

Сила связи слов и оценка релевантности текста единице представления знаний в открытых тестах

Михайлов Д. В., Козлов А. П., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

11-я Международная конференция
«Интеллектуализация обработки информации» (ИОИ-2016),

10–14 октября 2016 г.

г. Барселона, Испания

Единица знаний, оцениваемая открытым тестом

Определяется множеством семантически эквивалентных фраз предметно-ограниченного естественного языка.

Оптимальная передача смысла

Обеспечивается теми фразами из исходного множества эквивалентных по смыслу, которые при минимальной символьной длине имеют максимум слов, наиболее употребимых во всех исходных фразах.

Основные проблемы:

- выделение единиц знаний из текстов тематического корпуса;
- отбор текстов в корпус анализом релевантности исходной фразе.

Формирование тематического корпуса для открытых тестов: основные проблемы

- 1 Отбор текстов в корпус, как правило, субъективен и зависит от эксперта.
- 2 При выборе критерия отбора текстов необходимо одновременно учитывать как уровень сложности текста, так и его значимость для составления теста по заданным фрагментам экспертного знания (например, с точки зрения тематической рубрикации).
- 3 Значимость текста в решаемой задаче может определять выбор меры его близости исходной фразе и в общем случае безотносительна к образу, представляемому исходной фразой в анализируемых текстах.
- 4 Сама исходная фраза лишь в единичных случаях соответствует эталону для сопоставления.

- Фрагмент анализируемого текста, отвечающий составляющей образа, отождествим с некоторой смысловой связью слов в исходной фразе.
- Сила связи слов каждого такого фрагмента всегда больше силы связи любого слова данного фрагмента и слова, не принадлежащего ему.
- Сочетания общей лексики и терминов, преобладающих в корпусе, в анализируемом тексте можно отнести к составляющим искомого образа только при наличии фрагментов с большей силой связи слов.
- В общем случае не выдвигается требование наличия в тексте строго заданной части составляющих образа исходной фразы.

Выделение образа исходной фразы в текстах формируемого корпуса

- Исследование встречаемости как отдельных слов, так и их сочетаний.
- Оценка «силы» связи слов относительно текста и корпуса..

Основная задача

Изучение вариантов численной оценки значимости («силы») связи слов и их использования для выделения составляющих образа исходной фразы.

Инструменты:

- частота L -грамм (по К. Шеннону);
- частота и фильтрация по тэгам;
- математическое ожидание и дисперсия.

Методы оценки статистической значимости словосочетания:

- t -критерий Стьюдента;
- критерий согласия Пирсона (так называемый критерий χ^2);
- критерий отношения правдоподобия.

Проблемы

- Требуется синтаксически размеченный корпус текстов для выделения биграмм, с которыми здесь ассоциируются словосочетания.
- Синтаксическая разметка текстов корпуса не поддаётся полной автоматизации и требует существенных временных затрат.
- Существующие корпуса в большинстве случаев не содержат требуемых данных по биграммам из анализируемых текстов.

- 1 Оценка значимости словосочетаний, выделяемых в тексте из n фраз [Biemann C., 2004]:

$$\text{sig}(A, B) = x - k \log(x) + \log k!, \quad (1)$$

где $x = \frac{ab}{n}$, a — число фраз, которые содержат слово A ,
 b — слово B , k — A и B одновременно.

Слабые стороны:

- для корректного применения оценки (1) каждое из слов пары (A, B) должно присутствовать минимум в одной фразе анализируемого текста;
 - как идейно близкая G -тесту для распределения Пуассона, оценка (1) может давать неточный результат при ожидаемом числе фраз документа менее 5.
- 2 Дистрибутивно-статистический метод построения тезаурусов — связь между совместно встречающимися словами:

$$K_{AB} = \frac{k}{a + b - k}. \quad (2)$$

Замечание

В целях предотвращения деления на ноль в случае, когда A и B не встречаются во фразах анализируемого текста отдельно друг от друга, k значению в знаменателе формулы (2) следует прибавлять единицу.

Пусть

D — исходное текстовое множество.

X — упорядоченная по убыванию последовательность ненулевых $\text{sig}(A, B)$ либо K_{AB} относительно документа $d \in D$ для пар слов (A, B) , которым в исходной фразе соответствуют синтаксические связи.

H_1, \dots, H_r — последовательность кластеров, на которые разбивается X алгоритмом, содержательно близким алгоритмам класса FOREL.

Центром масс кластера H_i возьмём среднее арифметическое всех $x_j \in H_i$.

Функцию ранжирования документов определим как

$$W(d) = K_{\Sigma}(d) \frac{K_1(d)}{K_{\Sigma}(d)} = K_1(d), \quad (3)$$

где $K_{\Sigma}(d)$ — суммарная величина оценки «силы» относительно d для всех найденных в исходной фразе связей (A, B) ;

$K_1(d)$ — то же самое для связей, отнесённых к кластеру H_1 наибольших значений «силы».

Пусть

D разбивается на кластеры по аналогии с X , но по значению функции (3);
 $D' \subset D$ — кластер наибольших значений оценки (3).

Следуя терминологии поисковых систем, назовём далее поиск фраз, близких исходной, в документах $d \in D'$ *построением аннотации*.

Варианты отбора фраз в аннотацию:

- по числу найденных во фразе связей, отвечающих кластеру H_1 ;
- по суммарному значению «силы» указанных связей.

Сравнение с идейно близкими подходами

- Поиск нечетких дубликатов документов — мерой сходства двух документов служит отношение числа общих подстрок фиксированной длины (в нашем случае эта длина была бы равна двум) к размеру документа (в словах) [Manber U., 1994; Heintze N., 1996].
- Отличие от контекстно-зависимого аннотирования [Яндекс, 2008] — одна аннотация строится сразу для нескольких документов.

Пусть L есть последовательность биграмм — пар синтаксически связанных слов (A, B) исходной фразы, упорядоченная по убыванию силы связи относительно некоторого документа $d \in D$, $\{(A_1, B_1), (A_2, B_2)\} \subset L(d)$.

Определение 1

Биграммы (A_1, B_1) и (A_2, B_2) войдут в одну n -грамму $T \subseteq L(d)$, если

$$((A_1 = A_2) \vee (B_1 = B_2) \vee (A_1 = B_2) \vee (B_1 = A_2)) = \text{true}.$$

Оценка суммарной силы связи слов в составе T относительно d

$$N(T, d) = \frac{\sqrt{\sum_{i=1}^{\text{len}(T)} [S_i(d)]^2}}{\sigma(S_i(d)) + 1}, \quad (4)$$

где $S_i(d)$ — сила связи слов i -й биграммы относительно d ;

$\sigma(S_i(d))$ — среднеквадратическое отклонение указанной величины;

$\text{len}(T)$ — длина n -граммы T (в биграммах).

Обозначим далее множество n -грамм $\{T: T \subseteq L(d)\}$ как $\mathbb{T}(d)$.

Пусть D — исходное текстовое множество.

Функция ранжирования документов $d \in D$ оценкой найденных n -грамм:

$$W(d) = \frac{1}{|\mathbb{T}(d)|} \left[\sum_{T \in \mathbb{T}(d)} N(T, d) \right] \left[|\mathbb{T}(d)| - \max_{T \in \mathbb{T}(d)} \text{len}(T) \right] \frac{\min_{T \in \mathbb{T}(d)} N(T, d)}{\max_{T \in \mathbb{T}(d)} N(T, d)}. \quad (5)$$

Множество D разбивается на кластеры по значению функции (5).

Пусть $D' \subset D$ — кластер наибольших значений оценки (5).

Аналогично по значению функции (4) разбивается $\mathbb{T}(d)$ для $\forall d \in D'$, $\mathbb{T}'(d)$ — кластер наибольших значений оценки (4) по заданному d .

Для каждой фразы s каждого $d \in D'$ вводится оценка

$$Q(s) = \left| \{w \in b : \exists T \in \mathbb{T}'(d), b \in T\} \right| \quad (6)$$

как основа разбиения на кластеры всего множества $\{s : s \in d \mid d \in D'\}$.

Фразы аннотации

Составляют первый кластер из полученных по значению оценки (6).

Основные критерии

- Исходные фразы формулируются независимо друг от друга разными экспертами.
- Исходные множества текстов подбираются так, чтобы сравнить образы исходной фразы, выделяемые в текстах на основе оценки силы связи встречающихся в их фразах слов исходной фразы и на основе меры TF-IDF этих слов.
- Максимально полная и наглядная иллюстрация выявления в текстах контекстов использования как слов-терминов, так и общей лексики, обеспечивающей синонимические перифразы исходной фразы.
- Число фраз в текстовом документе — не менее пяти.

- 1 статья в журнале «Вестник Российского экономического университета им. Г. В. Плеханова (Вестник РЭУ)»;
- 1 статья в журнале «Философия науки»;
- материалы тезисов четырёх докладов на 4-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (ИИ ФМИ, 2010 г.);
- материалы тезисов двух секционных и одного пленарного доклада на 7-й Всероссийской конференции ИИ ФМИ, 2013 г.;
- материалы одного пленарного доклада на 8-й Всероссийской конференции ИИ ФМИ, 2014 г.;
- 1 статья в сборнике трудов 9-й Всероссийской конференции ИИ ФМИ, 2015 г.;
- 1 статья в журнале «Таврический вестник информатики и математики (ТВИМ)».

Примечание

Число слов в документах исходного множества здесь варьировалось от 618 до 3765, число фраз — от 38 до 276.

№ Исходная фраза

- 1 *Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода.*
- 2 *Под знанием понимается система суждений с принципиальной и единой организацией, основанная на объективной закономерности.*
- 3 *С точки зрения искусственного интеллекта знание определяется как формализованная информация, на которую ссылаются или используют в процессе логического вывода.*
- 4 *Факты обычно указывают на хорошо известные обстоятельства в данной предметной области.*
- 5 *Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики.*
- 6 *Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.*
- 7 *Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.*
- 8 *Отличительными чертами логических моделей являются единственность теоретического обоснования и возможность реализации системы формально точных определений и выводов.*
- 9 *Язык представления знаний на основе фреймовой модели наиболее эффективен для структурного описания сложных понятий и решения задач, в которых в соответствии с ситуацией желательно применять различные способы вывода.*

- 3 статьи в журнале «Таврический вестник информатики и математики»;
- 2 статьи в сборниках трудов конференций «Интеллектуализация обработки информации» 2010 и 2012 гг.;
- 1 статья в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов» (2011 г.);
- материалы тезисов двух докладов на 13-й Всероссийской конференции «Математические методы распознавания образов» (2007 г.);
- материалы тезисов четырнадцати докладов на 16-й Всероссийской конференции «Математические методы распознавания образов» (2013 г.);
- материалы тезисов двух докладов на конференции «Интеллектуализация обработки информации» 2014 г.;
- материалы одного научного отчёта (Михайлов Д. В., 2003 г.).

Примечание

Число слов в документах исходного множества здесь варьировалось от 218 до 6298, число фраз — от 9 до 587.

Исходное множество текстов: тематика отбираемых работ для варианта 2

- математические методы обучения по прецедентам (К. В. Воронцов, М. Ю. Хачай, Е. В. Дюкова, Н. Г. Загоруйко, Ю. Ю. Дюличева, И. Е. Генрихов, А. А. Ивахненко);
- модели и методы распознавания и прогнозирования (В. В. Моттль, О. С. Середин, А. И. Татарчук, П. А. Турков, М. А. Суворов, А. И. Майсурадзе);
- интеллектуальный анализ экспериментальных данных (С. Д. Двоенко, Н. И. Боровых);
- обработка, анализ, классификация и распознавание изображений (А. Л. Жизняков, К. В. Жукова, И. А. Рейер, Д. М. Мурашов, Н. Г. Федотов, В. Ю. Мартьянов, М. В. Харинов).

Некоторые технические детали

- Для вычисления предлагаемых оценок приведение слов к начальной форме выполнялось с помощью функции *getNormalForms* в составе [библиотеки русской морфологии](#).
- Выделение синтаксических связей реализовано на основе правил, задействованных в работе [Царьков С. В., *Естественные и технические науки*, 2012, № 6].
- Распознавание границ предложений в тексте по знакам препинания — с помощью обученной модели классификатора, построенного с применением интегрированного пакета [Apache OpenNLP](#).
- Обучение распознаванию границ предложений — на основе размеченных данных из [Leipzig Corpora](#) (газетные тексты на русском языке, 2010 г., всего 10^6 фраз).

№ Исходная фраза

- 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.*

Программная реализация и результаты экспериментов

Отбираемая фраза	Что представляет	«Наиболее сильные» связи	Оценка
Хачай М. Ю., ММРО-16			
<p><i>Системы ограничений, возникающие в задачах принятия решений, оптимизации, распознавания образов и анализа часто являются несовместными, подразумевающими те или иные подходы к их коррекции, связанной с обобщением классического понятия решения</i></p>	<p><i>Связь обобщения классического понятия решения и выбора правила принятия решений</i></p>	<p><i>распознавание – с, принятие – решение</i></p>	K_{AB}
<p><i>Современная теория комитетных решений и тесно связанных с ними комитетных методов обучения распознаванию опирается на фундаментальные результаты, полученные Вл. Д. Мазуровым</i></p>	<p><i>Связь понятия распознавание из исходной фразы с понятием обучение</i></p>		K_{AB}
Дюличева Ю. Ю., ТВИМ 2003 №2			
<p><i>Эмпирический решающий лес повысил эффективность распознавания объектов, не участвовавших ранее в обучении, по сравнению с одним решающим деревом, при использовании одного и того же критерия ветвления</i></p>	<p><i>Рассуждение о решающем дереве и лесе как способах представления решающих правил</i></p>	<p><i>распознавание – с, принятие – решение</i></p>	$\text{sig}(A, B)$
<p>Здесь:</p>			
K_{AB}	— фразы найдены только по максимуму числа «наиболее сильных» связей;		
$K_{AB}, \text{sig}(A, B)$	— как по максимуму числа «наиболее сильных» связей, так и по их суммарной силе.		

Кластеры по TF-IDF для отбора фраз		Оценка	«Наиболее сильные» связи
Воронцов К. В., ТВИМ 2004 №1, слова, представленные в кластерах		Дюличева Ю. Ю., ММРО-13	
H_1	<i>алгоритм,</i> <i>обобщать, способность</i>	K_{AB}	<i>увеличение – обобщать,</i> <i>увеличение – способность,</i> <i>обобщать – способность</i>
$H_{r/2}$	<i>κ, классификатор,</i> <i>увеличение</i>		
H_r	<i>вести</i>		
Воронцов К. В., ММРО-15, слова, представленные в кластерах		Воронцов К. В., ТВИМ 2004 №1	
H_1	<i>алгоритм</i>	$\text{sig}(A, B)$	<i>обобщать – способность</i>
$H_{r/2}$	<i>рост, композиция</i>		
H_r	<i>неограниченный,</i> <i>базовый, увеличение</i>		

Для сравнения: фраза, отобранная по TF-IDF и не выделенная по $\text{sig}(A, B)$: Наиболее общая теория *алгоритмических композиций* разработана в алгебраическом подходе κ построению корректных алгоритмов, предложенном академиком РАН Ю. И. Журавлёвым и активно развиваемом его учениками.

Не отнесены к «наиболее сильным»: композиция – алгоритм, вести – κ

Отбор релевантных фраз: сравнение с методом на основе TF-IDF

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
<i>на основе TF-IDF слов исходной фразы</i>										<i>по числу «наиболее сильных» связей по величине sig(A, B)</i>								
N	1	1	1	1	3	2	4	1	40	1	1	11	11	5	20	9	10	19
N_1	1	1	1	1	0	0	0	0	7	1	1	1	2	0	1	0	0	2
N_2	0	1	1	1	3	0	0	0	6	0	1	1	1	1	1	1	0	1
N_3	0	0	0	0	1	1	1	0	8	0	0	4	4	0	0	5	1	7
<i>по числу «наиболее сильных» связей по величине K_{AB}</i>										<i>по суммарной «силе» для «наиболее сильных» по sig(A, B)</i>								
N	1	1	15	15	5	11	1	1	1	9	9	1	1	1	1	6	3	8
N_1	1	1	3	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0
N_2	0	1	2	2	1	9	0	0	1	0	0	0	0	0	0	0	0	0
N_3	0	0	7	4	0	4	0	1	0	0	0	0	1	0	0	1	1	2
<i>по суммарной «силе» для «наиболее сильных» по K_{AB}</i>										<i>N — общее число отобранных фраз:</i>								
N	10	9	2	2	8	6	2	2	1	<i>N_1 — фраз, представляющих</i>								
N_1	0	0	0	0	1	0	0	0	1	<i>выразительные средства языка;</i>								
N_2	0	0	0	0	1	4	0	0	1	<i>N_2 — представляющих синонимы;</i>								
N_3	1	0	1	0	1	2	0	2	0	<i>N_3 — представляющих связи понятий.</i>								

Кластеры по значению меры TF-IDF для отбора фраз:

Янковская А. Е., ТВИМ 2004 №1, слова, представленные в кластерах	
H_1	различный
$H_{r/2}$	применять, модель, наиболее, ситуация, соответствие
H_r	с, решение, понятие, сложный, который, вывод, фреймовый, на, задача, в, и, основа, для, знание

Документы, лучшие по критерию (3), и связи слов исходной фразы:

Оценка	«Наиболее сильные» связи для отбора фраз
Русанов В. В., Вестник РЭУ 2012 №1	
K_{AB}	язык – на, язык – сложный, на – основа, представление – с, язык – фреймовый, представление – в, представление – для, представление – понятие, язык – основа
$\text{sig}(A, B)$	язык – на, на – основа, язык – сложный, язык – фреймовый, основа – с
Лекторский В. А., ИИ ФМИ, 2014 г.	
K_{AB}	язык – задача, представление – способ, основа – модель, модель – для, модель – применять, сложный – понятие, в – знание, описание – применять, решение – различный, на – описание
Крымская Е. Ю., ИИ ФМИ, 2010 г.	
K_{AB}	решение – задача, решение – с, задача – в, на – решение, решение – для
Янковская А. Е., ТВИМ 2004 №1	
$\text{sig}(A, B)$	на – основа, решение – задача

Отбираемая фраза	Что представляет	Оценка	
<p>Специфика структурно-фреймовой организации состоит в том, чтобы во фрейме (а он представляет собой достаточно сложную концептуальную конструкцию, записанную средствами программной части вычислительной (информационной) системы) все понятия, относящиеся к охватываемой данным фреймом предметной области, имели внутреннюю интерпретацию, т.е. были наделены смыслом на соответствующем языке представления знаний</p>	<p>Русанов В. В., Вестник РЭУ 2012 №1</p>	<p>Связи понятий сложная концептуальная конструкция – сложное понятие – внутренняя интерпретация и структурное описание – язык представления знаний</p>	<p>K_{AB}, sig (A, B)</p>
<p>Фреймовые структуры реализуются на базе языков программирования высокого уровня, позволяющих человеку работать с информационной системой, используя лингвистические средства, близкие к языку межлического общения</p>	<p>Связь понятий структурное описание и язык программирования высокого уровня, перифраза на основе \iff на базе</p>	<p>K_{AB}</p>	
<p>Здесь:</p> <p>$\text{sig}(A, B)$ — фразы найдены только по максимуму числа «наиболее сильных» связей;</p> <p>K_{AB} — фразы найдены только по суммарной силе «наиболее сильных» связей;</p> <p>K_{AB} — как по максимуму числа «наиболее сильных» связей, так и по их суммарной силе.</p>			

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
<i>на основе TF-IDF слов исходной фразы</i>										<i>по суммарной «силе» для «наиболее сильных» по K_{AB}</i>								
N	5	8	14	9	1	1	29	5	10	1	12	15	1	1	2	2	1	11
N_1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	1
N_2	0	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	2
N_3	2	1	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	4
<i>по числу «наиболее сильных» связей по величине K_{AB}</i>										<i>по числу «наиболее сильных» связей по величине $\text{sig}(A, B)$</i>								
N	2	4	1	3	2	1	6	1	5	3	2	32	1	2	1	18	1	3
N_1	0	1	0	1	2	1	0	0	0	0	0	0	0	0	0	1	0	0
N_2	0	0	0	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0
N_3	1	2	0	0	0	0	2	0	1	1	2	1	0	0	0	2	0	1

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области.

Сравнение наиболее значимых связей и n -грамм для отбора фраз (оценка K_{AB} , максимум числа «наиболее сильных» связей)

№ исх. фразы	Слова, не вошедшие в наиболее значимые связи	n -граммы
	Философия и методология инженерии знаний	
3	<i>с, информация, который, на</i>	<i>точка, зрения</i>
4		<i>в, факт, данный</i>
6	<i>при</i>	
9	<i>который, вывод, структурный, соответствие, различный, способ, ситуация</i>	
	Математические методы обучения по прецедентам	
3	<i>заниженность</i>	
4	<i>заниженность, являться</i>	

Примечание

В данной иллюстрации сравнение ведётся по тем документам, которые вошли в число наиболее релевантных исходной фразе при использовании обоих вариантов (3) и (5) функции ранжирования.

Сравнение наиболее значимых связей и n -грамм для отбора фраз (оценка K_{AB} , максимум числа «наиболее сильных» связей)

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
<i>максимизацией числа «наиболее сильных» связей по K_{AB}</i>										<i>анализом n-грамм на найденных связях слов</i>								
Философия и методология инженерии знаний																		
N	2	4	1	3	2	1	6	1	5	2	1	2	4	6	1	6	2	1
N_1	0	1	0	1	2	1	0	0	0	0	0	0	1	1	0	1	0	0
N_2	0	0	0	2	2	1	0	0	0	0	0	0	2	4	0	0	0	0
N_3	1	2	0	0	0	0	2	0	1	0	1	2	2	5	1	2	0	1
Математические методы обучения по прецедентам																		
N	1	1	15	15	5	11	1	1	1	2	4	1	1	3	1	2	1	1
N_1	1	1	3	2	0	0	0	0	1	0	1	1	1	0	0	0	0	0
N_2	0	1	2	2	1	9	0	0	1	0	0	1	1	3	1	0	0	0
N_3	0	0	7	4	0	4	0	1	0	1	2	0	0	0	1	0	1	1

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области.

Альтернативное решение: поиск фраз на готовом синтаксически размеченном текстовом корпусе

Слова и их сочетания для отбора фраз из Национального корпуса русского языка:

№ Слова и сочетания слов

Философия и методология инженерии знаний

- 1 модель – представление – знание, механизм – логический – вывод
- 2 система – суждение, объективный – закономерность
- 3 процесс – логический – вывод
- 4 данный – предметный – область
- 5 эвристика, данный – предметный – область
- 6 метазнание, свойство – знание, структура – знание, способ – получение – знание, способ – использование – знание, задача – искусственный – интеллект
- 7 представление – знание, управление – вывод, механизм – логический – вывод, управление – знание
- 8 теоретический – обоснование – модель, логический – модель, система – вывод, система – определение, точный – вывод
- 9 язык – представление – знание, фреймовый – модель, способ – вывод

№ Слова и сочетания слов

Математические методы обучения по прецедентам

- 1 переобучение, эмпирический – риск
- 2 эмпирический – риск
- 3 эмпирический – риск
- 4 эмпирический – риск
- 5 ошибка – средний
- 6 частота – ошибка, контрольный – выборка
- 7 оценка – частота, контрольный – выборка
- 8 ошибка – распознавание, правило – принятие – решение
- 9 базовый – классификатор

№	1	2	3	4	5	6	7	8	9
<i>Философия и методология инженерии знаний</i>									
N	13	67	2	15	29	30	79	224	20
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	2	5	0	1	1	2	3	2	2
<i>Математические методы обучения по прецедентам</i>									
N	56	1	1	1	24	17	21	5	2
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	0	0	0	0	0	0	0	1	0

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области.

- 1 Основной *результат* настоящей работы — *метод* формирования тематического корпуса текстов, релевантных по описываемым фрагментам знаний исходной фразе, с выделением составляющих её образа в виде слов и их сочетаний.
- 2 По сравнению с поиском совокупностей указанных составляющих на синтаксически размеченном текстовом корпусе, предложенный метод *позволяет* в среднем в **15** раз сократить выход фраз, не релевантных исходной ни по описываемому фрагменту знания, ни по языковым формам его выражения.
- 3 Для предметных областей, в текстах которых доля общей лексики сравнима с долей терминов, *выделение n-грамм* на найденных связях слов при этом *повышает* выход фраз, представляющих связи понятий.

- 1 Выделение составляющих образа исходной фразы в текстах анализом встречаемости её слов, отвечающих кластеру наибольших значений TF-IDF, совместно с n -граммами на найденных связях слов.
- 2 Интерпретация меры TF-IDF для указанных n -грамм.
- 3 Оценка точности выделения границ предложений для разных вариантов обучения классификатора.