеся в виде базы возможных подсказок, каждая из которых является тем контекстом, в который погружается цепочка вида (1).

Что касается процедуры извлечения подсказок, то для решения задачи использовался лингвистический процессор системы IHS Goldfire [5], обеспечивающий необходимый автоматический анализ текста. Процесс этой обработки включает в себя следующие этапы: лексический, лексико-грамматический, синтаксический и семантико-синтаксический анализ текста, что в совокупности обеспечивает построение его формальной семантико-синтаксической структуры, ориентированной на распознавание в тексте знаний основных типов: объектов, фактов и причинно-следственных отношений [5].

Задача собственно погружения цепочки (1) в контекст и формирования списка наиболее релевантных подсказок не является предметом рассмотрения в данной работе, будучи существенно алгоритмической проблемой, которая эффективно решается с использованием известных строковых алгоритмов [6].

2. Разработка базы подсказок

2.1 Грамматические структуры, лежащие в основе базы подсказок

Поскольку одним из основных результатов предварительной автоматической обработки текста указанным выше лингвистическим процессором является множество распознанных объектов, фактов и причинно-следственных отношений, естественным представляется то, что грамматические структуры, которые выступают носителями этих компонентов, можно считать основными с точки зрения решаемой задачи.

Экспертный анализ вышеуказанных грамматических структур для русского языка с учетом наиболее частотных шаблонов запросов позволил выделить в качестве базовых следующие структуры: *именная группа* (Noun Group, NG); *субъект + предикат* (S+P); глагольные группы: *инфинитив с прямым объектом в косвенном падеже, косвенным объектом с предлогом, или комбинацией таких объектов* (Inf+O, In+IO, Inf+O+IO), *форма 3 лица единственного или множественного числа глагола с прямым объектом в косвенном падеже, косвенным объектом с предлогом или комбинацией таких объектов* (VB3_Sing+O, VB3_PI+O, VB3_Sing+IO, VB3_PI+IO, VB3_Sing+O+IO, VB3_PI+O+IO); причастные группы: *именная группа + зависимое от нее причастие и – опционально – прямой объект в косвенном падеже или косвенный объект с предлогом* (NG+participle, NG+participle+O, NG+participle+IO), однословные подсказки: *глагольные инфинитивы, наречия, прилагательные, местоимения, числительные*.

Примеры конкретных наполнений всех этих структур, извлекаемых из текстов, и потенциальных пользовательских запросов, которые могут быть автоматически завершены на этапе ввода на основе этих наполнений-подсказок, приведены в таблице 1. Здесь показаны примеры текстового наполнения каждой из структур, а в последнем столбце проиллюстрирована сама процедура автодополнения: жирным шрифтом выделена автоматически дополненная часть запроса, тогда как набранный пользователем префикс запроса записан обычным шрифтом.

База подсказок членится на несколько разделов в соответствии с типами структур: разделы, соответствующие именным структурам, глагольным, атрибутивным и некоторым другим – в зависимости от того, в запросе какого семантического типа они скорее всего будут фигурировать, какой из префиксных паттернов будут продолжать (п. 2.3). Вводится также ограничение на длину подсказки, определенное эмпирическим путем.

Наполнение структур последнего типа – единичные словоформы – используется для автодополнения в рамках единичного слова, если для него не нашлось релевантного контекста из базы распространенных грамматических структур. В этот раздел попадают глагольные инфинитивы (*вытачивать, расплавить*), наречия (*высоко, ярко*), прилагательные (*деревянный, резбовой*), информативные местоимения (*любой, сам*), числительные (*восемьдесят, впятером*), длинные предлоги (*вследствие, навстречу*).

2.2 Фильтрация базы подсказок по критерию информативности

Не всякое наполнение вышеперечисленных грамматических структур можно считать приемлемым в качестве контекста запроса. Поэтому на этапе предобработки кандидаты на добавление в базу подсказок проходят фильтрацию по многим критериям, определяющим информативность определенного поля потенциального кандидата. Например, из множества потенциальных подсказок отфильтровываются кандидаты, содержащие в своих полях: 1) даты, значимые лишь в контексте денежные суммы, числовые маркировки и прочая

Таблица 1. Грамматические структуры, составляющие базу подсказок

Тип грамм. структуры	Подтип грамм. структуры	Пример наполнения	Пример запроса с автодополнением на основе указанной подсказки
Именные группы	NG	лазерная хирургия	лазерная хирургия
	Главное слово NG	хирургия	хирургия глаза
Предикативный центр / грамм. основа предложения	S+P	ломается двигатель	Почему ломается двигатель?
Глагольные группы	Inf+O	определить уровень инфляции	Как определить уровень инфляции ?
	VB3_Sing + O	предотвращает коррозию металла	Что предотвращает коррозию металла ?
	VB3_PI + IO	добывают в соляных шахтах	Что добывают в соляных шахтах ?
	Inf + O + IO	восстановить гладкость поверхности с помощью лазерной шлифовки	Как восстановить гладкость поверхности с помощью лазерной шлифовки ?
	VB3_Sing + O + IO	утверждает антикоррупционные законопроекты при парламентской системе	Кто утверждает антикоррупционные законопроекты при парламентской системе ?
	VB3_PI + O + IO	производят заменитель каучука на химических заводах	Из чего производят заменитель каучука на химических заводах ?

численная информация, не несущая смысловой нагрузки в отрыве от полного текста предложения: *деталь 21, 513 рублей*; 2) «языковые переменные», сигнификаты с изменяющимся контекстно-зависимым денотатом – местоимения и местоименные атрибуты: *она решила, мой кошелек, некоторый процент*; 3) малоинформативные атрибуты: порядковые числительные (не в устойчивых фразах и вне общепринятых терминов), притяжательные прилагательные и др.: *дедушкин дом, пятый квадрат*; 4) обценная лексика; 5) неинформативные аббревиатуры: *т. е., в т. ч., и др.*; но не *США, ОАО*; 6) наречия категории степени, интенсификации: *очень, весьма* и пр.

Фактически большей частью здесь идет речь о кандидатах, для которых характерно отсутствие самостоятельной смысловой нагрузки вне контекста.

2.3 Корреляция семантического типа запроса и раздела базы подсказок

Поисковые запросы пользователей в вопросительной форме или как минимум содержащие глагольную группу, оказалось возможным разделить на несколько семантических типов, объединенных по принципу общности префиксного паттерна и грамматического типа следующего за ним словосочетания, иными словами, типа раздела базы подсказок. Например, именные группы и именные группы с зависимым причастием будут использоваться в запросе похожим образом и потому предполагаются к хранению в одном разделе базы подсказок, а именно: в 1 раздел будут помещены именные группы и именные группы с причастием, во 2 – глагольные группы с формой

3 лица множественного числа и структуры типа «Субъект + Предикат», в 3 – глагольные группы с инфинитивом, в 4 – глагольные группы с формой 3 лица единственного числа, в 5 – атрибутивные (определятельные) элементы.

Такая структуризация помогает осуществлять поиск наиболее релевантной подсказки быстрее и эффективнее за счет сокращения пространства поиска на этапе ввода запроса.

Примеры префиксных паттернов и предлагаемых подсказок из соответствующего раздела базы подсказок приведены в таблице 2.

3 Заключение

Представленный в работе метод решения задачи автодополнения пользовательского запроса обладает рядом существенных достоинств. В частности, он: а) ориентирован на самую общую постановку задачи, более того, погружение запроса в контекст с использованием базы подсказок на уровне алфавита позволяет одновременно эффективно решать также задачу автокоррекции запроса; б) является универсальным по отношению к языку запроса; обеспечивает решение задачи, ориентированное на идеологию семантического поиска; в) обеспечивает решение задачи в условиях отсутствия либо невозможности использования истории запросов, опираясь при формировании базы потенциальных подсказок на индексированные для последующего поиска тексты, что обеспечивает не только более высокую вероятность соответствия подсказки ожиданиям пользователя, но и гарантированно релевантную реакцию поисковой системы; г) дает пользователю возможность точнее сформулировать свою информационную потребность.

Таблица 2. – Примеры префиксов запросов и соответствующих им разделов базы подсказок

Префикс запроса	№ раздела	Примеры
Как (+ можно, нужно и т. д.)	2, 3, 4	как можно определить заболевание на ранней стадии ; как возникает коррозия ; каким образом эффективно удалить налет
Что/Кто	4	что вызывает гипоксию ; кто принимает решение о выписке пациента
Что/Кто может/ помогает/позволяет	3, 4	что помогает отстирать вино ; кто может читать вслепую
Какой/чей/который/каков (+ падежные формы)	1, 2	какой сорт пшеницы используют в промышленном изготовлении хлеба ; каков текущий показатель рождаемости в Венгрии
Почему/где/зачем/откуда/если/из чего	3, 4	почему возникает коррозия ; где приобрести редкие реактивы
Третья форма глагола + ли	1	вызывает ли влага коррозию ; оплачивает ли страховая компания протезирование
Префикс первого слова самой подсказки – может продолжаться подсказками из любого раздела, проиллюстрируем раздел 5: холо...	5	холодная война, холодный климат

Особенность решения задачи QTA для русского языка при таком подходе заключается в широком комбинаторном разнообразии потенциальных шаблонов запросов и, соответственно, потенциальных подсказок, принимая во внимание флективный строй языка, его развитую падежную систему, богатую омонимию, почти свободный порядок слов, разнообразное глагольное управление и др. Тем не ме-

нее, как оказалось на практике, трудоемкость решения рассматриваемой задачи не столь высока, поскольку предложенный метод получения базы подсказок и существующие возможности алгоритмической реализации позволяют осуществлять автодополнение с достаточно небольшими временными затратами и высокой степенью соответствия подсказки ожиданию пользователя.

ЛИТЕРАТУРА

1. Search entry system with query log autocomplete: US Patent 20080065617 A1 / Eric Paul Burke, Duke Fan, Alan Wada, Jawahar Malhotra, Brian Coe. (2005/2008)
2. Search query suggestions based on personal information US Patent 9317585 B2 / Maureen Heymans, Ashutosh Shukla, Harish Rajamani, Matthew E. Kulick, Bryan C. Horling, Jennifer E. Fernquist, Weniger". (2013/2014)
3. Incremental query refinement: US Patent 9245052 B2 / Brett D. Brewer, Oliver Hurst-Hiller (2012/2014)
4. Question-answering system and method based on semantic labeling of text documents and user questions: US Patent 8,666,730 / James Todhunter, Igor Sovpel, Dzianis Pastanohau. (2009/2010)
5. Чеусов, А. В. Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора : дис. ... канд. техн. наук: 05.13.17 / А. В. Чеусов. – Минск, 2013. – 116 с.
6. Huizhong Duan. Online spelling correction for query completion / Huizhong Duan, Bo-June (Paul) Hsu // WWW '11 Proceedings of the 20th international conference, Pages 117–126, ACM New York, NY, USA – 2011.

REFERENCES

1. Search entry system with query log autocomplete: US Patent 20080065617 A1 / Eric Paul Burke, Duke Fan, Alan Wada, Jawahar Malhotra, Brian Coe. (2005/2008)
2. Search query suggestions based on personal information US Patent 9317585 B2 / Maureen Heymans, Ashutosh Shukla, Harish Rajamani, Matthew E. Kulick, Bryan C. Horling, Jennifer E. Fernquist, Weniger". (2013/2014)
3. Incremental query refinement: US Patent 9245052 B2 / Brett D. Brewer, Oliver Hurst-Hiller (2012/2014)
4. Question-answering system and method based on semantic labeling of text documents and user questions: US Patent 8,666,730 / James Todhunter, Igor Sovpel, Dzianis Pastanohau. (2009/2010)
5. Cheusov, A. V. Razrabotka algoritmov i tekhnologii postroyeniya mnogoyazychnogo bazovogo lingvisticheskogo protsessora : dis. ... kand. tekhn. nauk: 05.13.17 / A. V. Cheusov. – Minsk, 2013. – 116 s.
6. Huizhong Duan. Online spelling correction for query completion / Huizhong Duan, Bo-June (Paul) Hsu // WWW '11 Proceedings of the 20th international conference, Pages 117–126, ACM New York, NY, USA – 2011.