УДК 004.9:371.261

Н.Б. Шаховская, Р.Ю. Нога

МЕТОДЫ И СРЕДСТВА АНАЛИЗА ТЕКСТОВ ПУБЛИКАЦИЙ ДЛЯ ИССЛЕДОВАНИЯ ДЕЯТЕЛЬНОСТИ НАУЧНЫХ ШКОЛ

Постановка задачи

Необходимой предпосылкой повышения эффективности и инновационной отдачи научных исследований является полноценное и оперативное обеспечение общества новейшей информацией. При оценке результативности научной деятельности важное место отводится наукометрии — направлению исследований, которое изучает когнитивные коммуникации в науке по частоте цитирования научных работ и их авторов. Объектом исследований наукометрии является научная школа и результаты ее функционирования.

Научная школа — неформальный творческий коллектив исследователей разных поколений, объединенных общей программой и стилем исследовательской работы, которые действуют под руководством признанного лидера [1]. Деятельность научной школы заключается в выработке научных знаний, их распространении, подготовке одаренных специалистов. Одним из вариантов представления результата выработки научных знаний является научная публикация, представленная в виде слабоструктурированного или неструктурированного текста [2]. Наличие научных школ — один из определяющих факторов развития региона, поскольку это непосредственно указывает на инновационную деятельность, а также позволяет спрогнозировать, какую сферу деятельности целесообразно развивать. Однако большое количество публикаций в Интернете и снижение интереса к науке и ее развитию в последние годы значительно усложнило процедуру определения научных школ и анализа их деятельности.

Для обработки текстов используются различные методы поиска, рубрикации или кластеризации научных текстов. Необходимо отметить работы в области математической и прикладной лингвистики и лексикографии [3], экстракции значащих признаков [4] и рубрикации текстов [5].

В 1980-х гг. основными методами кластеризации текстов являлись экспертные методы, основанные на использовании экспертных оценок для определения тематики документов. Сегодня этот подход эффективен при решении задач, которые требуют принятия нетривиальных решений об отнесении тех или иных текстов к одному кластеру. Однако вместе с тем ручные методы кластеризации имеют ряд особенностей, что существенно ограничивает возможности их использования:

- ручная кластеризация применима лишь для относительно небольших массивов документов, что в современных условиях встречается довольно редко (по исследованиям Digital Universe Study 2013 года, суммарный объем данных составил 6,3 Zбайт (2^{70} байт), 43 % из них — объем данных, содержащих результаты научных исследований);
- ручные методы требуют продолжительного времени работы, так как эксперту необходимо время на принятие решения по каждому тексту.

Указанные особенности методов ручной кластеризации массивов текстов привели к разработке полуавтоматических, а позднее и автоматических методов

текстовой кластеризации [6, 7]. Для текстов используют такие алгоритмы, как Expectation maximization, Fuzzy Codok, k-means [8] и др. Однако основная проблема алгоритмов текстовой кластеризации состоит в определении сходства текстов и высокой вычислительной сложности [9, 10].

Актуальность темы исследования обусловлена такими факторами:

- популярность междисциплинарных исследований, усложняющая отнесение публикации только к одной научной школе;
- рост количества исследователей (средний процент роста количества защищенных диссертационных работ в Украине на протяжении последних четырех лет — 1,8 %);
- динамичность науки, быстрое старение информации (согласно данным Digital Universe Study, скорость старения информации за последние пять лет увеличилась вдвое);
- большое количество публикаций в Интернете усложняет выявление основоположников и участников научной школы существующими методами кластеризации, что делает невозможным налаживание связей между исследователями.

Поэтому на сегодня задача разработки методов и средств анализа текстов публикаций для определения научных школ и исследования их деятельности является актуальной.

Цель работы — разработка методов и средств определения научных школ на основе анализа текстов научных публикаций.

Формальное определение научной школы

Определение 1. Научная школа *S* характеризуется множеством публикаций *Sch* научного направления, которое определено множеством ключевых слов *Key*, множеством авторов *Author* и множеством основоположников школы *Main*:

$$S = \langle Main, Sch, Rate \rangle, Main \in Author,$$

$$Sch_i = \langle Key, Author, T_i, Publish, IFactor_i, Type \rangle,$$

$$Author_i = \langle Surname_i, Name_i, Degree_i, Organization_i, Post_i \rangle,$$

$$Rate = f_1(.IFactor, \Delta Degree),\ f_2 : T \to Key.$$

Публикация Sch_i характеризуется множеством ключевых слов Key и авторов $Author$, полным текстом T_i, изданием $Publish_i$, рейтингом $IFactor_i$ и типом $Type_i$; автор $Author_i$, соответственно, — такими характеристиками, как фамилия, имя, научная степень, организация, должность; $Rate = f_1(IFactor, \Delta Degree)$ — это функция определения рейтинга научной школы с учетом индекса публикаций и прироста количества авторов с научными степенями $\Delta Degree$; $IFactor$ — показатель цитирования журналов — определяет их информационную значимость:

$$Rate = \frac{\sum_{i}^{n} k_i IFactor_i}{m} \Delta Degree,$$

где $m = |Sch|$ — количество публикаций в научной школе, k_i — количество публикаций в издании с рейтингом $IFactor_i$.

$$Degree_{t_i} = 100Degree(Doctor) + Degree(Cand),$$

$$\Delta\, Degree = \frac{Degree_{tyear} - Degree_{t0}}{year},$$

$Degree_{t_i}$ — количество кандидатов $Degree(Cand)$ и докторов наук $Degree(Doctor)$ на t_i году наблюдения за развитием школы, $year$ — количество лет наблюдения за школой. Этот показатель дает возможность обеспечить анализ наличия звена «учитель–ученик» в научной школе.

$f_2 : T \to Key$ — функция получения ключевых слов на основе анализа текста T научной публикации. С использованием этой функции построен метод экстракции ключевых слов из текстов научных публикаций.

Определение научных школ в [12] происходит с помощью построенной семантической сети с учетом цитирования публикаций авторов. После выделения соавторов определяются их публикации и общие ключевые слова. Для выделения терминов использовался классический статистический метод TFIDF. Подобным образом определены родственные научные исследования на основе анализа текстов публикаций [2, 13]. С этой целью строилась сеть соавторов с определением веса узла.

В статье предлагается анализировать тексты публикаций и определять школы с учетом результатов анализа. Статистический метод TFIDF предполагает, что публикация — это набор слов. Однако такой подход к научной статье недопустим. Анализ семантической сети также не позволяет определить динамику изменения количественных и качественных характеристик школы (количество публикаций, защит и т.п.), что, в свою очередь, не дает возможности определить такие параметры, как $\Delta Degree$ и $Rate$.

Для определения научной школы и анализа ее деятельности предлагается такой подход.

Метод выделения составляющих текстового документа. Входной информацией для определения принадлежности публикации к научной школе является текстовый файл любого формата с содержимым публикации. Из файла необходимо извлечь базовые элементы публикации: автор(ы) публикации; научное учреждение; название публикации; ключевые слова; основной текст.

Кластеризация. Базовые элементы публикации являются входными параметрами метода кластеризации пубикаций. Результат применения этого метода — определение научных школ.

Классификация. Новые научные публикации анализируются (выделяются базовые характеристики) и относятся к существующим научным школам.

Прогнозирование изменения динамики публикации. Автоматически определяем количество защит ученых степеней среди участников школы. На основании количества защит и количества научных публикаций анализируется деятельность научной школы.

Метод выделения составляющих текстового документа

Научные публикации представляют собой слабоструктурированные электронные документы (ЭД). Элемент ЭД ОСНОВНОЙ ТЕКСТ также имеет внутреннюю структуру, элементы которой разделены заголовками.

Слова, длина которых не больше трех букв, выполняют в тексте служебную роль и не влияют существенным образом на семантику предложения. Для выделения из контента необходимой информации осуществляется загрузка ЭД, реферирование ЭД, экстракция элементов.

Метод выделения составляющих текстового документа (функция f_2) базируется на понятии веса предложения и слова (словосочетания). Основу анализа составляет процедура присвоения весовых коэффициентов каждому блоку текста согласно таким характеристикам:

- расположение блока в оригинале,
- частота появления в тексте,
- частота использования в ключевых предложениях,
- показатели статистической значимости.

Сумма индивидуальных весов слов и предложения, определенная после дополнительной модификации согласно специальным параметрам налаживания, связанным с каждым весом, дает общий вес предложения U:

$$Weight(U) = WordsWeight(U) + 10 * Place(U) + 10 * Format(U) \quad (1)$$

Для формирования реферата выделяются предложения из основной части.

Основная часть, в свою очередь, делится на разделы и подразделы, введенные авторами. Предполагается, что предложения из вступительной части и выводов имеют более высокое информативное значение, нежели предложения из основного текста.

Введем понятие веса предложения. Для этого формализуем элементы формулы (1).

Коэффициент расположения определяется как:

$$Place(U) = \begin{cases} 0, & \left(\frac{n}{n_{count}} > 0{,}9\right) \vee \left(\frac{n}{n_{count}} < 0{,}1\right) \\ 1, & \left(0{,}1 \le \frac{n}{n_{count}} < 0{,}3\right) \vee \left(0{,}7 < \frac{n}{n_{count}} <= 0{,}9\right) \\ 2, & \left(0{,}3 \le \frac{n}{n_{count}} \le 0{,}7\right) \end{cases}, \quad (2)$$

где n — номер предложения, а n_{count} — общее количество предложений в документе. Начало и конец текста оцениваются наименьшими значениями (поскольку это, преимущественно, вступление и выводы) 0 и 1, а основной текст — 2. Если документ содержит аннотацию, то ей присваивается $Place(U) = 4$.

Коэффициент форматирования предложения U определяется так:

$$Format(U) = \begin{cases} 0, & \text{выровнять слева или справа,} \\ 1, & \text{выровнять по ширине,} \\ 2, & \text{выровнять по центру.} \end{cases} \quad (3)$$

Коэффициент $WordsWeight(U)$ определяется как средний вес слова в предложении (сумма весов всех ключевых слов, входящих в предложение, которая разделена на количество ключевых слов в предложении). Таким образом, длинные предложения не имеют преимущества перед короткими.

Вес срока Q определяется по формуле:

$$Weight(Q) = Frequency(Q) + 1000 * Place(Q) + 10 * Format(Q) + 1000 * User(Q). \quad (4)$$

Частотный коэффициент $Frequency(Q)$ — отношение числа вхождения некоторого слова (word) к общему количеству слов (words) документа. Таким образом, оценивается значимость слова в пределах отдельного документа:

$$Frequency(Q) = \frac{word}{words}. \quad (5)$$

Коэффициент расположения $Place(Q)$ определяется как функция принадлежности к предложению, где встречается слово одной из ключевых фраз: «Ключевые слова:», «Key words:». Если такая фраза встречается, то коэффициент расположения равен 5, если слово находится в заголовке, то коэффициент равен 4, если в списке ключевых слов, присущих теме, — 3, если во внутренней ссылке — 2, если в основном тексте — 1.

Коэффициент форматирования слова $Format(Q)$ определяется в зависимости от того, как выделено слово: жирным, курсивом или подчеркнуто. Если слово совсем не отформатировано, то коэффициент равняется 0, если применен один формат, то — 1, если два — 2, если три — 3.

Показатель $User(Q)$ формируется на основе оценки слова пользователем. Этот показатель определяется как средневзвешенный вес всех весов, установленных пользователем.

Весовые коэффициенты из формул (1) и (4) получены эмпирически. В данном случае ставится задача не точного определения этих коэффициентов, а установления веса аддитивных параметров. Поэтому для этих коэффициентов важен порядок числа, а не его значение.

Входной информацией для отнесения публикации к научной школе является текстовый файл любого формата с содержимым публикации. Из файла необходимо определить базовые характеристики публикации:

- автор(ы) публикации (A);
- научное учреждение (B);
- тема публикации (C);
- ключевые слова (D);
- текст статьи.

Результатом метода выделения составляющих текстового документа является вектор, в котором для таких характеристик, как автор, научное учреждение, используются бинарные признаки, а для ключевых слов — веса.

Метод кластеризации публикаций по научным школам

Для группирования публикаций по направлениям, авторам и учреждениям разработан метод кластеризации публикаций. Результат метода — группа публикаций, которая и является научной школой.

Один из самых больших недостатков метода *k*-средних и ему подобных (например, fuzzy *c*-mean) — предварительное задание количества кластеров, от которого сильно зависит кластерное решение. Поэтому в работе решено модифицировать этот метод.

Модифицированный метод *k*-средних состоит из выполнения таких шагов [11].

1. Задаем количество кластеров k, $N >= k >= 2$, где N — количество публикаций.

Поскольку признаки кластеризации (автор, научное учреждение, название, ключевые слова) неотсортированы, используем метрику *d* изолированных точек:

$$l(X.x, Y.x) = \begin{cases} 1, & X.x = Y.x, \\ 0, & X.x \neq Y.x, \end{cases}$$

$$d(X, X_i) = \sum_{i}^{p} l(X.A_i, Y.A_i) + \sum_{j}^{r} l(X.D_j, Y.D_j) + \sum_{t}^{w} l(X.B_t, Y.B_t) + l(X.C, Y.C),$$

где функция l возвращает 1, если оба ее параметра имеют одинаковые значения, и 0 — в противном случае; X, Y — электронные версии текстов научных публикаций, p — количество авторов в текстах публикаций X, Y; r — суммарное количество ключевых слов; w — суммарное количество научных учреждений; $X.A_i$ — значение автора X_i публикации X; $X.C$ — значение названия C научной статьи X.

Изолированной точкой множества E является $x \in R^n$, если любая окрестность этой точки не содержит других точек E, кроме самой x:

$$d(x, y) = \begin{cases} 0, & x = y, \\ 1, & x \neq y. \end{cases}$$

Любая точка соприкосновения множества E является или предельной, или изолированной.

2. Выбираем k самых отдаленных объектов, которые будем считать центрами соответствующих кластеров (центроидами). Положим номер шага $t = 0$.

3. Формируем вектор центроидов $\langle cx_1^t, cx_2^t, ..., cx_k^t \rangle$.

Для каждого объекта находим расстояние ко всем центроидам. Для нахождения расстояния используем евклидовую метрику.

4. Ищем матрицу расстояний к центроидам кластеров:

$$\min \left[\sum_{j=1}^{k} \sum_{i}^{N} \| x_i - cx_j \|^2 \right],$$

где N — количество публикаций, cx_j — центроид кластера с номером j.

После расчетов матрицы расстояний ищем сильные связи объекта с кластером.

Определение 2. Сильной назовем связь между объектами X_i и X, если расстояние между названиями публикаций меньше трети максимального:

$$d_s(X, X_i) \leq \frac{\max d(X, X_i)}{3}.$$

5. Ищем стоимость разбивки:

$$Cost = \sum_{i=1}^{k} \sum_{j=1}^{S_i} d_{ij} d_s(x_j, cx_i),$$

где k — количество кластеров, S_i — количество объектов в кластере u, d_{ij} — расстояние к центру кластера u.

6. Ищем новые центроиды кластеров:

$$cx_i = \frac{1}{S_i} \sum_{x_j \in S_i} x_j.$$

Если $\| CX^t \| \neq \| CX \|$, то $t = t + 1$. Переходим к шагу 3.

7. Если $Cost$ не удовлетворяет условиям локального оптимума, k=k+1, переходим к шагу 3.

При наличии сильной связи применяется метод определения общих признаков в названии публикации. Для этого введено понятие меры расстояния между названиями. Признаки в названии статей назовем термами.

Пусть S — множество научных школ, выделенных из коллекции текстов Q. Множество E дополняется множеством значащих термов коллекции Q, которые выбираются в соответствии со следующим определением.

Пусть в коллекции Q n — общее количество термов во всех документах, n_i — количество термов в документах, в которых встречается терм i. Пусть общее число термов j во всех текстах — N_j, а количество термов j в документах, которые содержат терм u, — N_{ij}. Тогда величина

$$\rho_{ij} = \frac{\left(\frac{n_i}{n}\right)^{N_j}\left(1-\frac{n_i}{n}\right)^{N_j - N_{ij}} N_j!}{(N_j - N_{ij})!}$$

является мерой корреляции между термами i и j. Чем она меньше, тем больше коррелированы эти термы. Тогда сила связи термов i и j при $\rho = \max(\rho_{ij}, \rho_{ji})$ служит мерой корреляции термов i и j в случае $\rho_{ij} \neq \rho_{ji}$.

Терм t текстовой коллекции Q называется значащим (характерным) на уровне β, если различие между частотой, с которой терм t встречается в коллекции Q, и средней частотой, с которой он появляется во множестве научных публикаций, превышает β.

Для оценки качества кластеризации использовался показатель вероятности верной классификации и определено понятие ошибки классификации.

Основными результатами экспериментов являются величины ошибок, полученные в различных условиях.

Ошибки классификации бывают двух типов:

- ошибка первого рода — классификатор не заметил, что документ относится к текущему классу;
- ошибка второго рода — классификатор некорректно относит документ к текущему классу.

Выделены такие показатели:

- *TP* (*true positive*) — количество ЭД, правильно отнесенных к категории.
- *FP* (*false positive*) — ошибка второго рода — количество ЭД, неправильно отнесенных к категории.
- *FN* (*false negative*) — ошибка первого рода — количество ЭД, которые неправильно отброшены.
- *TN* (*true negative*) — количество ЭД, которые правильно отброшены.

Показатели *TP* и *TN* рассчитываются по формуле:

$$TP = Np - fn,$$

$$TN = Nn - fp,$$

где *Np* — количество «правильных» ЭД, а *Nn* — количество «неправильных» ЭД.

Дальше показатели *TP*, *TN*, *FN, FP* нормируются.

Классификация

Рассмотрим алгоритм классификации научных публикаций.

Для этого устанавливается релевантность определенного документа определенному классу (научной школе).

Шаг 1. Нормализация. Представляет собой способ уменьшения абсолютного значения веса индексных термов, выявленных в ЭД. Выбрана косинусная нормализация. При использовании этого метода нормализации вес каждого индексного терма делится на евклидову длину вектора оцениваемого документа.

Евклидова длина вектора определяется как

$$L = \sqrt{w_1{}^2 + w_2{}^2 + ... + w_{n^2}},$$

где $w_i = Weight(Q)$ — вес *i*-го терма (Q) в документе; определяется по формуле (4).

Рассмотрим формулу для вычисления веса (w) терма Q в документе с учетом косинусного фактора нормализации:

$$W = \frac{Weight(Q)}{\sqrt{w_1^2 + w_2^2 + ... + w_n^2}}.$$

Термы, отсутствующие в тексте документа, имеют нулевой вес. В запрашиваемом списке документы представлены в порядке уменьшения этого численного значения.

Шаг 2. Расчеты условных вероятностей. Для представления научных публикаций используется векторная модель, в которой любой документ характеризуется бинарным вектором $x = x_1, x_2, ..., x_n$, где $x_i = 0$ или 1, в зависимости от того, присутствует ли в тексте *i*-й индексный терм.

Рассматриваются два взаимно исключающихся события:

- w_1 — документ относится к научной школе x_i;
- w_2 — документ не относится к научной школе x_i.

Для каждой научной публикации вычислены условные вероятности $P(w_1 \mid x_i)$ и $P(w_2 \mid x_i)$, чтобы определить, какие документы относятся к определенной научной школе, а какие — нет:

$$P(w_i \mid x) = \frac{P(x \mid w_i) P(w_i)}{P(x)}, \quad i = 1, 2,$$

где $P(w_1)$ — первоначальная вероятность соответствия ($i = 1$) или несоответствия ($i=2$) запросу, величина $P(x \mid w_i)$ пропорциональна вероятности соответствия или несоответствия научной школы заданному x.

Шаг 3. Определение вероятности отнесения к классу. С этой целью используем теорему Байеса:

$$P(x) = \sum_{i=1}^{2} P(x \mid w_i) P(w_i),$$

где $P(x)$ — фактор, который нормализует $P(w_1 \mid x) + P(w_2 \mid x) = 1$.

Для определения релевантности документа определенной научной школе использовано правило: если $P(w_1 \mid x_i) > P(w_2 \mid x_i)$, то научная публикация принадлежит к научной школе x_i. Для множества научных школ определим вектор значений $P(w_1 \mid x_i)$.

Прогнозирование изменения динамики публикации

Следующий показатель формальной модели научной школы — прирост публикаций ΔSch и прирост количества защит докторских и кандидатских диссертаций $\Delta Degree$ представителями школы. Для вычисления ΔSch разработан метод тематического моделирования научных публикаций по научным школам. При этом анализируются не все слова ЭД (научной публикации), а только ключевые.

Вероятностная модель появления пары «школа–ключевое слово» представлена как

$$p(d,k)=\sum_{s\in S}p(s)\,p(k\mid s)\,p(d\mid s)=\sum_{s\in S}p(s)\,p(k\mid s)\,p(s\mid d)=\sum_{s\in S}p(k)\,p(s\mid k)\,p(d\mid s),$$

где S — множество школ; $p(s)$ — неизвестное априорное распределение школ во всей коллекции; $p(d)$ — априорное распределение на множестве научных публикаций, эмпирическая оценка $p(d)=n_d/n,$ где $n=\sum_d n_d$ — суммарная длина всех публикаций; $p(k)$ — априорное распределение на множестве ключевых слов, эмпирическая оценка $p(k)=n_k/n,$ где n_k — число вхождений ключевого слова k во все публикации.

Множество научных публикаций содержит для каждой публикации d дополнительную информацию, так называемую метаинформацию:

- список авторов публикации A;
- список публикаций $d',$ на которые ссылается d;
- список авторов A, на которых ссылается d;
- список публикаций, в которых ссылаются на d;
- список авторов, которые ссылаются на d;
- список научных школ, к которым относится d.

Искомые вероятности распределения $p\ (k|s)$, $p\ (s|d)$ выражаются как $p\ (s|k)$, $p(d|s)$ по формуле Байеса:

$$p(k\mid s)=\frac{p(s\mid k)p(k)}{\sum_{w'}p(s\mid k')p(k')};\ \ p(s\mid d)=\frac{p(d\mid s)p(s)}{\sum_{s'}p(d\mid s')p(s')},$$

где k', s' — список ключевых слов и научная школа соответственно, определенные из публикаций, на которые ссылается d.

Для идентификации параметров тематической модели (школы) по коллекции научных публикаций применяется принцип максимума правдоподобия, который приводит к задаче минимизации функционала:

$$\sum_{d\in D}\sum_{k\in d}n_{dk}\log p(d,k)\to\min,$$

$$\sum_k p(k\mid s)=1,\ \sum_s p(s\mid d)=1,\ \sum_s p(s)=1,$$

где n_{dk} — число вхождений ключевого слова k в публикацию d.

Прогнозирование изменения динамики публикации осуществлено с помощью временных рядов, а именно методом скользящего среднего. Задачей прогнозирования является нахождение зависимости между количеством публикаций по каждой из найденных научных школ, частотой появления новых ключевых слов и частотой получения научных степеней представителями школ. Динамика изменения количества ключевых слов зависит от базисного наблюдения и величины изменения соседних уровней.

В качестве статистических характеристик временного ряда $Y_i,\ \ i=\overline{1,n}$, использовано среднее арифметическое число публикаций $Y=\frac{1}{N}\sum_{j=1}^{N}Y_j$ и средний абсолютный прирост количества публикаций по школам $\Delta Sch=(Y_n-Y_1)/(N-1),$ где N — количество уровней ряда, Y_i — уровень ряда. Согласно методу проверки истинности различия средних, начальный временной ряд разбивается на две одинаковые части, после чего проверяется гипотеза о существенном различии средних для этих частей. Проверка однородности данных выполнена на основе критерия Ирвина, базирующегося на сравнении соседних значений ряда. Согласно этому критерию рассчитывается характеристика $ts=\frac{Y_t-Y_{t-1}}{\Delta Sch}$.

Анализ автокорреляции выполнен с помощью графика и критических значений коэффициентов, установленных экспертно. Параметры этого уравнения находят по методу наименьших квадратов. Среднее в выбранном интервале определено как взвешенное среднее всех предыдущих уровней. Метод наименьших квадратов использован также для поиска зависимости между приростом количества публикаций в научных школах по годам и приростом количества защит диссертационных работ $\Delta Degree$.

Для этого осуществлена загрузка файлов из сайта МОН Украины (приложения). Структура файлов определяется постоянным форматированием и состоит из таких компонентов:

- научная степень (доктор, кандидат);
- науки;
- учебное заведение (научное учреждение);
- специализированный ученый совет (не учитывается);
- фамилия, имя, отчество (ФИО), специальность (последняя характеристика не учитывается).

Критерием оптимизации принята минимизация суммы квадратов отклонений случайной величины от функции f:

$$\sum_{j=1}^{N} (\Delta Degree_j - f(\Delta Sch_j, t))^2 \to \min,$$

где j — номер научной школы, ΔSch — прирост публикаций в научной школе j за время t.

Анализ результатов

Разработана информационная система кластеризации научных публикаций. Построена архитектура, схема базы данных и основные программные модули. Программа состоит из таких модулей: база данных; подсистема графического представления; подсистема кластеризации научных статей по научным школам; подсистема определения весомости и скорости роста научной школы (рис. 1).

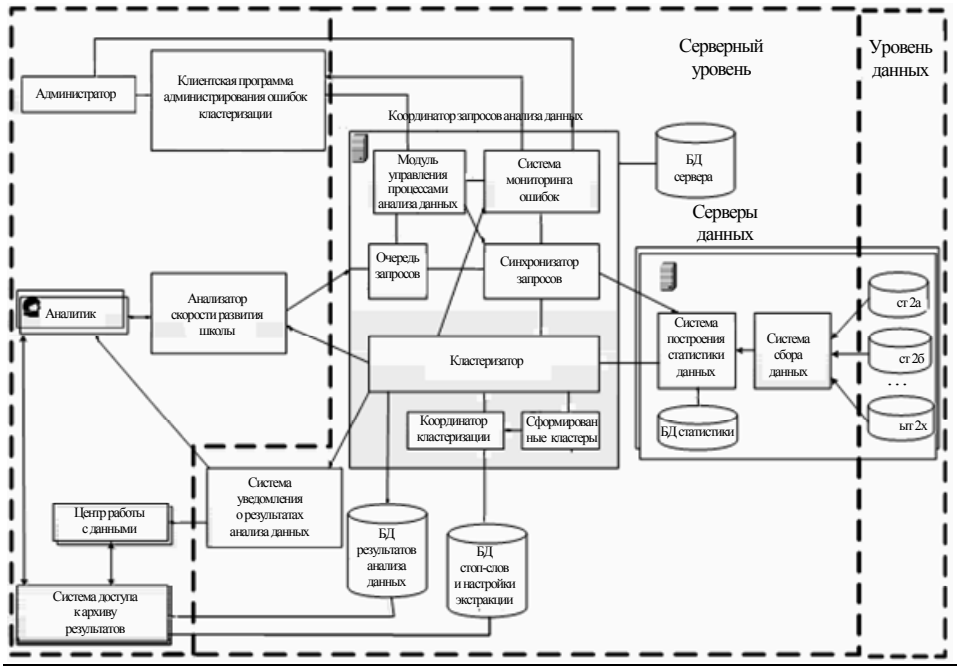

Рис. 1

Для тестирования работы системы обработаны 134 файла научных публикаций. В первую очередь анализировалась правильность кластеризации. Правильная рубрика текстовых документов известна заведомо и установлена экспертно (табл. 1). Здесь анализировались только ключевые слова (без учета авторов).

Среднее нормированное значение правильно рубрицированных документов составляет 94 %. Проанализирована зависимость качества кластеризации от объема публикаций в рубрике. Если рубрика универсальна, то ее трудно кластеризовать (рис. 2).

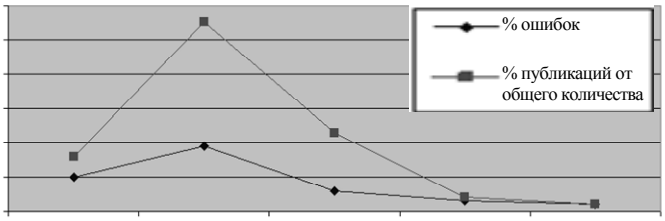

Рис. 2

Проанализировано качество кластеризации в зависимости от количества ключевых слов в каждой школе, а также от степени их пересечения. Алгоритм тестировался на четырех коллекциях входных данных с одинаковым количеством объектов в каждом классе, но с разным количеством ключевых слов и разным количеством общих для разных классов ключевых слов. Результаты анализа представлены в табл. 2.

Таблица 1

Класс	*ntp*, %	*nfp*, %	*nfn*, %	*ntn*, %
База данных	93	11	7	33
Информатика	93	13	7	25
Программирование	96	2	4	50
Сеть	94	6	6	60
Системный анализ	93	7	7	50

Таблица 2

Класс	Коллекция 1		Коллекция 2	
	К-во ключевых слов	*ntp*, %	К-во ключевых слов	*ntp*, %
База данных	7	87	16	88
Информатика	11	67	26	62
Программирование	12	69	19	67
Сеть	3	93	7	91
Системный анализ	4	94	5	89

Определено качество кластеризации для разных методов. Для сравнения проанализированы результаты работы трех алгоритмов на тех же коллекциях (табл. 3). Таким образом, разработанный алгоритм продемонстрировал лучшие результаты для *ntp* на текстовых коллекциях по сравнению с другими рассмотренными алгоритмами.

Далее проанализировано, действительно ли выделенные кластеры принадлежат научным школам. Для этого сравнивались множества публикаций, сформированные разработанным методом на основе анализа текстов и их кластеризации, и публикации

научных работников официально признанных научных школ. «Правильность» кластеров известна и оценена, так же как и качество рубрицирования. В отличие от рубрицирования, во время кластеризации учитываются также сведения об авторах публикаций. Анализировались статьи авторов, которые принадлежали разным научным школам (табл. 4).

Таблица 3

Метод кластеризации	*ntp*, %
Разработанный метод	92
Островная кластеризация	86
k-средних	71
Средней связи	78

Таблица 4

Публикации авторов разных научных школ, %	Ошибки, %
0	3
4	12
9	19
18	27

Для определения перспективности школы в течение трех лет анализировались файлы с информацией о защитах кандидатских и докторских диссертаций. Выполнен анализ вероятности появления новых публикаций в выделенных научных школах в зависимости от разных параметров. Спрогнозировано появление новых публикаций по школам (рис. 3).

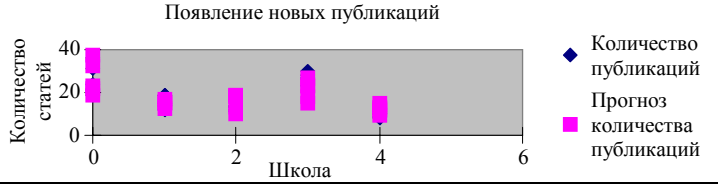

Рис. 3

В работе решена актуальная научная задача разработки математических методов и программных средств анализа текстов научных публикаций для выявления научных школ и исследования их результатов функционирования, что дает возможность повысить качество принятия решений относительно целесообразности поддержки научных исследований за счет выявления новых знаний в слабоструктурированных документах.

Н.Б. Шаховська, Р.Ю. Нога

МЕТОДИ ТА ЗАСОБИ АНАЛІЗУ ТЕКСТІВ ПУБЛІКАЦІЙ ДЛЯ ДОСЛІДЖЕННЯ ФУНКЦІОНУВАННЯ НАУКОВИХ ШКІЛ

Проаналізовано методи опрацювання текстової інформації з багатьох розрізнених інформаційних ресурсів. Удосконалено метод екстракції даних з наукової публікації, а також метод кластеризації *k*-середніх для поділу наукових статей за науковими школами. Визначено метрику якості кластерного рішення. Апробовано розроблені методи для електронної бібліотеки та наукової установи.

 ISSN 0572-2691

N.B. Shakhovska, R.Yu. Noha

METHODS AND TOOLS FOR TEXT ANALYSIS OF PUBLICATIONS TO STUDY THE FUNCTIONING OF SCIENTIFIC SCHOOLS

Thre are considered the methods of processing text information from a plurality of disparate information resources. The method of extraction of data from scientific publications is improved as well as the method of *k*-means clustering to subdivide research papers with respect to scientific schools. There is defined the quality metric of cluster solution. The developed methods were tested for e-libraries and for academic institutions.

1. *Захарчук Т.В.* Научные школы в библиографоведении: особенности формирования // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. — 2011. — № 1. — С. 19–25.
2. *Литвинова Л.А.* Наукові школи національної бібліотеки України ім. В.І. Вернадського в інформаційно-комунікаційному просторі України // Наукові праці Національної бібліотеки України імені В.І. Вернадського. — 2014. — Вип. 40. — С. 87–100.
3. *Широков В.А., Шевченко І.В., Рабулець О.Г.* Природномовна індексація як засіб вдосконалення пошукового апарату інформаційних систем // НТІ. — 2000. — № 3. — С. 23–25.
4. *Кунгурцев А.Б., Тыхан И.В.* Формирование онтологии на базе словаря предметной области // Реєстрація, зберігання і обробка даних. — 2014. — **16**, № 2. — С. 114–121.
5. *Данилюк І.Г.* Технологія автоматичного визначення тематики тексту // Лінгвістичні студії: Зб. наук. праць / Укл. А. Загнітко (наук. ред.) та ін. — Донецьк : ДонНУ, 2008. — Вип. 17. — С. 290–293.
6. *Larkey L.S., Croft W.B.*, Combining classifiers in text categorization // Proc. of SIGIR-96, 19th ACM International conference on research and development in information retrieval. — Zurich, CH, 1996. — P. 289–297.
7. *Erk K*. Vector space models of word meaning and phrase meaning: A survey // Language and linguistics compass. — 2012. — **6**, N 10. — P. 635–653.
8. *Alsabti K., Ranka S., Singh V.* An efficient k-means clustering algorithm / Proc. first workshop high performance data mining. — 1998. — P. 94–105.
9. *Дерецький В.О.*, *Богданова М.М., Ремарович С.С.* Підхід та засоби аналітичної обробки текстової інформації на основі агентної технології // Проблемы программирования. — 2002. — № 1–2. — С. 396–403.
10. *Киселев М.В.* Метод кластеризации текстов, основанный на попарной близости термов, характеризующих тексты, и его сравнение с метрическими методами кластеризации // Сб. работ участников конкурса Интернет-математика 2007. — Екатеринбург : Изд-во Урал. ун-та, 2007. — С. 74–83.
11. *Shakhovska N., Noha R.* One method of analysis of research publications' elements // MEST Journal. — 2014. — **2,** N 1. — P. 94–102; http://mest.meste.org/MEST_Najava/III_shakhovska.pdf
12. *Chappin E.J.L.*, *Ligtvoet A.* Transition and transformation: A bibliometric analysis of two scientific networks researching socio-technical change // Renewable and sustainable energy reviews. — 2014. — **30**. — P. 715–723.
13. *Ланде Д.В., Балагура І.В.* Наукометричні дослідження мереж співавторства по базі даних «Україніка наукова» // Реєстрація, зберігання і обробка даних. — 2012. — **14**, № 4. — С. 41–51.

Получено 25.03.2015
После доработки 04.06.2015