

ОБ ОПИСАНИИ НАУЧНЫХ ОБЪЕКТОВ В ИНФОРМАЦИОННОЙ СИСТЕМЕ АНАЛИЗА НАУЧНОЙ ДЕЯТЕЛЬНОСТИ

Губанов Д.А., Чхартишвили А.Г.

Институт проблем управления им. В.А. Трапезникова РАН,
Москва, Россия

dmitry.a.g@gmail.com, sandro_ch@mail.ru

Аннотация. В докладе изложен метод описания научных объектов (публикаций, авторов, журналов, организаций, конференций), используемых в Информационной системе анализа научной деятельности (ИСАНД) в области теории управления. Показано, что эти объекты можно представить как точки в метрическом пространстве.

Ключевые слова: анализ научной деятельности, теория управления, информационная система, классификация, онтология, тематический профиль, тематическое пространство, метрическое пространство.

Введение

В последние десятилетия наблюдается стремительный рост объема научных публикаций, что обусловило необходимость создания компьютерных систем, способных автоматизировать работу с большими массивами научной информации. Такие системы должны включать как базы публикаций, так и инструменты для их регулярного пополнения и сопровождения. В некоторых случаях системы включают и дополнительный функционал – например, возможность запрашивать и получать тексты статей, задавать вопросы и отвечать на них. Набор аналитических возможностей в каждой системе определяется задачами, поставленными ее разработчиками. Наиболее известные платформы – Web of Science, Scopus, РИНЦ, Google Scholar, ResearchGate, OpenAlex и др. – ориентированы преимущественно на анализ цитируемости, на основе которого формируются наукометрические показатели публикаций, авторов (например, индекс Хирша), а также научных журналов (импакт-фактор).

Значительно более сложным направлением является анализ содержания научных текстов. Таких систем сравнительно немного. Среди них можно выделить американскую платформу Semantic Scholar (ориентированную на компьютерные науки и медицину), а также разработку НИУ ВШЭ – iFORA, охватывающую не только научные публикации, но и патенты, рыночные данные и др.

Информационная система анализа научной деятельности (ИСАНД), разрабатываемая в Институте проблем управления им. В. А. Трапезникова РАН, нацелена на анализ тематического содержания научных публикаций в области теории управления. Ключевой задачей здесь выступает позиционирование текстов в едином тематическом пространстве теории управления, что отличает классификатор ИСАНД от традиционных подходов (УДК, классификатор OECD, классификаторы Российского научного фонда и др.), в которых теория управления представлена в укрупненном виде.

Классификатор ИСАНД обладает многомерной структурой и основан на современных принципах онтологического моделирования. В его основу положена разработанная ранее трехмерная онтология наук об управлении (см. [1]). Наличие такой онтологии позволяет формировать тематические профили основных объектов научной деятельности – публикаций, авторов, журналов, организаций и конференций – в терминах тематического пространства. Эти профили становятся инструментом решения практических задач анализа. К примеру, исследователь может использовать систему для поиска релевантных публикаций; руководство научной организации – для подбора специалистов с определенными компетенциями; редакции журналов и организаторы конференций – для подбора рецензентов или участников.

1. Онтология научного знания в области теории управления

Онтология – это формальная спецификация согласованного описания (концептуализации) предметной области (по Т. Груберу [2]), разрабатываемая группой экспертов и интерпретируемая как машиной, так и человеком. Иными словами, онтология представляет собой формализованное описание согласованных экспертами понятий в определенной предметной области, разработанное для однозначного понимания как людьми, так и компьютерами.

Онтология научного знания предназначена для систематизации и классификации знаний в области теории управления. Предлагаемый классификатор ИСАНД (см. описание в работах [3, 4]) является «системой координат» тематического пространства, позволяющей реализовать взгляд на совокупность

научных направлений с определенной точки зрения, а также отразить возможную многотемность научного документа, журнала или конференции, многообразие научных интересов ученого или организации. Характеристикой объекта в этом пространстве является вектор, называемый тематическим профилем. Отметим, что классификатор ИСАНД частично отражен в публикации [5].

Онтология научного знания ИСАНД имеет четырехуровневую структуру, уровни которой (кроме нижнего) представляют собой дерево. Уровни нумеруются числами от 0 до 3. Нулевой уровень содержит вершины «Математический аппарат», «Предметная область», «Сфера применения». Предполагается, что при возможных корректировках онтологии в будущем этот уровень не изменится. Он отражает не конкретные темы теории управления, а различные аспекты научных исследований: математический аппарат, используемый в исследовании («Комбинаторика», «Теория оптимизации», ...), предметную область, т. е. некоторую прикладную теорию («Автоматическое управление», «Навигация и управление движением», ...), и конкретную сферу применения («Авиация», «Робототехника», ...). Вершины нулевого уровня будем называть *метафакторами*.

Каждая из вершин нулевого уровня является корнем тематического поддерева, раскрывающего ее содержание. Например, поддерево «Математический аппарат» содержит вершину «Теория оптимизации» (первый уровень) и детализирующие ее вершины второго уровня: «Выпуклая оптимизация» и др. Соответственно, поддерево «Предметная область» среди прочих содержит вершину «Автоматическое управление» и детализирующие ее вершины второго уровня, например, «Стохастические системы управления», а поддерево «Сфера применения» содержит вершины «Робототехника» (первый уровень) и «Роботы водного базирования» (второй уровень). Каждый фактор нижнего (второго) уровня характеризуется фиксированным набором терминов.

2. Описание научных объектов в ИСАНД

Основой для классификации научных объектов в системе ИСАНД является онтология научного знания теории управления, описанная ранее в п. 1. Она позволяет единообразно рассматривать отдельные научные объекты (публикации, ученых, организации, журналы, конференции). Характеристикой каждого из этих объектов в тематическом пространстве является вектор, называемый (тематическим) *профилем*.

Напомним (см. п. 1), что онтология теории управления имеет четырехуровневую структуру (метафакторы, факторы, подфакторы, термины). Уровни нумеруются числами от 0 до 3.

Множество вершин первого уровня, называемых факторами, обозначим через $V = \{v_1, \dots, v_n\}$. При этом i -я вершина первого уровня связана с множеством $V_i = \{v_{i1}, \dots, v_{in_i}\}$ вершин второго уровня – подфакторов. Обозначим через m общее число подфакторов: $m = \sum_{i \in N} n_i$.

Третий уровень – это вершины-термины, характеризующие подфакторы. Каждый термин, как правило, характеризует один подфактор (т. е. в некоторых случаях древовидность онтологии может нарушаться).

Приведем далее алгоритм расчета профилей научных объектов в соответствии с работами [3, 4]. Обозначим:

K – множество ученых;

L – множество публикаций;

Δ_{lij} – сумма числа вхождений в l -ю публикацию базовых терминов ij -го подфактора;

$$\omega(k, l) = \begin{cases} 1, & \text{если } k\text{-й ученый является автором } l\text{-ой публикации;} \\ 0, & \text{в противном случае;} \end{cases}$$

$r(l)$ – количество авторов l -ой публикации.

Определим *профиль второго уровня публикации* l :

$$x_l = (x_{l1}, \dots, x_{lij}, \dots, x_{lm}), \text{ где } x_{lij} = \frac{\Delta_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \Delta_{lij}}, \quad l \in L, \quad j \in N_i, \quad i \in N.$$

Очевидно, что этот вектор является стохастическим, т.е. $\sum_{i,j} x_{lij} = 1$.

Для нахождения *профиля первого уровня публикации* l просуммируем для каждого фактора значения компонент профиля второго уровня, отвечающих связанным с ним подфакторам:

$$X_i = (X_{i1}, \dots, X_{li}, \dots, X_{in}), \text{ где } X_{li} = \sum_{j \in N_i} x_{lij}, \quad l \in L, \quad i \in N,$$

где $X_{li} = \sum_{j \in N_i} x_{lij}$, $l \in L$, $i \in N$.

Наконец, для нахождения *профиля нулевого уровня публикации* просуммируем для каждой из трех вершин нулевого уровня значения компонент профиля первого уровня, отвечающих связанным с ней вершинам первого уровня.

Таким образом, каждая публикация характеризуется трехмерным вектором профиля нулевого уровня, n -мерным вектором профиля первого уровня, m -мерным вектором профиля второго уровня. Все три вектора являются стохастическими.

На основании профилей публикаций можно определить профили других научных объектов, связанных с публикациями.

На основе аддитивного принципа агрегирования определим *профили второго и первого уровня k -го ученого*, используя массив его публикаций

$$y_{ij}^k = \frac{\sum_{l \in L} \omega(k, l) \frac{x_{lij}}{r(l)}}{\sum_{i \in N} \sum_{j \in N_i} \sum_{l \in L} \omega(k, l) \frac{x_{lij}}{r(l)}},$$

$$k \in K, \quad j \in N_i, \quad i \in N,$$

$$Y_i^k = \sum_{j \in N_i} y_{ij}^k, \quad k \in K, \quad i \in N.$$

Профиль нулевого уровня определяется суммированием для каждой из трех вершин нулевого уровня значений компонент профиля первого уровня, отвечающих связанным с ней вершинам первого уровня.

Далее определим профили *журнала*, в котором опубликованы работы ученых. Пусть $U \subset L$ – множество работ, опубликованных в журнале $p \in P$, где P – множество журналов. Тогда профили определяются по формулам

$$w_{ij}^p = \frac{\sum_{l \in U} x_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \sum_{l \in U} x_{lij}}, \quad p \in P, \quad j \in N_i, \quad i \in N,$$

$$W_i^p = \sum_{j \in N_i} w_{ij}^p, \quad p \in P, \quad i \in N.$$

Размерности профилей журнала также равны m (для профиля второго уровня) и n (для профиля первого уровня).

Наряду с тематическими профилями важной характеристикой журнала является количество опубликованных в нем работ, т. е. количество элементов в множестве U .

Аналогично профилю журнала можно рассчитать профиль *научной конференции*.

3. Расстояние между научными объектами

Для анализа данных о научной деятельности часто требуется находить расстояние между научными объектами, которые отображены в виде стохастических векторов или, что, по сути, то же самое, точек на стандартном симплексе соответствующей размерности. Например, прикладная задача поиска рецензентов для заданной статьи связана с нахождением расстоянием между тематическим профилем текста статьи и тематическими профилями других статей или их авторов.

Предлагается использовать следующее расстояние между стохастическими векторами $\mathbf{x} = (x_1, \dots, x_k)$ и $\mathbf{y} = (y_1, \dots, y_k)$:

$$d(\mathbf{x}, \mathbf{y}) = 1 - \sum_{j=1}^k \min(x_j, y_j) = \frac{1}{2} \sum_{j=1}^k |x_j - y_j|. \quad (1)$$

Отметим, что справедливость второго равенства в (1) для стохастических векторов легко показать:

$$\begin{aligned} \frac{1}{2} \sum_{j=1}^k |x_j - y_j| &= \frac{1}{2} \sum_{j=1}^k (\max(x_j, y_j) - \min(x_j, y_j)) = \frac{1}{2} \left(2 - \sum_{j=1}^k \min(x_j, y_j) - \sum_{j=1}^k \min(x_j, y_j) \right) \\ &= 1 - \sum_{j=1}^k \min(x_j, y_j). \end{aligned}$$

Отметим (см. [3]), что в данной метрике расстояние между профилями первого уровня всегда не больше расстояния между профилями второго уровня и не меньше расстояния между профилями нулевого уровня тех же объектов.

Метрика (1) предполагает одинаковые (равные 1) расстояния между любыми двумя вершинами одного и того же уровня онтологии. Однако возможно обобщение этой метрики, в котором значения расстояний между вершинами являются заданными (из каких-либо соображений) и, вообще говоря, различными. В этом случае набор базисных стохастических векторов следует определить таким образом, чтобы попарные расстояния между ними совпадали с заданными значениями¹. Итак, пусть задан набор k базисных векторов $\{\mathbf{e}_j = (e_{j1}, \dots, e_{jm})\}_{j=1, \dots, k}$, каждый из которых соответствует одному фактору. Тогда научный объект $\mathbf{x} = \sum_{j=1}^k x_j \mathbf{e}_j$, $\sum_{j=1}^k x_j = 1$, в естественном базисе характеризуется m -компонентным стохастическим вектором $(\sum_{j=1}^k x_j e_{j1}, \dots, \sum_{j=1}^k x_j e_{jm})$. При этом расстояние между факторами \mathbf{x} и \mathbf{y} задается следующим выражением, являющимся обобщением (1):

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{l=1}^m \left| \sum_{j=1}^k e_{jl} (x_j - y_j) \right|. \quad (2)$$

Выражение (2) переходит в (1) при $k = m$, $e_{jl} = 1$ при $j = l$, $e_{jl} = 0$ при $j \neq l$ (т.е. когда базис является естественным).

Нетрудно убедиться, что расстояние (2) также является метрикой, т.е. для него выполняются аксиомы метрического пространства, в частности неравенство треугольника:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \frac{1}{2} \sum_{l=1}^m \left| \sum_{j=1}^k e_{jl} (x_j - y_j) \right| = \frac{1}{2} \sum_{l=1}^m \left| \sum_{j=1}^k e_{jl} (x_j - z_j + z_j - y_j) \right| \leq \\ &\leq \frac{1}{2} \sum_{l=1}^m \left| \sum_{j=1}^k e_{jl} (x_j - z_j) \right| + \frac{1}{2} \sum_{l=1}^m \left| \sum_{j=1}^k e_{jl} (z_j - y_j) \right| = d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}). \end{aligned}$$

Таким образом, предлагаемый метод расчета расстояний между научными объектами может использоваться для анализа данных о научной деятельности в различных тематических областях, в том числе тех, для которых попарные расстояния между вершинами онтологии на одном и том же уровне могут различаться.

4. Заключение

Разработка информационной системы ИСАНД, позволяющей автоматизировать решение научно-организационных задач в области теории управления (подбор экспертов и рецензентов с определенными компетенциями, поиск публикаций по определенной тематике, анализ тематики научного коллектива и ее эволюции и т.д.) требует построения различных моделей, как концептуальных (например, построение онтологии), так и математических. В данном докладе описана модель единообразного представления научных объектов (публикаций, авторов, организаций, журналов, конференций) в тематическом пространстве теории управления и предложен метод расчета расстояний между научными объектами для случая, когда попарные расстояния между вершинами онтологии на одном и том же уровне могут различаться.

¹ Отметим, что задача определения координат точек, для которых заданы попарные расстояния, имеет аналогии в одном из методов статистического анализа – многомерном шкалировании (см., напр., [6]).

Литература

1. Кузнецов О.П., Суховеров В.С. Онтологический подход к оценке тематики научного текста // *Онтология проектирования*. – 2016. – Т. 6., № 1. – С. 55–66.
2. Gruber T.R. A translation approach to portable ontology specifications // *Knowledge acquisition*. – 1993. – Vol. 5, № 2. – P. 199–220.
3. Губанов Д.А., Кузнецов О.П., Суховеров В.С., Чхартишвили А.Г. О построении профилей в тематическом пространстве теории управления // *Материалы 9-й Международной конференции «Знания-Онтологии-Теории» (ЗОНТ-2023)*. – Новосибирск, 2023. – С. 89–94.
4. Губанов Д.А., Новиков Д.А., Чхартишвили А.Г., Курако Е.А., Кузнецов О.П., Лемтюжникова Д.В. Информационная система анализа научной деятельности (ИСАНД) в области теории управления // *Проблемы управления*. 2024. – № 3. – С. 42–65.
5. *Теория управления: словарь системы основных понятий*. – М.: ЛЕНАНД, 2024. – 128 с.
6. Толстова Ю.Н. *Основы многомерного шкалирования*. – М.: КДУ, 2006. – 160 с.