ISSN 2618-0332

Міністерство освіти і науки України

Херсонський національний технічний університет

ПРИКЛАДНІ ПИТАННЯ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Т. 3, № 2.2

Рекомендовано до друку Вченою радою
Херсонського національного технічного університету
(протокол № 8 від 29 червня 2020 року)

Журнал включений до Реєстру наукових фахових видань України категорії Б
на підставі Наказу МОН України від 17 березня 2020 року № 409.

Журнал включено до наукометричних баз, електронних бібліотек та репозитаріїв:
Google Scholar, Index Copernicus International Journal Master List,
CiteFactor Academic Scientific Journals, National Library of Ukraine (Vernadsky).

Херсон 2020

122 – Комп’ютерні науки

Борисенко В.Д.	**д.т.н., професор**
Ванін В.В.	**д.т.н., професор**
Вірченко Г.А.	**д.т.н., професор**
Гнатушенко В.В.	**д.т.н., професор**
Гумен О.М.	**д.т.н., професор**
Корчинський В.М.	**д.т.н., професор**
Литвиненко В.І.	**д.т.н., професор**
Мартин Є.В.	**д.т.н., професор**
Найдиш А.В.	**д.т.н., професор**
Несвідомін В.М.	**д.т.н., професор**
Пилипака С.Ф.	**д.т.н., професор**
Тулученко Г.Я.	**д.т.н., професор**
Устенко С.А.	**д.т.н., професор**
Шоман О.В.	**д.т.н., професор**

126 – Інформаційні системи та технології

Аль-Амморі А.Н.	**д.т.н., професор**
Баклан І.В.	**к.т.н., доцент**
Бень А.П.	**к.т.н., доцент**
Левикін В.М.	**д.т.н., професор**
Литвиненко О.І.	**к.т.н., доцент**
Мороз Б.І.	**д.т.н., професор**
Стеценко І.В.	**д.т.н., професор**
Шерстюк В.Г.	**д.т.н., професор**

151 – Автоматизація та комп’ютерно-інтегровані технології

Алексєєв М.О.	**д.т.н., професор**
Бардачов Ю.М.	**д.т.н., професор**
Головко В.І.	**д.т.н., професор**
Кондратець В.О.	**д.т.н., професор**
Мещеряков Л.І.	**д.т.н., професор**
Омельчук А.А.	**к.т.н.**
Осадчий С.І.	**д.т.н., професор**
Рожков С.О.	**д.т.н., професор**
Рудакова Г.В.	**д.т.н., професор**

Інші спеціальності

Мельник І.В.	**д.т.н., професор**
Розов Ю.Г.	**д.т.н., професор**

ISSN 2618-0332

Министерство образования и науки Украины

Херсонский национальный технический университет

ПРИКЛАДНЫЕ ВОПРОСЫ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Т. 3, № 2.2

Рекомендовано к печати Ученым советом
Херсонского национального технического университета
(протокол № 8 от 29 июня 2020 года)

Журнал включен в Реестр научных специализированных изданий Украины
категории Б на основании Приказа МОН Украины от 17 марта 2020 года № 409.

Журнал включен в наукометрические базы, электронные библиотеки и репозитарии:
Google Scholar, Index Copernicus International Journal Master List,
CiteFactor Academic Scientific Journals, National Library of Ukraine (Vernadsky).

Херсон 2020

Редакционный совет

Члены редакционной коллегии по специальностям:

Иностранные специалисты

113 – Прикладная математика

122 – Компьютерные науки

Борисенко В.Д.	**д.т.н., профессор**
Ванин В.В.	**д.т.н., профессор**
Вирченко Г.А.	**д.т.н., профессор**
Гнатушенко В.В.	**д.т.н., профессор**
Гумен Е.Н.	**д.т.н., профессор**
Корчинский В.М.	**д.т.н., профессор**
Литвиненко В.И.	**д.т.н., профессор**
Мартин Е.В.	**д.т.н., профессор**
Найдыш А.В.	**д.т.н., профессор**
Несвидомин В.Н.	**д.т.н., профессор**
Пилипака С.Ф.	**д.т.н., профессор**
Тулученко Г.Я.	**д.т.н., профессор**
Устенко С.А.	**д.т.н., профессор**
Шоман О.В.	**д.т.н., профессор**

126 – Информационные системы и технологии

Аль-Аммори А.Н.	**д.т.н., профессор**
Баклан И.В.	**к.т.н., доцент**
Бень А.П.	**к.т.н., доцент**
Левыкин В.М.	**д.т.н., профессор**
Литвиненко Е.И.	**к.т.н., доцент**
Мороз Б.И.	**д.т.н., профессор**
Стеценко И.В.	**д.т.н., профессор**
Шерстюк В.Г.	**д.т.н., профессор**

151 – Автоматизация и компьютерно-интегрированные технологии

Алексеев М.А.	**д.т.н., профессор**
Бардачев Ю.Н.	**д.т.н., профессор**
Головко В.И.	**д.т.н., профессор**
Кондратец В.А.	**д.т.н., профессор**
Мещеряков Л.И.	**д.т.н., профессор**
Омельчук А.А.	**к.т.н.**
Осадчий С.И.	**д.т.н., профессор**
Рожков С.А.	**д.т.н., профессор**
Рудакова А.В.	**д.т.н., профессор**

Другие специальности

Мельник И.В.	**д.т.н., профессор**
Розов Ю.Г.	**д.т.н., профессор**

ISSN 2618-0332

Ministry of Education and Science of Ukraine

Kherson National Technical University

APPLIED QUESTIONS OF MATHEMATICAL MODELLING

V. 3, № 2.2

Recommended for publication by the Academic Council of
Kherson National Technical University
(Minutes № 8 on 29th June 2020)

The journal is included in the Register of scientific specialized publications of Ukraine of category B on the basis of Minutes of the Ministry of Education and Science of Ukraine dated March 17, 2020 № 409.

The journal is included in the scientometric bases, electronic libraries and repositories: Google Scholar, Index Copernicus International Journal Master List, CiteFactor Academic Scientific Journals, National Library of Ukraine (Vernadsky).

Kherson 2020

Editorial Board

Members of Editorial Board by specialities:

Foreign Specialists

113 – Applied Mathematics

122 – Computer Science

Borysenko V.D.	Doctor of Engineering Science, Professor
Vanin V.V.	Doctor of Engineering Science, Professor
Virchenko H.A.	Doctor of Engineering Science, Professor
Hnatushenko V.V.	Doctor of Engineering Science, Professor
Humen O.M.	Doctor of Engineering Science, Professor
Korchynskyi V.M.	Doctor of Engineering Science, Professor
Lytvynenko V.I.	Doctor of Engineering Science, Professor
Martyn Ye.V.	Doctor of Engineering Science, Professor
Naidysh A.V.	Doctor of Engineering Science, Professor
Nesvidomin V.M.	Doctor of Engineering Science, Professor
Pylypaka S.F.	Doctor of Engineering Science, Professor
Tuluchenko H.Ya.	Doctor of Engineering Science, Professor
Ustenko S.A.	Doctor of Engineering Science, Professor
Shoman O.V.	Doctor of Engineering Science, Professor

126 – Information Systems and Technologies

Al-Ammori A.N.	Doctor of Engineering Science, Professor
Baklan I.V.	Ph.D., Associate Professor
Ben A.P.	Ph.D., Associate Professor
Levykin V.M.	Doctor of Engineering Science, Professor
Lytvynenko O.I.	Ph.D., Associate Professor
Moroz B.I.	Doctor of Engineering Science, Professor
Stetsenko I.V.	Doctor of Engineering Science, Professor
Tomashevskyi V.M.	Doctor of Engineering Science, Professor
Sherstiuk V.H.	Doctor of Engineering Science, Professor

151 – Automation and Computer Integrated Technologies

Aleksieiev M.O.	Doctor of Engineering Science, Professor
Bardachov Yu.M.	Doctor of Engineering Science, Professor
Holovko V.I.	Doctor of Engineering Science, Professor
Kondratets V.O.	Doctor of Engineering Science, Professor
Meshcheriakov L.I.	Doctor of Engineering Science, Professor
Omelchuk A.A.	Ph.D.
Osadchyi S.I.	Doctor of Engineering Science, Professor
Rozhkov S.O.	Doctor of Engineering Science, Professor
Rudakova H.V.	Doctor of Engineering Science, Professor

Other Specialties

Melnyk I.V.	Doctor of Engineering Science, Professor
Rozov Yu.H.	Doctor of Engineering Science, Professor

ЗМІСТ

СОДЕРЖАНИЕ

CONTENTS

УДК 004.91 + 81'32

К.С. ГАЙДУК, О.Г. ШЕВЧЕНКО, В.А. СВЯТНЫЙ
Донецкий национальный технический университет МОН Украины

ОЦЕНКА ТОЧНОСТИ ИЗВЛЕЧЕНИЯ КОНЦЕПТОВ И ПОНЯТИЙ НА ОСНОВАНИИ МЕР АССОЦИАЦИИ

В работе представлены результаты оценки качества двоичной классификации пар слов (*биграмм*) *на основании различных мер ассоциации, в ходе которой выполнялось разделение биграмм на классы «концепты и понятия» и «прочие биграммы». Показано, что обычное ранжирование объектов на основании значений меры ассоциации, с последующим применением пороговой фильтрации (либо отбором фиксированного количества первых элементов сортированного списка), позволяет получить лишь некоторую вершину рейтинга, но не позволяет достичь эффективного решения задачи классификации.*

Предложенный авторами подход основан на пороговой фильтрации не значений меры ассоциации, но вероятности принадлежности биграммы классу «концепты и понятия» при заданном значении меры ассоциации. Указанная вероятность рассчитывается на основании значений функций плотности вероятности (*ФПВ*), *соответствующих распределениям меры ассоциации как случайной величины в обоих классах. Построение эмпирических ФПВ выполнено посредством анализа размеченной обучающей выборки.*

Определение порогового значения вероятности сведено к решению одномерной задачи оптимизации, в ходе которой максимизируется отношение количества объектов, идентифицированных как «концепты и понятия», к количеству объектов, отнесенных к классу «прочие биграммы». Определение характера статистического распределения большинства рассмотренных мер ассоциации вызывает затруднение (*отклонение нулевой гипотезы для основных известных распределений по итогам* χ^2-*теста*), *в силу чего была использована аппроксимация ФПВ методом окна Парзена-Розенблатта. Подобное решение позволило существенно увеличить качество классификации* (*прирост* F_1-*меры до 58% для отдельных мер ассоцации*).

Выполненный корреляционный анализ мер ассоциации позволил выделить два кластера: меры, ориентированные на силу связи в коллокации, и меры, ориентированные на частоту встречаемости коллокации. Функция логарифмического правдоподобия и критерий Стьюдента примерно в равной степени учитывают оба указанных фактора.

Установлено, что применение функции логарифмического правдоподобия (*как меры ассоциации*), *совместно с предложенным алгоритмом пороговой фильтрации, позволяет достичь классификации с единичным значением* F_1-*меры* (*по данным, полученным для использованных обучающей и тестовой выборок*).

Ключевые слова: выделение понятий и концептов; коллокации; меры ассоциации; классификация; функция логарифмического правдоподобия; метод KDE.

К.С. ГАЙДУК, О.Г. ШЕВЧЕНКО, В.А. СВЯТНИЙ
Донецький національний технічний університет МОН України

ОЦІНКА ТОЧНОСТІ ВИДІЛЕННЯ КОНЦЕПТІВ І ПОНЯТЬ НА ОСНОВІ МІР АСОЦІАЦІЇ

В роботі наведено результати оцінки якості двійкової класифікації пар слів (*біграм*) *на підставі різних мір асоціації, в ході якої виконувався поділ біграм на класи*

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

«концепти і поняття» та «інші біграми». Показано, що звичайне ранжування об'єктів на підставі значень мір асоціації, з подальшим застосуванням порогової фільтрації (або відбором фіксованої кількості перших елементів сортованого списку), дозволяє отримати лише деяку вершину рейтингу, але не дозволяє досягти ефективного вирішення задачі класифікації.

Запропонований авторами підхід заснований на пороговій фільтрації не значень міри асоціації, але ймовірності приналежності біграми класу «концепти і поняття» при заданому значенні міри асоціації. Вказана ймовірність розраховується на підставі значень функцій густини ймовірності (ФГЙ), що відповідають розподілам міри асоціації як випадкової величини в обох класах. Побудову емпіричних ФГЙ виконано шляхом аналізу розміченої навчальної вибірки.

Визначення порогового значення ймовірності зведено до вирішення одновимірної задачі оптимізації, в ході якої максимізується відношення кількості об'єктів, ідентифікованих як «концепти і поняття», до кількості об'єктів, віднесених до класу «інші біграми». Визначення характеру статистичного розподілу більшості розглянутих мір асоціації викликає труднощі (відхилення нульової гіпотези для основних відомих розподілів за результатами χ^2-тесту), з урахуванням чого було використано апроксимацію ФГЙ методом вікна Парзена-Розенблатта. Подібне рішення дозволило істотно збільшити якість класифікації (приріст F_1-міри до 58% для окремих мір ассоціації).

Виконаний кореляційний аналіз мір асоціації дозволив виділити два кластери: міри, орієнтовані на силу зв'язку в колокації, та міри, орієнтовані на частоту зустрічаємості колокації. Функція логарифмічної правдоподібності та критерій Стьюдента приблизно в рівній мірі враховують обидва зазначені чинники.

Встановлено, що застосування функції логарифмічної правдоподібності (як міри асоціації), спільно із запропонованим алгоритмом порогової фільтрації, дозволяє досягти класифікації з одиничним значенням F_1-міри (за даними, отриманими для використаних навчальної та тестової вибірок).

Ключові слова: виділення понять та концептів; колокації; міри асоціації; класифікація; функція логарифмічної правдоподібності; метод KDE.

K.S. HAIDUK, O.H. SHEVCHENKO, V.A. SVIATNYI
Donetsk National Technical University MES of Ukraine

ASSESSMENT OF THE ACCURACY OF NOTION AND CONCEPT EXTRACTION BASED ON MEASURES OF ASSOCIATION

The paper presents the results of assessing the quality of the binary classification of pairs of words (bigrams) on the basis of various measures of association, during which the bigrams were divided into classes 'concepts and notions' and 'other bigrams'. It is shown that the usual ranking of objects based on the values of the association measure, followed by the use of threshold filtering (or selection of a fixed number of the first elements of the sorted list), allows you to get only a certain top of the rating, but does not allow you to achieve an effective solution to the classification problem.

The approach proposed by the authors is based on the threshold filtering not of the values of the association measure, but the probability of the bigram belonging to the class 'concepts and notions' for a given value of the association measure. The indicated probability is calculated based on the values of the probability density functions (PDFs) corresponding to the distributions of the association measure as a random variable in both classes. The construction of empirical PDFs was performed by analyzing the labeled training sample.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

Determination of the threshold value of the probability is reduced to solving a one-dimensional optimization problem, during which the ratio of the number of objects identified as 'concepts and notions' to the number of objects classified as 'other bigrams' is maximized. Determination of the nature of the statistical distribution of most of the considered association measures is difficult (*rejection of the null hypothesis for the main known distributions based on the results of the* χ^2*-test*), *due to which the PDF was approximated by the Parzen-Rosenblatt window method. Such a solution made it possible to significantly increase the quality of the classification* (*an increase in the* F_1*-measure up to 58% for certain association measures*).

The performed correlation analysis of measures of association made it possible to distinguish two clusters: measures focused on the strength of connection in a collocation, and measures focused on the frequency of occurrence of collocation. The logarithmic likelihood function and Student's t test take into account both of these factors approximately equally.

It was found that the use of the log-likelihood function (as a measure of association), together with the proposed threshold filtering algorithm, makes it possible to achieve a classification with a value of the F_1*-measure equal to one* (*according to the data obtained for the training and test samples used*).

Keywords: extraction of notions and concepts; collocations; measures of association; classification; function of logarithmic likelihood; KDE method.

Постановка проблемы

Коллокации являются важными объектами исследования компьютерной лингвистики, и представляют собой устойчивые словосочетания, состоящие из слов-коллокатов. Под устойчивостью в данном случае подразумевается ограниченное количество слов, с которыми может встречаться в паре данное, а также регулярность появления соответствующей комбинации в текстах, что делает возможным использование для выявления коллокаций статистических мер (мер ассоциации) [1]. В роли коллокаций могут выступать [2–3]: ключевые слова – слова или словосочетания, в своей совокупности обеспечивающие высокоуровневое описание содержания текста, и отражающие его тематику; понятия - отражения в мышлении объектов и явлений на основании их существенных признаков; концепты – в отличие от понятий, учитывают также несущественные признаки объектов и явлений; термины – названия понятий определенной области; именованные сущности – слова или словосочетания, обозначающие явление или предмет определенной категории; устойчивые обороты речи (описание некоторого действия или события определенными словами), идиомы и др. С учетом сказанного, возникает проблема идентификации выделенных из текста коллокаций - их интерпретации как понятий, терминов, ключевых слов и пр., что актуально для специалистов различных областей (информационный поиск, извлечение информации, инженерия знаний, лингвистика и др.) [1–2].

Меры ассоциации обладают различными возможностями в контексте выделения коллокаций того или иного вида [1–3], что широко используется для выделения ключевых слов, терминов и оборотов речи [1, 4–6], однако проблема выделения понятий и концептов на сегодняшний день всё еще остается актуальной [7–8]. Эффективным в данном контексте представляется совместное использование различных мер ассоциации, лингвистических шаблонов, систем продукций и методов машинного обучения [4, 8–10].

Анализ последних исследований и публикаций

Вопросам изучения мер ассоциации, а также их сравнительному анализу, посвящен ряд работ, среди которых [3, 6, 11–14] и др. Возможности использования

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

статистического и лингвистического подходов к извлечению коллокаций рассмотрены в [15]. В [8] представлен подход к извлечению концептов из текстов медицинской тематики, основанный на использовании сверточных нейронных сетей. Обширный обзор методов и алгоритмов извлечения ключевых слов приведен в [9].

Цель исследования

Целью данной работы является сравнительный анализ мер ассоциации, а также оценка качества (точность, полнота, F_1-мера) извлечения понятий и концептов из корпуса текстов посредством бинарной классификации на основании мер ассоциации.

Изложение основного материала исследования

Меры ассоциации

Ниже рассмотрены наиболее часто встречающиеся в литературе [3, 12, 13, 15] меры ассоциации, используемые для выделения коллокаций:

Мера Дайса [1]:

$$Dice(x,y) = \frac{2f(x,y)}{f(x)+f(y)}, \quad (1)$$

где $f(x,y)$ – частота встречаемости в корпусе текстов упорядоченной пары слов x и y, $f(x)$, $f(y)$ – частоты встречаемости слов x и y соответственно (здесь и далее подразумеваются абсолютные, а не относительные значения частот).

Если слова x и y встречаются исключительно парами вида xy, то

$$f(x) = f(y) = f(x,y), \quad (2)$$

и значение меры будет максимальным: $sup\ Dice(x,y) = 1$. Если слова x и y никогда не встречаются парами вида xy, то значение меры будет минимальным: $inf\ Dice(x,y) = 0$. Даже если в корпусе текстов пара xy встречается лишь один раз, и выполнено условие (2), то значение меры для такой биграммы будет максимальным. Т. о., на вершине рейтинга могут быть пары сильно связанных слов (например, имя и фамилия), но не имеющих ценности в плане отражения смысла текста.

Модифицированная мера Дайса [2]:

$$Dice'(x,y) = log_2\left(\frac{2f(x,y)}{f(x)+f(y)}\right). \quad (3)$$

Пусть, имеется некоторое множество биграмм мощностью M, и $p(x,y) = Dice(x,y)$ - вероятность того, что биграмма xy является искомым объектом (ключевым словом, термином и пр.). В таком случае, имеем систему с M возможными состояниями, а (3) – это частная энтропия соответствующего состояния, взятая со знаком минус. Область значений: $ran\ Dice'(x,y) = (-\infty;\ 0]$.

Коэффициент взаимной информации (mutualinformation, MI) [1–2]:

$$MI(x,y) = log_2\left(\frac{f(x,y)*N}{f(x)*f(y)}\right) = log_2\left(\frac{f(x,y)}{f(x)*f(y)}\right) + log_2(N), \quad (4)$$

где N – общее количество слов в корпусе.

Очевидно, что умножение на $N = const$ не имеет смысла, и дает только смещение значения меры.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

Область значений функции $MI(x,y)$: $ran\ MI(x,y) = (-\infty;\ -log_2(f(x,y))]$ (без умножения на N в числителе).

Поточечная взаимная информация (point wise mutual information, PMI) [16]:

$$PMI(x,y) = log_2\left(\frac{p(x,y)}{p(x)*p(y)}\right), \quad (5)$$

где $p(x) = f(x)/N$, $p(y) = f(y)/N$, а $p(x,y)$ приравнивают к $f(x,y)$ [17].

Действительно, рейтинг биграмм не зависит от делителя частоты $f(x,y)$ (единица, либо иной), однако называть частоту вероятностью, всё же, некорректно. Кроме того:

$$PMI(x,y) = log_2\left(\frac{f(x,y)}{\frac{f(x)}{N}*\frac{f(y)}{N}}\right) = log_2\left(\frac{f(x,y)}{f(x)*f(y)}N^2\right) = log_2\left(\frac{f(x,y)}{f(x)*f(y)}\right) + log_2(N^2). \quad (6)$$

Откуда следует, что рейтинги, сформированные на основании (4), и на основании (5-6), не будут отличаться.

Нормализованная поточечная информация (normalized point wise mutual information, NPMI) [16]:

$$NPMI(x,y) = \frac{log_2\left(\frac{p(x,y)}{p(x)*p(y)}\right)}{-log_2(p(x,y))}, \quad (7)$$

При условии расчета $p(x,y)$ в соответствии с выражением $f(x,y)/N$, получим $ran\ NPMI(x,y) = (-1;1]$, откуда и название «нормализованная».

$sup\ NPMI(x,y) = 1$ следует из выражения для области значений $MI(x,y)$, а $inf\ NPMI(x,y) = -1$ легко находится как предел:

$$\lim_{p(x,y)\to 0}\frac{log_2\left(\frac{p(x,y)}{p(x)*p(y)}\right)}{-log_2(p(x,y))} = \lim_{p(x,y)\to 0}\frac{[log_2(p(x,y)) - log_2(p(x)) - log_2(p(y))]'}{[-log_2(p(x,y))]'} = \frac{\frac{1}{p(x,y)ln2}}{-\frac{1}{p(x,y)ln2}} = -1. \quad (8)$$

Мера Миколова [18]:

$$m - score(x,y) = \frac{f(x,y)-\delta}{f(x)*f(y)}, \quad (9)$$

где $\delta \geq 0$ – некое пороговое целочисленное значение, позволяющее отсеивать биграммы с частотами $f(x,y) \leq \delta$.

В случае $\delta = 0$, получим $log_2(m - score(x,y)) = MI(x,y)$.

Мера Жаккара (Жаккарда, Джаккарда) [19]:

$$K_J(A,B) = \frac{|A\cap B|}{|A\cup B|}, \quad (10)$$

где A и B – два некоторых множества.

Чем больше общих элементов содержат множества, тем выше значение меры сходства $K_J(A,B)$. $ran\ K_J(A,B) = [0;1]$.

Если спроецировать (10) на задачу выделения коллокаций, то можно получить меру вида (11):

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

$$K_J'(x,y) = \frac{f(x)*f(y)}{f(x)+f(y)}. \quad (11)$$

В работе [20] также приведен вариант (12):

$$K_J'(x,y) = \frac{f(x,y)}{f(x)+f(y)-f(x,y)}. \quad (12)$$

Несложно заметить, что логика расчета рассмотренных мер ассоциации весьма схожа. Выражение вида $f(x)+f(y)$ можно интерпретировать как аналог вероятности появления слова x ИЛИ y (при условии несовместности x и y), выражение $f(x)*f(y)$ – как аналог вероятности совместного появления двух независимых событий x И y, $f(x)+f(y)-f(x,y)$ – как аналог вероятности $P(x$ ИЛИ $y)$ для совместных событий.

Если частоту встречаемости биграммы $f(x,y)$ рассматривать как случайную величину Z, подчиняющуюся биномиальному закону распределения [12, 15], то получим

$$Z = Bin(n,p), \quad (13)$$

где n – некое верхнее граничное значение для частоты, p – вероятность встречи биграммы xy в корпусе текстов.

Т. о., получаем интервал возможных значений случайной величины $\{0,\dots,n\}$, и соответствующее распределение вероятностей, с которыми Z может принимать значения из заданного интервала.

Если отобрать первые n биграмм из корпуса, то вероятность встречи среди них k биграмм xy будет оцениваться как

$$P(Z=k) = \binom{n}{k} p^k (1-p)^{n-k}. \quad (14)$$

На основании (14) можно построить функцию логарифмического правдоподобия [3, 12, 13, 15, 21]

$$\begin{aligned} LL(a,b,c,d) = {} & a\cdot log(a+1) + b\cdot log(b+1) + c\cdot log(c+1) + d\cdot log(d+1) - \\ & -(a+b)\cdot log(a+b+1) - (a+c)\cdot log(a+c+1) - \\ & -(b+d)\cdot log(b+d+1) - (c+d)\cdot log(c+d+1) + \\ & +(a+b+c+d)\cdot log(a+b+c+d+1), \end{aligned} \quad (15)$$

причем [11–12]

$$LL(a,b,c,d) = \frac{L(H_1)}{L(H_0)} \quad (16)$$

где a – частота заданной пары слов, b – сумма частот пар с той же левой леммой (нормализованной формой слова), c – сумма частот пар с той же правой леммой, d – сумма частот пар, отличных от a, $L(H_1)$ – вероятность гипотезы о наличии статистической связи между словами в биаграмме a, $L(H_0)$ – вероятность гипотезы об отсутствии статистической связи между словами в биграмме a.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

Аппроксимируя дискретную случайную величину Z некой непрерывной случайной величиной X с нормальным законом распределения, для оценки случайности совместного нахождения слов в биграмме, можно использовать критерий Стьюдента t [1–2]:

$$t = \frac{\bar{x}-\mu}{\sqrt{s^2/n}} = \frac{\bar{x}-\mu}{\sqrt{s^2/(N-1)}}, \quad (17)$$

где $\bar{x}$ – выборочное среднее, s^2 – выборочная дисперсия, N – количество слов в корпусе, $n = N - 1$ – размер выборки (общее количество биграмм, а не количество уникальных биграмм), μ - генеральное среднее. Стоит отметить, что, при больших размерах корпуса ($N > 10^5$) различиями между N и n можно пренебречь.

Зная параметры биномиального распределения, можно записать:

$$\bar{x} = np = n\frac{f(x,y)}{n} = f(x,y),$$

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{npq}{n}} = \sqrt{pq} = \sqrt{p(1-p)} = \sqrt{\frac{f(x,y)}{n}(1-\frac{f(x,y)}{n})} \approx \sqrt{\frac{f(x,y)}{n}} = \sqrt{f(x,y)}/\sqrt{n}.$$

Проецируя (17) на задачу выделения коллокаций, $\bar{x}$ именуют наблюдаемой (observed) частотой, а μ - ожидаемой (expected) частотой. Последняя рассчитывается следующим образом (из предположения независимости слов x и y):

$$\mu = \hat{f}(x,y) = n * p(x \,\&\, y) = n(\frac{f(x)}{N}\frac{f(y)}{N}) \approx \frac{f(x)*f(y)}{N}.$$

Учитывая то, что деление на $\sqrt{n}$ знаменателя не влияет на рейтинг биграмм, получаем [2, 12]:

$$t - score(x,y) = \frac{f(x,y)-\frac{f(x)*f(y)}{N}}{\sqrt{f(x,y)}}. \quad (18)$$

$$ran\ t - score(x,y) = (-\infty; \sqrt{f(x,y)}(1-\frac{f(x,y)}{N})].$$

Чем больше значение критерия $t - score(x,y)$, тем меньше вероятность нулевой гипотезы о независимости слов x и y.

Мера C-value [5]:

$$C - value(a) = \begin{cases} log_2(|a|) * f(a), \text{если терм не вложен в другие} \\ log_2(|a|) * \left(f(a) - \frac{1}{|S_a|}\sum_{b \in S_a} f(b)\right), \text{в противном случае'} \end{cases} \quad (19)$$

где a – n-грамма, для которой выполняется расчет меры, $|a|$ – количество слов в n-грамме, $f(a)$ – частота a, S_a – множество n-грамм, в которые входит данная (например, выполняя расчет меры для биграммы, можно оценивать количество ее вхождений в некоторое множество отобранных триграмм), $|S_a|$ – мощность множества S_a, $f(b)$ – частота b– той n-граммы из S_a.

Если не выполнять оценку количества вхождений a в n-граммы с количеством слов больше $|a|$, то $C - value$ вырождается просто в частоту встречаемости, масштабированную на $log_2(|a|)$. Если же n-грамма является составной частью более

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

сложных коллокаций, ее вес понижается. Умножение на $log_2(|a|)$ имеет смысл лишь при расчете меры для n-грамм различной мощности.

Существуют также расширения ряда рассмотренных мер для n-грамм с $n > 2$ [2].

Модель TF*IDF

В случае использования модели TF*IDF, корпус из n документов, содержащий m уникальных слов, представляется матрицей размером $m \times n$, элементами которой являются произведения значений локальной функции $TF(w, d)$ и глобальной функции $IDF(w, d)$, рассчитанных для соответствующих слова w (в общем случае, - n-граммы) и документа d. Произведение $TF(w, d) * IDF(w, d)$ называется также TF*IDF-мерой, определяющей вес слова w в документе d.

Существуют различные подходы к расчету функций $TF(w, d)$ и $IDF(w, d)$, а также нормализации значений TF*IDF-меры в пределах столбца матрицы. Ряд наиболее известных подходов объединен в рамках т. н. системы информационного поиска SMART [22].

Норма вектора, соответствующего слову (биграмме) в матрице TF*IDF, в данной работе использована как дополнительная мера ассоциации. Корпус документов при этом делился на K условных документов примерно равного размера.

Формирование обучающей и тестовой выборок

Для исследований использован корпус из шести книг тематики «программирование на языке Си». Предварительная обработка текстов в себя включала следующие этапы: токенизация (разделение на слова, знаки пунктуации и пр.); удаление небуквенных символов; приведение к нижнему регистру; удаление слов с длиной менее двух букв, а также слов, содержащих латиницу; фильтрация стоп-слов на основании вспомогательного словаря; лемматизация (приведение слов к их словарной форме); частеречная разметка; отбор имен существительных и прилагательных. В результате выполнения перечисленных этапов, были сформированы корпус размером 208689 словоупотреблений и словарь, содержащий 5154 слова. Количество уникальных биграмм в корпусе составило 79316. Стоит отметить, что процесс лемматизации текстов является достаточно ресурсоемким, и на машине с 2 Гб ОЗУ и двухъядерным ЦП (для обработки использовано лишь одно) корпус указанного размера, являющийся крайне малым по общим меркам, обрабатывался порядка 6 мин. Временная сложность алгоритма лемматизации составляет $O(n)$.

На основании полученного корпуса сформирована размеченная выборка размером 508 биграмм, из которых половина идентифицирована как «концпеты и понятия». Полученная выборка разделена на обучающую и тестовую в соотношении 4:1.

Расчет значений порогов фильтрации

Пусть, имеется размеченная выборка биграмм, представленная матрицей $S_{m \times n+2}$, в которой каждая строка соответствует вектору вида $\{class, bg, M_1, \dots, M_n\}$, где bg - биграмма, $class$ – код класса ($c_1 = 1 -$ «концепты», $c_2 = 0 -$ «не концепты»), $M_1, \dots, M_n$ - значения мер ассоциации. Количество биграмм каждого класса одинаково, и равно $m/2$.

Расчет вероятности принадлежности каждой биграммы классу c_1 на основании значения меры M_i выполняется на основании формулы (20):

$$P(bg \in c_1 | M_i = x) = \frac{p_1(x)}{p_1(x)+p_2(x)}, \quad (20)$$

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

где x – вещественное значение меры, p_1 и p_2 – функции плотности вероятности, соответствующие распределению значений меры в классах c_1 и c_2.

Бинаризация значений мер выполняется в соответствии с выражением (21):

$$M_i^b(bg|M_i = x) = \begin{cases} 1, if\ P(bg \in c_1|M_i = x) > T_i; \\ 0, if\ P(bg \in c_1|M_i = x) \leq T_i; \end{cases} \quad (21)$$

где $M_i^b(bg|M_i = x)$ – бинаризованное значение меры M_i для биграммы bg, T_i – порог бинаризации для меры M_i.

В результате бинаризации будут получены вектор-столбцы X_1^i и X_2^i, содержащие двоичные оценки меры для биграмм из классов c_1 и c_2 соответственно, где i – номер меры.

На основании полученных данных может быть рассчитана величина r_i, являющаяся отношением сумм элементов в X_1^i и X_2^i:

$$r_i = \frac{sum(X_1^i)}{sum(X_2^i)}. \quad (22)$$

Чем больше r_i, тем больше биграмм со значением $M_i^b = 1$ будет отнесено к классу c_1, в сравнении с количеством биграмм, отнесенных к c_2. Определение порогового значения T_i выполняется путем максимизации функционала $r_i(T_i)$.

Определение функций распределения вероятностей

В случае затруднительности определения закона распределения случайной величины X_{ic}, можно прибегнуть к ядерной оценке плотности (Kernel Density Estimation, KDE), также именуемой методом окна Парзена-Розенблатта) [23], аппроксимировав гистограмму эмпирического распределения X_{ic} функционалом вида (23), с последующим нормированием:

$$\hat{p}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x-x_i}{h}), \quad (23)$$

где x – значение случайной величины (СВ), для которой рассчитывается значение функции плотности вероятности, x_i – i-ое значение СВ из выборки размером N элементов, $h > 0$ – т.н. ширина полосы, K – функция, именуемая взвешенным ядром (существуют различные варианты данной функции [23]).

При слишком малом значении h, $\hat{p}(x)$ будет содержать много случайных выбросов, при слишком большом h, - будет чрезмерно сглажена.

Результаты и обсуждение

1. Оценка качества выделения понятий и концептов на основании меры TF*IDF.

С целью исследования влияния гиперпараметров модели TF*IDF на качество извлечения понятий и концептов (ПКТ) из корпуса текстов, была произведена оценка на основании первых 50-ти биграмм рейтинга, формируемого для каждого набора параметров (табл. 1). В качестве гиперпараметров рассматривались различные SMART-функции fun [22], а также количество условных документов N, на которое делится корпус D.

На основании табл. 1 можно заключить, что субоптимальными для решения задачи выделения ПКТ наборами мнемоник SMART являются тройки вида $(p_1p_2p_3) \in \{n, a\} \times \{t\} \times \{n, c, u\}$. Также видно, что N должно иметь порядок 3-5.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

2. Корреляционный анализ взаимосвязи между мерами ассоциации.

Выполним кодирование мер: *M*1 – мера Дайса, *M*2 – модифицированная мера Дайса, *M*3 – коэффициент взаимной информации (MI), *M*4 – нормализованная поточечная информация (NPMI), *M*5 – мера Миколова, *M*6 – критерий Стьюдента, *M*7 – логарифмическая функция правдоподобия, *M*8 – мера Жаккара, *M*9 – C-value, *M*10 – частота встречаемости, *M*11 – TF*IDF. Результат визуализации корреляционной матрицы для перечисленных мер представлен на рис. 1:

Таблица 1

Зависимость точности классификации от гиперпараметров модели TF*IDF

TF(w,d)		IDF(w,d)		Normalization		$N = \|D\|$	
fun	P	fun	P	fun	P	N	P
b	0.66	n	0.78	n	0.80	2	0.74
n	0.80	f	0.76	c	0.80	3	0.80
a	0.80	t	0.80	u	0.80	5	0.80
l	0.70	p	0.68	b	0.78	10	0.76
L	0.66						
d	0.70						

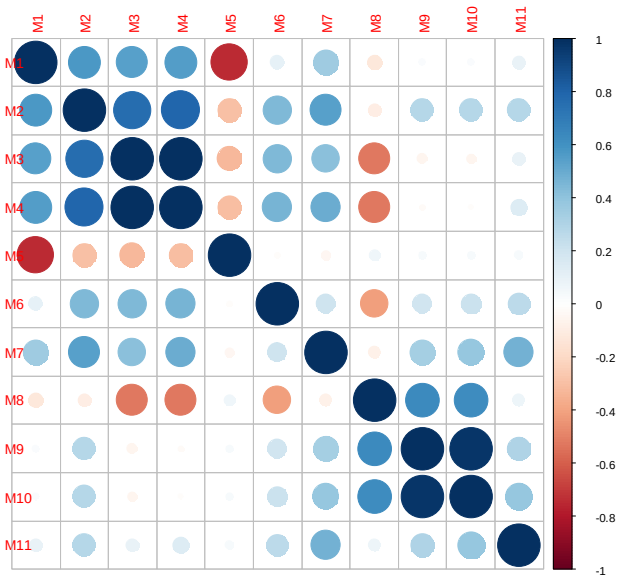

Рис. 1. Визуализация корреляции между мерами ассоциации.

Из рис. 1 видно, что наиболее сильно связь выражена между мерами, рассчитываемыми схожим образом. Можно выделить следующие условные кластеры: *M*1–*M*4 (см. (1), (3), (4), (7)) и *M*8–*M*10 (см. (11), (19)). Несмотря на то, что частота встречаемости биграммы не используется при расчете меры Жаккара (M8, расчет на основании (11)), наблюдается выраженная корреляция с *M*9–*M*10. Функция логарифмического правдоподобия (*M*7) приблизительно в равной мере коррелирует с элементами из кластеров *M*1–*M*4 и *M*9–*M*10, а также с TF*IDF (*M*11), что говорит о равном учете как частоты биграммы, так и силы связи слов в ней. Несмотря на отличия в логике расчета критерия Стьюдента (*M*6, (18)), видна выраженная корреляция с *M*1–*M*4. Мера Миколова (*M*5) практически не коррелирует с *M*6–*M*11, и находится в выраженной антикорреляции по отношению к *M*1–*M*4.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

Расчеты производились на основании размеченной выборки размером 508 биграмм.

3. Определение функций плотности вероятности.

Обозначим через $X_{ij} = \{M_i | c_j\}, i = \overline{1,11}, j = \overline{1,2}$ случайную величину (СВ), соответствующую значениям меры M_i при условии принадлежности соответствующей биграммы классу c_j. Т. о., имея размеченную выборку биграмм, получим 22 выборки соответствующих случайных величин X_{ij}.

По итогам χ^2-теста, нулевая гипотеза о нормальном характере распределения X_{ij} при уровне значимости $\alpha = 0.1\%$ подтвердилась лишь для четырех СВ из 22-х ($X_{2,1}$, $X_{3,1}$, $X_{4,1}$, $X_{2,2}$). При том же уровне значимости, было получено подтверждение нулевых гипотез для ряда иных распределений и СВ: бета-распределение: $X_{1,1}$, $X_{8,1}$; гамма-распределение: $X_{2,2}$, $X_{8,1}$; распределение Вейбула: $X_{2,2}$, $X_{8,1}$; экспоненциальное: $X_{8,1}$. Полученные результаты свидетельствуют о затруднительности определения характера распределения случайных величин $X_{ij}, i = \overline{1,11}, j = \overline{1,2}$, что дает основание для использования метода окна Парзена-Розенблатта (KDE). Пример результата применения указанного метода приведен на рис. 2, где $p_1(x)$ – эмпирическая функция плотности вероятности (ФПВ), $p_2(x)$ – аппроксимация посредством ФПВ нормального распределения, $p_3(x)$ – аппроксимация посредством KDE ($h = 0.02$), с использованием ядра Лапласа [23]. Видно, что $p_3(x)$ меньше отличается от $p_1(x)$, нежели $p_2(x)$, и является более гладкой, нежели $p_1(x)$.

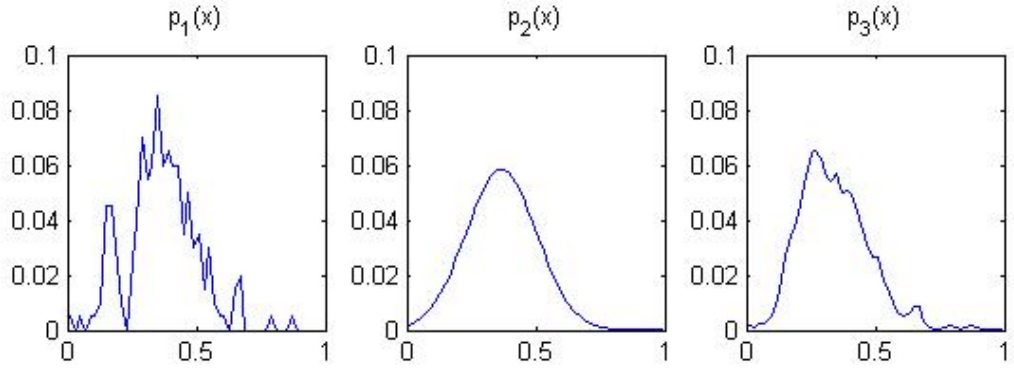

Рис. 2. Аппроксимация эмпирической функции плотности вероятности для $X_{3,1}$.

4. Результаты пороговой фильтрации.

Расчет пороговых значений фильтрации выполнялся в соответствии с выше изложенным алгоритмом на основании функций плотности вероятности $p_i(x), i = \overline{1,3}$ (в случаях $p_1(x)$ и $p_3(x)$ использована интерполяция по ближайшему соседу). Для максимизации значения (22) применен метод золотого сечения. Результаты классификации на основании одного признака приведены в табл. 2 (определение параметров распределений выполнено на основании обучающей выборки, классификация – на основании тестовой). Важно отметить, что используемая классификация базируется не на расчете порогового значения меры ассоциации, но на расчете порогового значения вероятности (20), на основании которого выполняется бинаризация значений меры.

Из табл. 2 видно, что использование $p_3(x)$ вместо $p_2(x)$ позволяет в ряде случаев существенно (более чем в 15 раз для M_2) повысить качество классификации, что особенно выражено для мер M_1, M_2, M_5 и M_8. Возможным является также ухудшение качества классификации (M_3, M_4, M_6), особенно выраженное для меры M_6,

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

однако уменьшение ширины полосы до $h = 0.001$ позволяет достичь классификации на основании M_6 с $F_1 = 1.0$, но двое ухудшает качество классификации по M_7. Данное обстоятельство указывает на целесообразность индивидуального подбора ширины окна h для каждой меры.

Таблица 2

Результаты бинарной классификации на основании одного признака

ФПВ	Оценка	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
$p_1(x)$	P	-	0,17	0,93	0,78	0,50	1,00	1,00	0,85	-	-	-
	F_1	-	0,03	0,63	0,64	0,67	0,11	1,00	0,64	-	-	-
$p_2(x)$	P	0,50	0,33	0,96	0,95	0,50	1,00	1,00	0,95	1,00	1,00	0,75
	F_1	0,67	0,04	0,98	0,97	0,67	0,41	1,00	0,53	0,04	0,04	0,10
$p_3(x)$, $h = 0.01$	P	0,77	0,76	0,92	0,87	0,83	1,00	1,00	0,89	1,00	1,00	0,67
	F_1	0,79	0,62	0,90	0,93	0,91	0,04	1,00	0,71	0,04	0,04	0,13

P – точность, F_1- F_1-мера.

Прочерки в оценках для $p_1(x)$ обусловлены отношением по итогам классификации всех объектов к одному классу («не концепты»). Данные табл. 2 также указывают на высокую селективную способность меры M_7 (функция логарифмического правдоподобия) в отношении концептов и понятий.

Для сравнения, в табл. 3 приведены оценки точности мер ассоциации на основании первых 50-ти биграмм, отобранных из соответствующих рейтингов. Разместив меры в порядке убывания соответствующей точности, получим ряд M_8, M_{10}, M_9, M_{11}, M_6, M_5, M_7, M_1, M_2, M_4, M_3, в котором мера M_7 находится на седьмой позиции. Это объясняется тем, что в данном случае отбор осуществлялся исключительно на основании ранжирования значений меры ассоциации, без учета характера распределений.

Эмпирические и аппроксимированные функции плотности вероятности для случайных величин $X_{7,1}$ и $X_{7,2}$, соответствующих мере M_7, показаны на рис. 3 (эмпирические показаны в виде гистограмм, аппроксимированные – сплошной линией).

Таблица 3

Оценка точности мер ассоциации на основании первых 50-ти биграмм рейтинга

Мера	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
Точность P	0.14	0.14	0.04	0.08	0.72	0.74	0.66	0.92	0.80	0.82	0.76

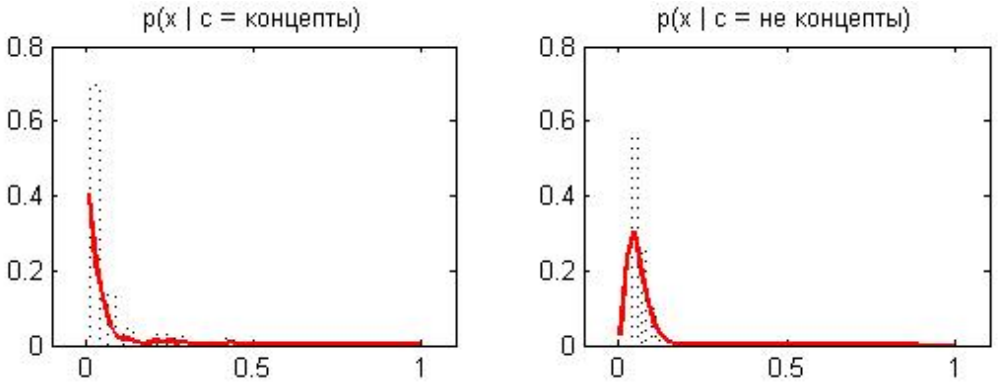

Рис. 3. Эмпирические и аппроксимированные распределения СВ $X_{7,1}$ и $X_{7,2}$.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

Рассмотрим также распределение значений мер между классами, выполнив суммирование значений каждой меры по отдельным классам и последующую нормализацию полученных сумм (рис. 4). Полученная гистограмма согласуется с результатами табл. 2-3 и заключением о том, что различающим признаком является не столько само значение меры, сколько характер ее статистического распределения.

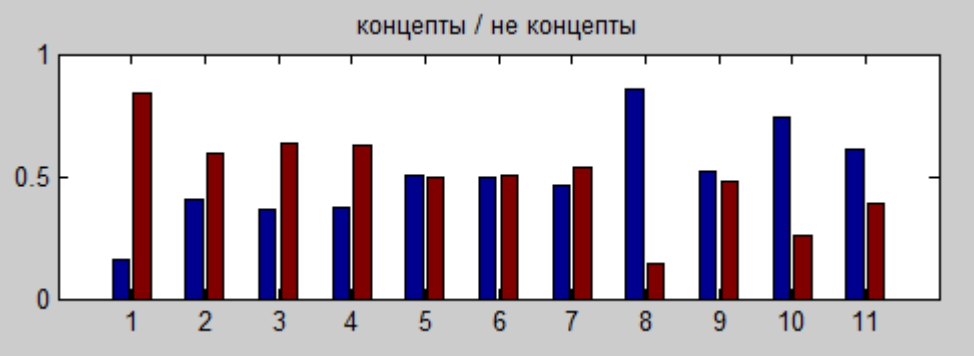

Рис. 4. Распределение значений мер между классами.

Уместным является вопрос о том, имеется ли смысл в фильтрации по порогу вероятности (20-21), или же достаточно выполнять бинаризацию на основании проверки неравенства

$$P(bg \in c_1 | M_i = x) > P(bg \in c_2 | M_i = x). \quad (24)$$

Результаты в табл. 4 показывают, что бинаризация на основании (24) понижает качество классификации для мер M_1, M_3, M_4, M_5 и M_7, и повышает для прочих семи. Однако, понижение F_1-меры для M_7 с 1.00 до 0.70 (например) является более критичным, нежели ее увеличение с 0.04 до 0.21 для M_9 (например).

Таблица 4

Результаты классификации (значения F_1-меры) при разных подходах к бинаризации значений мер ассоциации

Критерий	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
на основании порога	0,79	0,62	0,90	0,93	0,91	0,04	1,00	0,71	0,04	0,04	0,13
на основании соотношения вероятностей	0,68	0,69	0,69	0,70	0,67	0,26	0,70	0,91	0,21	0,21	0,31

Выводы

1. В контексте извлечения коллокаций из текста, первостепенным является не столько значение используемой меры ассоциации, сколько характер ее распределения (см. табл. 2 и рис. 4).

2. На основании корреляционного анализа, рассмотренные меры можно разделить на два кластера: ориентированные на силу связи в коллокации (*M*1–*M*4) и ориентированные на частоту встречаемости коллокации (*M*9–*M*11). Меры *M*6–*M*7 примерно в равной степени учитывают оба данных фактора (хотя, в случае пороговой фильтрации на основании (20–21), мера *M*6 показывает низкую точность

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

классификации). Мера *M*8 (11) явно не учитывает частоту встречаемости коллокации, однако находится в том же кластере, что и *M*9–*M*11. Мера *M*5 находится в антикорреляции с *M*1–*M*4, и не коррелирует с прочими, однако обладает достаточно высокой селективной способностью (табл. 2–3).

3. Бинаризация значений меры в соответствии с (20–21) обеспечивает б*о*льшую точность классификации на основании мер M_1, M_3, M_4, M_5 и M_7, нежели в соответствии с (24). Последний вариант повышает точность классификации для прочих мер ассоциации, однако получаемый уровень точности нельзя считать приемлемым.

4. Целесообразным является индивидуальный подбор ширины полосы h в методе KDE для каждой меры ассоциации. Если при $h = 0.01$ получено $F_1 = 1.0$ для меры *M*7, то при $h = 0.001$ – для меры *M*6.

5. Более эффективной является фильтрация не на основании значения самой меры, но на основании вероятности (20). Первый вариант позволяет получить «топ» коллокаций (табл. 3), но не выполнить классификацию (табл. 2).

6. Мера *M*7 существенно отличается от прочих рассмотренных учетом контекста коллокации. Однако, существуют подходы, позволяющие учитывать контекст также при использовании иных мер [2].

Перспективы дальнейших исследований: устранение переносов и ошибок на этапе предварительной обработки текстов, выбор требуемого варианта леммы на основании контекста слова, сопряжение коллокатов (род, число, падеж), распараллеливание процесса лемматизации.

Список использованной литературы

1. Баранов В. А. Опыт создания модуля *n*-грамм системы «Манускрипт» и оценки эффективности его использования для поиска коллокаций в корпусе М. В. Ломоносова. *Интеллектуальные системы в производстве*. 2016. №4. С. 124–131.
2. Большакова Е. И., Клышинский Э. С., Ландэ Д. В. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ, 2011. 272 с.
3. Lyse G. I., Andersen G. Collocations and statistical analysis of *n*-grams: Multiword expressions in newspaper text. *Exploring Newspaper Language*. Amsterdam, New York: John Benjamins, 2012. P. 79–109.
4. Виноградова Н. В., Иванов В. К. Современные методы автоматизированного извлечения ключевых слов из текста. *Информационные ресурсы России*. 2016. № 4. С. 13–18.
5. Lossio-Ventura J. A., Jonquet C., Roche M. et al. Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction. Proceedings of the *LBM: Languages in Biology and Medicine: 5th International Symposium*, (Japan, Tokyo, December 12-13, 2013). Tokyo, 2013, pp. 1–6.
6. Evert S., Krenn B. Using Small Random Samples for the Manual Evaluation of Statistical Association Measures. *Computer Speech & Language*. 2005. Vol. 19. P. 450–466.
7. Wei C.-H., Allot A., Leaman R. & Lu Z. PubTator central: Automated Concept Annotation for Biomedical Full Text Articles. *Nucleic Acids Research*. 2019. Vol. 47. P. 587–593.
8. Gehrmann S., Derenoncourt F., Li Y. et al. Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives. *PLoSOne*. 2018. Vol. 13. Issue 2. P. 1–19.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

9. Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов. *Новые информационные технологии в автоматизированных системах*. 2016. №.19. С. 85–93.
10. Мозжерина Е. С. Автоматическое построение онтологии по коллекции текстовых документов. *Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 13-й Всероссийской научной конференции.* (Россия, Воронеж, 19-22 октября 2011 г.) Воронеж: Издательство Воронежского государственного университета, 2011. С. 293–298.
11. Christopher D. M., Hinrich S. Foundations of Statistical Natural Language Processing. Cambridge, Mass.: MIT Press, 1999. P. 178–183.
12. Thanopoulos A., Fakotakis N., Kokkinakis G. Comparative Evaluation of Collocation Extraction Metrics. Proceedings of the *Third International Conference on Language Resources and Evaluation (LREC'02)*. (Canary Islands – Spain, Las Palmas, May, 2002). Luxembourg: European Language Resources Association (ELRA), 2002. P. 620–625.
13. Kolesnikova O. Survey of Word Co-occurrence Measures for Collocation Detection. *Computacion y Sistemas*. 2016. Vol. 20. № 3. P. 327–344. DOI: 10.13053/CyS-20-3-2456.
14. Hoang H. H., Kim S. N., Kan M.-Y. A Re-examination of Lexical Association Measures. Proceedings of the *Identification, Interpretation, Disambiguation and Applications: Workshop on Multiword Expressions (MWE 2009)*. (Singapore, Singapore, August, 2009). Stroudsburg: Association for Computational Linguistics, 2009. P. 31–39.
15. Pazienza M. T., Pennacchiotti M., Zanzotto F. B. Terminology extraction: an analysis of linguistic and statistical approaches. *Studies in Fuzziness and Soft Computing*. 2006. Vol. 185. P. 255–279.
16. Bouma G. Normalized (Pointwise) Mutual Information in Collocation Extraction. Proceedings of the *Biennial GSCL Conference*. 2009. P. 1–11.
17. Calculate Pointwise Mutual Information (PMI)/ URL: https://polmine.github.io/polmineR/reference/pmi.html.
18. Mikolov T., Sutskever I., Chen K. et al. Distributed Representations of Words and Phrases and their Compositionality. Proceedings of the *Neural Information Processing Systems 2013: conference*. (USA, Lake Tahoe, 2013). In *Advances in Neural Information Processing Systems*. 2013. 9 p.
19. Когай В. Н., Пак В. С. Алгоритмическая модель компьютерной системы выделения ключевых слов из текста на базе онтологий. *Проблемы современной науки и образования*. 2016. № 16(58). С. 33–40.
20. Damani O. Improving Pointwise Mutual Information (PMI) by Incorporating Significant Co-occurrence. Proceedings of the *Seventeenth Conference on Computational Natural Language Learning*. (Bulgaria, Sofia, August 8-9, 2013). Madison: Omnipress, 2013. P. 20–28.
21. Андреев И. А., Башаев В. А., Клейн В. В. и др. Комбинирование статистического и лингвистического методов для извлечения двухсловных терминов из текста. *Автоматизация процессов управления*. 2013. № 4. С. 61–70.
22. SMART Information Retrieval System. URL: https://en.wikipedia.org/wiki/SMART_Information_Retrieval_System.
23. Поршнев С. В., Копосов А. С. Использование аппроксимации Розенблатта-Парзена для восстановления функции распределения непрерывной случайной величины с ограниченным одномодальным законом распределения. *Научный журнал КубГАУ*. 2013. № 92. С. 1–14.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

References

1. Baranov, V. A. (2016). Opyit sozdaniya modulya n-gramm sistemyi «Manuskript» i otsenki effektivnosti ego ispolzovaniya dlya poiska kollokatsiy v korpuse M. V. Lomonosova. *Intellektualnyie sistemyi v proizvodstve*. **4**, 124–131.
2. Bolshakova, E.I., Klyishinskiy, E.S., & Lande, D. V. i dr. (2011). Avtomaticheskaya obrabotka tekstov na estestvennom yazyike i kompyuternaya lingvistika. M.: MIEM…
3. Lyse, G. I. & Andersen, G. (2012). Collocations and statistical analysis of *n*-grams: Multiword expressions in newspaper text. *Exploring Newspaper Language*. Amsterdam, New York: John Benjamins, pp. 79–109.
4. Vinogradova, N. V., & Ivanov, V. K. (2016). Sovremennyie metodyi avtomatizirovannogo izvlecheniya klyuchevyih slov iz teksta. *Informatsionnyie resursyi Rossii*. **4**, 13–18.
5. Lossio-Ventura, J. A., Jonquet, C., & Roche, M. et al. (2013). Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction. Proceedings of the *LBM: Languages in Biology and Medicine: 5th International Symposium*, (Japan, Tokyo, December 12-13, 2013). Tokyo, pp. 1–6.
6. Evert, S., & Krenn, B. (2005). Using Small Random Samples for the Manual Evaluation of Statistical Association Measures. *Computer Speech & Language*. **19**, 450–466.
7. Wei, C.-H., Allot, A., Leaman, R. & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*. **47**, 587–593.
8. Gehrmann, S., Derenoncourt, F., & Li, Y. et al. (2018). Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives. *PLoSOne*. **13**, 2, 1–19.
9. Vanyushkin, A. S., & Graschenko, L. A. (2016). Metodyi i algoritmyi izvlecheniya klyuchevyih slov. *Novyie informatsionnyie tehnologii v avtomatizirovannyih sistemah*. **19**, 85–93.
10. Mozzherina, E. S. (2011). Avtomaticheskoe postroenie ontologii po kollektsii tekstovyih dokumentov. Proceedings of the *Elektronnyie biblioteki: perspektivnyie metodyi i tehnologii, elektronnyie kollektsii: Trudyi 13-y Vserossiyskoy nauchnoy konferentsii*. (Rossia, Voronezh, October 19-22, 2011). Voronezh: Izdatelstvo Voronezhskogo gosudarstvennogo universiteta, pp. 293–298.
11. Christopher, D. M., Hinrich, S. (1999). Foundations of Statistical Natural Language Processing. MIT Press: Cambridge, Mass., pp. 178–183.
12. Thanopoulos, A., Fakotakis, N., & Kokkinakis, G. (2002). Comparative Evaluation of Collocation Extraction Metrics. Proceedings of the *Third International Conference on Language Resources and Evaluation* (*LREC'02*). (Canary Islands – Spain, Las Palmas, May, 2002). Luxembourg: European Language Resources Association (ELRA), pp. 620–625.
13. Kolesnikova, O. (2016). Survey of Word Co-occurrence Measures for Collocation Detection. *Computacion y Sistemas*. **20**, 327–344. DOI: 10.13053/CyS-20-3-2456.
14. Hoang, H. H., Kim, S. N., & Kan, M.-Y. (2009). A Re-examination of Lexical Association Measures. Proceedings of the *Identification, Interpretation, Disambiguation and Applications: Workshop on Multiword Expressions* (*MWE 2009*). (Singapore, Singapore, August, 2009). Stroudsburg: Association for Computational Linguistics, pp. 31–39.
15. Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. B. (2006). Terminology extraction: an analysis of linguistic and statistical approaches. *Studies in Fuzziness and Soft Computing*. **185**, 255–279.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7

16. Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. Proceedings of the *Biennial GSCL Conference*, pp. 1–11.
17. Calculate Pointwise Mutual Information (PMI). Retrieved from: https://polmine.github.io/ polmineR/reference/pmi.html.
18. Mikolov, T., Sutskever, I., & Chen, K. et al. (2013). Distributed Representations of Words and Phrases and their Compositionality. Proceedings of the *Neural Information Processing Systems 2013: conference.* (USA, Lake Tahoe, 2013). In *Advances in Neural Information Processing Systems*. 9 p.
19. Kogay, V. N., & Pak, V. S. (2016). Algoritmicheskaya model kompyuternoy sistemyi vyideleniya klyuchevyih slov iz teksta na baze ontologiy. *Problemyi sovremennoy nauki i obrazovaniya*. **16** (58), 33–40.
20. Damani, O. (2013). Improving Pointwise Mutual Information (PMI) by Incorporating Significant Co-occurrence. Proceedings of the *Seventeenth Conference on Computational Natural Language Learning*. (Bulgaria, Sofia, August 8-9, 2013). Madison: Omnipress, pp. 20–28.
21. Andreev, I. A., Bashaev, V. A., & Kleyn, V. V. i dr. (2013) Kombinirovanie statisticheskogo i lingvisticheskogo metodov dlya izvlecheniya dvuhslovnyih terminov iz teksta. *Avtomatizatsiya protsessov upravleniya*. **4**, 61–70.
22. SMART Information Retrieval System. Retrieved from: https://en.wikipedia.org/wiki/SMART_Information_Retrieval_System.
23. Porshnev, S. V., & Koposov, A. S. (2013). Ispolzovanie approksimatsii Rozenblatta-Parzena dlya vosstanovleniya funktsii raspredeleniya nepreryivnoy sluchaynoy velichinyi s ogranichennyim odnomodalnyim zakonom raspredeleniya. *Nauchnyiy zhurnal KubGAU*. **92**, 1–14.

Гайдук Кирилл Сергеевич – аспирант кафедры компьютерной инженерии Донецкого национального технического университета, e-mail: kyrylo.haiduk@donntu.edu.ua, ORCID:0000-0002-8040-9062.

Шевченко Ольга Георгиевна – старший преподаватель кафедры компьютерной инженерии Донецкого национального технического университета, e-mail: olha.shevchenko@donntu.edu.ua, ORCID:0000-0002-1056-2571.

Святный Владимир Андреевич – д.т.н., профессор, профессор кафедры компьютерной инженерии Донецкого национального технического университета, e-mail: volodymyr.svyatnyy@donntu.edu.ua, ORCID:0000-0003-4550-3616.

https://doi.org/10.32782/KNTU2618-0340/2020.3.2-2.7