

УДК 51-74

ВЫЧИСЛИТЕЛЬНАЯ СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ТЕКСТОВ НАУЧНО-ТЕХНИЧЕСКОГО СТИЛЯ

Вишняков Ю.М., Вишняков Р.Ю.

НАН ЧОУ ВО «Академия маркетинга и социально-информационных технологий – ИМСИТ»,
Краснодар, e-mail: y_vishnyakov@inbox.ru

В предлагаемой работе развивается формальный подход к семантической интерпретации текстов научно-технического стиля на основе вычислительной процедуры, которая представляется в виде семантической схемы. Для этого вводятся операция контекстного вычисления смысла, контекстная связка, представления фрагментов текста в обратной польской записи. Предлагается семантическую близость фрагментов текстов определять на основе результатов сравнения их семантических схем, используя сконструированные в работе критерии семантической близости и процедуру сравнения. Показывается, что семантический критерий близости включает в себя в качестве частного случая традиционный частотный критерий. Кроме того, в предлагаемой работе конструируется лингвистическая переменная близости документов, правила вывода и процедура включения релевантных документов в информационную выдачу. Семантическое сравнение иллюстрируется конкретными примерами, демонстрирующими результативность предлагаемого подхода.

Ключевые слова: семантическая интерпретация, семантическая близость, семантическая схема, обратная польская запись, информационный поиск, классификация, нечеткая математика

COMPUTATIONAL SEMANTIC INTERPRETATION OF SCIENTIFIC AND TECHNICAL TEXTS

Vishnyakov Yu.M., Vishnyakov R.Yu.

Non-state Academy of Marketing and Social Information Technologies – IMSIT,
Krasnodar, e-mail: y_vishnyakov@inbox.ru

The paper presents a formal approach to the semantic interpretation of scientific and technical texts. The identification of text semantics is performed by means of a computational procedure represented as a semantic scheme. For this purpose, the special concepts such as operation of semantic value context definition, context bunch as well as text chunk representation as expressions of Reverse Polish Notation were introduced. The approach suggests defining semantic similarity of text chunks by results of comparison of their semantic schemes, by applying the criterion of semantic similarity and comparison procedure. In this paper we show that semantic similarity criteria include a traditional frequency criterion as a special case. Furthermore, we constructed a linguistic variable of documents similarity, decision rules and selection procedure of relevant documents. Also, we provide semantic comparisons of text chunks and illustrate them in the examples. These examples show the performance of the described approach.

Keywords: semantic interpretation, semantic proximity, semantic scheme, Reverse Polish notation, information retrieval, classification, fuzzy mathematics

В настоящее время в частотной парадигме релевантности по эффективности обработки информации фактически достигнут предел, однако рост объемов информации и возрастающие требования к качеству ее обработки усиливают внимание к использованию семантических подходов. Данное обстоятельство особенно явно проявляется в областях информационного поиска и классификации документов. В то же время использованию семантических подходов препятствует отсутствие приемлемой формализации, что обусловлено сложностью и неоднозначностью естественных языков (ЕЯ), порождающих многочисленные исследования в данном направлении, например [3, 4]. Предлагаемая работа отражает часть наших исследований по семантической интерпретации текстов научно-технического стиля, обобщая и развивая, в частности, результаты работ [1, 2, 5].

Метод

Суть предлагаемого подхода состоит в формировании смысла текстового фрагмента (ТФ) с помощью вычислительной процедуры над смысловыми значениями его слов. Процедура конструируется в виде семантической схемы (СемСх), которая положена в основу сравнения ТФ на смысловую близость, а для сравнения специально сконструирован семантический критерий близости (СКБ).

А. Основные термины и понятия

Пусть в ЕЯ задана некоторая цепочка q слов a_i вида

$$q = a_1 a_2 \dots a_n, \quad (1)$$

представляющая целостный по смыслу ТФ. Целостность означает, что между словами цепочки q существует отношение непосредственного подчинения, оно задается словосочетаниями и является однозначным.

Если в некотором словосочетании a – главное слово, а b – зависимое, то эту зависимость представим выражением (записью) вида

$$\overrightarrow{a:b}, \quad (2)$$

в котором стрелка указывает на направление зависимости слов.

Если a – некоторое слово, входящее главным словом в несколько словосочетаний с зависимыми словами $b_1, b_2, \dots, b_p$, тогда зависимость запишем выражением вида

$$a: \{b_1, b_2, \dots, b_p\}, \quad (3)$$

которое назовем его контекстной связкой.

Обозначим через $S(a_i)$ множество смысловых значений слова a_i из q (1), а множество смысловых значений ФТ (1) представим функционалом вида

$$S(q) = \Phi(S(a_1), S(a_2), \dots, S(a_n)). \quad (4)$$

Если в ТФ q слово a_i является главным, то для (4) справедливо соотношение

$$S(q) \subset S(a_i). \quad (5)$$

$$S(\overrightarrow{a:(b_1, \dots, b_p)}) = S(\overrightarrow{a:b_1}) \cap S(\overrightarrow{a:b_2}) \cap \dots \cap S(\overrightarrow{a:b_p}). \quad (9)$$

Теперь с учетом (8) и (9) для ТФ можно уже строить множество смысловых значений (4). Поясним рассуждения примером ТФ q , где слова закодированы латинскими буквами:

$$q = \underbrace{\text{“международное”}}_a \underbrace{\text{признание”}}_b \underbrace{\text{образовательных”}}_c \underbrace{\text{программ”}}_d \underbrace{\text{российских”}}_e \underbrace{\text{вузов”}}_f$$

Используя (8) и (9) для примера ТФ q , получим выражение его смыслового значения:

$$S(q) = (S(a) \bar{\cap} S(b)) \cap (S(b) \bar{\cap} ((S(d) \bar{\cap} S(c))) \cap (S(d) \bar{\cap} (S(e) \bar{\cap} S(f))). \quad (10)$$

В. Обратная польская запись функционала текстового фрагмента, семантическая схема

Основную особенность обратной польской записи (ОПЗ) представляет способ ее вычисления, он же однозначно определяет структуру вычислительной процедуры. Используем данное обстоятельство и распространим нотацию ОПЗ на функционал смысла.

В нотации ОПЗ операция контекстного уточнения смысла (8) будет иметь вид

$$S(\overrightarrow{a:b}) = S(b)S(a)\bar{\cap}. \quad (11)$$

В (11) для сохранения зависимости слов стрелка над операцией меняет направление. Также ОПЗ контекстной связки (9) с учетом (11) принимает следующий вид:

$$S(\overrightarrow{a:(b_1, \dots, b_p)}) = \underbrace{S(b_1)S(a)}_1 \bar{\cap} \underbrace{S(b_2)S(a)}_2 \bar{\cap} \dots \underbrace{S(b_p)S(a)}_p \bar{\cap} p \bar{\cap}, \quad (12)$$

где p соответствует арности операции пересечения.

Применяя (11) и (12) к ТФ q (10), получим следующую ОПЗ:

$$S(q) = S(a)S(b)\bar{\cap} S(c)S(d)\bar{\cap} S(f)S(e)\bar{\cap} S(d)\bar{\cap} \cap S(b)\bar{\cap} \cap. \quad (13)$$

Таким образом, смысловое значение словосочетания есть подмножество смысловых значений его главного слова, задается оно смыслом зависимого слова, образующего контекст:

$$S(\overrightarrow{a:b}) = S(a)|_{S(b)}. \quad (6)$$

Для (6) справедливы соотношения

$$S(\overrightarrow{a:b}) \subset S(a); \quad S(\overrightarrow{a:b}) \neq S(\overrightarrow{b:a});$$

$$S(\overrightarrow{a:a}) = S(a). \quad (7)$$

На словосочетании следующим образом определим операцию контекстного уточнения смысла главного слова зависимым:

$$S(\overrightarrow{a:b}) = S(a) \bar{\cap} S(b), \quad (8)$$

где $\bar{\cap}$ – операция контекстного уточнения смысла, а стрелка над ней задает направление зависимости слов в словосочетании. Распространяя (8) на контекстную связку (9), получим

Промежуточные выражения пошагового вычисления ОПЗ (13) имеют вид

1. $S(q) = \underbrace{S(a)S(b) \bar{\cap} S(c)S(d) \bar{\cap} S(f)S(e) \bar{\cap} S(d) \bar{\cap} S(b) \bar{\cap} \cap}_{r1}.$
2. $S(q) = r_1 \underbrace{S(c)S(d) \bar{\cap} S(f)S(e) \bar{\cap} S(d) \bar{\cap} S(b) \bar{\cap} \cap}_{r2}.$
3. $S(q) = r_1 r_2 \underbrace{S(f)S(e) \bar{\cap} S(d) \bar{\cap} S(b) \bar{\cap} \cap}_{r3}.$
4. $S(q) = r_1 r_2 r_3 \underbrace{S(d) \bar{\cap} S(b) \bar{\cap} \cap}_{r5}.$
5. $S(q) = r_1 r_5 \underbrace{S(b) \bar{\cap} \cap}_{r6}.$
6. $S(q) = r_1 r_6 \cap_{r7}.$
7. $S(q) = r_7.$

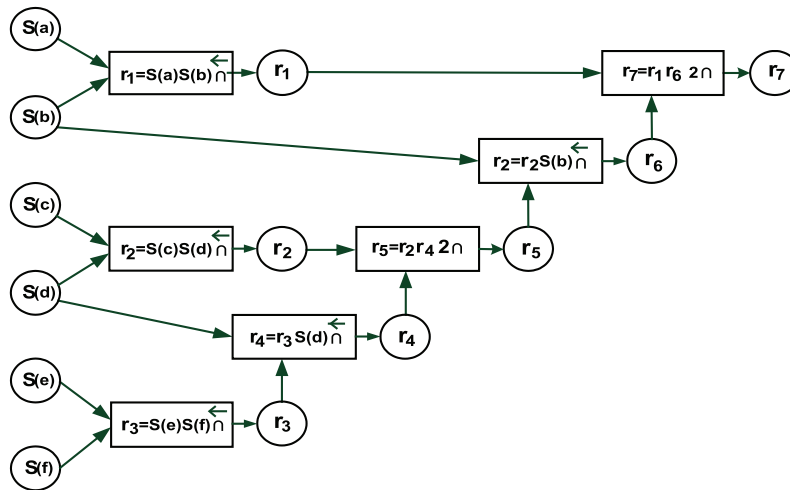


Рис. 1. Семантическая схема ТФ q

Здесь переменные r_i являются временными и предназначены для хранения промежуточных результатов каждого шага. Пока важен только факт их существования.

Поставим в соответствие каждой операции графический элемент, в котором прямоугольник представляет операцию, круглые вершины слева – входные данные, справа – результат. Такое представление назовем элементом смысла. Объединяя элементы смысла выражения (13), получим графическое представление всей вычислительной процедуры, которое назовем семантической схемой (СемСх). Вычислительная схема, построенная для примера ТФ в виде ОПЗ (13), представлена ниже на рис. 1 и имеет следующий вид:

С. Критерий сравнения текстовых фрагментов на семантическую близость

Пусть задан ТФ q в виде (1), который назовем образцом сравнения. Данный

образец сравним на семантическую близость с ТФ t вида

$$t = b_1 b_2 \dots b_m, \quad (14)$$

для чего введем критерий семантической близости (КСБ) следующим образом:

$$Cprox(S(q), S(t)) \in D; D = [0...1]. \quad (15)$$

Здесь $Cprox$ – КСБ, $S(q)$ – множество смысловых значений образца, $S(t)$ – множество смысловых значений ТФ t , а D – интервал значений КСБ $Cprox$. Если $Cprox = 0$, то близость между q и t отсутствует, при $Cprox = 1$ имеет место полное семантическое совпадение.

Отметим, что в общем случае всегда справедливо соотношение

$$Cprox(S(q), S(t)) \neq Cprox(S(t), S(q)). \quad (16)$$

Представим КСБ в виде доли совпадающих элементов смысла в СемСх образца q

и сравниваемого ТФ t к общему числу элементов смысла образца:

$$Cprox(S(q), S(t)) = p/m. \quad (17)$$

Здесь m – число элементов смысла в СемСх образца q , p – число совпадающих элементов смысла q в СемСх t соответственно. Частным случаем (17) является частотный критерий релевантности (φ), когда m представляет число слов в ТФ q , а p – число слов из q , совпадающих со словами в t .

Поясним семантическое сравнение образца q и на примерах конкретных ТФ. Одновременно для примеров сравним КСБ и частотный критерий релевантности.

Пример 1. Пусть задан первый ТФ_1 в виде предложения:

Правительство России проводит политику, направленную на международное признание образовательных программ российских вузов.

Здесь образец q является подстрокой ТФ_1 (выделено курсивом) и его СемСх полностью укладывается в СемСх ТФ_1. Соответственно $Cprox(S(q), S(t)) = 1$ и $\varphi = 1$.

Пример 2. Пусть задан второй ТФ_2 вида

Образовательные программы российских вузов нуждаются в международном признании.

В ТФ_2 подчеркнутым курсивом выделены подстроки, совпадающие с фрагментами образца.

Выделим совпадающие элементы смысла на СемСх образца жирными линиями,

тогда СемСх образца примет вид, показанный на рис. 2.

СемСх образца содержит 5 совпадающих элементов смысла, поэтому $Cprox(S(q), S(t)) = 5/7$ и его значение еще достаточно велико. В то же время критерий близости $\varphi = 1$, он изменения смысла не чувствует, поскольку все слова образца входят в ТФ_2.

Пример 3. Пусть задан третий ТФ_3 вида

Международное признание образовательных программ поднимает репутационный рейтинг российских вузов.

После сравнения СемСх образца представлена на рис. 3. Здесь нетрудно заметить, что СемСх образца и ТФ_3 имеют всего три совпадающих элемента смысла, поэтому $Cprox(S(q), S(t)) = 3/7$. Действительно, смысловая близость образца и ТФ_3 имеет именно эту оценку. Однако $\varphi = 1$ и он снова не чувствует изменения смысла.

Пример 4. Пусть задан четвертый ТФ_4 вида

Международное признание вузов увеличивает возможности России по привлечению иностранных студентов.

Результат сравнения образца q и ТФ_4 показан на рис. 4.

Из рис. 4 видно, что СемСх образца имеет один совпадающий элемент смысла и $Cprox(S(q), S(t)) = 1/7$ достаточно мал. В то же время значение $\varphi = 0,5$ еще высоко, поскольку половина слов образца входит во фрагмент текста.

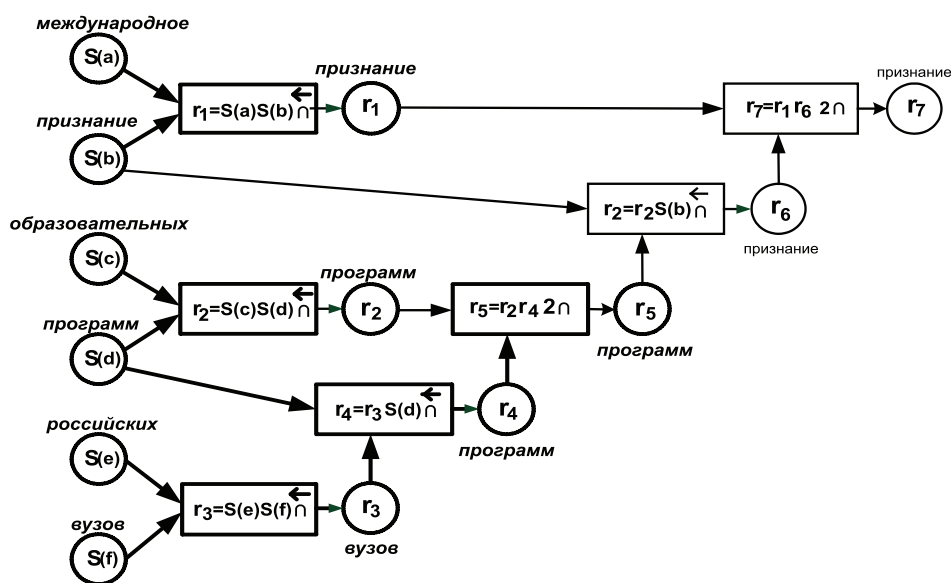
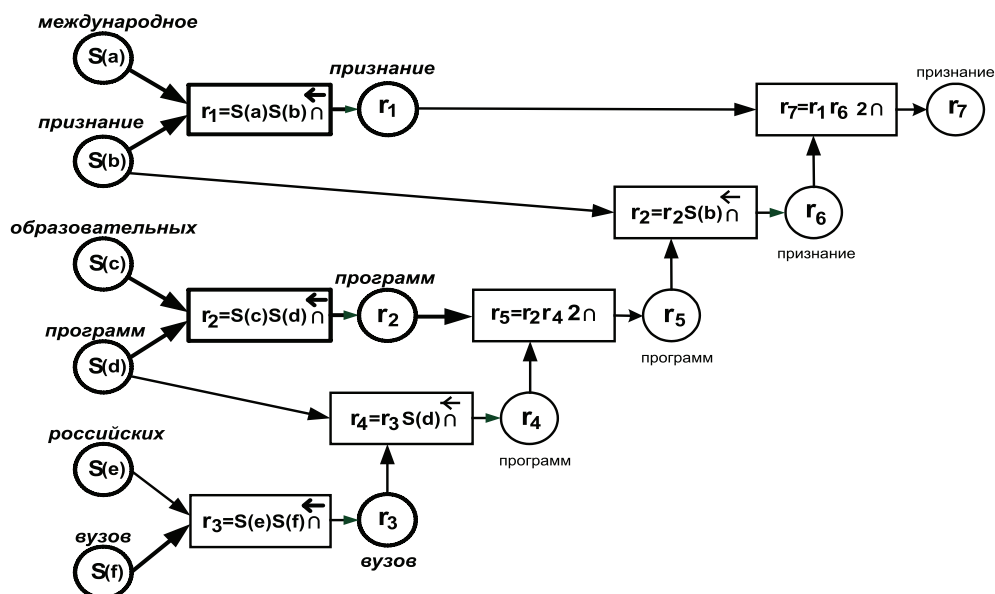
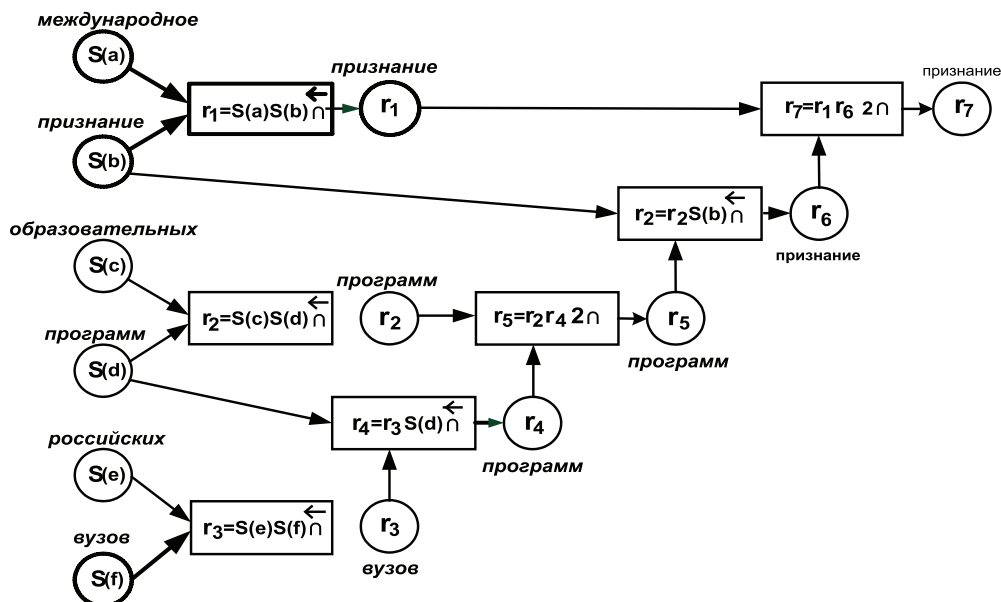


Рис. 2. Результат сравнения на близость образца q и ТФ_2

Рис. 3. Результат сравнения на близость образца q и TF_3 Рис. 4. Результат сравнения на близость образца TF_4

Пример 5. Рассмотрим пятый TF_5 : Международные программы академических обменов и участие в них российских ученых способствуют улучшению образовательного процесса вузов России.

Результат сравнения на близость $SемCх$ образца и $SемCх$ TF_5 показан ниже на рис. 5.

$SемCх$ образца не имеет совпадающих элементов смысла с $SемCх$ TF_5 , поэтому $Cprox(S(q), S(t)) = 0$. В то же время $\varphi = 0,83$, поскольку 5 слов образца из 6 входит в TF_5 .

Из приведенных примеров видно, что СКБ реально отражает семантическую близость образца и TF , в то время как значение частотного критерия не всегда отражает реальную семантическую близость. Особенно это явно проявляется в пятом примере.

D. Семантический информационный поиск и классификация документов.

Семантическую близость документов образцу (поисковому запросу) определим лингвистической переменной $SемProx$ вида

$$T = \{ \text{"Слабая"}, \text{"Относительно слабая"}, \text{"Достаточно сильная"}, \text{"Сильная"} \}, \quad (18)$$

где каждый терм терм-множества T является нечеткой переменной подынтервалов $[0...p_1)$, $[p_1...p_2)$, $[p_2...p_3)$, $[p_3...1]$ шкалы D соответственно.

Построим процедуру определения семантической близости документа запросу

следующим образом. Для каждого предложения документа определим значение $Cprox$. Далее для каждого подынтервала шкалы D подсчитаем число попавших в него значений $Cprox$, после чего построим из этих чисел характеристический вектор документа вида

$$\langle f_{\text{слабая}}, (f_{\text{относительно слабая}}, (f_{\text{достаточно сильная}}, (f_{\text{сильная}}) \rangle$$

Вычисление лингвистической переменной реализуем процедурой *Semantic_proximity* (Doc, *SemProx*), в которой Doc – анализируемый документ, *SemProx* – возвращаемое значение лингвистической близости. Вычисление проведем по решающим правилам вида

IF $((f_{\text{слабая}} > (f_{\text{достаточно сильная}} + f_{\text{сильная}})) \& (f_{\text{слабая}} > f_{\text{относительно слабая}}))$ THEN *SemProx* := “Слабая”;

IF $((f_{\text{относительно слабая}} > (f_{\text{достаточно сильная}} + f_{\text{сильная}})) \& (f_{\text{относительно слабая}} > f_{\text{слабая}}))$ THEN *SemProx* := “Относительно слабая”;

IF $((f_{\text{достаточно сильная}} > (f_{\text{слабая}} + f_{\text{относительно слабая}})) \& (f_{\text{достаточно сильная}} > f_{\text{сильная}}))$ THEN *SemProx* := “Достаточно сильная”;

IF $((f_{\text{сильная}} > (f_{\text{слабая}} + f_{\text{относительно слабая}})) \& (f_{\text{сильная}} > f_{\text{достаточно сильная}}))$ THEN *SemProx* := “Сильная”.

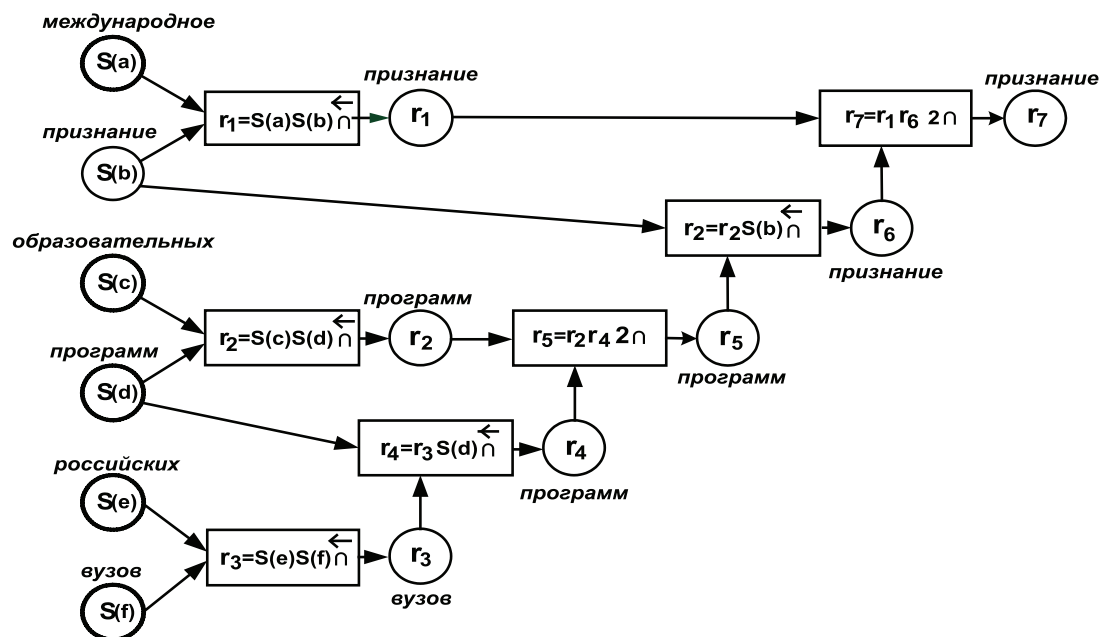


Рис. 5. Результат сравнения на близость образца q и $T\Phi_5$

Само включение документа в ВЫДАЧУ R реализуем простым способом:

BEGIN

WHILE (конец потока документов)

DO BEGIN

Semantic_proximity (Doc, *SemProx*);

IF (*Semprox*=(условие включения)) then $D \rightarrow R$;

END

END.

Здесь $\rightarrow$ включение документа D в ВЫ-ДАЧУ R . Введение условий для разных R превращает процедуру в семантический документный классификатор.

Заключение

Вычислительная модель семантической интерпретации на основе семантических схем позволяет перейти к формально математическому сравнению текстов на семантическую близость и организовать на этой основе высокоточный информационный поиск и классификацию документов. Особенно это актуально при семантическом различении текстов, использующих одну и ту же терминологию.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-07-00241\16.

Список литературы

1. Вишняков Ю.М., Вишняков Р.Ю. «Вычислительное» представление смысла текстовых фрагментов

на основе обратной польской записи // Известия ЮФУ. Технические науки. Тематический выпуск «Интеллектуальные САПР». – Таганрог: Изд-во ЮФУ. 2013. – № 7 (144). – С. 141–147.

2. Вишняков Ю.М., Вишняков Р.Ю. Интерпретационная модель смысла текстового фрагмента // Известия ЮФУ. Технические науки. Тематический выпуск «Интеллектуальные САПР». – Таганрог: Изд-во ЮФУ, 2013. – № 7 (144). – С. 152–156.

3. Ефименко И.В. Обработка естественно языковых текстов: онтологичность в лингвистике и дискурсивность в извлечении знаний // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006: тезисы докл. нац. конф. (Обнинск, 25–28 сент. 2006 г.) – М.: Физматлит, 2006. – Т. 2. – С. 230–234.

4. Dobrov B., Loukachevitch N. Multiple Evidence for Term Extraction in Broad Domains. Proceedings of the 8th Recent Advances in Natural Language Processing Conference (RANLP 2011). – Hissar, Bulgaria, 2011. – P. 710–715.

5. Yury M. Vishnyakov*, Renat Yu. Vishnyakov / The Linguistic Proximity in Information Retrieval and Document Classification. 14th IEEE International Symposium on Computational Intelligence and Informatics to be held on November 19–21, 2013 in Budapest, Hungary. – P. 131–134.