

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 6

Москва 2016

ОБЩИЙ РАЗДЕЛ

УДК 001.102 : 001.89

С.М. Оленев

Научная информация: современные гносеологические проблемы в контексте этоса науки

Рассматриваются современные задачи гносеологического анализа научной информации, научно-информационной и научно-исследовательской деятельности. Автор представляет дискуссионную точку зрения относительно процессов производства – потребления научной информации, идеологических и прикладных приоритетов науки. Предлагаются основания для оценки и систематизации научной информации в связи с ее праксеологическими характеристиками в русле актуальных задач реформирования научной инфраструктуры страны.

Ключевые слова: наука, научная информация, научное исследование, гносеология, классификация наук

Роль науки в жизни современного общества трудно переоценить. Фактически, любой самый тривиальный предмет домашнего обихода в той или иной степени предстает результатом применения наукоемких технологий, научные знания являются содержанием обучения с самых первых лет жизни современного человека, и авторитет науки – высок. Общество настолько свыклось с привыч-

ным представлением о естественном включении научного знания, научных исследований, научной инфраструктуры в его жизнь, что эффективность научного познания мира и точность репрезентации его гносеологического образа редко ставятся под сомнение. Однако научное знание неоднородно, как неоднородна и научная инфраструктура, а потому ценность и объективность научной информа-

ции исключительно сильно варьируются от одной дисциплинарной сферы к другой.

В современном науковедении частное понятие «научно-техническая информация» (далее – НТИ) применяется значительно чаще, чем общее – «научная информация» или, тем более, рядоположенные с НТИ понятия «естественнонаучная информация» или «социально-гуманитарная научная информация». Подобное игнорирование крупных сегментов научной информации в полной мере противоречит тому факту, что в обществе заметно шире, чем НТИ, распространена как раз социально-гуманитарная научная информация – явление, недостаточно осмысленное теоретически и не обеспеченное общепризнанным научным термином.

Научная информация обычно трактуется как логически организованная информация, полученная в результате научного познания, и отображающая явления и законы природы, общества и мышления [2]. Придерживаясь подобной трактовки, можно сказать, что явления и законы природы, общества и мышления признаются объектами научного познания, в процессе которого возникает научная информация. Это определение, с одной стороны, подчеркивает статус образа (понятие «образ» крайне полисемично и включено в различные философские концепции; так, в философской теории отражения диамата, весьма значимо различие между отражением и отображением), а с другой – выводит мышление из подсистемы средств в область объектов научного познания. При том, что мышление может и должно быть объектом научного исследования, его функциональную роль источника и средства научного познания мира, и как следствие – научной информации, нельзя отрицать. Именно это и вносит неоднозначность в гносеологическую функциональность научной информации. Но научная информация является неоднородным конгломератом не только по признаку ее гносеологического статуса, но и по своим праксеологическим характеристикам, а также по достоверности и ценностно-целевой значимости в структуре социальных процессов.

Учитывая указанные выше проблемы, можно констатировать необходимость исследования научной информации как центрального и основного функционального средства науки, обеспечивающего само существование научной деятельности и, что самое важное, – место науки в обществе, культуре и сознании людей.

Последние несколько столетий в европейском обществе непрерывно возрастала значимость науки, во многом вытеснившей религию в менталитете образованных людей. Но эта, казалось бы, прогрессивная тенденция, оказалась чреватой обратным явлением – приобретением наукой того же иррационального пиетета, которым до тех пор обладала религия. Пропагандистские усилия гуманитарных деятелей XVIII-XIX вв., убедивших широкую общественность в ценности просвещения и возвысивших человека науки над необразованной массой, сформировали трудно осознаваемую тенденцию отношения к науке как к социальной силе, превосходящей возможности отдельного индивида и заслуживающей не критического отношения. Уже к началу XX в. проблема этоса

науки перешла почти исключительно в нравственную плоскость, эффективно избавившись от двух своих важнейших составляющих – оценки достоверности любого научно-информационного субстрата, и выявления целевой направленности научного поиска. Фактически, научная инфраструктура утратила самокритику и возможность социально полезного управления своим развитием.

Безусловное приятие социальной ценности значительной части результатов научной деятельности распространилось на науку в целом. Научное сообщество в силу этого приобрело возможности суверенного существования, весьма слабо связанного с потребностями общества и культуры, и лишь в определенных точках контролируемого государством с помощью экономических и политических рычагов в тех аспектах, которые являлись и являются наиболее болезненными и/или ценными для групп политического или экономического влияния. Даже в тех государствах, которые де факто обладали и/или обладают жесткой системой администрирования, заметная автаркия научного сообщества сохранялась, что делало статус научного работника привлекательным именно в качестве защиты от социальных и политических перипетий. Это сформировало совершенно уникальное своеобразие научного сообщества, пользующегося привилегией независимого производства когнитивного материала и определения содержательных приоритетов производимой научной информации.

Наиболее заметным это положение стало в сфере социально-гуманитарной научной информации, в которой ценностно-целевые приоритеты ученых наиболее очевидно определяют результаты исследовательской работы. Фактически, идеология научно-исследовательской деятельности оказалась неподконтрольной действующей инфраструктуре управления наукой, но и не неподконтрольной кому-либо вообще: научное сообщество является неотъемлемой частью общества, и все идеологические тенденции, проявляющиеся вне научной среды, также оказывают свое воздействие и на нее. Поэтому говорить о беспристрастности, объективности и самостоятельности ученого как минимум наивно или нечестно – ученый является таким же точно проводником, адресатом и автором идеологического воздействия, как и любой член общества. Разница заключается только в том, что производимый ученым результат вызывает иррациональный пиетет у большинства.

Идеологическая составляющая научной информации относится к детерминантным факторам науки и производным от нее культурным явлениям не только в социально-гуманитарной сфере. Даже в таких областях, как НТИ и естественнонаучная информация, идеологические установки создателей когнитивного продукта определяют следующие характеристики:

- мировоззренческую позицию и средовые и/или индивидуальные вариации в содержании научной картины мира (которую в наши дни считается совершенно допустимым разбавлять религиозной эклектикой);
- области знания/незнания, обоснованной веры/неверия, а также игнорирования наличествующих знаний или имитации их наличия;

- представления об индивидуально-субъективной или социальной ценности данной области исследований;
- концептуальные установки исследователей;
- выбор объекта исследования;
- выбор качественно своеобразной системной методологии исследования и/или подмену системной методологической схемы имитацией методологии или хаотическим конгломератом методологических оснований;
- установку на определенный результат исследования;
- организационную форму проведения исследования, включая подбор исследовательского коллектива, источников финансирования (с их собственными целевыми установками), способов организации исследования;
- обоснование прикладной значимости результата исследования;
- форму представления результата исследования и способы доведения сведений о нем до заинтересованных лиц.

Как можно убедиться, один лишь идеологический фактор в науке уже сам по себе может определить высокую степень достоверности и ценности научной информации, или же свести их даже не к нулю, а к самоотрицанию. В сфере социально-гуманитарной научной информации это выражено значительно жестче, чем в НТИ. Как же обеспечить максимальную беспристрастность научного поиска, достоверность и ценность научной информации?

Безусловно, необходима выработка системы критериев оценки научной информации, позволяющая очистить знаниевый результат научной деятельности от тех трансформирующих влияний, которые обеспечивают становящийся все более заметным тренд нарастающей инфляции знания, создаваемого наукой. Учитывая тот факт, что наука неотделима от общества, подобная система критериев должна учитывать как структуру и таксономию видов научной информации, так и способы функционирования научной информации в обществе, ее роли, проявляющиеся в ценностно-целевой подсистеме научной деятельности.

Многочисленными исследователями информационных явлений – А.Д. Урсулом, К.К. Колиным и другими [3, 4], неоднократно приводились свойства информации как таковой – атрибутивные, прагматические и динамические. Научная информация является разновидностью информации по определению, но именно частный статус научной информации определяет неодинаковость значимости ее различных свойств. В настоящей статье не ставится задача исчерпывающего описания объекта, но декларируется значимость опоры именно на те свойства научно-информационного материала, которые в современной когнитивной ситуации являются критически важными. Отметим, что в большинстве своем они связаны с прагматическим и динамическим кластерами.

Намечая таксономию типов и видов научной информации, необходимо назвать основания деления, которые могут использоваться для классификации компонентов научной информации как системы.

Наиболее очевидным является такое основание деления, как качество объекта научного познания, определяющее существование естественных, технических и социально-гуманитарных наук, а также коррелирующее со всей дисциплинарной структурой научного знания. Соответственно, данное основание деления формирует три сегмента научной информации – социально-гуманитарный, научно-технический и естественнонаучный (включая медицинские науки), который можно было бы считать наименее подверженным искажениям, если бы не упомянутые выше идеологические факторы.

Затем, безусловно, необходимо привести такое основание деления как достоверность, определяющее существование собственно научной информации, квазинаучной, псевдонаучной, лженаучной и антинаучной информации. Конвенциональность научного знания, так хорошо изученная в русле философского науковедения, культурологии, социологии, политологии и даже семиотики, в значительной степени нивелирует возможности точного отнесения любого данного фрагмента научной информации к одному из названных типов. Фактически, исследование научной информации в связи с данным основанием деления чаще свидетельствует о конкретных идеологических установках лиц, осуществляющих анализ. Только применение такого критерия достоверности, как практика, способно снизить идеологическую зависимость суждений по данному поводу, но не исключить ее как таковую.

Важным основанием деления становится в наши дни аутентичность или исходность информации – в условиях дублирования исследований, недопустимых заимствований, массовых искажений и пр., это основание приобретает все большую значимость. Человек эпохи Античности, Средневековья или Возрождения, интенсивно обучаясь, мог овладеть если не всем доступным ему в свое время знаниевым багажом, то, по крайней мере, стать универсальным эрудитом. В наши дни изобилие информационных источников сделало это совершенно невозможным. Поэтому разрозненное и распределенное научное сообщество не просто многократно повторяет одни и те же исследования, но повторяет их на основе различных идеологических установок, делающих результаты исследований в чем-то отличающимися, а в чем-то сходными. Это определяет еще одну особенность научной информации – версифицированность. Материал, касающийся какой-либо одной темы, оказывается не стройным блоком сведений, раскрывающих данную тему, а чаще представляет собой широкий спектр мнений разной степени обоснованности, и требующих в силу этого, доксового анализа.

Некоторые части материала оказываются аутентичными, некоторые – производными, причем прямой зависимости прикладной ценности от аутентичности не наблюдается. В этом качестве неоднозначной становится проблема старения научной информации – только что произведенный материал может оказаться дублированием давно известного и широко распространенного знания, или даже – точным воспроизведением источника, ограниченная доступность которого обеспечила недостаточную распространенность

сведений о нем. Помимо этого, научная идея или разработка, казавшаяся устаревшей или неактуальной, может по-новому звучать с появлением инновационной концепции, меняющей локальные приоритеты научной дисциплины, а уж тем более – при наступлении очередной научной революции или смене моды.

Автор исходного материала может раскрыть тему полностью, сделав тем самым производные работы несущественными, а может совершить фундаментальную ошибку, обнаружить которую сможет только комментатор комментариев на комментариях. Немаловажным аспектом здесь оказывается и эргономика научного исследования – многократные и тем более неточные повторы одного и того же материала – пропорционально снижают эффективность науки: те же самые ресурсы могли бы быть брошены на «точки роста», но оказываются парализованы в многократном повторении пройденного. Однако нельзя сказать, что благополучие научного сообщества страдает от подобного положения дел: имитация движения вперед также может быть экономически эффективным занятием, обеспечивающим как отдельных ученых, так и целые коллективы.

Комплекс проблем, возникающих в связи с данным основанием деления, может быть разрешен только в условиях существования максимально практико-ориентированной метатеории, с высокой точностью определяющей области накопленного достоверного знания и сферы перспективных исследований, междисциплинарные и межтематические гиперссылки и сегменты структурного и содержательного тождества – такой метатеории на данный момент еще ни разу не было создано, ни в масштабах отдельной науки, ни в науке в целом. Неоднократно предпринимавшиеся попытки создать «теорию всего», «глобальный метанарратив», «теорию поля» и т.д., выдают бесплодность подобных умопостроений в условиях концептуальной и идеологической разнородности науки, а самое главное – подмены практики установками. Существующее, безусловно, может быть описано с минимальными неточностями (само существование структурированного нечто определяет возможность подбора аналогичной ему структуры), но только ориентир на практику при условии постоянной коррекции возникающих ошибок позволит сделать это эффективно. Данная проблема переводит настоящее исследование в новую (но кажущуюся хорошо исследованной) плоскость – плоскость дисциплинарной структуры научного знания и, соответственно, выражающей его научной информации.

Структура научной информации в идеале должна быть тождественна (или максимально подобна) структуре описываемого объекта – реальности. Но именно здесь и коренится наиболее тяжелая проблема научной информации и этоса науки: исследуемый объект далеко не всегда реален, а вместе с ним далеко не всегда реалистичны и методы его исследования и описания. Идеологи постмодернизма – М. Фуко, Р. Барт и пр. [1, 5], объясняли это особенностями языка как единственного средства и одновременно субъекта описания чего-либо. Согласиться с этим можно при условии распространения названного яв-

ления на эволюционно предшествующие языку психические процессы, сформировавшие бессознательную основу представлений, которые могут быть выражены речью: быть сознательным и говорящим можно только имея устойчивую опору в бессознательном. Другими словами, постоянная путаница между реально существующими и вымышленными объектами исследований основывается на проецировании бессознательного психического содержания не только на сам объект, но и на все многообразие средств его интерпретации. Стоит ли специально объяснять уязвимость и условность любой схемы дисциплинарного деления научной информации?

Но даже априори условная, дисциплинарная структура научной информации необходима, и без нее в принципе невозможно существование науки. Мы уже упоминали практико-ориентированную метатеорию, структура которой могла бы быть подобной предельному реальному объекту. Соответственно, такая метатеория уже определила бы реальные основания дисциплинарного деления научной информации и науки как таковой, но она же и потребовала бы четкого отделения объекта реального от вымышленного. На данный момент эта философская задача неразрешима никакими средствами, кроме последовательного самообмана, что, впрочем, не делает ее неразрешимой в перспективе.

В подобных условиях построение реалистической дисциплинарной структуры научной информации может быть охарактеризовано как стремление к объективности, т. е. не результат (с современным эпистемологическим инструментарием результат достигим ровно в той же мере, что и победа Ахиллеса над черепахой в апории Зенона), а процесс, способный каждой своей следующей фазой ближе подойти к созданию адекватного описания, создать более соразмерный объекту гносеологический образ. Если в эпоху Возрождения дисциплинарная структура наук и искусств была проста и ясна, то уже К. Маркс и Ф. Энгельс столкнулись с неоднозначностью решения проблемы дифференциации научного знания. Еще сложнее пришлось Б.М. Кедрову, работавшему над подобной задачей во второй половине XX века и Э.Р. Сукиасяну, строившему по модели Б.М. Кедрова библиотечно-библиографическую классификацию знаний [2]. На данный момент взрывообразный рост междисциплинарных областей и теоретических построений, претендующих на дисциплинарный статус, в полной мере хаотизировал предметную структуру научной информации. Но хаос – предтеча и источник следующей полуфазы порядка, естественно, отличающейся от предыдущей типом организованности.

Важнейшей гносеологической характеристикой научной информации является ее прикладная ценность. Трактовать ценность как применимость для решения какой-либо задачи недостаточно, хотя и такая трактовка уже добавляет характеристик каждому научно-информационному материалу: открывается возможность выявления прямых и вторичных целевых ориентиров каждого фрагмента знания. Любой фрагмент в потенциале способен решать какие-либо задачи и обеспечивать совершенно определенные направления в реализации научных исследований, в

связи с чем древний вопрос «*Qui prodest?*» встает вполне закономерно – он затрагивает ценностно-целевой, т. е. базовый аспект праксеологии научной информации. (Отметим только, что «*qui*» – кому, не всегда очевидное, значительно удобнее заменять на «*quid*» – чему: какой тенденции, влиянию, концепции, картине мира или заблуждению.)

Однако прикладная ценность измеряется еще и местом данного научно-информационного материала в процессе реализации научного исследования, его функциональностью в деле достижения научного результата. Ни для одного работника научной инфраструктуры не станет открытием тот факт, что в процессе осуществления научного исследования крайне редко удастся производить исключительно целесообразные действия, прямо приближающие субъекта к результату. Очень много исследовательских работ дают побочные эффекты, часть которых относится просто к «информационному шуму», а часть – открывает иные темы, непосредственно или косвенно связанные с исходной. В отличие от акына, баяна или барда, ученому не следует «растекаться мыслью (в оригинале иначе) по древу» и отказываться от контроля за своим познавательным процессом. Но на практике дело часто обстоит именно так, и особенно – в гуманитарной сфере.

В точном соответствии с представлениями Р. Барта о природе языка [1], дискурсивная практика приобретает самостоятельность и движет исследователем так же, как и процессом порождения новых фрагментов научного дискурса. Здесь мы сталкиваемся с проблемой, производной от недоработанности дисциплинарной структуры научной информации, – с хаотизмом представлений о предметно-содержательной структуре науки. Наука, сохраняющая статус науки, не может и не должна обладать признаками ризоморфизма, а значит, и не имеет права следовать свободному смыслообразованию, лишённому базиса в практике и в предметной структуре накопленного достоверного опыта. Редкие образцы классификации предметов исследований в рамках дисциплин, полагания их логических взаимоотношений, построения карты знаний, такие, как библиотечно-библиографическая или универсальная десятичная классификации, при всей их ценности, устаревают тем быстрее, чем более они детальны и проработанны. Однако подобные науковедческие ориентиры самим фактом своего существования уже конституируют как научную картину мира, так и структурно-методологические основания метатеорий, способных значительно рационализировать любые научные исследования. Остается только сожалеть о том, что подавляющее большинство «людей науки» имеет смутное представление о гносеологическом потенциале этих построений.

В связи со сказанным, праксеология научной информации остается загадочной областью, порождающей волюнтаристические декларации, обоснованные зачастую лишь личностными амбициями авторов. Проецируясь, например, в квалификационные признаки диссертаций, вариативная праксеология научной информации приобретает форму голословных утверждений о теоретической и практической значи-

мости работ, неясных положений, выносимых на защиту и т.д., и характеризуется отсутствием практико-ориентированной связи с потребностями обслуживаемого вида деятельности. В подобном контексте особенно обращает на себя внимание методологическая культура исследователей, все реже выбирающих адекватный инструментарий и стремящихся подогнать процесс исследования под желаемый результат.

Описанная гносеологическая ситуация, с одной стороны, может считаться неизбежным следствием роста массовой доли научно-исследовательской деятельности в структуре профессиональной деятельности населения развитых стран (что, безусловно, снижает качественные требования к отбору претендентов на вступление в ряды представителей научного сообщества), но, с другой, – не может быть признана препятствием непреодолимой силы, перед которым должно смириться все общество как создатель и потребитель продукта научной деятельности. Такие явления, как многократно дублирующиеся исследования, мелкотемье, распространение псевдонаучных идей, гипертрофия значимости неприменимого ни к чему материала, расходование бюджетных средств на имитацию науки, снижение качества профессиональной подготовки по наукоемким направлениям и специальностям, и многое другое – все это может и должно быть изжито с целью восстановления высокого статуса и ведущей роли науки в обществе. Как же должна производиться оценка качества научной информации?

Для этого, во-первых, кардинальной трансформации требует организационная инфраструктура производства, хранения и предоставления научной информации. Разрушение системы ГСНТИ, снижение роли отраслевых центров научной информации, стихийная передача значительной части научно-информационных функций «энтузиастам» Интернета, самостоятельным околонаучным сообществам и зарубежным информационным службам, как государственным, так и коммерческим, например, Thomson Reuters и пр., определили организационную невозможность централизации и систематизации массивов научной информации. Даже при условии наличия высокоэффективных методов упорядочения и оценки научной информации (что еще не достигнуто), подобное состояние исключает эффективное развитие науки и беспристрастное отношение к научно-информационным материалам. Решение этой проблемы – в руках высшей администрации страны, и остается только надеяться, что оно будет скорым.

Во-вторых, уже упомянутая ранее насущная потребность в формировании системы многокритериальной оценки научной информации, учитывающей ее идеологические, ценностно-целевые и праксеологические характеристики, должна реализоваться в простой и наглядной методологической схеме анализа любого научно-информационного материала, позволяющей беспристрастно ответить на вопросы о его концептуальном базисе, праксеологических приоритетах, направлениях научных исследований и социальных процессов, которым этот материал содействует, об уровне и качественных особенностях методологического обеспечения, с помощью которо-

го он получен, о соблюдении или нарушении внутренней логики материала, об его аутентичности или производности и пр. Это – непростая задача, но до ее решения современные науки, в особенности гуманитарные, так и останутся в состоянии укрепляющегося хаоса и экспоненциального роста доли наукообразных работ.

СПИСОК ЛИТЕРАТУРЫ

1. Барт Р. Избранные работы: Семиотика. Поэтика. – М.: Прогресс; Универс, 1994. – 616 с.
2. Кедров Б.М. Классификация наук : В 3 кн. / Академия общественных наук при ЦК КПСС. – М.: Изд-во ВПШ и АОН, 1961–1985.
3. Колин К.К. Философия информации и проблема формирования современного научного мировоззрения / Институт проблем информатики РАН // Вестник Международной академии наук (русская секция) . – 2013 . – № 1 . – С. 73–76.

4. Урсул А.Д. Природа информации: философский очерк / Челябинская государственная академия культуры и искусств; Российский государственный торгово-экономический университет. – 2-е изд. – Челябинск, 2010. – 231 с.
5. Фуко М. Археология знания / пер. с фр. М.Б. Раковой, А.Ю. Серебрянниковой; вступ. ст. А.С. Колесникова. – СПб: Гуманитарная Академия; Университетская книга, 2004. – 416 с.

Материал поступил в редакцию 19.04.16.

Сведения об авторе

ОЛЕНЕВ Святослав Михайлович – доктор философских наук, профессор, начальник научно-методического отдела ФГБОУ ВО «Московская государственная академия хореографии»
e-mail: nmo-mgah@yandex.ru

Информационное обеспечение: факторы развития, управление, эффективность

Рассматриваются макрофакторы, определяющие уровень информационного обеспечения сферы «наука-производство». Анализируются тенденции развития наиболее перспективных направлений российской информатики. Рассматривается методология и модель управления информационным обеспечением на этапах цикла исследование-разработка-производство (И–Р–П). Показано, что разработка типового проектного решения на базе модели позволит сократить временной лаг и качественно повысить интегральный уровень информационной поддержки всех этапов цикла И–Р–П. Сформулирован ряд предложений по преодолению тенденций инерционного развития национальной информационной системы.

Ключевые слова: информационное обеспечение, информационные технологии, цифровые информационные ресурсы, социальные научные сети, управление информационным обеспечением, информационный аутсорсинг, компьютерное моделирование, постобработка информации, эффективность информационного обеспечения, технологии Больших Данных, широкополосный интернет

ВВЕДЕНИЕ

Представление о перспективных (будущих) методах и новых технологиях информационного обеспечения часто является заложником противоречий между преемственностью и переменами. В основном внимание обращается на известное прошлое и игнорируются слабые тенденции, указывающие на будущее. Мы исходим из привычных понятий, неизменного характера процессов информационного обеспечения и понимания грядущих изменений (перспектив) как экстраполяции уже наметившихся процессов. Однако смена парадигмы в представлениях о месте и роли информации в современной постиндустриальной, инновационной экономике и возникновении новой экономической категории – национальные информационные ресурсы – требуют более глубокого понимания сущности процессов конвергенции человеческого потенциала, потенциала новых информационных технологий и цифровых информационных ресурсов.

Опережающий рост высокотехнологичных секторов экономики, развитие глобальной информационной среды, появление новых информационных технологий показывает, что ключевое значение для повышения эффективности информационного обеспечения сферы наука-производство имеют не отдельные направления, а прежде всего их комбинация (синергетическое взаимодействие), использование кластеров взаимоувязанных наиболее перспективных методов, средств и технологий. С некоторой долей осторожности можно утверждать, что это есть ос-

новной, неявный, тренд развития современной информатики.

Задачи повышения эффективности информационной поддержки исследований, разработок, трансфера технологий чрезвычайно актуализируются в современных сложных условиях российской экономики и стимулируют поиск новых подходов и инновационных решений.

В целом, несколько упрощая, можно сформулировать пять основных направлений комплекса работ и мероприятий, направленных на повышение уровня информационного обеспечения научно-промышленной сферы:

- разработка и адаптация новых технологий обработки (в т.ч. аналитической постобработки и моделирования) информации;
- развитие национальной информационной инфраструктуры;
- создание эффективных методов и средств управления процессами подготовки и реализации информационной поддержки цикла исследование-разработка-производство;
- минимизация и/или нейтрализация объективных факторов (детерминант) неэффективного использования научно-технической информации;
- ускоренное формирование национальных цифровых информационных ресурсов.

В последующих разделах рассматриваются факторы, прямо или опосредованно оказывающее значимое влияние на уровень (эффективность) информационного обеспечения работ на этапах цикла исследование-разработка-производство.

СОВРЕМЕННЫЕ ТРЕНДЫ РАЗВИТИЯ КАК ФАКТОР МОДЕРНИЗАЦИИ НАЦИОНАЛЬНОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Цифровые информационные ресурсы. По оценкам International Data Corporation (IDC), мировой объем информации удваивается каждые два года. В 2011 г. только текстовой информации появилось больше, чем за все существование человечества. По некоторым оценкам суммарный объем всех научных журналов в мире за один год составляет более 1 терабайта информации (1 Тбайт = 1024 Гбайт) [1]. Развитие и усложнение глобальной цифровой среды ведет к формированию целого поколения новых сервисов для работы с информацией – систем интеллектуального поиска, систем информационной навигации и мгновенного смыслового перевода.

В качестве значимых трендов для российской информатики можно выделить следующие:

- электронная каталогизация и оцифровка библиотечных и архивных фондов;
- доступные из Интернета полнотекстовые базы данных (с научной периодикой), научные электронные библиотеки и коллекции [2];
- базы данных результатов выполненных исследований и разработок научных фондов;
- электронные депозитарии промежуточных и заключительных отчетов по выполненным проектам федеральных научно-технических программ;
- рост популярности модели «открытого доступа» (Open Access) для публикации и распространения научной информации в интернет-пространстве.

Открытость позволит повысить в перспективе прозрачность науки, сократить нерациональные затраты, существенно снизить издержки от финансирования дублирующих исследований. В некоторых научных областях наблюдается удвоение цитирования публикаций при переходе к открытому доступу [3]. По плану Европейской комиссии в рамках 7-летней программы Horizon-2020 (бюджет \$80 млрд), уже к 2016 г. 60% всех публикаций европейских учёных, проводящих свои исследования за государственный счёт, будут размещаться в журналах открытого доступа [4]. В России реализуется пилотный проект КиберЛенинка, где по лицензионному договору размещено около 500 научных журналов открытого доступа. В списке имеющихся в eLIBRARY журналов открытого доступа – 3108 наименований. Проблема открытого доступа обсуждалась на Второй научно-практической конференции «Технология создания, агрегации и использования научного и образовательного контента», которая проходила в ноябре 2014 г. в ВЦ РАН. Открытый доступ позволит в значительной степени снять ценовые и правовые проблемы при обеспечении доступа к научной информации.

Развитие социальных научных сетей. Быстрое развитие цифровой информационной среды предопределило формирование и активное использование социальных сетевых информационных ресурсов, выполняющих функции преобразователя коммуникаций и распространителя разноплановой информации. Устойчивой тенденцией является развитие социальных научных сетей для повышения уровня информацион-

ного взаимодействия ключевых аудиторий в сегменте исследований, разработок, трансфера технологий. В последнее десятилетие увеличивается доля работ, прежде всего междисциплинарного характера, выполненных коллективами авторов, принадлежащих к различным научным и исследовательским организациям. В целом развитие современной сферы исследований и разработок во многом определяется возможностью общения и взаимодействия представителей этой сферы между собой, в первую очередь, для поиска коллег, соавторов, возможных рецензентов. Интернет дает уникальные возможности для быстрого и качественного информационного обмена в решении вопросов, возникающих при проведении исследований и разработок. К наиболее значимым российским научным сетям в RuNet следует отнести: а) **SciPeople.ru** (<http://scipeople.ru/>). При создании этой сети подразумевалось, что учёные будут выкладывать и обсуждать свои научные публикации, работать над совместными проектами. В настоящий момент аудитория сети составляет более 44 тыс. учёных; б) сеть «Учёные России» (**Russian-scientists.ru**) создана в 2010 г. по инициативе Российской академии естествознания (<http://russian-scientists.ru/>). Зарегистрировано свыше 13 тыс. пользователей из различных научных областей; в) сеть **Scientbook.com** создавалась как свободная информационная площадка для научного общения, где можно было представлять свои научные исследования, обмениваться знаниями, обсуждать научные публикации (<http://scientbook.com/>). Сеть запущена в 2012 г. Поддержка публикаций и общения осуществляется на русском и английском языках (<http://internet.cnews.ru/news/line/index.shtml?2012/03/06/480255>); г) **Machintbook** – специализированная социальная сеть для коммерческих и промышленных предприятий (около 12 тыс. пользователей). Это деловая сеть нового поколения, проводит линию объединения в единое сообщество промышленных предприятий различных отраслей и секторов экономики (<http://machintbook.ru>).

Основное направление развития научных социальных сетей – это совершенствование существующих и создание принципиально новых онлайн-сервисов, ориентированных непосредственно на ученых и разработчиков. Например визуализация химических структур молекул, инфракрасных спектров углеродных соединений, представление NMR-спектров и т. п. Большими возможностями обладают социальные сети нового поколения: как правило, они реализуются в формате 3D, предлагают на своей платформе аудио и видео чаты, онлайн-конференции, например, Youiverseworld (<http://yoniverseworld.ru>) [5].

Системы и технологии информационной поддержки наукоемкого производства

1. Федеральный портал по науке и инновациям (<http://www.sci-innov.ru/sci-dev/>). Портал обеспечивает навигацию и доступ к серверам информационных систем, хранящих полнотекстовую информацию о результатах исследований и разработок, в том числе, имеющих перспективу дальнейшей промышленной коммерческой реализации. Включает раздел под-

держки делового сотрудничества в области инновационных проектов, ориентированного на потенциальных инноваторов и инвесторов, заинтересованных в поиске партнеров и организации гибких и эффективных форм взаимодействия и сотрудничества для реализации перспективных идей и разработок. Концептуальным прототипом этого направления является система *CORDIS* – интерактивная информационная платформа в области европейских инноваций, исследований и разработок [6].

2. Разрабатываемая Государственная информационная система промышленности [7], содержащая информацию: а) об использовании ресурсосберегающих технологий и возобновляемых источников энергии в процессе промышленной деятельности; б) о прогнозе развития отраслей и субъектах деятельности в сфере промышленности; в) о прогнозах выпуска основных видов промышленной продукции, ее характеристиках, а также об объеме импорта промышленной продукции в Российскую Федерацию (по видам промышленной продукции); г) о государственных и муниципальных программах, разрабатываемых в целях формирования и реализации промышленной политики; д) об информационно-технических справочниках по наилучшим доступным технологиям и о методических рекомендациях по их применению.

3. Автоматизированная технология непрерывной информационной поддержки работ по этапам цикла исследование-разработка-производство (И–Р–П) [8, 9]. Базовой основой технологии является система автоматизированного проектирования (САПР) подготовки и реализации информационного обеспечения по выделенным этапам цикла И–Р–П. Реализация такой технологии повлечет переход от спонтанно-дискретного к управляемому процессу непрерывного информационного обеспечения всего цикла И–Р–П, в том числе в распределенной производственной среде [10], а также позволит на предпроектной стадии оценивать и осуществлять подготовку информационных (вычислительных, трудовых, финансовых ресурсов), необходимых для реализации информационной поддержки инновационного цикла (во взаимосвязи с актуализирующимися задачами и действующими производственными планами); в) сократит временной лаг и качественно повысит интегральный уровень информационной поддержки процессов создания новой наукоемкой продукции.

Информационный аутсорсинг. Термин аутсорсинг означает использование внешних ресурсов. Информационный аутсорсинг – это передача внешней организации, располагающей для этого необходимыми ресурсами (информационному центру), функций по информационной поддержке хозяйственной деятельности (научно-технической, производственной и др.) предприятия или организации на основе долгосрочных соглашений. Информационный аутсорсинг является развивающимся видом как качественного информационного обслуживания, так и оптимизации деятельности предприятий [11]. Для предприятия цель информационного аутсорсинга – использование актуальной проблемно-ориентированной научно-технической и технико-экономической информации для решения своих задач экономического развития

(при высоком качестве услуг и минимальных затратах). Широко известный режим ИРИ (избирательного распространения информации) является локальным прототипом данного вида аутсорсинга. Можно прогнозировать, что интерес предприятий и организаций к информационному аутсорсингу будет расти не только из-за возросших возможностей информационных центров, центров обработки данных, научно-технических библиотек, но и в силу объективных противоречий современной экономической среды между:

- «информационным насыщением» экономической среды, в которой функционируют предприятия, и реалиями «информационной недостаточности» с позиции доступности необходимой информации по конкретным задачам в конкретный период времени;
- возрастающей ролью информационных ресурсов для развития инновационной экономики и неразвитостью методов их эффективного использования, отсутствием высокотехнологичной информационной поддержки инновационной деятельности.

Во многом модель информационного аутсорсинга уже используется региональными центрами научно-технической информации. Применение ими технологии с использованием САПР [9] информационной поддержки позволит на качественно более высоком уровне осуществлять подготовку и реализацию информационного обеспечения предприятий и организаций.

Постобработка информации и компьютерное моделирование. Постобработка больших массивов научно-технической и технико-экономической информации (например, БНД ВИНТИ РАН содержит свыше 35 млн записей) с использованием статистических методов, методов анализа данных позволяет выявлять статистические закономерности, выражающие зависимости между распределениями различных параметров исследуемых систем и процессов и характер изменения распределений во времени. Как отмечалось в [12, 13], использование статистических методов при постобработке информации представляется весьма перспективным для решения ряда задач, в числе которых: анализ структуры отечественной и мировой науки; определение тенденций и процессов, происходящих в науке; выявление наиболее актуальных или, напротив, теряющих свою актуальность научных направлений; отслеживание генезиса конкретных научных идей и истории их развития; определение продуктивности работы исследователей (и эффективности материальных затрат) в конкретной научной области; анализ структуры научного сообщества и изучение науки как социального организма.

Под компьютерным моделированием в самом широком смысле понимаются процессы создания и исследования моделей: различных технологических процессов; химических реакций и биологических процессов; прогнозирования демографической и/или экологической ситуации в стране и мире; экономического развития предприятий, отрасли, страны и т.д. [14]. В медицинской информатике активно используются модели и методы автоформализации профессиональных знаний. Для поддержки научных исследова-

дований в химии, медицине, социологии используется ДСМ-метод – метод автоматического порождения гипотез (средство автоматизированной формализации знаний о предметной области на основе квазиаксиоматических теорий) [15, 16].

Следует отдельно остановиться на тенденциях, связанных с развитием технологий, объединяемых термином Big Data (Большие Данные). Это направление и сам термин появились примерно лет пять назад и стали следствием цифровизации и информационного взрыва последнего десятилетия (с выраженным экспоненциальным развитием). До сих пор нет общепотребительного определения для технологий Больших Данных, их отличительные особенности: а) обработка разнородной неструктурированной и структурированной информации (доля структурированной информации не превышает ~40%); б) высокая скорость поступления данных; в) высокая скорость обработки больших объемов информации в режиме реального времени. В обозримом будущем следует ожидать создания и широкого применения автоматизированных систем поддержки и принятия решений на основе Больших Данных с расширенными возможностями многомерного статистического анализа (что позволит находить корреляции между различными характеристиками, параметрами, событиями). По экспертным оценкам, развитие технологий Больших Данных ведет к появлению в скором времени распределенных самообучающихся систем когнитивных вычислений, качественно новых и эффективных методов прогнозирования научно-технических, инновационно-технологических, экономических и социальных процессов. Технологической основой Больших Данных являются высокоскоростные телекоммуникации и суперкомпьютерные технологии. В настоящее время общепризнано, что Большие Данные – это ведущий тренд экономического и технологического развития. В определенной мере технологии Big Data – это ответ на качественно новые задачи в промышленности и науке. Они обеспечивают создание новых, на порядки более эффективных, чем раньше, методов прогнозирования экономических, научно-технических и социальных процессов [1].

Развитие телекоммуникационной сети и сетевых сервисов. По данным компании Cisco, только мобильный трафик данных в мире в начале 2013 г. достиг ~900 петабайт/месяц (один петабайт = 10 в 15 степени байт). К 2016 г. доля пользователей, применяющих гаджеты для выхода в Интернет, составит ~48%, а ежегодный объем глобального IP-трафика – 1,3 зеттабайта (один зеттабайт = миллиарду гигабайт) [1]. Активно развивается широкополосный Интернет (≥ 1 Гбит/с). Количество абонентов широкополосного доступа в Интернет в США > 80 млн человек, в Китае ~100 млн человек, в России ~30 тыс. человек (по данным J'son&Partners Consulting на начало 2015 г.) [17]. Интернет2 – некоммерческий консорциум из 230 американских университетов, создающий передовые сетевые приложения и технологии [18]. Отличительные особенности Интернет2 – использование протокола передачи данных IPv6, а также использование высокоскоростных (>10 Гбит/с) магистральных каналов. Сеть GEANT – образова-

тельный проект мульти-гигабайтной сети, соединяющий более 3500 образовательных учреждений в Европе (более 30 стран). Скорость передачи данных от 1 до 12 Гбит/с. [19]. Пользователями Интернета в России являются 82 млн человек (66% населения); 50 млн пользователей выходят в Интернет с помощью мобильных устройств (55% – смартфоны, 41% – планшеты) [20]. С высокой степенью вероятности можно прогнозировать рост числа обращений с использованием мобильных устройств (и сервисов) к базам данных с научно-технической информацией.

Повышение эффективности информационного обеспечения. Постоянный рост удельного веса высокотехнологичных секторов экономики, интенсификация инновационной деятельности все более актуализируют вопрос эффективности информационного обеспечения. Вместе с тем, в некоторой степени парадоксальной является ситуация, когда при этом отсутствует сама постановка задачи управления информационным обеспечением (подавляюще превалирует постановка задачи информационного обеспечения управления). Однако при отсутствии управления (методологии, систем, подходов) и четко сформулированных критериев эффективности обсуждать проблему повышения эффективности информационного обеспечения, по меньшей мере, некорректно. Предметом и задачами управления информационным обеспечением, в общем случае, являются следующие:

- управление процессами подготовки и реализации информационного обеспечения;
- планирование комплекса функциональных задач информационного обеспечения, т.е. выявление и формализация потребностей в информации (в соответствии с конкретными задачами и спецификой НИИ, КБ, предприятия);
- проектирование комплекса компонент реализации по каждой задаче информационного обеспечения (источники, ресурсы, сроки предоставления, глубина ретроспективного поиска и т.д.);
- оценка уровня подготовки и реализации каждой задачи информационного обеспечения и/или всего комплекса задач по конкретному этапу работ (инновационного цикла);
- оценка эффективности планирования и расходования средств на реализацию информационного обеспечения;
- мониторинг отдельных параметров внешней информационной среды (с целью снижения затрат и повышения уровня информационного обеспечения);
- оценка эффекта мультипликативности результатов информационного обеспечения.

Далее рассмотрим общий подход и модель управления процессами подготовки и реализации информационного обеспечения работ по этапам цикла исследование-разработка-производство.

МОДЕЛЬ УПРАВЛЕНИЯ ИНФОРМАЦИОННЫМ ОБЕСПЕЧЕНИЕМ НА ЭТАПАХ ЦИКЛА ИССЛЕДОВАНИЕ-РАЗРАБОТКА-ПРОИЗВОДСТВО

Задачи организации и реализации обеспечения исследований и разработок научно-технической информации при создании новой продукции, в на-

стоящее время решаются не системно, а локально для каждой стадии цикла И–Р–П без должной увязки в рамках единой информационной технологии на основе качественно и количественно определенных целей, задач и ресурсных ограничений. Организационно-управленческие проблемы и задачи, связанные с повышением уровня информационного обеспечения и имеющие в своей организационно-технологической основе часто рациональные, но обособленные решения, не находят должного и эффективного воплощения в конкретной научно-информационной практике НИИ, инновационных центров, предприятий.

Актуальным является разработка и внедрение методов и технологий автоматизированного проектирования и управления всем комплексом работ по информационному обеспечению всех этапов цикла И–Р–П. Разработка и внедрение такой системы управления позволит: 1) обеспечивать полноту и своевременность предоставления необходимой научно-технической информации руководителям и специалистам на каждом из этапов цикла И–Р–П; 2) определять уровень и оптимизировать комплексное информационное обеспечение работ по этапам цикла И–Р–П; 3) оценивать размеры информационных, вычислительных, трудовых и финансовых ресурсов, необходимых для реализации информационной поддержки; 4) управлять процессами информационного обеспечения всего цикла этапов работ [8].

Методологическая основа новой технологии информационного обеспечения базируется на следующих основных положениях: автоматизированное проектирование и управление комплексным информационным обеспечением цикла И–Р–П; формирование системы целенаправленных функциональных задач информационного обеспечения (ИО) и

автоматизированного планирования их ресурсообеспечения для каждого этапа жизненного цикла изделия; количественная оценка уровня организации ИО отдельных этапов и всего цикла работ в целом. Функциональная задача информационного обеспечения ($З_j$) в общем случае включает в себя континуум единичных поисковых запросов (d_j), $З_j = \{d_j\}$, где $j = 1 \div m$. Решение задачи оценки уровня организации информационного обеспечения позволит выявить узкие места и «белые пятна» в процессах предоставления информации по каждому этапу инновационного цикла и вовремя принять действенные меры по их устранению. Детально решение этой задачи рассмотрено в отчете по НИР (ВИНИТИ, 2006 г.) [8] и в работе [9].

Базовой основой модели управления является матрица $n \times m$ (фрейм – в терминах систем искусственного интеллекта (СИИ)). Матричная модель является инвариантной и может быть представлена в графическом виде (см. табл. 1)

Элемент матрицы $\hat{a}_{ij}$ – комплекс взаимосвязанных компонент (проблемно-ориентированных характеристик организации информационного обеспечения), определяющих качество реализации j -й функциональной задачи ИО на i -м этапе цикла И–Р–П (R_{ij}). Элемент $\hat{a}_{ij} = \langle a_{ij} R_{ij} \rangle$ (слот, в терминах СИИ) является функционально зависимым от этапа жизненного цикла и функциональных задач информационного обеспечения, т.е.

$$\hat{a}_{ij} = \varphi(\Theta_i, Z_j), \quad (1)$$

где Z_j , например, задача предоставления информации по технико-экономическим параметрам отечественных и зарубежных аналогов разрабатываемого изделия.

Таблица 1

Функциональные задачи, j Этапы жизненного цикла изделий, i	$З_1$	$З_2$	$З_3$	...	$З_m$
Θ_1	$\langle a_{11} R_{11} \rangle$		$\langle a_{13} R_{13} \rangle$		$\langle a_{1m} R_{1m} \rangle$
Θ_2		$\langle a_{22} R_{22} \rangle$			
Θ_3			$\langle a_{33} R_{33} \rangle$		
...				$\langle a_{ij} R_{ij} \rangle$	
Θ_{n-1}	$\langle a_{n-1} R_{n-1} \rangle$				
Θ_n		$\langle a_{n2} R_{n2} \rangle$	$\langle a_{n3} R_{n3} \rangle$		$\langle a_{nm} R_{nm} \rangle$

Каждый элемент матрицы $\hat{a}_{ij} = \langle a_{ij} R_{ij} \rangle$ представляет собой гипотетический кортеж двоек (компонент-оценка реализации) вида

$$\hat{a}_{ij} = \langle (W, r_{ij}^W), (A, r_{ij}^A), (I, r_{ij}^I), (T, r_{ij}^T), (Q, r_{ij}^Q), (G, r_{ij}^G), (C, r_{ij}^C), (N, r_{ij}^N), (F, r_{ij}^F), (V, r_{ij}^V), (P, r_{ij}^P), (H, r_{ij}^H), (D, r_{ij}^D), (E, r_{ij}^E) \rangle, \quad (2)$$

где: W – перечень научно-производственных подразделений (специалистов, руководителей) участвующих в работах по i -му этапу; A – тематическая направленность информации (выражаемая в виде разделов рубрикаторов, УДК, профилей, классификаторов, вербальной модели); I – источники получения научной, технической, экономической и другой информации по интересующей проблематике; T – календарные сроки предоставления информации или время реакции информационной системы на заявленную информационную потребность; Q – ориентировочный объем информации по j -й функциональной задаче информационного обеспечения; G – требуемая (реализуемая) глубина ретроспективного поиска информации; C – оценочные затраты и источники финансирования работ по реализации j -й функциональной задачи информационного обеспечения; N – позиции номенклатуры информационной продукции (информационные карты НИОКР, периодические издания, обзоры, переводы, фрагменты БД (ЭБ) и т.д.); F – режимы поиска, формы и средства представления информации; V – аналитическая и логикоматематическая обработка информации (выявление неявных данных в информационном потоке, зависимостей и связей между характеристиками, экстраполирование и др.), агрегирование и сжатие; P – технические, материальные, людские, финансовые ресурсы, необходимые для реализации информационного обеспечения (по j -й задаче на i -м этапе); H – подразделение системы НТИ или информационные работники (инженеры-кураторы), ответственные за реализацию j -й задачи информационного обеспечения; D – характеристики моделей информационных потребностей специалистов и руководителей научно-производственных подразделений; E – адрес (электронные реквизиты) предоставления подготовленной информации; r_{ij}^l – требуемый уровень подготовки i -го компонента для реализации j -й функциональной задачи информационного обеспечения на i -м этапе.

В матрице, описанной в работе [9], макроалгоритм управления подготовкой и реализацией информационного обеспечения включает в себя: [определение этапов и работ цикла И–Р–П, требующих информационного обеспечения] ► [определение (выбор) необходимых для реализации информационной поддержки работ каждого i -го этапа функциональных задач ИО] ► [спецификация (номинация) и формирование требований к компонентам ($W, A, I, T, Q, G, C, N \dots$) элементов $\{\hat{a}_{ij}\}$ матрицы] ► [сопоставительная проблемно-ориентированная оценка сформированных спецификаций и требований с конкретными возможностями реализации] ► [априор-

ная оценка уровня организации решения j -й задачи информационного обеспечения – R_{ij} , и уровня организации информационного обеспечения i -го этапа – R_i] ► [расчет и планирование затрат на реализацию информационного обеспечения] ► [разработка и реализация мероприятий по достижению необходимого (максимально возможного) уровня организации информационного обеспечения i -го этапа] ► [апостериорная оценка R_{ij} , R_i ; интегральная оценка уровня подготовки и реализации выделенных этапов – R_Σ] ► [апостериорная оценка эффективности планирования и расходования средств Σ_i для реализации информационного обеспечения работ i -го этапа].

Проблемно-ориентированная оценка уровня организации информационного обеспечения заключается в определении (измерении) соответствия компонент – r_{ij}^l , требуемому уровню их реализации в заданный период времени. Оценки макрореlevантности (соответствия) R_{ij} в каждом отдельном случае формируются путем обобщения частных оценок соответствия выделенных компонент r_{ij}^l , $0 \leq r_{ij}^l \leq 1$, $l=1 \div n$. Обобщенный результат измерения $R_{ij} = 1$, когда все предпосылки для требуемой реализации компонент j -й функциональной задачи в интересах информационного обеспечения специалистов имеются ($0 \leq R_{ij} \leq 1$). Техника оценки r_{ij}^l и R_{ij} базируется на сборе информации и последовательном анализе: а) объективных (апостериорных) данных об опыте и результатах реализации j -й функциональной задачи ИО при решении различных научно-технических задач; б) результатов оценки специалистов; в) оценок, сделанных внешними экспертами. Наиболее конструктивным подходом при получении R_{ij} является использование метода обобщенного критерия (применяются различные функции свертки: мультипликативная, аддитивная, смешанная). В простейшем случае уровень организации решения j -й функциональной задачи ИО на i -м этапе может быть определен в результате свертки компонент r_{ij}^l в совокупную макрореlevантность R_{ij} как

$$R_{ij} = \left(\sum_{l=1}^n r_{ij}^l d_j^l \right) \cdot \left(\sum_{l=1}^n d_j^l \right)^{-1}, \quad (3)$$

где d_j^l – важность l -го компонента для решения j -й функциональной задачи на i -м этапе.

Для содержательной оценки R_{ij} используются логически идентичные шкалы оценки макрореlevантности, приведенные в таблице 2.

Оценки макрореlevантности дают возможность проводить сравнительный анализ уровня организации ИО работ различных этапов, выявлять изъяны и диспропорции в подготовленности отдельных задач этого обеспечения уже на предплановой стадии НИОКР, эффективнее использовать ресурсы информационных систем, планомерно и целенаправленно воздействовать на весь процесс организации и управления информационным обеспечением этапов работ цикла И–Р–П.

Уровень стратификации	Содержательная оценка уровня стратификации	Вербальное определение уровня соответствия R_{ij}	Интервал значения уровней организации
1	Отсутствие соответствия	Отсутствие информационной продукции по требуемой тематической направленности и отсутствие данных о ее наличии в отечественных и зарубежных центрах – источниках НТИ	(0; 0,2]
2	Низкий уровень соответствия	Орган НТИ, реализующий информационное обеспечение, не обладает требуемой информационной продукцией или ресурсами для ее восприятия, но имеет информацию о ее наличии в центрах – источниках НТИ. Решение j -й задачи информационного обеспечения в заданные сроки требует значительных ресурсных затрат	(0,2; 0,4]
3	Средний уровень соответствия	Уровень соответствия компонент a_{ij} имеет предпосылки решения j -й задачи информационного обеспечения по i -му этапу. Для обеспечения более полного соответствия необходим приемлемый объем дополнительных ресурсов	(0,4; 0,7]
4.	Высокий уровень соответствия	Фактическое состояние компонент a_{ij} соответствует условиям решения j -й задачи информационного обеспечения i -го этапа и может быть обеспечено повышения уровня соответствия по мере выполнения работ	(0,7; 0,9]
5	Полное соответствие	Состояние компонент a_{ij} достаточно для решения j -й задачи информационного обеспечения i -го этапа в заданные сроки	(0,9; 1,0]

Уровень подготовки и реализации информационного обеспечения i -го этапа цикла И–Р–П определяется на основе свертки R_{ij}

$$R_i = \frac{1}{L_i} \cdot \sum_{j=1}^{L_i} \lambda_j \cdot R_{ij}, \quad (4)$$

где λ_j – коэффициент относительной важности j -й функциональной задачи ИО для реализации работ i -го этапа, $\lambda_j = 0 \div 1$; L_i – число задач информационного обеспечения на i -м этапе цикла И–Р–П.

Используя (3), можно получить апостериорную оценку эффективности планирования и расходования средств Θ_i для реализации информационного обеспечения работ i -го этапа

$$\Theta_i = R_i \cdot Z'_i, \quad (5)$$

где

$$Z'_i = \frac{100\% \cdot Z_i}{Z_i^{\max}} - \text{реальные затраты (нормиро-}$$

ванные),

$Z_i^{\max}$ – максимально допустимые (запланированные) затраты по i -му этапу, $Z' \in [0; 100\%$.

С учетом (3) и (4) и временных ограничений в качестве критерия оптимизации информационного обеспечения работ i -го этапа может быть выбран,

$$\begin{cases} \sum_{i=1}^{L_i} R_{ij} \rightarrow \max; \\ \sum_{i=1}^{L_i} Z_i \rightarrow \max. \end{cases} \quad (6)$$

Общая интегральная оценка R_{Σ} уровня организации информационного обеспечения всех этапов K (блоков работ) цикла И–Р–П может быть представлена как

$$R_{\Sigma} = \sum_{i=1}^{k-1} \int_i^{i+1} f(\mu_i \cdot R_i) dR_i, \quad (7)$$

где $i=1, K \geq 1$; μ_i – убывающая функция вида

$$\mu_i = \frac{M}{K} - i^x,$$

где

$$M = 1 + \sum_{i=1}^K i^x.$$

Коэффициент μ введен с учетом того, что каждый последующий этап использует информацию и зависит от результатов предыдущего, таким образом можно априори считать важность предыдущего этапа большей по отношению к последующему, т.е. $\mu_1 > \mu_2 > \mu_3 \dots$.

Учет затрат и распределения дополнительных ресурсов на увеличение макрореlevантности проблемно-ориентированных характеристик организации информационного обеспечения определяется как

$$Zoi = \int_{t^1}^{t^2} [ZT_i(t) + Z'D_i(t)] dt, \quad (8)$$

где Zoi – суммарные затраты на организацию информационного обеспечения i -го этапа в интервале времени $t^1 \div t^2$;

$ZT_i(t)$, $Z'D_i(t)$ – исходные и дополнительные затраты на организацию комплексного информационного обеспечения i -го этапа цикла И–Р–П:

$$ZT_i(t) = \sum_{j=1}^{L_i} ZT_{ij}(t), \quad Z'D_i(t) = \sum_{j=1}^{L_i} Z'D_{ij}(t), \quad (9)$$

где $ZT_{ij}(t)$, $Z'D_{ij}(t)$ – исходные и дополнительные затраты на реализацию j -й функциональной задачи информационного обеспечения на i -м этапе.

Теоретический и практический интерес представляет решение задачи количественной оценки экономической эффективности информационного обеспечения. Однако постановка и решение такого рода задачи содержит значительные методологические трудности объективного характера. Основными факторами являются: а) значительный временной лаг получения конечных результатов; б) непрогнозируемый уровень получаемых доходов на вложенный капитал; в) неопределенность конъюнктуры рынка на период финальной стадии внедрения; г) неопределенность с ценообразованием и востребованностью новой продукции; д) множеством привходящих факторов на этапах цикла И–Р–П экономического, технического, организационного характера; е) отсутствие разработанной и апробированной методологии.

Проблема количественной оценки экономической эффективности информационного обеспечения особенно сложна в сфере инновационной деятельности и венчурного финансирования. По американской статистике, только один из восьми инновационных проектов приносит прибыль. В целом проблема оценки экономической эффективности требует отдельных целевых исследований и серьезной разработки, что, безусловно, положительно повлияет на развитие всей сферы прикладной информатики.

ЗАКЛЮЧЕНИЕ

Несколько выходя за рамки проблематики, рассматриваемой в настоящей статье, можно сформулировать следующие выводы и заключения.

1. Очевидно, что в современных экономических условиях проблема преодоления тенденций инерционного развития национальной информационной системы требует для своего решения новых идей и новых концептуальных подходов. Реальные перспективные точки роста есть. Государственная система

научно-технической информации, как отрасль народного хозяйства, располагает значительным научным и информационным потенциалом.

2. Реализация высокого уровня информационного обеспечения исследований, разработок, производства является основной задачей Национальной информационной системы. Решение этой задачи обеспечивается рациональным, сбалансированным развитием информационной инфраструктуры, информационных ресурсов, информационных технологий. Следует констатировать, что такие перспективные направления как *технологии Большие данные (Big Data)* и *широкополосный интернет* существенно отстают от мирового уровня. Технологической основой Больших данных являются высокоскоростные телекоммуникации и суперкомпьютерные технологии. В последние годы Россия не входит в список Top-500 суперЭВМ. А, как отмечалось выше, количество абонентов широкополосного доступа в Интернет: в США > 80 млн чел., в Китае ~ 100 млн чел., в России ~ 30 тыс. чел. (по данным J'son&Partners Consulting на начало 2015 г.)

3. Разработка на базе рассмотренной Модели типового проектного решения управления информационным обеспечением работ по этапам цикла И–Р–П позволит: а) реализовать непрерывный процесс подготовки и реализации информационной поддержки; б) количественно оценивать уровень информационной поддержки; в) предпроектно оценивать размеры информационных, вычислительных, финансовых и других ресурсов, необходимых для реализации информационной поддержки; г) оценивать эффективность планирования и расходования средств на реализацию информационной поддержки по выделенным работам; д) сократить временной лаг и качественно повысить интегральный уровень информационной поддержки всех этапов создания высокотехнологичной продукции.

4. Для повышения эффективности информационного обеспечения сферы наука–производство необходимо комплексное развитие информационной инфраструктуры, информационных (цифровых) ресурсов, внедрение новых методов и компьютерных технологий информационной поддержки исследований и разработок. Первоочередные шаги в этом направлении должны включать:

- разработку комплексной федеральной Программы развития (и структурной оптимизации) национальной информационной системы (на период 4 – 5 лет);
- решение задачи повышения эффективности управления: развитие информационной инфраструктуры, информационных ресурсов; внедрение новых технологий информационного обеспечения.

Необходимой предпосылкой успешного решения этих задач является концентрация полномочий и ответственности по модернизации национальной информационной системы в рамках одного федерального ведомства.

В целом, в заключение, можно констатировать, что повышение эффективности управления информационным обеспечением интенсифицирует процессы исследований и разработок в научно-промышленной сфере – приоритетной задаче экономического развития страны.

СПИСОК ЛИТЕРАТУРЫ

1. Сюттюренько О.В. Цифровая среда: тренды и риски развития // Научно-техническая информация. Сер. 1. – 2015. – № 2. – С.1-7; Syunturenko O.V. The Digital Enviroment: The Trends and Risks of Development // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 1. – P. 24-29.
2. Сюттюренько О.В., Булычева О.С. Электронные библиотеки: перспективы и риски развития // Сборник научных трудов Президентской библиотеки им. Б.Н. Ельцина. Сер. «Электронные библиотеки». Вып. 6: Интегрированные цифровые ресурсы: организационно-технологические и научно-методические основы развития. – 2015. – 271 с.
3. Родионов И.И., Цветкова В.А., Родина И.В., Каширская Н.Ф. Распространение научных периодических изданий: состояние и перспективы // Научно-техническая информация. Сер. 1. – 2015. – № 7. – С. 22-27.
4. Европа откроет доступ к научным статьям. – URL: <http://www.postovet.ru/blog/russia/347784.html>.
5. Сюттюренько О.В. Перспективы использования интернет-СМИ, журналов открытого доступа и социальных медиа в научно-технической сфере // Научно-техническая информация. Сер. 1. – 2015. – № 6. – С. 30-36.
6. EUROPA – CORDIS: Community Research and Development. – URL: <http://cordis.europa.eu/>.
7. Федеральный закон от 31.12.2014 № 488-ФЗ (ред. от 13.07.2015) «О промышленной политике в Российской Федерации», статья 14. Государственная информационная система в промышленности.
8. Арский Ю.М., Борисова Л.Ф., Гиляревский Р.С., Сюттюренько О.В. и др. Отчет по НИР «Разработка и обоснование организационной структуры национальной информационной системы научной, научно-технической и инновационной деятельности (НИСНИД). Организация мониторинга информационных ресурсов НИСНИД для национальной инновационной системы и подготовка предложений по оптимизации состава указанных информационных ресурсов» (шифр №2006-РН-15.0/003), 2006, гос. регистрационный № 012006.07926, инв. № 02200607015 (ЦИТиС), 192 с.
9. Сюттюренько О.В., Булычева О.С. Концептуальный облик перспективного технологического пакета информационной поддержки наукоемкого производства // Научно-техническая информация. Сер. 2. – 2016. – № 4. – С. а-б.
10. Моисеева Е.В. Особенности наукоемких производств и специфика управления себестоимостью наукоемкой продукции // Бизнес. Образование. Право. Вестник Волгоградского института бизнеса. – 2011. – № 1 (14). – С. 78-81.
11. Черенков В., Зимченко Ю. Риски и шансы глобального ИТС-аутсорсинга для российских предприятий // Инновации. – 2007. – № 1. – С. 35-43.
12. Борисова Л.Ф., Сюттюренько О.В. Реферативный банк данных ВИНТИ РАН: перспективы постобработки информации с использованием методов анализа данных // Научно-техническая информация. Сер. 1 – 2007. – № 11. – С. 6-11.
13. Биктимиров М.Р., Гиляревский Р.С., Сюттюренько О.В. Новая концептуальная основа развития информационной деятельности ВИНТИ РАН // Научно-техническая информация. Сер. 1. – 2016. – № 1. – С. 1-8.
14. Новосельцев В.Н. Математическое моделирование в век компьютеров // ЭМЖ (Электронный медицинский журнал). 2009 – URL: <http://emm.informed.su/articles/8-matematichskoe-modelirovar>.
15. Автоматическое порождение гипотез в интеллектуальных системах / сост. Е.С. Панкратова, В.К. Финн. – ЛИБРОКОМ, 2009. – 528 с.
16. Финн В.К. Об особенностях ДСМ-метода как средства интеллектуального анализа данных // Научно-техническая информация. Сер. 2. – 2001. – № 5. – С. 1-4.
17. Интернет статистика. В мире. – URL: http://www.bizhit.ru/index/polzovateli_interneta_v_mire/0-404
18. Северная Америка (опорные сети)Internet2, ESnet, NLR, NISNandCANARIE. – URL: <https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%82%D0%B5%D1%80%D0%BD%D0%B5%D1%822>
19. GEANT is the pan-European data network for the research and education community. – URL: <https://ru.wikipedia.org/wiki/GEANT>
20. Интернет статистика. Россия. – URL: http://www.bizhit.ru/index/users_count/0-151

Материал поступил в редакцию 08.02.16.

Сведения об авторе

СЮНТЮРЕНКО Олег Васильевич – доктор технических наук, профессор, ведущий научный сотрудник ВИНТИ РАН, Москва
e-mail: olegasu@mail.ru

В.А. Лекае, С.В. Моздор, Е.Н. Ольхов

Организация инфосферы большого научно-промышленного комплекса как объекта информационной индустрии

Представлена разработанная на базе системной методологии модель инфосферы ОПК как большого научно-промышленного комплекса. Обоснованы ближайшие задачи развития инфосферы ОПК и проект ее модернизированного облика.

Ключевые слова: облик инфосферы ОПК, системное проектирование инфосферы, цели и задачи инфосферы ОПК, информационная индустрия

ВВЕДЕНИЕ

Подходы к моделированию и анализу инфосферы большого научно-промышленного комплекса рассмотрены в [1]. В настоящей статье на основе предложенных в [1] методик мы представляем результаты моделирования инфосферы оборонно-промышленного комплекса (ОПК) как наиболее яркого представителя научно-промышленных комплексов России и делаем попытку определить контуры облика этой инфосферы ближайшего будущего.

С нашей точки зрения, основным ключевым звеном инфосферы ОПК, способным вывести не только ее, но и экономику России в целом из кризиса, является инновационная составляющая – Национальная инновационная система (НИС) ОПК. Это важнейшая часть, инфосферы ОПК. Ее идее, истории развития, состоянию и перспективам посвящена монография [2].

Центром Национальной инновационной системы должна стать система *Государственной регистрации и учета НИОКР*, охватывающая все планируемые, проводящиеся и законченные разработки, финансируемые по Государственному оборонному заказу (ГОЗ) или выполняемые инициативно.

Другое ключевое звено инфосферы ОПК нового поколения – *организация коммерческой реализации результатов интеллектуальной деятельности (РИД) и иной гражданской продукции ОПК* [2].

Согласно [1], анализ инфосферы ОПК в целом отражает:

- объектное пространство, во взаимодействии с которым инфосфера ОПК должна функционировать;
- основные направления развития инфосферы ОПК, которые должны координироваться с общим направлением развития инновационных процессов России;

- тенденции развития параметров рынков заказчиков и потребителей оборонной продукции;
- исследование и прогнозирование тенденций развития параметров рынка поставщиков.

МОДЕЛЬ ИНФОСФЕРЫ ОБОРОННО-ПРОМЫШЛЕННОГО КОМПЛЕКСА

Первый результат проектирования – конструкция модели инфосферы ОПК, сформированная по рассмотренной в [1] методике В. Г. Овчинникова [3] – представлена на рис. 1 в форме двух конусов (биконусов), напоминающих «песочные часы». Модель отражает, соответственно, внешнюю среду рассматриваемой инфосферы и внутреннюю структуру инфосферы ОПК.

Обсуждение и уточнение конструкции представленной модели не входит в задачи настоящей статьи. Суть компонентов нижней, обеспечивающей части, состоящей из набора моделей системного анализа В. Г. Овчинникова [3], рассмотрена в [1].

ПРОЕКТ МОДЕРНИЗИРОВАННОГО ОБЛИКА ИНФОСФЕРЫ ОПК

Логическая структура проектирования

По нашему мнению, проектирование задач по расширению и совершенствованию организации инфосферы ОПК нового поколения в границах, задаваемых информационными потребностями субъектов инфосферы (этапы формирования которых представлены в верхнем конусе рис. 1), должно стать последовательным процессом, логическая структура которого показана на рис. 2.

Рассмотрим проектирование некоторых компонент *информационной составляющей* инфосферы ОПК.

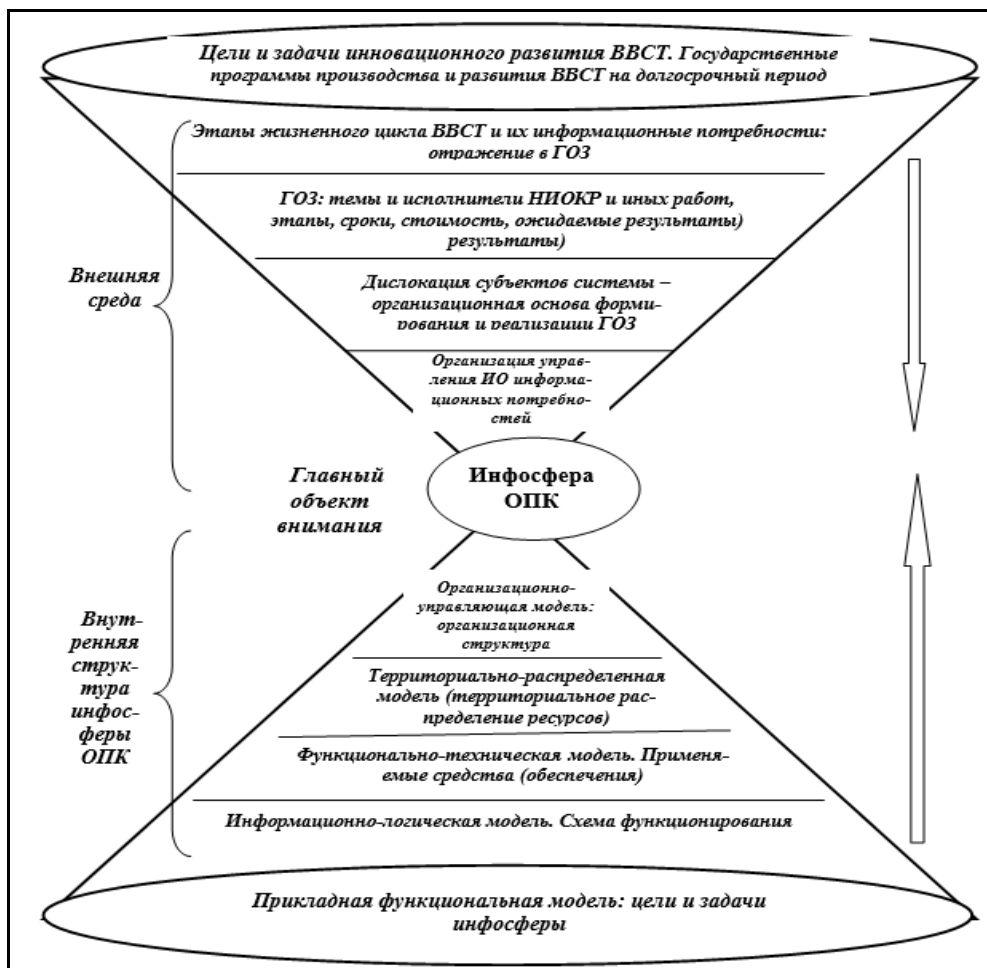


Рис. 1. Конструкция модели инфосферы ОПК (ВВСТ – вооружение, военная и специальная техника)

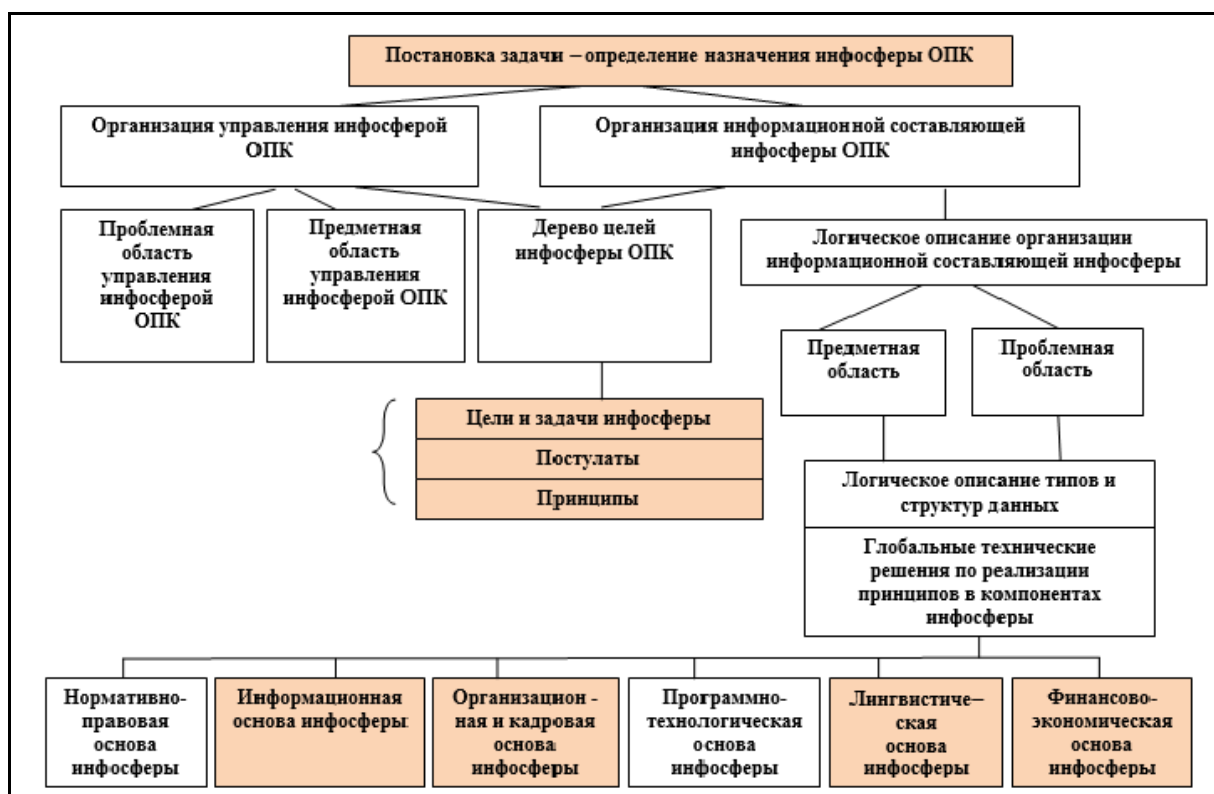


Рис. 2. Логическая структура проектирования облика инфосферы ОПК

Общее дерево целей и задач инфосферы ОПК

Определим назначение инфосферы ОПК, ее цели, задачи и принципы организации. Все это должно представлять **философский аспект** проектирования инфосферы ОПК, что в работе [1] рассматривалось по отношению к инфосфере вообще.

Общую цель модернизируемой инфосферы оборонно-промышленного комплекса можно сформулировать следующим образом.

Обеспечение комплексного информационного сопровождения процессов организации и управления оборонно-промышленным комплексом, прежде всего в части формирования и реализации Государственного оборонного заказа, и иных процессов жизнедеятельности оборонно-промышленного комплекса, включая коммерческую реализацию продукции.

Можно выделить две подцели инфосферы ОПК:

1) **информационное обеспечение управления функционированием и развитием оборонно-промышленного комплекса;**

2) **информационное обеспечение собственного функционирования и развития инфосферы оборонно-промышленного комплекса.**

Основные проблемы по реализации подцели «Информационное обеспечение управления функционированием и развитием оборонно-промышленного комплекса».

Состав задач этой группы очевиден, поскольку для любого научно-промышленного комплекса все укрупненные функции и основные процессы известны и определены как этапы жизненного цикла изделий (ЖЦИ). При реализации каждого этапа жизненного цикла изделий осуществляется управление стратегическим планированием. Оно включает планирование и контроль за аналитическими исследованиями, разработкой и производством изделий, их хранением, логистикой (включая закупки), а также управление качеством всех процессов и изделий, сбытом и сервисом, утилизацией, кадровой политикой и кадрами, охраной труда и здоровья трудящихся, финансами, документооборотом, недвижимым имуществом, защитой окружающей среды. Очевидно, что инфосфера ОПК должна обеспечивать информационное сопровождение всех этих функций, а также их безопасность.

Задачи названной группы отражают два кластера информационных технологий: а) технологии проектирования и изготовления изделий и б) технологии сопровождения хозяйственной деятельности – от маркетинга, закупок и поставок комплектующих, управления процессами производства до управления финансами и кадрами. Оба кластера задач реализуются на базе прикладных автоматизированных информационных систем (АИС).

Назовем основные проблемы, порождающие соответствующие задачи инфосферы ОПК.

Первое. Разнообразие программных средств, используемых в ОПК, вызывает необходимость стандартизации используемых типов программных

средств для повышения эффективности их взаимодействия и интеграции разного рода систем, включая лингвистику и форматы.

Второе. Территориальная и ведомственная разбросанность информационных ресурсов и принципиальная, с нашей точки зрения, невозможность собрать эти ресурсы в одном вычислительном комплексе делает целесообразным и даже необходимым накопление специальной интегрированной технологической базы данных справочно-навигационного типа (метаданных), отражающей мини-информацию о первичных информационных источниках.

Третье. Ощущается необходимость продолжения адаптации стандартов ИСО (Международной организации по стандартизации) в области CALS¹ и иных технологий.

Четвертое. Недостаточно уделяется внимания анализу прототипов (прежде всего зарубежных) разрабатываемой продукции, особенно из источников «серой»² информации. Поиск такой информации требует финансовых затрат и навыка. Ранее это реализовывалось в низовых и отраслевых органах научно-технической информации. Огромная масса нынешних исполнителей, хотя и «живут» в Интернете, давшем им мощнейшие поисковые машины (Gagle, Yandex и т.д.), «развращены» кажущейся простотой и быстротой поиска, не подстегивающей к изучению глубоких «эшелонов выдачи» информации об источниках. По существу, многие специалисты не умеют это делать в силу специфики поиска, а информационные работники, если и не исчезли как класс, также недостаточно качественно могут проводить эту работу ввиду большого тематического разброса.

Возникает проблема организации аналитического центра для проведения такой работы с использованием Интернета, создания рабочих групп информационных работников и специалистов разработчиков. Функции такого центра могли бы быть использованы и для более серьезной аналитической работы – при проведении государственных закупок и при приёмке работ. Организация такого центра позволит начать мощное движение вперёд для совершенствования инфосферы ОПК и создания не на бумаге, а на деле Единого информационного пространства ОПК.

Кроме основных, комплексных проблем можно назвать и более локальные задачи по подцели «Информационное обеспечение управления функционированием и развитием оборонно-промышленного комплекса»: неполный охват планируемых, проводящихся и оконченных НИОКР в системе Госрегист-

¹ CALS-технологии (англ. *Continuous Acquisition and Lifecycle Support* – непрерывная информационная поддержка поставок и жизненного цикла изделий), или ИПИ (информационная поддержка процессов жизненного цикла изделий) – подход к проектированию и производству высокотехнологичной и наукоёмкой продукции, заключающийся в использовании компьютерной техники и информационных технологий на всех стадиях жизненного цикла изделия

² Под «серой» информацией понимается информация о различных мероприятиях – выставках, конференциях и т.д.

рации НИОКР (в настоящее время только 10% отчетов по НИОКР проходят обязательный государственный учет), а также отсутствие:

- законодательного закрепления ответственности за непредставление отчетных документов по НИОКР и результатам научно-технической деятельности (РНТД);
- комплексной системы учета, хранения и выдачи НТИ (отчетов по НИР и ОКР и РНТД);
- единого центра хранения документации по результатам интеллектуальной деятельности (РИД);
- системы проведения комплексной экспертизы научно-технических решений и инновационных разработок;
- мониторинга и анализа вклада НИОКР в разработку вооружений, военной и специальной техники;
- единой системы поиска документов, принятых для отражения в информационных потоках результатов интеллектуальной и научно-технической деятельности, полученных при выполнении Государственного оборонного заказа;
- инвентаризации НИОКР;
- инфраструктуры трансфера технологий из военной сферы в гражданскую, и наоборот;
- четкого механизма постановки на учет и баланс результатов научно-технической деятельности, полученных при выполнении Государственного оборонного заказа.

Основные проблемы по реализации подцели «Информационное обеспечение собственного функционирования и развития инфосферы оборонно-промышленного комплекса».

Проблемы, порождающие задачи обеспечения собственного функционирования и развития оборонно-промышленного комплекса, следующие.

Первое. Поскольку в настоящее время на российском рынке существует много продуктов, обеспечивающих разработку систем управления производством, а вопрос о создании единой платформы для организации Единого информационного пространства инфосферы ОПК не решен, необходимо провести исследование и осуществить выбор такой платформы с целью ее стандартизации для использования в инфосфере ОПК.

Второе. Текущее состояние инфосферы оборонно-промышленного комплекса показывает необходимость разработки и внедрения более адресного информационного обеспечения для всех этапов управления инновационными процессами в ОПК, чем в имеющихся на сегодняшний день проектах, где основное внимание уделено аппаратной и программной частям, предназначенным для использования в инфосфере оборонно-промышленного комплекса.

В результате проведенного анализа, частично представленного в настоящей статье, сформирован общий перечень задач инфосферы ОПК по выделенным выше подцелям (см. таблицу), который, конечно, требует дальнейшей проработки и актуализации и может стать предметом дискуссий.

Задачи построения инфосферы оборонно-промышленного комплекса

Подцели инфосферы ОПК	Группы задач по подцелям инфосферы ОПК
Информационное обеспечение управления функционированием и развитием оборонно-промышленного комплекса	Информационное сопровождение основных процессов разработки, создания и производства изделий оборонно-промышленного комплекса на всех стадиях жизненного цикла изделий, включая хранение, утилизацию, сбыт
	Информационное сопровождение процессов управления инновационной составляющей инфосферы оборонно-промышленного комплекса
Информационное обеспечение собственного функционирования и развития инфосферы оборонно-промышленного комплекса	Развитие информационного сопровождения основных функций разработки, производства и реализации вооружений, военной и специальной техники и иных изделий оборонно-промышленного комплекса
	Оптимизация организации и развитие инструментария инфосферы оборонно-промышленного комплекса

Формирование организационной структуры инфосферы ОПК

Организационная структура Единой системы информационного обеспечения оборонных отраслей (ЕСИО) и ее автоматизированной части – Межотраслевой информационной автоматизированной системы (МИАС «Созвездие») соответствовала процессам государственного управления экономикой того периода.

В работе Р.А. Фатхутдинова [4] рассмотрены различные типы организационных структур, которые можно применить к информационным системам ОПК разного периода. Так, в советский период был реализован вариант «звездной структуры» – разомкнутой конфигурации, характеризующейся четкой централизацией и отсутствием периферийных связей, которые, разветвляясь, шли вниз, к предприятиям (нисходящие потоки) и от предприятий по тем же каналам вверх (восходящие потоки). Такой была организационная структура МИАС «Созвездие», развитие которой при сохранении прежней экономической формации и уровня режимных требований смогло превратить ее в «колесо» – замкнутую централизованную конфигурацию, сформированную на основе сужающих/расширяющих соединений, с появляющимися периферийными связями. Ослабленное централизованное управление, действующее в то время в ОПК, позволяло бы субъектам системы иметь периферий-

ные функциональные связи, выражающиеся в информационном обмене между субъектами всех уровней.

В переходный период экономики России инфосфера ОПК деформировалась в «двойное кольцо» – все еще замкнутую конфигурацию, сформированную на основе сужающих и расширяющих соединений. В данной структуре отсутствует выраженная централизация. Но и вполне децентрализованной такая структура не являлась, поскольку все еще имелся относительный центр, заключенный во внутреннем кольце (остатки отраслевых центров НТИ), и относительная периферия, заключенная во внешнем кольце (информационные подразделения предприятий, активно разрушающиеся).

Подобные структуры характерны для организаций, управление которыми осуществляет совет, каждый член которого курирует какое-то определенное направление деятельности. Правда «совет» – управляющий орган того времени доживал в своей предыдущей структуре последние дни.

С нашей точки зрения, конфигурация общей инфраструктуры инфосферы ОПК должна в дальнейшем трансформироваться скорее всего в структуру по типу «колесо с двойным ободом». Это сочетание «двойного кольца» (замкнутой структуры с периферийными связями) со «звездной структурой» дает рациональную и широко распространенную конфигурацию, обладающую в отличие от «двойного кольца» четкой централизацией.

Все связи нынешних предприятий оборонно-промышленного комплекса сами по себе не сформируют системно организованного Единого информационного пространства, этот процесс должен быть управляемым. Смеем утверждать, что даже реальный переход экономики страны на рыночный путь развития не отменит определенную обособленность и закрытость ОПК в силу следующих причин:

режимный характер большинства разработок, ограниченный или полностью закрытый доступ к информационным каналам и информационным массивам, отражающим ход разработки вооружений, военной и специальной техники (ВВСТ);

невозможность полного исключения государственного регулирования инновационных процессов и производства, даже при реализации на коммерческой основе открытой части ВВСТ и продукции двойного назначения;

даже при определенной «вольнице» коммерческая составляющая в информационном обмене будет сопровождаться достаточно жестким контролем за реализацией в гражданских отраслях и тем более за рубежом фрагментов разработок, финансируемых в составе Государственного оборонного заказа.

Другой принцип реализации Национальной инновационной системы оборонно-промышленного комплекса как центральной части инфосферы ОПК, следующий из анализа зарубежного опыта, формулируется так: *исследовательские и проектные работы в части, осуществляющейся за государственный счет, должны быть упорядочены и представлены в согласованных документах, в форматах, предполагающих эффективное использование их резуль-*

татов, удобное хранение в поисковых системах, контроль за сохранением государственной и коммерческой тайны.

Основными движителями функционального объединения информационных ресурсов инфосферы ОПК в Единое информационное пространство ОПК должна стать Государственная регистрация НИОКР и результатов интеллектуальной деятельности (РИД), а также поставленная на надежные правовые рельсы коммерциализация реализации фрагментов оборонной продукции в гражданских отраслях.

Для НИОКР, финансируемых из государственного бюджета, должно быть обязательным представление сведений о планировании и результатах этих работ в директивно устанавливаемой череде информационных документов – регистрационных (РК) и информационных (ИК) картах, результатах интеллектуальной деятельности (РИД), частных технических решениях (ЧТР), отчетах. Порядок регистрации и отчетности по НИОКР, включая формирование информационной базы данных (регламенты, инструкции, форматы представления данных), должен быть определен обновленными директивными документами.

Для всех других работ предприятий и организаций оборонно-промышленного комплекса участие в формировании документов, включаемых в каналы системы и ее БД, является добровольным, но желательным, основывается на экономической выгоде и заинтересованности в продвижении своей продукции, обосновании потребностей в решении проблем, поиске партнеров, выявлении конкурентов, оперативной связи с торговыми и финансовыми органами, получении сведений о новейших научных, технических, экономических решениях, ноу-хау, опыте конверсии и приватизации, получении рекомендаций независимой экспертизы и др.

В первом случае информационные документы нормализованы и являются обязательными для любых предприятий и организаций. Во втором случае формы документов, содержащих сведения, вводимые в систему, должны быть более «либерализованы» при обязательном обеспечении требований достоверности, полноты и оперативности (письма, бланки, фирменные материалы, рекламные описания, телеграммы и т. д.).

Уже на начальном этапе обмена информацией могут возникнуть деловые контакты горизонтального уровня в среде разработчиков и создателей продукции ОПК. Безусловно, даже на современном этапе развития инфосферы ОПК информация об эффективных разработчиках и успешных предприятиях уже используется специалистами, готовыми к ее восприятию. Но охват специалистов всех уровней в создаваемой системе необходим во много раз большим. Предполагается дать возможность обмена информацией не только о готовых и ресурсоемких проектах, но и о технологиях постепенного «выращивания» высокотехнологичного производства малого бизнеса оборонно-промышленного комплекса. Сюда должна попадать информация о фрагментах инноваций, которые авторы планируют передавать и тиражировать с выгодой для предприятия и для себя лично.

Готовность к созданию систем электронного документооборота, к автоматизации процесса управления собственным производством должна стать основным побудительным мотивом руководителей предприятий.

Оптимальные управляющие решения на уровне руководителей различных уровней производства и разработки изделий также могут стать предметом обмена в едином пространстве. Продолжением контактов могут стать объединение производственных и людских ресурсов на отдельных этапах разработки и производства оборонной продукции в единое виртуальное «предприятие», хотя и сопровождаться повышением локальной конкуренции путем расширения круга исполнителей Государственного оборонного заказа.

В процессе разработки и создания изделий с характеристиками, которые превосходят существующие аналоги, обязательно будет присутствовать комплекс новых подходов и технических решений, которые и позволят создавать более эффективные образцы техники. Эти частные технические решения (ЧТР) и могут стать основным предметом коммерческого обмена внутри отрасли, основой для создания отдельных направлений развития техники, а принципы, на которых они основаны, могут быть использованы при создании или усовершенствовании новых средств сдерживания для других исполнителей Государственного оборонного заказа. Пример наших зарубежных партнеров говорит ровно об этом (см. действующий с 1986 г. Закон «О передаче федеральных технологий»). Не грех вспомнить и положительный

опыт ВИМИ по созданию каналов распространения информации о частных технических решениях.

Резюмируя представленные рассуждения, можно сказать, что строго регламентируемые государством процессы государственной регистрации и учета и процессы коммерческой реализации отдельной продукции оборонных предприятий (изделий двойного применения, частных технических решений), являющейся предметом двойной собственности (государства, финансирующего разработку и производство в составе Государственного оборонного заказа, и разработчика), должны составлять два направления государственной стратегии в области интеллектуальной собственности.

Перспективная функциональная модель развивающейся инфосферы ОПК

Функциональная модель МИАС «Созвездие» (технологический аспект рассмотрения инфосферы) представлена в [5]. Эта система составляла костяк инфосферы оборонно-промышленного комплекса значительного периода. Модель показывала, что реализация основных форм информационного обеспечения осуществляется с помощью системных обеспечений, реализующих технологические операции, организованные в этапы единого технологического процесса системы.

В перспективной функциональной модели инфосферы ОПК (рис. 3) в роли обеспечений выступают все модели, представленные ранее на рис. 1.

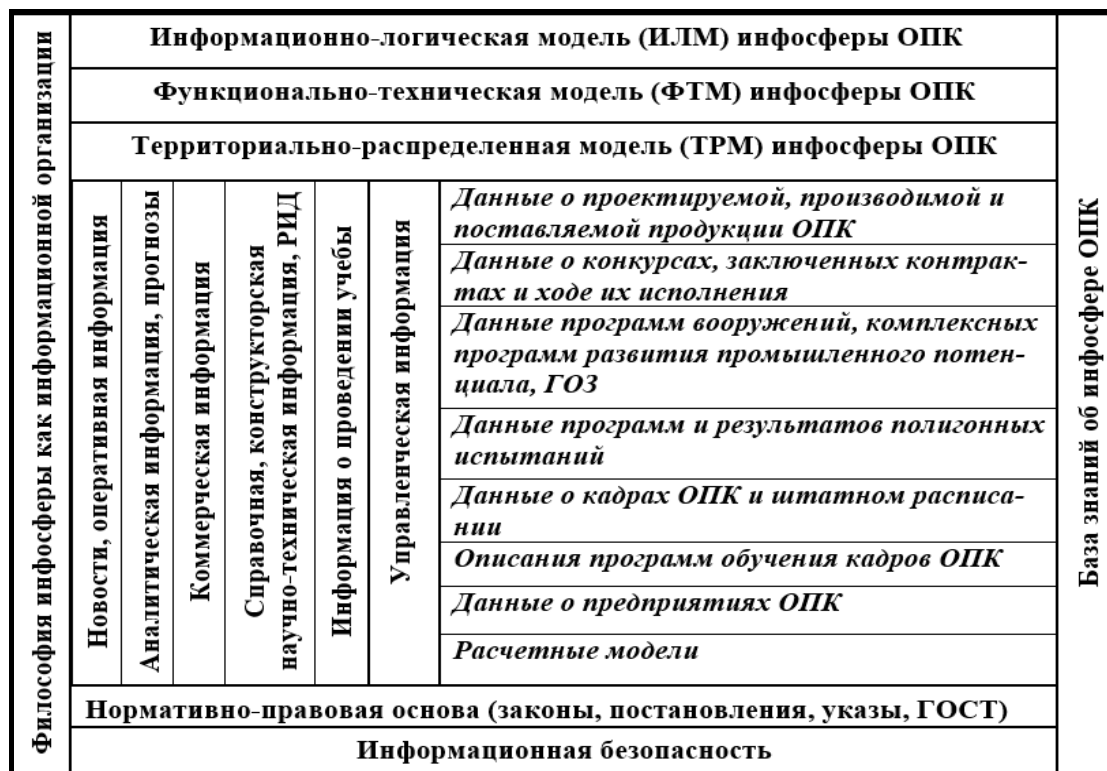


Рис. 3. Перспективная интегрированная модель инфосферы ОПК

Все компоненты модели объединены философией инфосферы как информационной организации и базой знаний о системе.

Каждый элемент модели может быть развернут в детальную предметную область при реальном проектировании.

ЗАКЛЮЧЕНИЕ

Облик инфосферы ОПК ближайшего будущего, фрагменты которой представлены в настоящей статье, отражает общий взгляд авторов на развитие нынешней ситуации и в определенной степени является дискуссионным.

Детальное рассмотрение проблемы организации инфосферы ОПК, прежде всего анализ истории ее развития и текущего состояния, актуальные проблемы коммерческой реализации инновационной и иной продукции ОПК требуют иных издательских возможностей.

В самом общем виде организация справочно-поискового аппарата (лингвистический аспект рассмотрения инфосферы) представлена в работе [6].

Экономический аспект рассмотрения инфосферы оборонно-промышленного комплекса, заложенный в статьях Г.Т. Артамонова [7] и кратко рассмотренный нами в работе [1], и анализ действия законов прагматической информатики, представленных в названных статьях Г.Т. Артамонова, еще ждут своего исследователя.

СПИСОК ЛИТЕРАТУРЫ

1. Антопольский А.Б., Евстафьев В.Ф., Моздор С.В. Инфосфера больших научно-производственных комплексов как форма организации: методика анализа // Научно-техническая информация. Сер.2. – 2016. – № 3. – С.1-13.
2. Моздор С.В., Ольхов Е.Н., Иванов М.О. Межотраслевой обмен инновациями ОПК и передача достижений. Актуальные проблемы. LAP LAMBERT Academic Publishing. – URL: <https://www.lap-publishing.com/catalog/details/store/gb/book/978-3-659-58698-9/%D0%9C%D0%B5%D0%B6%D0%BE%D1%82%D1%80%D0%B0%D1%81%D0%BB%D0%B5%D0%B2%D0%BE%D0%B9-%D0%BE%D0%B1%D0%BC%D0%B5%D0%BD-%D0%B8%D0%BD%D0%BD%D0%BE%D0%B2%D0%B0%D1%86%D0%B8%D1%8F%D0%BC%D0%B8-%D0%9E%D0%9F%D0%>

9A-%D0%B8-%D0%BF%D0%B5%D1%80%D0%B5%D0%B4%D0%B0%D1%87%D0%B0-%D0%B4%D0%BE%D1%81%D1%82%D0%B8%D0%B6%D0%B5%D0%BD%D0%B8%D0%B9.

3. Овчинников В.Г. Методологическое введение в информатику. – М.: Компания Спутник+, 2004. – 217 с.
4. Фатхутдинов Р.А. Организация производства: учебник. 3-е изд., перераб. и доп. – М.: ИНФОРМ-М, 2007. – 504 с.
5. Кузнецов М.В., Моздор С.В. Методические основы проектирования и эксплуатации информационных систем. Часть 2. Общие подходы к логическому проектированию информационной системы // Межотраслевая информационная служба. – 2012. – № 3 (160). – С. 36-43.
6. Головатый И.Н., Евстафьев В.Ф., Моздор С.В. Информационная среда оборонного комплекса. Часть 2. Концепция развития информационной инфраструктуры Рособоронзаказа // Межотраслевая информационная служба. – 2012. – № 2 (159). – С. 3-13.
7. Артамонов Г.Т. Информатика: теория и практика (заготовки к книге) // Научно-техническая информация. Сер 1. – 1997. – № 8. – С. 30–33; 1998. – №1. – С. 29–34; № 4. – С. 31–36; 1998. – № 6. – С. 31–35; 1998. – № 12. – С. 30–33; 1999. – № 6. – С. 36–43.

Материал поступил в редакцию 21.01.16.

Сведения об авторах

ЛЕКАЕ Владимир Алексеевич – кандидат химических наук, доцент Института информационных наук и технологий безопасности (ИИНТБ РГГУ), Москва
e-mail: valek41@yandex.ru

МОЗДОР Сергей Владимирович – кандидат технических наук, доцент, консультант Всероссийского научно-исследовательского института межотраслевой информации (ВИМИ), Москва
e-mail: dr44@inbox.ru

ОЛЬХОВ Евгений Николаевич – начальник отдела ВИМИ
e-mail: oen@vimi.ru

Алгоритм информационного проецирования на аффинное подпространство*

Исследован алгоритм поиска ближайшей точки аффинного подпространства в положительном ортанте к исходной точке относительно расстояния Кулбэка-Ляйблера. Эту задачу решает классический алгоритм Дарроча-Рэтклиффа. Наше изложение алгоритма использует идеи информационной геометрии, созданной Н.Н. Ченцовым и И. Чисаром. Основная теорема работы доказывает сходимость этого алгоритма способом, отличным от известных ранее. Предложенный алгоритм может применяться, например, к нахождению оценок максимального правдоподобия в экспоненциальном семействе, как описано в заключении.

Ключевые слова: КЛ-расстояние, положительный ортант, аффинное подпространство, экспоненциальное семейство, МП-оценка

ВВЕДЕНИЕ

Во многих задачах обработки информации исходный объект описывается точкой с положительными координатами (элементом положительного ортанта), получаемой суммированием каких-то характеристик. Часто при этом теория требует, чтобы такие объекты удовлетворяли семейству линейных равенств (и неравенств). Увеличением размерности (положительного) ортанта (введением дополнительных координат) выпуклое множество точек, определяемое системой равенств и неравенств, может быть заменено на часть аффинного подпространства, лежащую в ортанте. Таким образом, возникает задача замены исходного объекта (не удовлетворяющего теоретическим ограничениям) на точку аффинного подпространства, в некотором смысле ближайшую к нему. При этом обычное (евклидово) расстояние не является удовлетворительным, так как ортогональная проекция может вывести из ортанта (этот эффект продемонстрирован в примере 1). Для преодоления описанной трудности естественно использовать расширение расстояния Кулбэка-Ляйблера на положительный ортант. С вычислительной точки зрения этот алгоритм напоминает алгоритм Дарроча-Рэтклиффа [1].

Информационная геометрия, созданная в работах Н.Н. Ченцова [2] и И. Чисара [3], позволяет придать этой задаче наглядный вид. Проекция на аффинное подпространство совпадает с единственной точкой пересечения подпространства и ортогонального ему экспоненциального семейства точек, определяемого

исходной точкой и матрицей, задающей подпространство. Поэтому алгоритм информационного проецирования может быть описан, как движение точки вдоль экспоненциального семейства от начальной точки до пересечения с аффинным подпространством. Но возникает вопрос: каковы должны быть шаги такого движения?

Часто полезно предполагать, что вектора координат точек нормированы (по норме l_1), что позволяет рассматривать их как распределения вероятностей, заданные на конечном множестве (номеров координат). При этом происходит дополнительное ограничение геометрии на единичный симплекс (множество точек с единичной l_1 -нормой) в положительном ортанте, а к системе уравнений добавляется уравнение $x_1 + \dots + x_n = 1$. Наличие этого уравнения позволяет нормировать матрицу и столбец свободных членов, задающих аффинное подпространство, так, чтобы суммы всех столбцов были равны 1.

Получившаяся нормированная матрица задает линейное отображение исходного ортанта внутрь положительного ортанта из пространства ее столбцов. Но на образе имеется свое расстояние Кулбэка-Ляйблера! При этом все точки аффинного подпространства отображаются в одну точку симплекса-образа, задаваемую нормированным столбцом свободных членов.

Шаг алгоритма информационного проецирования будет соответствовать расстоянию Кулбэка-Ляйблера от образа текущей точки до фиксированной точки образа аффинного подпространства. При этом, если это расстояние будет равно нулю, то и расстояние от текущей точки до аффинного подпространства будет равно

* Работа выполнена при поддержке гранта РФФИ (проект 14-07-00856а).

нулю, что означает, что мы нашли информационную проекцию. Эта идея и реализуется в доказательстве теоремы о сходимости предложенного алгоритма.

В первом разделе мы определим и докажем основные факты из информационной геометрии. Автор здесь не претендует ни на какую оригинальность, включение этого материала в работу диктуется желанием автора сделать изложение максимально замкнутым.

Второй раздел посвящен изложению алгоритма, а третий – доказательству теоремы о сходимости. В заключение мы укажем на возможность применения описанного алгоритма для вычисления оценок максимального правдоподобия в экспоненциальном семействе (здесь опять излагается классический результат, изложенный в книге И. Чисара и П. Шилдса [4]).

1. БАЗИСНЫЕ ОПРЕДЕЛЕНИЯ, ФАКТЫ И ПРИМЕРЫ

КЛ-расстояние $P = (p_1, \dots, p_n) \in \mathbb{R}_+^n$ от $Q = (q_1, \dots, q_n) \in \mathbb{R}_+^n$ определяется как число

$$\delta(P|Q) = \sum_{j=1}^n p_j \cdot \log \frac{p_j}{q_j} - \log_2 e \cdot \sum_{j=1}^n p_j + \log_2 e \cdot \sum_{j=1}^n q_j$$

Полагаем, что

$$0 \cdot \log_2 0 = 0 \cdot \log_2 \frac{0}{0} = 0 \text{ и } p \cdot \log_2 \frac{p}{0} = +\infty$$

при $p > 0$. При условии $\sum_{j=1}^n q_j = \sum_{j=1}^n p_j = 1$ это определение совпадает с классическим. В литературе классическое КЛ-расстояние имеет множество других имен: информационная дивергенция, относительная энтропия, и т.п. Следует сразу указать, что это расстояние не удовлетворяет многим свойствам метрики, кроме одного, доказанного в Лемме 1 ниже. Множители $\log_2 e$ (e – основание натуральных логарифмов) используются как нормировка с учетом использования в классическом КЛ-расстоянии двоичных логарифмов.

Лемма 1.

Для любых $P \in \mathbb{R}_+^n$ и $Q \in \mathbb{R}_+^n$ имеем $\delta(P|Q) \geq 0$. Более того, $\delta(P|Q) = 0$ тогда и только тогда, когда $\forall j [p_j = q_j]$.

Доказательство. Заметим, что

$$\delta(P|Q) = \sum_{j=1}^n p_j \cdot \log_2 e \cdot \left(\frac{q_j}{p_j} - 1 - \ln \frac{q_j}{p_j} \right).$$

Но $x - 1 \geq \ln x$ при $x > 0$, и $x - 1 = \ln x$, если и только если $x = 1$.

Носитель $P \in \mathbb{R}_+^n$ определяется как множество $S(P) = \{j | p_j > 0\}$. Говорят, что $Q \in \mathbb{R}_+^n$ доминирует $P \in \mathbb{R}_+^n$ ($P \ll Q$), если $S(P) \subseteq S(Q)$. Точка $P \in \mathbb{R}_+^n$, для которой $S(P) = \{1, \dots, n\}$, или эквивалентно $\forall j (p_j > 0)$, называется внутренней ($P \in \text{Int}(\mathbb{R}_+^n)$). Легко доказать, что для любых

$P \in \mathbb{R}_+^n$ и $Q \in \mathbb{R}_+^n$ выполняется $\delta(P|Q) < \infty$ тогда и только тогда, когда $P \ll Q$.

Для вектора-столбца $P = (p_1, \dots, p_n) \in \text{Int}(\mathbb{R}_+^n)$ определим *ковектор-стоку* как $P = [\log_2 p_1, \dots, \log_2 p_n] \in \mathbb{R}^n$.

Легко проверить, что *градиент* $\nabla_1 \delta(P|Q)$ имеет

$$\text{компоненты } \frac{\partial}{\partial p_j} \delta(P|Q) = \log_2 \frac{p_j}{q_j} = \log_2 p_j - \log_2 q_j.$$

Поэтому *приращение* $\delta(P|Q)$ вдоль направления $V = (v_1, \dots, v_n)$ равно свертке

$$\langle \tilde{P} - \tilde{Q} | V \rangle = \sum_{j=1}^n \frac{\partial}{\partial p_j} \delta(P|Q) \cdot v_j = \sum_{j=1}^n \log_2 \frac{p_j}{q_j} \cdot v_j.$$

И-проекция точки $Q \in \mathbb{R}_+^n$ на выпуклое замкнутое подмножество $\Gamma \subseteq \mathbb{R}_+^n$ определяется как единственная точка $\Pi_\Gamma(Q) \in \Gamma$, для которой $\delta(\Pi_\Gamma(Q)|Q) \leq \delta(P|Q)$ для любой $P \in \Gamma$.

Лемма 2.

Пусть $Q \in \text{Int}(\mathbb{R}_+^n)$ и $\Gamma \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$. Тогда для $R = \Pi_\Gamma(Q)$ выполняется $\langle \tilde{R} - \tilde{Q} | P - R \rangle \geq 0$ для любого $P \in \Gamma$.

Доказательство легко следует из минимальности $\delta(P|Q)$ вдоль направления $V = P - R$.

Теорема 1.

Пусть $Q \in \text{Int}(\mathbb{R}_+^n)$ и $\Gamma \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$. Тогда $\Pi_\Gamma(Q) \in \Gamma \cap \text{Int}(\mathbb{R}_+^n)$. Точка $R \in \Gamma \cap \text{Int}(\mathbb{R}_+^n)$ совпадает с $\Pi_\Gamma(Q)$ тогда и только тогда, когда для любого $P \in \Gamma$ выполняется неравенство Пифагора $\delta(P|Q) \geq \delta(P|R) + \delta(P|Q)$.

Доказательство. Из условия $P \in \Gamma \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$ следует, что $\delta(P|Q) < \infty$. Тогда тем более $\delta(\Pi_\Gamma(Q)|Q) < \infty$, что влечет $\Pi_\Gamma(Q) \in \text{Int}(\mathbb{R}_+^n)$. Прямые вычисления показывают, что $\delta(P|Q) = \delta(P|R) + \delta(R|Q) - \langle \tilde{Q} - \tilde{R} | P - R \rangle$. Теперь теорема следует из Леммы 2 и легко проверяемого факта, что $\forall P \in \Gamma [\delta(P|Q) \geq \delta(P|R) + \delta(R|Q)]$ влечет $R = \Pi_\Gamma(Q)$.

Аффинное подпространство определяется как $\Lambda_{M,V} = \{P \in \mathbb{R}_+^n | M \cdot P = V\}$, где M – $m \times n$ -матрица, а V – вектор с m компонентами. *Ортогональное к $\Lambda_{M,V}$ экспоненциальное семейство* (э.с.) определяется как

$$\Upsilon_{M,Q} = \{R \in \text{Int}(\mathbb{R}_+^n) | \exists \Theta = [\theta_1, \dots, \theta_m] (\tilde{R} = \Theta \cdot M + \tilde{Q})\},$$

где $\tilde{Q}$ – ковектор для $Q \in \text{Int}(\mathbb{R}_+^n)$. Ортогональность здесь понимается в смысле утверждения равенства свертки нулю из Леммы 3 с заменой Q на любую точку $\Upsilon_{M,Q}$.

Лемма 3.

Пусть $Q \in \text{Int}(\mathbb{R}_+^n)$ и $\Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$.

Тогда для $R = \Pi_{\Lambda_{M,V}}(Q)$ выполняется

$$\langle \tilde{R} - \tilde{Q} | P - R \rangle = 0$$

для любого $P \in \Lambda_{M,V}$.

Доказательство. Как в теореме 1 условие $\Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$ влечет, что

$$R = \Pi_{\Lambda_{M,V}}(Q) \in \Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n).$$

Берем произвольное $P \in \Lambda_{M,V}$. По Лемме 2

$$\langle \tilde{R} - \tilde{Q} | P - R \rangle \geq 0.$$

Но так как $R = \Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$, найдется такое $t < 0$, что $t \cdot P + (1-t) \cdot R \in \Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n)$. Опять по Лемме 2 $\langle \tilde{R} - \tilde{Q} | (t \cdot P + (1-t) \cdot R) - R \rangle \geq 0$. Но $\langle \tilde{R} - \tilde{Q} | (t \cdot P + (1-t) \cdot R) - R \rangle = t \cdot \langle \tilde{R} - \tilde{Q} | P - R \rangle \geq 0$.

Теорема 2.

Пусть $Q \in \text{Int}(\mathbb{R}_+^n)$ и $\Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$. Тогда $\Pi_{\Lambda_{M,V}}(Q) \in \Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n)$. Точка $R \in \Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n)$ совпадает с $\Pi_{\Lambda_{M,V}}(Q)$ тогда и только тогда, когда для любого $P \in \Lambda_{M,V}$ выполняется тождество Пифагора $\delta(P|Q) = \delta(P|R) + \delta(P|Q)$. При этом $\{ \Pi_{\Lambda_{M,V}}(Q) \} = \Upsilon_{M,Q} \cap \Lambda_{M,V}$.

Доказательство. Опять $\Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$ влечет, что $\Pi_{\Lambda_{M,V}}(Q) \in \Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n)$. Тождество $\delta(P|Q) = \delta(P|R) + \delta(R|Q) - \langle \tilde{Q} - \tilde{R} | P - R \rangle$ и Лемма 3 доказывают тождество Пифагора. Если $R \in \Upsilon_{M,Q} \cap \Lambda_{M,V}$, то

$$\begin{aligned} \langle \tilde{Q} - \tilde{R} | P - R \rangle &= \langle \Theta \cdot M | P - R \rangle = \\ &= \langle \Theta | M \cdot (P - R) \rangle = \langle \Theta | V - V \rangle = 0 \end{aligned}$$

Пример 1.

Для $M = \|1, 2\|$ и $Q = (q_1, q_2) \in \text{Int}(\mathbb{R}_+^2)$ выполняется

$$\Upsilon_{M,Q} = \left\{ R = (r_1, r_2) \in \text{Int}(\mathbb{R}_+^2) \mid \exists \theta_1 \begin{cases} r_1 = q_1 \cdot 2^{\theta_1}; \\ r_2 = q_2 \cdot 2^{2\theta_1} \end{cases} \right\}.$$

Другими словами, $\Upsilon_{M,Q}$ содержит точки $(r_1, r_2) \in \text{Int}(\mathbb{R}_+^2)$, лежащие на квадратичной параболе $r_2 = c \cdot (r_1)^2$, где $c = q_2 / (q_1)^2$. И-проекция $Q = \left(1, \frac{15}{2}\right)$ на $\Lambda_{\|1,2\|(2)} = \{P \in \mathbb{R}_+^2 \mid p_1 + 2 \cdot p_2 = 2\}$ есть

$$R = \left(\frac{1}{3}, \frac{5}{6}\right) - \text{решение системы } \begin{cases} r_2 = \frac{15}{2} \cdot (r_1)^2 \\ r_1 + 2 \cdot r_2 = 2 \end{cases}.$$

$$\text{ние } X = \left(-\frac{9}{5}, \frac{19}{10}\right) \text{ системы } \begin{cases} 2 \cdot (x_1 - 1) = (x_2 - \frac{15}{2}) \\ x_1 + 2 \cdot x_2 = 2 \end{cases}$$

является евклидовой проекцией. Но эта точка находится вне положительного квадранта!

И-проекция распределения вероятностей (р.в.) Q на $\{1, \dots, n\}$ ($\sum_{j=1}^n Q(j) = 1$) на аффинное подпространство р.в. вида

$$\Lambda_{M,V} = \left\{ P \in \mathbb{R}_+^n \mid M \cdot P = V; \sum_{j=1}^n P(j) = 1 \right\},$$

где M – $m \times n$ -матрица, а V – вектор с m компонентами определяется как единственная точка $\Pi_{\Lambda_{M,V}}(Q) \in \Lambda_{M,V}$, для которой

$$\delta(\Pi_{\Lambda_{M,V}}(Q) | Q) \leq \delta(P | Q)$$

для любой $P \in \Lambda_{M,V}$. Ортогональное к $\Lambda_{M,V}$ экспоненциальное семейство – это

$$\Xi_{M,Q} = \left\{ R \in \text{Int}(\mathbb{R}_+^n) \mid \exists \theta_{m+1} \exists \theta = (\theta_1, \dots, \theta_m) \begin{bmatrix} \tilde{R} = \Theta \cdot M + \theta_{m+1} + \tilde{Q} \end{bmatrix} \right\}.$$

Лемма 4.

Пусть р.в. $Q \in \text{Int}(\mathbb{R}_+^n)$ и $\Lambda_{M,V} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$. Тогда $\{ \Pi_{\Lambda_{M,V}}(Q) \} = \Xi_{M,Q} \cap \Lambda_{M,V}$. Для любого р.в. $P \in \Lambda_{M,V}$ выполняется тождество Пифагора

$$\delta(P|Q) = \delta(P | \Pi_{\Lambda_{M,V}}(Q)) + \delta(\Pi_{\Lambda_{M,V}}(Q) | Q).$$

Доказательство. Расширим матрицу M до $(m+1) \times n$ матрицы $\tilde{M}$, добавив снизу строку $[1, 1, \dots, 1]$ длины n . Вектор V добавлением последнего элемента 1 расширим до вектора $\tilde{V}$. Тогда

$$\begin{aligned} \Lambda_{M,V} &= \left\{ P \in \mathbb{R}_+^n \mid M \cdot P = V; \sum_j P(j) = 1 \right\} = \\ &= \left\{ P \in \mathbb{R}_+^n \mid \tilde{M} \cdot P = \tilde{V} \right\} \end{aligned}$$

Ясно, что $\Xi_{M,Q} = \Upsilon_{\tilde{M},Q}$. Теперь лемма сводится к Теореме 2.

Пример 2.

Рассмотрим аффинное подпространство распределений вероятностей

$\Lambda_{\|1-10\|(0)} = \{P \in \mathbb{R}_+^3 \mid p_1 = p_2; p_1 + p_2 + p_3 = 1\}$. Э.с. для $M = \|1-10\|$ и р.в. $Q = (q_1, q_2, q_3)$ равно

$$\Xi_{M,Q} = \left\{ R = (r_1, r_2, r_3) \in \text{Int}(\mathbb{R}_+^3) \mid \exists \theta_1 \exists \theta_2 \begin{cases} r_1 = q_1 \cdot 2^{\theta_1} \cdot 2^{\theta_2}; \\ r_2 = q_2 \cdot 2^{-\theta_1} \cdot 2^{\theta_2}; r_3 = q_3 \cdot 2^{\theta_2} \end{cases} \right\}.$$

Таким образом, $\Xi_{M,Q}$ содержит точки $(r_1, r_2, r_3) \in \text{Int}(\mathbb{R}_+^3)$ конуса $(r_3)^2 = c \cdot r_1 \cdot r_2$, где $c = \frac{(q_3)^2}{q_1 \cdot q_2}$.

2. АЛГОРИТМ ИНФОРМАЦИОННОГО ПРОЕКТИРОВАНИЯ

Как было показано в предыдущем параграфе, для аффинного подпространства р.в.

$$\Lambda_{M,V} = \left\{ P \in \mathbb{R}_+^n \mid M \cdot P = V; \sum_j P(j) = 1 \right\}$$

имеет место равенство $\Lambda_{M,V} = \left\{ P \in \mathbb{R}_+^n \mid \tilde{M} \cdot P = \tilde{V} \right\}$,

где $(m+1) \times n$ матрица $\tilde{M}$ получается из матрицы M добавлением последней строки $[1, 1, \dots, 1]$ длины

n , а вектор $\tilde{V}$ получается из V добавлением последней координаты, равной 1. Но точки подпространства $P \in \Lambda_{M,V}$ определяются не только как решения системы линейных уравнений $\tilde{M} \cdot P = \tilde{V}$, но

и любой другой, получаемой из $\tilde{M} \cdot P = \tilde{V}$ удалением (и добавлением) линейно-зависимых уравнений, делением любого уравнения на ненулевую константу и прибавлением к любому уравнению произвольного кратного любого другого уравнения. Подберем теперь такую систему уравнений $W \cdot P = U$ минимального ранга k , чтобы $\Lambda_{M,V} = \Lambda_{W,U}$, $\forall j \left[\sum_{i=1}^k w_{i,j} = 1 \right]$,

$$\forall i \forall j [w_{i,j} > 0] \text{ и } \sum_{i=1}^k u_i = 1, \forall i [u_i > 0].$$

Во-первых, так как $\Lambda_{\tilde{M},\tilde{V}} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$, то ранг r расширенной $(m+1) \times (n+1)$ матрицы $\|\tilde{M}\tilde{V}\|$

совпадает с рангом $(m+1) \times n$ матрицы $\tilde{M}$. Поэтому выполнено неравенство $r \leq \min \{n, m+1\}$. Ясно, что ранг не изменится, если мы добавим к каждой строке такое кратное последней строки $[1, 1, \dots, 1]$ длины $n+1$, чтобы выполнялись условия $\forall i \forall j [\tilde{w}_{i,j} > 0]$ и $\forall i [\tilde{u}_i > 0]$.

Во-вторых, выкидывая, если нужно, линейно зависимые строки (уравнения), не трогая последнюю, можно считать, что число k строк $\|\tilde{W}\tilde{U}\|$ совпадает с ее рангом r . При этом случай $n = m+1$ не интересен, так как тогда $\Lambda_{M,V} = \Lambda_{\tilde{W},\tilde{U}}$ состоит из одной точки, которая и будет $\Pi_{\Lambda_{M,V}}(Q)$ для любого распределения вероятностей $Q \in \text{Int}(\mathbb{R}_+^n)$. Значит, можно считать, что $k \leq m+1 < n$.

В-третьих, деля каждый элемент $\|\tilde{W}\tilde{U}\|$ на

$$d = \frac{1}{1-\varepsilon} \max \left\{ \sum_{i=1}^k \tilde{w}_{i,1}, \dots, \sum_{i=1}^k \tilde{w}_{i,n}, \sum_{i=1}^r \tilde{u}_i \right\},$$

получим $\Lambda_{M,V} = \Lambda_{\tilde{W},\tilde{U}}$ с $\forall j \left[\sum_{i=1}^r \hat{w}_{i,j} \leq 1-\varepsilon \right]$,

$$\forall i \forall j [\hat{w}_{i,j} > 0] \text{ и } \left[\sum_{i=1}^r \hat{u}_i \leq 1-\varepsilon \right], \forall i [\hat{u}_i > 0].$$

Положим

$$\begin{aligned} \hat{w}_{k+1,1} &= 1 - \sum_{i=1}^k \hat{w}_{i,1}, \dots, \hat{w}_{k+1,n} = \\ &= 1 - \sum_{i=1}^k \hat{w}_{i,n}, \hat{u}_{k+1} = 1 - \sum_{i=1}^r \hat{u}_i. \end{aligned}$$

Очевидно, что

$$\hat{w}_{k+1,1} \geq \varepsilon > 0, \dots, \hat{w}_{k+1,n} \geq \varepsilon > 0, \hat{u}_{k+1} \geq \varepsilon > 0.$$

По определению $\forall j \left[\sum_{i=1}^{k+1} \hat{w}_{i,j} = 1 \right]$ и $\sum_{i=1}^{k+1} \hat{u}_i = 1$. Из

этого следует, что эта система является линейно зависимой, так как суммы по столбцам, равные 1, пропорциональны уравнению $p_1 + \dots + p_n = 1$, которое добавили при переходе от $\|MV\|$ к $\|\tilde{M}\tilde{V}\|$. Удаляя

предпоследнюю строку $\left[\frac{1}{d} \dots \frac{1}{d} \right]$, ему соответствующую, и умножая остальные строки на $\frac{d}{d-1}$, получим

расширенную матрицу $\|WU\|$ минимального ранга k , для которой $\Lambda_{M,V} = \Lambda_{W,U}$, $\forall j \left[\sum_{i=1}^k w_{i,j} = 1 \right]$,

$$\forall i \forall j [w_{i,j} > 0] \text{ и } \sum_{i=1}^k u_i = 1, \forall i [u_i > 0].$$

Пример 3.

Рассмотрим аффинное подпространство распределений вероятностей

$$\Lambda_{\|1-10\|,(0)} = \left\{ P \in \mathbb{R}_+^3 \mid p_1 = p_2; p_1 + p_2 + p_3 = 1 \right\}.$$

Начинаем с матрицы $M = \|1-10\|$ и вектора

$$V = (0). \text{ Добавленные матрица } \tilde{M} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

и столбец $\tilde{V} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ преобразуются сначала в

$$\tilde{W} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \text{ и } \tilde{U} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \text{ потом (при } \varepsilon = 0.2 \text{ и}$$

$$d = 5) \text{ в } \tilde{W} = \begin{bmatrix} 0.6 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.2 \end{bmatrix} \text{ и } \tilde{U} = \begin{pmatrix} 0.4 \\ 0.2 \end{pmatrix}. \text{ Затем}$$

$$\begin{bmatrix} 0.6 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.4 \end{bmatrix} \text{ и } \begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} \text{ вычеркиванием предпослед-$$

ней строки и умножением на $\frac{d}{d-1} = 1.25$ сводятся к

$$W = \begin{bmatrix} 0.75 & 0.25 & 0.5 \\ 0.25 & 0.75 & 0.5 \end{bmatrix} \text{ и } U = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

Алгоритм.

Положим $R_0 = Q$ и $\tilde{R}_{i+1} = \tilde{R}_i + (\tilde{U} - \tilde{W}\tilde{R}_i) \cdot W$. Вычисления останавливаются, когда $\delta(U|WR_i) \leq \delta$.

Лемма 5.

Пусть $k \times n$ -матрица $W = \|w_{i,j}\|$ и вектор-столбец $U = (u_1, \dots, u_k)$ удовлетворяют условиям

$$\text{нормировки: } \forall j \left[\sum_{i=1}^k w_{i,j} = 1 \right], \quad \forall i \forall j [w_{i,j} > 0],$$

$$\forall i [u_i > 0] \text{ и } \sum_{i=1}^k u_i = 1. \text{ Для } Q \in \text{Int}(\mathbb{R}_+^n) \text{ определим}$$

$$\tilde{S} = \tilde{Q} + (\tilde{U} - \tilde{W}\tilde{Q}) \cdot W. \text{ Тогда для любого } a > 0 \text{ имеем } \tilde{S} = \alpha \tilde{Q} + (\tilde{U} - \tilde{W} \cdot \alpha \tilde{Q}) \cdot W.$$

Доказательство. Так как $W \cdot \alpha Q = \alpha \cdot WQ$, то $\widetilde{W} \cdot \alpha \widetilde{Q} = \widetilde{WQ} + (\log_2 \alpha) \cdot [1, \dots, 1]$. Поэтому верно $\tilde{S} = \tilde{Q} + (\tilde{U} - \widetilde{WQ}) \cdot W = \tilde{Q} + (\tilde{U} - \widetilde{W \cdot \alpha Q}) \cdot W + (\log_2 \alpha) \cdot [1, \dots, 1] \cdot W = \tilde{Q} + (\tilde{U} - \widetilde{W \cdot \alpha Q}) \cdot W + (\log_2 \alpha) \cdot [1, \dots, 1] = \alpha \tilde{Q} + (\tilde{U} - \widetilde{W \cdot \alpha Q}) \cdot W$.

3. ДОКАЗАТЕЛЬСТВО СХОДИМОСТИ АЛГОРИТМА

Лемма 6.

Пусть $k \times n$ -матрица $W = \|w_{i,j}\|$ и вектор-столбец $U = (u_1, \dots, u_k)$ удовлетворяют условиям нормировки: $\forall j \left[\sum_{i=1}^k w_{i,j} = 1 \right]$, $\forall i \forall j [w_{i,j} > 0]$, $\forall i [u_i > 0]$ и $\sum_{i=1}^k u_i = 1$. Для $Q \in \text{Int}(\mathbb{R}_+^n)$ определим $\tilde{S} = \tilde{Q} + (\tilde{U} - \widetilde{WQ}) \cdot W$. Тогда для любого р.в. $P \in \text{Int}(\mathbb{R}_+^n) \cap \Lambda_{W,U}$ выполняется $\langle \tilde{P} - \tilde{S} | P \rangle \geq 0$.

Доказательство. Имеем $s_j = q_j \cdot 2^{\langle \tilde{U} - \widetilde{WQ} | We_j \rangle}$. Поэтому

$$\begin{aligned} \langle \tilde{P} - \tilde{S} | P \rangle &= \sum_{j=1}^n \left(\log_2 \frac{p_j}{s_j} \right) \cdot p_j = \\ &= \sum_{j=1}^n \left(\log_2 \left(\frac{p_j}{q_j} \right) - \langle \tilde{U} - \widetilde{WQ} | We_j \rangle \right) \cdot p_j = \\ &= \sum_{j=1}^n \log_2 \left(\frac{p_j}{q_j} \right) \cdot \left(\sum_{i=1}^k w_{i,j} \right) \cdot p_j - \\ &- \sum_{j=1}^n \left(\sum_{i=1}^k \log_2 \left(\frac{u_i}{\sum_{l=1}^n w_{i,l} \cdot q_l} \right) \cdot w_{i,j} \right) \cdot p_j = \\ &= \sum_{i=1}^k \sum_{j=1}^n \log_2 \left(\frac{w_{i,j} \cdot p_j}{w_{i,j} \cdot q_j} \right) \cdot (w_{i,j} \cdot p_j) - \\ &- \sum_{i=1}^k \sum_{j=1}^n \log_2 \left(\frac{\sum_{l=1}^n w_{i,l} \cdot p_l}{\sum_{l=1}^n w_{i,l} \cdot q_l} \right) \cdot (w_{i,j} \cdot p_j) = \\ &= \sum_{i=1}^k \sum_{j=1}^n \log_2 \left(\frac{w_{i,j} \cdot p_j / \sum_{l=1}^n w_{i,l} \cdot p_l}{w_{i,j} \cdot q_j / \sum_{l=1}^n w_{i,l} \cdot q_l} \right) \cdot (w_{i,j} \cdot p_j) = \\ &= \sum_{i=1}^k \left[\left(\sum_{j=1}^n w_{i,j} \cdot p_j \right) \cdot \log_2 \left(\frac{w_{i,j} \cdot p_j / \sum_{l=1}^n w_{i,l} \cdot p_l}{w_{i,j} \cdot q_j / \sum_{l=1}^n w_{i,l} \cdot q_l} \right) \right] \geq 0 \end{aligned}$$

Лемма 1 влечет, что

$$\sum_{j=1}^n \log_2 \left(\frac{w_{i,j} \cdot p_j / \sum_{l=1}^n w_{i,l} \cdot p_l}{w_{i,j} \cdot q_j / \sum_{l=1}^n w_{i,l} \cdot q_l} \right) \geq 0.$$

Условие $P \in \Lambda_{W,U}$ позволяет заменить u_i на

$$\sum_{i=1}^n w_{i,l} \cdot p_l.$$

Лемма 7.

Пусть $k \times n$ -матрица $W = \|w_{i,j}\|$ и вектор-столбец $U = (u_1, \dots, u_k)$ удовлетворяют условиям нормировки: $\forall j \left[\sum_{i=1}^k w_{i,j} = 1 \right]$, $\forall i \forall j [w_{i,j} > 0]$, $\forall i [u_i > 0]$ и $\sum_{i=1}^k u_i = 1$. Пусть $P \in \text{Int}(\mathbb{R}_+^n) \cap \Lambda_{W,U}$. Для р.в. $Q \in \text{Int}(\mathbb{R}_+^n) \cap \Xi_{W,P}$ определим

$$\tilde{S} = \tilde{Q} + (\tilde{U} - \widetilde{WQ}) \cdot W.$$

Тогда если $\langle \tilde{P} - \tilde{Q} | P \rangle$ достаточно мало, то $\sum_{j=1}^n s_j \leq 1$.

Доказательство. $\tilde{S} = \tilde{Q} + (\tilde{U} - \widetilde{WQ}) \cdot W$ в координатах имеет вид $s_j = q_j \cdot 2^{\langle \tilde{U} - \widetilde{WQ} | We_j \rangle}$. Ясно, что при фиксированном $U = WP$ малость $\langle \tilde{P} - \tilde{Q} | P \rangle$ эквивалентна малости $\varepsilon = \sum_{j=1}^n \left| \langle \tilde{U} - \widetilde{WQ} | We_j \rangle \right|$. Тогда

$$\begin{aligned} \sum_{j=1}^n s_j - 1 &= \sum_{j=1}^n s_j - \sum_{j=1}^n q_j = \\ &= \sum_{j=1}^n q_j \cdot \left(2^{\langle \tilde{U} - \widetilde{WQ} | We_j \rangle} - 1 \right) = \\ &= \sum_{j=1}^n q_j \cdot \left(1 + \ln 2 \cdot \langle \tilde{U} - \widetilde{WQ} | We_j \rangle + o(\varepsilon) - 1 \right) = \\ &= \ln 2 \cdot \sum_{j=1}^n q_j \cdot \langle \tilde{U} - \widetilde{WQ} | We_j \rangle + o(\varepsilon) = \\ &= \ln 2 \cdot \langle \tilde{U} - \widetilde{WQ} | WQ \rangle + o(\varepsilon) = \\ &= -\ln 2 \cdot \langle \widetilde{WQ} - \tilde{U} | WQ \rangle + o(\varepsilon) = \\ &= -\ln 2 \cdot \delta(WQ|U) + o(\varepsilon) < 0 \end{aligned}$$

Неравенство следует из Леммы 1.

Теорема 3.

Пусть $k \times n$ -матрица $W = \|w_{i,j}\|$ и вектор-столбец $U = (u_1, \dots, u_k)$ удовлетворяют условиям нормировки: $\forall j \left[\sum_{i=1}^k w_{i,j} = 1 \right]$, $\forall i \forall j [w_{i,j} > 0]$, $\forall i [u_i > 0]$ и $\sum_{i=1}^k u_i = 1$. Для р.в. $Q = (q_1, \dots, q_n) \in \text{Int}(\mathbb{R}_+^n)$

(с $\sum_{j=1}^n q_j = 1$) и $\Lambda_{W,U} \cap \text{Int}(\mathbb{R}_+^n) \neq \emptyset$ имеем $R_t \rightarrow \Pi_{\Lambda_{W,U}}(Q)$ при $t \rightarrow \infty$, для $R_0 = Q$ и $\widetilde{R}_{t+1} = \widetilde{R}_t + (\widetilde{U} - \widetilde{WR}_t) \cdot W$, где $R_t = (r_1^{(t)}, \dots, r_n^{(t)})$.

Доказательство. По Лемме 6 имеем

$$0 \leq \langle \tilde{P} - \widetilde{R}_{t+1} | P \rangle.$$

Выполняется также

$$\begin{aligned} \langle \tilde{P} - \widetilde{R}_{t+1} | P \rangle &= \langle \tilde{P} - (\widetilde{R}_t + (\widetilde{U} - \widetilde{WR}_t) \cdot W) | P \rangle = \\ &= \langle \tilde{P} - \widetilde{R}_t | P \rangle - \langle (\widetilde{U} - \widetilde{WR}_t) \cdot W | P \rangle = \\ &= \langle \tilde{P} - \widetilde{R}_t | P \rangle - \langle (\widetilde{U} - \widetilde{WR}_t) | WP \rangle = \\ &= \langle \tilde{P} - \widetilde{R}_t | P \rangle - \langle (\widetilde{U} - \widetilde{WR}_t) | U \rangle \end{aligned}$$

Так как

$$\left\{ (x_1, \dots, x_n) \in \mathbb{R}_+^n : 0 \leq \langle \tilde{P} - \tilde{X} | P \rangle \leq \langle \tilde{P} - \tilde{Q} | P \rangle = \right. \\ \left. = \delta(R|Q) \right\}$$

компактно, то найдется такая подпоследовательность, что $R_{t_l} \rightarrow R$ для некоторого R при $l \rightarrow \infty$.

Переходя к пределу $l \rightarrow \infty$ получим $\langle \tilde{P} - \tilde{R} | P \rangle = \langle \tilde{P} - \tilde{R} | P \rangle - \langle \tilde{U} - \widetilde{WR} | U \rangle$ для любого

$P \in \Lambda_{W,U}$. Поэтому $\langle \tilde{U} - \widetilde{WR} | U \rangle = 0$, тогда $WR = U$,

то есть $R \in \Lambda_{W,U}$. Лемма 4 влечет $R = \Pi_{\Lambda_{W,U}}(Q)$. Для $R_t = (r_1^{(t)}, \dots, r_n^{(t)})$ обозначим через $\|R_t\|_1 = \sum_{j=1}^n r_j^{(t)}$ и

введем $\overline{R}_t = \frac{R_t}{\|R_t\|_1} = \left(\frac{r_1^{(t)}}{\|R_t\|_1}, \dots, \frac{r_n^{(t)}}{\|R_t\|_1} \right)$. $\langle \tilde{R} - \overline{R}_t | R \rangle$ бу-

дет малым для достаточно больших l . По Лемме 5

(для $a = \frac{R_t}{\|R_t\|_1}$) имеем

$$\widetilde{R}_{t+1} = \widetilde{R}_t + (\widetilde{U} - \widetilde{WR}_t) \cdot W = \widetilde{R}_t + (\widetilde{U} - \widetilde{W}\overline{R}_t) \cdot W.$$

Для $t < t_l$ Лемма 7 влечет $1 - \sum_{j=1}^n r_j^{(t+1)} \leq 0$.

Тогда

$$\begin{aligned} \delta(R|R_{t+1}) &= \langle \tilde{P} - \widetilde{R}_{t+1} | P \rangle - \\ &- (\log_2 e) \cdot \left(1 - \sum_{j=1}^n r_j^{(t+1)} \right) \leq \langle \tilde{P} - \widetilde{R}_{t+1} | P \rangle = \\ &= \langle \tilde{P} - (\widetilde{R}_t + (\widetilde{U} - \widetilde{WR}_t) \cdot W) | P \rangle = \\ &= \langle \tilde{P} - \widetilde{R}_t | P \rangle - \langle (\widetilde{U} - \widetilde{WR}_t) \cdot W | P \rangle = \\ &= \langle \tilde{P} - \widetilde{R}_t | P \rangle - \langle (\widetilde{U} - \widetilde{WR}_t) | WP \rangle = \\ &= \delta(R|R_t) + (\log_2 e) \cdot (\|R\|_1 - \|R_t\|_1) - \langle \tilde{U} - \widetilde{WR}_t | WP \rangle = \\ &= \delta(R|R_t) - \left(\langle \tilde{U} - \widetilde{WR}_t | U \rangle - (\log_2 e) \cdot (\|U\|_1 - \|WR_t\|_1) \right) = \\ &= \delta(R|R_t) - \delta(U|WR_t) \end{aligned}$$

Лемма 1 влечет $0 \leq \delta(R|R_{t+1}) < \delta(R|R_t) < \delta(R|R_{t_l})$ при $t_l < t < t_{l+1}$. Переходя к пределу $l \rightarrow \infty$, получим $R_t \rightarrow R = \Pi_{\Lambda_{W,U}}(Q)$ при $t \rightarrow \infty$, что и требовалось доказать.

ЗАКЛЮЧЕНИЕ

Опишем теперь классическое приложение описываемого алгоритма – поиск оценки максимального правдоподобия в экспоненциальном семействе распределений вероятностей.

Оценка *максимального правдоподобия* (МП-оценка) для последовательности $x_1^n = (x_1, \dots, x_n) \in A^n$

(и эмпирического распределения $P_{x_1^n} = \left(\frac{n_1}{n}, \dots, \frac{n_{|A|}}{n} \right)$,

где $|\{k : x_k = a_j\}| = n_j$) внутри семейства Φ р.в. на

$A = \{a_1, \dots, a_{|A|}\}$ – это такое р.в. $P \in \Phi$, что

$$P^n(x_1, \dots, x_n) = \max_{Q \in \Phi} Q^n(x_1, \dots, x_n).$$

Лемма 8.

МП-оценка $\arg \max_{P \in \Phi} P^n(x_1, \dots, x_n) = \arg \min_{P \in \Phi} \delta(P_{x_1^n} | P)$,

где эмпирическое распределение $P_{x_1^n} = \left(\frac{n_1}{n}, \dots, \frac{n_{|A|}}{n} \right)$

определяется формулами $n_j = |\{k : x_k = a_j\}|$.

Доказательство – легкая проверка.

Теорема 4.

Пусть

$$\Phi = \left\{ R \in \text{Int}(\mathbb{R}_+^n) \mid \exists \theta_{m+1} \exists \Theta = \right. \\ \left. = (\theta_1, \dots, \theta_m) [\tilde{R} = \Theta \cdot M + \theta_{m+1} + \tilde{Q}] \right\}$$

для некоторой $m \times n$ -матрицы M и р.в. Q на $\{1, \dots, n\}$ такого, что $\forall j [Q(j) > 0]$. Тогда

МП-оценка $\arg \max_{P \in \Phi} P^n(x_1, \dots, x_n) = \Pi_{\Lambda_{M,V}}(Q)$ для

$V = M \cdot P_{x_1^n}$, если $\text{Int}(\mathbb{R}_+^n) \cap \Lambda_{M,V} \neq \emptyset$.

Доказательство. По заданию V имеем $P_{x_1^n} \in \Lambda_{M,V}$.

По Лемме 4 выполняется тождество Пифагора

$$\delta(P_{x_1^n} | Q) = \delta(P_{x_1^n} | \Pi_{\Lambda_{M,V}}(Q)) + \delta(\Pi_{\Lambda_{M,V}}(Q) | Q).$$

Если $S \in \Phi$, т.е.

$$\exists \vartheta_{m+1} \exists \bar{\Theta} = (\vartheta_1, \dots, \vartheta_m) [\tilde{S} = \bar{\Theta} \cdot M + \vartheta_{m+1} + \tilde{Q}],$$

то $\Phi = \Xi_{M,S}$, так как

$$\begin{aligned} \tilde{R} &= \Theta \cdot M + \theta_{m+1} + \tilde{Q} = \tilde{R} = \\ &= (\Theta - \bar{\Theta}) \cdot M + (\theta_{m+1} - \vartheta_{m+1}) + \tilde{S} \end{aligned}$$

По Лемме 4

$$\{ \Pi_{\Lambda_{M,V}}(Q) \} = \Xi_{M,Q} \cap \Lambda_{M,V} = \Xi_{M,S} \cap \Lambda_{M,V} = \{ \Pi_{\Lambda_{M,V}}(S) \}.$$

Тождество Пифагора

$$\delta(P_{x_1^n} | S) = \delta(P_{x_1^n} | \Pi_{\Lambda_{\bar{M}, \bar{Y}}}(Q)) + \delta(\Pi_{\Lambda_{\bar{M}, \bar{Y}}}(Q) | S)$$

влечет $\delta(P_{x_1^n} | \Pi_{\Lambda_{\bar{M}, \bar{Y}}}(Q)) \leq \delta(P_{x_1^n} | S)$ для любого $S \in \Phi$.

Применение Леммы 8 завершает доказательство.

Представляется интересным сравнить скорость сходимости предложенного алгоритма с градиентным спуском и применить его программную реализацию для поиска оценок максимального правдоподобия и решения других задач статистического анализа данных.

* * *

Автор благодарит М.А. Михеенкову за поддержку, проф. Е.М. Бениаминова и проф. Г.Б. Шабата за полезные обсуждения.

СПИСОК ЛИТЕРАТУРЫ

1. Darroch J., Ratcliff D. Generalized iterative scaling for log-linear models // Annals Math. Statist. – 1972. – Vol. 43. – P. 1470-1480.

2. Ченцов Н.Н. Статистические решающие правила и оптимальные выводы. – М.: Наука, 1972.
3. Csiszar I. I-divergence geometry of probability distributions and minimization problems // Annals Probab. – 1975. – Vol. 3. – P. 146-158.
4. Csiczar I., Shields P. Information theory and statistics: a tutorial. – Boston: Now Publishers Inc., 2004.

Материал поступил в редакцию 08.04.16.

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН и Российского государственного гуманитарного университета, Москва
e-mail: vinogradov.d.w@gmail.com

К определению семантической близости на основе связей объединенного тезауруса

Рассматриваются вопросы использования тезаурусов для нечеткого сопоставления понятийных образов. Предложены мера семантической близости, при расчете которой используются иерархические и ассоциативные связи тезауруса, и алгоритм формирования семантического пересечения понятийных образов по принципу совпадающих максимумов. На массиве текстов и поисковых понятийных образов диссертаций проведены экспериментальные исследования, которые показали, что за счет использования лексики различных предметных областей политематического тезауруса для пар понятийных образов удастся более адекватно установить семантическую близость. При этом мощность пересечения образов значительно возрастает за счет связанных мерой семантической близости пар дескрипторов, однако средняя степень попарного пересечения увеличивается всего на 1–2%, что говорит о несущественном «расширении» понятийного образа в случае его использования в качестве поискового при формировании выдачи в механизмах автоматизированного информационного поиска.

Ключевые слова: онтология, тезаурус, имманентные отношения, ситуативные отношения, синтагматические отношения, парадигматические отношения, мера семантической близости, принцип совпадающих максимумов

ВВЕДЕНИЕ

Информационный поиск, равно как и любой процесс поиска в человеческой деятельности, в общем случае выполняется опосредованно – через сопоставление образов, а не собственно реальных объектов. То есть образ объекта – кандидата на отбор – соотносится с образом целевого объекта (возможно, существующего только гипотетически). В основной деятельности человека такого рода образами обычно являются описания объектов или процессов (научные, технические публикации, математические модели, проектная документация и т.п.), в той или иной форме представляющие определенные свойства и поведение объекта с полнотой и точностью, обусловленной характером решаемой практической задачи. В ИПС это поисковые образы документов – дескриптивное описание основных понятий и, возможно, связей, обычно отражающее особенности не только предметной области (ПрО), но и поискового информационного ресурса.

Такой подход позволяет реализовать эффективную (по используемым ресурсам) процедуру: объект с потенциально бесконечным разнообразием свойств и связей заменяется в операции сопоставления компактным, хорошо формализованным образом – иден-

тификатором, свойства которого представлены в форме, адекватной операционной среде.

Описание любого объекта является некоторой моделью, отражающей только те его свойства и поведение, которые являются существенными для конкретной задачи и рассматриваемой ПрО. Но при этом, для того чтобы описание было адекватно воспринято (был «узнан» искомый объект) и понято¹, оно должно быть представлено понятиями той степени общности/детальности и зафиксировано теми терминами, которые будут иметь одинаковые значения как для генератора, так и для приемника сообщения. И если унифицированность понятийно-терминологической составляющей более или менее успешно решается путем нормализации лексики (словари, тезаурусы), то представление связей и поведения описываемого

¹ Отметим, что успешность восприятия может быть оценена состоянием, когда термины полученного сообщения имеют эквивалентное отображение в понятийно-терминологической системе приемника и образуют логически непротиворечивую структуру. То есть, когда с помощью построенной по полученному сообщению понятийной системы можно воспроизвести гипотетический объект, обладающий предполагаемыми свойствами и поведением (в частности, свойствами объекта-оригинала).

объекта до сих пор не имеет устойчивого и рационального решения.

Задача информационно-поискового языка (ИПЯ) – иметь эффективные (сравнительно простые и выразительные) средства идентификации не только самих объектов и процессов, но и их отношений. И описательные возможности ИПЯ определяются тем, насколько поисковый образ семантически соответствует оригиналу, и тем, насколько адекватным будет способ соотнесения образов.

Формализации и сравнительной оценке мер семантической близости посвящено довольно много исследований. Спектр задач обработки естественного языка, в которых они нашли применение, довольно широк – устранение неоднозначности [1], обнаружение синонимов [2], автоматическая проверка орфографии, обнаружение и исправление ошибок [3], автоматическое индексирование [4], расширение поискового запроса [5]. Можно отметить работы, содержащие довольно развернутые обзоры, посвященные мерам семантической близости [6, 7]. Однако в каждом конкретном случае формализация меры и ее исследование основаны на конкретном (по большей части – лингвистическом) ресурсе: в качестве такого ресурса в основном рассматривается WordNet, называемый в ряде публикаций тезаурусом или онтологией с таксономическими отношениями.

Цель настоящей статьи – предложить и обосновать метод определения семантической близости, используя параметры и структурные свойства семантических сетей информационно-поисковых тезаурусов.

1. СЕМАНТИЧЕСКИЕ ОТНОШЕНИЯ И ИХ ИДЕНТИФИКАЦИЯ

Задача повышения смысловыразительных возможностей ИПЯ решается уже несколько десятилетий. Средства выражения синтагматических отношений присутствовали уже в ИПЯ первых поколений автоматизированных ИПС. ИПЯ СИНТОЛ [8], разработанный в Национальном центре научных исследований Франции (автор – Жан-Клод Гарден), оперировал как имманентными, так и ситуативными отношениями и имел в качестве языковой единицы синтагму, обозначающую ситуацию или ситуативный класс предметов. Грамматика ИПЯ RX-кодов (Институт кибернетики АН УССР) позволяла фиксировать не только факт наличия ситуативных отношений между предметами, но и их характер [9]. Однако эти языки, для того чтобы адекватно выполнять свою функцию идентификации содержания, должны были быть грамматически ориентированы на конкретную предметную область или отрасль наук, т.е. контекст использования терминологии должен был быть зафиксирован.

В [9] предложена классификация ИПЯ, в основе которой рассматривается наличие в языке средств формирования данных об отношениях между предметами. Классификация опирается на два типа отношений – имманентные и ситуативные.

Имманентные отношения (парадигматические – на языковом уровне) отражают постоянные, стабильные для рассматриваемой предметной области связи между предметами. В практике информационной

деятельности такие отношения фиксируются в информационно-поисковых тезаурусах и делятся на три основных подтипа – иерархические, ассоциативные, отношения эквивалентности (синонимия). Иерархические отношения отражают на вещном уровне имманентные связи «класс–подкласс» и реализуются лексико-семантическими связями между понятиями типа «род–вид» и «часть–целое». Ассоциативные отношения, согласно [10], включают все типы связей, кроме «род–вид» и синонимии (например, «причина–следствие», «процесс–субъект», «процесс–объект»). Следует отметить, что иерархические отношения в тезаурусе связывают понятия, выраженные словами естественного языка, а не сами объекты или предметы (что справедливо для таксономии), поэтому нарушается правильная структура дерева – один и тот же термин может иметь несколько «родителей» – вышестоящих терминов на предыдущем уровне. На рис. 1 изображен фрагмент тезауруса INIS, представляющий собой понятийную сеть по отношению к связям² с вышестоящими дескрипторами. Дескрипторы «Силастик», «Силиконы», «Резины», «Полиизопрен», «Пластмассы», «Неопрен», «Полиэфиры», «Органические полимеры», «Гидрофильные полимеры» входят в несколько иерархий, причем некоторые из них (например, «Гидрофильные полимеры» и «Силиконы») в различных иерархиях находятся на разных уровнях.

Ситуативные отношения, в отличие от имманентных, воспринимаются как временные, возникшие в определенной ситуации. И если имманентные отношения присущи предметной области в целом, то ситуативные связывают объекты ПрО в рамках фрагмента «реального мира» (соответствующего некоторому процессу), к тому же и изменяющегося во времени. Этим объясняется сложность типизации ситуативных отношений: модели, создаваемые для представления «реального мира», зависят от целей и характера решаемых задач, от того, как формулируется и ограничивается фрагмент ПрО, т.е. от конкретной концептуальной модели этой области.

Отдельный документ информационного массива представляет собой описание конкретной ситуации (взаимодействующих объектов реального мира). При этом он может (и должен) быть отнесен к одной или нескольким предметным областям, представляющим этот фрагмент в том или ином аспекте. Формирование идентифицирующего (например, поискового) образа документа основано на сведениях содержания, представленного средствами естественного языка, к хорошо формализованной структуре, для представления (фиксации) которой используется тот же либо другой язык. При этом структурная редукция должна привести документ к форме, наиболее удобной для манипулирования в вычислительной среде, а семантическая – по возможности сохранить (или, по крайней мере, обозначить наличие в случае идентифицирующего образа) исходные, прежде всего существенные для этой ПрО смыслы.

² Здесь и далее используются следующие обозначения связей: BT (Broader Term) – вышестоящий дескриптор, RT (Related Term) – ассоциативный дескриптор.

Наиболее распространенные ИПС (в том числе и поисковые машины Интернет) ориентируются на линейную (векторную) модель формирования поискового образа, в котором присутствуют независимые (не связанные между собой) термины (слова или словосочетания) как идентификаторы содержания документа. Термины (в том числе и на естественном языке) обозначают объекты и процессы фрагмента реального мира, представленного в конкретном документе, и могут отражать состояние ПрО в целом или носить ситуативный характер. В первом случае – это термины, обозначающие

устоявшиеся понятия, присутствующие в специализированных словарях, учебниках, учебных пособиях и т.п., во втором – термины, обозначающие понятия, которые либо имеют эпизодический характер, либо претендуют на новизну в рамках ПрО.

В такой модели для реализации имманентных отношений могут быть использованы информационно-поисковые тезаурусы как сети понятий ПрО, обозначаемых дескрипторами. Это позволит расширить представление семантики поискового образа за счет парадигматических связей между дескрипторами.

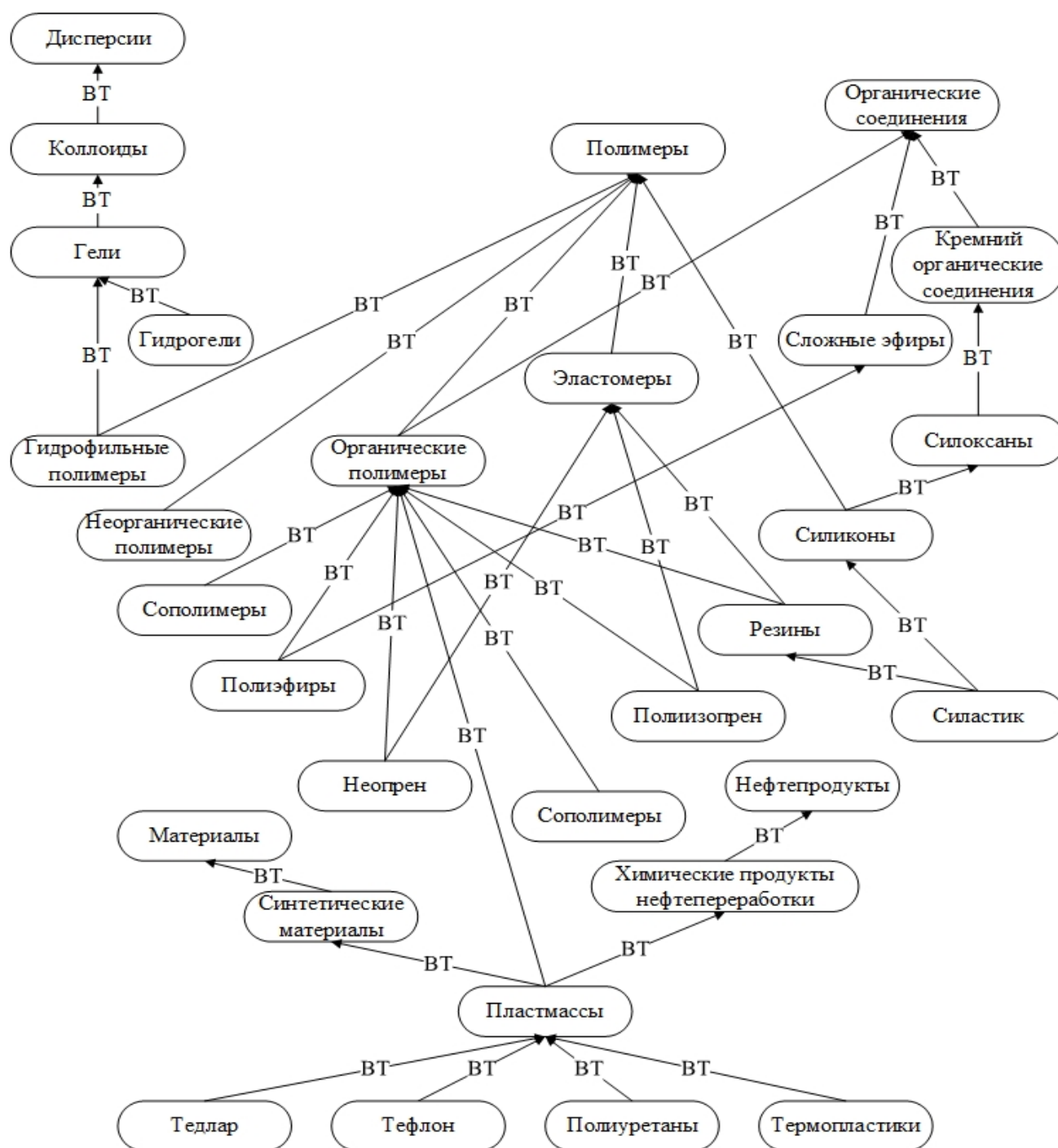


Рис. 1. Фрагмент тезауруса INIS

Однако любой документ, отражая реальную действительность, в первую очередь интересен именно как отражение отношений ситуативных, и его поисковый образ при этом должен фиксировать не только связь документа с известным знанием, но и содержать элементы новизны, по которым в основном он и должен извлекаться, например, в процедурах тематического поиска. В лингвистике такие отношения, отражающие на языковом уровне ситуативные отношения ПрО, называют синтагматическими.

На рис. 2 и 3 представлены фрагменты графов, отражающих синтагматические (сложившиеся в конкретном контексте диссертации на тему «Синтез и структура пленок на основе гидроксиапатита») и парадигматические (в тезаурусе INIS) отношения между понятиями «Стоматология», «Костная ткань» и «Синтетические материалы». И если в тексте документа между этими понятиями существует непосредственная функциональная связь, то в тезаурусе между ними в лучшем случае могут быть проложены маршруты разной длины (если учитывать, что симметрично связи к вышестоящему дескриптору в тезаурусе присутствует связь к нижестоящему, не обозначенная в явном виде на рисунке).

Идентифицирующий образ как информационный объект, представляющий собой модель некоторой предметной области, с необходимостью должен отразить те свойства и связи объекта, которые являются существенными для конкретной задачи и обстоятельств, связанных с целями и методами создания или использования объекта, причем с учетом того, что задачи и обстоятельства могут меняться. То есть отдельная ПрО может быть представлена разными понятийными и терминологическими множествами, отражающими её с разной степенью полноты и детальности.

В этом смысле онтологический³ подход к формированию образа (описания) знаний представляется конструктивным, поскольку позволяет идентифицировать объект (информацию, знания, предметы и т.п.) не только на основе его свойств, но и его «поведения» в концептуальном и историческом пространствах.

В [11] предложено определение онтологии с точки зрения идентификации содержания документа при информационном поиске. Онтология рассматривается как совокупность трех систем – функциональной, понятийной и терминологической. Назначение функциональной системы – представить на знаковом уровне «рабочий интерфейс» онтологии, т.е. существенные с точки зрения задач пользователя объекты и способы их совместного существования и использования через ситуативные отношения. «Функциональность» системы обусловлена конкретной моделью выделения ситуативных отношений и их идентификации посредством синтагматических связей. Понятийная система онтологии как логико-семантический базис фиксирует устойчивые понятия ПрО и имманентные связи, выполняя роль «посредника» в операционном пространстве: при сопоставлении или преобразовании онтологий их необходимо рассматривать в рамках общего понятийного контекста, сложившегося в предметной области. В задачах информационного поиска в качестве такого базиса применяются информационно-поисковые тезаурусы. Соответственно, терминологическая система, основное назначение которой – зафиксировать на носителе смысловой образ, отражает свойства языка на уровне знаков – терминов, которые могут быть связаны отношениями эквивалентности (синонимии в естественном языке) и включения (образования словосочетаний по правилам естественного языка).



Рис. 2. Фрагмент графа диссертации

³ Отметим, что термин «онтология» в дальнейшем изложении будет использован в узком смысле поисковых задач – как адаптивный композиционный идентификатор, отражающий основные свойства объекта в сопоставимой форме.

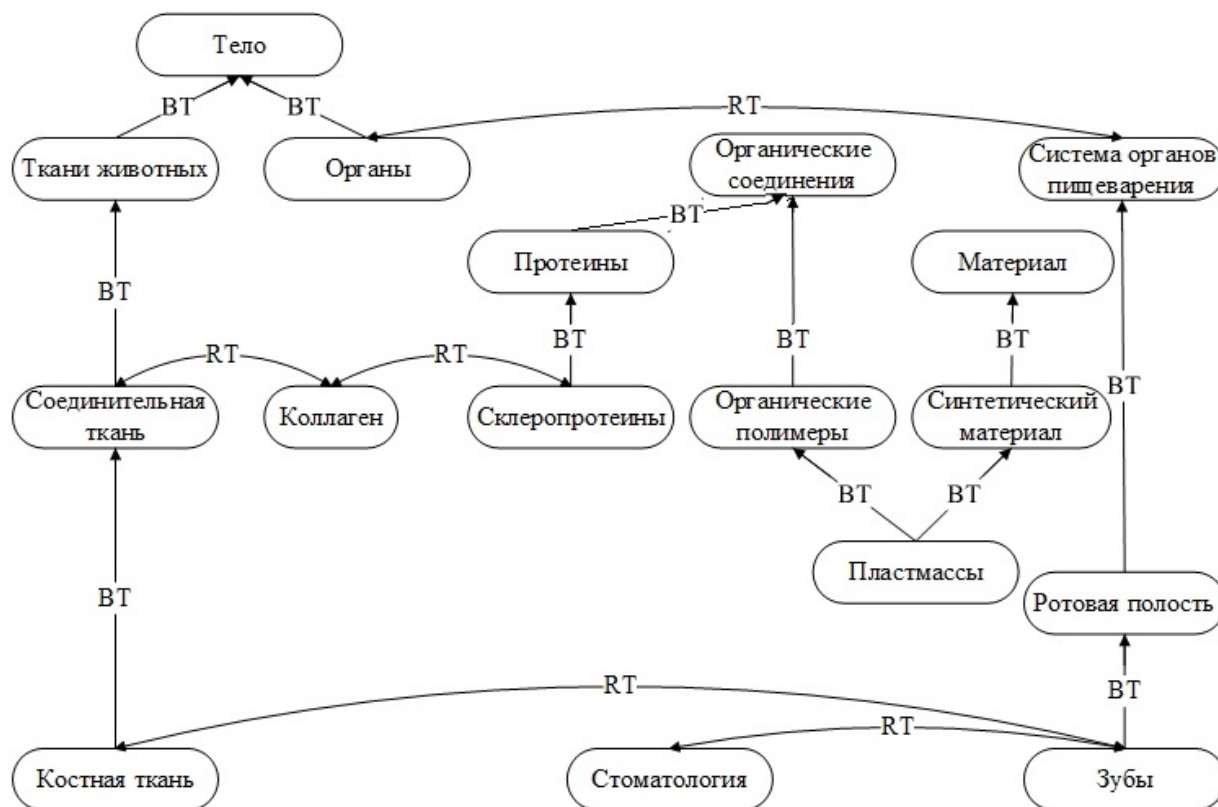


Рис. 3. Фрагмент тезауруса INIS для дескрипторов «Стоматология», «Костная ткань» и «Синтетические материалы»

Введенные в [11] операции над онтологиями используют при выделении семантически схожих элементов нечеткую функцию подобия объектов и отношений как меру семантической близости элементов знакового уровня. Нечеткость функции подобия обусловливается тем, что при сравнении терминов, обозначающих предметы, рассматриваются не только случаи их полного тождества, лексикографического включения или синонимии естественного языка, но и ситуации наличия между терминами парадигматических связей. Именно поэтому операции над онтологиями должны рассматриваться в рамках конкретной ПрО, а соотносимые онтологии должны иметь общую понятийную и терминологическую системы.

2. МЕРЫ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ

Рассматривая меру близости как численную величину, показывающую степень схожести двух объектов, необходимо определить понятие семантической близости, когда в качестве объектов выступают описания, построенные с помощью терминов, обозначающих предметы (классы предметов) – знаки устойчивых понятий предметной области. По аналогии с [8] будем говорить, что между понятиями существует семантическая близость, если между обозначаемыми ими классами предметов существуют имманентные отношения. Наличие таких отношений фиксируется в специализированных структурах, к которым можно отнести таксономии, информационно-поисковые тезаурусы, онтологии.

На понятийном уровне имманентные отношения представлены парадигматическими связями. Парадигматические связи условно разделим на два класса – «вертикальные» (иерархические – в основном, «род-вид» и «часть-целое») и «горизонтальные» (ассоциативные – прочие, устойчивые для предметной области). В соответствии с такой схемой деления рассмотрим следующую типологию существующих на текущий период подходов к определению (расчету) меры семантической близости (рис. 4).

При расчете мер с использованием только «вертикальных» связей без множественного наследования (что характерно для таксономии) применяются следующие параметры:

- глубина понятий в иерархии;
- глубина ближайшего общего понятия в иерархии;
- множество (количество) потомков каждого понятия;
- множество (количество) листьев среди потомков понятия;
- множество (количество) родителей каждого понятия;
- длина иерархического пути.

Расчет мер с использованием «вертикальных» связей с возможностью множественного наследования в основном опирается на следующие параметры:

- длина иерархического пути;
- множество (количество) родителей каждого понятия;
- множество (количество) родителей ближайшего общего понятия.

В качестве одной из основных моделей, обеспечивающих использование всех видов связей, можно рассматривать теоретико-множественную модель признаков, предложенную в [12, 13].

По методу Арапова-Ратцевой мера семантической близости для двух терминов вычисляется как отношение числа общих семантических множителей к числу всех семантических множителей. В качестве семантических множителей, например, могут рассматриваться термины, связанные с рассматриваемыми терминами парадигматическими отношениями.

В основе теоретико-множественного подхода Тверски – сопоставление общих ($A \cap B$) и различных ($A - B$, $B - A$) признаков объектов a и b :

$$S(a, b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$

где f – некоторая функция; θ, α, β – неотрицательные веса для общих и различных свойств, которые позволяют создать ассиметричную меру близости.

Нормализованная модель Тверски (ratio-model) позволяет установить интервал изменения меры (от 0 до 1):

$$S(a, b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}$$

Такие модели могут не учитывать видовые различия связей (считая все связи равноправными признаками), но могут и использовать вклад отдельного вида связей в аддитивной свертке по n критериям:

$$S(a, b) = \sum_{i=1}^n \omega_i S_i(a, b), \quad (1)$$

где S_i – мера близости по отдельному критерию, а ω_i – вес критерия, $\sum_{i=1}^n \omega_i = 1$.

В табл. 1 приведены формулы некоторых мер семантической близости в соответствии с предложенной выше типологией.

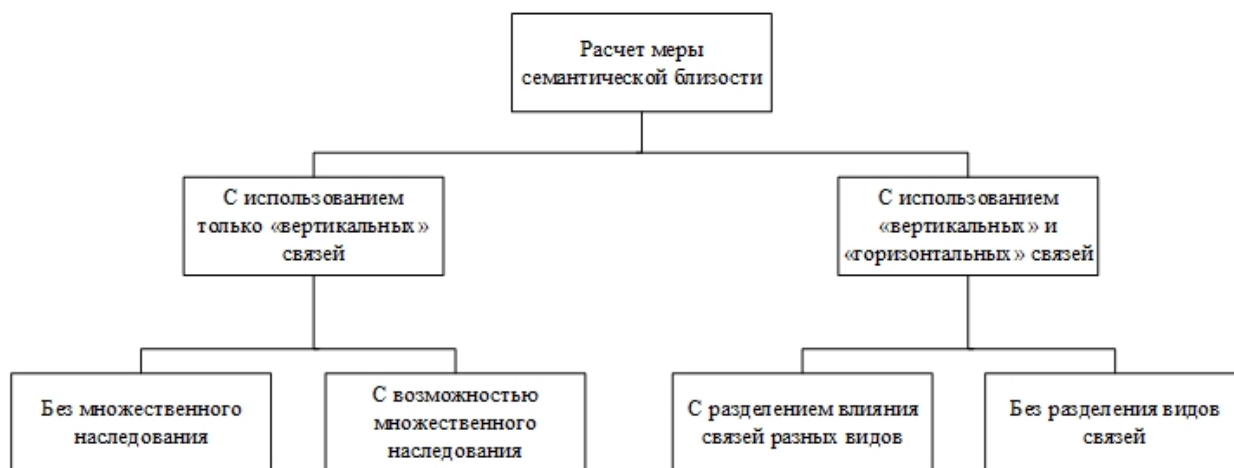


Рис. 4. Типология методов расчета мер семантической близости

Таблица 1

Меры семантической близости

№	Мера	Комментарий
Только «вертикальные» связи		
1	Без множественного наследования Принятые обозначения: $LCS(a, b)$ – ближайший общий родитель понятий a, b ; $IC(a)$ – информационное содержание понятия a ; $hypo(a)$ – множество потомков понятия a ; max_nodes – общее количество понятий в таксономии; $depth(a)$ – глубина понятия a в таксономии; max_depth – максимальная глубина таксономии; $Leaves(a)$ – множество листьев среди потомков понятия a ; max_leaves – общее количество листьев в таксономии; $length(a, b)$ – длина иерархического пути между понятиями a, b	

№	Мера	Комментарий
1a	$S(a, b) = \frac{2 \cdot \text{depth}(LCS(a, b))}{N_1 + N_2 + 2 \cdot \text{depth}(LCS(a, b))}$ N_1 и N_2 - длина пути до ближайшего общего родителя от понятия a и b соответственно	Для расчетов использовалась иерархическая сеть, построенная авторами [14]
1b	$S(a, b) = IC(LCS(a, b))$ Информационное содержание вычисляется по одной из формул 1d – 1f	Для экспериментов был использован лингвистический ресурс WordNet [15]
1c	$S(a, b) = \frac{2IC(LCS(a, b))}{IC(a) + IC(b)}$ Информационное содержание вычисляется по одной из формул 1d – 1f	Для экспериментов был использован лингвистический ресурс WordNet [2]
1d	$IC(a) = 1 - \frac{\log(\text{hypo}(a) + 1)}{\log(\text{max_nodes})}$	В [16] расчет информационного содержания основан только на иерархической структуре WordNet.
1e	$IC(a) = k \frac{\log(\text{hypo}(a) + 1)}{\log(\text{max_nodes})} + (1 - k) \frac{\log(\text{depth}(a))}{\log(\text{max_depth})}$	Для экспериментов был использован лингвистический ресурс WordNet, $k = 0,5$ [17]
1f	$IC(a) = k \frac{\log(\text{depth}(a))}{\log(\text{max_depth})} \left(1 - \frac{\log(\sum_{c \in \text{hypo}(a)} \left(\frac{1}{\text{depth}(c)} \right) + 1)}{\log(\text{max_nodes})} \right)$	Для экспериментов был использован лингвистический ресурс WordNet [18]
1g	$S(a, b) = -\log \left(\frac{\text{length}(a, b)}{2 \cdot \text{max_depth}} \right)$	Для экспериментов был использован лингвистический ресурс WordNet [19]
1h	$S(a, b) = e^{-\alpha \cdot \text{length}(a, b)} \frac{e^{\beta \cdot \text{depth}(LCS(a, b))} - e^{-\beta \cdot \text{depth}(LCS(a, b))}}{e^{\beta \cdot \text{depth}(LCS(a, b))} + e^{-\beta \cdot \text{depth}(LCS(a, b))}}$ α, β – параметры, определяющие влияние на меру длины пути и глубины общей вершины соответственно	Для экспериментов был использован лингвистический ресурс WordNet; $\alpha = 0,2$; $\beta = 0,6$ [20]
1i	$S(a, b) = \left(1 - \frac{\text{length}(a, b)}{\text{length}(a, b) + \text{depth}(LCS(a, b)) + \beta} \right) \times$ $\times \left(\frac{\text{depth}(LSC(a, b))}{\text{length}(a, b) + \frac{\text{depth}(LCS(a, b))}{2} + \alpha} \right)$	Для экспериментов был использован лингвистический ресурс WordNet; $\alpha = 0$, $\beta = 1$ [21]
2	С возможностью множественного наследования	
2a	$S(a, b) = \max_{i=1 \dots m} (\text{spec}^{s(P_i)} \text{gen}^{g(P_i)})$ $g(P)$ – количество ребер в направлении обобщения («выше»); $s(P)$ – количество ребер в направлении детализации («ниже»); $\text{gen}, \text{spec} \in [0, 1]$ – параметры, соответственно, характеризующие вклад в направлении обобщения и детализации. $\{P_1, P_2, \dots, P_m\}$ – множество возможных путей между понятиями a и b.	Рассматривается множество всех возможных иерархических путей между понятиями a и b ($P = \{c_1, c_2, \dots, c_n\}$, $c_1 = a$, $c_n = b$), с учетом множественного наследования [22].
2b	$S(a, b) = \frac{ UC(c_1, H^c) \cap UC(c_2, H^c) }{ UC(c_1, H^c) \cup UC(c_2, H^c) }$ $UC(c_i, H^c) = \{c_j \in H^c \mid \exists m = (c_i, x_{i_1}, x_{i_2}, \dots, x_{i_k}, c_j) \cup (c_i = c_j)\}$ m – маршрут, включающий в себя переходы $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ к вышестоящим понятиям	Предложена в [23] для кластеризации метаданных, описывающих ресурсы Semantic Web на основе онтологии

№	Мера	Комментарий
С использованием «вертикальных» и «горизонтальных» связей		
3	С разделением влияния связей разных видов	
3a	$S(a,b) = C - path(a,b) - k * d$ $path(a,b)$ – путь между понятиями a, b ; d – количество перегибов в пути, C и k – константы	В качестве лингвистического ресурса использовался тезаурус Роже ($C = 5, k = 1$) [24]
4	Без разделения видов связей	
4b	$S(a,b) = \rho \frac{ A(c_1) \cap A(c_2) }{ A(c_1) } + (1 - \rho) \frac{ A(c_1) \cap A(c_2) }{ A(c_2) }$ $\rho \in [0, 1]$ $A(c)$ – вычисляется итеративно: $A = \{c\};$ $extend(A) = A \cup \{y x \in A \wedge y \notin A \wedge r(x,y) \in T\}.$ где r – множество допущенных семантических отношений; T – множество связей между элементами онтологии.	Мера близости на основе контекстных множеств, которые определяются итеративно: в контекстное множество понятия включаются те понятия онтологии, которые достигнуты из данного [25]

2.1. Мера семантической близости, определяемая с учетом связей в информационно-поисковых тезаурусах

Приведенный выше обзор показывает, что аналитическое выражение меры семантической близости зависит от вводимых признаков самой семантической близости. В случае, когда мера вычисляется с использованием специализированных понятийных или лингвистических ресурсов, такие признаки могут быть связаны с видами отношений.

В информационно-поисковом тезаурусе, согласно [10], понятийная сеть формируется в основном иерархическими и ассоциативными отношениями (т.е. для каждого дескриптора тезауруса можно построить семантическое окружение из дескрипторов, имеющих с ним иерархические и/или ассоциативные связи).

Рассмотрим в качестве критериев в аддитивной мере (1) наличие у дескриптора двух видов связей и будем строить тезаурусную меру семантической близости между дескрипторами d_1 и d_2 в соответствии с выражением:

$$S(d_1, d_2) = \alpha \cdot S_H(d_1, d_2) + \beta \cdot S_A(d_1, d_2) \quad (2)$$

где $S_H(d_1, d_2)$ и $S_A(d_1, d_2)$ – меры семантической близости по иерархическим и ассоциативным связям соответственно, α и β – коэффициенты, определяющие вес связи ($\alpha + \beta = 1$).

При вычислении $S_H(d_1, d_2)$ необходимо учесть следующее:

1) иерархические связи в тезаурусе обладают свойством транзитивности;

2) понятийные деревья тезауруса могут пересекаться, что приводит к наличию нескольких «родителей» (т.е. нескольких вариантов перехода вверх по иерархии – см. рис. 1) и обуславливает невозможность применения мер для «вертикальных» связей, не учитывающих множественное наследование.

Рассмотрим предложенный в [23] (№ 2b в табл. 1) подход к вычислению семантической близости. Такой подход может быть сведен к методу Арапова-Ратцевой (или к теоретико-множественной модели Тверски с единичными весами), если считать семантическим множителем дескриптор, стоящий выше по иерархии. Таким образом, с учетом транзитивности иерархической связи необходимое множество семантических множителей формируют все вышестоящие дескрипторы (причем сам дескриптор как источник связей тоже включается в это множество), и мера будет определяться следующим выражением:

$$S_H(d_1, d_2) = \frac{|UC(d_1, H^{d_1}) \cap UC(d_2, H^{d_2})|}{|UC(d_1, H^{d_1}) \cup UC(d_2, H^{d_2})|} \quad (3)$$

Множество $UC(d_i, H^{d_i})$ содержит дескриптор d_i , а также всех его предков в иерархической цепочке – H^{d_i} :

$$UC(d_i, H^{d_i}) = \{d_j \in H^{d_i} | \exists m = (d_i, x_{i_1}, x_{i_2}, \dots, x_{i_k}, d_j) \cup (d_i = d_j)\}$$

где m – маршрут, соединяющий d_i и d_j через вышестоящие дескрипторы $x_{i_1}, x_{i_2}, \dots, x_{i_k}$.

Рассмотрим возможность применения аналогичного подхода для расчета $S_A(d_1, d_2)$. Ассоциативная связь в тезаурусе не транзитивна, следовательно, множество семантических множителей будут формировать только дескрипторы, находящиеся на расстоянии от d_1 (d_2), равном одной ассоциативной связи.

Согласно [10], ассоциативная связь обладает свойством симметричности, однако, поскольку тип связи не задан точно, конкретная семантика симметрии не может быть определена. Для построения множества семантических множителей рассмотрим ассоциативную связь как однонаправленную (выходящую), сформировав при этом «семантический дубль» дескриптора (рис. 5a, 5б).

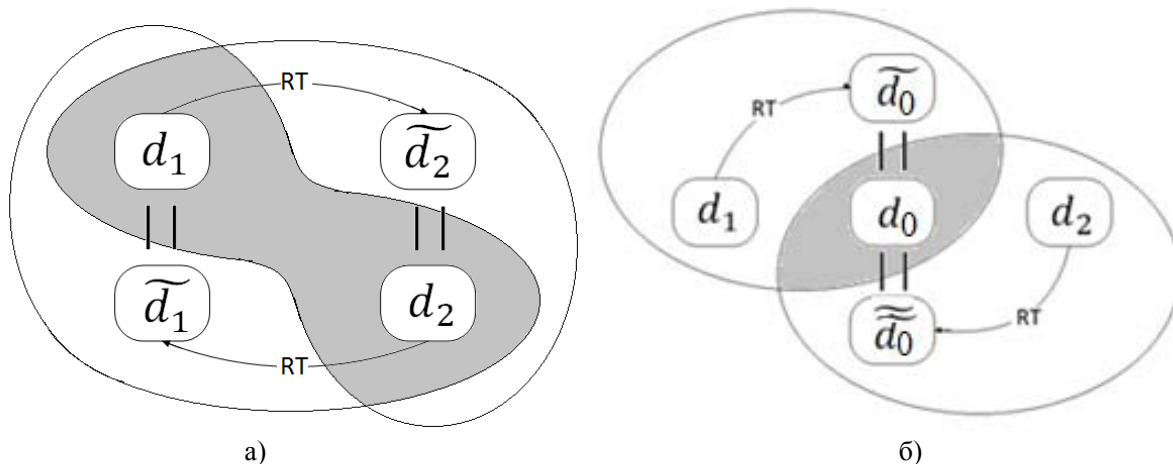


Рис. 5. Образование семантического дубля: а) объединение и пересечение семантических множителей для пары ассоциативных дескрипторов; б) объединение и пересечение семантических множителей для двух дескрипторов, ассоциативных третьему

Тогда ассоциативная связь на тезаурусном графе будет представлена как исходящая для дескриптора и как входящая – для его «семантического дубля». Множество семантических множителей $UC(d_i, A^{d_i})$ для ассоциативных связей по аналогии с (4) может быть построено следующим образом:

$$UC(d_i, A^{d_i}) = \{d_j \in A^{d_i} \mid \exists a = (d_i, \tilde{d}_j) \cup d_j \cup (d_i - d_j)\},$$

где a – ассоциативная связь, соединяющая d_i и $\tilde{d}_j$ ($\tilde{d}_j$ – «семантический дубль» дескриптора d_j). Расчет $S_A(d_1, d_2)$ в этом случае будет проводиться по формуле:

$$S_A(d_1, d_2) = \frac{|UC(d_1, A^{d_1}) \cap UC(d_2, A^{d_2})|}{|UC(d_1, A^{d_1}) \cup UC(d_2, A^{d_2})|}. \quad (5)$$

Для определения значений коэффициентов α и β рассматривается «вертикальное» и «горизонтальное» развитие конкретного информационно-поискового тезауруса:

$$\alpha = \frac{S_H}{S_H + S_A}, \quad \beta = \frac{S_A}{S_H + S_A}$$

где

$$S_H = \max_{d_1, d_2 \in T} S_H(d_1, d_2),$$

$$S_A = \max_{d_1, d_2 \in T} S_A(d_1, d_2),$$

T – множество дескрипторов тезауруса.

2.2. Модификация меры семантической близости для случая объединенного тезауруса

Тесная связь функциональной и понятийной систем (в рассматриваемом определении онтологии) требует обязательного задания предметной области

как для формирования, так и для использования онтологий. Однако в настоящее время, скорее, наблюдается устойчивая тенденция к созданию и использованию больших политематических баз данных, для логико-семантической поддержки которых необходимо иметь и политематический тезаурус.

В [26] приведены результаты структурно-статистического анализа лингвистического ресурса, построенного путем объединения информационно-поисковых тезаурусов различных областей знаний: науковедения, политологии, религиоведения, информатики, лингвистики, экономики, ядерной энергетики и др. Анализ лексики показал, что около 10% дескрипторов входят более чем в один тезаурус. Для таких дескрипторов предложено различать два типа полисемии: *парадигматическую полисемию*, т.е. многозначность на уровне соответствующих термину понятий, и *синтагматическую полисемию* – многоаспектность его использования.

В объединенном тезаурусе каждый дескриптор, с одной стороны, имеет связь с одним или несколькими тезаурусами, в которых он зафиксирован как понятие, а с другой стороны – входит в обобщенную понятийную сеть. Это означает, что связи дескриптора, принадлежащего более чем одному тезаурусу, могут представлять семантику различных предметных областей. Таким дескрипторам поставлено в соответствие множество содержащих его тезаурусов, разбитое на непересекающиеся подмножества по признаку парадигматической полисемии:

$$T(d_j) = U_{i=1}^k T_j^i, \quad T_j^l \cap T_j^m = \emptyset, \quad l \neq m,$$

т.е. для дескриптора рассматривается k различных семантических значений, каждое из которых дескриптор принимает в контексте⁴, определяемом соответствующим подмножеством тезаурусов.

⁴ Контекст определяется терминологическим окружением дескриптора.

Возможность дескриптора иметь несколько семантических значений учитывается при формировании множества семантических множителей: рассматриваются связи, принадлежащие только тезаурусам из подмножества, соответствующего текущему семантическому значению.

Пусть дескриптор d_1 имеет k семантических значений $v_1^1, v_1^2, \dots, v_1^k$, а дескриптор d_2 – p семантических значений $v_2^1, v_2^2, \dots, v_2^p$. Тогда выражения (3) и (5) примут следующий вид:

$$S_H(d_1, d_2) = \frac{|\bigcup_{i=1}^k UC(d_1, H^{v_i^1}) \cap \bigcup_{j=1}^p UC(d_2, H^{v_j^2})|}{|\bigcup_{i=1}^k UC(d_1, H^{v_i^1}) \cup \bigcup_{j=1}^p UC(d_2, H^{v_j^2})|}$$

$$S_A(d_1, d_2) = \frac{|\bigcup_{i=1}^k UC(d_1, A^{v_i^1}) \cap \bigcup_{j=1}^p UC(d_2, A^{v_j^2})|}{|\bigcup_{i=1}^k UC(d_1, A^{v_i^1}) \cup \bigcup_{j=1}^p UC(d_2, A^{v_j^2})|}$$

То есть, в условиях неопределенности при вычислении меры семантической близости учитываются все возможные парадигматические развития дескриптора.

Множество $UC(d, H^{v^i})$ для дескриптора d формируется по следующим правилам:

- 1) семантическому значению v^i ставится в соответствие множество тезаурусов (ST^i), парадигматические связи которых учитываются при определении вышестоящих понятий;
- 2) для дескриптора d множество ST^i равно множеству T^i ;
- 3) при перемещении вверх по иерархии множество ST^i может быть расширено, если семантическому значению дескриптора верхнего уровня соответствует множество тезаурусов

$$T_H : ST^i \cap T_H \neq T_H \rightarrow ST^i = ST^i \cup T_H.$$

При формировании множества $UC(d, A^{v^i})$ справедливы правила 1 и 2.

На рис. 6 приведен подграф объединенного тезауруса, содержащий все вышестоящие понятия для дескрипторов «Второе лицо» и «Бедренная кость».

Семантическая близость между этими дескрипторами имеет нулевое значение: по правилам формирования множества семантических множителей для единственного семантического значения дескриптора «Второе лицо» (в рамках рассматриваемого объединенного тезауруса – грамматическая категория глагола) в это множество не включается дескриптор «Голова», который мог бы обеспечить непустое пересечение семантических множителей.

Если же вычислять значение семантической близости между дескрипторами «Лицо» и «Бедренная кость», необходимо учесть оба семантических значения дескриптора «Лицо» (грамматическая категория и передняя часть головы человека), тогда мера семантической близости по «вертикальным» связям равна 7/27.

Для текущего актуального состояния объединенного тезауруса были рассчитаны коэффициенты формулы (2): $\alpha = 0,66$; $\beta = 0,34$.

3. АЛГОРИТМ ПОСТРОЕНИЯ СЕМАНТИЧЕСКОГО ПЕРЕСЕЧЕНИЯ ПОНЯТИЙНЫХ ОБРАЗОВ

Использование мер семантической близости всегда сопровождается проблемой установления ее порогового значения. Чаще всего эта проблема решается эмпирически, причем получаемые практические результаты существенно зависят от величины порога. Кроме того, определение порогового значения связано и с поставленной задачей, в рамках которой применяется мера.

Вернемся к исходной цели – использованию понятийной системы онтологий для реализации операций на уровне функциональных систем. Ориентируясь на парадигматический характер тезаурусных отношений, меру семантической близости можно рассматривать как некоторую оценку соответствия контекстов использования понятий (т.е., если в одном документе встретилось понятие, семантически близкое понятию в другом документе, то с определенной долей уверенности можно говорить о возможном наличии общего тематического контекста). Кроме того, присутствие этих понятий на функциональном уровне свидетельствует об их участии в формировании ситуативных отношений.

Назовем понятийным онтологическим образом документа пустой подграф мультиграфа функциональной системы, состоящий из вершин мультиграфа, совпадающих с дескрипторами тезауруса, принимаемого в качестве понятийной системы. Другими словами, такой образ представляет собой множество дескрипторов:

$$D = (d_1, d_2, \dots, d_n).$$

Рассмотрим задачу построения семантического пересечения образов $D_1 = (d_1^1, d_2^1, \dots, d_n^1)$ и $D_2 = (d_1^2, d_2^2, \dots, d_m^2)$. Для всех пар дескрипторов в тезаурусе (в том числе и в объединенном) может быть рассчитана матрица семантической близости W размерности $n \times m$: $W = (w_{ij})$, где $w_{ij} = S(a_i, b_j)$, $i = \overline{1...n}$, $j = \overline{1...m}$.

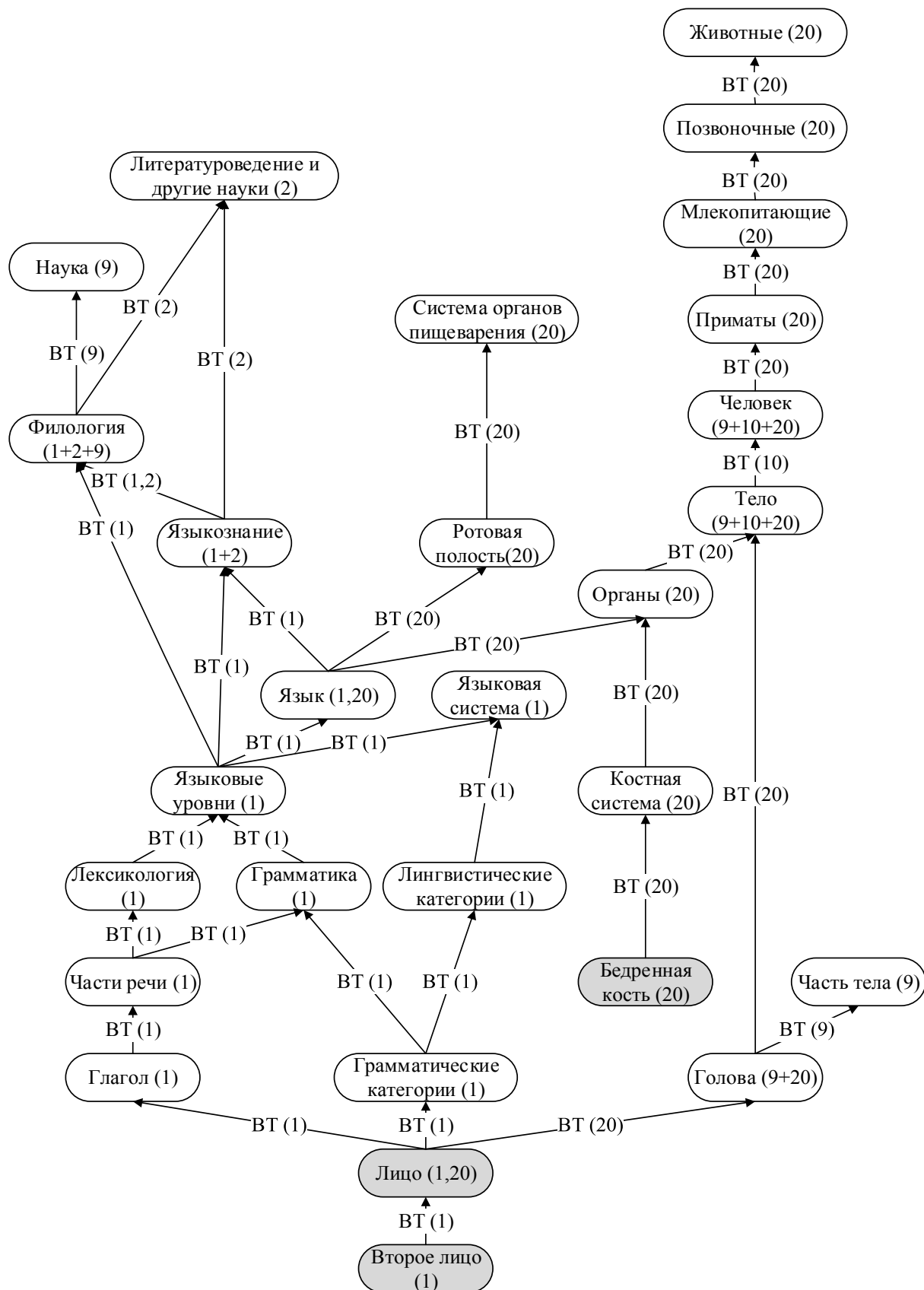


Рис. 6. Фрагмент объединенного тезауруса с «вертикальными» связями

Приняты следующие обозначения: 1 – лингвистический тезаурус, 2 – тезаурус по литературе, 9 – общелингвистический тезаурус Российского общесемантического словаря, 10 – тезаурус по религиоведению, 20 – тезаурус INIS. Знаком «+» соединены тезаурусы, в которых дескриптор имеет одинаковое значение.

Матрица семантической близости

	Активизация инвестиционного процесса	Амортизационные отчисления	Бюджетное финансирование	Бюджетные инвестиции	Внеэкономическая деятельность	Выпуск облигационного займа	Государственная политика
Анализ данных	0	0	0,37	0,25	0,25	0,37	0
Анализ деятельности	0	0	0,37	0,25	0,25	0,37	0
Бюджетная политика	0	0	0,25	0	0	0,25	0,7
Бюджетная система	0	0	0	0	0	0	0
Бюджетное финансирование	0	0	1	0,87	0,37	0,62	0
Бюджеты субъектов федерации	0	0	0	0	0	0	0
Внебюджетное финансирование	0	0	0,62	0,37	0,25	0,37	0
Государственное регулирование	0,1	0	0	0	0	0	0,8
Государственное финансирование	0	0	0,62	0,37	0,25	0,37	0

Достаточная величина предлагаемой меры семантической близости для включения пары дескрипторов в семантическое пересечение в этом случае может быть определена не через заданный фиксированный порог, а на основании *локального* контекста, задаваемого дескрипторами каждого образа, т.е. семантическое пересечение формируется из тех дескрипторов, для которых выполняется условие совпадения максимумов:

$$\max_{j=\overline{1,n}}(w_{ij}) = \max_{i=\overline{1,m}}(w_{ij}) .$$

В табл. 2 приведен пример матрицы семантической близости для двух множеств дескрипторов.

Несмотря на то, что в таблице немало довольно высоких значений меры, множество пересечения составили всего две пары: «Бюджетное финансирование» – «Бюджетное финансирование» и «Государственное регулирование» – «Государственная политика».

4. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

В целях экспериментального исследования предложенных подходов рассматривались множества понятий онтологических образов диссертаций (понятийные образы), построенные с использованием объединенного тезауруса. Были получены и исследованы попарные пересечения понятийных образов диссертаций в отрасли педагогических (10 диссертаций) и экономических (20 диссертаций) наук:

- 1) по точному совпадению дескрипторов;
- 2) с использованием матрицы семантической близости и алгоритма, описанного в п. 3.

Для оценки попарного пересечения было вычислено количество совпадающих дескрипторов и количество пар, сформированных по принципу совпаде-

ния максимумов меры семантической близости. В табл. 3 приведены усредненные результаты (в %) прироста числа пар в пересечении при использовании матрицы семантической близости.

Как следует из таблицы, число пар терминов, попавших в пересечение понятийных образов, в среднем существенно увеличивается. Сильный рост попарного пересечения понятийных образов разных предметных областей (педагогические науки – экономические науки) обусловлен политематичностью понятийной системы: незначительное пересечение по совпадению дескрипторов следует из различий понятийной лексики, а появление понятийных «мостов», количество которых превышает мощность множества совпадающих дескрипторов более чем в два раза, говорит о наличии понятий, семантически связанных в других предметных областях.

Для этих же текстов было рассчитано среднее увеличение средней степени попарного пересечения понятийных образов при использовании принципа совпадающих максимумов по формуле:

$$\Delta C = \frac{1}{n \cdot m} \cdot \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^m \left(\frac{\sum_{(d_i, d_j) \in D_{12}} S(d_i, d_j)}{|D_i \cup D_j|} - \frac{|D_i \cap D_j|}{|D_i \cup D_j|} \right),$$

где $D_i = (d_1^i, d_2^i, \dots, d_n^i)$ и $D_j = (d_1^j, d_2^j, \dots, d_m^j)$, D_{12} – множество пар дескрипторов, вошедших в семантическое пересечение понятийных образов.

Результаты расчета (табл. 3) показывают, что, несмотря на существенное (в среднем) увеличение числа семантически связанных пар в пересечении, степень попарного пересечения в среднем увеличивается менее чем на 2% (независимо от отрасли наук).

Результаты экспериментов

Среднее увеличение числа пар в пересечении, %	Среднее увеличение степени попарного пересечения, %	Педагогические	Экономические
Педагогические	60	1,6	1,2
Экономические	225	110	1,7

Таблица 4

Примеры пересечений понятийных образов

Совершенствование инструментария управления реализацией федеральных целевых программ инновационной направленности	Формирование элементарных математических представлений у детей старшего дошкольного возраста на основе этнокультурных традиций тувинского народа	
высшее образование	общее образование	0,44
высшие учебные заведения	образовательные учреждения	0,07
качество продукции	качество работ	0,33
количественная оценка	количественная оценка	1
перспективы развития	уровень развития	0,2
рабочее место	социальное развитие	0,36
учебные заведения	система образования	0,05
Разработка и реализация рекламных кампаний продовольственных товаров: механизм, технологии, модели	Формирование элементарных математических представлений у детей старшего дошкольного возраста на основе этнокультурных традиций тувинского народа	
доходы населения	социальное развитие	0,41
качество товара	качество работ	0,38
национальные особенности	национальные особенности	1
образ жизни	образ жизни	1
общественные науки	история культуры	0,16
оценка эффективности	количественная оценка	0,25
перспективы развития	уровень развития	0,2
пищевая промышленность	уборка урожая	0,03

В табл. 4 приведены примеры пересечений понятийных образов для диссертаций по педагогическим и экономическим наукам, где в первой и второй колонках – пары дескрипторов, сформированные с помощью матрицы семантической близости по принципу совпадающих максимумов. Жирным шрифтом выделены заглавия диссертаций. В третьей колонке – значение меры.

ВЫВОДЫ

Обзор способов расчета мер семантической близости показывает, что в основе аналитических выражений семантической близости, как правило, рассматриваются родо-видовые (а также таксономические) отношения,

которые образуют иерархические деревья. Это накладывает определенные структурные ограничения на формирование семантических связей предметной области.

Напротив, понятийные сети информационно-поисковых тезаурусов ориентированы на множественное наследование: отдельное понятие даже в рамках одной предметной области может иметь несколько аспектов развития, обозначаясь при этом одним и тем же термином. Существенный вклад в формирование понятийной сети вносят и ассоциативные связи. Предлагаемый подход к построению меры семантической близости позволяет учесть специфику связей информационно-поисковых тезаурусов (в со-

ответствии с [10]) и вычислить веса критериев в зависимости от особенностей конкретного тезауруса.

Использование предложенной меры в процедуре попарного пересечения понятийных образов путем построения матрицы семантической близости и формирования пар дескрипторов по принципу совпадения максимумов позволяет решить задачу определения порогового значения для вычисления семантической близости.

Эксперименты, проведенные на текстах диссертаций, показали, что с использованием политематического тезауруса удается построить для пары понятийных образов пересечение, включающее лексику различных предметных областей. Причем, несмотря на значительное количественное увеличение мощности пересечения за счет связанных мерой семантической близости пар, средняя степень попарного пересечения понятийных образов увеличивается всего на 1–2%, что говорит о несущественном «расширении» понятийного образа в случае его использования в качестве поискового при формировании выдачи в механизмах автоматизированного информационного поиска.

Формирование на основе семантической близости понятийных «мостов» между онтологиями позволяет также выделять тематические кластеры. Так, дескриптор «Математическая статистика» в диссертации по специальности 13.00.08 (Теория и методика профессионального образования) послужил «мостом» к онтологиям диссертаций по специальностям 08.00.05 (Экономика и управление народным хозяйством), 13.00.01 (Общая педагогика, история педагогики и образования), 08.00.13 (Математические и инструментальные методы экономики), соединившись с дескрипторами «Функция плотности», «Функция распределения», «Степень свободы», «Регрессионный анализ», «Статистические данные». Диссертации в результате вошли в тематический кластер «Использование методов математической статистики и теории вероятностей».

СПИСОК ЛИТЕРАТУРЫ

1. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета. – 2011.
2. Lin D. An information-theoretic definition of similarity // Proc. 15th Int. Conf. on Machine Learning. – Massachusetts: Morgan Kaufmann, 1998. – Vol. 98. – P. 296–304.
3. Budanitsky A., Hirst G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures // Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics. – Pittsburg, 2001.
4. Fiorini N. et al. USI: a fast and accurate approach for conceptual document annotation // BMC bioinformatics. – 2015. – Vol. 16, № 1, Issue 83.
5. Hsu M.H., Tsai M.F., Chen H.H. Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison // Information Retrieval Technology. – Springer Berlin Heidelberg, 2006. – P. 1–13.
6. Крюков К.В., Панкова Л.А., Пронина В. А., Суховеров В.С., Шипилина Л.Б. Меры семантической близости в онтологии // Проблемы управления. – 2010. – № 5. – С. 51–92.
7. Нгок Н.Б., Тузовский А.Ф. Модель информационного поиска на основе семантических метаописаний // Управление большими системами: сборник трудов. – 2013. – № 41. – С. 51–92.
8. СИНТОЛ // Сборник переводов по вопросам информационной теории и практики. – М.: ВИНТИ, 1968. – № 10. – 177 с.
9. Скороходько Э.Ф. Лингвистические проблемы обработки текстов в автоматизированных информационно-поисковых системах // Вопросы информационной теории и практики. – 1974. – № 25. – С. 5–120.
10. ГОСТ 7.25 – 2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. – М.: Изд-во стандартов. – 2001.
11. Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска // Научно-техническая информация. Сер. 2. – 2012. – № 5. – С. 1–9.
12. Ратцева И.И. Проблема выбора значения слова и смысловые расстояния // Научно-техническая информация. – 1966. – № 5.
13. Tversky A. Similarity features // Psychological Review. – 1977. – Vol. 84. – P. 327–352.
14. Wu Z., Palmer M. Verbs semantics and lexical selection // Proceedings of the 32nd annual meeting on Association for Computational Linguistics. – Association for Computational Linguistics, 1994. – P. 133–138.
15. Resnik P. Using information content to evaluate semantic similarity in a taxonomy // Proceedings of the 14th International Joint Conference on Artificial Intelligence. – 1994. – P. 448–453.
16. Seco N., Veale T., Hayes J. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. // 16th European Conference on Artificial Intelligence, ECAI 2004. – 2004. – P. 1089–1090.
17. Zhou Z., Wang Y., Gu J. A new model of information content for semantic similarity in WordNet // Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. – IEEE. – 2008. – Vol. 3. – P. 85–89.
18. Meng L., Gu J., Zhou Z. A new model of information content based on concept's topology for measuring semantic similarity in WordNet // International Journal of Grid and Distributed Computing. – 2012. – Vol. 5, № 3. – P. 81–94.
19. Leacock C., Chodorow M. Combining local context and WordNet similarity for word sense identification // WordNet: An electronic lexical database. – 1998. – Vol. 49, № 2. – P. 265–283.

20. Li Y., Bandar Z. A., McLean D. An approach for measuring semantic similarity between words using multiple information sources // Knowledge and Data Engineering, IEEE Transactions on. – 2003. – Vol. 15, № 4. – P. 871–882.
21. Hao D. et al. An approach for calculating semantic similarity between words using WordNet // Digital Manufacturing and Automation (ICDMA), 2011 Second International Conference on. – IEEE, 2011. – P. 177–180.
22. Bulskov H., Knappe R., Andreasen T. On measuring similarity for conceptual querying // Proc. 5th Int. FQAS Conf. LNCS. – Berlin: Springer, 2002. – Vol. 2522. – P. 100–111.
23. Maedche A., Zacharias V. Clustering ontology-based metadata in the semantic web // Principles of Data Mining and Knowledge Discovery. – Springer Berlin Heidelberg, 2002. – P. 348–360.
24. Hirst G., St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms // WordNet: An electronic lexical database. – 1998. – T. 305. – P. 305–332.
25. Knappe R. Measures of semantic similarity and relatedness for use in ontology-based information retrieval. PhD thesis. – Roskilde University, 2006.
26. Голицына О.Л., Максимов Н.В. Сравнительный структурно-статистический анализ лексик и связей информационно-поисковых тезаурусов // Научно-техническая информация. Сер. 2. – 2015. – № 6. – С. 14–28.

Материал поступил в редакцию 01.04.16.

Сведения об авторах

ГОЛИЦЫНА Ольга Леонидовна – кандидат технических наук, доцент кафедры Системного анализа Национального исследовательского ядерного университета «МИФИ»
e-mail: OLGolitsina@YANDEX.RU

МАКСИМОВ Николай Вениаминович – доктор технических наук, профессор кафедры Системного анализа Национального исследовательского ядерного университета «МИФИ»
e-mail: NV-MAKS@YANDEX.RU

ФЕДОРОВА Вера Александровна – студент кафедры Системного анализа Национального исследовательского ядерного университета «МИФИ»
e-mail: Bamsikasper@MAIL.RU