

ЭКОНОМИКА И ОРГАНИЗАЦИЯ НАУЧНОЙ И ХОЗЯЙСТВЕННОЙ ДЕЯТЕЛЬНОСТИ

ФОРМИРОВАНИЕ РАЦИОНАЛЬНОЙ ОРГАНИЗАЦИИ НАУЧНЫХ КОНКУРСОВ НА ОСНОВЕ ДАННЫХ ЛЕКСИЧЕСКОГО АНАЛИЗА ТЕКСТОВЫХ ИСТОЧНИКОВ

В.Л. Белоусов, В.О. Мелихов, Н.Е. Салькова

Работа посвящена описанию применения информационных технологий семантического распознавания текстов для целей формирования оптимальной организации научно-исследовательской деятельности на конкурсной основе. Количество квот победителей конкурсного отбора предлагается определять, сопоставляя тематическую направленность конкурсных проектов с составом государственного перечня научно-технологических приоритетов.

Ключевые слова: распознавание текстовых документов, научно-исследовательская деятельность, научные приоритеты.

В настоящее время различные приложения, связанные с семантическим распознаванием текстовых документов, занимают все более заметное место на рынке информационных технологий. Причиной тому становятся объективно складывающиеся условия, при которых из-за постоянного роста информационных ресурсов любой исследователь оказывается стеснен в своих возможностях как в плане получения полной информационной выборки, релевантной поставленному запросу; так и в плане времени, необходимого для анализа содержания уже найденных документов.

В частности, одной из проблем, для решения которых указанные тенденции развития современного информационного общества достаточно критичны, является актуализация национальных приоритетов в научно-технической сфере, имеющая принципиально важное значение для обеспечения целенаправленного развития отечественной науки и образования. На сегодняшний день наиболее массовым и эффективным механизмом реализации целей государственного управления и кадровой подготовки в научной сфере стала организация научно-исследовательской деятельности (НИД) на конкурсной основе. Характерным примером реализации такого механизма можно назвать Конкурс грантов Президента Российской Федерации для государственной поддержки молодых российских ученых и научных школ [3]. Однако, несмотря на широкое распространение конкурсной организации НИД, принципиально важным для управления научной отраслью является вопрос, связанный с определением квот победителей среди участников конкурса.

Пусть квоты тематически широкого научного конкурса определяются вектором $X = \{x_i, i = 1, \dots, N\}$, где N – количество разделов, по которым в конкурсе происходит независимый отбор победителей. В любом случае, даже если количество победителей конкурса *a priori* не оговаривается, то неявно эта величина ограничивается ценой конкурсной программы – D , а также величиной гранта d , предоставляемого победителям для проведения НИД. Такое условие при равномерном распределении грантов может быть описано соотношением

$$\sum_{i=1, \dots, N} x_i \leq D/d, \quad (1)$$

а при неравномерном – соотношением

$$\sum_{i=1, \dots, N} d_i \cdot x_i \leq D. \quad (2)$$

Вместе с тем, в плане реализации задач государственного управления и кадровой подготовки в сфере науки одинаковые по значению выражений (1) или (2) решения качественно могут существенно различаться. Таким образом, принимая в расчет *ориентацию конкурсной деятельности на решение приоритетных научных задач, эффективность организации конкурса следует понимать в смысле определения значений квот по его тематическим разделам в рамках установленной конкурсной стоимости.*

Не затрагивая в целом исследование сложнейшего вопроса по выделению закономерностей развития государственного научного и кадрового потенциала, рассмотрим ниже только некоторые из его аспектов, связанные с установлением квот в конкурсной организации.

Очевидно, что помимо непосредственного (прямого) исполнения проектов по тематике декларированных государственных приоритетных направлений и критических технологий, целевому развитию научно-технического и кадрового потенциала могут служить проекты, имеющие лишь частичное или косвенное отношение к приоритетной тематике. Например, к этому типу НИД относятся проекты фундаментальных исследований, результаты которых, как правило, не получают немедленного воплощения в составе приоритетных на сегодняшний день проблем, но важны для решения соответствующих задач в будущем. Более того, развитию всех составляющих научного потенциала (включая научные кадры) в определенной степени способствуют и науки, тематическое направление которых отсутствует в системе государственных приоритетов. Примером этому могут служить гуманитарные предметы, обуславливающие общественную выполнимость любого научно-технического развития, но не входящие в состав государственных научных приоритетов. Таким образом, в целях формализации целевых установок поставленной задачи следует считать, что тематическое соответствие конкурсных проектов проблематике приоритетной НИД может иметь частичный характер.

Заметим, что моментом, в определенной степени усложняющим понимание «тематического соответствия НИР приоритетным направлениям», становится также неопределенность содержания понятия научных приоритетов, так как на практике оно декларируется в максимально общей постановке, детализируется и приобретает конкретные черты часто лишь в процессе проведения конкретных разработок.

Все это указывает на то, что решение задачи следует искать в области смыслового сопоставления результатов исполняемых НИР с целевыми установками государственных научных приоритетов. Будем считать, аналогично работам [6], что содержание конкурсных проектов и содержание декларированных научных приоритетов может быть описано (проиндексировано) в терминах некоторой ограниченной вербальной системы. Например, в качестве вербальной системы, позволяющей индексировать содержание как отдельных проектов, так и групп проектов, выражающих описание некоторого раздела конкурса, могут служить области знания (ОЗ) из Государственного рубрикатора научно-технической информации (ГРНТИ) [1] или другие системы тематической кодификации (УДК, ББК), а также непосредственно сами списки Приоритетных направлений (ПН) и Критических технологий (КТ) [4].

Пусть используемая вербальная система включает Z классификационных признаков $\{p_1, p_2, \dots, p_Z\}$, характеризующих состав государственных научно-технологических приоритетов. Для определения тематического соответствия конкурса государственным приоритетным целям и задачам в сфере науки проведем по проектам-победителям в рамках установленной системы смысловую индексацию текстовых документов, характеризующих выполненные исследования.

В широком смысле получение показателей смысловой индексации относится к процедурам распознавания, общим моментом в которых является классификация свойств исследуемого объекта путем сопоставления его характеристик с эталонными образцами. Исходя из того, что выделяемый в текстовых документах проектов состав словоформ для тематически сходных текстов должен быть также сходен, определение смысловой (тематической) направленности анализируемых текстов может производиться на основе сопоставления словесной совокупности исходного анализируемого документа с классифицированной совокупностью словоформ, смысловое наполнение которой уже известно. В данном случае, для сопоставления тематической направленности конкурсных проектов с государственными целями развития научного потенциала по приоритетным направлениям, источником эталонной для смыслового распознавания информации стали результаты экспертного опроса (ЭО) по проблеме формирования Перечня приоритетов высшей школы, приведенные в работе [2] .

Пусть $S = \{s_j\}$ – список словоформ, полученных из текстов содержательных описаний Предложений в составе Перечня отраслевых научных приоритетов и классифицированных (экспертами) по системе классификационных признаков $P = \{p_1, p_2, \dots, p_m\}$. Словоформы, образованные из текстовых описаний Предложений в состав Перечня отраслевых приоритетов, формируются в тезаурус эталонной совокупности, обеспечивающий указание на принадлежность каждой известной словоформы $s \in S$ семантике некоторого классификационного признака p . При этом принадлежность словоформы классификационному признаку согласно [5] предлагается оценивать величиной *информационной значимости*

$$I_s^p = \frac{f_s^{s|p}}{f_s^S}, \quad (3)$$

понимаемой как отношение $f_s^{s|p}$ частоты употребления слов, соответствующих словоформе s в связи с классификационным признаком p , к общей частоте употребления таких слов f_s^S . Таким образом, тезаурус эталонной совокупности словоформ образуют «тройки» вида $\langle p, s, I_s^p \rangle$.

На основе исследуемой совокупности текстов отчетов по конкурсным проектам $\{T_1, T_2, \dots, T_k\}$ также формируется тезаурус словоформ, которые сопоставляются со словоформами эталонной совокупности, с одновременным накоплением для каждого из текстов суммарной информационной значимости по каждому из признаков. При этом уровень принадлежности текста T_k признаку p_i может оцениваться величинами

$$\mathfrak{Z}_{T_k}^{p_i} = \sum_{\forall t \in T, s=t} I_s^{p_i}, \quad (4)$$

объединение которых для всей исследуемой совокупности текстов позволяет (в нормализованной форме) оценивать долю каждого из классификационных признаков, т. е. характеризовать уровень принадлежности всей анализируемой совокупности по каждому из признаков.

Заметим, что в ходе проведения классификации тезаурус эталонной совокупности словоформ может быть расширен дополнением, состоящим из ранее неклассифицированной совокупности словоформ с соответствующей оценкой их информационной значимости.

$$\left\langle p_i, s_{new}, \frac{\sum_{T_k} \mathfrak{Z}_{T_k}^{p_i}}{\sum_{p_i} \mathfrak{Z}_{T_k}^{p_i}} \right\rangle. \quad (5)$$

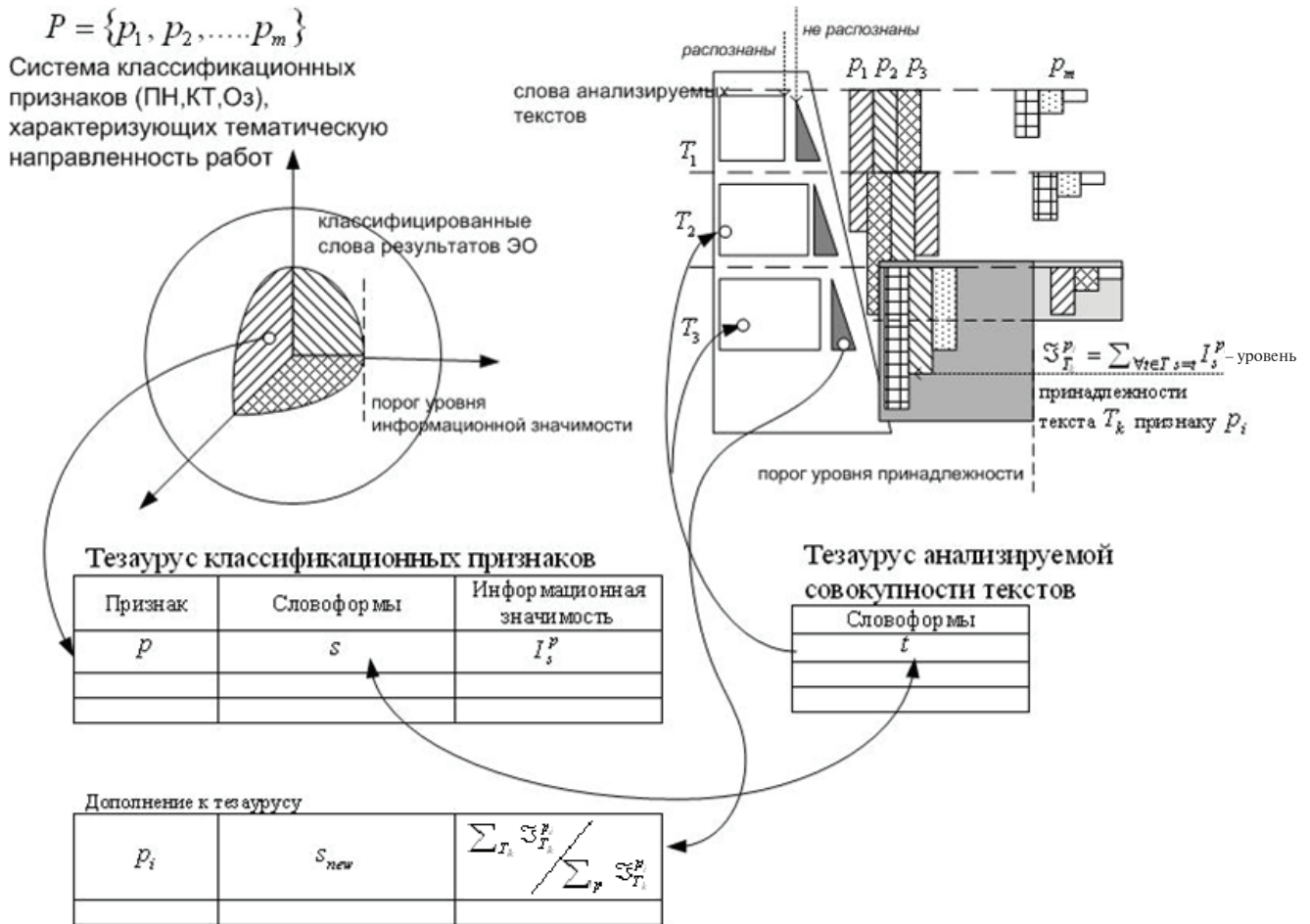


Рис. 1. Процесс лексического выделения тематической направленности работ

Таким образом, представленный механизм имеет свойства адаптивности, а его использование сопровождается расширением лексических возможностей инструмента.

Представленный на рис. 1 механизм позволяет установить для проектов – победителей конкурса способ сопоставления их содержания с приоритетами в научной сфере.

Полученные значения могут быть усреднены на любую группу НИР. Если для работ, относящихся к j -му разделу конкурса $j = 1, \dots, N$, установлено что уровень их принадлежности к тематике направлений p_i $i = 1, \dots, Z$ (включая и направление, не относящееся ни к одному из приоритетных) определяется нормированной величиной $c_{ij} = \mathfrak{Z}_{T_j}^{p_i}$, где

$$0 \leq c_{ij} \leq 1 \text{ и } \forall j = 1, \dots, N \sum_{i=1, \dots, Z} c_{ij} = 1, \quad (6)$$

то такой же величиной характеризуется и уровень достижений (например, в плане подготовки) исследователя, работающего по этому разделу. Тогда при известных значениях квот конкурса $X = \{x_j, j = 1, \dots, N\}$ совокупные достижения в рамках конкурса целей государственной научной политики по каждому из приоритетных направлений будут характеризоваться величиной

$$\forall i = 1, \dots, Z F(X) = \sum_{j=1, \dots, N} c_{ij} \cdot x_j, \quad (7)$$

для которой при ограничениях (1) или (2) можно говорить об оптимальном квотировании состава победителей.

Список литературы

1. **Государственный** рубрикатор научно-технической информации. 4-е изд. М.: ВИНТИ, 1992.
2. **Мелихов В.О.** Модели, методы и средства выделения научных приоритетов высшей школы. М.: Компания Спутник +, 2006. 223 с.
3. **Официальный** сайт Федерального государственного учреждения «Научно-исследовательский институт – Республиканский информационный научно-консультационный центр экспертизы (ФГУ НИИ РИНКЦЭ)». Режим доступа: <http://www.extech.ru/>.
4. **Приоритетные** направления развития науки, технологий и техники Российской Федерации (Пр-577 от 30 марта 2002 г.; Пр-843 от 21 марта 2006 г.).
5. **Салькова Н.Е.** Совершенствование механизмов формирования приоритетов в научно-технической сфере высшей школы. Автореф. дис. ... канд. экон. наук по специальности 08.00.05 Экономика и управление народным хозяйством (сфера услуг). На правах рукописи., М.: МГУПИ, 2007.
6. **Соловьева Н.С., Сомин Н.В.** Рубрицирование текстов как информационная технология. Системы и средства информатики. Вып. 11. М.: Наука, 2001.