The system for automatic suggestions generation for the purpose of autocompletion of Russian-language user search queries

Julia Haliak
Belarusian State University
IHS Markit Ltd.
Minsk, Belarus
julia.haliak@ihsmarkit.com

***Abstract*—The paper presents a solution to the problem of predictive input of user queries for information systems working with full-text databases. In contrast to the traditional approach, it is based on the preliminary automatic construction of a set of suggestions that are being recognized in the search space itself. The main advantages of the obtained solution, implemented in well-known information system, are presented.**

***Keywords*—query type-ahead, predictive query input, search query, query autocompletion, linguistic processor**

I. Introduction

The majority of modern information systems with an interactive natural language user interface have the functionality of automatic completion of a user query. The most common solutions are based on the generated history of the previous search, use query logging and/or some information about the user [1]–[3]. The solution proposed in our case is focused on information systems working with full-text databases (FTDBs) and is based on the possibility of preliminary (before the system exploitation) formation of a "history"of not yet carried out, but intended search, in the form of a set of autosuggestions that are automatically recognized in advance in the full-text database, since the user queries are going to be addressed to FTDB.

II. The proposed solution

The general formulation of the problem of autocompletion of user queries in our case states as follows. Let the query entered by the user at a certain point in time be represented as the chain:

$W_1\ W_2\ ...\ W_n$,

(1)

where each W_i, $i = 1, \ldots, n-1$, $n > 1$, is a natural language word, and W_n, $n \geq 1$, is a word or a word prefix. The task of completing a query of the form (1), which to a certain extent already reflects the user's information need, is to automatically generate a list of such, ideally, grammatically and semantically correct natural language word sequences that include members of the chain (1). That means we are dealing not only with sequential completion of the string of the entered query, considered as its prefix, but with a more complex procedure, which, in the general case, involves immersing the part of the query already typed by the user into the context, namely, its addition at the beginning, at the end, or even inside. Moreover, the proposed sequences may not contain all words from (1), but in any case, they should be ranked according to their relevance to the original chain. In addition, the autocompletion procedure can be applied to the entered query repeatedly (fragmentarily). Proceeding from the fact that the majority of users follow the already established practice of formulating search queries and that they are traditionally more or less focused on keyword search, the following steps for solving the problem of generating a set of suggestions to autocomplete user queries can be proposed in described case:

1) expert analysis based on open source available datasets of the most frequent search queries and classification of their types;
2) classification of syntactic structures of queries for each of their types;
3) automatic linguistic analysis of the FTDB in order to recognize in its text documents the syntactic structures obtained at the previous stage and the selection of all sorts of corresponding lexical content from the FTDB, which constitutes the set of autosuggestions P.

There is the classification of the main types of search queries (the autosuggestions) obtained on the basis of the expert analysis, and the classification of syntactic structures for each of their types given in [4]. The following types of autosuggestions are defined as the most relevant:

1) "simple noun phrase"(повреждения металла (metal damage); защитный слой (protective layer)),
2) "extended noun phrase with prepositional-nominal construction"(установка оборудования в цехах (equipment installation in workshops); обработка поверхностей в условиях высокой температуры (surface treatment in high temperature conditions)),
3) "extended noun phrase with participle

phrase"(покрытие, создающее защитный слой (coating creating protective layer); повреждения, вызванные коррозией (damage caused by corrosion)),

4) "verb phrase"(создавать защитный слой (create protective layer); предотвратить повреждения металла (prevent metal damage)),
5) "sentence subject and predicate"(покрытие предотвращает (coating prevents); коррозия вызывает (corrosion causes)),
6) "lexicon"(антикоррозийный (anticorrosive); соосно (coaxially)).

There are also requirements formulated for the linguistic processor that provides automatic recognition of the above-mentioned syntactic structures in texts from FTDB. Such recognition was made possible due to the use of the basic linguistic processor (BLP) [5] of the information system IHS Goldfire [6], which carries out automatic linguistic analysis of the input text by stage-by-stage processing:

- text formatting and normalization: here the text is converted into a certain unified format that preserves the stylistic and structural markup of documents as much as possible; in addition, the text is divided into paragraphs; headings, subheadings and sections are recognized;
- lexical text analysis: here the words and sentences boundaries are defined, the problems of recognition of proper names, abbreviations, e-mail addresses, digital and other sign complexes are solved partially or completely;
- lexical-grammatical text analysis: here the system identifies the lexical-grammatical category of each word, taking into account its morphology and context in accordance with a given classifier;
- syntactic and semantic text analysis: here the syntactic relations are recognized in each sentence and presented, as a rule, in the form of a functional or syntactic tree, in which the words of the sentence obtain the identification of their grammatical function and the type of syntactic connection between them is determined; on this stage of analysis the system also recognizes the relationships between concepts expressed by noun phrases, within the so-called SAO-structure [7]: Subject - Action (Predicate) - Object, and each element in this structure has its own attributes [8], [9].

Consider the following sentence as an example:

Антикоррозийное покрытие аэрозольного нанесения, создающее защитный слой, предотвращает повреждения металла, вызываемые коррозией. (Anti-corrosion spray coating that creates a protective layer prevents metal damage caused by corrosion.)

Having a linguistic analysis of this sentence performed, BLP recognizes the following three SAO-relations presented in Table 1, where HW stands for Headword (words that serve as a link to a parent relation – usually Predicate or Noun Phrase, but links to other fields are possible), C – Conjunction that relates to the whole relation, S –Syntactic Subject, P – Predicate (verb infinitive), NP – Nominal Predicate, O – Direct Object, D – Object in Dative, I – Object in Instrumental, Pr – Preposition that introduces syntactic Indirect Object, IO – Indirect Object, A – Adverbial modifier, In – Introduction phrase, AO – Action Original, original form of a predicate in text.

Таблица I
Table 1. SAO Fields

Fields	SAO 1	SAO 2	SAO 3
HW	-	повреждения металла металла	анти-коррозийное покрытие аэрозольного нанесения
C	-	-	-
S	анти-коррозийное покрытие аэрозольного нанесения	коррозия	анти-коррозийное покрытие аэрозольного нанесения
P	предотвращать	вызывать	создавать
NP	-	-	-
O	повреждения металла	повреждения металла	защитный слой
D	-	-	-
I	-	-	-
Pr	-	-	-
IO	-	-	-
Adv	-	-	-
In	-	-	-
AO	предотвращает	вызываемые	создающее

The SAO-structure is he source material on the basis of which, taking into account a certain expansion of the BLP functionality, algorithms for the automatic construction of autosuggestions of all the above-mentioned types are being built. These algorithms are based on the synthesis of the required autosuggestions from the content of the fields corresponding to each of their types. In particular, the source for generating autosuggestions of «simple noun phrase» type are the following fields: Headword, Subject, Nominal Predicate, Direct Object, Object in Dative, Object in Instrumental, Indirect Object; for the «extended noun phrase with prepositional-nominal construction» type – Headword, Preposition and Indirect Object fields; for

the «extended noun phrase with participle phrase» type – Headword, Action Original, Direct Object, Object in Dative, Object in Instrumental, Preposition and Indirect Object fields; for the «verb phrase» type – Predicate, Action Original, Direct Object, Object in Dative, Object in Instrumental, Preposition and Indirect Object fields; for the «subject and predicate» type – Subject and Action Original fields; and all the SAO fields – for the «lexicon» type.

From the autosuggestions obtained in this way derivative suggestions are additionally synthesized by truncating noun phrases with coordinated attributes (антикоррозийное покрытие -> покрытие (anti-corrosion coating -> coating)), and considering only the main words of noun phrases (коррозия металла -> коррозия (metal corrosion -> corrosion)). In addition, these algorithms take into account the following important circumstance. One of the main criteria for the relevance of a suggestion for the automatic completion of a query is its informativeness, as a separate query or its part, in a given subject domain [10]. For example, noun phrases obtained from introduction constructions, such as «в том числе» (amongst other things), «в большинстве случаев» (in majority of cases) – «то число» (e.g. other things), «большинство случаев» (majority of cases), – firstly, cannot be a separate query, and secondly, due to their obviously poor informativeness, are also a bad part of a query. This fact is confirmed by the absence of such bigrams and trigrams in the lists of user requests available in open sources [11]–[14]. In addition, certain filtering is required, for example, for such noun phrases as « т.н. желтая ржавчина» (so-called yellow rust) because elements like certain abbreviations, which are more typical for formal style, are low-frequent in user queries [11]–[14]. The studies carried out in this aspect allowed, as a result, to develop a set of rules for identifying such cases in order to exclude certain SAO-relations or their fields from the list of candidates for generating autosuggesions, as well as the necessary further filtering of the candidates remaining in the list and their possible transformation.

Assuming the example sentence is included in the FTDB, the following list of autosuggestions will be obtained from it:

1) покрытие аэрозольного нанесения (aerosol application coating); аэрозольное нанесение (aerosol application);
2) нанесение (application);
3) повреждения металла (metal damages);
4) повреждения (damages);
5) металл (metal);
6) коррозия (corrosion);
7) защитный слой (protective layer);
8) слой (layer);
9) антикоррозийное покрытие аэрозольного нанесения, создающее защитный слой (anticorrosive spray coating creating a protective layer);
10) повреждение металла, вызываемые коррозией (metal damage causing by corrosion);
11) антикоррозийное покрытие аэрозольного нанесения, создающее (anticorrosive spray coating creating);
12) повреждения металла, вызываемые (metal damage causing); покрытие, создающее защитный слой (coating creating protective layer);
13) вызываемые повреждения металла (causing metal damage); покрытие, создающее (coating creating);
14) повреждения, вызываемые (damage causing);
15) предотвращать повреждения металла; вызывать повреждения металла (prevent metal damage);
16) создавать защитный слой (create protective layer);
17) предотвращает повреждения металла (prevents metal damage);
18) предотвращать повреждения (prevent damage);
19) предотвращает повреждения (prevents damage);
20) вызывать повреждения (cause damage);
21) создавать слой (create layer);
22) предотвращать повреждения металла, вызываемые коррозией (prevent metal damage causing by corrosion);
23) предотвращает повреждения металла, вызываемые коррозией (prevents metal damage causing by corrosion);
24) предотвращать повреждения, вызываемые коррозией (prevent damage causing by corrosion);
25) предотвращает повреждения, вызываемые коррозией (prevents damage causing by corrosion);
26) антикоррозийное покрытие аэрозольного нанесения предотвращает (anti-corrosion spray coating prevents);
27) предотвращает антикоррозийное покрытие аэрозольного нанесения (anti-corrosion spray coating prevents – inversed);
28) покрытие аэрозольного нанесения предотвращает (spray coating prevents);
29) покрытие предотвращает (coating prevents);
30) предотвращает покрытие аэрозольного нанесения (spray coating prevents – inversed);
31) предотвращает покрытие (coating prevents – inversed);
32) антикоррозийное (anti-corrosion);
33) аэрозольное (aerosol);
34) вызываемые (caused);
35) защитный (protective);
36) создающее (crating);
37) предотвращать (prevent);
38) вызывать (cause);
39) создавать (create).

III. Conclusions

The proposed method for solving the user search query autocompletion problem has a number of significant

advantages.

- It is focused on the most general formulation of the problem; and moreover, immersing a query into context using a database of autosuggestions at the alphabet level allows to simultaneously and effectively solve the problem of user query autocorrection.
- It is universal in relation to the query language due to the universality of the SAO-relation itself; it also provides a solution to the problem focused on the ideology of semantic search.
- It gives a solution to the problem in the absence or inability to use the history of queries, relying on the texts to be indexed for the later search when forming a database of potential autosuggestions, which provides not only a higher probability of matching the autosuggestion to the user's expectations, but also a guaranteed relevant response of the search engine.
- It enables the user to more accurately formulate his informational need.

As for the problem of actually immersing the query prefix in the context provided by autosuggestions, it is being effectively solved using the well-known string algorithms [15]. The presented results were incorporated into the IHS Goldfire information system and showed their relevance and effectiveness.

References

[1] 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)

[2] Search entry system with query log autocomplete: US Patent 20080065617 A1. Eric Paul Burke, Duke Fan, Alan Wada, Jawahar Malhotra, Brian Coe. (2005/2008)

[3] 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)

[4] Haliak J.D. Printsipial'naya skhema resheniya zadachi avtodopolneniya pol'zovatel'skikh poiskovykh zaprosov na russkom yazyke i ee analiz [Conceptual scheme of the solution for the problem of user search query autocompletion in Russian language and its analysis]. Uchenye zapiski VGU imeni P. M. Masherova: sb. nauch. Trudov [Scientific notes of Vitebsk State University named after P.M. Masherov: collection of scientific papers]. Vitebsk, VSU P. M. Masherova, 2020, Vol. 31, pp. 142-147.

[5] System and method for automatic semantic labeling of natural language texts: Patent EP2406731A4. J. Todhunter, I. Sovpel, D. Pastanohau. (2010).

[6] IHS Goldfire. Available at: https://ihsmarkit.com/products/enterprise-knowledge.html (accessed 2021, May).

[7] Cheusov, A.V. Razrabotka algoritmov i tekhnologii postroeniya mnogoyazychnogo bazovogo lingvisticheskogo protsessora. Diss. k-ta tekhn. nauk [Development of algorithms and technologies for building a multilingual basic linguistic processor. PhD thesis]. Minsk, 2013. 116 p.

[8] Haliak, J.D., Sovpel, I.V. Avtomaticheskoe dopolnenie russkoyazychnykh pol'zovatel'skikh zaprosov, formuliruemykh v vide imennykh grupp [Autocompletion of russian language user queries formulated as noun phrases]. Vestnik MGLU. Filologiya [MSLU Bulletin. Philology], Vol 1, 2021, No 3 (112), pp. 101-110.

[9] Haliak, J.D., Sovpel, I.V. Avtomaticheskoe dopolnenie russkoyazychnykh pol'zovatel'skikh zaprosov na osnove podskazok tipa «glagol'naya gruppa», «grammaticheskaya osnova predlozheniya» i «leksikon» [Autocompletion of Russian language user queries based on suggestions of «verb phrase», «sentence subject and predicate» and «lexicon» types]. Vestnik MGLU. Filologiya [MSLU Bulletin. Philology], Vol 1, 2021, No 4 (113) (in press)

[10] Voronkov, N.V. Metody, algoritmy i modeli sistem avtomaticheskogo referirovaniya tekstovykh dokumentov. Diss. k-ta tekhn. nauk [Methods, algorithms and models of systems for automatic summarization of text documents. PhD thesis]. Minsk, 2007, 165 p.

[11] Kpapamih: search-queries. Available at: https://www.kaggle.com/kpapamih/search-queries (accessed 2021, May).

[12] Wordstat.Yandex. Available at: https://wordstat.yandex.ru (accessed 2021, May).

[13] Reidsma, M. (2016). Summon Topic Explorer Results by Search Query [Data set]. Zenodo. http://doi.org/10.5281/zenodo.47723 [Electronic resource]. Available at: https://zenodo.org/record/47723 (accessed 2020, Apr).

[14] Rishiraj. (2018). Supplementary Material for Nested Segmentation of Web Search Queries [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1137746 [Electronic resource]. Available at: https://zenodo.org/record/1137746 (accessed 2020, Apr).

[15] Irzhavskii, P. A. eds. Teoriya algoritmov: ucheb. posobie [Algorithm theory: schoolbook]. Minsk, BSU, 2013, 159 p.

Система автоматического построения подсказок с целью автодополнения русскоязычных пользовательских запросов

Ю. Д. Голяк

В работе представлено решение задачи предиктивного ввода пользовательских запросов для информационных систем, работающих с полнотекстовыми базами данных. Оно, в отличие от традиционного подхода, основано на предварительном автоматическом построении множества подсказок, распознаваемых в самом поисковом пространстве. Приводятся основные достоинства полученного решения, внедренного в состав известной информационной системы.

Received 21.07.2021