Conference Paper

Modeling of Conceptual and Terminological Structures Based on AML/CFT Texts for Solving Problems of Semantic Search

Gavrilkina A. S.

Post-graduate student at National Research Nuclear University MEPhI
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, Moscow, 115409, Russia

Abstract

The paper proposes method of conceptual and terminological structures construction for semantic search problems solving. The method based on automatic term extraction and analysis of their joint appearance dependencies using statistical and morphological terms characteristics.

Corresponding Author:
A. S. Gavrilkina
asgavrilkina@yandex.ru

Received: 11 December 2017
Accepted: 20 January 2018
Published: 13 February 2018

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the FinTech and RegTech: Possibilities, Threats and Risks of Financial Technologies Conference Committee.

1. Introduction

Investigation of financial crimes involves the search and collection of data, including the processing of large arrays of text documents. One of the methods, which seeks to improve search accuracy, is the semantic search, which takes into account the contextual meaning of terms. Terminology structures can be used to form a conceptual context, its main purpose is to reflect the immanent relationships, i.e. permanent, stable for the subject area links between objects and the situational relationships, which link the objects in a subject area within its fragment (corresponding to a certain process).

Despite the abundance of different approaches to the automatic formation of conceptual and terminological structures [1-4], it is possible to select some general stages of their construction: the preparation of a collection of texts that adequately and objectively reflect the state of the subject area; extraction of terms from texts; building relationships on extracted terms.

The paper proposes method of construction conceptual and terminological structures, which based on automatic term extraction and analysis of the dependencies of their joint appearance using statistical and morphological characteristics of terms.

Experiments were carried out with the collection of AML / CFT texts, in particular, conventions, orders, resolutions, guidelines, presidential decrees, federal laws, resolutions of the Government of the Russian Federation. A prototype of software, which was developed with the participation of the author, was used to build examples.

2. Materials and methods

2.1. Automatic terminology extraction

The task of automatically extracting terminology from the text is usually solved with the help of statistical methods [5-7] which are based on counting the frequency of occurrence of terms in a text or an array of texts, methods with the use of morphological patterns [5, 8], in which for each word in the phrase assigns morphological characteristics, or their combinations [9-12].

The proposed approach is based on the analysis of the morphological characteristics of terms. From texts by using separators, which are punctuation marks and all parts of speech with the exception of adjectives, nouns and prepositions, constructions are extracted with a high degree of probability being nominal substantive phrases. These constructions are considered as word combinations, if they meet the following basic rules:

- adjective precedes the noun;
- in phrases with two or more nouns without prepositions, second and subsequent nouns should be in the genitive, dative or instrumental case (place of names is not considered);
- nouns related by prepositions may be in any case.

The phrases selected in this way are arranged according to the principle of lexicographic inclusion in the form of a hierarchical dictionary, in which high-frequency nouns are used at the first level of the hierarchy (Fig. 1).

3. Construction of conceptual and terminological structures

There are many papers about this problem. Note the most constructive ones reflecting approaches to the terms links construction by using automatic text analysis. In [1],

ИМУЩЕСТВО (property)
  ДВИЖЕНИЕ ИМУЩЕСТВА (movement of property)
  ИЗЪЯТИЕ ИМУЩЕСТВА (exemption of property)
  ИНФОРМАЦИЯ ОБ ИМУЩЕСТВЕ (property information)
  ИСПОЛЬЗОВАНИЕ ИМУЩЕСТВА (use of property)
  КОНКРЕТНОЕ ИМУЩЕСТВО (specific property)
    КОНФИСКАЦИЯ КОНКРЕТНОГО ИМУЩЕСТВА (confiscation of specific property)
  КОНТРОЛЬ НАД ИМУЩЕСТВОМ (control over property)
  КОНФИСКАЦИЯ ИМУЩЕСТВА (confiscation of property)
    ФОРМА КОНФИСКАЦИИ ИМУЩЕСТВА (form of confiscation of property)
  ЛИШЕНИЕ ИМУЩЕСТВА (depreciation of property)
  ПЕРЕДАЧА ИМУЩЕСТВА (transfer of property)
  ПЕРЕДВИЖЕНИЕ ИМУЩЕСТВА (movement of property)
  ПРАВО НА ИМУЩЕСТВО (the right to the property)
  РАСПОРЯЖЕНИЕ ИМУЩЕСТВОМ (disposition of property)
  СТОИМОСТЬ ИМУЩЕСТВА (cost of property)

Figure 1: Lexicographical neighborhood for term PROPERTY.

an approach to building an ontology based on lexical-syntactic patterns for information retrieval is described, the author [2] proposes a model of automatic construction of ontologies in the form of the production system and the application of genetic and automatic programming to create the necessary models, in [3] for constructing ontology, statistical methods for the analysis of texts in natural language are mainly used, in [4] a combined approach using statistical methods and methods based on the production system is described.

In the present work, the following types of relations are used to construct a conceptual-terminological structure:

- quasi-hierarchical, arising when the term enters a longer one;
- typed, i.e. unified set of relations, characteristic for the subject area, extracted from the text on the basis of existence between a pair of terms of the connecting structure corresponding to the template;
- untyped, based on the calculation of the measure of relation for terms, in particular, the determination of the correlation between the appearance of terms in text fragments.

Quasi-hierarchical relationships reflect inclusion relationships that are used in constructing a hierarchical vocabulary.

Typed relationships arise between terms associated in the text with a connecting structure according to patterns, examples of which are given in Table 1.

Untyped relations are used when linking terms as a result of calculating the communication measure.

Table 1: Patterns for extraction of bonding structures.

Pattern	Example
Термин+глагол+термин (term+verb+term)	ТЕХНИЧЕСКИЕ ОГРАНИЧЕНИЯ препятствуют ПЕРЕДАЧЕ ПОЛНОЙ ИНФОРМАЦИИ
Термин+глагол+инфинитив+термин (term+verb+infinitive+term)	КОМПЕТЕНТНЫЕ ОРГАНЫ должны вести ВСЕСТОРОННЮЮ СТАТИСТИКУ; ФИНАНСОВОЕ УЧРЕЖДЕНИЕ должно идентифицировать КЛИЕНТА
Термин+глагол+предлог+термин (term+verb+preposition+term)	ФИНАНСОВОЕ УЧРЕЖДЕНИЕ сомневается в ИСТИННОСТИ
Предлог+термин+глагол+термин (preposition+term+verb+term)	в БОЛЬШИНСТВЕ СТРАН существуют ПРАВОВЫЕ ПРЕПЯТСТВИЯ
Термин+предлог+термин (term+ preposition+term)	ДОХОД от ПРЕСТУПЛЕНИЯ, ТРЕБОВАНИЯ по ИДЕНТИФИКАЦИИ КЛИЕНТОВ

To calculate the measure of communication, the text is divided into n intervals by the Sturges' rule (1):

$$n = 1 + [\log_2 N], \tag{1}$$

where N is the number of sentences in the text.

The presence of a connection between terms is determined using the Pearson correlation coefficient (2):

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2}$$

where x_i is the number of sentences containing the first term in the interval i, y_i is the number of sentences containing the second term in the interval i, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the mean sample values.

To assess the strength of communication, the Cheddock scale is used.

The algorithm for constructing a conceptual and terminological structure can be represented as follows:

- selection of the root term;
- constructing a hierarchy that includes the selected term;
- the construction of relations based on the calculation of the measure of communication: only those terms that are at least once encountered together within the same sentence are linked and have a strong bond;
- search of connecting structures by templates for related pairs of terms.

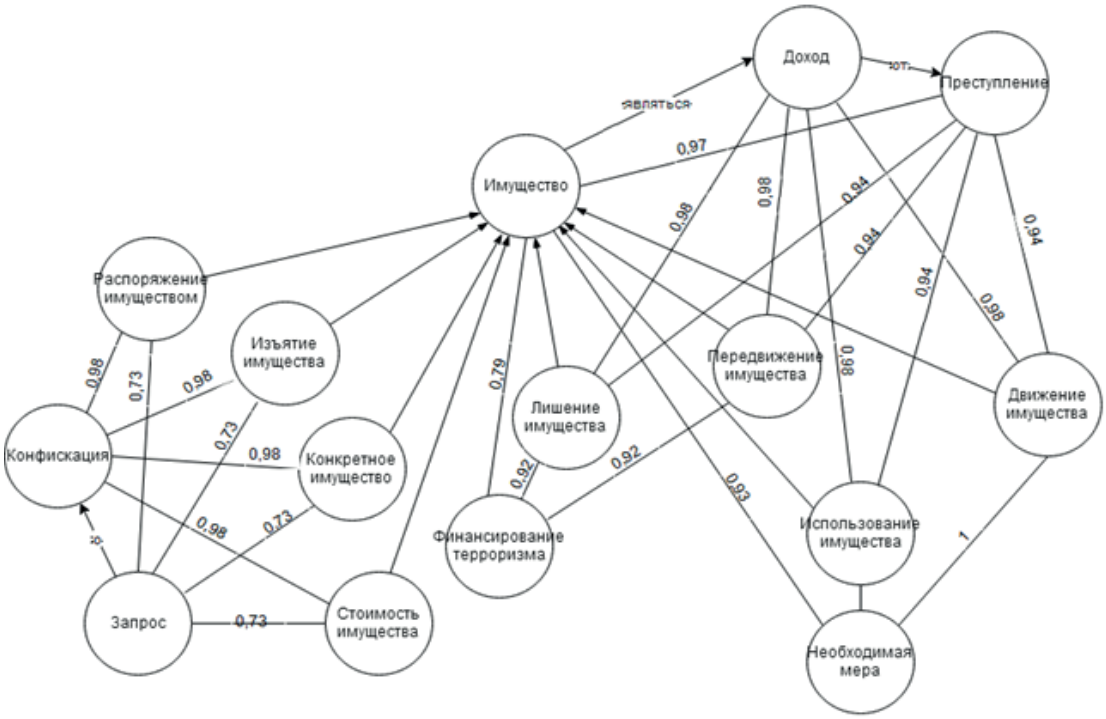

Figure 2: Fragment of the conceptual terminology structure.

4. Results

Figure 2 shows a fragment of the terminological structure in the form of a graph for the term PROPERTY.

Oriented arcs without indication of weight connect pairs of terms that are in quasi-hierarchical relations, and are directed from a longer term to the term entering into it; typed relations between terms are depicted on the graph by means of oriented arcs over which the connecting structure is indicated as the weight; untyped relations are reflected by non-oriented arcs indicating the communication measure as the weight.

5. Conclusion

The proposed methodology is focused on the automatic formation of conceptual and terminological structures for semantic search. The constructed word-combinations and relations can be used to supply existing conceptual and terminological systems or create new ones, in particular, thesauruses, taxonomies, ontologies of the subject area, etc.

Acknowledgements

This work was supported by Competitiveness Growth Program of the Federal Autonomous Educational Institution of Higher Education National Research Nuclear University MEPhI (Moscow Engineering Physics Institute).

References

[1] Rabchevsky EA, Automatic construction of ontologies based on lexico-syntactic patterns for information retrieval // Electronic Libraries: Advanced Methods and Technologies, Digital Collections: Proceedings of the 11th All-Russian Scientific Conference RCDL'2009. (Petrozavodsk, Russia, September 17-21, 2009). - Petrozavodsk: KarRC RAS, 2009. - 487 p.

[2] Nayhanova L.V. Methods and models of automatic construction of ontologies on the basis of genetic and automatic programming: Author's abstract. dis. Doct. those. sciences. - Krasnoyarsk, 2008. - 36 p.

[3] Mozzherina E.S. Automatic construction of ontology on the collection of text documents // Electronic Libraries: Advanced Methods and Technologies, Electronic Collections - Voronezh, 2011. - P. 293 - 298.

[4] Ahmadeeva I.R. Development of automation tools for building a taxonomic kernel of ontology based on the body of texts: WRC - Novosibirsk, 2013. - 23 p.

[5] Tabarcha A.I. Search for morphological templates for stable word combinations of arbitrary length // Postgraduate student and competitor. - 2010. - No. 6. - P. 138-139.

[6] Moshkin VS Study of the terminology representation in the linguistic support of CAD on the basis of integration of fuzzy ontologies and logical inference: - Ulyanovsk, 2017. - 23 p.

[7] Zakharov VP, Khokhlova M.V. Isolation of terminological word combinations from special texts on the basis of various measures of association // Internet and Contemporary Society "IMS-2014": a collection of scientific articles of the XVII All-Russian Joint Conference, 2014. - P. 290-293.

[8] Bolshakova EI Language of lexico-syntactic patterns LSPL: experience of use and ways of development // Software systems and tools: a thematic collection. - Moscow: department of the faculty of the Higher School of Management of Moscow State University; MAX Press, 2014. - P. 15-26.

[9] Braslavsky P., Sokolov E. Comparison of five methods of extraction of terms of arbitrary length / / Proceedings of the international conference "Dialogue 2008". -

Bekasovo: 2008. - P. 67-74

[10] Velichko V. Automated thesaurus creation of domain terms for local search engines / Velichko V., Voloshin P., Svitla S. // "Knowledge – Dialogue – Solution" International Book Series "INFORMATION SCIENCE & COMPUTING", Number 15. – FOI ITHEA Sofia, Bulgaria. – 2009. – pp. 24–31

[11] Nayhanova L V, Osnovnye aspekty postroeniya ontologiy verkhnego urovnya i predmetnoy oblasti [Main aspects of construction of high level ontologies and subject area], Internet Portals: Content and Tehnologies, 2005, Moscow, Informika, Prosveshchenie, pp.452-479

[12] Dobrov B. V., Lukashevich N. V., Syromyatnikov S. V. Formation of the base of terminological word-combinations on the subject domain texts // Proc. - 2003. - S. 201-210.