Светлов Н.М., Зуев А.А. Перспективы автоматизации семантического анализа документов по аграрно-экономической тематике в системе управления АПК // Материалы международной научной конференции (декабрь 2001 г.): Сб.науч.тр.: Вып.8. М.: Изд-во МСХА, 2002. – С.231-236.

Перспективы автоматизации семантического анализа документов по аграрно-экономической тематике в системе управления АПК

***Н.М. Светлов**, доцент кафедры экономической кибернетики, к.э.н.;*

***А.А. Зуев**, старший лаборант научно-методического кабинета экономического факультета*

Современный этап научно-технического прогресса характеризуется взрывным ростом объёма научных информационных ресурсов, в том числе и ресурсов аграрно-экономической направленности. Рост сопровождается расширением тематического разнообразия документов и снижением их качества. Отслеживать документы, содержащие существенные элементы научной новизны, интересующие конкретного получателя информации (ПИ), становится всё труднее.

Существует два основных подхода к отбору документов, тематика которых интересует данного ПИ, — по идентификационным кодам, назначаемым экспертами, либо по содержащимся в них ключевых словам.

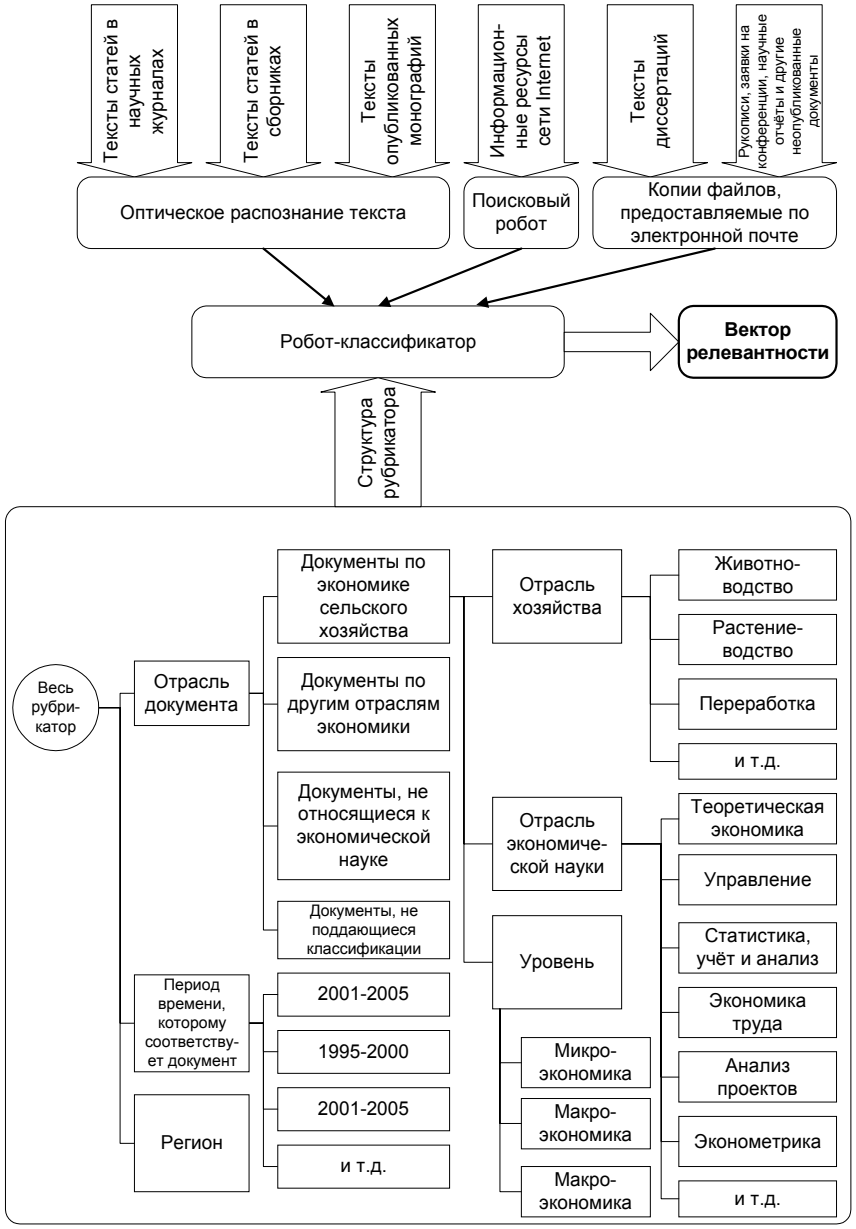
Первый подход предполагает рост численности квалифицированных экспертов-библиографов, пропорциональный росту информационного потока, а это практически невозможно. Второй подход перспективнее. Но и у него есть два недостатка.

Во-первых, точность отбора научных документов по ключевым словам оставляет желать лучшего: по опыту авторов, при поиске в сети Интернет документов, признаки которых не известны достоверно, лишь один документ из 15…50 найденных соответствует цели поиска. Вследствие диверсификации информационных ресурсов этот показатель с течением времени ухудшается, а производительность труда научных работников и преподавателей при поиске требуемой информации падает. Ситуация усугубляется тем, что эту работу, как правило, невозможно перепоручить техническому персоналу.

Во-вторых, отбор по ключевым словам неэффективен для обнаружения результатов новых научных исследований, в которых вводятся новые термины, заранее не известные ПИ.

Процесс управления АПК нуждается в своевременном обеспечении оперативной информацией о методиках и результатах научных исследований. Эта потребность особенно остра в период освоения новой системы хозяйствования. Вышеназванные проблемы приводят одновременно к повышению затрат на управление АПК и снижению его эффективности вследствие несвоевременного либо неполного информационного обеспечения. В условиях ограниченности финансовых ресурсов на расширение штата управленческого персонала невозможность обработать поток экономической информации приводит к дальнейшему отступлению отечественного АПК с занимаемых им конкурентных позиций как на мировом, так и на внутреннем рынках.

Могут ли быть решены или, по крайней мере, смягчены перечисленные проблемы? По нашему мнению, да. Современная информатика располагает достаточно мощным научным и алгоритмическим обеспечением, которое может воплотиться в инструментальные средства, радикально повышающие степень автоматизации семантического анализа, классификации и отбора научных публикаций по проблемам экономики АПК.



Светлов Н.М., Зуев А.А. Перспективы автоматизации семантического анализа документов по аграрно-экономической тематике в системе управления АПК //
Материалы международной научной конференции (декабрь 2001 г.): Сб.науч.тр.: Вып.8. М.: Изд-во МСХА, 2002. – С.231-236.

В частности, для этой цели могут использоваться методы формализации семантики, основанные на теории графов; представление знаний о признаках релевантности документа конкретной тематике в форме условных вероятностей; использование методов автоматического обучения[1].

На рисунке приведена принципиальная схема предлагаемой автоматизированной системы семантического анализа документов. Анализ документов, поступающих из самых разных источников — издательств, библиотек, научных учреждений, Интернета — осуществляется специализированным инструментальным средством — роботом-классификатором. Робот, анализируя формализуемые атрибуты документа, вычисляет показатели его соответствия каждой рубрике тематического рубрикатора, организованного специальным образом. Эти показатели приписываются документу и определяют его ранг в каждой рубрике до тех пор, пока не изменится либо структура рубрикатора, либо правила расчёта показателей соответствия. Идентификаторы (библиографические описания или сетевые пути) документов вкупе с показателями соответствия каждой рубрике образуют библиографическую базу данных, предоставляемую конечным пользователям.

Рассмотрим алгоритм работы робота.

Иерархический рубрикатор можно формально представить графом Q вида «дерево», который задаёт множество term(Q)[2] компонентов семантики анализируемых документов. Множество term(Q) определяет стандартный симплекс T в ортанте $\mathbb{R}^{+}_{\#(\mathrm{term}(Q))}$, представляющий собой множество значений семантики документа над графом Q. Семантика конкретного документа d над графом Q описывается #(term(Q))-компонентной вектор-функцией $\mathbf{r}(d,Q)$, обладающей свойством $\mathbf{r}(d,Q)\mathbf{i} = 1$. Значение этой функции называется *вектором релевантности*. Задача семантического анализа документа сводится к отысканию статистической оценки вектора $\mathbf{r}(d,Q)$ на основе формализуемых атрибутов документа d.

Для решения этой задачи можно использовать представление знаний о семантике документов в форме набора продукционных правил, отображающих атрибуты документов на значения вероятности соответствия документа данному правилу при условии его принадлежности к данной рубрике. Если задана априорная вероятность принадлежности документа к той или иной рубрике (априорный вектор релевантности), то вектор вероятности принадлежности документа каждой рубрике, определённый применением формулы Байеса к заданной априорной вероятности и условным вероятностям, соответствующим множеству продукционных правил, которым соответствует документ, обладает всеми свойствами вектора релевантности, указанными выше. Разность априорного и апостериорного, т.е. вычисленного с использованием правил, вектора релевантности характеризует информативность использованных правил и, как следствие, качество классификации. Как вектор априорных вероятностей, так и вероятности событий «рубрика+правило» можно определить статистически на основе представительной выборки из анализируемого потока документов.

[1] Землянский А.А., Светлов Н.М. Теоретические основы формализации линейного экономико-математического моделирования // Современные информационные технологии в экономике: Сборник научных трудов / Моск. эконом.-стат. ин-т. — М., 1992. — с.85-100; Berger J. Bayesian analysis. Amsterdam: North-Holland, 1994; Нейлор К. Экспертные системы: принципы работы и примеры. — М., 1987; Лорьер Ж.-Л. Системы искусственного интеллекта. М., 1991.

[2] Поясним некоторые математические обозначения, использованные в статье: term(Q) — отображение ориентированного графа Q на множество его терминальных вершин; #(Q) — оператор числа элементов множества; $\mathbb{R}^{+}_{n}$ – неотрицательный ортант n-мерного евклидова пространства; $\mathbf{i}$ — единичный вектор.

Качество вектора релевантности и полезность всей системы зависят от того, насколько удачен используемый набор правил. Можно рекомендовать следующие группы правил.

♦ Оценивающие принадлежность документа к классу научных текстов: источник получения (в порядке снижения релевантности — рецензируемый научный журнал, научное издательство, научно-исследовательское учреждение, учебное заведение, научная библиотека, специализированный файловый архив в Интернете, другие источники); отсутствие в тексте диалогов и разговорной лексики; отсутствие ссылок на информационные ресурсы заведомо ненаучного содержания; соотношение численности различных знаков препинания; соотношение объёмов текста, таблиц и формул; наличие библиографического списка, аннотации, ключевых слов, раздела «выводы»; наличие слов и фраз во всех словоформах, характерных для научных текстов — в целом, в аннотации, в выводах; средняя длина предложения; частотное распределение предложений по синтаксической структуре; частота орфографических ошибок.

♦ Принадлежность документа конкретной отрасли экономики сельского хозяйства оценивается при посредстве специальной базы данных, содержащей сведения об авторах, работающих в каждой области, об известных публикациях либо сайтах по данной проблематике, о терминах, специфических для данной предметной области, о периодических научных изданиях по данной теме. Документ оценивается на основании наличия и частоты ссылок: на имеющиеся в этой базе данных публикации (сайты); на другие публикации известных системе авторов; на публикации в известных системе периодических изданиях. Менее значимый признак — частота использования соответствующих терминов в целом по тексту и особенно в аннотации и выводах. Кроме того, учитываются векторы релевантности документов, ранее классифицированных роботом, на которые имеются ссылки в данном документе; специальности упоминаемых в документе диссертаций и авторефератов; упоминания конкретных научно-исследовательских и финансирующих организаций. Значения признаков, оценивающих научный характер текста, также могут варьировать по рубрикам — это дополнительная информация для классификации. Аналогичные признаки позволяют с достаточной точностью определить региональную релевантность документа.

♦ Соответствие документа определённому периоду времени определяется на основе: даты его публикации; дат, упоминаемых в аннотации и в выводах; датировки данных в таблицах; датировки ссылок в библиографическом списке; дат, встречающихся в тексте; периодов научной деятельности упоминаемых авторов; базы данных о «времени жизни» научных терминов.

База знаний должна непрерывно совершенствоваться в двух направлениях — расширение списка правил и уточнение вероятностей. Уточнение вероятностей производится автоматически при вмешательстве эксперта в процесс классификации (для этого в системе должен быть предусмотрен соответствующий интерфейс) на основе хорошо известных методов — см., например, упомянутую книгу К. Нейлора. Расширение списка правил производится по результатам анализа причин систематических ошибок робота: указывается и формализуется правило, по которому можно отличить документы, правомерно либо ошибочно помещаемые роботом в данную рубрику. После добавления нового правила условные вероятности опять-таки вычисляются автоматически. Робот также самостоятельно корректирует хранимые в базе знаний условные вероятности при изменении структуры рубрикатора и указании экспертами коэффициентов релевантности представительного множества документов данной рубрике. Для остальных документов робот вычислит коэффициенты релевантности новой рубрике самостоятельно.

Светлов Н.М., Зуев А.А. Перспективы автоматизации семантического анализа документов по аграрно-экономической тематике в системе управления АПК // Материалы международной научной конференции (декабрь 2001 г.): Сб.науч.тр.: Вып.8. М.: Изд-во МСХА, 2002. – С.231-236.

Принципиально важна структура рубрикатора. Он должен содержать следующие специальные рубрики: документы, не поддающиеся классификации; документы, не относящиеся к экономической науке; документы экономической тематики, не относящиеся к аграрной экономике. В первую из них попадают документы[3], не содержащие достаточного объёма текстовой информации для математического анализа семантики. Эти документы передаются для классификации экспертам. Во вторую — несомненно, самую многочисленную — попадают документы, не имеющие отношения к задачам системы и исключаемые из дальнейшего рассмотрения. Их идентификаторы хранятся в этой рубрике лишь как документальное подтверждение того факта, что их оценка состоялась. Периодически новые поступления в эту рубрику должны просматриваться экспертами на предмет обнаружения случайно попавших сюда документов аграрно-экономического содержания. Каждый подобный факт должен внимательно изучаться и приводить к уточнению правил вычисления показателей соответствия. В третью группу попадают документы по экономической тематике, не имеющие отношения к сельскому хозяйству. Они не подлежат дальнейшей классификации, но сведения об этих документах целесообразно предоставлять аграрным экономистам, которые могут найти в них научные результаты, приложимые к проблемам сельского хозяйства, пользуясь традиционным поиском по ключевым словам и датам.

Названные группы должны быть терминальными, т.е. не должны сами делиться на рубрики.

Основные рубрики определяют отрасль экономической науки, которой соответствует документ, отрасль сельскохозяйственного производства, регион, период времени. При необходимости эти рубрики могут иметь сколь угодно сложную иерархию подрубрик. На схеме представлен один из возможных наборов подрубрик.

Экономистам – пользователям системы доступно несколько стратегий поиска требуемого документа. Простейшая и в большинстве случаев достаточная — выбор одного или нескольких документов, в наибольшей степени соответствующих рубрике, отражающей научные интересы данного исследователя. Более сложная предполагает учёт соответствия сразу нескольким рубрикам — например, конкретизирующим научное направление, отрасль сельского хозяйства, период времени и регион. В этом случае для каждого документа вычисляется мультипликативный или комплементарный показатель соответствия требуемым рубрикам. Наконец, в рамках обеих названных стратегий возможно задание дополнительных критериев поиска по ключевым словам. Пользователи смогут подписаться на рассылку по электронной почте информации о вновь поступивших в систему документах с показателями соответствия тем или иным рубрикам, превышающими указанные пользователем значения.

Автоматизированная оценка и классификация документов позволяет решить задачу углубления семантического анализа: единицей классификации может быть не только документ, но глава, раздел и (в отдельных случаях) даже абзац. Можно сформулировать математический критерий максимально достижимой глубины семантического анализа: минимальная разность априорного и апостериорного векторов релевантности. Переход на более низкий уровень семантического анализа позволит значительно повысить производительность научного труда аграрных экономистов — особенно при работе с монографиями, диссертациями и научными отчётами.

[3] Говоря, что документ отнесён к некоторой рубрике, мы имеем в виду, что показатель его соответствия данной рубрике выше, чем показатели соответствия другим рубрикам.