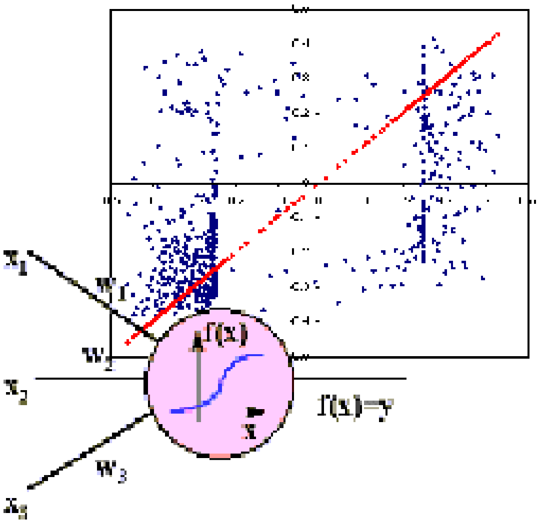
Міністерство освіти і науки України
Національна Академія наук вищої освіти України
Донбаська державна машинобудівна академія (Україна)
Academy of Professional Studies Šumadija – Kragujevac (Serbia)
Факультет інженерної механіки університета Штроссмайера (Хорватія)
Зеленогурський університет (Польща)
"American Jurnal Neural Network and Aplication" (USA)
Міжнародний університет безперервної освіти (Україна)
Інститут проблем штучного інтелекту (Україна)
ПрАТ «Новокраматорський машинобудівний завод» (Україна)
ПрАТ «Краматоський завод важкового верстатобу-дування» (Україна)
Проблемна лабораторія мобільних інтелектуальних технологічних машин (Україна)
Міжнародний університет безперервної інноваційної освіти (Україна)
ГО «Юнацький технопарк» (Україна)



ЗБІРНИК НАУКОВИХ ПРАЦЬ

XIX Міжнародної наукової конференції

«НЕЙРОМЕРЕЖНІ ТЕХНОЛОГІЇ ТА ЇХ ЗАСТОСУВАННЯ НМТіЗ-2020»

Краматорськ 2020

Міністерство освіти і науки України
Національна Академія наук вищої освіти України
Донбаська державна машинобудівна академія (Україна)
Academy of Professional Studies Šumadija – Kragujevac (Serbia)
Факультет інженерної механіки університета Штроссмайера (Хорватія)
Зеленогурський університет (Польща)
"American Jurnal Neural Network and Aplication" (USA)
Міжнародний університет безперервної освіти (Україна)
Інститут проблем штучного інтелекту (Україна)
ПрАТ «Новокраматорський машинобудівний завод» (Україна)
ПрАТ «Краматоський завод важкового верстатобу-дування» (Україна)
Проблемна лабораторія мобільних інтелектуальних технологічних машин (Україна)
Міжнародний університет безперервної інноваційної освіти (Україна)
ГО «Юнацький технопарк» (Україна)

НЕЙРОМЕРЕЖНІ ТЕХНОЛОГІЇ ТА ЇХ ЗАСТОСУВАННЯ НМТіЗ-2020

ЗБІРНИК НАУКОВИХ ПРАЦЬ

XIX Міжнародної наукової конференції

За загальною редакцією
д-ра техн. наук, проф. С. В. Ковалевського

Краматорськ 2020

УДК 004.032.26+621(061.3)
Н46

Рецензенти:

Рамазанов С.К., докт.техн.наук, докт.екон.наук, професор, Київський національний університет імені Тараса Шевченка;

Суботін С. О., докт. техн. наук, професор, Запорізький національний технічний університет

Рекомендовано
вченою радою Донбаської державної машинобудівної академії
(протокол № 5 від 26.11.2020)

Програмний комітет конференції

Н46 Нейромережні технології та їх застосування НМТіЗ-2019: збірник наукових праць XIX Міжнародної наукової конференції «Нейромережні технології та їх застосування НМТіЗ-2020» / за заг. ред. С.В.Ковалевського. - Краматорськ: ДДМА, 2020. – 169 с.

ISBN 972-966-379-965-0

У збірнику праць представлені перспективні теоретичні та практичні розробки в області нейромережних технологій, виконані в 2020 р. науковими школами України і світу. Розглядається можливість застосування нейронних мереж для управління об'єктами в режимі реального часу і особливості нейронного керування динамічними об'єктами. Наводиться ряд розробок по застосуванню нейронних мереж в різних областях практичної і науково-дослідної діяльності та створенню інтелектуальної системи для підвищення швидкості та зниження трудомісткості технологічної підготовки виготовлення нових виробів.

Для наукових працівників широкого профілю та фахівців.

УДК 004.032.26+621(061.3)

ISBN 972-966-379-965-0

УДК 004.89

Слюсар В.И. (Центральний науково-дослідний інститут озброєння та військової техніки ЗС України)

ПРИМЕНЕНИЕ ТОРЦЕВОГО ПРОИЗВЕДЕНИЯ МАТРИЦ В ЗАДАЧАХ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Рассмотрены варианты применения торцевого произведения матриц инцидентности для решения задач анализа текстов, в частности, определения частоты встречаемости трех, четырех и более слов в предложениях отдельно взятого корпуса текста.

Variants of using the face-splitting product of incidence matrices for solving problems of text analysis, in particular, determining the frequency of occurrence of three, four or more words in sentences of a separate text corpus are considered.

Обработка естественного языка (Natural Language Processing, NLP) в последнее время стала одним из важных направлений применения технологий искусственного интеллекта. При этом многие варианты NLP базируются на использовании теории графов. В последнее время такой подход получил дальнейшее развитие на основе применения предложенного автором в 1996 году торцевого произведения матриц [1 - 4]. Первой публикацией такого рода применительно к NLP стал препринт [5], в котором для уменьшения количества вычислений при анализе текстов была описана общая идея замены кронекеровского произведения матриц на торцевое. Такой метод позволил получить в [5] достаточно обнадеживающие результаты, хотя некоторые важные аспекты применения торцевого произведения остались за пределами указанной публикации. К тому же, допущенная автором [5] опечатка в примере с матрицей совместной встречаемости (*co-occurrence matrix*) может стать причиной заблуждений читателей в контексте правильного использования соответствующего математического аппарата.

Целью работы является изложение результатов применения торцевого произведения матриц для решения задач анализа текстов естественного языка.

Чтобы сохранить преемственность в демонстрации соответствующих возможностей торцевого произведения матриц, ограничимся для примера использованием данной матричной операции применительно к решению задачи анализа текстового фрагмента из трех предложений, рассмотренных в [5]:

1) I like math; 2) You like math; 3) I like you.

Составим для этого набора предложений так называемую *матрицу инцидентности*. Ее строки будут соответствовать конкретному предложению, а столбцы - отдельно взятому слову. При этом числовые элементы в каждой строке будут указывать на то, сколько раз то или иное слово встречается в данном предложении.

Количество столбцов должно соответствовать максимальному количеству слов в отдельно взятом предложении рассматриваемого текста. В указанном фрагменте наибольшее количество слов – четыре, – имеет второе предложение. Прежде чем непосредственно перейти к матрице инцидентности, составим с учетом сказанного для наглядности таблицу, ей соответствующую.

Таблица 1

Порядковый номер предложения	X_1=I	X_2=like	X_3=math	X_4=you
1	1	1	1	0
2	0	1	1	1
3	1	1	0	1

Как видно, все слова в указанном корпусе текста встречаются в предложениях не более одного раза. Отсюда, получим матрицу инцидентности вида:

$$G = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}.$$

Квадратичная форма на основе указанной матрицы типа G^TG именуется *матрицей совместной встречаемости (co-occurrence matrix)* [5]. Для рассматриваемого примера она будет следующей:

$$G^TG = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & 3 & 2 & 2 \\ 1 & 2 & 2 & 1 \\ 1 & 2 & 1 & 2 \end{bmatrix}.$$

Данная матрица симметрична относительно своей главной диагонали. Согласно теории NLP, элементы на главной диагонали полученной матрицы показывают, сколько раз отдельно взятое слово встречается в анализируемом тексте. Кстати, получить эту информацию можно, сложив все строки инцидентной матрицы, что эквивалентно операции умножения ее на вектор-строку единиц

$$\mathbf{1}^T G = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} G = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix},$$

или аналогично, умножив транспонированную матрицу G на вектор единиц

$$G^T \mathbf{1} = G^T \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}^T \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 2 \end{bmatrix}$$

Таким образом, для получения информации об элементах главной диагонали матрицы совместной встречаемости можно обойтись без формирования самой матрицы, что упрощает дело. Однако необходимость в формировании такой матрицы обусловлена наличием в ней дополнительной информации. Чтобы пояснить это, перейдем от матрицы G^TG к эквивалентной таблице, с помощью которой можно наглядно представить назначение полученных матричных элементов (табл. 2).

Таблица 2

	X_1=I	X_2=like	X_3=math	X_4=you
X_1=I	2	2	1	1
X_2=like	2	3	2	2
X_3=math	1	2	2	1
X_4=you	1	2	1	2

Нетрудно заметить, что элементы строк и столбцов указывают, сколько раз конкретное слово встречается в тексте в сочетании с другим отдельно взятым словом (отсюда происходит название матрицы совместной встречаемости). Например, из первой строки следует, что слово X_1=I дважды совместно присутствует в тексте в одном предложении со словом X_2=like и по одному разу – с остальными. Это можно также трактовать как количество соответствующих пар слов (биграмм) в тексте.

Числа на пересечениях одноименных столбцов и строк соответствуют, как уже отмечалось, частоте появления того или иного слова в анализируемом фрагменте. Это свойство рассматриваемой матрицы позволяет применять ее для расшифровки текстов по частоте появления слов, сопоставляя полученный результат со специальным словарем наиболее встречающихся слов данного языка. Аналогичная процедура может быть проделана и для букв (символов), использованных в тексте.

След от квадратичной формы $tr(G^TG)$, равный сумме диагональных элементов квадратной матрицы, соответствует полному числу слов в тексте (без учета их содержания). В данном случае эта величина равна 9. Чтобы получить распределение общего количества слов по отдельным предложениям, следует рассчитать другую квадратичную форму - GG^T. В рассматриваемом примере она имеет вид:

$$GG^{T}=\begin{bmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{bmatrix}\begin{bmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{bmatrix}^{T}=\begin{bmatrix}3&2&2\\2&3&2\\2&2&3\end{bmatrix}$$

Таким образом, элементы на главной диагонали матрицы GG^{T} соответствуют количеству слов в предложениях, равному в данном случае 3. Элементы в строках или столбцах указывают на количество одинаковых слов, используемых в паре соответствующих предложений. Например, из первой строки следует, что первое предложение имеет два слова, совпадающих со вторым предложением (пересечение первой строки и второго столбца) и два слова – с третьим из предложений (соответственно первый элемент третьего столбца матрицы GG^{T}).

Разобравшись с основными возможностями классического матричного аппарата, применяемого для анализа текстов, перейдем далее к более сложной задаче анализа большего количества сочетаний слов. Как указано в [5], для этого необходимо воспользоваться *торцевым произведением матриц*. В частности, для анализа парных сочетаний слов отправной точкой должна стать вторичная матрица инцидентности, сформированная из исходной матрицы G для отдельно взятых слов с помощью выражения:

$$I2=G\ \square\ G,$$

где □ – символ торцевого произведения матриц.

Применительно к рассматриваемому случаю получим:

$$I2=G\ \square\ G=\begin{bmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{bmatrix}\ \square\ \begin{bmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{bmatrix}=$$

$$=\begin{bmatrix}1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0\\0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1\\1&1&0&1&1&1&0&1&0&0&0&0&1&1&0&1\end{bmatrix}.$$

Для пояснения смысла элементов полученной матрицы снова обратимся к табличной форме представления данных (см. табл. 3).

Таблица 3

№ предл	X_1				X_2				X_3				X_4			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
2	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
3	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1

Строка таблицы, а соответственно и матрицы инцидентности, показывает в данном случае, сколько раз та или иная пара слов встречается в конкретном предложении. При этом по паре, образованной дублированием одного и того же слова (X_mX_m, m= 1; 2; 3; 4), по–прежнему можно судить о том, сколько раз такое слово включено в разные предложения.

Согласно [5], матрица совместной встречаемости для анализа парных словосочетаний может быть сформирована на основе исходной матрицы инцидентности и ее версии в виде торцевого произведения:

$$C=G^{T}(G\ \square\ G)=\begin{bmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{bmatrix}^{T}\begin{bmatrix}1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0\\0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1\\1&1&0&1&1&1&0&1&0&0&0&0&1&1&0&1\end{bmatrix}=$$

$$=\left[\begin{array}{cccc:cccc:cccc:cccc}2&2&1&1&2&2&1&1&1&1&1&0&1&1&0&1\\2&2&1&1&2&3&2&2&1&2&2&1&1&2&1&2\\1&1&1&0&1&2&2&1&1&2&2&1&0&1&1&1\\1&1&0&1&1&2&1&2&0&1&1&1&1&2&1&2\end{array}\right]$$

Полученная в результате матрица имеет размерность 16×4 и может быть представлена в виде 4 блоков, каждый из которых соответствует одному из слов в парных словосочетаниях (см. табл. 4). На главной диагонали каждого из этих блоков расположены числа, показывающие, сколько раз во всем тексте в отдельно взятом предложении встречается та или иная пара слов. Например, из первого блока видно, что слово I (X_1X_1) представлено в тексте дважды. Также 2 раза встречается пара слов I like (X_1X_2), а по одному разу в предложениях присутствуют словосочетания I math (X_1X_3) и I you (X_1X_4). При этом необязательно, чтобы указанные слова в паре располагались подряд. Кроме того, повторяющиеся слова в комбинации засчитываются как одно слово, например $X_1X_2X_2$

Таблица 4

	X_1=I				X_2=like				X_3=math				X_4=you			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
X_1	2	2	1	1	2	2	1	1	1	1	1	0	1	1	0	1
X_2	2	2	1	1	2	3	2	2	1	2	2	1	1	2	1	2
X_3	1	1	1	0	1	2	2	1	1	2	2	1	0	1	1	1
X_4	1	1	0	1	1	2	1	2	0	1	1	1	1	2	1	2

Существенно, что элементы 4-го квартета столбцов табл. 4 и соответствующей блочной матрицы отличается от приведенного в [5] выражения, содержащего опечатки.

Цифры вне главной диагонали матрицы C характеризуют частоту появления конкретных троек слов в исследуемом тексте без учета их порядка в последовательности. Например, из первого блока матрицы следует, что набор слов I like math ($X_1X_2X_3$) встречается однажды, равно как и I like you ($X_1X_2X_4$). Кроме того, повторяющиеся слова в комбинации засчитываются как одно слово, например, набор $X_1X_2X_2$ эквивалентен сочетанию X_1X_2, а $X_mX_mX_m= X_m$ для любого m.

След, взятый для каждого из блоков как сумма элементов на их главной диагонали, показывает, сколько парных сочетаний образует слово в данном тексте с другими словами и с самим собой. К примеру, согласно первому блоку, след которого равен 6, слово I образует 6 парных сочетаний, из которых две пары соответствуют непосредственно этому слову в первом и третьем предложениях.

Альтернативой формированию матрицы совместной встречаемости при анализе парных словосочетаний является расчёт квадратичной формы $I2^TI2$:

$$I2^TI2= (G \square G)^T (G \square G)= (G^T \blacksquare G^T) (G \square G),$$

где ■ – символ произведения Хатри-Рао.

При этом получаем квадратную матрицу размерностью 16×16, элементы которой представлены в табл. 5. След данной матрицы $tr(I2^TI2)=27$ соответствует полному количеству пар во всех предложениях с учетом взаимной обратимости порядка следования слов в паре и сочетания слова с самим собой. Особенностью этой блочной матрицы является расположение на ее главной диагонали блоков рассмотренной выше матрицы C. Кроме того, блоки, расположенные симметрично по обе стороны относительно главной диагонали, совпадают. Эти блоки вне главной диагонали позволяют получить информацию о квартетах слов, что является важнейшим преимуществом данной квадратичной формы. При этом элементы главной диагонали блока, расположенного на главной диагонали блочной матрицы, характеризуют частоту появления одного слова и их пары, остальные элементы этих боков характеризуют распространенность в тексте пар и троек слов. Блоки, расположенные вне главной диагонали блочной матрицы, на своей главной диагонали содержат элементы, описывающие количество тех или иных троек слов в тексте, а остальные элементы этих блоков охватывают квартетные комбинации слов. Например, как следует из табл. 5, дублирующей в удобном виде содержимое матрицы $I2^TI2$, в рассматриваемом фрагменте текста отсутствуют комбинации $X_1X_2X_3X_4$, $X_1X_3X_4X_m$ (элементы матрицы, находящиеся на пересечении соответствующих строк, столбцов, блок-строк и блок-столбцов равны нулю). След от блока, расположенного вне главной диагонали матрицы, также указывает на общее количество соответствующих комбинаций символов с учетом их повторяемости.

Таблица 5

		X_1=I				X_2=like				X_3=math				X_4=you			
		X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
X_1	X_1	2	2	1	1	2	2	1	1	1	1	1	0	1	1	0	1
	X_2	2	2	1	1	2	2	1	1	1	1	1	0	1	1	0	1
	X_3	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
	X_4	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1
X_2	X_1	2	2	1	1	2	2	1	1	1	1	1	0	1	1	0	1
	X_2	2	2	1	1	2	3	2	2	1	2	2	1	1	2	1	2
	X_3	1	1	1	0	1	2	2	1	1	2	2	1	0	1	1	1
	X_4	1	1	0	1	1	2	1	2	0	1	1	1	1	2	1	2
X_3	X_1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
	X_2	1	1	1	0	1	2	2	1	1	2	2	1	0	1	1	1
	X_3	1	1	1	0	1	2	2	1	1	2	2	1	0	1	1	1
	X_4	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
X_4	X_1	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1
	X_2	1	1	0	1	1	2	1	2	0	1	1	1	1	2	1	2
	X_3	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
	X_4	1	1	0	1	1	2	1	2	0	1	1	1	1	2	1	2

Другая квадратичная форма имеет вид:

$$I2\ I2^T = (G \square G)\,(G \square G)^T = (G \square G)(G^T \blacksquare G^T) = GG^T \circ GG^T = \begin{bmatrix} 9 & 4 & 4 \\ 4 & 9 & 4 \\ 4 & 4 & 9 \end{bmatrix}.$$

Ее след совпадает со следом предыдущей квадратичной формы, т. е.

$$tr[(G \square G)\,(G \square G)^T] = tr[(G \square G)(G^T \blacksquare G^T)] = tr[(G^T \blacksquare G^T)\,(G \square G)] = 27.$$

При этом отдельно взятый элемент главной диагонали соответствует количеству всевозможных парных сочетаний слов, которые могут быть образованы внутри данного предложения с учетом взаимной перестановки порядка слов и сочетания слова с самим собой. В данном случае эта величина становит 9. Цифры вне главной диагонали указывают общее количество таких словопар для двух сравниваемых предложений (равно 4). Чтобы получить лишь значения элементов главной диагонали матрицы совместной встречаемости, достаточно умножить слева матрицу инцидентности на вектор строку единиц:

$$\mathbf{1}^T G2 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} G2 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} (G \square G) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} \square \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} \right) =$$

$$= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} =$$

$$= \left[2 \quad 2 \quad 1 \quad 1 \,\vdots\, 2 \quad 3 \quad 2 \quad 2 \,\vdots\, 1 \quad 2 \quad 2 \quad 1 \,\vdots\, 1 \quad 2 \quad 1 \quad 2 \right]$$

Аналогично с помощью торцевого произведения можно выйти на подсчет количества совпадений сразу трех, четырех и более слов в фразах.

Дальнейшее расширение функциональных возможностей обработки текста в рассматриваемом контексте задач позволяет получить переход к тройному торцевому произведению матриц инцидентности:

$$I3 = G \square G \square G,$$

где □ – символ торцевого произведения матриц.

Для анализируемого фрагмента текста такое произведение приводит к 16-блочной матрице, состоящей из 4 четвёрок блоков:

$$I3=G \square G \square G=\begin{vmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{vmatrix} \square \begin{vmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{vmatrix} \square \begin{vmatrix}1&1&1&0\\0&1&1&1\\1&1&0&1\end{vmatrix}=$$

$$=\left[\begin{array}{cccc:cccc:cccc:cccc:}1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\1&1&0&1&1&1&0&1&0&0&0&0&1&1&0&1\end{array}\right.$$

$$\begin{array}{:cccc:cccc:cccc:cccc:}1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0\\0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1\\1&1&0&1&1&1&0&1&0&0&0&0&1&1&0&1\end{array}$$

$$\begin{array}{:cccc:cccc:cccc:cccc:}1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0\\0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\end{array}$$

$$\left.\begin{array}{:cccc:cccc:cccc:cccc}0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1\\1&1&0&1&1&1&0&1&0&0&0&0&1&1&0&1\end{array}\right]$$

Значения элементов матрицы I3 соответствуют количеству троек символов, присутствующих в конкретном предложении. Например, нулевая первая строка в последней четверке блоков обусловлена отсутствием в первом из предложений слова you. Аналогично нулевая третья срока в третьем квартетном блоке, расположенном выше предыдущего, соответствует отсутствию слова math в третьем из предложений. Существенно, что в полученной структуре блочной матрицы при индексации столбцов внешнее слева число в нижнем индексе меняется наиболее медленно:

$X_1X_1X_1$ ... $X_1X_1X_4$ ... $X_1X_2X_1$ ... $X_1X_2X_4$ ... $X_1X_4X_1$ ... $X_1X_4X_4$ ...
$X_2X_1X_1$ ... $X_2X_1X_4$ ... $X_2X_2X_1$ ... $X_2X_2X_4$ ... $X_2X_4X_1$ ... $X_2X_4X_4$ ... и т.д.

Подробно это пояснено в табл. 6.

Чтобы получить собственно матрицу совместной встречаемости, следует сформировать матричное произведение G^TI3 размерностью 64×4. Его результат позволяет охватить не только тройки, но и четверки слов (символов).

Важный вывод, который следует сделать из представленных результатов, состоит в возможности понижения размерности анализируемого набора данных за счет применения свойств торцевого произведения:

$$diag[(G^T \blacksquare G^T)(G \square G)]= \mathbf{1}^T (G \square G);\ bdiag[(G^T \blacksquare G^T)(G \square G)]=G^T(G \square G);$$
$$diag[(G^T \blacksquare G^T\blacksquare G^T)(G \square G\square G)]= \mathbf{1}^T (G \square G\square G);$$
$$bdiag[(G^T \blacksquare G^T\blacksquare G^T)(G \square G\square G)]= G^T(G \square G\square G),$$

где diag(A) – вектор-строка из диагональных элементов матрицы A,

bdiag(A) – блочная вектор-строка, сформированная из блоков главной диагонали матрицы A.

Таблица 6

№ предл	X_1=I (словосочетания $X_1X_nX_m$)															
	X_1=I				X_2=like				X_3=math				X_4=you			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1
№ предл	X_2=like (словосочетания $X_2X_nX_m$)															
	X_1=I				X_2=like				X_3=math				X_4=you			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
2	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
3	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1
№ предл	X_3=math (словосочетания $X_3X_nX_m$)															
	X_1=I				X_2=like				X_3=math				X_4=you			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
2	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
№ предл	X_4=you (словосочетания $X_4X_nX_m$)															
	X_1=I				X_2=like				X_3=math				X_4=you			
	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1
3	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1

ВЫВОДЫ

В целом можно заключить, что максимальное количество слов N, доступное для анализа с помощью квадратичных форм типа $I2^TI2$, $I3^TI3$ и т.д. равно удвоенной величине кратности торцевого произведения, использованного для формирования матрицы инцидентности, то есть:

$$N=2K,$$

где K - количество сомножителей в торцевом произведении первичных матриц инцидентности.

Для совместного анализа нескольких корпусов текста может быть использовано блочное торцевое произведение [2 - 4].

Литература

1. Слюсар В.И. Торцевые произведения матриц в радиолокационных приложениях// Известия высших учебных заведений. Радиоэлектроника.- 1998. - Том 41, № 3.- С. 71 - 75.
2. Слюсар В.И. Семейство торцевых произведений матриц и его свойства// Кибернетика и системный анализ. – 1999.- Том 35; № 3.- С. 379-384.- DOI: 10.1007/BF02733426
3. Слюсар В.И. Обобщенные торцевые произведения матриц в моделях цифровых антенных решеток с неидентичными каналами.//Известия высших учебных заведений. Радиоэлектроника.- 2003. - Том 46, № 10. - С. 15 - 26.
4. Основы военно-технических исследований. Теория и приложения. Том. 2. Синтез средств информационного обеспечения вооружения и военной техники. / А.И. Миночкин, В.И. Рудаков, В.И. Слюсар. – Киев:«Гранмна, 2012. – С. 7 – 98; 354 – 521.
5. Bryan Bischof. Higher order co-occurrence tensors for hypergraphs via face-splitting. Published 15 February, 2020, Mathematics, Computer Science, - https://arxiv.org/abs/2002.06285 ArXiv.