Мера TF-IDF и формирование единиц представления знаний для открытых тестов

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

Новгородский государственный университет
имени Ярослава Мудрого

Всероссийская конференция с международным участием
«Математические методы распознавания образов» (ММРО-17),

19–25 сентября 2015 г.

г. Светлогорск, Калининградская обл.

Цель исследований, исследуемая проблема

Единица знаний, оцениваемая открытым тестом

Определяется семантически эквивалентными (СЭ) фразами предметно-ограниченного естественного языка (ЕЯ).

Проблема

Как найти вариант наиболее рациональной передачи смысла ?

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.

Основная цель исследований

Разработка и теоретическое обоснование методов и алгоритмов поиска оптимального варианта передачи смысла между экспертами и обучаемыми в системе контроля знаний на основе открытых тестов.

Формирование единиц экспертных знаний

Наиболее актуальные задачи

1. Тематическая рубрикация текстовых документов.
2. Представление предметных областей в виде специализированных тезаурусов и онтологий.

Задачи эксперта, требующие автоматизации

1. Поиск эквивалентных по смыслу форм выражения отдельного фрагмента фактического знания в заданном естественном языке. При этом фрагмент фактического знания эксперта отвечает некоторому факту предметной области.
2. Сопоставление фрагментов собственных знаний эксперта с наиболее близкими фрагментами знаний других экспертов.

Требования к решению

1. Выделение из текста понятий и отношений между ними.
2. Выявление в текстовом корпусе контекстов использования общей лексики, обеспечивающей синонимичные перифразы.

Мера TF-IDF

Согласно классическому определению, данная мера есть произведение TF-меры (отношения числа вхождений слова к общему числу слов документа) и инверсии частоты встречаемости в документах корпуса (IDF).

TF-мера оценивает важность слова t_i в пределах отдельного документа d и определяется как

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

где n_i — число вхождений слова t_i в документ d,
а в знаменателе — общее число слов в документе.

IDF (inverse document frequency) — обратная частота документа, является единственной для каждого уникального слова в корпусе D и равна

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

где в числителе представлено общее число документов корпуса,
а $|D_i \subset D|$ есть число документов, где t_i встретилось хотя бы раз.

Базовые предположения

1. Наиболее уникальные слова в документе (с наибольшими значениями TF*IDF) будут относиться к терминам его предметной области.
2. Наличие синонимов у слова-термина ведёт к снижению значения TF относительно документа в случае, когда синонимы встречаются в этом же документе.
3. Термины, преобладающие в корпусе, а также слова общей лексики будут иметь значения IDF, близкие к нулю.
4. Слова-синонимы, уникальные для отдельных документов корпуса, будут иметь более высокие значения IDF.

   Пример — слова общей лексики, задающие конверсивные замены: *«приводить $\Leftrightarrow$ являться следствием»*.

Пусть

X — упорядоченная по убыванию последовательность $\mathrm{tf}\,(t,d)\cdot\mathrm{idf}\,(t,D)$ для всех слов t исходной фразы относительно документа $d\in D$.

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

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

Оценку качества кластеризации слов исходной фразы определим как

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

где $\mathrm{diam}\,(H_i)$ — диаметр кластера H_i;

$\min\,(F)$ и $\max\,(F)$ — минимальное и максимальное, соответственно, значения диаметра кластера из представленных в списке F;

$\mathrm{len}\,(F)$ — длина списка F.

Пусть

D разбивается на кластеры по аналогии с X, но по значению функции (3);

$D' \subset D$ — кластер наибольших значений оценки (3).

Требуется отобрать фразы из документов $d \in D'$ по максимуму слов, представленных в кластерах $\{H_1, H_{r/2}, H_r\} := Cl$:

H_1 — *слова-термины* исходной фразы, *наиболее уникальные* для d;

$H_{r/2}$ — *общая лексика*, обеспечивающая *синонимические перифразы*, и *термины-синонимы*;

H_r — *слова-термины*, *преобладающие* в корпусе.

Оценка представленности слов фразы $s \in d$, $d \in D'$, в кластерах из Cl

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

где первое слагаемое в знаменателе — среднеквадратическое отклонение числа слов фразы документа d, представленных в кластере из списка Cl.

Исходные данные эксперимента: состав текстового корпуса

- 3 статьи в журнале «Таврический вестник информатики и математики (ТВИМ)»;
- 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 корпуса варьировалось от 218 до 6298.

Исходные данные эксперимента: тематика отбираемых работ

- 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).

Исходные фразы для формирования единиц экспертных знаний

№	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

Пример: поиск фраз, максимально близких исходной фразе №9

Кластеры для отбора фраз:

Воронцов К. В., ТВИМ 2004 №1, слова, представленные в кластерах	
H_1	*алгоритм, обобщать, способность*
$H_{r/2}$	*классификатор, увеличение, число*
H_r	*вести*
Воронцов К. В., ММРО-15, слова, представленные в кластерах	
H_1	*алгоритм*
$H_{r/2}$	*рост, композиция*
H_r	*неограниченный, базовый, увеличение*

Результаты (содержат слова *обобщать, способность, алгоритм*):

Отбираемая фраза	Что представляет
Обобщающая способность определяется как вероятность ошибки найденного алгоритма, либо как частота его ошибок на неизвестной контрольной выборке, также случайной, независимой и одинаково распределённой	**Связь определения** *обобщающей способности алгоритма* **с понятиями** *вероятность ошибки* **и** *частота ошибок на контрольной выборке*
Результатом обучения является не только сам алгоритм, но и достаточно точная оценка его обобщающей способности	*ведёт к* $\Longleftrightarrow$ *является результатом*

Пример поиска синонима для слова-термина в исходной фразе №4

Кластеры для отбора фраз:

Воронцов К. В., ТВИМ 2004 №1, слова, представленные в кластерах	
H_1	*риск, эмпирический*
$H_{r/2}$	*заниженность, являться, переподгонка*
H_r	*нежелательный*
Воронцов К. В., ММРО-15, слова, представленные в кластерах	
H_1	*риск*
$H_{r/2}$	*результат*
H_r	*нежелательный, заниженность, переподгонка*
Дюличева Ю. Ю., ТВИМ 2002 №1, слова, представленные в кластерах	
H_1	*переподгонка*
$H_{r/2}$	*являться*
H_r	*нежелательный, заниженность, риск*

Отобранная фраза (*эмпирический, риск, являться, заниженность*): *Причиной является всё то же переобучение, которое приводит к заниженности эмпирического риска.*

Синонимы-термины: *переподгонка* $\Longleftrightarrow$ *переобучение*

Вариант конверсивной замены: *результат* $\Longleftrightarrow$ *причина*

Значения TF, IDF и TF-IDF (для сравнения)

Кластеры для отбора фраз:

Воронцов К. В., ТВИМ 2004 №1, диапазоны значений 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	$0{,}0000 \ldots 0{,}0000$
Воронцов К. В., ММРО-15, диапазоны значений 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	$0{,}0000 \ldots 0{,}0000$
Дюличева Ю. Ю., ТВИМ 2002 №1, диапазоны значений 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	$0{,}0000 \ldots 0{,}0000$

Значения TF (Воронцов К.В., ТВИМ 2004 №1) и IDF слов исходной фразы №4:

слово	нежела-тельный	заниженность	переподгонка	являться	результат	эмпири-ческий	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	$0{,}0000$	$2{,}1839 \cdot 10^{-4}$	$1{,}4386 \cdot 10^{-4}$	$1{,}7347 \cdot 10^{-4}$	$4{,}2392 \cdot 10^{-4}$	$0{,}0020$	$0{,}0026$

Кластеры для отбора фраз:

Воронцов К. В., ТВИМ 2004 №1, диапазоны значений TF-IDF		
H_1	*оценка, ошибка*	0,0019 . . . 0,0029
$H_{r/2}$	*заниженность*	$2{,}1839 \cdot 10^{-4}$ . . . $2{,}1839 \cdot 10^{-4}$
H_r	*с, принятие*	0,0000 . . . 0,0000
Дюличева Ю. Ю., ТВИМ 2002 №1, диапазоны значений TF-IDF		
H_1	*ошибка*	0,0068 . . . 0,0068
$H_{r/2}$	*решение, распознавание, принятие*	$3{,}0603 \cdot 10^{-4}$ . . . $3{,}7303 \cdot 10^{-4}$
H_r	*заниженность, с, связанный*	0,0000 . . . 0,0000
Дюличева Ю. Ю., ТВИМ 2003 №2, диапазоны значений TF-IDF		
H_1	*решение, распознавание, принятие*	0,0017 . . . 0,0018
$H_{r/2}$	*правило*	$4{,}2541 \cdot 10^{-4}$ . . . $4{,}2541 \cdot 10^{-4}$
H_r	*заниженность, с*	0,0000 . . . 0,0000

Отобранная фраза:

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 [Дюличева Ю. Ю., ТВИМ 2002 №1].

Возможные альтернативные решения и их недостатки

1. Поиск слов, связанных по смыслу с заданными, на основе известных семантических отношений и форм их выражения в текстах.
   Система «Серелекс»:
   - по исходной фразе №8 найдена единственная связь *«решение — с»*;
   - по исходной фразе №9 связей не найдено.

   Задействованные коллекции документов:
   - заголовки статей Википедии ($2{,}026 \cdot 10^9$ словоформ, $3\,368\,147$ лемм);
   - текстовый корпус ukWaC ($0{,}889 \cdot 10^9$ словоформ, $5\,469\,313$ лемм).

   Недостаток:
   - не предусмотрена предметная классификация лексики, что затрудняет использование реализуемых системой лексико-синтаксических шаблонов для выделения требуемых фрагментов текстов тематического корпуса.
2. Тезаурус типа WordNet:
   - внутри каждой группы синонимов (синсета) степень синонимии слов зависит от их предметной ориентации.
3. Суммарное значение TF-IDF слов исходной фразы, встречающихся во фразе документа, как альтернатива оценке (4).

   Недостаток: малая (менее 2%) доля общей лексики, реализующей синонимичные перифразы исходной фразы, в составе отбираемых фраз.

Основные выводы

1. Основной *результат* настоящей работы — *метод* поиска в текстовом корпусе описаний близких фрагментов знаний и языковых форм их выражения.
2. Помимо подготовки открытых тестов, важнейшая *сфера приложения* данного метода — построение специализированных тезаурусов, идейно близких «Чёрному квадрату», развиваемому ВЦ РАН.
3. По сравнению с известными подходами, предложенный метод *позволяет* решить задачу выделения понятий предметной области и отношений между ними на основе меньших обучающих выборок и без ориентации на определённые типы связей слов исходных фраз.

Открытые проблемы

1. Выработка численной оценки, которая учитывала бы одновременно:
   - качество выделения тем — совокупностей специальных терминов предметной области, совместно встречающихся в документах;
   - характер распределения терминов в теме;
   - характер распределения тем в документе.
2. Предсказуемость появления слов во фразе документа и её связь с составом выделяемых кластеров по значению TF-IDF для слов исходной фразы.
3. Для многозначных слов — учёт потенциального синтаксического контекста.