От редакции. *Продолжаем начатую в №1 публикацию серии статей, знакомящих читателей с возможностями новых перспективных информационных систем.*

Технологии обработки неструктурированных массивов информации,

используемые для анализа, выявления неявных и аномальных связей между субъектами политической и экономической деятельности

(на примере СМИ и других открытых источников, в том числе сайта zakupki.gov.ru)*

Т.В. ИГНАТОВА, В.А. ИВИЧЕВ, ООО «Медиалогия», Москва.
E-mail: tignatova@mlg.ru, vivichev@mlg.ru

В статье рассмотрены основные вопросы, связанные с прозрачностью и возможностями анализа государственных закупок, поставлены задачи, от которых зависит существенное улучшение качества процесса государственных закупок. Предложены технологии решения указанных задач. Основным их достоинством является автоматизированный процесс оперативного предоставления полной аналитической информации различным категориям потребителей: участникам закупок, органам власти, осуществляющим контроль как за исполнением законодательства в сфере государственных закупок, так и их эффективностью, а также представителям общественных и политических организаций, журналистам, блогерам и др.

Ключевые слова: государственные закупки, прозрачность и доступность информации, выявление нарушений, кластеризация, информационная система

Отношения в сфере размещения заказов на поставки товаров, оказание услуг, выполнение работ для государственных и муниципальных нужд (далее – госзакупки) регулируются федеральными законами № 94-ФЗ и № 223-ФЗ, в соответствии с которыми информация о размещаемых заказах и заключаемых контрактах должна публиковаться в открытом доступе на официальном сайте Российской Федерации для размещения информации о размещении заказов http://www.zakupki.gov.ru.

* Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации. УДК 004.031.42.

Публикуемая информация о размещаемых заказах и заключаемых контрактах представляет непосредственный интерес для широкого круга потребителей, включая коммерческие структуры, государственных заказчиков, контролирующие органы, ведомства, а также представителей широкой общественности, озабоченных проблематикой государственных закупок.

Наличие большого количества потенциальных потребителей такой информации обусловливает необходимость структуризации публикуемых данных, а также их агрегации с целью дальнейшего формирования аналитической отчетности. Эти данные представляют собой значительный массив первичной информации, потенциально позволяющей на основании объемов и соотношений затраченных средств, длительности контрактов, взаимоотношений контрагентов и других атрибутов проводить исследования рынка госзакупок. Но ресурс http://www.zakupki.gov.ru **не предоставляет развернутых средств по анализу структуры госзакупок, выявлению закупок, оформленных с нарушениями, а также по поиску закупок, поэтому практически не содержит возможностей для организации конкурентной среды.**

Отсутствие должного уровня прозрачности и возможностей анализа, наряду с коррумпированностью в сфере государственных закупок, приводит к существенным потерям государства и общества от низкой эффективности процесса размещения государственного и муниципального заказа. Эти потери условно можно разделить на четыре вида:

- финансовые – заключение сделок на невыгодных для государства условиях (завышение цен закупаемой продукции по сравнению с текущим рыночным уровнем, включение в условия государственных контрактов предоплаты вместо отсрочки платежа и т.п.);
- количественные – завышение или занижение объема поставляемых материалов или оказанных услуг по сравнению с необходимым количеством; приобретение товаров или услуг в личных целях ответственных чиновников, а не для удовлетворения государственных нужд, и т.п.;
- качественные – заключение сделок с нарушением требуемых технических условий, таких как поставка товаров, выполнение работ или оказание услуг ненадлежащего качества (худшие условия гарантийного и послегарантийного обслуживания;

недостаточные требования по контролю качества выполнения работ и услуг и пр.);

• политические – ухудшение инвестиционного климата, потеря доверия со стороны граждан к государственным структурам и государству в целом, расшатывание экономической и финансовой системы страны, нарушение принципов свободной конкуренции и т.п.

Отечественные разработки в области анализа государственных закупок представлены коммерческими системами – агрегаторами заказов типа http://www.magelan.su; http://trade.su; http://ist-budget.ru, а также системой Seldon2010. Данные системы ориентированы на предоставление и поиск информации о проводимых на государственных и коммерческих торговых площадках закупках. Функциональность поиска реализована с переменным успехом и разной степенью удобства использования, однако ни у одной из этих систем нет поиска с использованием механизмов кластеризации и лингвистического анализа вложенных документов.

Разработки в области противодействия коррупции и повышения прозрачности в сфере госзаказа, как правило, ограничены решениями энтузиастов (активными пользователями блогов) и не представляют расширенной функциональности по поиску и мониторингу закупок.

Таким образом, **в настоящий момент на рынке отсутствуют решения, которые объединяли бы функции по поиску информации о проводимых на государственных и коммерческих торговых площадках закупках и функции, предназначенные для повышения прозрачности и уменьшения нарушений в сфере государственного заказа**.

В связи с крайней важностью задач повышения прозрачности и формирования конкурентной среды в сфере госзакупок в 2010 г. были проведены научно-исследовательские работы, нацеленные на апробацию подходов к решению указанных задач. В качестве отправной точки был использован международный опыт по раскрытию подобной информации – ресурс http://usaspending.gov/, предоставляющий пользователям анализ данных о расходах госбюджета по различным срезам, включая отраслевую и территориальную составляющие.

Результатом проведенных работ явился ресурс РосГосЗатраты, открытый для общего доступа по адресу

http://rosspending.ru/. Он предоставляет аналитическую отчетность, полученную на основе информации единственного источника – реестра государственных контрактов на основе агрегации данных с возможностью их детализации до отдельных контрактов за 2007–2010 г. Анализ контрактов позволил определить основные виды нарушений при размещении документации государственного заказа, прохождении конкурсных процедур и выполнении заключенных контрактов, определить основные тренды выполнения государственных контрактов.

Следующим этапом являются описываемые в рамках данной статьи работы, направленные на **создание информационной системы, реализующей следующие задачи**:

• сбор, агрегация и лингвистическая обработка информации из открытых источников (федеральных печатных источников, федеральных телевизионных каналов, региональных СМИ, интернет-СМИ, сайтов органов государственной власти, интернет-блогов и форумов;

• сбор, агрегация и лингвистическая обработка информации с сайта zakupki.gov.ru;

• семантико-фактографическая обработка собранной и обработанной информации, выявление субъектов и объектов экономической деятельности, нарушений при проведении закупочных процедур и реализации государственных контрактов, неявных и аномальных связей между субъектами экономической деятельности, определение индикаторов и показателей качества государственных закупок.

Используемые технологии

В целях сбора, агрегации и лингвистической обработки неструктурированной информации открытых источников используются технологии мониторинга и оперативного получения неструктурированных данных (обновленных фактически в режиме реального времени). Применяются технологии упорядочивания, систематизации и структуризации неструктурированных материалов, лингвистическая обработка, включающая в себя выделение информационных объектов (физических и юридических лиц, географических понятий и брендов), тематическая и жанровая классификация текстов, выявление групп информационных событий и автоматическая кластеризация поступающих информационных материалов,

выделение прямой и косвенной речи информационных объектов, ранжирование важности упоминания информационного объекта в тексте сообщения (главная, второстепенная или эпизодическая роль), определение количества эфирного времени с сюжетами, в которых освещается информационный объект, определение характера упоминания объекта и др.

Для решения задачи сбора, агрегации и лингвистической обработки структурированной и неструктурированной информации с сайта zakupki.gov.ru используются технологии мониторинга и получения открытой информации с сайта в режиме реального времени, а также структуризации и лингвистической обработки, включающие классификацию (рубрикацию) – разбиение объектов (закупок) по существующим в информационной системе сайта справочникам, автоматическая категоризация закупок на группы в рамках автоматически создаваемых в справочниках подкатегорий посредством кластеризации, выделения и анализа неструктурированных документов заказа (технических заданий, запросов котировок, аукционной документации и др.).

Для семантико-фактографического преобразования собранной и обработанной информации, выявления субъектов и объектов экономической деятельности, нарушений при проведении закупочных процедур и реализации государственных контрактов, неявных и аномальных связей между субъектами экономической деятельности, определения индикаторов и показателей качества государственных закупок используются **технологии выделения информационных объектов** (участников, фактов нарушений, предметов и цен закупок), установления неявных и аномальных связей между субъектами экономической деятельности.

Методы кластеризации

Как показывает практика, одна из основных проблем при работе различных групп участников на рынке государственных закупок – выявление интересующих закупок на основе совокупности поисковых критериев либо прочих неявных признаков, а также мониторинг. Основная проблема заключается в высокой вероятности пропуска релевантных закупок из-за большого количества анализируемой информации и ее представления в неструктурированном виде.

В основе существующих информационных систем лежит принцип категоризации закупок на основе фиксированного рубрикатора, который позволяет отнести закупку к той или иной отрасли экономической деятельности на основании ее параметров. Недостатками данного подхода являются необходимость ручной обработки большого количества закупок, большие объемы реально используемых рубрикаторов, трудности с классификацией отдельных закупок и, как следствие, большой процент ошибок. Описанные проблемы ведут к тому, что процент «упущенных» заинтересованными участниками закупок остается достаточно высоким.

Для решения проблемы предлагается **использование инновационного подхода, связанного с так называемой «кластеризацией» закупок**. Кластеризация представляет собой автоматический процесс объединения закупок в схожие группы – кластеры, или темы – на основании публикуемой в них информации. Закупки в рамках одной темы представляют собой закупки схожих товаров и услуг, проводятся у одного заказчика или имеют другие схожие параметры. Кластеризация проводится автоматически, с использованием алгоритмов лингвистического анализа приложенных к закупке неструктурированных документов, что позволяет максимально повысить качество группировки и предоставить пользователям принципиально новый уровень работы с государственными закупками.

Кластеризация (англ. Data clustering) – процесс разбиения заданного множества объектов на кластеры так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Множеством объектов является поток информации о государственых закупках, публикуемых на сайте http://www.zakupki.gov.ru, а объектами – собственно закупки. Существует множество методов кластеризации, опишем используемый в статье метод гравитационной кластеризации. Суть метода: вводится величина – радиус притяжения (совокупность параметров, которым должна соответствовать закупка), далее если «расстояние» между двумя закупками меньше радиуса притяжения (закупки схожи), то закупки образуют кластер. Соответственно, если новая закупка «притягивается» к любому уже существующему кластеру, она включается в него.

Объединение закупок в кластеры (темы) происходит по содержимому приложенной к ним документации (извещения, конкурсная документация, протоколы и т.п.). Выявление ключевых

фрагментов в документации происходит посредством применения лингвистических алгоритмов.

Закупки добавляются в кластер, если произошло совпадение (полное или частичное, в определенном процентном соотношении) по объектам системы с другими закупками кластера.

При создании кластера выбирается главная закупка, формируется заголовок кластера (повторно выбираются при перекластеризации).

Главной закупкой кластера становится закупка с наибольшим весом следующих параметров:

- связанность с максимальным количеством сообщений в кластере по сравнению с другими статьями;
- дата и время закупки (самые актуальные).

Если удалить главную закупку из кластера, то он не «развалится». Кластер будет оставаться без главной закупки до тех пор, пока к нему не добавится очередная закупка, прошедшая кластеризацию, и он (кластер) не отправится на перекластеризацию, во время которой произойдет очередной пересчет (сравнение закупок кластера по вышеперечисленным критериям) на предмет выявления главной закупки.

Проблема предотвращения объединения кластеров (если между ними есть «дорожка из схожих закупок», но кластеры не являются близкими по содержанию) решается следующим образом: вводится дополнительный параметр – связанность текста и связанность кластера. Проверка закупки перед присоединением к кластеру такова: количество текстов, с которыми связан вновь добавляемый в кластер текст, делится на общее количество текстов в кластере, полученное значение является связанностью текста, оно должно быть больше/равно по величине связанности кластера – усредненной связанности текстов в кластере.

Процесс кластеризации фактически состоит из этапов:

- этап кластеризации (непрерывен): анализ вновь поступивших закупок и включение их в уже имеющиеся темы, проверка схожести с независимыми закупками (которые после аналогичной проверки не были включены в какой-либо кластер для последующего объединения в новый кластер или определены как независимые);
- этап «перекластеризации» дискретен. Он имеет свое расписание: из имеющихся кластеров выбираются те, которые были обновлены (была добавлена еще одна закупка) с момента последнего процесса «перекластеризации», выбранные закупки проверяются на возможность объединения или разбиения (когда в кластере появляется вторая главная закупка).

Представление результатов анализа неструктурированных данных

Разрабатываемая в рамках настоящего проекта система может быть применена в различных отраслях экономической деятельности. Основными ее потребителями, очевидно, будут являться сотрудники коммерческих структур. Данные компании на рынке государственных закупок выступают в качестве поставщиков. Главная цель этих пользователей – получение объективной информации о состоянии рынка, актуальных заказах, доле и конкурентоспособности коммерческих компаний в структуре госзаказа.

В качестве примера приведем опыт продажи своей продукции предприятиям и компаниям госсектора региональным молокозаводом. Сотрудники отдела продаж завода получают уникальную возможность выхода на рынок государственных закупок с минимальными трудозатратами на поиск новых заказов. Использование мониторинга рынка закупок по поисковым атрибутам, сформированным информационной системой (и отсутствующим в настоящее время на сайте zakupki.gov.ru), позволит всегда быть в курсе изменений на рынке, публикаций новых заказов на поставки молочной продукции и подобной информации. Размещенная в системе аналитическая отчетность позволит оценить темпы роста рынка молочной продукции в регионах и станет мощным средством поддержки принятия решений о выходах на новые рынки. Такие возможности анализа позволят значительно повысить эффективность работы предприятий-поставщиков в области госзаказа.

Другая категория потребителей информации – сотрудники организаций, выступающих на рынке госзакупок в качестве заказчиков, получат объективную полную и актуальную информацию о состоянии рынка с точки зрения реализованных и реализующихся государственных заказов, смогут оценить надежность и конкурентоспособность поставщиков, использовать данные системы для подготовки и размещения новых заказов.

Для заказчиков (равно как и для поставщиков) представляет интерес поисковая и аналитическая функциональность разрабатываемой системы, позволяющая быстро и эффективно искать схожие закупки и знакомиться с аналогичными

предложениями. Данная информация также позволит определять эффективность проводимых заказчиком закупок относительно всего рынка госзакупок.

Другая группа потребителей – это сотрудники органов власти, осуществляющие контроль как исполнения законодательства в сфере госзакупок, так и их эффективности. Их задача – выявление и пресечение нарушений законодательства в процессе размещения государственных заказов и заключения государственных контрактов. Для удовлетворения потребностей данной группы реализован целый спектр аналитических отчетов, позволяющих отслеживать изменения отдельных показателей, таких как экономия госбюджета на проводимых закупках, эффективность конкурентных процедур и т.п., а также оперативно выявлять нарушения и попытки сокрытия информации о размещаемых заказах. В настоящее время система позволяет выявлять следующие виды нарушений.

Некачественное заполнение информации о закупке

Признаки нарушения:

Присутствуют слова, содержащие одновременно как кириллические, так и латинские символы в названии заказа или любого из лотов заказа.

В названиях заказов ключевые слова написаны в виде отдельных букв, разделенных пробелами (например, «Закупка Б Ы Т О В О Й техники»).

Заголовок заказа не содержит ни одного ключевого слова, которое могло бы пояснить его суть.

Код ОКДП не имеет достаточной детализации.

Неверная длина кода ИНН или КПП организации-поставщика.

Отсутствие кода ИНН или КПП организации-поставщика.

Ошибки в коде ИНН организации-поставщика.

Использование в коде ИНН или КПП символов, отличных от цифр.

Отсутствие страны в адресе поставщика.

Отсутствие имени либо телефона контактной персоны поставщика.

Отсутствие или искажение наименования организации-поставщика.

Отсутствие региона организации-поставщика.

Нарушение сроков размещения информации в реестре контрактов.

Нарушение сроков заключения контрактов.

Нецелевое использование средств

Признаки нарушения:

Предметы роскоши: в предмете заказа или одного из лотов указан товар, который присутствует в списке «Роскошь», и стои-

мость закупки превышает указанную для этой позиции в списке «Роскошь» цену.

Несоответствие статьи бюджетных расходов сути закупки: закупки по неконтрактуемым бюджетным кодам КОСГУ.

Нарушение конкуренции

Признаки нарушения:

Похожий документ заказа уже был использован в других закупках (при этом отягчающее обстоятельство – победитель тех закупок участвует и в данной закупке).

Неадекватно короткий срок исполнения работ (рейтинг по стоимости одного дня).

Единственный участник закупки при высокой стоимости заказа.

Участник с максимальной ценой становится победителем.

Наличие жалоб в ФАС, «Рособоронзаказ» на поставщиков заказа.

Отмененные закупки.

Похожее техническое задание уже было использовано в закупках, при этом победитель закупок с похожим техническим заданием закупок участвует и в данной закупке.

Завышение цен

Признаки нарушения:

Цена единицы товара существенно отличается от цены аналогичных товаров в других закупках.

Коды ОКДП контракта не соответствует кодам ОКДП лота.

Для данной категории потребителей будет представлена информация о неявных и аномальных связях поставщиков и заказчиков, базирующаяся на сведениях из открытых источников (федеральных печатных источников, федеральных телевизионных каналов, региональных СМИ, интернет-СМИ, сайтов органов государственной власти, интернет-блогов и форумов).

Кроме того, ограниченный доступ к системе будут иметь представители широкой общественности – физические лица, сотрудники информационных агентств, блогеры и т.д. Данная группа пользователей проявляет озабоченность состоянием рынка госзакупок и осуществляет опосредованный контроль над прозрачностью и соблюдением законодательства в этой сфере.

Направлениями дальнейших исследований являются совершенствование применяемых технологий, а также распространение предложенных решений для других сфер политической и экономической жизни государства.