TF-IDF metrics and formation of units for knowledge representation in open tests

Mikhaylov D., Kozlov A., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

All-Russian Conference with International Participation
«Mathematical Methods for Pattern Recognition» (MMPR-17),

September 19–25, 2015

Svetlogorsk, Kaliningrad obl., Russian Federation

Research purpose and considered problem

Knowledge unit estimated by means of open form test assignment

Is defined by a set of natural-language (NL) phrases equivalent-by-sense (i. e. semantically equivalent, SE) relatively to the subject area considered.

Actual problem

How to find the most rational variant to express the meaning ?

Optimal sense transfer

Is provided by those phrases from initial set of equivalent-by-sense which are of minimal character length under a maximum of words most frequently used in all initial phrases.

Main purpose of research

Development and theoretical reasoning of methods and algorithms for seeking an optimal variant for sense transfer among experts and testees in a knowledge-control system that implements the open tests.

The most actual tasks

1. Thematic categorization of text documents.
2. Representation of topical areas by means of thesauruses and ontologies.

Expert tasks to be automated

1. Search for semantically equivalent forms for description of reality fragment in the given natural language. Here the fragment of actual expert knowledge corresponds to some fact of topical area.
2. Comparison of knowledge of given expert with the closest knowledge fragments of another experts.

Requirements for the solution

1. Revelation of concepts and relations between them in a given text.
2. Extraction from texts of corpus the usage contexts of general vocabulary by means of which synonymic paraphrases can be formed.

According to classic definition, TF-IDF is the product of two statistics: term frequence (TF) and inverse document frequency (IDF).

Term frequency estimates the significance of word t_i within the document d and can be defined as

$$\mathrm{tf}\left(t_i, d\right) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

where n_i is the number of times that t_i occurs in document d, and denominator contains the total number of words for d.

The value of IDF is unique for each unique word in corpus D and can be determined as follows:

$$\mathrm{idf}\left(t_i, D\right) = \log\left(\frac{|D|}{|D_i|}\right), \tag{2}$$

where numerator represents the total number of documents in corpus, and $|D_i \subset D|$ is a number of documents where the word t_i appears.

1. The words, which are the most unique in document and have the largest values of TF*IDF, must be related to terms of document's topical area.
2. The fact that the term has synonyms at the same document means the decrease of TF metrics for this word relatively to given document.
3. For words of general vocabulary and for those terms which are prevail in corpus the value of IDF tends to zero.
4. Synonyms, unique for some documents of corpus, will have a higher values of IDF.

   For example: general-vocabulary words which are define the conversive replacements, like *«приводить* $\Leftrightarrow$ *являться следствием»* (in Russian).

Clustreing the vocabulary of initial phrase by TF-IDF metrics

Let

X be an ordered descending sequence of $\mathrm{tf}\,(t,d)\cdot\mathrm{idf}\,(t,D)$ values for all words t of initial phrase relatively to document $d \in D$.

F be the sequence of clusters $H_1, \ldots H_r$ as a result of splitting the initial X by means of algorithm close to FOREL class taxonomy algorithms.

As the mass center of cluster H_i the arithmetic mean of all $x_j \in H_i$ is taken.

The *estimation of clustering quality* here can be defined as

$$Q\,(F) = \frac{\sum_{i=1}^{r} \mathrm{diam}\,(H_i)}{\mathrm{len}\,(F)} \Big(\mathrm{len}\,(F) - \max\,(F)\Big) \frac{\min\,(F)}{\max\,(F)}, \qquad (3)$$

where $\mathrm{diam}\,(H_i)$ is the width of cluster H_i;

$\min\,(F)$ and $\max\,(F)$ are minimal and maximal values of width for clusters represented in F;

$\mathrm{len}\,(F)$ is the length of F.

Let

D be clustered by analogy with X, but according to the values of function (3);
$D' \subset D$ be the cluster of greatest values of (3).

It is required to select phrases from documents $d \in D'$ according to the criteria of maximum number of words presented in clusters $\{H_1, H_{r/2}, H_r\} := Cl$:

H_1 — the *terms* from initial phrase which are the *most unique* for d;

$H_{r/2}$ — *general vocabulary* as a basis of *synonymic paraphrases*, and those *terms* which have *synonyms*;

H_r — those *terms* which are *prevail* in corpus.

Representation estimation of words of phrase $s \in d$, $d \in D'$, in clusters from Cl

$$\mathrm{N}\left(s, Cl\right) = \frac{\sqrt{\sum_{j \in \{1, r/2, r\}} \left|\left\{t_i \in s \colon \mathrm{tfidf}\left(t_i, d, D\right) \in H_j\right\}\right|^2}}{\sigma\left(\left|\left\{t_i \in s \colon \mathrm{tfidf}\left(t_i, d, D\right) \in H_j\right\}\right|\right) + 1}, \qquad (4)$$

where the first summand in denominator is the root-mean square deviation of number of words presented in cluster from Cl and related to a phrase from d.

The test corpus for proposed method includes Russian papers published in:

- Taurida journal of computer science theory and mathematics (TJCSTM, 3 papers);
- Proceedings of International conferences «Intelligent Information Processing» IIP-8 and IIP-9 (2 papers);
- Proceedings of All-Russian Conference with International Participation on Mathematical Methods for Pattern Recognition (MMPR-15, 1 paper);
- Proceedings of the Conference MMPR-13 (2 papers);
- Proceedings of the Conference MMPR-16 (14 papers);
- Proceedings of the Conference IIP-10 (2 papers).

Remark 1

In addition, the corpus included the text of a scientific report prepared in 2003 by Dmitry Mikhaylov.

Remark 2

The number of words in corpus documents varied from 218 to 6298.

- mathematical methods for learning by precedents (K. Vorontsov, M. Khachay, E. Djukova, N. Zagoruiko, Yu. Dyulicheva, I. Genrikhov, A. Ivakhnenko);
- methods and models of pattern recognition and forecasting (V. Mottl, O. Seredin, A. Tatarchuk, P. Turkov, M. Suvorov, A. Maysuradze);
- intelligent processing of experimental information (S. Dvoenko, N. Borovykh);
- image processing, analysis, classification and recognition (A. Zhiznyakov, K. Zhukova, I. Reyer, D. Murashov, N. Fedotov, V. Martyanov, M. Kharinov).

№	Initial phrase
1	*Переобучение приводит к заниженности эмпирического риска.*
2	*Переподгонка приводит к заниженности эмпирического риска.*
3	*Переподгонка служит причиной заниженности эмпирического риска.*
4	*Заниженность эмпирического риска является результатом нежелательной переподгонки.*
5	*Переусложнение модели приводит к заниженности средней ошибки на тренировочной выборке.*
6	*Переподгонка приводит к увеличению частоты ошибок дерева принятия решений на контрольной выборке.*
7	*Переподгонка приводит к заниженности оценки частоты ошибок алгоритма на контрольной выборке.*
8	*Заниженность оценки ошибки распознавания связана с выбором правила принятия решений.*
9	*Рост числа базовых классификаторов ведёт к практически неограниченному увеличению обобщающей способности композиции алгоритмов.*

software implementation and experimental results

Clusters for phrases selection:

K. Vorontsov, TJCSTM 2004 №1, words presented in clusters	
H_1	*алгоритм, обобщать, способность*
$H_{r/2}$	*классификатор, увеличение, число*
H_r	*вести*
K. Vorontsov, MMPR-15, words presented in clusters	
H_1	*алгоритм*
$H_{r/2}$	*рост, композиция*
H_r	*неограниченный, базовый, увеличение*

Results (contain the words *обобщать, способность, алгоритм*):

Selected phrase	Expressed relations
Обобщающая способность ***определяется как*** *вероятность ошибки найденного алгоритма,* ***либо как*** *частота его ошибок на неизвестной контрольной выборке, также случайной, независимой и одинаково распределённой*	**The definition of** *generalizing capability of algorithm* **relates with the concepts of** *error probability* **and** *rate (frequency) of errors in control sample*
Результатом обучения является не только сам алгоритм, но и достаточно точная оценка его обобщающей способности	*ведёт к* $\Longleftrightarrow$ *является результатом*

Clusters for phrases selection:

K. Vorontsov, TJCSTM 2004 №1, words presented in clusters	
H_1	*риск, эмпирический*
$H_{r/2}$	*заниженность, являться, переподгонка*
H_r	*нежелательный*
K. Vorontsov, MMPR-15, words presented in clusters	
H_1	*риск*
$H_{r/2}$	*результат*
H_r	*нежелательный, заниженность, переподгонка*
Yu. Dyulicheva, TJCSTM 2002 №1, words presented in clusters	
H_1	*переподгонка*
$H_{r/2}$	*являться*
H_r	*нежелательный, заниженность, риск*

Selected phrase (contains the words *эмпирический, риск, являться, заниженность*)**:**
Причиной является всё то же переобучение, которое приводит к заниженности эмпирического риска.

Synonymic terms: *переподгонка* $\Longleftrightarrow$ *переобучение*

Variant of conversive replacement: *результат* $\Longleftrightarrow$ *причина*

Clusters for phrases selection:

K. Vorontsov, TJCSTM 2004 №1, ranges of values for TF-IDF	
H_1	$0{,}0020 \ldots 0{,}0026$
$H_{r/2}$	$1{,}4386 \cdot 10^{-4} \ldots 2{,}1839 \cdot 10^{-4}$
H_r	$0{,}0000 \ldots 0{,}0000$
K. Vorontsov, MMPR-15, ranges of values for TF-IDF	
H_1	$0{,}0021 \ldots 0{,}0021$
$H_{r/2}$	$4{,}3890 \cdot 10^{-4} \ldots 4{,}3890 \cdot 10^{-4}$
H_r	$0{,}0000 \ldots 0{,}0000$
Yu. Dyulicheva, TJCSTM 2002 №1, ranges of values for TF-IDF	
H_1	$0{,}0040 \ldots 0{,}0040$
$H_{r/2}$	$1{,}7015 \cdot 10^{-4} \ldots 1{,}7015 \cdot 10^{-4}$
H_r	$0{,}0000 \ldots 0{,}0000$

TF (concerning [K. Vorontsov, TJCSTM 2004 №1]) and IDF values for words of initial phrase №4:

word	нежелательный	заниженность	переподгонка	являться	результат	эмпирический	риск
TF	0,0000	$1{,}5623 \cdot 10^{-4}$	$1{,}5623 \cdot 10^{-4}$	0,0031	0,0022	0,0033	0,0028
IDF	1,3979	1,3979	0,9208	0,0555	0,1938	0,6198	0,9208
TF-IDF	0,0000	$2{,}1839 \cdot 10^{-4}$	$1{,}4386 \cdot 10^{-4}$	$1{,}7347 \cdot 10^{-4}$	$4{,}2392 \cdot 10^{-4}$	0,0020	0,0026

Clusters for phrases selection:

K. Vorontsov, TJCSTM 2004 №1, ranges of values for TF-IDF		
H_1	*оценка, ошибка*	0,0019 . . . 0,0029
$H_{r/2}$	*заниженность*	$2{,}1839 \cdot 10^{-4} \ldots 2{,}1839 \cdot 10^{-4}$
H_r	*с, принятие*	0,0000 . . . 0,0000
Yu. Dyulicheva, TJCSTM 2002 №1, ranges of values for TF-IDF		
H_1	*ошибка*	0,0068 . . . 0,0068
$H_{r/2}$	*решение, распознавание, принятие*	$3{,}0603 \cdot 10^{-4} \ldots 3{,}7303 \cdot 10^{-4}$
H_r	*заниженность, с, связанный*	0,0000 . . . 0,0000
Yu. Dyulicheva, TJCSTM 2003 №2, ranges of values for TF-IDF		
H_1	*решение, распознавание, принятие*	0,0017 . . . 0,0018
$H_{r/2}$	*правило*	$4{,}2541 \cdot 10^{-4} \ldots 4{,}2541 \cdot 10^{-4}$
H_r	*заниженность, с*	0,0000 . . . 0,0000

Selected phrase:

Сравнивая прогнозируемый коэффициент ошибки t с ошибками ветви $T(t)$ и наибольшей из ветвей с корнем в дочерней вершине вершины t, принимается решение о том оставлять без изменений $T(t)$, редуцировать или наращивать в вершине t [Yu. Dyulicheva, TJCSTM 2002 №1].

1. Usage the additional knowledge about semantic relations and their textual expressional forms to search the words related with the given ones.

   Experiment with the Serelex system:

   - for initial phrase №8 a single link *«решение — с»* was found;
   - for initial phrase №9 it was not found any link.

   The collection of documents involved by system included:

   - headers of Wikipedia articles ($2{,}026 \cdot 10^9$ word forms and $3\,368\,147$ lemmas);
   - ukWaC text corpus ($0{,}889 \cdot 10^9$ word forms and $5\,469\,313$ lemmas).

   Disadvantage:

   - the *subject-oriented classification* of vocabulary *is not provided*, what *complicates to apply* the *lexico-syntactic patterns* implemented by system, for selection of required fragments in corpus texts.

2. WordNet-like thesaurus:

   - the *synonymy degree* within each *set of cognitive synonyms* (synset) is actually *depends* on *subject orientation of words* which constitute it.

3. Usage of summary TF-IDF for initial phrase's words occurring at the document phrase, as an alternative of estimation (4).

   Disadvantage: too low percent (less than 2%) of general vocabulary in selected phrases to release the synonymic paraphrases of initial phrase.

1. The main *result* of current work is the *search method* for descriptions of close knowledge fragments and their linguistic expressional means represented in text corpus.
2. Besides the design of open tests, another important *application scope* of the offered method is the problem-oriented thesauruses conceptually close to «Black Square» developed by Dorodnicyn Computing Centre of RAS.
3. In comparison with known approaches the offered method *allows* to reveal concepts and relations between them concerning the given topical area on the base of lesser training samples and without predefined orientation on the certain types of relations of words in initial phrases.

1. Elaboration of the numerical estimation which respects simultaneously:
   - the quality of extraction of themes (topics) as a sets of topical area's terms which are co-occur in documents;
   - the singularity of term distribution for a topic;
   - the singularity of topic distribution for a document..
2. How does predictability of occurrence for words in document phrases can be related with the structure of clusters that are formed according to TF-IDF values for words of initial phrase ?
3. Taking into account the potential syntactic contexts for multiple-valued words.