Знания-Онтологии-Теории (ЗОНТ-15)

Информационная система «ИСТИНА» как Big Data – инструментарий в области управления на основе анализа наукометрических данных

В.А. Садовничий, В.А. Васенин, С.А. Афонин, А.С. Козицын, Д.Д. Голомазов

МГУ имени М.В. Ломоносова, Ленинские горы, д. 1, г. Москва, 119991, Россия.

info@rector.msu.ru, vasenin@msu.ru, serg@msu.ru, alexanderkz@mail.ru, denis.golomazov@gmail.com

Аннотация. *Доклад посвящен анализу функциональных возможностей и перспектив развития информационно-аналитической системы «ИСТИНА», которая активно используется в МГУ имени М.В. Ломоносова для управления научно-образовательным процессом на основе анализа наукометрических данных. Представлены требования к Системе, ее цели, задачи и свойства, позволяющие отнести ее к классу систем с «большими данными». Описаны принципы, положенные в основу ее создания, основные архитектурно-технологические особенности Системы, предоставляемые сервисы и перспективы развития. Особое внимание уделено вопросам применения в Системе семантических методов, в частности - онтологий для тематической классификации статей. Отдельно рассмотрены задачи построения и классификации портретов подразделений и научных коллективов, кластеризации графов соавторства и выделения терминов из публикаций. Система служит полигоном для тестирования теоретически значимых моделей, методов и алгоритмов обработки текстов на естественном языке, в том числе – использующих семантику, а также - анализа графов. Практическую значимость и эффективность Системы подтверждает расширяющийся перечень функций, которые возлагаются на нее в сфере подготовки принятия управленческих решений в Московском университете, а также активный интерес к ней со стороны других научно-образовательных учреждений России.*

Ключевые слова: наукометрические данные, информационно-аналитические системы, системы подготовки принятия решений, оценка эффективности научной деятельности, семантические модели, онтологии, тематическая классификация

1 Введение

В последние 2-3 года потребность в создании информационно-аналитической системы подготовки принятия решений на основе анализа данных научно-инновационного и образовательного содержания настоятельно обозначается на всех уровнях национальной инфраструктуры управления образованием и наукой. Систему подобного назначения (далее для краткости изложения – Систему) хотели бы иметь как отдельные организации (вузы, НИИ, научно-производственные центры и т.п.), особенно крупные, претендующие на высокий национальный и международный рейтинги, субъекты управления в регионах, так и Министерство образования и науки России. Основной мотив такой потребности - создать Систему, инструментальные средства которой позволяли бы **оперативно** анализировать (по различным срезам и, соответственно, запросам) **надежно верифицируемые** данные как **персональные** и **обобщенные по отдельным научным коллективам и организациям** (вузам, НИИ и т.п.), так и данные, обобщенные **по регионам и по России в целом. Целью такого анализа является получения с высокой степенью объективности оценки показателей результативности**

и/или тенденции (тренды) научно-педагогической и инновационно-внедренческой деятельности. Полученные с помощью Системы **оценки и результаты анализа трендов** научно-технической, образовательной и инновационной деятельности могут быть положены **в основу**: различного рода **отчетной документации**; **принятия управленческих** (финансово-экономических, кадровых, структурных и т.п.) **решений**; **определения рейтинговых показателей** субъектов подобной деятельности на всех уровнях. С учетом перечисленных требований критериями готовности целевой Системы к решению поставленных перед ней задач должны быть следующие:

1. **в Системе должны присутствовать максимально точные** («очищенные» и верифицируемые) **в должном объеме** (по отношению к численности подлежащей анализу группы - организация, регион, страна в целом) данные в Системе о научно-педагогической и инновационно-внедренческой деятельности отдельных, участвующих в ней персон;
2. **данные**, которые аккумулируются в Системе, в первую очередь, **должны включать,** как **сведения библиографического характера**, так и **сведения** (лучше на русском и английском языках) **аннотационного характера, включая ссылки** на другие источники, которые связаны с источниками первичными;
3. в Системе должны присутствовать программные механизмы, которые реализуют современные модели и алгоритмы анализа данных, учитывающие аналитику предметных областей, к которым относятся те или иные запросы пользователей;
4. для реализации требований к системе в полном объеме **должны быть по возможности полно (в рамках авторских прав) представлены результаты деятельности персоналиев в полнотекстовом виде.**

С учетом изложенного, подлежащий созданию продукт, обладает всеми признаками нового класса систем, именуемых Big Data –системами с «большими данными». Термин «большие данные» в рассматриваемом случае характеризует Систему не только по объему аккумулируемых в ней данных, но и их гетерогенный характер и, как следствие – сложную структуру сбора, хранения и обработки таких данных. Такие системы характеризуются целым рядом свойств, в том числе - следующими, которые требуют их исследования и определяют дополнительные требования, которые должны приниматься во внимание разработчиками новых информационных систем.

- Данные хранятся и могут поступать из разных источников (от персоналиев, источников в Интернет, из различных, уже сложившихся реляционных баз и т.д.). Нужно иметь возможность (механизмы) их эффективно собирать и систематизировать для последующей первичной обработки и хранения.
- Данные, как правило, имеют разную структуру. В этой связи нужно иметь механизмы их конвертации в форму, удобную для первичной (предварительной) обработки с целью последующей верификации и организации хранения.
- Кроме того, что данные имеют разную структуру, их очень много, что приводит к сложности поиска тех из них, которые необходимы для анализа, а также к трудностям обработки поисковых запросов. Необходимо иметь механизмы эффективного поиска и анализа данных, а значит – предварительной систематизации (кластеризации) и поиска эффективных механизмов обработки запросов.

Следует заметить, что отмеченные выше механизмы имеют разную природу. Это и новые модели, и реализующие их средства автоматизации процессов сопровождения данных, и некоторые меры, механизмы административного характера. Последние, как правило, очень важны для эффективного сопровождения подобного класса систем, что будет отмечено далее.

Одно из направлений научно-практической деятельности в Московском университете, которое тесно связано с проблемой больших наукометрических данных и с вопросами эффективного управления исследованиями на их основе, является создание и развитие Информационно-аналитической системы «Наука МГУ», которая построена на основе моделей, механизмов и инструментальных средств Интеллектуальной Системы Тематического Исследования Научно-технической информации – «ИСТИНА» (далее – Система). Не вдаваясь в детали ее реализации, с которыми можно ознакомиться по монографии [1], отметим лишь общие архитектурные особенности, принципы, модели и алгоритмы, положенные в основу разработки Системы и программно реализованные на настоящее время, а также перспективы её развития на основе моделей семантического анализа.

2 Архитектура

Взаимодействие пользователей с Системой осуществляется через веб-интерфейс. Архитектура Системы является типовой для современных веб-приложений. В основе Системы лежит сервер приложений. Главным языковым средством для разработки приложений в Системе на настоящее время является программное средство Django, которое написано на языке Python. Архитектура сервера приложений Системы (рис. 1) отражает требования к ее функциональным возможностям.

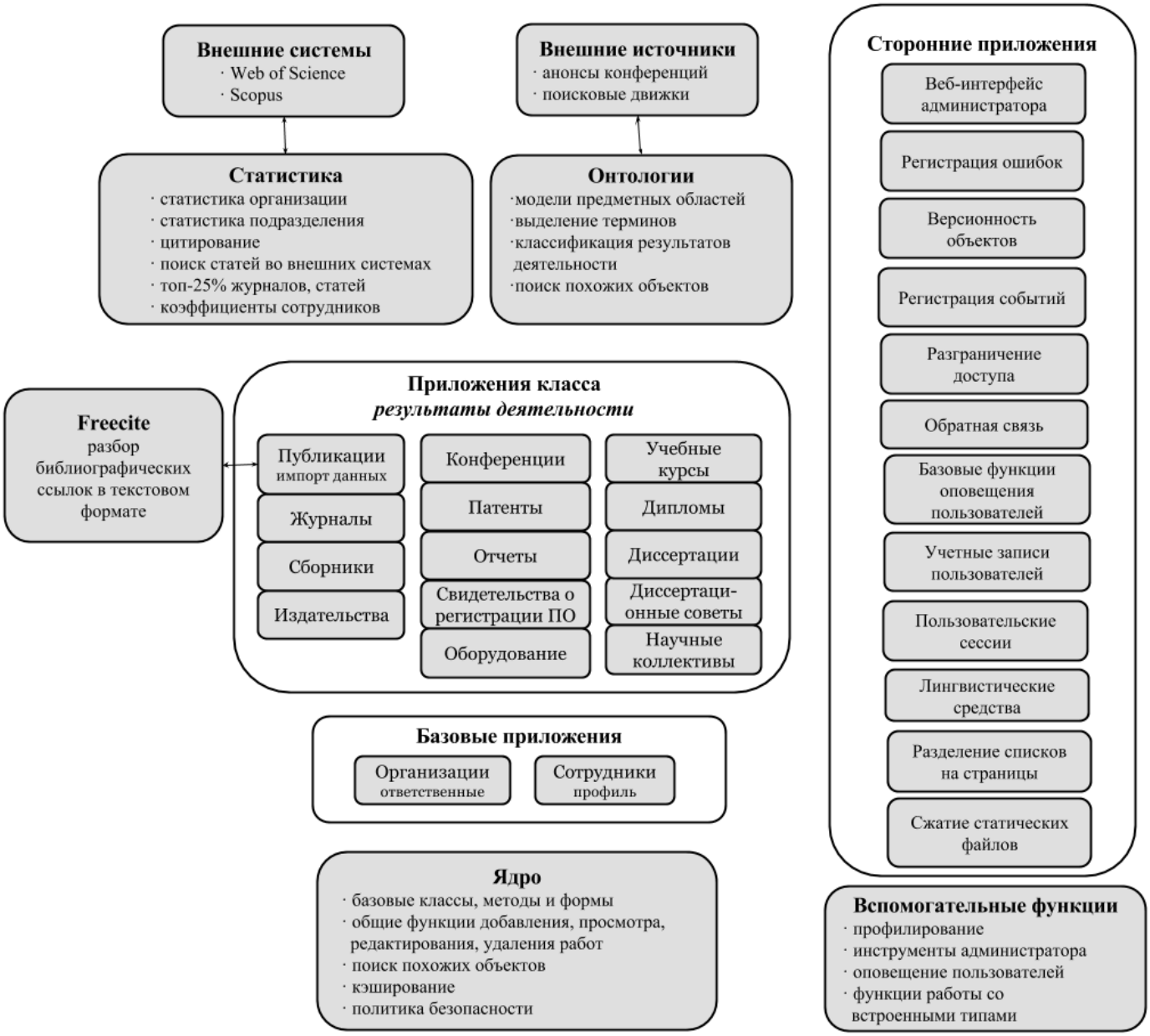

Рис. 1. Архитектура сервера приложений системы ИСТИНА.

Основу Системы составляет ядро, реализующее базовые (системообразующие) функции, которые напрямую определяют и показатели качества Системы в целом. Ядро обеспечивает единый интерфейс и базовый каркас сбора, просмотра, редактирования и удаления результатов научно-педагогической

деятельности сотрудников. В ядре реализуются функции поиска «похожих»[1] объектов, которые используются для подбора, в частности, «похожих» сотрудников, журналов и статей. Важной частью ядра является механизм кэширования. Он используется для ускорения доступа к различным данным Системы. В ядре реализуются также и другие общие функции, например, поддерживающие механизмы разграничения доступа к различным категориям данных Системы.

На следующем уровне архитектуры расположены два базовых приложения, модуль организаций и модуль сотрудников. Модуль организаций включает классы и логику действий, связанную с данными на уровне организаций, в частности, механизм управления информацией о сотрудниках, которым делегирована роль ответственных от организации и отдельных ее подразделений за сопровождение информации в Системе. Приложение «сотрудники» содержит базовые классы, связанные с сотрудниками, в частности, их профиль в Системе и список работ.

Основной архитектурный уровень составляют приложения класса «результаты деятельности». Каждое приложение отвечает за отдельный тип результатов научно-инновационной и преподавательской деятельности сотрудников (например, публикации) или за общую сущность, связанную с таким типом (например, журналы). Программисту необходимо лишь настроить Систему на конкретный тип результатов деятельности, указав специфицирующие его свойства. Отметим, что наиболее сложным приложением из рассматриваемого класса являются публикации. Оно использует модуль разбора библиографических ссылок в текстовом формате Freecite, который был усовершенствован с целью предоставления пользователю возможности обработки текстов на русском языке.

На следующем уровне иерархии архитектуры Системы расположены два модуля, отвечающие за анализ данных, накопленных на предыдущих уровнях. Модуль статистики содержит функции преимущественно количественного анализа данных Системы на уровне организации, подразделения и отдельного сотрудника. В этом модуле также реализован простой тематический анализ показателей результативности сотрудников. Приложение «статистика» включает функции получения и обработки показателей цитирования отдельных статей из Web of Science, из Scopus, а также поиск статей в этих системах. Кроме этого, в модуль статистики входят механизмы расчета принятых в подразделениях организации коэффициентов эффективности деятельности сотрудников на основе результатов их деятельности.

Приложение «онтологии» имеет целью выполнение более глубокого тематического анализа данных в Системе. В частности, программные механизмы этого модуля основаны на моделях предметных областей, которые строятся в автоматизированном режиме. Для их построения используются анонсы научных конференций, а также результаты запросов к поисковым системам, таким как Bing. Эти модели содержат термины, характерные для соответствующей предметной области. По этим терминам затем выполняется классификация результатов деятельности сотрудников (например, статей), а также подбор похожих объектов для удобной навигации по Системе. Отметим, что большинство функций этого модуля пока находятся в стадии исследовательской разработки. Однако эти исследования рассматриваются разработчиками Системы, как одно из самых перспективных направлений ее развития.

3 Принципы и основные направления развития Системы

Система ориентирована на эффективное сочетание с одной стороны - интересов организации, в которой она эксплуатируется, и сотрудников её административно-управленческого аппарата, с другой - с интересов отдельных ученых и педагогов. Приоритет в процессах сбора и верификации данных отдается потокам, которые именуются потокам «снизу-вверх», а именно - от отдельных персоналиев в базу данных системы. В этой связи – одно из главных требований к Системе - предельно удобный, дружественный для взаимодействия с конечным пользователем интерфейс и эффективные разноплановые механизмы верификации (очистки) данных. В первую очередь, при этом принимается во внимание тот факт, что только сами ученые могут точно описать и сверять с оригиналом, если это необходимо, свои научные результаты. Здесь следует отметить, что этот принцип не исключает возможности использования данных, которые могут быть экспортированы, в том числе – в массовом порядке, из других источников с последующей их верификацией. Однако такой способ пополнения

1 Под термином «похожий» объект в Системе имеются в виду реквизиты (сведения) о результатах научной деятельности и авторского коллектива его исполнителей, которые могут быть по ошибке (в т.ч. и с позиции принятого стандарта) внесены в Систему.

Системы данными является вспомогательным, а их верификация требует дополнительных моделей и программных механизмов верификации в автоматическом и/или автоматизированном с участием пользователей режиме. К их числу относятся: «поиск двойников», сверка названий результатов деятельности и т.п.

Наличие большого массива данных в Системе и возможности их разноплановой обработки заставляют разработчиков активно создавать эффективные процедуры оценки достоверности представленных в ней данных. Кроме перечисленных выше автоматизированных программных механизмов, такая процедура включает иерархически организованный институт ответственных от подразделений МГУ за сопровождение информации в Системе, а также механизмы разграничения доступа к данным разного уровня конфиденциальности. Институт ответственных с четким обозначением сферы их компетенции в работе с данными является важным звеном, реализующим, наряду с программными механизмами, меры административного характера, способствующие подержанию высокого уровня достоверности данных в Системе. Здесь сразу следует отметить, что в настоящее время возможности внедрения (использования) таких механизмов в процессе взаимодействия с базами наукометрических данных (и не только зарубежных) ограничены. Однако, это обстоятельство, по мнению авторов, не должно являться препятствием на пути создания отечественных, отвечающих российским реалиям подходов к адекватной оценке научно-инновационной и педагогической деятельности. На направлении разграничения доступа к данным, в том числе – агрегированным, с разным уровнем конфиденциальности, в Системе задействованы современные модели и программные механизмы, в том числе - новые, авторские, разработанные в МГУ [1].

В настоящее время научно-инновационная педагогическая деятельность ученых характеризуется более чем 20 параметрами, в числе которых: публикации в научных журналах, сборниках статей и тезисы докладов; патенты и свидетельства на интеллектуальный продукт, курсы лекций и семинары; руководство дипломными работами и диссертациями.

Другой важный принцип, которым разработчики руководствуются в процессе развития системы, это высокий уровень открытости информации, доступ к которой либо предоставляется самими пользователями, либо разрешается соответствующими нормативно законодательными документами РФ. Все сведения о результатах научной работы сотрудников МГУ, например, доступны в сети Интернет. Для каждого сотрудника университета автоматически формируется его открытый научный профиль, который отражает всю научно-педагогическую деятельность. Такой уровень открытости информации не только повышает статус ученого в научном мире, но и способствует повышению ее достоверности, что очень важно в современных условиях.

Система активно используется для оценки степени востребованности результатов ученых МГУ имени М.В.Ломоносова в мире. Главным критерием оценки в данном случае являются индексы цитирования Web of Science и Scopus. Безусловно, они не отражают все аспекты научной деятельности ученого, особенно в гуманитарных областях. Однако индексы цитирования являются основой для многих международных рейтингов учебных заведений и нет оснований не ориентироваться на эти показатели, в числе других.

Данные о более чем 12,5 тысячах ученых и преподавателей университета об их научно-инновационной и педагогической деятельности, 353283 статей, 80472 докладов на конференциях, 34668 книг, сведения о 5745 научно-исследовательских работ - далеко не полный перечень, характеризующий объем данных в Системе. Это массив, результаты анализа которого и являются в настоящее время исходными для принятия административно-управляющих решений в МГУ.

В конце прошлого года в МГУ оперативно (за 3-4 недели) были проведены 3 конкурса, которые были призваны по итогам года определить ученых и преподавателей, внесших наиболее весомый вклад в достижения МГУ. Исходными для Конкурсной комиссии были результаты сравнительного анализа данных о научно-инновационной и педагогической деятельности конкурсантов. Анализ более чем 3-х тысяч заявок на конкурс был оперативно проведен по данным, представленным в Системе.

В университете разработан и запущен в эксплуатацию на основе системы «ИСТИНА» Официальный сайт 104 диссертационных советов МГУ. В нем аккумулированы и эффективно обрабатываются все соответствующие Положениям Минобрнауки и ВАК РФ данные по прохождению дел в диссертационных советах МГУ.

Годовая отчетная компания по научно исследовательской работе уже второй год проведена на основе анализа данных в системе «ИСТИНА».

В сфере кадровой политики – это анкеты из системы «ИСТИНА» на конкурсный отбор, в свете поощрения за результаты научных исследований – данные о публикациях (их рейтинговая ценность),

описание наиболее ярких достижений и другие. В настоящее время завершается работа по переходу на новые принципы конкурсного избрания на научные должности в МГУ. В их основе механизмы Системы «ИСТИНА», которые позволяют каждому ученому и руководителю подразделения оперативно посчитать:

- персональный рейтинг любого научного сотрудника подразделения;
- на основе агрегирования и определения (по медиане) этих рейтингов получить средние по подразделению рейтинги для отдельных категорий сотрудников (м.н.с., н.с., в.н.с., и гл.н.с.);
- гистограмму – зависимость числа сотрудников, имеющих тот или иной рейтинговый балл, от величины этого балла.

С этой целью в каждом из структурных подразделений МГУ (факультеты, институты, центры) разработана формула расчета показателя эффективности работы отдельного научного работника.

Для оценки эффективности работы научных сотрудников в рамках Системы построен "Конструктор формул расчета эффективности работы сотрудника". Этот конструктор позволяет ответственным пользователям Системы сформировать необходимую формулу расчета, которая может учитывать практически все виды работ, внесенных в Систему. Конструктор обладает достаточной гибкостью для адекватного учета специфики того или иного подразделения (факультета, института, центра). Подготовленная с помощью такого конструктора формула расчета эффективности может использоваться как для построения таблицы с указанием полученных баллов для всех сотрудников подразделения, так и для составления персонального отчета, который содержит перечень всех работ конкретного сотрудника, с указанием полученного числа баллов за каждую работу. Персональные отчеты позволяют проверить достоверность интегральных показателей эффективности.

Как следствие, у каждого сотрудника появляется возможность увидеть на своей, закрытой от других пользователей странице в ИСТИНЕ индивидуальный рейтинг, коррелирующий с данными о его научно-педагогической деятельности персональной страницы. Как следствие, каждый ученый может определиться с тем, как его повысить, если он ниже пороговых значений, которые он тоже может увидеть в ИАС «ИСТИНА». Здесь следует еще раз отметить, что с помощью программных механизмов логического разграничения доступа данные по персональному рейтингу уже являются конфиденциальными. Они доступны только пользователю – персоналию и научно-административному персоналу, стоящему выше его по управленческой иерархии университета.

4 Перспективы развития

Все представленные выше результаты создают достаточно прочную основу (базу) для решения большинства текущих задач, которые поставлены перед целевой Системой, с помощью предложенных моделей и программных механизмов. В достаточно полном объеме предъявляемых к ней требованиям решена только первая из перечисленных выше задач. Об этом свидетельствует полнота и уровень адекватности реалиям данных по МГУ, которые размещены в Системе. Задачи 2 и 4 пока не решены в данном объеме, но не по причинам технического характера, а в связи объективными сложностями их реализации согласно положениям нормативных документов РФ (соглашения с издательствами р возможности публикации и т.п.).

Однако, необходимо отметить, что методология, которая в настоящее время используется для анализа представленных в Системе данных (задача 3) не может рассматриваться, как основа для создания перспективных на будущее и эффективных в вычислительном плане программных средств такого анализа. В этой связи рассмотрим направления теоретических исследований и практических работ, которые направлены на устранение отмеченных «узких мест» (недостатков). Здесь следует заметить, что в настоящее время исследования по этим направлениям ведутся. Однако они находятся на разных этапах жизненного цикла создаваемых продуктов – от предпроектных исследований (создание системного проекта) для одних до стадии тестовых испытаний для других. С этих позиций следующие представление подобных задач можно расценивать, как призыв к активным действиям, направленным на их решение.

4.1 Онтологии в системе ИСТИНА

Для хранения и наглядного представления всех сущностей и связей, характеризующих те или иные сферы научной деятельности, в Системе (пока в тестовом режиме) используются на ограниченном наборе сфер деятельности решения на основе онтологий - формальной модели представления знаний, опирающейся на дескриптивную логику. Общая схема такого решения представлена на рис. 2.

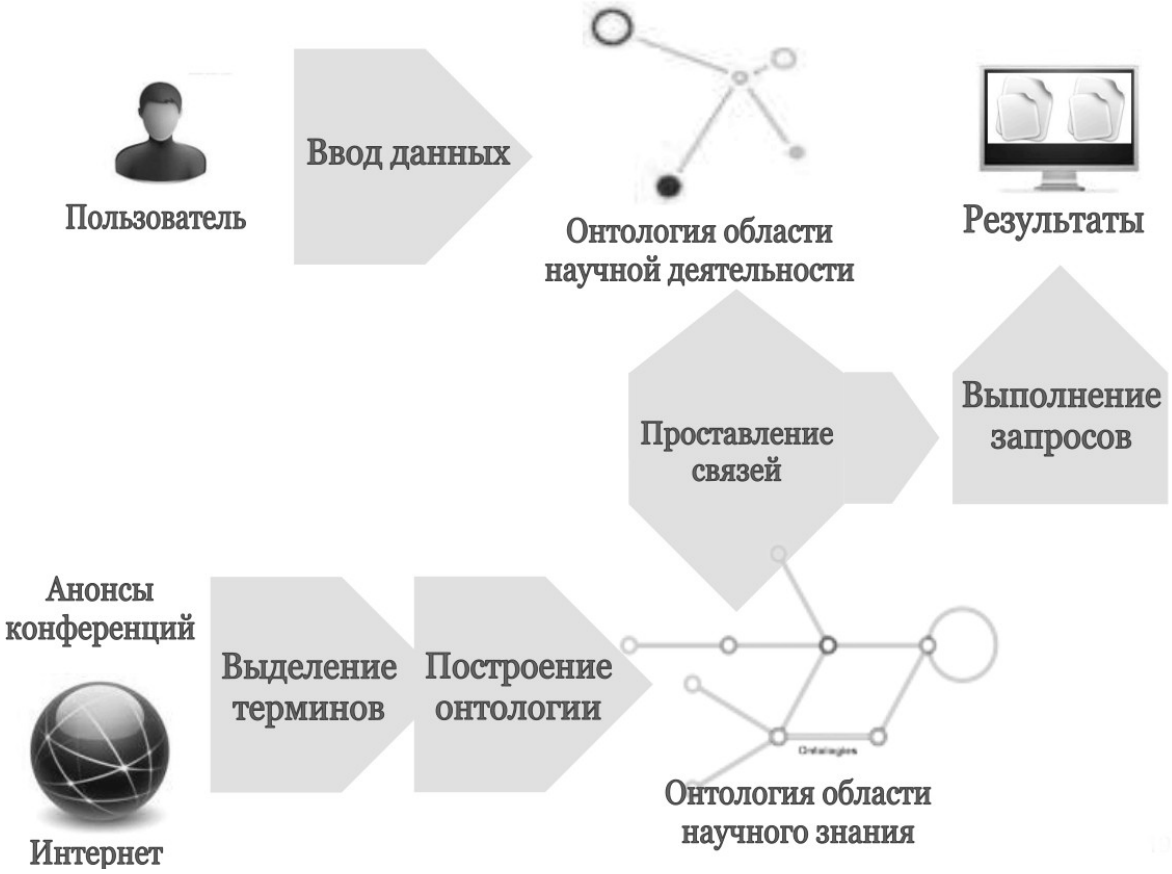

Рис.2. Общая схема использования онтологий в системе ИСТИНА.

Она включает онтологию научно-организационной составляющей деятельности (такие сущности, как «человек», «организация», «публикация», «конференция», «проект» и связи между ними), а также онтологии отдельных областей научного знания - моделей конкретных отраслей науки, описанных на языке OWL, например, математики, информатики или физики, которые содержат основные направления исследований в рамках этой области, используемые понятия и связи между ними. Для эффективного формирования таких онтологий разработан алгоритм Sonmake [2] автоматизированного построения онтологий областей научного знания на основе текстов анонсов научных конференций и информации из поисковых систем в Интернет. Заметим, что в нем используется созданный авторами алгоритм выделения терминов из полуструктурированных текстов Brainsterm [3].

Установление связей между перечисленными двумя типами онтологий позволит проводить тематическую классификацию сущностей Системы по терминам онтологий областей научного знания. Такая классификация может служить основой для решения целого ряда наукоемких и востребованных на практике задач. Например, с ее помощью можно осуществлять поиск экспертов по произвольной тематике, заданной набором ключевых слов. Система ищет пользователей, работы которых отмечены этими ключевыми словами, и ранжирует их по определенному алгоритму, учитывая при этом количество найденных терминов, и общую результативность работы ученого.

Онтологии области знания также помогут расширить возможности анализа деятельности подразделений и организаций по отдельным темам. В настоящее время в качестве источника информации о темах исследований в Системе используются рубрикаторы, причем тема публикации определяется по фиксированной тематике журнала, в котором опубликована статья. Использование онтологий вместо рубрикаторов позволит более точно определить тематическую направленность отдельной работы, так как предполагается, что онтология содержит большее число понятий и

отношений, чем рубрикатор, и к понятию из онтологии привязана каждая отдельная публикация, а не журнал в целом. Наконец, использование онтологий позволит добавить в Систему такие функции, как отображение списка работ, похожих на заданную, списка ученых, занимающихся похожей тематикой, списка похожих журналов и другие тематические запросы.

Построенные онтологии могут использоваться, в том числе, для уточнения и расширения запросов при осуществлении пользователем полнотекстового поиска в Интернет. Суть подхода в том, что пользователь задает набор ключевых слов для поиска, а затем этот набор автоматически дополняется терминами из онтологии, соответствующими его интересам. Например, запрос «определение группы» будет преобразован в «определение группы AND (математика OR физика OR вычислительные методы)». Такое уточнение позволит Системе не показывать пользователю тексты, которые содержат определения групп, но относятся к социологии или биологии. Указанное в запросе понятие дополняется списком связанных с ним дочерних понятий с учетом типа, степени и направления связи. Например, запрос «алгоритм RSA» будет дополнен понятиями «теорема Эйлера», «открытый ключ» и другими. Такое дополнение позволит найти документы, которые посвящены обсуждению связанных с алгоритмом RSA вопросов, однако не содержат его явного упоминания.

В настоящее время целями дальнейшего развития Системы на направлении использования онтологий являются: разработка и применение алгоритмов тематической классификации сущностей Системы по терминам онтологий областей научного знания; создание онтологий областей научного знания полуавтоматическим способом с привлечением экспертов в соответствующих предметных областях; разработка сервисов (например, поиска экспертов) на основе полученной тематически размеченной информации.

4.2 Тематическая классификация статей

Классификация (кластеризация) по отдельным тематическим группам является важным направлением исследований. Решение этой задачи позволит более точно определять распределение по этим темам результатов научных исследований. В настоящее время, как уже отмечалось ранее, тематические портреты подразделений строятся на основе принадлежности публикаций их участников к журналам, тематика которых определена, но достаточно широко. Такой подход не может быть достаточно точным и уточнение их тематического портрета может быть проведено с использованием методов тематического анализа текстов, например, ранее отмеченный онтологии.

Отметим, что кластеризацию тематических направлений на основе анализа публикаций, можно использовать не только для построения портрета подразделения, но и для построения тематических портретов журналов и конференций. Последнее может оказать существенную помощь авторам при подготовке публикаций и докладов, поскольку тематики журналов и конференций на сайтах сформулированы достаточно широко и не всегда можно однозначно оценить, насколько тематика предполагаемой к публикации работы или доклада соответствует кругу читателей.

4.3 Тематическая классификация портретов подразделений и научных коллективов

Построение более точных тематических портретов структурных подразделений МГУ, научных коллективов по результатам их исследований позволит определить наличие потенциальных связей между группами, работающими со схожей тематикой. Результаты такой классификации предоставят пользователям возможность выполнять более точный поиск научных коллективов и отдельных ученых, работающих в тех или иных направлениях исследований, уточняя, в дополнение к ключевым словам запроса, интересующую их область науки.

4.4 Кластеризация графов соавторства на предмет выделения устойчивых подграфов авторов – научных сообществ

Применение методов кластеризации к графам, описывающим авторство научных публикаций позволит определить устойчивые научные сообщества - соавторские группы (подграфы) и определять роли (вклад) отдельных авторов в этих сообществах. Для этого должны выделяются вершины (авторы), которые для устойчивых подгрупп являются «ключевыми». Такими, как правило, являются научные руководители (лидеры) этих научных сообществ. Может производиться анализ динамики развития (эволюции) таких групп, результаты которого совместно с тематическим анализом результатов их исследований, позволят получить представление об основных трендах в той или иной научной области, применительно к отдельной организации или к более широкому сообществу.

4.5 Автоматизированное расширение наборов ключевых слов для тематически связанных публикаций с целью повышения эффективности поиска.

Использование методов кластеризации к наборам ключевых слов для тематически связанных публикаций позволит автоматически расширять такие наборы, рекомендуя их авторам. Необходимость создания такого сервиса объясняется тем обстоятельством, что авторы, не задумываясь о потенциально возможном последующем анализе текста, не всегда указывают для своих публикаций достаточное качество ключевых слов. Заметим, что аналогичные методы используются в поисковых системах при расширении поисковых запросов. Основным отличием в рассматриваемом случае является отсутствие необходимости выполнять вычисления в реальном времени. Это дает возможность использовать более сложные алгоритмы анализа, а также возможность корректировать полученные гипотезы с использование текстов статей. Расширение таких списков позволит существенно улучшить полноту поиска с сохранением достаточной точности.

Литература

[1] Интеллектуальная система теоретического исследования научно-технической информации (ИСТИНА) / С.А. Афонин и др. Под ред. академика В.А. Садовничего. – М.: Издательство Московского университета, 2014. - 262 с.

[2] Васенин, В.А. К созданию системы управления научной ин-формацией на основе семантических технологий / В.А. Васенин, С.А. Афонин, Д.Д. Голомазов // Материалы Всероссийской конференции с международным участием "Знания - Онтологии – Теории" (ЗОНТ-2011), 3-5 октября 2011 г., том 1. — Институт математики им. С.Л. Соболева СО РАН, Новосибирск, 2011. — С. 78–87.

[3] Голомазов, Д.Д. Выделение терминов из коллекции текстов с заданным тематическим делением / Д.Д. Голомазов // Информационные технологии. — 2010. — № 2. — С. 8–13.