Автор: admin
26.10.2010 00:03 - Обновлено 18.10.2011 00:22

Применение частотного анализа в соционике. Новая методика определения авторства текста

Постановка задачи

Пусть дан текст на русском языке. Попробуем решить следующие задачи:

1. Определение автора из множества известных, книги которых у нас уже проанализированы.

2. Определение основных факторов оказывающих наибольшее влияние на распределение частот слов в тексте.

3. Определение соционического типа автора текста с помощью частотного анализа.

Новая методика определения авторства по тексту на естественном языке

Первая попытка создания методики для определения автора текста была сделана еще в самом начале 20 века Морозовым. Позднее она была подвергнута критике специалистом по теории вероятностей и математической статистике Марковым
.

Уже в наше время была предложена интересная методика определения авторства текста с использованием буквенной и грамматической информации
, которая использует формальную математическую модель последовательности букв (и любых других элементов) текста как реализации цепи Маркова.

Известный пример с определением авторства романа "Тихий Дон", об ответе на которой спорят несколько десятилетий, показывает, что данная проблема до сих пор актуальна. В настоящий момент, в связи с бурным развитием вычислительной техники встает вопрос о попытках автоматизировать этот процесс. В частности, математиком Хетсо была предложена методика
на основе следующих параметров:

- Средняя длина слова в буквах, вычисляемая на основании выборок размером 500 текстовых слов.
- Общее распределение длины слова.
- Средняя длина предложения в словах, вычисляемая на основании выборок размером в 30 предложений.
- Общее распределение длины предложения.
- Лексический спектр текста на уровня словаря.
- Лексический спектр текста на уровне текста.
- Индекс разнообразия лексики.

С помощью нее он провел компьютерный анализ текстов Шолохова, подтвердивший его авторство.

Известно, что клуб любителей творчества Пушкина собирал информацию о частотном распределении слов великого поэта. На это занятие им понадобилось несколько лет кропотливого труда. К счастью, с приходом новейших технологий, туже самую операцию компьютер способен сделать за несколько минут с гораздо большей точностью.

Методика, которая описывается в этой статье, была случайно получена мной в качестве побочного эффекта при исследовании возможности определения соционического типа автора текста на естественном языке. До этого я не читал материалов по данной теме.

Слова русского языка имеют огромную разницу в распределении частот. Например, слово "время" встречается в 500 раз чаще чем "удивительный". В качестве эталона распределения частот слов русского языка был взят частотный словарь Шарова (общее количество различных слов более 60000), который составлен на основе анализа 40 миллионов слов и является более адекватным чем аналогичный известный частотный словарь Засориной

, который был составлен в 1977 году и использовал для анализа всего лишь 1 миллион слов.

В базу данных Oracle были закачаны результаты частотного анализа 104 книг 38 человек (количество книг для каждого писателя было от 1 до 14) общим размером более 30 Мегабайт чистого текста, в которых использовано почти 6 миллионов слов. Для анализа были написаны несколько программ на PL/SQL.

Алгоритм

- Составление частотного словаря для каждой книги.
- На основе нескольких книг создается частотный словарь писателя.
- С помощью частотного словаря Шарова происходит нормализация. То есть полученные значения частоты употребления слов делятся на средние в русском языке.

- Вводится понятие расстояния между словарями, как сумма квадратов разностей частот между отдельными анализируемыми словами.
- При этом если слово есть в одном словаре, но совсем отсутствует в другом, то оно не учитывается (для чего это сделано объясняется ниже)
- Учитываются первые 5000-10000 наиболее употребляемых слов русского языка.
- В качестве результата берется словарь с минимальным расстоянием.

Если взять больше 10000 слов, то редкие слова оказывают слишком большое влияние на результат, если меньше, то информации становится недостаточно. Учитывая такое количество слов текст должен быть достаточно большим, желательно от 30 Kb, причем чем больше, тем лучше. На текстах малой длины частоты неустойчивы и сильно зависят от предметной области. К аналогичным выводам пришел польский исследователь Е. Ворончак в работе, посвященной математико-статистическому анализу устойчивости различных показателей, используемых в настоящее время в исследованиях языка и стиля произведения: "границей объема текста (ниже которой результаты не достоверны, а выше – достоверны) является пять тысяч словоформ".

По данному методу для всех 104 книг автор был определен верно в 102 случаях с двумя ошибками определения для Александра Пушкина "Том 7. История Пугачева. Исторические статьи и материалы" и "Том 9. Письма". Также при правильном определении авторства для для нескольких небольших рассказов Николая Гоголя разница между следующим писателем была не очень большой. Для книг, не

участвовавших в формировании словаря данный метод не проверялся, можно понять, что он будет работать на следующем примере.

Были взяты 38 словарей писателей. 104 словаря для книг, в том числе пять из них Льва Толстого. В нижеприведенной таблице показано расстояние по словарям для книги Льва Толстого "Юность", если произведение не указано имеется в виду частотный словарь писателя.

	Автор	Произведение	Расстояние
1	***Лев Толстой***	Юность	0
2	***Лев Толстой***	Частотный словарь автора	169
3	***Лев Толстой***	Детство	289
4	***Лев Толстой***	Война и мир. Том 2	307
5	Джек Лондон	Частотный словарь автора	373
6	Герман Гессе	Эссе	385
7	Николай Гоголь	Частотный словарь автора	393
8	Герман Гессе	Частотный словарь автора	396
9	Федор Достоевский	Частотный словарь автора	403
10	Федор Достоевский	Записки из мертвого дома	404
11	Иван Тургенев	Новь	406
12	***Лев Толстой***	Хаджи-Мурат	415
13	Иван Тургенев	Казаки	421
14	***Лев Толстой***	Частотный словарь автора	422
...	...	...	...
142	Жан-Поль Сартр	Частотный словарь автора	868

Отсюда видно, что все пять книг Толстого плюс словарь писателя попали на первые 14

позиций, то есть книги Толстого находятся достаточно близко друг к другу, поэтому чем больше будет проанализировано данных для составления частотных словарей писателей, тем более надежным будет результат.

Жанр произведений

Но вернемся к Пушкину и одной из ошибок определения автора по тому 7 "История Пугачева. Исторические статьи и материалы":

	Автор	Произведение	Расстояние
1	***Александр Пушкин***	Том 7 История Пугачева. Исторические статьи	0
2	Станислав Лем	Звездные дневники Йонаша Тихого	341
3	Джек Лондон	Частотный словарь автора	363
4	Станислав Лем	Частотный словарь автора	364
5	Антон Чехов	Частотный словарь автора	375
6	Герман Гессе	Эссе	397
...	...	...	...
39	***Александр Пушкин***	Частотный словарь автора	545
...	...	...	...
49	***Александр Пушкин***	Том 2 Стихотворения 1825-1836	575
...	...	...	...
71	***Александр Пушкин***	Том 4 Евгений Онегин Драматические произве	650
...	...	...	...
73	***Александр Пушкин***	Том 1 Стихотворения 1814-1822	657
...	...	...	...
129	***Александр Пушкин***	Том 9 Письма	2436

...	...	...	...

и к правильному определению автора по тому 2 "Стихотворения 1823-1936":

№	Автор	Произведение	Расстояние
1	***Александр Пушкин***	Том 2 Стихотворения 1823-1836	0
2	***Александр Пушкин***	Том 1 Стихотворения 1813-1822	53
3	***Александр Пушкин***	Том 3 Поэмы, сказки	143
4	***Александр Пушкин***	Том 4 Евгений Онегин. Драматические произведения	144
5	***Александр Пушкин***	Рассказы	162
6	Антон Чехов		420
...	...	...	...
35	***Александр Пушкин***	Том 7 История Пугачева. Исторические статьи и материалы	575
...	...	...	...
130	***Александр Пушкин***	Том 9 Письма	2456
...	...	...	...

Вывод, который напрашивается из этих двух таблиц: есть три достаточно далеко расположенные друг от друга группы произведений Пушкина: поэзия (Тома 1-4), письма (Том 9) и проза (Том 7 "История Пугачева. Исторические статьи и материалы"). Таким образом наглядно показано, что кроме собственно авторства частота слов в тексте очень сильно зависит от жанра произведений.

Еще одно подтверждения этого было получено, когда были проанализированы два ЖЖ-дневника (авторы имеют психологические типы СЭЭ и ЛСЭ) и сообщения на форуме (автор СЛИ). Казалась бы большое расхождение должны были бы дать разница в стиле, возрасте, образе жизни, психотипе и словарном запасе. Один из словарей был составлен по дневнику журналистки Таты Олейник (Почти_новая_горжетка), у которой словарный запас оказался самым большим по первым 80 книгам. Тем не менее по данной методике расстояния между этими тремя словарями получились относительно небольшими, для одного из словарей два других

оказались ближайшими из 37 словарей. Таким образом язык on-line дневников и форумов, близкий к разговорному сильно отличается от литературного и научного, причем это отличие можно выявить с помощью данной методики или ее модификации. А значит ее можно применять для больших социологических и психолингвистических исследований русского языка на больших корпусах текстов.

Предметная область

Полную версию таблицы расстояний между 28 словарями писателей, психологов и соционików в базе данных можно посмотреть здесь . В качестве психотипа стоит моя версия

Далее проанализируем полученную таблицу. Отсортируем список по возрастанию расстояния от словаря Агаты Кристи:

	Автор	ТИМ	Расстояние	Профессия
1	Агата Кристи	СЛИ	0	
2	Иван Тургенев	ЭСИ	242	
3	Станислав Лем	ИЛИ	256	
4	Антон Чехов	ЛИИ	285	
5	Федор Достоевский	ЭИИ	286	
6	Джек Лондон	ЛИЭ	322	
7	Теодор Драйзер	ЭСИ	350	
8	Виктор Гюго	ЭСЭ	351	

9	Николай Гоголь	ЭИЭ	355
10	Лев Толстой	СЭЭ	356
11	Жюль Верн	ЭСЭ	382
12	Пауло Коэльо	ЭИИ	386
13	Гарсия Маркес	ЭСЭ	401
14	Ги де Мопассан	СЛИ	420
15	Герман Гессе	ИЛИ	428
16	Зигмунд Фрейд	ИЛЭ	552
17	Карл Юнг	ИЛИ	574
18	Эрик Берн	ИЛЭ	713
19	Александр Пушкин	ИЭЭ	725
20	Иван Крылов	ИЛИ	742
21	Билл Гейтс	ЛИЭ	790
22	Абрахам Маслоу	ИЭЭ	850
23	Эрих Фромм	ЭИИ	932
24	Екатерина Филатова	ЭИИ	952

25	Жан-Поль Сартр	СЛИ	952
26	Виктор Гуленко	ЛИИ	1 040
27	Аушра Аугустинавичюте	ИЛЭ	2 604
28	Александр Лоуэн	СЭИ	3 381

Все писатели сверху! Отсюда следует, что профессия, а значит и предметная область существенно влияют на частотный анализ.

Аналогично отсортируем список по возрастанию расстояния от словаря Абрахама Маслоу:

	Автор	**ТИМ**	**Расстояние**	**Профессия**
1	Абрахам Маслоу	ИЭЭ	0	
2	Карл Юнг	ИЛИ	294	
3	Эрих Фромм	ЭИИ	295	
4	Зигмунд Фрейд	ИЛЭ	369	
5	Эрик Берн	ИЛЭ	479	
6	Пауло Коэльо	ЭИИ	653	
7	Станислав Лем	ИЛИ	654	

8	Антон Чехов	ЛИИ	691
9	Билл Гейтс	ЛИЭ	695
10	Герман Гессе	ИЛИ	707
11	Екатерина Филатова	ЭИИ	708
12	Лев Толстой	СЭЭ	719
13	Виктор Гюго	ЭСЭ	727
14	Джек Лондон	ЛИЭ	728
15	Жюль Верн	ЭСЭ	751
16	Иван Тургенев	ЭСИ	781
17	Теодор Драйзер	ЭСИ	793
18	Федор Достоевский	ЭИИ	830
19	Агата Кристи	СЛИ	850
20	Николай Гоголь	ЭИЭ	851
21	Ги де Мопассан	СЛИ	866
22	Гарсия Маркес	ЭСЭ	898

23	Виктор Гуленко	ЛИИ	914
24	Александр Пушкин	ИЭЭ	1 075
25	Иван Крылов	ИЛИ	1 238
26	Жан-Поль Сартр	СЛИ	1 581
27	Аушра Аугустинавичюте	ИЛЭ	1 707
28	Александр Лоуэн	СЭИ	2 968

Заметим, что наиболее близкими оказались словари практически всех психологов за исключением Александра Лоуэна.

Теперь проверим остается ли действовать это правило для соционików:

	Автор	**ТИМ**	**Расстояние**	**Профессия**
1	Аушра Аугустинавичюте	ИЛЭ	0	
2	Екатерина Филатова	ЭИИ	1 169	
3	Виктор Гуленко	ЛИИ	1 294	
4	Зигмунд Фрейд	ИЛЭ	1 578	
5	Эрих Фромм	ЭИИ	1 670	
6	Карл Юнг	ИЛИ	1 703	

7	Абрахам Маслоу	ИЭЭ	1 707
8	Эрик Берн	ИЛЭ	1 709
9	Пауло Коэльо	ЭИИ	2 143
10	Лев Толстой	СЭЭ	2 154
11	Антон Чехов	ЛИИ	2 272
12	Билл Гейтс	ЛИЭ	2 284
13	Герман Гессе	ИЛИ	2 307
14	Джек Лондон	ЛИЭ	2 346
15	Теодор Драйзер	ЭСИ	2 369
16	Иван Тургенев	ЭСИ	2 378
17	Станислав Лем	ИЛИ	2 401
18	Виктор Гюго	ЭСЭ	2 433
19	Ги де Мопассан	СЛИ	2 470
20	Николай Гоголь	ЭИЭ	2 505
21	Жюль Верн	ЭСЭ	2 510

22	Федор Достоевский	ЭИИ	2 529
23	Гарсия Маркес	ИЛИ	2 544
24	Александр Пушкин	ИЭЭ	2 591
25	Агата Кристи	СЛИ	2 604
26	Иван Крылов	ИЛИ	2 968
27	Жан-Поль Сартр	СЛИ	3 194
28	Александр Лоуэн	СЭИ	3 861

Соционики сверху, далее подряд все психологи, опять же за исключением Лоуэна. Таким образом наша гипотеза о существенном влиянии предметной области на распределение частот слов в тексте еще раз подтвердилась.

Хотя это правило выполняется не всегда, например, для Гуленко, словарь Филатовой опять же оказывается сверху, но словарь Аушры находится в конце списка.

Так почему же словарь Лоуэна расположен настолько далеко от остальных психологов?

Объем анализируемого текста

Для ответа на этот вопрос построим таблицу для самого Лоуэна:

Автор	**ТИМ**	**Расстояние**	**Профессия**

1	Александр Лоуэн	СЭИ	0
2	Зигмунд Фрейд	ИЛЭ	2 698
3	Карл Юнг	ИЛИ	2 778
4	Эрих Фромм	ЭИИ	2 928
5	Абрахам Маслоу	ИЭЭ	2 968
6	Гарсия Маркес	ЭСЭ	2 993
7	Пауло Коэльо	ЭИИ	2 998
8	Джек Лондон	ЛИЭ	3 009
9	Лев Толстой	СЭЭ	3 017
10	Станислав Лем	ИЛИ	3 023
11	Герман Гессе	ИЛИ	3 069
12	Николай Гоголь	ЭИЭ	3 087

13	Екатерина Филатова	ЭИИ	З 109
14	Ги де Мопассан	СЛИ	З 111
15	Эрик Берн	ИЛЭ	З 155
16	Виктор Гюго	ЭСЭ	З 162
17	Иван Тургенев	ЭСИ	З 171
18	Антон Чехов	ЛИИ	З 181
19	Виктор Гуленко	ЛИИ	З 248
20	Теодор Драйзер	ЭСИ	З 266
21	Жюль Верн	ЭСЭ	З 316
22	Федор Достоевский	ЭИИ	З 316
23	Агата Кристи	СЛИ	З 381

24	Билл Гейтс	ЛИЭ	3 390
25	Александр Пушкин	ИЭЭ	3 536
26	Иван Крылов	ИЛИ	3 738
27	Аушра Аугустинавичюте	ИЛЭ	3 861
28	Жан-Поль Сартр	СЛИ	4 037

Получается, что для самого Лоуэна словари психологов оказываются ближе, чем все остальные. Так в чем же дело? Мне кажется в данном случае есть две основные причины:

- Для анализа была взята книга "Вы и ваше тело" по телесно-ориентированной терапии, которая отличается от остальных психологических направлений своеобразием лексики достаточно сильно
- В последнем столбце указано количество анализируемых слов. Для Лоуэна, Аушры, Сартра, Крылова оно относительно мало. Хотя эта проблема частично решается пятым пунктом алгоритма, малое количество анализируемых слов делает словарь неустойчивым.

Еще одним фактор, который может воздействовать на частоту вхождения слов, время написания книги, подробно не анализировался.

А теперь переходим к самому интересному для соционików.

Попытка определения психотипа

Отсортируем список по возрастанию расстояния от словаря Германа Гессе:

	Автор	ТИМ	Расстояние	Профессия
1	Герман Гессе	***ИЛИ***	0	
2	Лев Толстой	СЭЭ	213	
3	Джек Лондон	ЛИЭ	234	
4	Иван Тургенев	ЭСИ	240	
5	Пауло Коэльо	ЭИИ	256	
6	Ги де Мопассан	СЛИ	261	
7	Станислав Лем	***ИЛИ***	265	
8	Виктор Гюго	ЭСЭ	283	
9	Антон Чехов	ЛИИ	302	
10	Теодор Драйзер	ЭСИ	305	
11	Федор Достоевский	ЭИИ	312	
12	Гарсия Маркес	ЭСЭ	336	
13	Николай Гоголь	ЭИЭ	348	

14	Жюль Верн	ЭСЭ	407
15	Агата Кристи	СЛИ	428
16	Александр Пушкин	ИЭЭ	450
17	Карл Юнг	***ИЛИ***	485
18	Зигмунд Фрейд	ИЛЭ	495
19	Эрик Берн	ИЛЭ	654
20	Эрих Фромм	ЭИИ	705
21	Абрахам Маслоу	ИЭЭ	707
22	Билл Гейтс	ЛИЭ	712
23	Иван Крылов	***ИЛИ***	723
24	Екатерина Филатова	ЭИИ	827
25	Жан-Поль Сартр	СЛИ	961
26	Виктор Гуленко	ЛИИ	1 021
27	Аушра Аугустинавичюте	ИЛЭ	2 307
28	Александр Лоуэн	СЭИ	3 069

Опять все писатели сверху, то есть влияние предметной области определяется достаточно точно.

Но если посмотреть на психотип ИЛИ, то он получается разбросанным по всей таблице. Аналогичные результаты видны и в остальных таблицах, приводимых выше. Я собрал версии о психотипах известных людей большинства известных соционików и построил эталонный список на основе их анализа. В нем, как представители типа интуитивно-логических интровертов (ИЛИ), оказались Герман Гессе, Станислав Лем и Гарсия Маркес (мое мнение - ЭСЭ), по поводу психотипа Карла Густава Юнга мнения соционików разделились между ИЛИ и ЛИИ. В любом случае при замене версий типов Маркеса и Юнга на более распространенные общая картина не меняется, то есть данная методика, использующая частотный анализ первых 5000-10000 наиболее употребляемых слов не может дать определение психотипа (точнее совпадения с наиболее вероятными версиями).

Итак, в целом частотные словари оказались достаточно устойчивыми на больших массивах информации. То есть **каждый их нас обладает своим неповторимым частотным словарем и аналогично почерку его можно идентифицировать с достаточно большой вероятностью** .

Это дает надежду возможности определения психотипа на основе его анализа.

О семантическом подходе в соционике писали Вайсбанд, Филимонов, Ритчик, Шепетько, Аушра. Прокофьева, Ермак, питерская группа соционików, а также авторы этого сайта (я и Елена Заманская) составили свои семантические словари по каждой из функций.

Первые же идеи, которая приходят в голову для модификации данной методики: отфильтровать слова русского языка и рассматривать только те, которые относят к наполнению соционических функций, а также попробовать использовать при типировании основные дихотомии Юнга и признаки Рейнина. Данное исследование было проведено. О его результатах читайте в следующей статье.

Заключение

Итак в данной статье:

1. Показано, что частотный словарь человека достаточно устойчив на больших объемах текста и неустойчив на малых.
2. Была предложена новая методика определения автора текста на естественном языке. Основными плюсами данной методики являются ее надежность, простота и возможность автоматического использования. К минусам можно отнести то, что анализируемый текст должен быть достаточно большим для надежного определения авторства. Возможно в дальнейшем удастся синтезировать ее с методикой Хетсо.

3. Показано, что на частоту употребления слов существенно влияет не только автор, но также предметная область, жанр и размер анализируемого текста.
4. Переводчик оказывает гораздо меньшее влияние на распределение частот.
5. С помощью частотного анализа по наиболее употребительным словам не удается определять соционический тип без дополнительной фильтрации по семантическим словарям.

Полученные результаты показывают, что психотип влияет на частоту употребления слов в русском языке в целом меньше, чем предметная область, жанр и размер анализируемого текста.

Сам анализируемый текст должен быть достаточно большим, иначе выводы будут ненадежными!

Данная статья не претендует на полноценное исследование, так как, например, для оценки надежности новой методики определения авторства нужно обработать гораздо большое число книг и источников информации. Возможно я это сделаю в будущем.

Автор - Хрулёв Олег

Список литературы

1	Н.А.Морозов	Морозов орудие объективного исследования древних документов
2	А.А. Марков	Об одном применении статистического метода
3	Г. Хетсо	Методика, основанная на методах математиче
4	Л.И. Бородкин	Математические методы и компьютер в задача
5	О.В. Кукушкина, А.А. Поликарпов, Д.В. Хмелёв	Определение авторства текста с использован
6	С.А. Шаров	Частотный словарь Шарова
7	Л.Н. Засорина	Частотный словарь Засориной
8	Р.М. Фрумкина	Психолингвистик