ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 68 за 2020 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Белова К.М., Судаков В.А.

Исследование эффективности методов оценки релевантности текстов

Рекомендуемая форма библиографической ссылки: Белова К.М., Судаков В.А. Исследование эффективности методов оценки релевантности текстов // Препринты ИПМ им. М.В.Келдыша. 2020. № 68. 16 с. http://doi.org/10.20948/prepr-2020-68
URL: http://library.keldysh.ru/preprint.asp?id=2020-68

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

К.М. Белова, В.А. Судаков

Исследование эффективности методов оценки релевантности текстов

Москва — 2020

Белова К.М., Судаков В.А.

Исследование эффективности методов оценки релевантности текстов

На примере задачи сопоставления аннотаций статей и описаний планируемых научно-исследовательских работ анализируются различные методы оценки релевантности текстов и определяется их эффективность. Для получения рациональной оценки релевантности предложена процедура комплексирования результирующих оценок. Разработано программное обеспечение, реализующее предложенный подход. Апробация предложенной процедуры проведена на примере извещений о грантах и государственных закупках и аннотаций научных статей, размещенных в российском индексе научного цитирования.

Ключевые слова: релевантность, косинусное сходство, коэффициент Танимото, Okapi BM25, недоминируемые альтернативы

Kristina Mikhailovna Belova, Vladimir Anatolyevich Sudakov

Effectiveness of methods for assessing the text relevance

Using the example of the task of comparing the abstracts of articles and descriptions of planned research projects, various methods for assessing the relevance of texts are analyzed and their effectiveness is determined. To obtain a rational assessment of relevance, a procedure for integrating the resulting estimates is proposed. Software has been developed that implements the proposed approach. Approbation of the proposed procedure was carried out on the example of notices of grants and public procurement and abstracts of scientific articles posted in the Russian Science Citation Index.

Key words: relevance, cosine similarity, Tanimoto index, Okapi BM25, undominated alternatives

Работа выполнена при поддержке внутреннего гранта РЭУ им. Г.В. Плеханова.

Введение

На сегодняшний день существует проблема, связанная с трудностью поиска необходимой информации о потенциальных исполнителях сложных наукоемких работ. При наличии больших объемов данных сложно вручную найти и подобрать те сведения, которые максимально соответствовали бы критериям поиска. Для решения столь важной проблемы предлагается исследовать возможность использования различных методов оценки релевантности текстов. Данные методы способствуют осуществлению рационального отбора требуемой информации.

Аналогичные задачи поиска соответствия исполнителей и работ в инновационных отраслях экономики встречаются достаточно часто [1]. В данном исследовании аннотации научных трудов должны соответствовать указанному запросу, например, определенной тематике научно-исследовательской работы (НИР). Существование большого числа научных статей требует обеспечения максимальной эффектности и качественного поиска.

В настоящее время существует множество методов, которые используются для оценки релевантности информации. Одним из них является метод, основанный на использовании функции ранжирования Okapi BM25. Данная функция позволяет упорядочить документы по их релевантности поисковому запросу. Она является одной из самых распространенных функций, которые применяются в различных программных средствах, используемых для поиска информации (в том числе и в коллекциях научно-технических текстов) [2]. Функции, подобные Okapi BM25, основываются на вычислении весов слов и степенью релевантности информации, что позволяет точнее установить взаимосвязь между документами и запросом. Кроме того, Okapi BM25 позволяет учитывать обратную связь по релевантности.

Другим методом, который рассматривается в данной работе, является метод, основанный на использовании коэффициента Танимото (расширенный коэффициент Жаккара). Коэффициента Танимото позволяет описать степень схожести двух множеств, которые могут представлять собой совокупности объектов из различных сфер. Поэтому данная мера схожести применяется во многих направлениях (информатика, экология, молекулярная биология и др.). Этот коэффициент также позволяет оценить степень релевантности текстов, где элементами сравниваемых множеств являются слова, а в некоторых работах – предложения [3]. Кроме того, коэффициента Танимото позволяет спроектировать программный продукт, дающий рекомендации, которые основаны на заранее известных критериях [4].

Третий метод оценки релевантности текста, который изучается в настоящей работе, основывается на косинусном сходстве. Косинусное сходство относится к мерам схожести, применяемым к векторным моделям, которые представляют коллекции документов в виде векторов, относящихся к единому

векторному пространству. Косинусное сходство позволяет провести эффективную оценку релевантности, так как учитывает только ненулевые веса слов в документах и не зависит от длины документа. Также использование данной меры схожести позволяет решать задачи классификации по заданным критериям (в том числе и для коллекций научных и экономических работ) [5].

Все перечисленные методы обеспечивают оптимальный релевантный поиск текстов, упрощающий аналогичный процесс, производимый вручную. Каждый метод имеет свои преимущества и недостатки, которые рассматриваются в дальнейшем с целью определения эффективности работы каждого из них.

Аннотации научных работ и тематики НИР, используемые в настоящей работе для осуществления релевантного поиска, необходимо подготовить для анализа и последующей работы. Для этого требуется:

- Перевести текст на русский язык, что осуществляется посредством использования Google Translate API. В данной работе перевод производится только для документов, которые состоят из латинских символов;
- Преобразовать текст в единый регистр для обеспечения грамотного сравнения слов без разделения на прописные и заглавные буквы. В данной работе текст приводится в нижний регистр;
- Произвести лемматизацию (процесс приведения словоформы к лемме – её нормальной форме). Лемматизация позволяет оставить только смысловую составляющую текста, убрав из него грамматическую информацию (числа, падежи, род, времена и т.д.), которая препятствует качественной оценке релевантности;
- Удалить знаки препинания и различные стоп-слова (предлоги, междометия, частицы и др.), которые не несут смысловой нагрузки, а лишь добавляют много шума, препятствую точности определения релевантности.

Все эти действия необходимо производить с любым набором данных, чтобы убедиться в их качественном анализе и обеспечить эффективную и безопасную работу методов.

Метод с использованием Okapi BM25

Okapi BM25 является функцией ранжирования, которая относится к TF-IDF-подобным функциям. В таких функциях TF (term frequency) отвечает за частотность слова в документе (насколько часто слово встречается в нем) и позволяет исключить влияние длины документа на значимость этого слов. IDF (inverse document frequency) представляет собой обратную частоту слова в документе, которая помогает перераспределить вес значимости (редкие слова имеют значимость выше, чем часто употребляемые). Данные факторы позволяют использовать функцию Okapi BM25 для определения релевантности

документов определенному запросу. Она имеет множество модификаций, которые используются в контексте определённых задач.

Функция Okapi BM25 состоит из нескольких других функций, которые содержат различные параметры и компоненты. Оценка релевантности документа D запросу Q определяется Okapi BM25 по формуле:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i)\, TF(q_i),$$
$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i)\, \frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{|D|}{avgdl})}, \quad (1)$$

где Q состоит из слов $q_1, \dots, q_n$, $TF(q_i)$ – частотность слова q_i в документе D, $f(q_i, D)$ – частота употребления слова q_i в документе D, $|D|$ – количество слов в документе (длина документа), $avgdl$ – средняя длина документа в коллекции, k_1 и b – свободные коэффициенты (чаще всего их выбирают как 2.0 и 0.75 соответственно), $IDF(q_i)$ – обратная документная частота слова q_i.

$IDF(q_i)$ имеет несколько вариаций расчета. Чаще всего используют формулу:

$$IDF(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}, \quad (2)$$

где N – количество документов в коллекции, $n(q_i)$ – количество документов, содержащих слово q_i.

Недостаток данной формулы заключается в том, что IDF может принимать отрицательные значения. Это происходит в том случае, когда слово встречается в более чем половине документов коллекции. Таким образом, слово, которое часто встречается в коллекции текстов, может некорректно повлиять на итоговую оценку релевантности, за счёт того, что документ, не содержащий данное слово, будет превалировать над документом с ним. Во избежание этого на IDF принято накладывать условие: если IDF меньше некоторого заданного числа ε, то IDF нужно считать равной этому числу ε.

Главное преимуществ функции Okapi BM25 состоит в эффективности ее использования. Она применяется для решения различных задач поиска и его улучшения посредством использования обратной связи релевантности. Также различные модификации и улучшения метода позволяют выбрать наиболее оптимальный вариант решения поставленной задачи [6].

Данный метод имеет также ряд недостатков, которые свойственны многим методам оценки релевантности текстов. Так, рассматриваемая функция ранжирования не учитывает близость слов, их взаимное расположение, а также позицию слова в документе. Помимо этого, данный метод не позволяет абстрагироваться от формы слова (проблема решается путем подготовки исходных данные и их преобразования в требуемый формат с применением лемматизации). Размер документа также может плохо сказываться на

производительности данной функции, так как большое количество слов в рассматриваемом тексте приводит к долгим вычислениям.

Метод с использованием коэффициента Танимото

Коэффициент Танимото (расширенный коэффициент Жаккара) представляет собой меру сходства, которая способна измерить степень схожести двух множеств. Данный коэффициент является частным случаем коэффициента Жаккара, где все элементы множества трактуются одинаково и имеют вес, принимающий двоичное значение (0 или 1). Коэффициент Танимото позволяет оценить меру сходства для векторов с непрерывными значениями (например, вес слова определяется мерой TF-IDF). Если веса принимают двоичные значения, индекс Танимото сводится к индексу Жаккара.

Коэффициент Танимото также используется и для оценки релевантности документов заданному запросу. Это возможно благодаря тому, что строки документов и запроса можно рассматривать в виде множеств, где элементами являются слова.

Существует несколько вариаций расчета коэффициента Танимото, в основе которых заложен принцип, построенный на отношении пересечения множеств к их объединению. Чаще всего для оценки релевантности он определяется по формуле:

$$k = \frac{c}{a+b-c}, \tag{3}$$

где k – коэффициент Танимото, принимающий значение от 0 до 1 (чем ближе значение к 1, тем больше сходство между множествами), a – количество элементов первого множества, b – количество элементов второго множества, c – количество элементов, принадлежащих обоим множествам одновременно.

При оценке релевантности с помощью индекса Танимото никакой из документов коллекции не зависит от другого. Так, рассматриваемый метод выдаёт оценку, основываясь исключительно на тексте запроса и документа без учета других элементов коллекции. В этом заключается главное достоинство коэффициента Танимото. Также рассматриваемый метод достаточно прост в понимании и применении.

Коэффициент Танимото также имеет ряд недостатков. Как и большинство методов оценки релевантности, рассматриваемый коэффициент не учитывает взаимное расположение слов в документе и их позицию в нем. Кроме того, на результат работы метода влияет форма слова, которая учитывает грамматическую информацию текста (проблема также решается путем использования лемматизации). Еще один недостаток индекса Танимото состоит в его зависимости от длины документа, что может существенно уменьшить настоящую оценку релевантности, хотя это не является обоснованным и целесообразным.

Метод с использованием косинусного сходства

Косинусное сходство представляет собой меру сходства, которая используется для векторных моделей. Векторная модель позволяет представить текст документа и запроса в виде векторов одного пространства, а степень схожести этих векторов выражается через косинус угла между ними.

Документы и запрос состоят из множества слов, которые в терминологии информационного поиска именуются термами. В векторной модели веса термов дают возможность описать документы в виде векторов, размерность которых совпадает с количеством различных термов в коллекции. Это можно представить в математическом виде:

$$d_j = (w_{1j}, w_{2j}, \dots, w_{nj}), \quad (4)$$

где d_j – векторное представление j-го документа коллекции, w_{ij} – вес i-го терма в j-ом документе коллекции, n – количество различных термов во всей коллекции.

Запрос в математическом виде определяется аналогичным образом:

$$q = (w_{1q}, w_{2q}, \dots, w_{nq}). \quad (5)$$

Такое представление позволяет определять расстояние между точками пространства, следовательно, и оценивать степень схожести документов (близкое расположение точек указывает на высокую релевантность текстов). На рисунке 1 представлен пример векторной модели запроса q и документов d_1 и d_2.

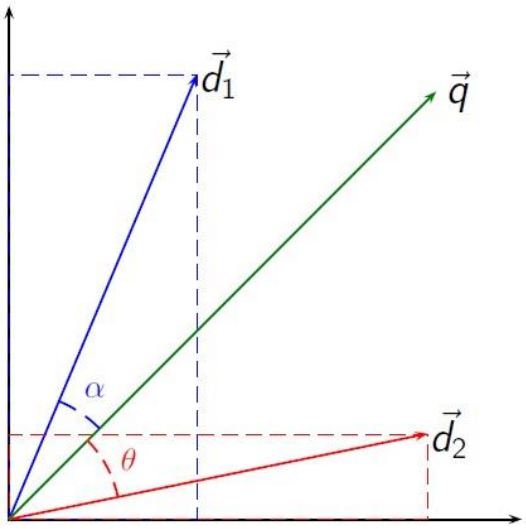

Рис. 1. Векторная модель

В векторной модели вес терма может определяться различными способами:

- булевский вес – вес терма равен 1, если терм входит в документ, иначе он равен 0;
- TF – вес терма зависит от частоты встречаемости терма в текста;
- TF-IDF – вес терма равен произведению частоты встречаемости терма в документе на обратную частоту этого документа.

В контексте рассматриваемой задачи используется совмещение двух способов определения веса терма. Для термов запроса используется булевский вес, а для термов документов – метод TF-IDF. Формула для расчета IDF совпадает с формулой, приведенной в методе Okapi BM25, а формула для TF имеет вид:

$$TF(t) = \frac{n_t}{\sum_k n_k}, \tag{6}$$

где n_t – количество вхождений терма t в документ, а знаменатель содержит общее число слов в этом документе.

Такой вариант определения весов выбран для того, чтобы термы документов имели бóльшую значимость, что позволит обеспечить корректную оценку релевантности документов запросу.

Оценка релевантности документа d_j запросу q определяется косинусным сходством по формуле:

$$\cos(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \cdot \|q\|},$$

$$\cos(d_j, q) = \frac{\sum_{i=1}^{N} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{N} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{N} w_{iq}^2}}, \tag{7}$$

где косинус между векторами документа d_j и запроса q определяется как отношение скалярного произведения векторов к произведению их длин (длиной вектора принято считать его евклидову норму).

При оценке релевантности косинусное сходство документа и запроса принимает значения из диапазона от 0 до 1. Это связано с тем, что веса термом не могут быть отрицательными (на веса TF-IDF накладывается условие нижней границы, аналогичное условию для функции Okapi BM25). Если значение косинуса равно 0, то вектора документа и запроса являются ортогональными и не сходятся между собой (термы запроса не встречаются в рассматриваемом документе). Вектора документа и запроса полностью совпадают, если значение косинуса равно 1.

Одно из преимуществ косинусного сходства заключается в его низкой сложности, так как данный метод исследует исключительно ненулевые

измерения [7]. Также рассматриваемый метод позволяет компенсировать влияние длины документа, так как он обеспечивает нормализацию векторов.

Косинусное сходство разделяет недостатки, присущие большинству метод оценки релевантности текстов. Так, оценка релевантности, являющая результатом использования данного метода, зависит от формы слов, которые участвуют в анализе текстов (данная проблема также устраняется путем применения лемматизации). Также, косинусное сравнение не рассматривает близость слов и их положение в документе. Большой размер исследуемых документов занимает много памяти и сильно влияет на производительность работы метода.

Парето-оптимальное решение и взвешенная оценка

Пусть каждый из трех методов – это отдельный критерий, который необходимо использовать для получения наилучшего решения [8]. Оценка, вычисленная методом, – это значение соответствующего критерия.

Рассмотрим Парето-оптимальное множество решений. Решение называется оптимальным по Парето, если не существует других решений, которые бы его доминировали. В контексте рассматриваемой задачи множеством таких решений являются аннотации, для которых не существует других аннотаций, у которых по каждому из трех критериев оценки не хуже, а по одному из них – строго лучше. То есть следующая система не должна выполняться:

$$\begin{cases} score_{method_i}(d_1) \geq score_{method_i}(d_2), i = (1,3) \\ \exists\, method_j\colon score_{method_j}(d_1) > score_{method_j}(d_2) \end{cases}, \tag{8}$$

где $method_i$ и $method_j$ выступают в роли одного из методов (Okapi BM25, коэффициент Танимото или косинусное сходство).

Чтобы рационально определить степень релевантности документа запросу, необходимо рассчитать общую оценку работы всех методов. Она представляет собой сумму трех нормированных оценок, которые складываются с заданным весом w_i:

$$score_{final} = \sum_i w_i\, score_{norm\, method_i},$$
$$score_{norm\, method_i} = \frac{score_{method_i} - \min_i(score_{method_i})}{\max_i(score_{method_i}) - \min_i(score_{method_i})}, \tag{9}$$

где $method_i$ выступает в роли одного из методов (Okapi BM25, коэффициент Танимото или косинусное сходство), $score_{norm\, method_j}$ нормализует оценку каждого метода, функции min и max определяют соответственно наименьшую и наибольшую оценку для каждого них.

Это позволяет снизить погрешности и шумы работы некоторых методов, а также увеличить точность работы других. Таким образом, полученная сумма будет служить итоговой взвешенной оценкой релевантности текстов, которая должна наиболее грамотно отражать сходство анализируемых текстов.

Выбор веса w_i – это не простая задача. В обычных задачах многокритериального анализа полагаются на мнение лица, принимающего решение (ЛПР) или эксперта. Часто используют процедуру парных сравнений, которая снижает информационную нагрузку на ЛПР. Также используются специализированные интерфейсы пользователя с представлением весов в виде столбчатой диаграммы или долей круга. Перспективным направлением дальнейшего развития методов определения w_i является использование методов машинного обучения, однако на настоящем этапе данного исследования пока отсутствует достаточный объем исходных данных для обучения машин. Текущие вычислительные эксперименты проводились с равными весами $w_1 = w_2 = w_3 = \frac{1}{3}$.

Результаты использования методов

Фрагменты результатов, полученных в ходе проведённого исследования, представлены в таблицах 1-3. Они отражают работу каждого из методов для конкретной темы НИР и позволяют сравнить аннотации, которые в ходе работы этих методов выбираются как наиболее релевантные. Аннотации, оценки которых принадлежат к Парето-оптимальному множеству решений, помечены звездочкой (*). Аннотации ранжированы в порядке уменьшения итоговой взвешенной оценки. Парето-оптимальному множеству решений, принадлежат более одной аннотации. Лучшие оценки по каждому критерию выделены жирным шрифтом. Максимум оценки по коэффициенту Танимото, давал альтернативу, которая совпадала со взвешенной оценкой при равных весах. Однако в таблице 2 видно, что коэффициент Танимото не смог различить 2 альтернативы, одна из которых получила более низкую оценку по другим критериям.

Можно заметить, что каждый из методов находит аннотации, которые действительно соответствуют заданному запросу, но иногда предложения системы не совпадают с ожидаемыми результатами. Примеры результатов работы программы позволяют заметить, что в некоторых случаях один метод работает лучше, чем другие, но в иных случаях - хуже. Это позволяет сделать вывод, что каждый описанный метод имеет свои достоинства и недостатки, которые по-разному проявляются при работе с конкретным набором данных.

Так, метод с коэффициентом Танимото увеличивает оценку коротких аннотаций, так как он зависит от длины документа. А методы Okapi BM25 и косинусное сходство не позволяют получить релевантную оценку, которая не зависела бы от оценок других документов коллекции. В то же время, косинусное сходство, в отличии от Okapi BM25, имеет верхнюю границу

допустимого значения, что ограничивает максимальную оценку релевантности для оптимальности поиска и анализа.

Таблица 1

Развитие информационно-аналитической системы оперативного мониторинга и оценки состояния и рисков научно-технического обеспечения развития сельского хозяйства (ИАС НТОР-СХ) в части создания подсистемы «Создание отечественного конкурентоспособного кросса мясных кур в целях получения бройлеров»

Аннотация	Okapi BM25	Коэффициент Танимото	Косинусное сходство	Итоговая взвешенная оценка
* Представлены подходы к созданию общей системы управления агропромышленным комплексом на основе одновременного формирования и развития подсистемы хозяйственного самоуправления и государственного регулирования	23.3403717	**0.15**	0.1580072	0.8394604
* Рассматриваются экономические аспекты обеспечения информационной безопасности в организации и предлагается новый подход к оценке и анализу информационных рисков	19.9765648	0.102564	**0.2571705**	0.8191516
* В статье дана оценка государственному регулированию аграрного сектора, государственной поддержке и программам развития сельского хозяйства в зарубежных странах	21.00501	0.1282051	0.1799463	0.7893144
* В статье предпринимается попытка оценить текущий уровень развития сельского хозяйства стран, исходя из принципов устойчивого развития. Данная оценка производилась на основе определения Евклидового расстояния между фактическими значениями страны и эталонными значениями показателей, характеризующих сельское хозяйство. В итоге была дана краткая характеристика уровня развития сельского хозяйства	**24.052538**	0.0806452	0.2117787	0.7642289
* В статье раскрывается понятие устойчивого развития сельского хозяйства. Выделяются основные отличительные черты устойчивого и неустойчивого развития сельского хозяйства. Определяются основные особенности присущее устойчивому развитию сельского хозяйства. Выявляются основные барьеры присущее устойчивому развитию сельского хозяйства. В итоге представлены основные мероприятия, способствующие достижению устойчивого развития сельского хозяйства...	25.8197137	0.0597015	0.2263507	0.7593894

Таблица 2

Международный опыт исследования состояния конкуренции на товарных рынках в условиях трансформации системы экономических связей для целей антимонопольного правоприменения

Аннотация	Okapi BM25	Коэффициент Танимото	Косинусное сходство	Итоговая взвешенная оценка
* В статье дана оценка перспектив формирования в России мегарегулятора финансового рынка, проведен анализ сложившихся условий и предпосылок мегарегулирования с учетом зарубежного опыта, современного состояния российского финансового рынка и системы его регулирования	**17.032513**	**0.1428571**	0.2291146	0.9773649
Помещены материалы, посвященные проблемам и перспективам регионального развития Российской Федерации, международным экономическим отношениям, проблемам энергетических ресурсов и экономической безопасности, вопросам состояния и развития финансового рынка	14.701038	0.1212121	0.2188059	0.8672524
* В статье рассмотрен экономический механизм региональной экономической системы, предложен системный подход исследования функционирования региональной экономической системы	11.832532	0.1111111	**0.2458061**	0.8241601
Проведено комплексное исследование различных подходов к трактовке дефинитивного аппарата в таможенной сфере в целях совершенствования национальной таможенной системы в условиях глобализации	12.455930	**0.1428571**	0.1265349	0.7486928
Рассматриваются вопросы устойчивого развития и реструктуризации предприятий. Уделяется внимание исследованию понятий гибкости и прочности связей между элементами производственно-экономической системы, то есть между подразделениями и структурными звеньями	12.255572	0.1290322	0.1402735	0.7311442

Таблица 3

Разработка стратегии поддержки экспорта

Аннотация	Okapi BM25	Коэффициент Танимото	Косинусное сходство	Итоговая взвешенная оценка
* Стратегия - интеграционная модель действий, предназначенных для достижения целей организации. Содержанием стратегии служит набор правил принятия решений, используемый для определения основных направлений эффективной ее деятельности. В статье рассматриваются основные понятия о стратегии, а также основы ее выбора, формирования, разработки и реализации в целях выполнения миссии организации. При этом	9.7401557	**0.0408163**	**0.3079591**	0.7618717

Аннотация	Okapi BM25	Коэффициент Танимото	Косинусное сходство	Итоговая взвешенная оценка
используются: количественные методы прогнозирования, разработка сценариев будущего развития, портфельный анализ и другие...				
В данной статье автором разработан критерий эффективности экспорта на основе методики, используемой Федеральным статистическим ведомством Германии. Согласно этой методике, экономическая эффективность экспорта может быть определена с помощью отношения совокупного объема экспорта к совокупному объему импорта в исследуемый период. Разработанный мной в рамках данной статьи критерий позволяет оценивать эффективность экспорта страны, в том числе, и в период локальных или мировых кризисов, а также в условиях нестабильности мировой...	9.6335018	0.0188679	0.2799978	0.676744
* В статье представлен анализ состояния внешней торговли России после введения санкций западными странами в связи с украинскими событиями 2014 г. Изучена динамика основных макроэкономических показателей страны: экспорта, импорта, внешнеторгового оборота. Особое внимание уделено изменению товарной структуры внешней торговли России со странами ЕАЭС и странами-санкционерами. Сделан вывод, что благодаря санкциям «замороженные» на долгое время статьи экспорта и импорта РФ начали изменяться, приобретать потенциал, и у России есть шанс двигаться дальше по пути повышения экспорта обрабатывающей промышленности и развития импортозамещения при условии комплексного применения мер поддержки...	**9.7411208**	0.0277778	0.2328088	0.6501392
Основной целью данной статьи является рассмотрение вопросов, связанных с развитием экспорта ведущих стран ЕС в условиях возникших экономических проблем внутри еврозоны. Авторами в рамках данной работы оценивается эффективность мер стимулирования экспорта ведущих стран Европейского Союза (Германии и Франции). В данной статье авторами предложена методика учета влияния социально - экономического состояния стран - импортеров на объемы экспорта ведущих стран - экспортеров ЕС. В статье дается авторская оценка динамики роста экспорта...	9.0267027	0.016129	0.2621746	0.6302973

Итоговая взвешенная оценка предоставляет свой набор результатов, полученный путём агрегирования всех оценок. Такой подход обеспечивает приемлемый результат работы системы. Часто модель не может самостоятельно оценить, каким словам предать бóльшую значимость, без смыслового понимания текста. Так, каждый метод пытается определить оценку по частоте совпадаемых слов, не уделяя должного внимания смысловой структуре и

контексту документов и запроса. Поэтому в практических задачах имеет смысл выдавать первые несколько предложений системы из ранжированного по итоговой оценки списка для окончательного принятия решения пользователем.

Заключение

Рассмотренные в настоящей работе методы оценки релевантности текстов позволяют находить аннотации, которые соответствуют заданной тематике НИР. Использование данных методов обеспечивает осуществление рационального поиска и значительно упрощает процесс отбора данных.

Каждый метод оценки релевантности имеет свои преимущества и недостатки, которые влияют на результат работы системы. Так, в некоторых случаях один из методов может оценивать аннотацию наиболее грамотно и точно, а в других – уступать остальным методам и предлагать аннотацию, которая не соответствует заданной тематике НИР. Не существует метода, который всегда дает оценку релевантности текстов лучше других. Поэтому целесообразно определить Парето-оптимальное множество решений, которое позволит найти наиболее эффективные оценки аннотаций.

Применение данной методики может быть расширено на задачи автоматического анализа резюме и мотивационных писем кандидатов на замещение вакансий по описанию вакансий в случае сложных слабоструктурированных требований к навыкам кандидатов.

Библиографический список

1. Судаков В.А. Автоматизация процесса управления разработкой корпоративной информационной системы // Вестник Московского авиационного института. 2010. Т. 17. № 1. с. 149-153

2. Савостин П.А., Ефремова Н.Э. Методы и средства поиска информации в коллекциях научно-технических текстов // Новые информационные технологии в автоматизированных системах, 2018. No. 21, 151-155 с. – Изд-во: Московский институт электроники и математики НИУ ВШЭ (Москва), ISSN: 2227-0973, URL: https://cyberleninka.ru/article/n/metody-i-sredstva-poiska-informatsii-v-kollektsiyah-nauchno-tehnicheskih-tekstov

3. Sura Mahmood Abdullah, Sura Mazin Ali, Mohammed Abduljaleel Makttof. Modifying Jaccard Coefficient for Texts Similarity // Opcion, Año 35, Nº Especial 19 (2019), p. 2899-2921 – ISSN 1012-1587, URL: https://www.researchgate.net/publication/340267006_Modifying_Jaccard_Coefficient_for_Texts_Similarity

4. Avi Rana, K. Deeba. Online Book Recommendation System using Collaborative Filtering (With Jaccard Similarity) // Journal of Physics: Conference Series 1362:012130 (2019). URL:

https://www.researchgate.net/publication/337310327_Online_Book_Recommendation_System_using_Collaborative_Filtering_With_Jaccard_Similarity

5. Putri Yuni Ristanti, Aji Prasetya Wibawa, Utomo Pujianto Electrica. Cosine Similarity for Title and Abstract of Economic Journal Classification, 2019. URL: https://www.researchgate.net/publication/339174965_Cosine_Similarity_for_Title_and_Abstract_of_Economic_Journal_Classification

6. Sergio Jimeneza, Silviu-Petru Cucerzan, Fabio A. Gonzalez, Alexander Gelbukh, George Dueñas. BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies // Journal of Intelligent and Fuzzy Systems POST-PRINT VERSION 34, No. 5 (2018) 2887–2899, IOS Press. URL: https://www.researchgate.net/publication/325231406_BM25-CTF_Improving_TF_and_IDF_factors_in_BM25_by_using_collection_term_frequencies

7. Сторожук Н.О., Коломойцева И.А. Анализ методов определения текстовой близости документов – Материалы студенческой секции IX Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование» (ИУСМКМ-2018). – Донецк: ДонНТУ, 2018. – С. 43-47.

8. Судаков В.А. Инструментарий поддержки решений в многокритериальных задачах развития высокотехнологичных отраслей промышленности – М.: Вега-Инфо, 2020. – 200 с. – ISBN 978-5-91590-056-0.

Оглавление