Информационные технологии

УДК 004.912

В.А. ЯЦКО
(Абакан)

ЭФФЕКТИВНОСТЬ ПРИМЕНЕНИЯ КОСИНУСНОЙ МЕТРИКИ ДЛЯ ОПРЕДЕЛЕНИЯ СМЫСЛОВОЙ БЛИЗОСТИ ДОКУМЕНТОВ*

Оценивается эффективность применения косинусной метрики определения смысловой близости документов для решения задачи авторской атрибуции текстовых документов. Исходными статистическими данными послужило распределение стоп-слов в трёх произведениях художественной литературы, два из которых были написаны одним автором. Показано, что более адекватный результат получается при применении метрики к отклонениям частотностей стоп-слов от распределения Ципфа при условии предварительного выравнивания входных текстов.

Ключевые слова: *смысловая близость текстов, косинусная мера, распределение Ципфа, стоп-слова, классификация документов.*

VYACHESLAV YATSKO
(Abakan)

EFFICIENCY OF THE USE OF COSINE MEASURE TO DETERMINE THE DEGREE OF DOCUMENT SIMILARITY

The article deals with the assessment of the efficiency of using the cosine metrics to determine documents similarity for solving the task of the author's attribution of text documents. The source statistic data were the distribution of stop words in three fiction works, the two of which were written by one author. There is demonstrated that a more adequate result is obtained while using this metrics applied to the deviations of stop words frequencies from Zipf distribution on condition that source texts are pre-aligned.

Key words: *text documents similarity, cosine measure, Zipf distribution, stop words, document classification.*

Современный этап развития общества характеризуется глубоким проникновением во все сферы его жизни лингвистических технологий. Выполняя поиск с помощью систем "Google" или «Яндекс», отдавая голосовые команды электронным устройствам, пользователи не подозревают, что функционирование этих систем является результатом большой работы по решению проблем автоматической обработки текстов на естественном языке.

Одной из таких проблем является устранение зависимости вычисления смысловой близости текстовых документов от размеров текстов [6]. Информационный поиск, в сущности, основывается на анализе распределения терминов запроса, сформулированного пользователем, в документах, которые находятся в базе данных информационно-поисковой системы. В результате такого анализа вычисляется степень смысловой близости между запросом и документом и находится уровень релевантности документа. Очевидно, однако, что распределение терминов запроса в том или ином документе напрямую зависит от размеров этого документа. В документах больших по размеру частотность терминов запроса будет выше, однако это не означает большую релевантность данного документа. В этой связи и возникает проблема нормализации размеров текстов документов. Эта проблема важна не только для информационного поиска, но и для других направлений, связанных с автоматической классификацией текстов, таких как распознавание плагиата, авторская атрибуция, категоризация, определение жанра. Решение данной проблемы является актуальной задачей всей предметной области, связанной с автоматической обработкой текстовых документов.

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-07-00124.

Цель настоящей статьи – оценить эффективность метода косинусного измерения смысловой близости, который был разработан в теории информационного поиска и в настоящее время широко применяется и в других областях. Нами будет проведено сопоставление эффективности авторской атрибуции на основе косинусной меры и разработанным нами ранее методом атрибуции на основе отклонений от распределения Ципфа.

Метод косинусного измерения смысловой близости основывается на представлении частотностей терминов документов в виде векторов и измерении косинуса угла между ними [3]. Чем меньше угол, тем больше смысловая близость между запросом и документом. Смысловая близость (Sim) между двумя текстами *t1*, *t2* на основе косинусной меры вычисляется по формуле:

$$Sim(t1,t2) = \frac{\sum f_{i\epsilon t1} \times f_{i\epsilon t2}}{\sqrt{\sum f_{i\epsilon t1}^2} \times \sqrt{\sum f_{i\epsilon t2}^2}} \text{ (1)},$$

где *f* – частотность термина *i*-го в соответствующем тексте. Величина Sim изменяется в пределах от 0 до 1 (100%). Стопроцентная близость указывает на полное совпадение двух текстов. Считается, что данная формула позволяет сглаживать разницы в размерах текстов.

В данной статье будет также использована модифицированная формула:

$$Sim(j,R) = \frac{\sum dZ_{ij} \times dZ_{iR}}{\sqrt{\sum dZ_{ij}^2} \times \sqrt{\sum dZ_{iR}^2}}, \text{ (2)},$$

где *j* – входной текст, *R* – эталонный текст (наиболее типичный для данного автора), *dZ* – отклонение от распределения Ципфа.

Отклонение находится по формуле:

$$\mathrm{dZ}(ij) = |Z(ij) - \mathrm{f(ij)}| \text{ (3)},$$

где *Z* – распределение Ципфа, *f* – частотность термина. Распределение Ципфа находится по формуле:

$$Z_{ij} = F(1_j)/R_{ij} \text{ (4)},$$

где $f(1_j)$ – частотность первого по рангу термина в *j*-ом тексте, а *R* – номер ранга термина [1].

Под терминами нами понимаются стоп-слова – служебные слова, которые встречаются часто в текстах различных жанров и не выражают значения вне контекста. К таким словам относятся предлоги, артикли, союзы, местоимения, частицы. Стоп-слова удобно использовать в качестве основы для классификации текстов, т. к. их можно найти в любом тексте, хотя частотность их использования различается у разных авторов и может служить классификационным параметром. Использование стоп-слов также позволяет существенно сократить количество обрабатываемых единиц текста и повысить быстродействие системы-классификатора.

В качестве материала исследования нами были взяты два текста, написанные Дж. Голсуорси: *“The Man of Property”, “Indian Summer of a Forsyte”,* а также роман Ч. Диккенса *“Oliver Twist”*. Для обозначения текстовых файлов были приняты следующие соответствующие условные обозначения: *G1, G2, D*. Первые два произведения написаны Дж. Голсуорси и относятся к одному циклу – “The Forsyte Saga”. Отбирая эти тексты, мы исходили из предположения о том, что смысловая близость между текстами Дж. Голсуорси и Ч. Дикенса должна быть существенно меньше, чем между текстами Дж. Голсуорси, что обуславливается не только принадлежностью двух первых текстов одному автору, но также и тем, что и они относятся к разным временным периодам. Работа Ч. Дикенса была создана в середине XIX в., а произведения Дж. Голсуорси – в начале XX в. В соответствии с принятой методикой, сначала сопоставлялось распределение терминов (в нашем случае – стоп-слов) в G2 и G1, затем – в D

и G1. Таким образом, файл G1 был принят как эталонный (это произведение считается наиболее типичным для Дж. Голсуорси), а два остальных файла – как входные тексты.

Выбор английских текстов был обусловлен тем, что стоп-слова в английском языке характеризуются отсутствием флексий, что позволяет не применять дополнительные алгоритмы предварительной обработки текстов (стемминг, лемматизацию). Были выбраны произведения художественной литературы по следующим критериям: 1) они характеризуются большим разнообразием лексики (в том числе и стоп-слов), своеобразием стилевых особенностей, что должно отражаться на распределении стоп-слов; 2) для произведений XIX – начала XX в. характерен большой объём, что позволяет получать достоверные результаты; 3) были отобраны тексты с сайта проекта Гутенберг (Project Gutenberg) [2], на котором размещаются произведения с истекшим сроком действия авторских прав. Эти произведения редактируются и вычитываются. Нами было проведено дополнительное редактирование, в результате которого удалена информация о самом проекте Гутенберг, которая занимает достаточно много места в конце и начале каждого произведения, размещаемого на сайте. Исходные статистические данные, полученные с помощью конкорданса AntConc, представлены в табл. 1.

Таблица 1

Статистические данные исходных текстов

Текст	Автор	Количество токенов	Количество уникальных слов	Количество стоп-слов
G1	Galsworthy	113149	9083	387
G2	Galsworthy	111891	8844	379
D	Dickens	163849	10451	390

Стоп-слова в текстах были найдены на основе списка К. Фокса, который считается эталонным для английского языка [5]. Особенность формул (1) и (2) состоит в том, что должны учитываться частотности одного и того же термина в двух текстах. В этой связи нами было проведено пересечение списков стоп-слов в G1 и G2, а также G1 и D, и были найдены стоп-слова, встречающиеся в обоих текстах. В G1 и G2 оказалось 366 общих стоп-слов; в G1 и D – 191. В табл. 2 указаны совпавшие стоп-слова и их частотности.

Таблица 2

Распределение стоп-слов (выборка)

Тексты	Стоп-слова (первые три)	Частотность в G1	Частотность в G2 и D
G1∩G2	the	5644	4696
	of	3461	2722
	and	2997	3354
G1∩D	the	5644	9597
	and	2997	5395
	to	2874	3944

Результаты вычислений по формуле (1): Sim (G1, G2)=0,9892 = **98,92%;** Sim (G1, D) = 0,96709=96,71%. Соответственно, разница составляет **2,21%**. Таким образом, смысловая близость между текстами, написанными разными авторами (Дж. Голсуорси и Ч. Диккенсом), на указанную величину меньше, чем между текстами одного автора (Голсуорси).

Результаты вычислений по формуле (2): Sim (G1, G2)=0,966801526=**96,68%**; Sim (G1, D)=0,90536**=90,54%**. Соответственно, разница составляет **6,14%.** Таким образом, смысловая близость

между текстами, написанными разными авторами, на указанную величину меньше, чем, между текстами одного автора. Напомним, что формула (2) основывалась на предложенной нами методике вычисления отклонений от распределения Ципфа. Она дала более адекватный результат, чем традиционная формула, т. к. по последней разница в смысловой близости между текстами не выходит за рамки статистической погрешности.

Одним из достоинств традиционной формулы считается то, что с её помощью сглаживаются разницы в размерах текстов. Мы решили протестировать это утверждение, выровняв входные тексты по нижнему пределу. В соответствии с этим методом, который в зарубежной литературе называют undersampling [4], входные тексты выравниваются по размеру самого маленького из них путём удаления части текстов, начиная с конца. В связи с тем, что в нашем случае входных текстов всего два, то выравнивался файл D по размеру (в токенах) меньшего файла G2 (см. табл. 1). По формуле (2) Sim (G1, D)=0,894951= **89,5%,** соответственно разница составляет **7,19%,** что на 1,5% лучше, чем на тексте, который не выравнивался.

Нами было проведено три оригинальных теста косинусной меры вычисления смысловой близости текстовых документов, основанных на анализе распределения стоп-слов в двух текстах, написанных одним автором и в третьем, написанным другим автором. В первом тесте анализировались сырые частотности стоп-слов в полных входных текстах; во втором – анализировались коэффициенты отклонения от распределения Ципфа в полных текстах; в третьем также анализировались данные коэффициенты, но в выровненных текстах. В результате проведённого анализа можно сделать следующие выводы. Косинусная мера позволяет находить наиболее близкие тексты, однако недостаточно адекватно определяется разрыв между текстами, относящимися к разным классам (в нашем случае – написанными разными авторами). Это затрудняет её использование с целью автоматической классификации текстов. Хотя общепринятым является представление о способности с помощью данной метрики сглаживать разницы в размерах текстов, на самом деле требуется дополнительное выравнивание текстов. Подтверждается мнение некоторых авторов о необходимости применения некоторых поправочных коэффициентов. Более соответствующим целям автоматической классификации текстов является предложенный нами метод анализа отклонений от распределения Ципфа, который позволяет находить разрыв между текстами, относящимися к разным классам, в 60 и более процентов [1]. Наш метод предполагает использование более простых вычислений по формулам (3), (4). Основным классификационным параметром выступает среднее квадратичное отклонение, которое находится для числового ряда, включающего значения отклонений, а смысловая близость определяется как разница по модулю этими параметрами во входных и эталонном текстах. Наш метод требует предварительного выравнивания входных текстов, однако, как было показано, этого требует и косинусная метрика.

Сказанное не означает однозначно негативной оценки применения рассматриваемой метрики. Она была разработана с целью оптимизации информационного поиска и не предназначена специально для решения классификационных проблем. Её использование в целях классификации документов требует проведения дополнительных исследований, что является задачей последующей работы.

Литература

1. Яцко В.А. Метод автоматической классификации текстов, основанный на законе Ципфа // Научно-техническая информация. Сер. 2.: Информационные процессы и явления. 2015. № 5. С. 19–24.

2. Free eBooks – Project Gutenberg. [Электронный ресурс]. URL: https://www.gutenberg.org/ (дата обращения: 05.07.2020).

3. Madylova A., Oguducu S.G. A taxonomy based semantic similarity of documents using the cosine measure // 24th International Symposium on Computer and Information Sciences. Guzelyurt, 2009. P. 129–134. DOI: 10.1109/ISCIS.2009.5291865.

4. Polydouri A., Vathi E., Siolas G. et al. An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection // Evolving systems. 2018. DOI: 10.1007/s12530-018-9232-1. [Электронный ресурс]. URL: https://www.researchgate.net/profile/Andrianna_Polydouri/publication/326383978_An_efficient_classification_approach_in_imbalanced_datasets_for_intrinsic_plagiarism_detection.pdf (дата обращения: 05.07.2020).

5. Sarica S., Luo J. Stopwords in technical language processing. [Электронный ресурс]. URL: https://arxiv.org/abs/2006.02633 (дата обращения: 05.07.2020).

6. Singhal A., Salton G., Mitra M., Buckley C. Document length normalization // Information processing & management. 1996. Vol. 32. Issue 5. P. 619–633.