О формализации понятий в анализе неструктурированной информации

Майсурадзе Арчил Иериевич,
Московский государственный университет им. М.В. Ломоносова

Введение

В последнее время потребности в анализе неструктурированной информации постоянно растут. Запрос к специалистам в области информационных технологий на автоматизированную поддержку такого анализа идет непосредственно от аналитических центров, а также транслируется через социологические научно-образовательные центры.

На российском рынке программного обеспечения уже присутствует ряд решений. Однако практически все эти решения, особенно с конкурентоспособной поддержкой русского языка, можно отнести к классу платформ для развертывания промышленных систем. Наблюдается серьезная нехватка решений для поддержки индивидуальной работы аналитиков. Обзор 38 программных систем (практически все зарубежные) для анализа неструктурированной информации и визуализации результатов такого анализа, проведенный летом 2010 года в АЦ при Правительстве РФ, показал, что все они существенно не соответствуют пожеланиям индивидуальных пользователей. Представляется, что основное препятствие для развития рассматриваемого класса ПО методологическое.

Группа специалистов в области социологии и информационных технологий из МГУ и ВШЭ разрабатывает персональное автоматизированное рабочее место для анализа неструктурированной информации, основными преимуществами которого являются: развитие системы от содержательных задач предметной области (а не от формальных методов), формализация и тиражирование экспертного знания методами машинного обучения.

Предлагаемая вниманию читателей работа демонстрирует ряд особенностей и сложностей, возникших в ходе совместной работы специалистов разных дисциплин над развитием общей методологии. В частности, будут рассмотрены разные подходы к пониманию, формализации и решению задачи «выделения концептов».

Трудности междисциплинарного общения

При решении задач, определенных данными, изложенными на естественном языке, возникает целый комплекс проблем. Такие проблемы уже с середины 60-х годов фиксируются и обсуждаются в публикациях. Но только сравнительно недавно стало понятно, что одна из основных проблем состоит в том, что «естественнонаучные» специалисты, создающие программные решения, не знают, какие задачи на самом деле решают их «гуманитарные» потребители – эксперты в предметных областях. Конечно, поиск взаимопонимания является неотъемлемой частью любого междисциплинарного исследования. К сожалению, в области анализа неструктурированной информации границы между научными дисциплинами совпали с границами между производителями и потребителями ПО. В итоге, развитие индустрии проходит гораздо менее сбалансировано между предложениями и потребностями, чем в области анализа структурированной информации. Например, в области интеллектуального анализа (структурированных) данных лидеры рынка ПО уже предлагают системы для решения содержательных задач, четко декларируют, что их ПО поддерживает соответствующие содержательным задачам методологии и индустриальные стандарты; объекты анализа точно формализованы, алгоритмы, реализующие методы ИАД, четко описаны.

Программные системы для анализа неструктурированной информации до сих пор предоставляют, по сути, лишь некоторые библиотеки функций, причем список этих функций сформирован самим производителем. Обычно для сторонних специалистов (даже в области информационных технологий) остается непонятным, почему пользователю был предложен именно такой список функций. Обычно неясно, как именно формализованы понятия, используемые в определении таких функций. А когда речь заходит о методах, лежащих в основе функциональности, а тем более о точных реализациях алгоритмов, тут уж сами производители ПО стараются сказать как можно меньше. Например, в ходе круглого стола по вопросам

анализа неструктурированной информации в социологических исследованиях, прошедшего в ВШЭ зимой 2011 года, неразрешимым вопросом для поставщиков ПО оказался следующий: «Укажите публикации по реализованным в Вашем ПО методам, на которые можно сослаться в научных работах».

Указанное положение дел неудобно для потребителей ПО, а для научной деятельности оно практически неприемлемо, так как методология научных исследований не поощряет рассуждения, в которых «неясно что» сделано «не пойми как». Даже если авторы методов, реализованных в коммерческом ПО, где-то публикуются (например, можно упомянуть труды конференции по компьютерной лингвистике «Диалог»), отсутствует связь между четким описанием метода и конкретным программным продуктом.

Представляется, что описанная ситуация сложилась именно из-за недостаточного развития методологии использования автоматизированных программных средств для решения содержательных задач. Думается, что сложности развития такой методологии обусловлены тем, что вообще методологии естественнонаучного и гуманитарного исследования несколько отличаются. Если в естественных науках принято практически всегда проводить логические рассуждения с использованием точно формализованных «объективных» понятий, то в гуманитарных дисциплинах важную роль играет сравнение «субъективных» точек зрения.

Как показало личное наблюдение автора за практикой взаимодействия социологов и специалистов в области информационных технологий, последние не сразу понимают, что первые хотят сделать. И тогда вместо обмена идеями происходит лишь обмен словами; обе стороны пытаются ухватиться за некоторый набор терминов, которые, как им кажется, они совместно используют. Трудность не только в том, что в этот набор попадают термины, имеющие различное определение в разных дисциплинах. Серьезная проблема в том, что алгоритмисты вынуждены дополнять этот набор словами, которые социологи упомянули в формулировке задач. Специалисты в области естественных наук вынуждены строить свои определения таких понятий, пришедших из предметной области.

В настоящее время можно говорить о следующих основных подходах к формализации понятий предметной области.

- Специалисты в области естественных наук сами предлагают какую-то формализацию, практически без участия потребителей. Сегодня это наиболее распространенный подход в области анализа неструктурированной информации. В настоящее время основная проблема этого подхода состоит в том, что формализация понятий почти всегда идет от некоторого метода. Методологически должно быть наоборот: сначала ставится задача, а уж потом рассматриваются разные методы её решения. Нередки жалобы на то, что как алгоритмист не понял, чего хотел потребитель, так и потребитель не понял, что предложил алгоритмист.
- Формализация фиксируется как набор готовых ответов для заданного набора конкретных ситуаций. Профессиональная подготовка набора готовых ответов занимает годы. При появлении новых ситуаций потребителю придется довольно долго ждать поддержки со стороны производителя. Потребитель обязан согласиться с предоставленными ему готовыми ответами. Типичный пример – словари, тезаурусы.
- Просим у пользователей обучающие примеры. Много разных несогласованных мнений. Не всегда понятны причины мнения. Мнение нестабильно (меняется во времени). Мнение зависит от текущей деятельности. Эксперт не хотел бы тратить время на подготовку примеров. Способ узнать мнение у эксперта опять-таки требует формализации. Можно индивидуально поддержать одновременно разные виды деятельности или разных экспертов. Оперативная актуализация системы.
- Извлекаем обучающие примеры из уже накопленной информации. В первую очередь речь идет о том, что в Интернет накоплено колоссальное количество примеров, которые можно использовать для обучения. Пока больше надежд, чем результатов. Подход дает возможность следить как за поставщиками информации, так и за её потребителями.

В настоящее время на рынке программных средств для анализа неструктурированной ин-

формации первые два подхода получили подавляющее распространение. В данной работе приводятся примеры работы в рамках подходов 3 и 4.

Задачи анализа текстов

Разумеется, когда речь идет об анализе информации, выраженной текстом на естественном языке, в самой общей постановке хотелось бы говорить о понимании машиной текста, автоматическом выделении смысла. В формулировке, приближенной к реальным возможностям современных систем, стоит задача восстановления отдельных объектов и их взаимосвязей, которые были описаны, либо упомянуты, либо подразумевались автором текста неявно. Это заставляет обращаться к работе с некоторыми объектами, отсутствующими в тексте в явном формализованном виде, но описанными автором и несущими реальное значение для решения содержательной задачи.

Рассмотрим некоторые примеры, в которых пытаются формализовать и решить задачу поиска смысловых понятий текста и взаимосвязей между ними.

В ходе построения информационно-поисковой Интернет-системы задача в первую очередь сводится к следующему. Машине необходимо составлять базу данных, в которой хранится информация о некоторых объектах, процессах и явлениях, описанных в текстах, и эти записи сопровождаются информацией о свойствах, качествах и взаимосвязях описанных в текстах объектах. При этом один и тот же объект может описываться с использованием различных слов (проблема использования синонимов). Объект может даже не описываться, а упоминаться косвенно, например, надо найти тексты о бейсболе, в которых слова «бейсбол» нет. Кроме этого, в различных текстах одинаковыми словами могут описываться различные объекты (проблема использования омонимов). В информационно-поисковой системе, с одной стороны, стоит задача сужения области поиска, исключения из нее документов, упоминающих ненужные пользователю объекты/события, а с другой стороны, возникает необходимость застраховаться от излишнего сужения, традиционно возникающего за счет того, что пользователь может спрашивать об объекте совсем не теми словами, которыми пользовался автор при описании.

Вслед за проблемой восстановления объектов, описанных в тексте, возникают проблемы восстановления взаимосвязей, отношений, характеристик и так далее. Если задача по выделению объектов в упрощенном виде решаема известными средствами — составление словарей объектов, словарей синонимов и т.д., — то восстановление связей, отношений и характеристик, особенно описываемых неявно и/или разрозненно, является весьма сложной задачей, не решаемой без применения интеллектуальных технологий, базируемых на проработанных моделях естественного языка, моделях построения текстов, моделях мышления.

Заслуживает упоминания задача дешифровки исторических систем письма. Имеется осмысленный текст, записанный на неизвестном языке. Нужно, исходя в первую очередь из самого текста, выяснить свойства неизвестного языка и уже затем, путем сопоставления с известными языками и привлечения с большой осторожностью внетекстовой информации (археологических, исторических, филологических и иных сведений), передать смысл неизвестного текста. Возможность исследовать текст формальными методами перерастает в необходимость, если требуется максимальным образом исключить субъективный анализ текста.

Схожие задачи стоят и перед разработчиками программного обеспечения машинного перевода текстов. Похожей проблемой занимаются разработчики систем создания и определения неестественного происхождения документа. Упомянем задачи распознавания спама по контексту, задачу построения семантических сетей, используемых в словарях и тезаурусах, задачи рубрикации текста, выделения тематики, автоматического аннотирования и реферирования. Существуют также более локальные задачи, например, задача автоматического поиска тавтологий в тексте или задача автоматического переноса слов (например, в английском языке омонимы переносятся по-разному).

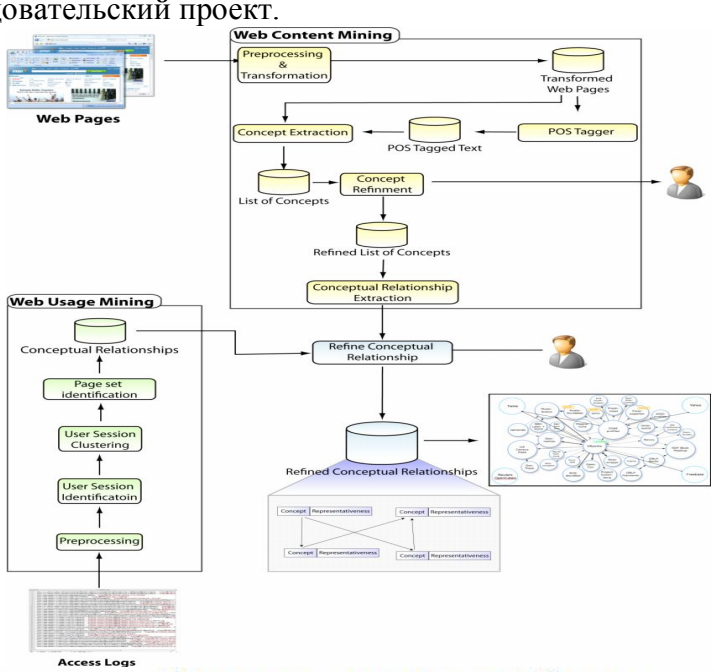
В последнее время стали появляться работы по автоматизации накопления и использования общих знаний об окружающем мире, необходимых для понимания текста. Кроме того, наметились подходы для анализа одного и того же текста, как с точки зрения автора, так и с

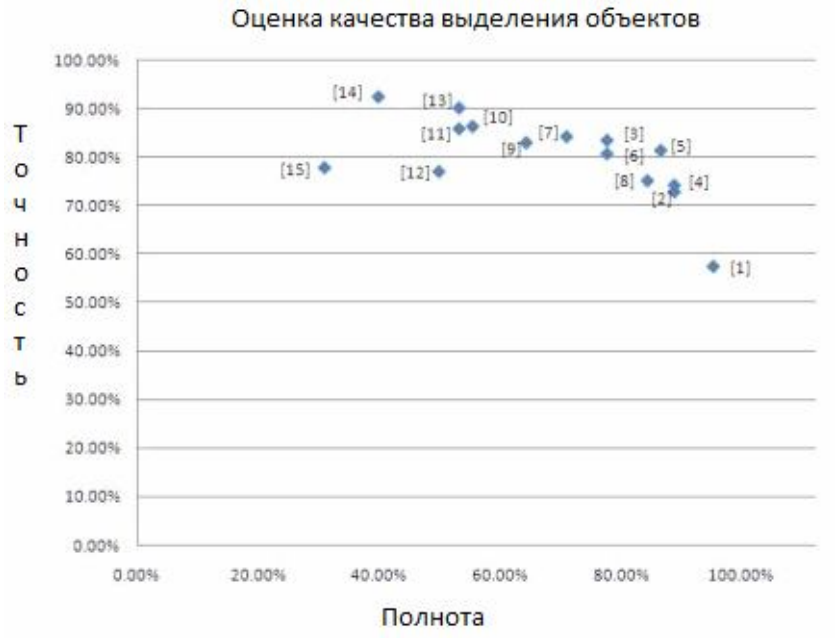
точки зрения читателя. В данной работе в качестве основного иллюстративного примера использована задача выделения концептов.

Использование накопленной информации

В настоящее время формализацию знаний о предметной области чаще всего реализуют в виде онтологий. Онтологии используются в процессе программирования как форма представления знаний о реальном мире или его части. Большинство используемых сегодня онтологий созданы вручную, их разработка и наполнение — это долгий и трудоёмкий процесс. Они оцениваются чаще с точки зрения применимости, чем полноты. Пользователь практически лишен возможности актуализировать накопленные в системе знания самостоятельно.

Ниже приводится архитектура системы, предназначенной для полуавтоматического наполнения онтологии информацией. Новизна системы определяется тем, что собирается как информация о текстах определенных авторов, так и информация, отражающая поведение читателей. Пример приводится именно для того, чтобы продемонстрировать сочетание двух потоков информации, а не для рассказа о методах. Указанная система разрабатывается как студенческий исследовательский проект.





В данной работе основное внимание уделяется задаче выделения концептов, что в теории

онтологий соответствует экземплярам и понятиям. Экземпляры — это отдельные объекты, понятия (классы) — абстрактные группы объектов. Онтология может обойтись без конкретных объектов или не различать экземпляры и понятия. Поэтому позволим себе использовать только термин концепт.

Система сочетает в себе два подхода: Web Content Mining и Web Usage Mining. Первый подход анализирует текст и структуру страниц, именно эта информация характеризуют автора. Второй подход анализирует информацию о поведении читателей. Интерес представляет именно совместное применение подходов.

При анализе текста концепты выделяется на основе взвешенных частот. Использовать непосредственно частоты не удается, так как размер каждого отдельного текста мал. Веса назначаются на основе HTML-тэгов, в которые концепт заключен. Кроме того, привлекается морфологическая информация, в особенности о частях речи. Это полезно для правильного отбора цепочек слов, обозначающих концепт.

Информация об использовании страниц позволяет выявить шаблоны поведения читателей. Разумно предположить, что пользователи перемещаются по сайту в соответствии со своим восприятием предлагаемого материала. Такой анализ группирует страницы, даже если между ними нет прямых связей.

Постараемся оценить качество результатов, даваемых системой. Трудность в том, что для онтологий не выработано стандартных показателей качества. Ближе всего в данной ситуации представляются понятия точности и полноты, используемые при оценке качества информационного поиска. Однако при этом приходится проводить сравнение результатов работы системы с некоторой эталонной онтологией. Результаты нескольких запусков системы приведены на диаграмме.

Использование экспертных примеров

В данном разделе рассматривается задача выделения «концептов» и «релевантных понятий» в одном тексте. Проведем частичную формализацию терминов «концепт» и «релевантное понятие». Концепт в тексте представлен одним или несколькими вхождениями слов или словосочетаний. Концепт — это некоторый набор вхождений; релевантные понятия — такие концепты, которые «соответствуют ключу информационной сферы». Такая частичная формализация была положена в основу опроса экспертов, чтобы собрать примеры «концептов» и «релевантных понятий». Было принято решение отобрать ряд русских текстов и предложить группе экспертов выделить в них группы вхождений, соответствующие концептам и релевантным понятиям.

Отбор текстов осуществлялся на основе следующих посылок. Во-первых, эксперты должны с текстами справиться, а значит, объем каждого текста и их количество в опросе должны быть ограничены. Несколько попыток показали, что для человеческого восприятия удобнее всего маленькие, порядка 1000 символов, тексты — это два-три небольших абзаца, или один большой. Во-вторых, каждый текст должен быть связным и завершенным, в нем должна четко прослеживаться основная идея. Это не должен быть «выдранный из контекста кусок». Лучше всего поставленным требованиям отвечают тексты в жанре журналистских заметок. Это своего рода резюме к большой статье, где акцент сделан уже на основных ключевых моментах текста, а потому экспертам довольно легко выделить искомые понятия.

Тематика для текстов, входящих в набор, выбиралась одна: «Последствия аномальной жары и лесных пожаров в России летом 2010 года». Это связано с тем, что наличие одной проблемы в корпусе практически наверняка гарантирует составление одного терминологического словаря. Такой стиль исследования называется работой в одном дискурсе. К тому же, выбор темы был обусловлен сравнительной простотой и равной актуальностью проблемы для всей экспертной группы.

Некоторые характеристики набора текстов. Максимальное число символов в одном тексте 1787, среднее число символов 1060, минимальное число символов 646. Максимальное число слов в тексте 231, среднее число слов 137, минимальное число слов 85. Максимальное число предложений в тексте 20, среднее число предложений 8, минимальное число предложений 3.

Максимальное число абзацев 4, среднее число абзацев 3, минимальное число абзацев 2.

Выбор экспертной группы основывался на следующих факторах. Каждый эксперт должен воспринимать тематику текстов, понимать, что требуется найти в предложенном тексте, иметь начальные профессиональные навыки обработки текстов с лингвистической точки зрения. В частности, перечисленным требованиям удовлетворяют люди с филологическим или лингвистическим образованием. В качестве экспертов были привлечены студенты факультета иностранных языков и регионоведения МГУ им. М.В. Ломоносова.

В данной работе после опроса экспертов мы хотим получить концепты как группы, состоящие из одного или нескольких вхождений, и помеченные среди них особые группы, которые мы называем релевантными. При обдумывании вопроса, в каком виде мы бы хотели получить информацию от экспертов, возник другой важный вопрос: а как экспертам было бы удобнее (интереснее) решать поставленные перед ними задачи. Оказалось, что заполнять какую-либо анкету по графам эксперты не хотят.

В результате нескольких проб сформировалась следующая процедура опроса. Экспертам предлагалось помечать вхождения, относящиеся к разным концептуальным группам, фломастерами разных цветов, а рядом с группой, которую они считают релевантной, ставить какой-нибудь отличительный символ, например, звездочку. На практике эксперт крайне редко выделял в одном тексте более 4 групп, так что количество доступных фломастеров (не менее 8) не было ограничением.

Дополнительно в анкете, предложенной экспертам, была отдельная графа под комментарии. В этой графе каждый эксперт мог в свободной форме указать, по какой причине он сгруппировал те или словосочетания, почему считает выделенные им группы релевантными, попытаться подумать о тех закономерностях, на которые можно было бы впоследствии ориентироваться при составлении алгоритмов.

Как показывает практика, даже с маленькими текстами экспертам утомительно работать, если их много. Поэтому было решено каждому эксперту предъявлять по 8 текстов. Каждый текст был предъявлен хотя бы 4 экспертам. Эксперты работали над текстами от 40 минут до полутора часов. Приведем результаты опроса на примере одного из текстов.

«Реальная площадь лесных пожаров с начала пожароопасного периода 2010 года в России примерно вдвое превышает данные субъектов РФ и составляет около трех миллионов гектаров, сообщил в понедельник журналистам начальник Авиалесоохраны Федерального агентства лесного хозяйства России Николай Ковалев.

По официальным данным Рослесхоза, на начало сентября текущего года во время пожароопасного периода на территории страны возникло более 30 тысяч природных пожаров общей площадью более чем 1,246 миллиона гектаров. Однако, эти данные по площади лесных пожаров, предоставленные регионами, не совпадают с данными аэрокосмического мониторинга, проведенного в то же время, и сильно занижены. Это связано с тем, что главы ряда субъектов не хотят показывать истинное положение дел».

№ эксперта	Группы концептуальных понятий	Релевантные понятия	Комментарии
1	Группа 1: *Реальная площадь лесных пожаров (1, 1–4), истинное положение дел (2, 98–100)*; Группа 2: *По официальным данным Рослесхоза (2, 38–41), данные по площади лесных пожаров, предоставленные регионами (2, 69–75).*	Группа 1	Концептуальными группами являются те, ядром которых служит существительное.
2	Группа 1: *Реальная площадь (1, 1–2), на территории страны (2, 51–53), общей площадью (2, 60–61), по площади (2, 70–71);* Группа 2: *лесных пожаров (1, 3–4), пожароопасного*	Группа 1 Группа 4	Группа 1 — мера площади и пространства Группа 2 — пожары и

	периода (1,7-8), пожароопасного периода (2, 49–50), природных пожаров (2, 58–59), лесных пожаров (2, 72–73); Группа 3: *2010 года (1, 9–10), трех миллионов (1, 22–23), 30 тысяч (2, 56–57), 1,246 миллиона (2, 64–65);* Группа 4: *вдвое превышает (1, 14–15), данные (1, 16), не совпадают (2, 76–77), сильно занижены (2, 88–89), не хотят показывать истинное положение дел (2, 98–103).*		все что их касается Группа 3 — по наличию цифр Группа 4 — несовпадение предоставляемой информации с реальностью
3	Группа 1: *Реальная площадь (1,1-2), около трех миллионов гектаров (1,21-24), площадью более чем 1, 246 миллиона гектаров (2,61-66);* Группа 2: *лесных пожаров (1,3-4), пожароопасного периода (1,7-8), пожароопасного периода (2,49-50), природных пожаров (2,58-59), лесных пожаров (2,72-73).*	Группа 1	
4	Группа 1: *Реальная площадь лесных пожаров (1, 1–4), составляет около трех миллионов гектаров (1, 20–24), на территории страны (2, 51–53), 30 тысяч природных пожаров общей площадью более чем 1,246 миллиона гектаров (2, 56–66), данные по площади лесных пожаров (2, 69–73);* Группа 2: *с начала пожароопасного периода 2010 года (1, 5–10), на начало сентября текущего года (2, 42–46).*	Группа 1	Группа 1 — площадь Группа 2 — время

Как и ожидалось, ответы экспертов частично разошлись. Интересно отметить, что некоторые группы сформированы по составу, например, по наличию цифр. Важно подчеркнуть, что кроме общей идеи выделения групп, сформулированной в начале раздела, никакое мнение экспертам не предлагалось. В некоторых случаях эксперты составили такие группы, о которых авторы исследования не предполагали. Следовательно, если бы алгоритмисты попытались сами формализовать задачу или метод решения, то неизбежно работали бы с неверным представлением о потребностях потребителей.


Реальная площадь лесных пожаров с начала пожароопасного периода 2010 года в России примерно вдвое превышает данные субъектов РФ и составляет около трех миллионов гектаров, сообщил в понедельник журналистам начальник Авиалесоохраны Федерального агентства лесного хозяйства России Николай Ковалев.

По официальным данным Рослесхоза, на начало сентября текущего года во время пожароопасного периода на территории страны возникло более 30 тысяч природных пожаров общей площадью более чем 1,246 миллиона гектаров. Однако, эти данные по площади лесных пожаров, предоставленные регионами, не совпадают с данными аэрокосмического мониторинга, проведенного в то же время, и сильно занижены. Это связано с тем, что главы ряда субъектов не хотят показывать истинное положение дел.

Был предложен способ визуализации результатов опроса, названный картой концептов. Работе 4 экспертов соответствуют цвет, фон, размер и эффект шрифта.

Для применения методов ИАД в дальнейшем была сформирована обучающая выборка, для чего консолидация мнения экспертов проводилась двумя способами: наложение групп и их разбиение. В итоге каждому тексту соответствовало несколько групп с разными весами.

Дискурс-анализ электронных СМИ с применением сетевого подхода (пример обсуждения вступления РФ во Всемирную торговую организацию)

Просянюк Дарья Вячеславовна, *НИУ ВШЭ*

Сегодня каждое принятое или готовящееся к принятию экономико-политическое решение находит отражение в информационном Интернет-пространстве. Здесь каждая страна имеет возможность собственного позиционирования и коммуникации с другими странами. Поэтому планирование долгосрочной идеологической перспективы действий в этом полифакторном пространстве представляется вопросом исключительной важности. Примером такого экономико-политического решения может служить вступление Российской Федерации во Всемирную Торговую Организацию. Данная проблема уже не первый год является крайне актуальной для российской внутренней и внешней политики и экономики.

Любое экономико-политическое решение (особенно такого масштаба как вступление в ВТО) должно получить легитимацию в различных слоях населения. Соответственно, информационное пространство — это визуализатор экономико-политических решений, которые принимаются и готовятся к принятию, это своего рода «поле», где взаимоотношения между странами становятся очевидными. Таким образом, информационное пространство является уникальным репрезентатором внутри- и межстрановых отношений: в нем апробируются способы аргументации и легитимации принимаемых «наверху» решений, а также с относительно большей полнотой, чем в других открытых источниках, эксплицируются неявные взаимоотношения и взаимосвязи между странами.

1. Методика

Шаг 1: выбор объекта. Всякий социально-значимый объект описывается в информационном пространстве в различных аспектах. На первом этапе построения эгоцентричной сетевой карты выбирается аспект, который за счет представленности в различных источниках информации обеспечит достаточный размер выборки и вариативность типов информационных сообщений. В нашем исследовании для построения эгоцентричной сетевой карты была выбрана тема «Вступление Российской Федерации во Всемирную торговую организацию».

Шаг 2: кодирование сообщений. Под сообщением мы понимаем всякий текст, имеющий отчетливо обозначенные границы. Традиционно в журналистике принято выделять три жанра написания текстов: информационный, аналитический и художественно-публицистический[1]. Для построения эгоцентричной сетевой карты используются тексты информационного и аналитического жанров. Характер содержания в текстах информационной группы эмпирический: они в первую очередь передают аудитории фактические сведения о действительности. Такие тексты позволяют выявлять и классифицировать информационные поводы и информационные события. В аналитических текстах содержится информация эмпирико-теоретического характера: здесь главным образом выполняется задача систематизации фактов, их объяснения и обобщения, в них содержатся реконструкции позиций и типов аргументации, которых могут придерживаться различные представители разных стран. Описанные группы относятся к профессиональным текстам.

При построении эгоцентричной сетевой карты могут представлять интерес и непрофессиональные тексты — блоги. Поскольку они репрезентируют неофициальный взгляд на проблему, интерпретация информационных поводов в них может значительно отличаться от официальных источников.

[1] Ким М.Н. Жанры современной журналистики. М.: Изд-во Михайлова В.А., 2004.