DOI 10.25741 / 2413-287X-2019-04-2-060
УДК 004.912

ТЕХНОЛОГИЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ ДЛЯ СТРАТЕГИЧЕСКОЙ АНАЛИТИКИ ОТРАСЛИ

И. КУЗЬМИНОВ, канд. геогр. наук, **И. ЛОГИНОВА**, **П. ЛОБАНОВА**,
Национальный исследовательский университет «Высшая школа экономики»
E-mail: ikuzminov@hse.ru, iloginova@hse.ru, plobanova@hse.ru

Для эффективного управления научно-технологическим развитием российского АПК необходимо опережающее отслеживание существующего информационно-аналитического контекста приоритетных сфер сельского хозяйства. Возрастающая потребность в получении и использовании объективных аналитических данных, без которых невозможно принятие стратегических решений на различных уровнях, формирует необходимость интеграции прикладных инструментов аналитики в существующие аналитические системы. Такие инструменты разнообразны и основаны преимущественно на методах автоматизированного анализа данных. Статья иллюстрирует возможности интеллектуального анализа данных на примере системы текст-майнинговой аналитики.

Ключевые слова: *большие данные, семантический анализ, анализ естественного языка, отраслевые исследования, стратегическая аналитика, АПК, агропромышленный комплекс, отраслевые рынки, региональные рынки, комбикормовая промышленность.*

Effective management of scientific and technological advancement of Russian agricultural production requires the anticipating monitoring of the existing informational and analytic media in the top-priority spheres of the agriculture. Increasing necessity in the calculation and application of objective and reliable analytical data for the strategic decision making at different levels is forcing the integration of applied analytical tools into analytical systems. These tools are versatile and primarily based on the automatic data processing. The analytical system of text mining is presented as an example of intellectual data analysis and its opportunities.

Keywords: *large data arrays, semantic analysis, analysis of natural language, branch-wise research, strategic analytics, agricultural production, branch markets, regional markets, feed milling industry.*

Диверсификация продуктовых портфелей и поддерживающих их бизнес-процессов, усложнение технологических цепочек на всех уровнях управления и производства в условиях цифровизации экономики, новой научной и промышленной (Индустрия 4.0) революции выступает серьезной угрозой в вопросах эффективной реализации стратегического планирования. Хотя существующие информационные системы предлагают инструменты практически полного и всестороннего сбора различных типов информации, возникающей в результате роста частоты наступления важных событий, одновременно с этим стремительно сокращаются возможности традиционных методов быстрой структуризации потоков количественных и качественных данных и последующего отделения критически важных событий от случайных и малозначимых, растет спрос на технологии искусственного интеллекта, в том числе семантические технологии.

Требования к форматам аналитики, отражающим обобщенные результаты анализа собранных данных, также меняются: традиционные краткие аналитические отчеты становятся все менее востребованными, так как не отвечают задачам эффективного восприятия извлеченных из данных знаний для быстрого принятия управленческих решений на стратегическом уровне.

Новые задачи, с которыми сегодня сталкивается высший менеджмент, актуализируют потребность в планировании и прогнозировании, что формирует направления развития новых подходов к анализу данных, включая исследования будущего, форсайт, долгосрочное прогнозирование и другие. Эти форматы нельзя назвать до конца сформированными, но, опираясь на те задачи, в целях достижения которых они применяются, возможно заключить, что методологически такие форматы базируются на схожих инструментах, одним из которых является тренд-споттинг — мониторинг и обнаружение возникающих трендов [1].

Для рефлексии процессов, происходящих в области трансформации требований к аналитическим продуктам

и форматам аналитики, в статье авторами используется термин «стратегическая аналитика», который можно определить как новый зарождающийся вид аналитики, в основе которого лежат автоматизированный трендспоттинг и тренд-анализ как наиболее востребованные и перспективные аналитические инструменты, используемые в целях производства тех объективных и релевантных данных, которые помогают выстроить систему принятия стратегических решений на основе четкого видения процессов внешней микро- и макросреды, понимания того, какие аналитические категории являются приоритетными в вопросах отраслевого стратегического планирования.

Одним из инструментов автоматизации процессов трендспоттинга и тренд-анализа является текст-майнинг (или семантический анализ). Этим термином называют метод извлечения ценной информации из больших объемов неструктурированных текстовых данных с использованием методов машинного обучения, обработки естественного языка и управления знаниями [2]. Назначение семантического анализа состоит в извлечении из текстов смысловых сущностей различного уровня сложности, например, трендов, взаимосвязей, смыслообразующих категорий и других [3]. Прикладные направления применения автоматизированной текстовой аналитики обширны: ее используют как в науке, так и в бизнесе и государственном управлении.

На основе методов и алгоритмов семантического анализа в ИСИЭЗ НИУ ВШЭ была разработана система интеллектуального анализа больших текстовых данных iFORA, производящая на основе алгоритмов машинного обучения и компьютерной лингвистики обнаружение и оценку сложных смысловых концепций — трендов, технологий, рынков, продуктов, актуальных тематик научных исследований, то есть тех содержательных категорий и их взаимосвязей, которые могут оказывать существенное влияние на развитие исследуемых отраслей. Актуальная база данных iFORA состоит из около трехсот пятидесяти миллионов документов, включающих материалы рыночной аналитики и специализированных отраслевых СМИ, тексты научных публикаций, данные о патентах, что обеспечивает комплексный охват и релевантность получаемых аналитических результатов.

В настоящее время модули системы iFORA позволяют с высокой степенью точности не только выявлять из неструктурированных текстовых массивов тренды, оценки рынков, прогнозы, оценки рисков, но также и осуществлять бенчмаркинг, семантический анализ документов госполитики и корпоративных стратегий, региональный анализ, выявление центров компетенций, сетевой анализ, выявление профессиональных компетенций и навыков, анализ образовательных программ, анализ закупок, поддержку проектного управления и другое.

Надежность методологии iFORA подтверждена ее международной апробацией по итогам представления исследовательских результатов в рамках научно-практического воркшопа «Семантический анализ для целей инновационной политики», организатором которого выступила Организация экономического сотрудничества и развития (Париж, 2018 г.).

Методология iFORA сочетает в себе два основных алгоритма: выявление совстречаемости — совместного употребления терминов (тематик) в предложениях, а также векторное представление выявленных терминов — присваивание каждому термину уникального числового вектора, полученного путем анализа контекста употребления термина в текстах. Термины, употребляющиеся в схожем контексте, имеют схожие числовые векторы и близко расположены в векторном пространстве.

В рамках статьи для углубленного исследования комбикормовой промышленности применяются оба подхода, реализованные в одних из базовых инструментах системы iFORA.

Первый из них проводит кластеризацию (разделение на тематические категории) выявленных терминов. Результат такого анализа визуализируется в виде семантической карты. Семантические карты позволяют сделать выводы о важнейших научно-технологических тематиках рассматриваемой области, которые наиболее полно освещаются в документах базы данных. Узлами (точками) на семантических картах обозначены отдельные тематики и их наименования, а цветами — кластеры тематик (устойчивые группы направлений, объединенные общностью взаимосвязей).

Второй инструмент осуществляет подсчет частоты встречаемости терминов в текстах по годам. Результат данного анализа визуализируется в виде тренд-карты. Основное ее назначение — определение трендов и их классификация по масштабу и зрелости. Тренд-карта представляет собой поле, разделенное на четыре квадранта исходя из значений показателей значимости (частоты встречаемости слов) и динамичности (темпов роста частоты встречаемости). Тематики в правом верхнем квадранте тренд-карты являются зрелыми трендами, они имеют высокие показатели значимости и динамичности и задают актуальную повестку научно-технологического развития; эти тренды распространены в проанализированных источниках и устойчиво растут в популярности. Тематики в левом верхнем квадранте — это старые тренды: они определяют существующую структуру секторов и отраслей, широко распространены, но их популярность почти перестала расти в последние годы. Тематики в правом нижнем квадранте являются зарождающимися трендами, их значимость сравнительно невысока, но довольно быстро растет в последние годы; с высокой вероятностью они будут определять развитие в долгосрочной перспективе. Тематики в левом нижнем квадранте — это нишевые тренды, они не являются популярными и динамичными; их следует изучать для выявления тематических направлений, по тем или иным причинам недооцененных, но потенциально важных. Отслеживание

и классификация трендов — ключевой, но не единственный способ применения тренд-карт. Например, другими востребованными аналитическими срезами, которые могут быть получены с помощью тренд-карты, являются оценка динамики развития организаций, соотнесение параметров рынков и научно-технических заделов с целью определения направлений проектного финансирования и другое.

Для построения семантических и тренд-карт требуется определение порядка отображения терминов на картах. Для семантических карт зачастую характерно визуальное «наложение» терминов из-за невозможности проецирования на карту всех выделенных по результатам анализа тематик (количество которых может превышать несколько тысяч), что на пользовательском уровне затрудняет восприятие карт. В связи с этим, семантические карты визуализируют не все многообразие тематик, а часть из них, отобранную на основании приоритета той или иной статистической метрики, присваиваемой тематикам; в рамках статьи был выбран алгоритм подсчета показателя частоты встречаемости слов — в случае «наложения» терминов на карте визуализируется термин, характеризующийся большей частотой встречаемости в документах базы данных. Для работы с невизуализированными терминами системой iFORA дополнительно формируется табличный перечень всех релевантных для отрасли терминов. Таким образом, результаты исследования мировой комбикормовой промышленности, приведенные в статье, включают в себя не только семантические и тренд-карты, но и выдержку из таблицы терминов, необходимую для комплексной оценки результатов. Следует обратить внимание на то, что автоматизированная обработка текстовых данных задействует алгоритмы приведения терминов к их простейшей форме, поэтому как на картах, так и в таблицах написание некоторых терминов может не соответствовать нормам английского языка (например, «modify crop» вместо «modified crop»).

Семантическая карта направления «Комбикормовая промышленность» за 2012—2017 гг. (рис. 1) позволяет обнаружить крупнейшие тематики (технологии, продукты, услуги, перспективные научные направления и т.д.) и связанные с ними тренды, к числу которых относятся: органическая сертификация (organic certification), воздействие микотоксинов (mycotoxin effect), премиксы (feed premix), использование пестицидов (pesticide use), содержание белка (protein content), модифицированные культуры (modify crop) и другие, и их объединение по тематических группам (кластерам).

Тренд-карта направления «Комбикормовая промышленность» за 2012—2017 гг. (рис. 2) указывает, в частности, на то, что основными трендами (тематиками, расположенными в правом верхнем углу тренд-карты) являются молочная промышленность (dairy industry), защита сельскохозяйственных культур (crop protection), здоровье животных (animal health), мясная промышленность (meat industry), генномодифицированные сельскохозяйственные культуры (engineer crop, GMO crop), животноводство (livestock farming), генномодифициро-

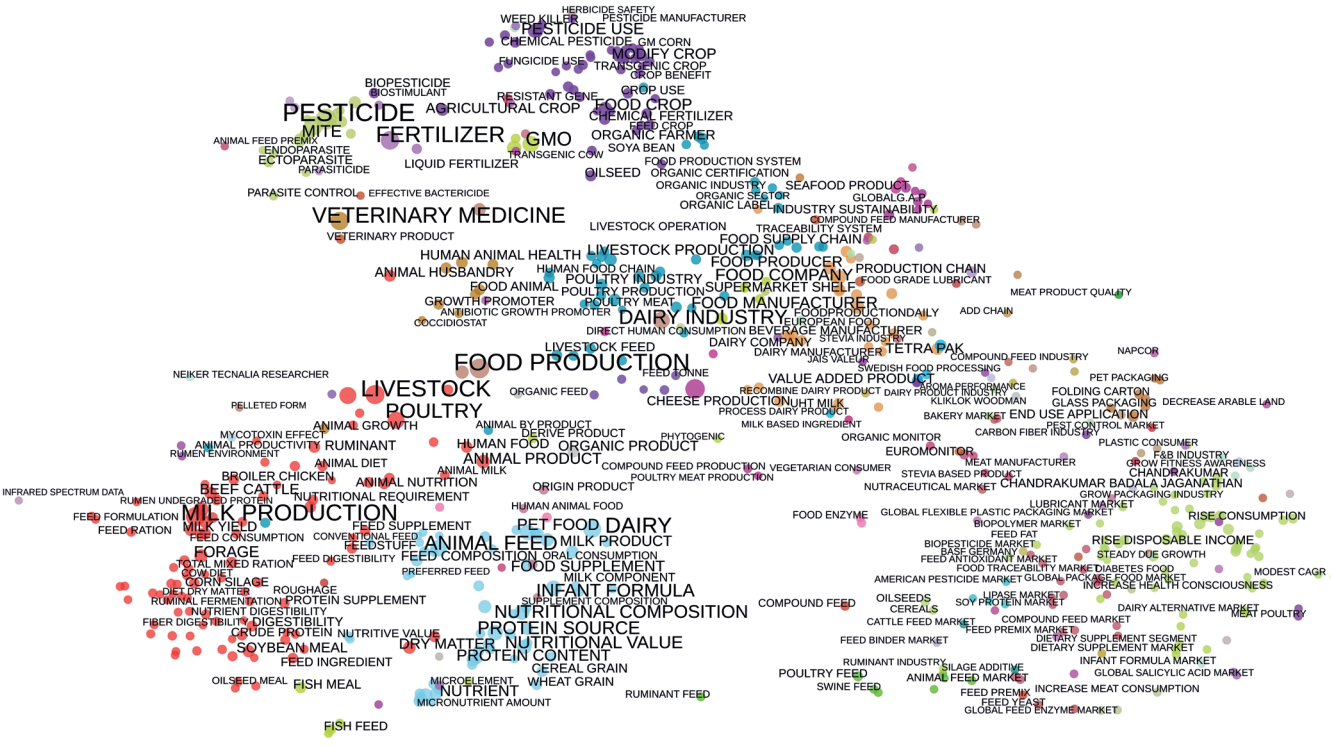

Рис. 1. Семантическая карта направления «Комбикормовая промышленность» за 2012—2017 гг.

(Источник: Система интеллектуального анализа больших данных iFORA; правообладатель — ИСИЭЗ НИУ ВШЭ)

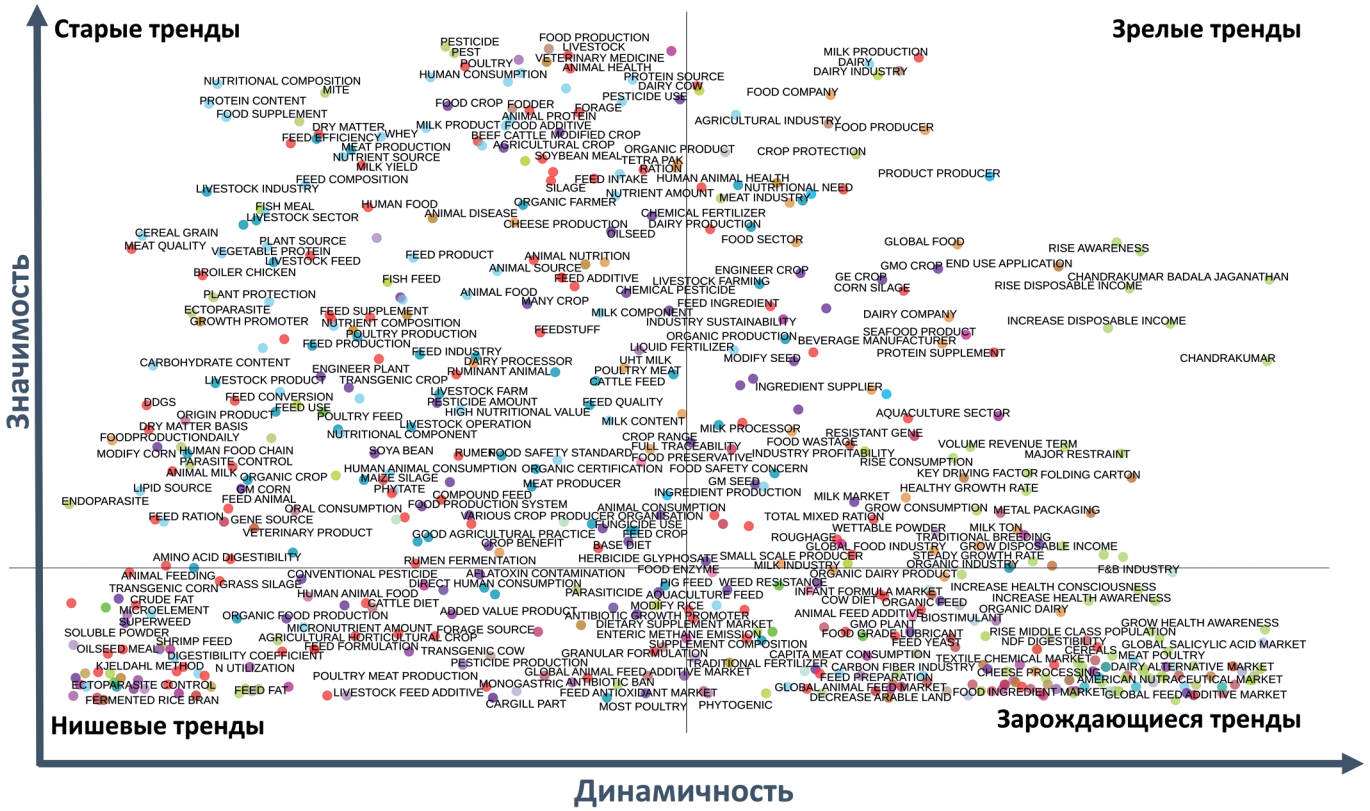

Рис. 2. Тренд-карта направления «Комбикормовая промышленность» за 2012—2017 гг.
(Источник: Система интеллектуального анализа больших данных iFORA; правообладатель — ИСИЭЗ НИУ ВШЭ)

ванные семена растений (modify seed, GM-seed), аквакультура (aquaculture sector), устойчивый ген (resistant gene), органическая промышленность (organic industry), поставка протеина (protein supplement), продовольственные отходы (food wastage).

Перечень терминов, ранжированных по критерию наибольшей частоты встречаемости, входящих в направление «Комбикормовая промышленность», обозначающих укрупненные, стратегически значимые тематики для отрасли, представлен в таблице.

Помимо построения семантических и тренд-карт, возможности системы iFORA включают в себя еще несколько десятков различных аналитических решений. Одно из них — матрицы для анализа совстречаемости терминов одной смысловой группы с терминами другой смысловой группы.

В качестве кейса, демонстрирующего работу этого модуля, в настоящей статье авторами осуществлена оценка совстречаемости (совместное упоминаемость) *компонентов, входящих в состав комбикормов для домашней птицы* (витамины, микроэлементы и другие), и *распространенных заболеваний птицы и их симптомов*. Выбор этого кейса обусловлен высокой значимостью тематики «домашняя птица» и расположением этого термина в правом верхнем квадранте тренд-карты, а также интересом к данной тематике со стороны научного сообщества [4, 5]. Построенная матрица представлена на рисунке 3.

Элементами матрицы являются располагающиеся на оси абсцисс наименования заболеваний или симптомов заболеваний домашней птицы, и на оси ординат — ключевые слова, относящиеся к компонентам комбикормов. Наличие круга на пересечении строки и столбца означает, что заболевание устойчиво упоминается совместно с определенным компонентом. Размер точки на пересечении отражает частоту совместной встречаемости. Отсутствие точки на пересечении указывает на то, что совстречаемость не превышает порога, установленного в системе iFORA, для идентификации устойчивых, статистически надежных закономерностей.

Характерной особенностью матриц является идентификация скрытых связей без определения направленности этих связей: например, компоненты комбикорма могут рассматриваться в контексте профилактики или лечения заболеваний, а могут и быть описаны в качестве причин ее развития.

Более того, в отдельных случаях совместная упоминаемость формируется научно-исследовательским, а не практико-ориентированным контекстом: отдельные текстовые источники (например, научные статьи) публикуют результаты исследования связи компонентов комбикорма с заболеваниями, и, даже в случае отсутствия

значимой связи по результатам научных экспериментов (например, [6]), совстречаемость все равно будет зафиксирована. В данном случае упоминаемость может свидетельствовать о научном интересе к связи компонента корма с болезнью, а значит требует повторного анализа в будущем. Описанная особенность работы с данным инструментом подчеркивает необходимость глубокого экспертного исследования для полноценной и корректной интерпретации матриц, а также совершенствования технологии за счет перехода от простой совстречаемости к высокоспецифичной (например, когда учитываются только факты совместной встречаемости с определенной структурой семантических связей, таких как «субъект — объект», «объект — действие», «объект — объект» и т.д.; либо когда рассчитываются отдельно совстречаемость в контекстах с позитивным сентиментом, то есть оценочной окраской суждения, и отдельно — с негативным сентиментом).

Несмотря на указанные ограничения, матрицы полезны для решения ряда актуальных для комбикормовой промышленности задач, в частности они могут использоваться для выявления слабых сигналов. Например, на матрице, представленной на рис. 3, таким сигналом выступает точка на пересечении железа и замедленной скорости роста. Рядом исследователей было показано, что добавление в корм домашней птицы железа в пропорциях 0,75 и 1,5 г на тонну корма ведет к значимому увеличению скорости роста [7].

Ранжированный по частоте встречаемости перечень терминов направления «Комбикормовая промышленность»

№	Термин	Перевод	Показатель суммарной встречаемости	Показатель прироста относительной встречаемости	Рейтинг термина по встречаемости и динамичности
1	Livestock	Домашний скот	5628	−0,38221	527
2	Milk production	Производство молока	5571	0,14914	689
3	Poultry	Домашняя птица	3724	−0,45003	474
4	Animal health	Здоровье животных	3648	−0,38220	522
5	Forage	Фураж	1230	−0,37427	506
6	Beef cattle	Мясной скот	870	−0,44215	446
7	Feed efficiency	Эффективность корма	865	−0,59308	358
8	Soybean meal	Соевая мука	697	−0,39803	468
9	Milk yield	Надои молока	641	−0,53504	381
10	Ruminant	Жвачные животные	579	−0,28169	529
11	Meat quality	Качество мяса	372	−0,80508	236
12	Digestibility	Удобоваримость	365	−0,57609	316
13	Broiler chicken	Цыплята-бройлеры	322	−0,66159	257
14	Feed additive	Кормовая добавка	321	−0,38776	423
15	Animal performance	Производительность животных	313	−0,37819	425
16	Corn silage	Кукурузный силос	300	0,09856	576
17	Feed supplement	Кормовая добавка	266	−0,55409	298
18	Crude protein	Сырой протеин	237	−0,59707	266
19	Protein supplement	Белковая добавка	222	0,63904	590
20	Feed conversion	Конверсия корма	178	−0,63344	215
21	Ddgs	Сухое сброженное зерно с растворимыми веществами	172	−0,77407	172
22	Milk replacer	Заменитель молока	164	0,11330	520
23	Nutrient requirement	Потребность в питательных веществах	159	−0,21943	441
24	Rumen	Рубец жвачных животных	134	−0,45157	294
25	Feed utilization	Затрата кормов	128	−0,63480	184
26	Nutrient digestibility	Переваримость питательных веществ	107	−0,50448	246
27	Total mixed ration	Общий смешанный рацион (метод TMR)	92	0,03510	458
28	Starch digestibility	Усвояемость крахмала	88	−0,44823	269
29	Rapeseed meal	Рапсовая мука	85	−0,21186	397
30	Neutral detergent fiber	Нейтрально расщепляемая клетчатка	71	−0,04408	432
31	Rumen fermentation	Пищеварение жвачных в рубце	66	−0,48822	225
32	Amino acid digestibility	Усвояемость аминокислот	65	−0,74260	108
33	Grass silage	Травяной силос	57	−0,64009	126
34	Milk fat content	Содержание жира в молоке	51	0,05445	427
35	Cattle diet	Рацион крупного рогатого скота	47	−0,53334	184

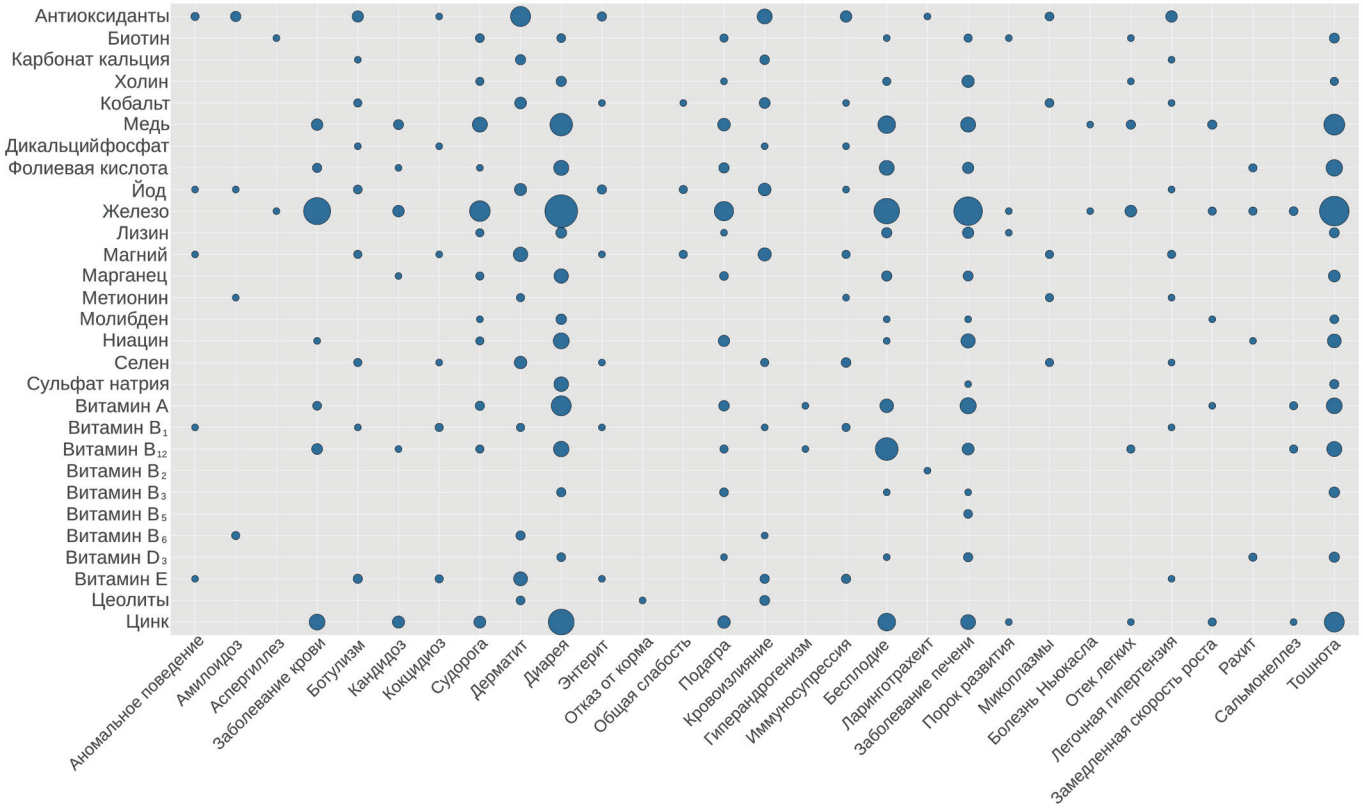

Рис. 3. Матрица совстречаемости компонентов, входящих в состав комбикормов, и распространенных заболеваний птицы и их симптомов

(Источник: Система интеллектуального анализа больших данных iFORA; правообладатель — ИСИЭЗ НИУ ВШЭ)

В качестве эффективной стратегии развития комбикормовой промышленности возможно внедрение управленческого подхода, основанного на непрерывном предиктивном мониторинге ключевых трендов, которые определят приоритеты развития отрасли. Такой мониторинг может быть реализован с помощью инструментов семантического анализа больших данных, востребованность которых в будущем будет расти в связи с масштабированием и усложнением существующего отраслевого и межотраслевого научно-технологического ландшафта. Ожидаемое ускоренное внедрение технологий и технологических решений, формирование новых перспективных научных направлений и рынков, рост структурной сложности существующих рынков актуализируют потребность в разработке интегрированного подхода к управлению, который будет включать не только экспертные методы, но и комплекс автоматизированных методов стратегического планирования и стратегической аналитики. Система интеллектуального анализа больших данных iFORA может выступить в качестве прикладного инструмента, обеспечивающего методическое, информационное и экспертно-аналитическое сопровождение принятия управленческих решений в области научно-технологической и инновационной политики в комбикормовой промышленности как в целях реализации государственных задач, так и в целях, приоритетных для агробизнеса.

Литература

1. *Andreassen, T. W.* Trend spotting and service innovation / T. W. Andreassen, L. Lervik-Olsen, G. Calabretta // Journal of Service Theory and Practice. — 2015. — Т. 25. — № 1. — P. 10–30.
2. *Berry, M. W.* Survey of text mining / M. W. Berry // Computing Reviews. — 2004. — Т. 45. — № 9. — P. 548.
3. A survey of emerging trend detection in textual data mining / A. Kontostathis [et al.] // Survey of text mining. — Springer, New York, NY, 2004. — P. 185–224.
4. EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). Scientific Opinion on the use of cobalt compounds as additives in animal nutrition // EFSA Journal. — 2009. — Т. 7. — № 12. — P. 1383.
5. *Surai, P. F.* Natural antioxidants in poultry nutrition: new developments / P. F. Surai // Proceedings of the 16th European symposium on poultry nutrition. — World Poultry Science Association, 2007. — P. 26–30.
6. Effects of copper, iron, zinc, and manganese supplementation in a corn and soybean meal diet on the growth performance, meat quality, and immune responses of broiler chickens / X. J. Yang [et al.] // Journal of Applied Poultry Research. — 2011. — Т. 20. — № 3. — P. 263–271.
7. Iron nanoparticles as a food additive for poultry / I. N. Nikonov [et al.] // Doklady Biological Sciences. — SP MAIK Nauka/ Interperiodica, 2011. — Т. 440. — № 1. — P. 328–331.