ПРИМЕНЕНИЕ КОМБИНИРОВАННЫХ СТАТИСТИЧЕСКИХ АЛГОРИТМОВ ДЛЯ ФОРМИРОВАНИЯ РЕФЕРАТОВ И ОЦЕНКИ РЕЛЕВАНТНОСТИ НАУЧНО-ТЕХНИЧЕСКИХ ПУБЛИКАЦИЙ

Тарасов А. Ф., Васильева Л. В., Морозов Д. А.

Проанализированы существующие методы автоматизированной обработки текста на основе применения комбинированных статистических алгоритмов для формирования рефератов и оценки релевантности статей. Проведено исследование группы статей, в результате которого выделены ключевые слова по тематике применения интенсивных пластических деформаций. С применением предложенного алгоритма выполнена автоматизированная обработка текстов статей на двух языках – английском и русском, что подтверждает универсальность принятого подхода к анализу и оценке научных текстов. Улучшен алгоритм статистической обработки научно-технической информации с учетом выделения разделов статьи, управления рядом параметров, характеризующих модель оценки релевантности текста, что позволило получать сжатые текстовые документы на выбранном языке.

Проаналізовано існуючі методи автоматизованої обробки тексту на основі застосування комбінованих статистичних алгоритмів для формування рефератів і оцінки релевантності статей. Проведено дослідження, в результаті якого виділені ключові слова по тематиці застосування інтенсивних пластичних деформацій. Досліджена обробка текстів статей двома мовами – англійською і російською, що підтверджує універсальність прийнятого підходу до аналізу наукових текстів. Поліпшено алгоритми, методи, системи обробки науково-технічної інформації, з урахуванням нелінійної і ієрархічної структури тексту, що дозволило отримувати стислі текстові документи обраною мовою.

The existing methods of automated text processing on the basis of the combined statistical algorithms for generating abstracts and assessment of relevant articles are analyzed. The study was carried out to highlight the key words on the subject of the application of intensive plastic deformation. Articles' words processing articles in two languages was made - in English and Russian, which confirms the universality of the approach taken to the analysis and assessment of scientific texts. Algorithm for statistical processing the scientific and technical information was improved, taking into account the non-linear and hierarchical structure of the text. It allowed to receive short text documents in the selected language.

Тарасов А. Ф. — д-р. техн. наук, проф. каф. КИТ ДГМА

Васильева Л. В. — канд. техн. наук, доц. каф. КИТ ДГМА
kit@dgma.donetsk.ua

Морозов Д. А. — магистрант каф. КИТ ДГМА

ДГМА – Донбасская государственная машиностроительная академия, г. Краматорск.

УДК 004.912

Тарасов А. Ф., Васильева Л. В., Морозов Д. А.

ПРИМЕНЕНИЕ КОМБИНИРОВАННЫХ СТАТИСТИЧЕСКИХ АЛГОРИТМОВ ДЛЯ ФОРМИРОВАНИЯ РЕФЕРАТОВ И ОЦЕНКИ РЕЛЕВАНТНОСТИ НАУЧНО-ТЕХНИЧЕСКИХ ПУБЛИКАЦИЙ

Современные информационные задачи потокового международного научного документооборота в научных библиотеках создают необходимость комплексного автоматизированного машинного анализа электронных (в основном pdf) версий публикуемых научных работ. В рамках такого анализа возникает необходимость автоматического распознавания обязательных элементов опубликованной научной или технической информации (НТИ), включающей как текстовую, так и графическую информацию различного вида.

На данный момент большое количество программных комплексов для обработки текстов разработано, в основном, специалистами университетских и научных центров [1–3]. В этих системах реализованы различные нетрадиционные решения, отличные от статистических методов, основанные на построении лексических цепочек, концептуальных графов, а также эффективных способов описания структуры текста. Однако все эти методы ориентированы на учет особенностей конкретных языков (в основном английского), поэтому не могут быть применены для автоматического анализа и реферирования текстов на русском или украинском языках. Кроме того, наиболее эффективные разработки имеют коммерческий характер, в связи с чем, принципы и алгоритмы их работы не раскрываются [4–5]. Научные статьи сопровождаются обычно иллюстрациями с описанием объекта исследования, схемами или чертежами, которые демонстрируют принципы действия тех или иных частей конструкции, а также результатами исследований в виде графических зависимостей ряда показателей. Результаты исследований могут быть представлены в виде диаграмм, графиков, изображений внешнего вида (форма заготовки, конструкция устройства) и внутренней структуры (микроструктура, молекулярное состояние) исследуемых материалов до и после исследований. Для обработки изображений, содержащихся в научных публикациях, разработан ряд методов и программных продуктов [6]. В том числе авторами данного исследования было разработано алгоритмическое и программное обеспечение сегментации изображений [7].

Таким образом, актуальным в данный момент является вопрос улучшения существующих или создания новых комбинированных алгоритмов, методов, систем обработки НТИ, учитывая нелинейную и иерархическую природу текста и позволяющих получать сжатые текстовые документы на необходимом языке.

Целью данной работы является совершенствование существующих методов автоматизированной обработки текста на основе применения комбинированных статистических алгоритмов для формирования рефератов и оценки релевантности статей.

Принятый алгоритм анализа и обработки текстовой информации в научно-технических публикациях содержал типовые этапы анализа текста, что создает основу для исследования различных алгоритмов и их модификаций. Текст статьи разбивался на структурные разделы и предложения, из каждого предложения удалялись стоп-слова. Первичным результатом автоматического анализа текстовой информации является частотный анализ документа, который позволяет сформировать список ключевых слов, вычислить их значимость в виде плотности вхождения слов в текст статьи.

Наличие ключевых слов позволило выполнить оценку предложений в целом и сформировать реферат, отражающий содержание исходного документа, с учетом ввода параметров обработки текста (раздел статьи, максимальное количество или минимальная оценка предложений, важность слов, регистр текста и др.). В данной работе анализ научно-

технических публикаций ограничен рассмотрением новых процессов обработки заготовок с применением интенсивных пластических деформаций [8].

Кроме того, конечным результатом может быть не только список ключевых слов, аббревиатур исходного документа и реферат, но и общая оценка статьи, что позволяет оценивать ее релевантность изучаемой теме. При этом реферат может рассматриваться вместо статьи, что снижает трудоемкость анализа.

В данной работе рассматривалась возможность выделения в процессе обработки научно-технических текстов эталонных статей и рефератов для последующего определения отношения произвольной статьи к заданной тематике, а также степени её близости к эталонной группе. Эталонные статьи С1 – С5 определялись по принятым наукометрическим соотношениям для расчета показателей цитируемости научных работ. Для проведения обработки текста были выбраны по пять статей на английском и русском языках как эталон, который позволяет получить оценку их коэффициентов значимости для последующего анализа релевантности других документов. На основе проведенного частотного анализа выделены по 8 слов, которые имеют наибольшой вес и являются значимыми для всех выбранных для анализа статей (табл. 1).

Таблица 1

Нормированная плотность вхождения слов в каждую статью

Русскоязычная статья						Англоязычная статья					
Слово	Плотность вхождения					Слово	Плотность вхождения				
	С1	С2	С3	С4	С5		С1	С2	С3	С4	С5
Деформации	0,70	0,38	1,00	0,61	0,21	Deformation	1,00	0,90	0,53	0,27	0,40
Пластической	0,25	0,30	0,44	0,30	0,23	Plastic	0,30	0,29	0,27	0,65	0,36
Материалов	0,39	0,24	0,18	0,19	0,63	Strain	0,10	0,10	0,33	0,62	0,28
Структуры	0,07	0,15	0,30	0,52	0,18	Stress	0,23	0,19	0,12	0,67	0,13
ИПД	0,11	0,10	0,36	0,02	0,16	Crack	0,00	0,00	0,18	0,55	0,00
Давлений	0,36	0,18	0,04	0,30	0,04	ECAP	0,81	0,70	0,64	0,00	0,02
Сдвиг	0,66	0,38	0,04	0,12	0,00	Fracture	0,05	0,00	0,18	0,49	0,02
Металл	0,15	0,14	0,40	0,10	0,44	Material	0,26	0,29	0,20	0,35	0,26

Оценку степени значимости слов производили путем присвоения им рангового номера. Слову, которое имеет наивысшую значимость, присваивали ранг 1, а равнозначным словам присваивается одинаковый ранговый номер. На основе данных частотного анализа составляется сводная матрица рангов. Так как в матрице имеются связанные ранги (одинаковый ранговый номер) в значимости слов, производят их переформирование без изменения значимости слов. Оценка средней степени согласованности для англоязычных и русскоязычных статей выполнялась с помощью коэффициента конкордации для случая, когда имеются связанные ранги:

$$W = \frac{S}{\frac{1}{12}m^2(n^3-n)-m\sum T_i},$$

где S = 420, n = 8, m = 5, $\boldsymbol{T_i}$ – число связок (видов повторяющихся элементов) в оценках i-й статьи, $\boldsymbol{t_i}$ – количество элементов в i-й связке для i-й статьи (количество повторяющихся элементов).

Оценка значимости коэффициента конкордации определялась по критерию Пирсона. Так как χ^2 расчетный 18,86 больше табличного (χ^2 = 14,067), то коэффициент конкордации – величина не случайная, а потому полученные результаты могут использоваться в дальнейших исследованиях.

На основе полученных значений рангов были рассчитаны показатели значимости русскоязычных (рис. 1) и англоязычных статей (рис. 2).

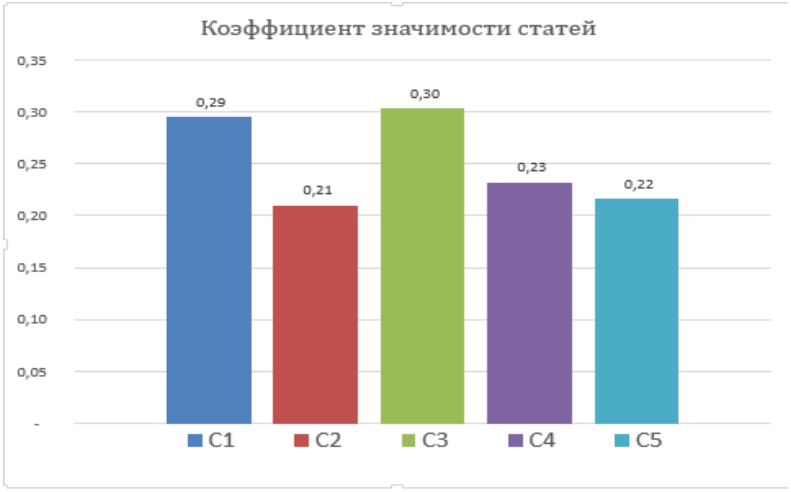

Рис. 1. Коэффициент значимости русскоязычных эталонных статей

Рис. 2. Коэффициент значимости англоязычных эталонных статей

Таким образом, на основе выделения ключевых слов в эталонных статьях для предметной области можно осуществить анализ нового документа и определить его «меру схожести» (релевантность) с эталоном. Для реализации такого анализа предлагается следующий алгоритм:

1. Определение плотности вхождения ключевых слов в изучаемую статью.
2. Нормирование значений плотности вхождения ключевых слов новой статьи с имеющимися эталонными.
3. Нахождение матрицы рангов весомости ключевых слов.
4. Определение расчетного значения весомости статьи с учётом коэффициента весомости ключевых слов.
5. Сравнение среднего коэффициента значимости эталона $\boldsymbol{k}_{\text{ср}}$ с рассчитанным коэффициентом значимости проанализированной статьи $\boldsymbol{k}_{\text{ан}}$.
6. Определение сходства с эталоном по формуле:

$$\boldsymbol{k}_{\text{схожести}} = \frac{\boldsymbol{k}_{\text{ан}}}{\boldsymbol{k}_{\text{ср}}} * \mathbf{100\%}.$$

Приведенный алгоритм был протестирован на научно-технических публикациях на двух языках (английском и русском) в пределах рассмотренной тематики. В результате расчетов получены коэффициенты значимости для русскоязычных (рис. 3) и англоязычных эталонных статей (рис. 4) и рассчитаны коэффициенты для новых анализируемых статей (табл. 2).

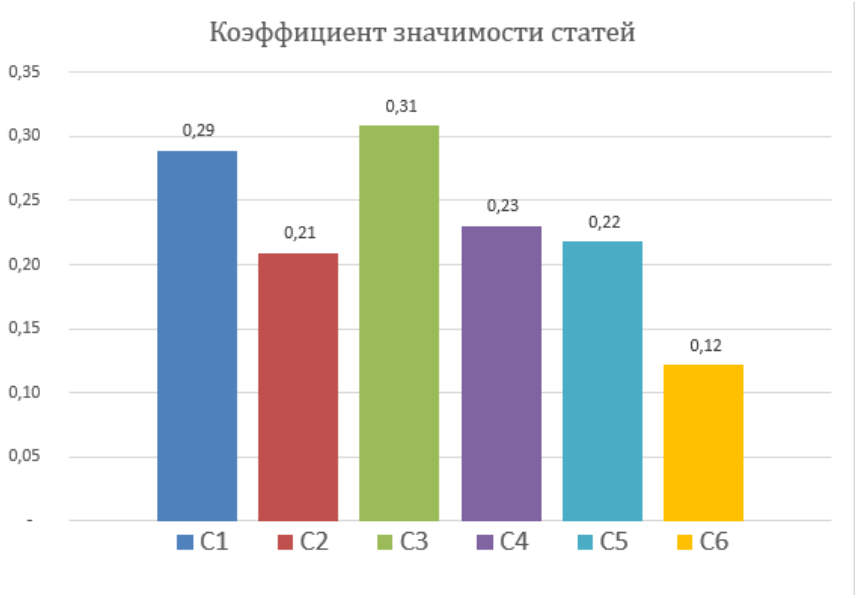

Рис. 3. Коэффициент значимости (оценка релевантности) новой русскоязычной статьи относительно группы эталонных статей

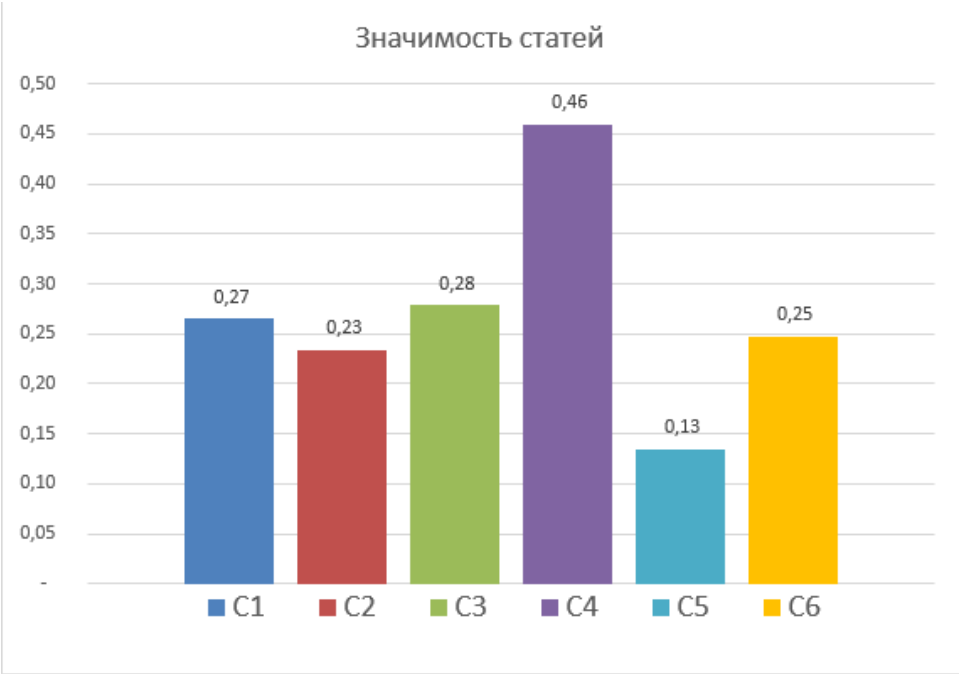

Рис. 4. Коэффициенты значимости (оценка релевантности) новой англоязычной статьи относительно группы эталонных статей

Таблица 2

Рассчитанные показатели проанализированных статей

Коэффициент значимости, $К_з$	Русскоязычные статьи	Англоязычные статьи
Средний $К_з$ эталонных статей	0,25	0,27
$К_з$ новой статьи	0,12	0,25
Коэффициент схожести новой статьи с эталонными, %	48,50	90,38

Результаты расчетов позволяют сделать вывод, что коэффициент значимости новой англоязычной статьи близок к эталонному, а русскоязычная статья только на 48,5% имеет сходство с эталоном. Таким образом, выполнена статистическая оценка значимости (релевантности) новых статей для пользователя.

Предложенный алгоритм анализа НТИ позволяет автоматизировать обработку текстов статей по заданной тематике на основе выделения ключевых слов и оценки значимости предложений при формировании рефератов статей. Модели оценки значимости могут применяться разные. При этом выделенные типовые этапы обработки текста являются основой для модификации анализа при рассмотрении групп статей. Применение данного подхода может быть эффективным и для поиска в уже накопленном у пользователя массиве публикаций по определенной теме исследования. В частности, если имеется группа статей, то возможно автоматическое включение их всех или нескольких произвольных в группу эталонных и последующее автоматизированное ранжирование статей на основе ограничения перечня интересующих пользователя ключевых слов и других параметров настройки процесса обработки. Естественно, что решение о релевантности статей в каждом случае будет ситуативным.

ВЫВОДЫ

Исследование применения комбинированных статистических алгоритмов для формирования рефератов и оценки релевантности научно-технических публикаций на двух языках показало возможность их применения как инструмента поиска нужной информации в массиве накопленной НТИ.

Разработанный программный комплекс является основой для изучения возможностей различных методов автоматизированной обработки НТИ и ситуативной оценки их релевантности в массиве накопленной НТИ.

Проведено исследование, в результате которого выделены ключевые слова по тематике применения интенсивных пластических деформаций. Изучена обработка текстов статей на двух языках – английском и русском, что подтверждает универсальность принятого подхода к анализу НТИ.

На основе расчета коэффициентов значимости для исследуемых документов определена степень релевантности статей эталонным публикациям.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1.Гинкул А.С. Сравнительный анализ существующих систем автоматического реферирования текста / А.С. Гинкул // Політ. сучасні проблеми науки – Киев, 2012. – С. 255.

2.Extractor – Метод извлечения ключевых фраз [Электронный ресурс]. – Режим доступа : http://www.legroom.net/software/uniextract

3.TextAnalyst – Система автоматического анализа текста [Электронный ресурс]. – Режим доступ : http://www.analyst.ru/index.php.

4.Луканин А.В. Автоматическая обработка естественного языка / А.В. Луканин; ЮУрГУ – Челябинск : Изд. центр ЮУрГУ, 2011. – 70 с.

5.Aliguliyev R. M. Automatic document summarization by sentence extraction // Institute of Information Technology of the National Academy of Sciences of Azerbaijan, Baku, October 2008, pp. 1-4.

6.Яне Б. Цифровая обработка изображений. /Пер. с англ. – М.: Техносфера, 2007. – 584 с. ISBN 987-5-94836-122-2.

7.Васильева Л. В. Разработка алгоритмического и программного обеспечения сегментации изображений / Л.В. Васильева, А.Ф. Тарасов, И.А. Гетьман // Вимірювальна та обчислювальна техніка в технологічних процесах. – Хмельницький : ХНУ — 2016. – № 3 (56). – С. 117-122.

8.Систематизация процессов интенсивного пластического деформирования (ИПД) объемных заготовок на основе онтологического подхода / А. В. Периг, А. Ф. Тарасов, А. В. Алтухов // Вісник Національного технічного університету «ХПІ». Серія: Нові рішення в сучасних технологіях: зб. наук. пр. – Харків: НТУ «ХПІ». – 2012. – № 46 (952). – С. 83-89. – Бібліогр.: с. 89. – ISSN 2079-5459.

Статья поступила в редакцию 20.04.2017 г.