МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ФОРМИРОВАНИЯ КРИТЕРИЯ ТЕСНОТЫ СВЯЗИ ДВУХ ТЕХНИЧЕСКИХ ТЕКСТОВ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Кекиш Н. И.

Дайняк И. В. – канд. техн. наук

Предложена математическая модель для определения тесноты связи двух технических текстов на основе статистической корреляционной оценки по выборке слов. Описана методика расчета безразмерного нормированного коэффициента тесноты связи, приведены его градации по уровням тесноты связи.

При оформлении результатов научной работы вначале автор готовит так называемые первичные документы, к которым относятся статьи, а также научные и научно-технические тексты в виде отчетов, диссертаций, монографий. Затем на основе первичного документа готовится вторичный документ в виде текста аннотации или реферата. Вторичный документ – это документ быстрого доступа, который должен формировать максимально полное представление о первичном тексте. При этом особенно важна взаимная обусловленность первичного и вторичного документов, характеризующаяся степенью взаимной адекватности рассматриваемых текстов.

Существующие методы и подходы лингвоанализа [1] позволяют оценить адекватность и степень тесноты связи, как правило, только на качественном, логическом уровне. Но без использования математических моделей, которые могут дать вполне объективные критерии и количественные характеристики, эта оценка будет субъективной и очень приблизительной, учитывающей только авторскую доминанту как прагматическую характеристику документа. Таким образом, разработка математических методов и моделей описания и анализа тесноты связи первичного и вторичного технических текстов является актуальной и весьма востребованной задачей, имеющей важное прикладное значение.

Главная идея предложенной математической модели состоит в возможности получения количественной оценки тесноты связи первичного и вторичного документов по формальному критерию, который имеет математическую однозначную интерпретацию. Для этого в работе предложено математическое описание первичного и вторичного документов, построенное на математической теории случайных функций и случайных процессов с целочисленным аргументом. Выборка слов при этом рассматривается как статистическая функция с целочисленным аргументом, в качестве которого выступает присвоенный слову порядковый номер в выборке. Значения функции задаются частотной характеристикой использования анализируемых слов в тексте статьи или ее рефератах в соответствии с их номером. Таким образом, получаем две частотных статистических функций для двух анализируемых документов. Эти частотные функции уже сами несут, существенную информацию по общим закономерностям в использовании тех или иных слов и их частотной зависимости в анализируемых документах. Из анализа частотной диаграммы можно выбрать некоторое количество наиболее значимых в документе слов и одновременно исключить из дальнейшего рассмотрения менее значимые слова. На основании такого лингвоанализа возможно также автоматизировать выбор ключевых слов для статьи. Но более существенной является возможность дальнейшего совместного анализа этих двух функций, представляемых в виде графических гистограмм, выводимых на экран компьютера. В результате их сопоставления может быть получена доминантная диаграмма, которая позволяет анализировать уровень соответствия двух рассматриваемых документов по разностной частотной характеристике слов.

В дальнейшем две полученные частотные функции были использованы при расчете количественного критерия тесноты связи двух рассматриваемых текстов, основанного на корреляционной связи между собой соответствующих частотных функций.

Для количественной оценки тесноты связи первичного и вторичного технических документов нами предложено использовать статистическую оценку несмещенного коэффициента взаимной корреляции $r_{xy}(k)$ при k=0 [1]. Так как в реальных условиях используется конечная выборка, то коэффициент взаимной корреляции $r_{xy}(k)$ [1] должен оцениваться по выборочной статистике $C_{xy}(k)$ [2]. Расчет выборочной оценки $C_{xy}(k)$ осуществляется по формулам:

$$C_{xy}(k)=\begin{cases}\dfrac{1}{n-k}\sum\limits_{j=1}^{n-k}(x_j-\overline{x})(y_{j+k}-\overline{y})\\ \dfrac{1}{n+k}\sum\limits_{j=1}^{n+k}(y_j-\overline{y})(x_{j-k}-\overline{x}),\end{cases} \quad (1)$$

где n – число анализируемых слов в выборке; k – текущий номер слова; x_j – частотная характеристика анализируемого слова первичного текста с номером j; y_j – частотная характеристика j-го слова вторичного текста.

Расчетные значения выборочных средних арифметических функций $x(t)$ и $y(t)$ при целочисленных значениях аргумента t от 1 до n вычисляется по формулам:

$$\overline{x}=\frac{1}{n}\sum_{j=1}^{n}x_j;\overline{y}=\frac{1}{n}\sum_{j=1}^{n}y_j \quad (2)$$

Расчетные зависимости для статистики $C_{xy}(k)$ позволяют рассчитать коэффициент взаимной корреляции двух текстов, названных соответственно *X* и *Y*. Этот коэффициент является безразмерным и измеряется в пределах от 0 до 1. Расчетное значение коэффициента $r_{xy}(k)$ определяется по формуле:

$$r_{xy}(k) = \frac{C_{xy}(k)}{S_x \cdot S_y}, \tag{3}$$

где $C_{xy}(k)$ – характеристика связи тесноты текстов *X* и *Y*, рассчитываемая по формуле (1); S_x и S_y – дисперсии частотных функций $x(t)$ и $y(t)$ для текстов *X* и *Y*. Расчетные значения выборочных дисперсий, соответственно, функций $x(t)$ и $y(t)$ при целочисленных значениях аргумента *t* от 1 до *n* определяется по формулам:

$$S_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}; S_y = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2} \tag{4}$$

Приведенные выше формулы предложенной математической модели позволили разработать вычислительный алгоритм нахождения выборочной оценки коэффициента корреляции первичного и вторичного текстов, в качестве которых в настоящей работе выступают текст технической статьи и ее реферат. Последовательность вычислительной процедуры, выполняемой в соответствии с предложенной моделью следующая:

1) осуществляется формирование частотных гистограмм основного текста и реферата;
2) проверяется параметрическое согласование рассматриваемых гистограмм;
3) вычисляются выборочные средние арифметических $\bar{x}$, $\bar{y}$ по формулам (2);
4) вычисляются выборочные оценки тесноты связи функций $x(t)$ и $y(t)$ по формулам (1);
5) вычисляются выборочные дисперсии функций $x(t)$ и $y(t)$ по формулам (4);
6) находится безразмерный нормированный коэффициент взаимной корреляции $r_{xy}(k)$.

Значение коэффициента взаимной корреляции $r_{xy}(k)$ достигает верхней границы, равной 1, если два анализируемых текста (первичный и вторичный) идентичны по анализируемым признакам, в частности, по частотным характеристикам анализируемых слов или выражений, которые в этом случае совпадают. Тексты могут быть по объему знаков разными.

Значение коэффициента $r_{xy}(t)$ достигает низшего уровня $r_{xy}(k)$ лишь в том случае, если в двух анализируемых текстах вообще нет никаких совпадений по анализируемым признакам. В этом случае первичный (основной текст) и вторичный текст (реферат или аннотация) абсолютно разные.

Эти два крайних случая позволяют установить полный диапазон изменения коэффициента тесноты связи, оцениваемого характеристикой $r_{xy}(k)$, при k=0, который лежит в пределах от 0 до 1. Очевидно, что значения характеристики r_{xy}, близкие к нулю, свидетельствуют о слабой или очень слабой тесноте связи анализируемых текстов, и наоборот, если значения коэффициентов r_{xy} приближаются к 1, то связь текстов *X* и *Y* значительно тесная.

Компьютерный эксперимент был проведен с использованием разработанной авторами программы «Лингвоанализатор» [3]. Был выполнен компьютерный лингвоанализ по разработанным математическим моделям и критериям тесноты связи различных технических публикаций. На основе проведенного анализа и математической обработки результатов предложено весь диапазон изменения коэффициентов $r_{xy}(k)$ от 0 до 1 разбить на пять уровней, представленных в следующей таблице.

Таблица – Уровни тесноты связи

Уровень тесноты связи	Коэффициент тесноты связи	Характеристика связи
I	0,8…1,0	очень сильная
II	0,6…0,8	сильная
III	0,4…0,6	средняя
IV	0,2…0,4	слабая
V	0,0…0,2	очень слабая

Проведенный компьютерный эксперимент показал, что для всех проанализированных публикаций из различных научно-технических журналов и сборников статей коэффициенты тесноты связи лежат в широком пределе (от 0,15 до 0,75). Среднее значение этого коэффициента по всем использованным публикациям равно 0,45. Согласно таблице весь диапазон исследованных публикаций имеет характеристику связи от очень слабой до сильной. Среднее значение 0,45 соответствует III уровню тесноты связи со средней характеристикой связи. Необходимо отметить, что для публикаций в рецензируемых зарубежных и отечественных изданиях коэффициент r_{xy} находится в диапазоне от 0,65 до 0,85. В то же время сборники нерецензированных статей и материалов конференций имеют коэффициент связи значительно ниже – в пределах 0,1…0,6.

Список использованных источников:

1. Пугачев, В. С. Основы теории случайных функций / В. С. Пугачев. – М. : Наука, 1982.
2. Карпович-Каспжак, О. С. Вспомогательные средства прагматического анализа технического текста / О. С. Карпович-Каспжак, А. А. Метлюк // Известия Белорусской инженерной академии. – 2005. – № 1(19)/1. – С. 89–93.
3. Карпович Каспжак, О. С. Вспомогательная программная среда для прагматического анализа технического текста / О. С. Карпович-Каспжак // Актуальные проблемы радиоэлектроники: научные проблемы, подготовка кадров : сб. науч. ст. – В 3 ч. – Ч.2 / М-во образования РБ, МГВРК. – Минск : МГВРК, 2005.