Санкт-Петербургский государственный университет

Экономический факультет

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по направлению 080500 – «Бизнес-информатика»

АВТОМАТИЗАЦИЯ ПРОЦЕССА СБОРА ПЕРВИЧНОЙ ИНФОРМАЦИИ В ИНФОРМАЦИОННОМ АГЕНТСТВЕ

Выполнила:

бакалавриант 4 курса, группы БИ-4

Трищ Евгения Геннадьевна

_____/Подпись/

Руководитель:

Доцент, к. ф.-м. н.

Комаров Игорь Иванович

_____/Подпись/

Санкт-Петербург

2016

СОДЕРЖАНИ

ВВЕДЕНИЕ

Настоящая выпускная квалификационная работа является результатом систематизации знаний, навыков, умений, полученных в ходе прохождения практики в

одном из крупных информационных агентств Санкт-Петербурга - информационно-аналитическом агентстве «INFOLine».

Информационные агентства (ИА) - специализированные информационные предприятия, основная функция которых - снабжать оперативной политической, экономической, социальной, культурной информацией редакции газет, журналов, телевидения, радиовещания, а также другие учреждения, организации, частных лиц, являющихся подписчиками на их продукцию. Деятельность ИА охватывает широкий спектр услуг по сбору, созданию, накоплению, обработке и распространению информации. Информация, публикуемая ИА впоследствии используется как средствами массовой информации, так и предприятиями всех отраслей для принятия управленческих решений и, реже, физическими лицами.

Несмотря на то, что первые ИА в привычном нам понимании появились на российском рынке относительно недавно - во время информационного бума 1990-х годов, в настоящее время на российском рынке существуют тысячи организаций, в той или иной степени называющих себя информагентствами, - и все они состоят в жесткой конкуренции друг с другом. Особенно ярко конкуренция выражена среди ИА, работающих в экономической, маркетинговой, бизнес-сфере. С одной стороны, такие ИА являются незаменимым источником информации для организаций, не имеющих возможности содержать собственный штат аналитиков, и, в случае успешности своей деятельности, имеют постоянный приток корпоративных клиентов. Однако с другой стороны, это накладывает на ИА повышенные требования к предоставляемой ими информации: от ее полноты, достоверности, оперативности предоставления напрямую зависит качество управленческих решений, принимаемых организациями. ИА, уступающие своим конкурентам в качестве и оперативности предоставляемых услуг, очень быстро лишаются доверия и теряют клиентов.

Таким образом, информация является для ИА основным продуктом деятельности. Качество финального информационного продукта зависит в первую очередь от первичной информации, собираемой аналитиками. Результаты исследования, проводимого по заведомо неверным или неактуальным данным, являются сомнительными и не могут быть использованы на практике. Именно поэтому каждое успешное ИА уделяет большое внимание процессу сбора первичной информации.

Процесс сбора информации заключается в сборе необходимых сведений, их сортировке, классификации и нахождении взаимосвязей. В рассматриваемом ИА этот процесс целиком производится вручную специалистами-аналитиками, занимая большую часть времени проведения исследования; основными источниками информации являются статистические сборники, средства массовой информации и сеть интернет.

Настоящая ВКР отражает актуальную необходимость оптимизации процесса сбора первичной информации при проведении аналитического исследования, современное состояние которого имеет несколько существенных проблем, вызванных, в первую очередь, низким уровнем автоматизации этого процесса.

Первая проблема заключается в отсутствии у ИА каких-либо средств автоматического сбора первичных данных для решения рутинных задач, повторяющихся от исследования к исследованию и не требующих от исполнителя использования профессиональных навыков аналитика. Ручное выполнение такого сбора данных занимает значительную часть рабочего времени специалистов-аналитиков, которое, во-первых, приводит к излишним затратам на заработную плату, а во-вторых, может быть посвящено более интеллектуально емким задачам.

Вторая и основная проблема заключается в сложности преобразования собранных данных в полезную информацию, которая позволяла бы делать выводы об объекте исследования. Для этого необходимы не только глубокое понимание объекта исследования и специальные аналитические навыки выполняющего эту задачу сотрудника, но и достаточно трудоемкая предварительная работа с самими данными: их сортировка, классификация, нахождение взаимосвязей, при этом алгоритмы работы с данными специфичны для каждой отдельной предметной области и требуют специальной разработки.

Третья проблема связана с особенностями хранения и поиска информации в сети интернет, который является основным источником первичной информации для ИА:

- во-первых, число источников в Сети чрезвычайно велико, и, по оценкам экспертов, только 20% информации, получаемой при поиске, оказывается полезной, остальные 80% составляет “информационный шум”.
- во-вторых, по данным корпорации Google[21], около 30% интернет-документов являются полными или близкими копиями друг друга.
- в-третьих, неструктурированные данные, главным образом текст, составляют не менее 90% информации и лишь 10% приходится на структурированные данные, загружаемые в реляционные СУБД.
- в-четвертых, объем информации в Сети не только огромен по объему, но еще и крайне изменчив: за доли минут в виртуальной сети появляются сотни новых или измененных документов, десятки перемещаются на новые адреса, а единицы - навсегда прекращают свое существование.

Естественным методом решения перечисленных проблем является автоматизация части процесса проведения исследования, связанная со сбором вторичной информации.

Автоматизация процесса сбора вторичной информации позволит:

- существенно сократить время выполнения процесса исследования;

- существенно сократить затраты на оплату труда путем высвобождения человеческих ресурсов;
- снизить нагрузку на сотрудников, освободить время для выполнения более важных, интеллектуально емких задач;
- исключить вероятность человеческого фактора и значительно увеличить массив обрабатываемых данных, тем самым повысив достоверность, актуальность и полноту данных – то есть. повысить качество собираемой информации и, как следствие, качество всего исследования в целом.

Необходимо также отметить, что технические препятствия для проведения автоматизации, ввиду низкого уровня автоматизации процесса, на данный момент отсутствуют либо решаются единожды и на все время функционирования информационной технологии при ее внедрении.

Таким образом, **целью работы** является сокращение временных и стоимостных затрат при проведении исследований в ИА, а также повышение качества этих исследований путем автоматизации процесса сбора вторичной информации.

Для достижения поставленной цели предполагается решение следующих частных задач:

- описание предметной области с краткой характеристикой специфики ИА и проводимых в нем аналитических исследований;
- описание универсальной технологии сбора информации, включая описание источников информации и обобщенного процесса сбора информации;
- разработка технологии информационного обеспечения ИА, включая сравнение существующих технологий, обоснование выбора конкретной технологии и ее адаптация к специфике исследований ИА;
- практическая реализации части технологии, оценка качества ее функционирования.

Объектом исследования является информационно-аналитическое агентство «ИНФОЛайн» и проводимые в ходе осуществления его деятельности аналитические исследования.

Предметом исследования является процесс сбора первичной информации в рамках проведения аналитического исследования в ИА «ИНФОЛайн».

1. ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Об информационном агентстве

Компания «INFOLine» - это информационно-аналитическое агентство, созданное в 1999 году для оказания услуг в сфере B2B. Первой услугой, оказываемой агентством, стала рассылка факс-сообщений с актуальными экономическими показателями в условиях нестабильной экономической ситуации 1999 года. Уже в следующем 2000 году компания запустила уникальные на тот момент “тематические новости” различных сфер бизнеса, а еще через год подготовила базу данных промышленных и торговых компаний Санкт-Петербурга, которая позволила Администрации города оценить состояние бизнес-сектора. В 2004 году компания начала работу в направлении маркетинговых исследований и анализа рынков, в том числе заказных исследований и отраслевых обзоров, а в 2006 запустила направление "PR-поддержка", в рамках которого начала оказывать услуги, направленные на выстраивание корпоративного бренда в СМИ и формирование позитивного информационного поля вокруг компании клиента.

В настоящее врем ИА «INFOLine» является одним из лидеров российского рынка информационно-аналитической поддержки бизнеса, осуществляя на постоянной основе информационную поддержку более 1000 компаний России и мира. ИА «INFOLine» ежедневно проводит мониторинг публикации в более 5000 СМИ и ежедневно ведет аналитическую работу по 80 отраслям реального сектора экономики РФ, самостоятельно и по партнерским программам ежедневно реализует десятки информационных продуктов.

Среди клиентов компании – известные финансовые и сервисные российские и зарубежные компании (такие как ОАО «Газпром», «Альфа-банк», Toshiba Corporation, «PepsiCo», «Mars»), солидные маркетинговые и консалтинговые агентства.

1.2. Характеристика исследований, проводимых ИА

ИА проводит отраслевые и маркетинговые исследования в сфере B2B, как по собственной инициативе, так и по заказу конкретных организаций и предприятий.

Заказные исследований приносят ИА большую часть прибыли, хоть и составляют сравнительно небольшую долю от общего количества проектов. Большей части своих постоянных клиентов ИА предоставляет стандартный комплекс информационных услуг на условиях аутсорсинга (аналитический мониторинг СМИ, конъектурный анализ, формирование баз данных и так далее), однако также выполняет и уникальные проекты в соответствии с индивидуальными требованиями заказчика.

Основную, рутинную часть деятельности ИА составляют инициативные исследования, проводимые ИА самостоятельно (по собственной инициативе), после чего

продаваемые клиентам в виде готового продукта. Такие исследования не только приносят ИА дополнительные стабильный доход, но и играют другую важную роль: именно по результатам инициативных исследований потенциальный клиент может оценить качество работы и исследовательские возможности ИА при поиске исполнителя для своего проекта. К тому же сам факт проведения инициативных исследований положительно характеризует И А как компанию, комплексно и ответственно подходящую к своей работе. Более половины инициативных исследований ИА составляют исследования рынков и отраслевые обзоры по более чем 80 тематикам, а также различные базы и справочники.

Отличительной особенностью таких исследований является то, что они, как правило, являются постоянными и выходят с регулярной периодичностью, как, например, ежегодные обзоры отраслевых рынков или регулярно обновляющиеся базы перспективных проектов в той или иной сфере, публикующиеся ежемесячно. Такие исследования имеют стандартный алгоритм, общий для большей части исследований ИА и существенно не меняющийся с течением времени для конкретного исследования. Более того, хотя формулировка технических заданий организаций-клиентов является коммерческой тайной и не разглашается ИА, можно с уверенностью утверждать, что большая часть заказных исследований также являются типовыми для ИА и алгоритм их проведения фактически не отличается для объектов одной отрасли.

1.3. Обобщенная характеристика процесса проведения исследования

Говоря о процессе проведения исследования в самом общем виде, в литературе, в том числе в учебнике «Бизнес-разведка» А.И. Доронина[5], обычно выделяют следующие его классические этапы:

1. Целеполагание и планирование:
    a. описание и постановка проблемы исследования (определение предмета исследования);
    b. определение конкретной цели и задач проведения исследования;
    c. определение информации, необходимой для достижения этих целей;
    d. определение совокупности возможных источников информационных массивов.
2. Сбор данных:
    a. сбор вторичной информации;
    b. сбор первичной информации.
3. Обработка данных - превращение их в информацию:
    a. систематизация - группировка информации по общности освещаемых вопросов, регистрация источников и время ее получения;

b. верификация - оценка в отношении надежности источника их получения, времени, прошедшего с того момента, когда они были получены, их достоверности и точности по сравнению с уже имеющейся информацией по данному вопросу;

4. Отбор релевантной информации.

5. Анализ и синтез информации превращение ее в знания.

6. Описание результатов исследования, формулировка выводов и рекомендации.

7. Распространение — получение знаний конечным потребителем.

Обобщенный процесс проведения исследования в ИА представлен на рис.1.

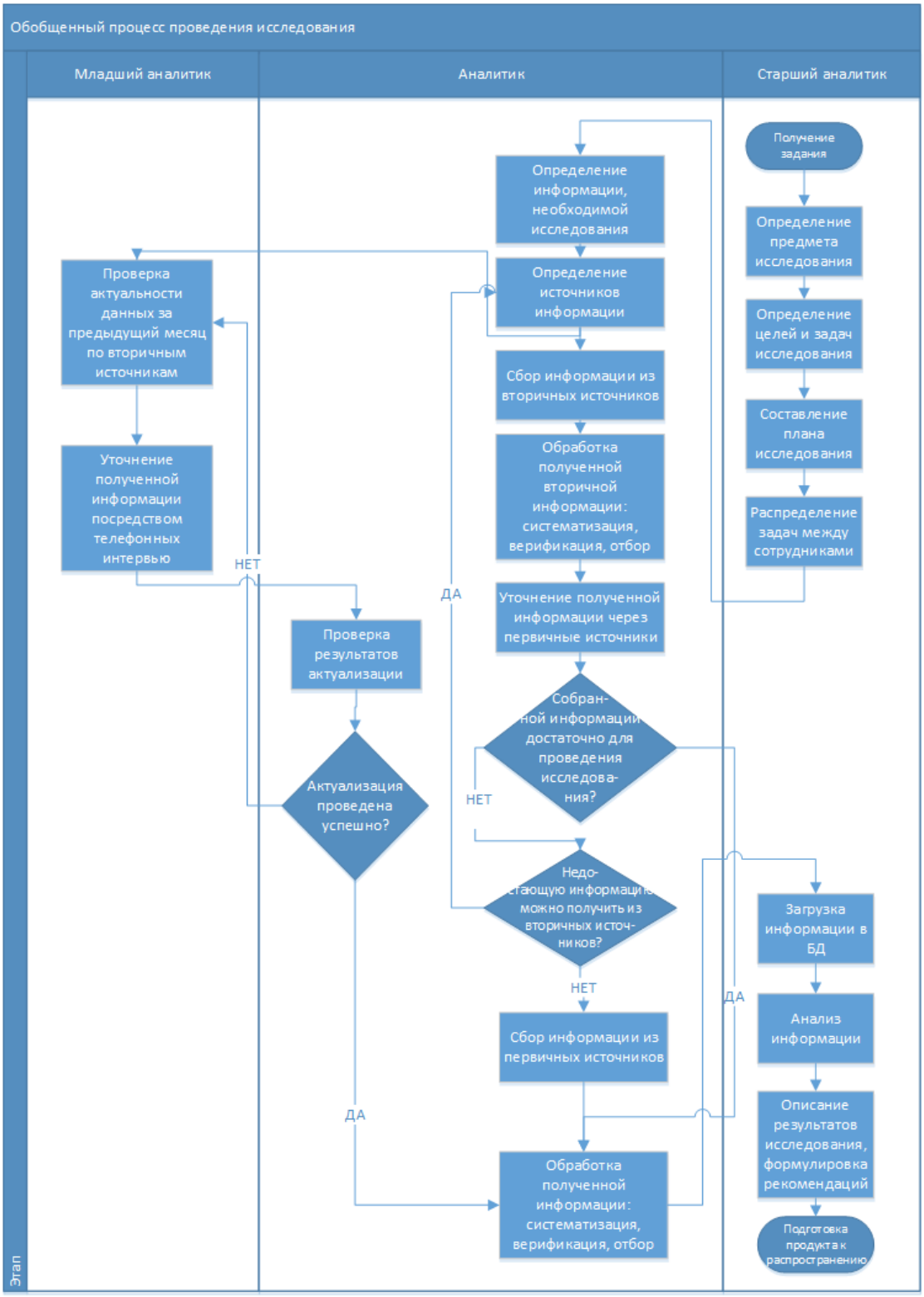


Рисунок 1 - Обобщенный процесс проведения исследования

2. УНИВЕРСАЛЬНАЯ ТЕХНОЛОГИЯ СБОРА ИНФОРМАЦИИ

2.1 Контекст сбора информации

Одной из ключевых особенностей проводимых ИА исследований является тот факт, что большая их часть являются типовыми, при этом часто - регулярными, и имеют единый стандартный алгоритм проведения, мало изменяющийся во времени.

Другой важной особенностью проводимых ИА исследований, помимо их регулярности, необходимо выделить тот факт, что практически вся информация, для них необходимая, уже существует в том или ином виде в общедоступных источниках, таких как публикации СМИ, статистические сборники, финансовая отчетность компаний, решения судов и так далее (так называемая первичная маркетинговая информация).

При этом, после того, как вся возможная информация, хранящаяся в открытых источниках, собрана и обработана, она обязательно проверяется на правильность по первичным источникам, то есть. напрямую у сотрудников исследуемых организаций или у экспертов посредством интервью. Таким образом, процесс поиска информации фактически дублируется и, соответственно, занимает в 2 раза больше времени.

Получение новой информации напрямую из первичных источников происходит гораздо реже, только в случаях, когда поиск в открытых источниках оказывается бесполезен либо неприменим изначально.

При этом как во время поиска информации в существующих источниках, так и при проведении интервью с сотрудником организации, чрезвычайно важно соблюдать следующие общепринятые требования к информации:

1. Достоверность. Это свойство означает, что информация должна правдиво, без искажений отражать состояние объекта исследования. Достоверность и точность информации определяет правильность сделанных выводов, а также эффективность принимаемых на их основе решений и рекомендаций.
2. Актуальность. Она отражает степень новизны информации и ее своевременность. Устаревшая информация не представляет ценности, поэтому разрыв времени от момента получения информации до ее использования должен быть минимальным.
3. Полнота. Это свойство указывает на то, что содержание информации должно обеспечивать необходимые и достаточные показатели для достижения цели исследования.
4. Релевантность. Один из главных терминов в сфере поиска, который определяет степень соответствия запроса полученным результатам.
5. Сопоставимость или иначе сравнимость - предполагает возможность сравнения данных предмета исследования и круга ключевых показателей. Сопоставимость достигается на основе единой методологии проведения исследования и способов измерения характеристик.
6. Доступность. Это свойство означает, что информация должна быть понятна для того, кому она предназначена, и представлена на удобном для него носителе.
7. Экономичность. Это свойство предполагает, что затраты на получение и обработку информации не должны превышать получаемый эффект от ее использования.

Соблюдение всех вышеперечисленных требований строго обязательно, нарушение хотя бы одного из них ставит под сомнение результаты всего исследования и делает их неприменимыми на практике для принятия управленческих решений.

Однако необходимо учитывать особую сложность соблюдения этих требований при поиске вторичной информации, обуславливаемую несколькими особенностями сложившейся в России экономической ситуации, сформулированных Б.Е. Токаревым[11]:

- отсутствие достоверной и точной информации по многим вопросам и, следовательно, необходимость своевременно ее проверять и актуализировать;
- большое количество источников информации и полное отсутствие ее структуры;
- отсутствие открытого доступа к информации как официальной, так и коммерческой;
- преобладание в анализе рыночной ситуации качественных методов над количественными.

Именно поэтому этап сбора информации по времени составляет вплоть до 90% всего процесса проведения исследования.

2.2 Характеристика источников информации

2.2.1 Источники первичной информации

Главным источником первичной информации являются непосредственно сотрудники организаций той отрасли, в которой проводится исследование, а также эксперты в различных областях. Основным методом получения такой информации является телефонное либо, в редких случаях, личное интервью.

Кроме того, ИА тесно сотрудничает с крупными торговыми организациями, что позволяет более оперативно получать информацию в ответ на интересующие вопросы. Взаимодействие с такими организациями обычно ведется посредством запросов по электронной почте.

2.2.1 Источники вторичной информации

Отбор релевантных источников вторичной информации требует от работника, осуществляющего его, глубоких знаний тематики проводимого исследования, а также навыков информационно-поисковой работы.

В некотором смысле этап отбора источников информации - более требовательный с точки зрения профессиональных качеств исполнителя, чем этап самого поиска, который, хоть и требует от специалиста понимания предмета исследования и навыков быстрого

поиска достоверной и релевантной информации, сам процесс поиска является рутиной, повторяющейся от исследования.

Сбор информации осуществляется из следующих источников:

- сайты организаций в сети интернет. Являются одним из основных источников информации о различных компаниях: контактные данные, масштаб, высший менеджмент, последние новости, а иногда и финансовое положение. Именно эта информация - основа для многочисленных баз данных, выпускаемых ИА в качестве готовых продуктов, а также при отраслевом анализе;
- средства массовой информации, в первую очередь региональные. Из них собирается информация об экономических процессах, происходящих в стране и отдельных отраслях, о деятельности крупных организаций и происходящих в них изменениях;
- официальная информация, публикуемая общественными организациями. В нее входят: отчеты, балансы, финансовые результаты деятельности;
- источники профессиональной информации о компаниях и персоналиях;
- специальные издания, в основном посвященные тематике экономики и маркетинга;
- государственная и отраслевая статистика;
- публикации учебных, научно-исследовательских, проектных институтов и общественно-научных организаций, симпозиумов, конгрессов, конференций;
- базы данных;
- интернет;

2.3. Описание процесса сбора информации

В крупных ИА процессом сбора информации для одного исследования занимаются одновременно несколько человек, распределяя ее между собой в зависимости от источника и требуемого опыта и профессиональных навыков. При этом рутинным мониторингом и, в случае регулярного исследования, актуализацией данных за прошлый период занимаются сотрудники низших уровней - стажеры и младшие аналитики. Вся собранная информация передается более опытному специалисту, который после чего проводит ее анализ и обрабатывает результаты.

Перед тем, как начинать непосредственно сбор данных, важно четко сформулировать объект исследования, что о нем уже известно и какую информацию нужно получить, после чего составить список ключевых терминов, названий, имен и отобрать только те источники, поиск по которым будет наиболее эффективным. Это позволяет ограничить объем прорабатываемой информации, тем самым сокращая затрачиваемое время и повышая качество собираемых данных. После обработки собранных данных, их систематизации, проверки на релевантность и очистки от всего

лишнего, необходимо пересмотреть параметры искомой информации и пул релевантных источников и повторить процесс сбора. Цикл повторяется до тех пор, пока не будет найдена вся необходимая информация или же пока дальнейший поиск информации перестанет быть возможным либо экономически оправданным.

Обобщенный алгоритм сбора информации представлен на рис. 2.

2.4. Модель хранения информации в ИА

Обобщенная модель хранения данных в ИА представлена на рис. 3.

Собранная из различных источников, обработанная и очищенная первичная информация помещается в единое хранилище данных, откуда потом выгружается сотрудниками в нужном объеме в виде файлов Microsoft Excel для проведения анализа или составления отчетов.

При этом необходимо отметить, что технологии хранения данных являются коммерческой тайной ИА и потому не подлежат разглашению и какому-либо внешнему анализу.

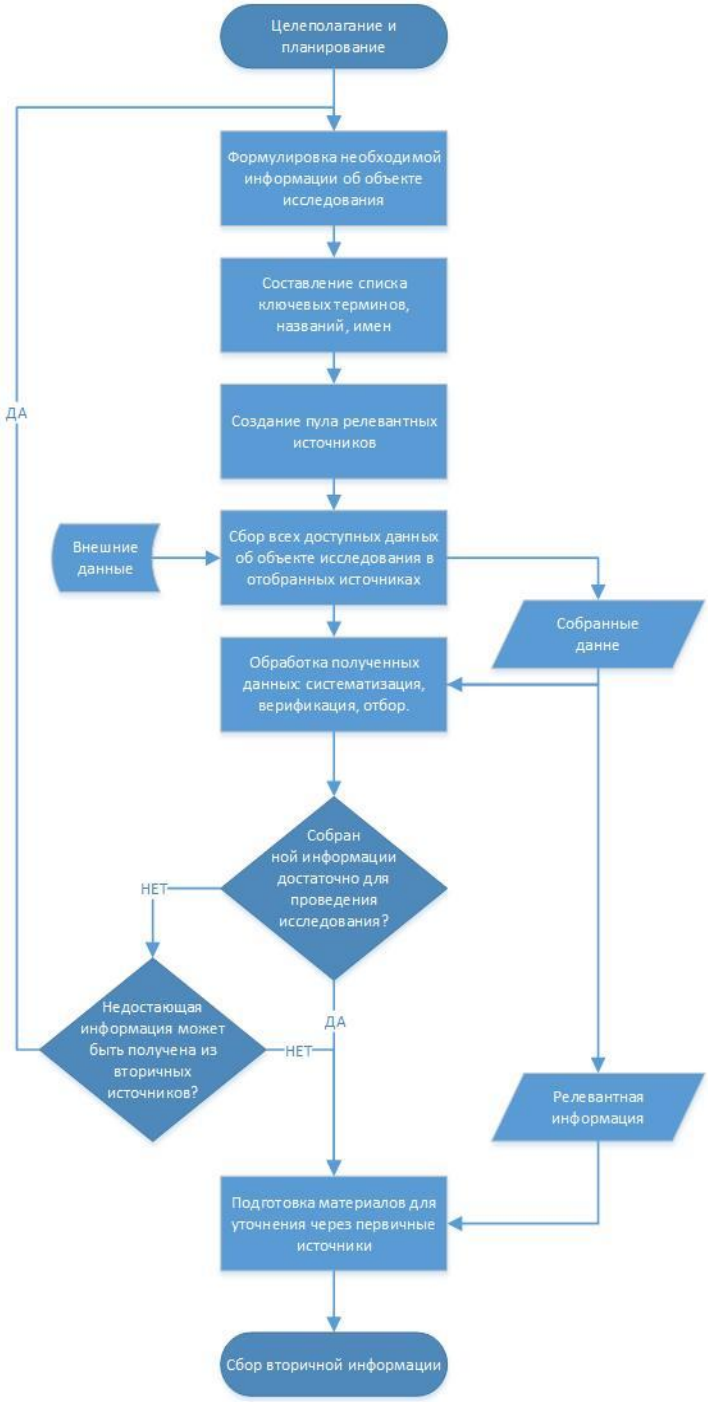


Рисунок 2 - Обобщенный алгоритм сбора информации

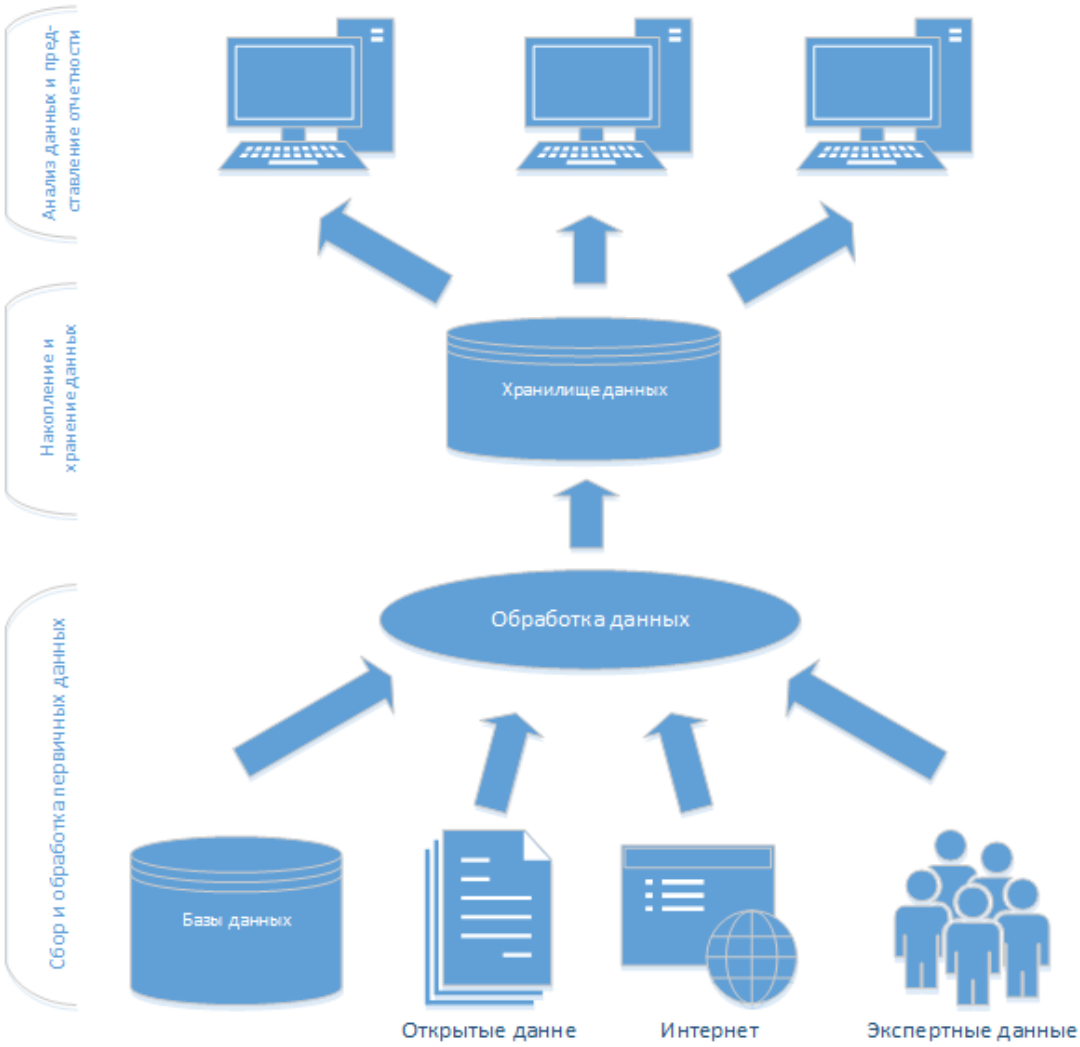


Рисунок 3 - Обобщенная модель хранения данных в ИА

3. РАЗРАБОТКА ТЕХНОЛОГИИ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ ИА

3.1. Выявление проблем информационного обеспечения ИА

Таким образом, в сложившемся на данный момент процессе проведения исследования в ИА можно выделить следующие недостатки:

- большую часть времени проведения исследования занимает сбор информации по вторичным источникам, который затрудняется большим объемом неструктурированной, противоречивой информации;
- почти вся найденная вторичная информация проверяется по первичным источникам, что, с одной стороны, значительно повышает ее достоверность, но с другой - почти в два раза увеличивает время проведения исследования;
- процесс сбора информации является рутинным и требует не столько глубокого понимания предмета исследования, сколько хороших навыков работы с источниками информации. Несмотря на это, выполнение данной задачи нельзя доверить неквалифицированному работнику, а поручать ее опытному специалисту, какие занимаются проведением исследований в ИА, представляется нецелесообразным с точки зрения затрачиваемых средств на оплату труда.

При этом необходимо отметить, что большая часть исследований ИА является типовыми и регулярными и при их проведении аналитики следуют стандартному выработанному алгоритму работы, мало подверженному изменениям во времени и единому для всех отраслей и направлений.

3.2. Способ повышения качества информационного обеспечения ИА

Естественным методом устранения вышеперечисленных недостатков является автоматизация части процесса проведения исследования, связанная со сбором вторичной информации.

Автоматизация процесса сбора вторичной информации позволит:

- существенно сократить время выполнения процесса исследования;
- существенно сократить затраты на оплату труда сотрудников;
- снизить нагрузку на сотрудников, освободить время для выполнения более важных, интеллектуально емких задач;
- исключить вероятность человеческого фактора и значительно увеличить массив обрабатываемых данных, тем самым повысив достоверность, актуальность и полноту данных - то есть повысить качество собираемой информации и, как следствие, качество всего исследования в целом.

Таким образом, **цель** данной работы является сокращение временных и стоимостных затрат при проведении исследований в ИА, а также повышение качества этих исследований путем автоматизации процесса сбора вторичной информации.

3.3. Обзор существующих программных продуктов автоматизации поиска информации

В настоящее время на рынке существует множество программных продуктов, направленных на автоматизацию процесса сбора информации. В зависимости от круга решаемых информационно-аналитических задач, они условно подразделяются на системы поиска информации и системы анализа информации.

Основной функцией информационно-поисковых систем является оперативный сбор и накопление релевантной потребностям пользователя информации из различных документов в соответствии с заданными правилами поиска. При этом поиск информации может проводиться как единовременно, так и автоматически в режиме мониторинга и с использованием специальных сценариев. Результатом поиска обычно является отсортированный по дате или релевантности список ссылок на оригинальные документы, сопровождаемые краткими характеристиками и аннотациями. Профессиональные информационно-поисковые системы предоставляют дополнительные инструменты для создания сложных и детализированных правил поиска информации и тонкой настройки под нужды конкретного пользователя.

Лидерами рынка профессиональных информационно-поисковых систем, по заявлениям практикующих специалистов конкурентной разведки, являются: система интернет-мониторинга и конкурентной разведки Avalanche, система поиска информации в интернете FileForFiles & SiteSputnik и система автоматического мониторинга веб-сайтов WebSite Watcher. Также следует упомянуть систему аналитического мониторинга российских СМИ “Медиалогия”, программу анализа рисков при выборе контрагентов X-Files.

Системы анализа текстовой информации рассматривают тексты документов как структурированные последовательности основных терминов и их связей с весовыми характеристиками, что позволяет использовать их при составлении списка ключевых слов, построении рефератов, отборе похожих и связанных документов, автоматической классификации и кластеризации документов, выявлении заимствованных фрагментов, а также при аннотировании, определении тональности текста по отношению к исследуемому объекту.

Из множества существующих на рынке систем стоит выделить систему интеллектуального анализа текста на естественном языке “Аналитический курьер”,

систему анализа текстовой информации **RCO Fact Extractor** SDK, систему анализа текстовых и структурированных данных **PolyAnalyst.**

3.4. Сравнение и обоснование выбора ИС

В ходе изучения процесса сбора информации в ИА, а также существующих в настоящий момент решений по автоматизации подобных процессов, были выделены следующие специфические требования к программному продукту.

Обязательные функциональные требования к системе:

- поиск информации в сети интернет и мониторинг изменений в поисковой выдаче;
- поиск информации внутри указанного веб-сайта и мониторинг изменений контента на указанных веб-сайтах;
- мониторинг новостных потоков, их рубрикация, составление дайджестов;
- гибкая настройка правил, по которым осуществляется поиск и мониторинг информации;
- возможность добавления собственных источников информации;
- ведение архива/базы данных информации;
- поддерживаемые языки источников: русский, английский.
- Желательные требования к функционалу системы:
- поиск информации о юридических лицах и мониторинг изменений информации о юридических лицах;
- поиск информации в невидимом интернете;
- возможность анализа информации;
- ведение досье для объектов мониторинга.

Кроме того, система не должна иметь большого количества излишнего дорогостоящего функционала, предназначенного для полного обеспечения всех процессов маркетинга и конкурентной разведки крупного предприятия.

Таким образом, в соответствии с требованиями, предъявляемыми к ИО, были более подробно рассмотрены представленные на рынке информационно-поисковые системы, изучен предоставляемый ими функционал, после чего были выбраны три основные системы, считающиеся лидерами в своих сегментах, и проведен сравнительный анализ выбранных систем.

3.4.1. Avalanche

Общее описание системы:

Система интернет-мониторинга Avalanche предназначена для отслеживания изменений, происходящих на веб-сайтах. Она производит сбор информации с интернет-страниц по заданному алгоритму, после чего информация помещается в собственную базу данных. Поиск информации пользователем происходит непосредственно из этой базы с

помощью операторов Булевой Алгебры, подобных тем, которые используются при формировании поискового запроса в Яндексе или Гугле.

Первоначальная версия программы была разработана компанией Андрея Масаловича в 2001 году по заказу Гарвардского Университета. В настоящий момент работает версия Avalanche 2.7, которая, согласно мнению ряда экспертов, является лучшей программой для мониторинга интернета.

Функциональная часть:

По заявлению самого разработчика программы, “технология Avalanche базируется на трех "китах": концепции "умных папок" (Smart Folders), автономном интеллектуальном поисковом роботе ("пауке") и встроенной базе данных, допускающей преобразование в "персональную энциклопедию". “Паук” (поисковый робот) осуществляет поиск в интернете по заранее прописанному алгоритму по заранее заданному расписанию и собирает информацию в единую базу данных. “Умные папки” - рубрикатор с расширенными возможностями, который самостоятельно сортирует принесенную поисковым роботом информацию и отображает ее в удобном для работы с ней виде. При этом поиск проходит как по “видимому”, так и по “невидимому” интернету, тем самым обеспечивая больший объём найденной информации по сравнению с обычными поисковыми системами.

Отличительные возможности программы:

- Различные типы поисковых роботов, позволяющие собирать информацию с различных интернет-ресурсов, сайтов, RSS-лент, социальных сетей и массовых сервисов.
- Отслеживание и удаление дубликатов из результатов поиска.
- Тонкие индивидуальные настройки алгоритмов поиска для поискового робота;
- Автоматический мониторинг по расписанию без участия человека.
- Средства генерации новостных лент и отчетов по результатам мониторинга.

Недостатки:

- Требует проведения больших объемов работы по настройке поисковых алгоритмов, поскольку, как правило, каждый источник информации требует индивидуальной настройки.
- В связи с тем, что сайты-источники динамичны и регулярно меняют свою структуру, требуется регулярная частая перенастройка поисковых алгоритмов.
- От специалиста, проводящего настройку, требуется, помимо знаний самого продукта, еще как минимум отличное знание HTML.

Стоимость и предоставляемые услуги:

Стоимость базового коммерческого решения начинается от 750 000 руб и быстро растет при адаптации системы под нужды заказчика.

Базовое решение включает в себя:

- Бессрочную лицензию на 5 пользователей клиентской части Avalanche 2.99.

- Начальную настройку.
- Обучение специалистов заказчика работе с системой.
- Поддержку и сопровождение в течение одного года.

Интегрированное решение по сравнению с базовым включает в себя средства управления расширяемым кластером на основе технологии Hadoop, объединяющим сервера Avalanche и Лавина Пульс, а также систему управления распределенной базой данных Elastic Search со встроенными средствами обеспечения отказоустойчивости. Интегрированное решение допускает «бесшовное» расширение до 50 серверов в кластере. Возможна доработка и расширение функционала по отдельным техническим заданиям заказчика.

3.4.2. FileForFiles SiteSputnik

Общее описание системы:

Программа FileForFiles SiteSputnik (СайтСпутник) предназначена для поиска, сбора, мониторинга и анализа информации, размещенной в “видимом” и “невидимом” интернете.

Программа разрабатывается с 2003 года программистом Алексеем Мыльниковым при поддержке Сообщества Практиков Конкурентной разведки и, имея такие неофициальные названия как “Программа для допроса интернета” и “Швейцарский ножик для поиска в интернете”, по признанию многих экспертов не имеет аналогов в глубоком поиске и мониторинге информации.

Функциональная часть:

Программа SiteSputnik носит модульный характер и базируется на базовом модуле “Pro”, включающем в себя поиск, сбор, рубрикацию и мониторинг информации. Поиск происходит на основе наборов правил и источников, задающихся пользователем. Отличительной особенностью программы является то, что она позволяет вести поиск по нескольким правилам, объединенным в пакеты, используя при этом до 11 поисковых систем одновременно, после чего применяется анализ и объединение результатов в единую выдачу, удаление дубликатов и сортировка по релевантности запросу.

Другие модули программы включают в себя:

- Objects - сбор информации о физических и юридических лицах по набору реквизитов, таких как: наименование, телефон или ИНН; установление связей между ними, поиск негативной информации для проверки надежности лица.
- News - осуществляет мониторинг информационных потоков: СМИ, RSS-лент, социальных сетей и других сайтов.
- Comments - осущетсвляет мониторинг комментариев к новостям, сообщениям, публикациям.

- WebSpider - мониторинг уже существующих страниц на предмет появления новой информации, соответствующей запросу.
- NewStreams - мониторинг сети интернет на предмет появления новых источников, потенциально полезных пользователю.
- Local - мониторинг файлов и папок локального компьютера и локальной сети.
- Station - позволяет создавать корпоративную сеть для организации круглосуточного мониторинга и коллективной обработки информации.
- Invisible - осуществляет базовый поиск информации в невидимом интернете.
- Contacts - по списку названий организаций собирает в сети интернет контактную информацию каждого из них, а именно: адрес, телефон, факс, e-mail и оформляет ее в виде таблицы.

Отличительные возможности:

- Метапоиск сразу в нескольких поисковых системах: Яндекс, Google, Yahoo, Рамблер, MSN (Bing), Mail, Апорт, сервисы поиска по блогам Яндекса и Google.
- Пакетный поиск с возможностью объединения и разделения результатов поиска.
- Аналитическое объединение, фактическая релевантность.
- Библиотека запросов и пакетов запросов, пакеты запросов с изменяемыми параметрами.
- Материализация видимого интернета: построение карт сайтов и объектов из найденных ссылок.

Недостатки:

- Непрезентабельный, достаточно сложный интерфейс.
- Наличие в общем доступе только базовых инструкций по работе с многочисленным функционалом системы, что затрудняет самостоятельное изучение программы и требует по крайней мере первичных консультаций со специалистами.

Стоимость и предоставляемые услуги:

Прайс-лист системы представлен на ее официальном сайте. Стоимость базового модуля Pro составляет 5500 руб. за 1 лицензию с дальнейшим усложнением комплектации вплоть до приблизительно 200 000 руб. за 1 лицензию. Оплата лицензии одноразовая, переходы на последующие версии бесплатны.

3.4.3. WebSite Watcher

Общее описание системы:

Разработанная в Германии программа, осуществляющая мониторинг веб-сайтов на предмет появления новой информации. Вся обнаруженная новая информация выделяется цветом, а веб-страницы сохраняются. Поддерживает большое количество настраиваемых фильтров, предназначенных для более тонкой точечной настройки контроля за изменениями определенной информации и исключения неинформативных сообщений об обновлении страницы.

Функциональная часть:

- Программа позволяет осуществлять мониторинг веб-страниц всех типов, независимо от расширения файла. Все изменения программа выделяет цветом, в зависимости от настроек пользователя: либо измененный фрагмент полностью, либо только ключевые слова.
- Мониторинг RSS-лент. Страницы RSS преобразуются в текстовый формат и позволяет работать с ними, как с обычными веб-страницами.
- Мониторинг веб-страниц, защищенных паролем. Для этого требуется единожды написать макрокоманду, которая в дальнейшем выполняется автоматически и страница проверяется на обновление.
- Отслеживание создания новых и обновления уже имеющихся тем на форумах. Сами форумы при этом преобразуются в обыкновенные текстовые страницы. Поддерживает большую часть популярных форумных движков.
- Мониторинг страниц с Javascript.
- Мониторинг изменений изображений на веб-сайтах, однако изменения в самих изображениях не выделяются.
- Отслеживание изменений в двоичных файлах, таких как файлы zip или exe.
- Мониторинг документов. Файлы pdf, word, excel автоматически преобразовываются в html-формат и в дальнейшем обрабатываются как текстовые файлы.
- Мониторинг любых локальных файлов, хранящихся на жестком диске или в локальной сети.
- Архивирование веб-страниц в самом WebSite Watcher не реализовано, однако разработчик предоставляет программу Local WebSite Archive, которая и осуществляет архивирование.

Стоимость и предоставляемые услуги:

При приобретении корпоративной версии на более чем 10 пользователей стоимость одной лицензии составляет 69 евро.

3.4.4. Сравнительная характеристика программных продуктов

Сравнительная характеристика программных продуктов представлена в табл. 1.

SiteSputnik имеет несколько характеристик, выгодно отличающих его от своих конкурентов, таких как:

- уникальный функционал: глубокий поиск и метапоиск по нескольким поисковым системам, а также возможность специфического поиска информации о юридических и физических лицах;
- средства аналитической обработки информации, семантический анализ;
- отсутствие излишнего дорогостоящего функционала;
- приемлемая цена.

Таблица 1 - Сравнительная характеристика программных продуктов

Параметр	Avalanche	SiteSputnik	WebSite Watcher
Язык интерфейса	русский	русский	русский
Возможность совместной работы нескольких пользователей	да	да	нет
Интеграция с ИС и СУБД заказчика	Универсальный интерфейс	Универсальный интерфейс	Универсальный интерфейс
Поиск по страницам выдачи поисковых машин	Да	Да	Да
Глубокий поиск информациии в сети интернет	Нет	Да	Нет
Поиск одновременно по нескольким источникам	Нет	Да	Нет
Метапоиск по нескольким запросам	Нет	Да	Нет
Аналитическое объединение результатов поиска	Нет	Да	Да
Мониторинг изменений в поисковой выдаче	Да	Да	Да
Поиск информации о физических и юридических лицах	Нет	Да	Нет
Поиск контактов	Нет	Да	Нет
Мониторинг новостных потоков	Да	Да	Да
Мониторинг изменений на сайтах	Да	Да	Да
Формирование, объединение новостных потоков, вычисление связей	Да	Да	Нет
Рубрикация, составление дайджестов	Да	Да	Да
Средства анализа информации	Да	Да	Нет
Графическая реализация результатов анализа	Да	Нет	Нет
Поддерживаемые источники новостных потоков	Сайты, RSS-потоки, блоги, форумы, социальные сети, FTP-сервера	Сайты, RSS-потоки, блоги, форумы, социальные сети, комментарии	Сайты, RSS-потоки, форумы, защищенные паролем страницы, файлы pdf, excel, word
Доступное количество источников	около 7000		
Возможность расширения списка источников	Да	Да	Да
Гибкая настройка правил и алгоритмов поиска и мониторинга	Да	Да	Да
Требуемые специальные знания при создании правил и алгоритмов	HTML	Языки поисковых	Встроенный язык

		запросов	сценариев
Ведение досье для объектов мониторинга	Да	Нет	Нет
Оповещение заказчика о важных событиях	да	Да	Да
Работа с “невидимым” интернетом	Да, одна из лучших на рынке	Да	Нет
Обучение программы	Да	Нет	Нет
Защищенность канала передачи данных	Да	Нет	Нет
Анонимность работы	Да	Нет	Нет
Архивирование	Да	Да	Да, отдельной программой
База данных	Да	Да	Да
Клиентские папки	Да	Да, менее проработанные	Нет
Дружелюбность интерфейса	5	3	4
Полнота информации на веб-сайте	3	5	4
Полнота предоставляемых инструкций по работе с программой	5	4	5
Стоимость	1 000+ тыс.руб. /5 лицензий	~150 тыс.руб. /лицензия	69 евро /лицензия

4. РЕАЛИЗАЦИЯ ТЕХНОЛОГИИ

Рассмотрим структуру типичного отраслевого исследования, выпускаемого ИА. Оно состоит из следующих трех частей:

1. Описание состояния отраслевого рынка: макроэкономические показатели, динамика развития, актуальные тенденции.
2. Описание крупнейших участников рынка, рейтинг участников по ряду показателей.
3. База данных участников рынка, включающая в себя контактную информацию, мощности, операционные и финансовые показатели, региональную представленность, ассортимент продукции и прочие показатели, индивидуальные для отрасли.

При этом для проведения такого исследования используются следующие источники информации:

- Основным источником данных для анализа состояния отраслевых рынков является Федеральная служба государственной статистики. Данные выгружаются напрямую с официального сайта ФСГС и не требует существенных усилий при поиске и подготовке к анализу.
- Основным источником информации об организациях и инвестиционных проектах является официальный сайт компании. В случае отсутствия информации на сайте, осуществляется ее поиск в других источниках: сети интернет, средствах массовой информации, каталогах тендеров, справочниках организаций, архивах судебных решений и др.

Одним из крупнейших инициативных исследований, регулярно проводимых ИА, является “Аналитическая база “700 торговых сетей FMCG России”. Расширенная версия”. Это исследование также является типичным для ИА - оно имеет стандартную структуру, характерную для большей части аналитических исследований, проходящих в ИА. Именно поэтому мы будем рассматривать реализацию технологии на его примере.

Как указано на официальном веб-сайте ИА[14], в состав исследования “Аналитическая база “700 торговых сетей FMCG России” входят:

1. Состояние рынка розничной торговли (около 55 страниц);
2. Рейтинг ТОП-100 ритейлеров FMCG (около 30 страниц);
3. Развитие сетей FMCG по форматам (около 40 страниц);
4. База данных "700 сетей и 550 складов сетей FMCG России" (около 190 страниц);
5. Бизнес-справки по ТОП-100 ритейлерам FMCG России (около 560 страниц);

Исследование начинается со сбора информации для раздела 4 База данных "700 сетей и 550 складов сетей FMCG России". На основе этой информации и данных из Федеральной службы государственной статистики создаются первые три раздела. Для создания последнего раздела бизнес-справок по ТОП-100 ритейлерам FMCG России

используются данные четвертого раздела, а также дополнительно мониторинг СМИ, ежедневно проводимый специалистами ИА.

База данных "700 сетей и 550 складов сетей FMCG России" включает в себя следующие строки:

1. Бренд сети
2. Юридическое название
3. Менеджмент сети:
   a. Генеральный директор
   b. Финансовый директор
   c. Директор по закупкам
   d. Директор по IT
   e. Директор по развитию
   f. Директор по логистике
4. Фактический адрес
5. Телефон
6. Факс
7. E-mail
8. Web-сайт
9. Интернет-магазин
10. Общее количество магазинов сети
11. Количество магазинов по форматам
12. Общая торговая площадь магазинов сети
13. Чистая выручка (без учета НДС) торговой сети в 2014-2015 гг., млрд. руб.
14. Региональная представленность (в каких регионах и городах размещены магазины сети с указанием их количества)
15. Количество РЦ/складов
16. Общая площадь РЦ/складов
17. Регионы присутствия РЦ/складов на

4.1. Процесс сбора информации: как есть.

Процесс сбора информации начинается с просмотра веб-сайта компании: большая часть организаций указывают на своем сайте бренды, под которыми проходят их магазины, фактический адрес, телефон, e-mail и количество либо список магазинов. Реже - юридическое название и имя генерального директора. Внутреннюю же информацию, такую как имена высших менеджеров или площадь имеющихся у компании складов, можно найти упомянутой в публикациях СМИ.

Схема процесса сбора информации представлен на рис. 4.

Подробный процесс сбора информации с указанием ее источников в случае, если у компании имеется веб-сайт, представлен на рис. 5 и рис. 6.

Подробный процесс сбора информации с указанием ее источников в случае, если у компании не имеется веб-сайта, представлен на рис 7.

4.2. Процесс сбора информации: как должно быть.

На рис. 8 представлена схема процесса сбора информации после внедрения системы SiteSputnik.

На схеме видно, что после внедрения ИС, весь процесс поиска информации осуществляется системой и при этом делится на четыре этапа:

1. Поиск информации о банкротстве компании или ее закрытии по какой-либо другой причине (в случае подозрения аналитиком ее неактивности в данный момент).
2. Поиск информации по сайту компании.
3. Поиск недостающей информации в иных источниках.
4. Поиск финансовой информации о компании в отчетах компании, в СМИ, а также в специализированных справочниках.

В стандартном случае аналитик редактирует готовые поисковые запросы, создаваемые при внедрении ИС для каждой отдельной группы сотрудников, в случае необходимости внося в запрос изменения или создавая дополнительные запросы.

4.3. Инструментарий SiteSputnik, используемый при поиске информации

4.3.1. Для компаний, имеющих веб-сайт.

1. Простой пакетный поиск. Реализует выполнение произвольного количества запросов одновременно с последующим аналитическим объединением или разделением результатов поиска.

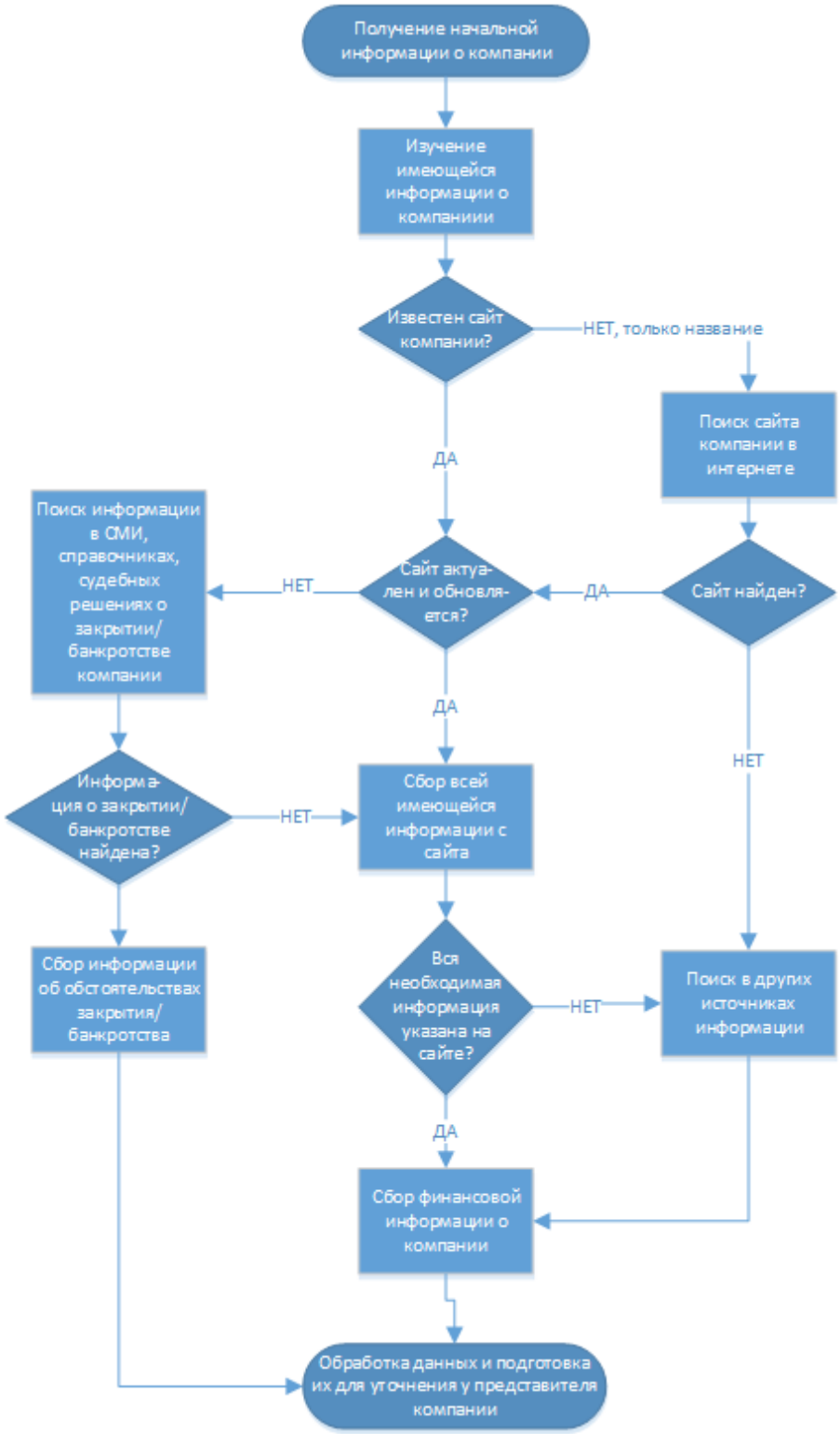


Рисунок 4 - Процесс сбора информации в ИА: как есть

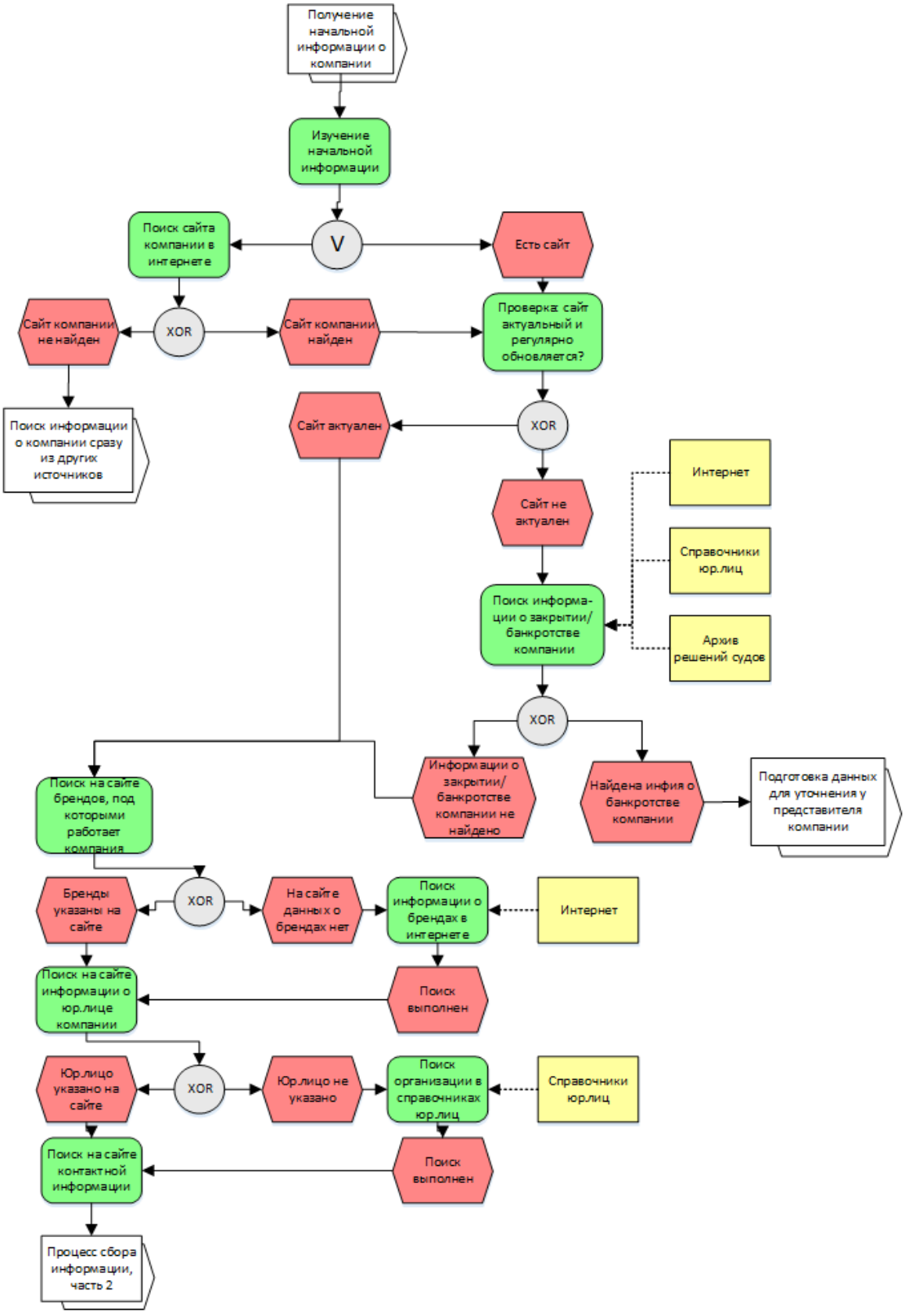


Рисунок 5 - Процесс сбора информации в случае, если у компании имеется веб-сайт

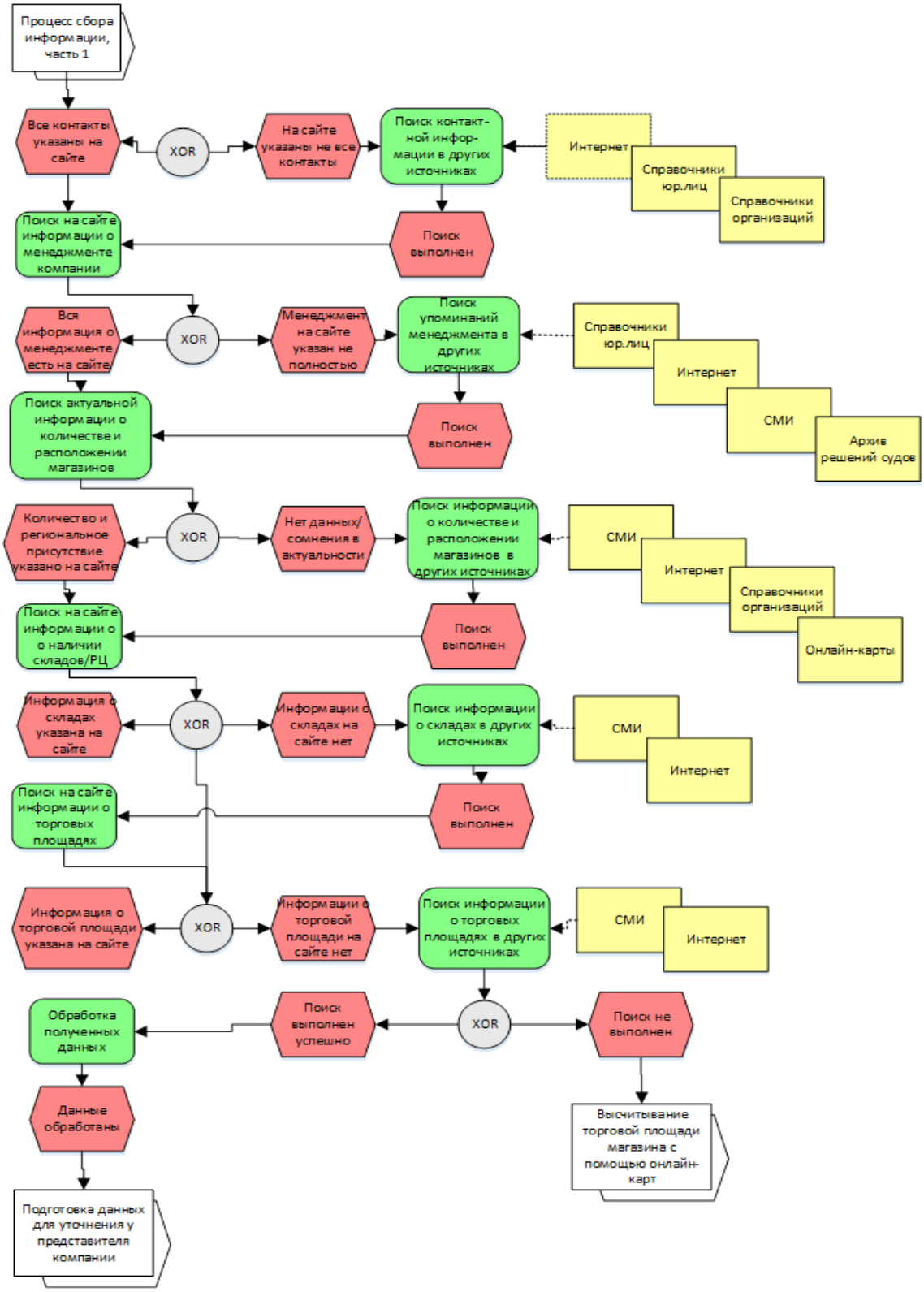


Рисунок 6 - Процесс сбора информации в случае, если у компании имеется веб-сайт - продолжение

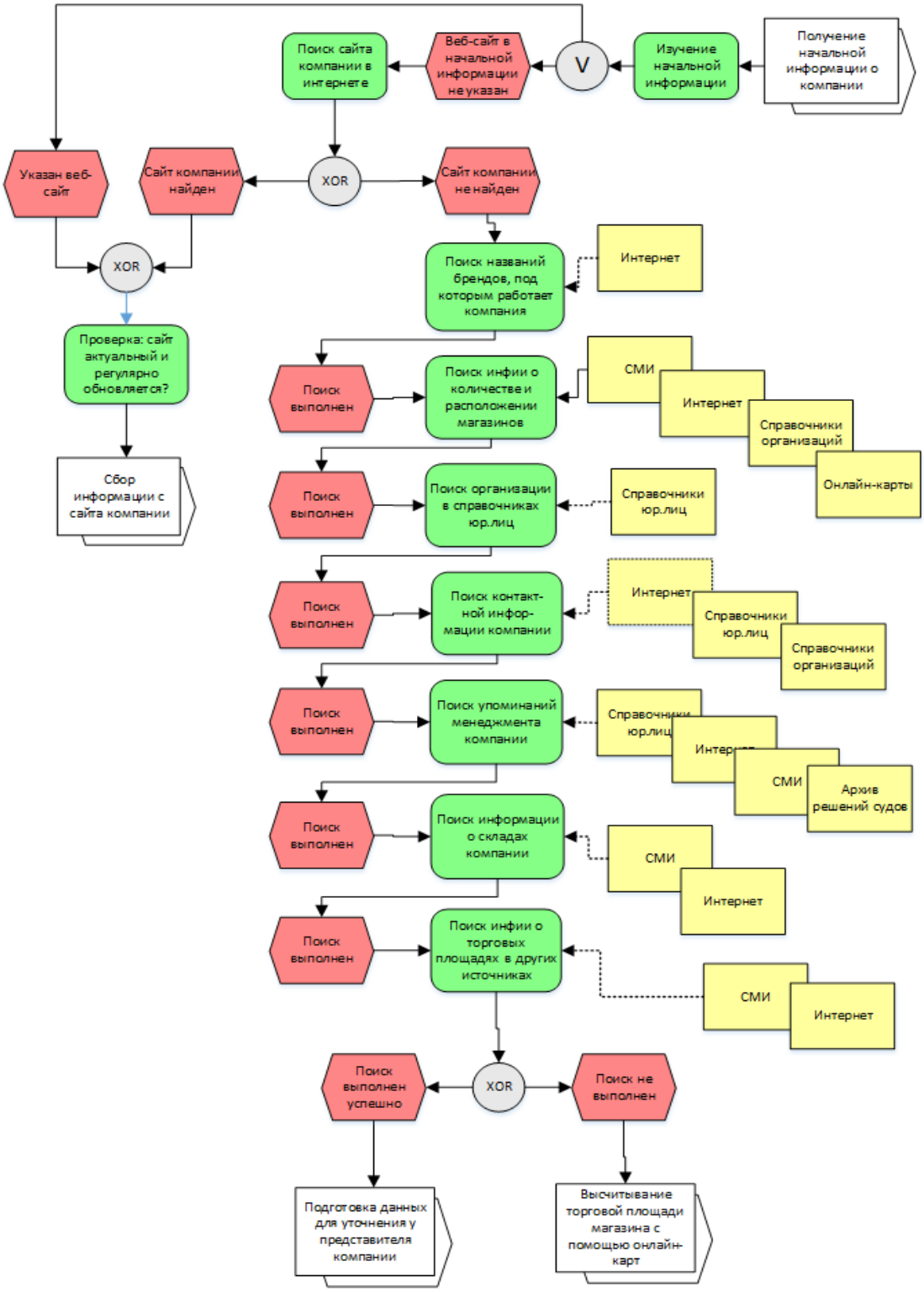


Рисунок 7 - Процесс сбора информации в случае, если у компании не имеется веб-сайта

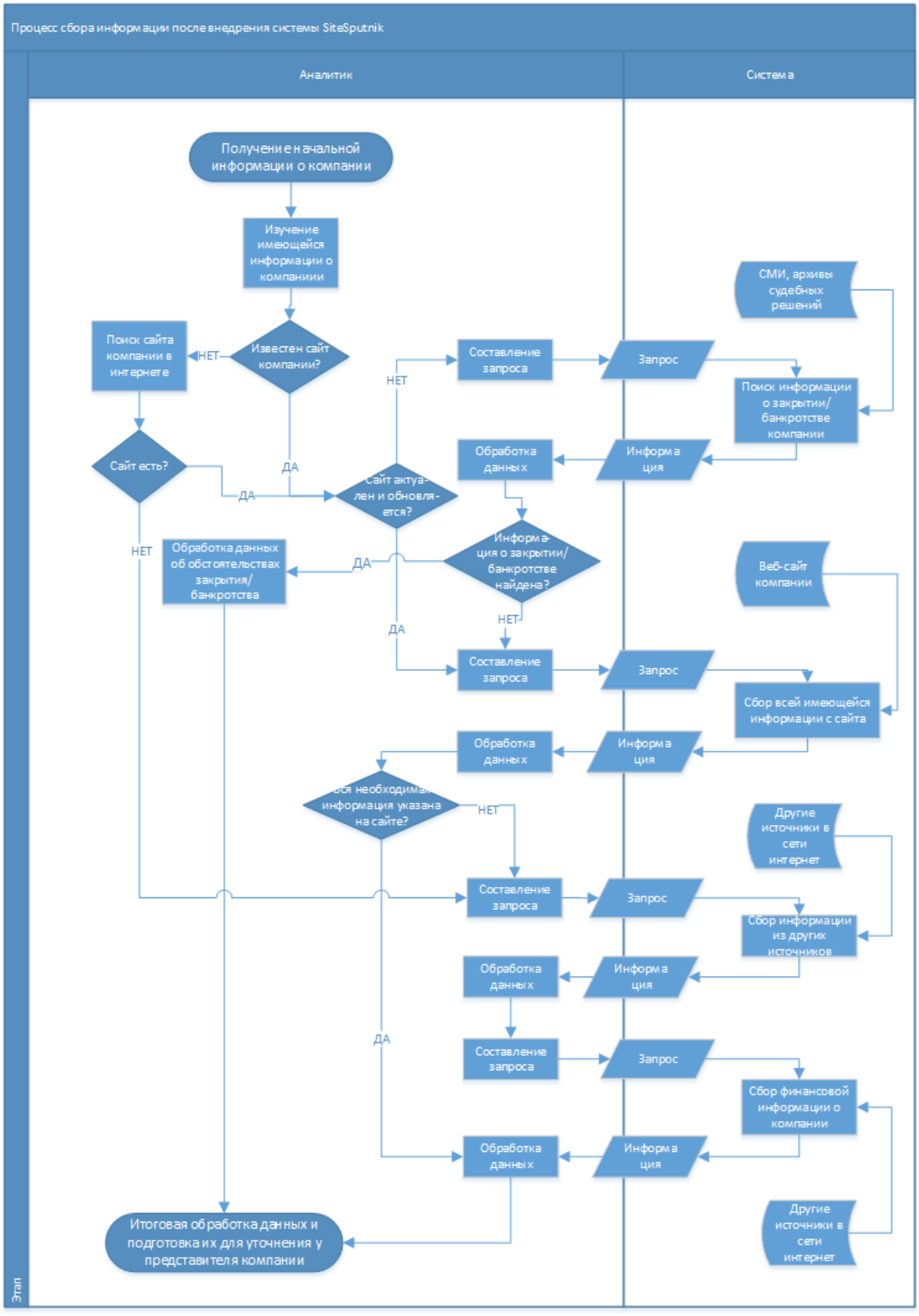


Рисунок 8 - Процесс сбора информации после внедрения системы SiteSputnik

При этом в процессе аналитического объединения результатов запросов происходит семантический анализ объектов и вычисление семантических связей, то есть нахождение интернет-страниц, на которых пересекаются искомые объекты. Это позволяет

автоматически удалять дубликаты выдачи не только по адресам интернет-страниц, но и по их смысловому содержанию и выводить всю собранную информацию по ее реальной релевантности относительно поставленной задачи, а не по релевантности поисковых систем.

2. Процедурный поиск в интернете.

Процедурный поиск позволяет проводить последовательный поиск по произвольному количеству поисковых систем, при этом для каждой поисковой системы прописывается свой запрос за соответствующем языке поисковых запросов, что дает уникальную возможность полностью использовать алгоритмическую мощность языков запросов каждой из поисковых систем. Объединенные в пакет несколько таких запросов позволяют максимально полно и точно запрограммировать и выполнить задачу поиска и сбора информации по искомому объекту.

Функция WebSpider. Используется для мониторинга обновлений интернет-источников. При этом под мониторингом подразумевается весь процесс обработки информации, а именно: скачивание и сохранение веб-страниц, анализ ее контента и сравнение с предыдущей версией, выделение новых и измененных фрагментов интернет-страниц и проверка удовлетворения их контентом заданному критерию - пост-запросу.

Пост-запрос позволяет задавать критерии интересующей пользователя изменяемой информации не посредством указания от каких HTML-тегов до каких HTML-тегов находятся интересующие его изменения на страницах, а проверкой соответствия измененной информации определенному семантическому правилу. Это позволяет еще на стадии сбора информации отсеивать “информационный мусор”, такой как, например, обновление на сайте рекламы или счетчика.

На рис. 9 и рис. 10 представлен процесс сбора информации с использованием ресурсов SiteSputnik при условии, что у компании в наличии есть веб-сайт.

4.3.2. Для компаний, не имеющих веб-сайта:

1. Прежде всего, применяется функция “**Contacts**”, позволяющая по списку наименований организаций извлекать из сети интернет контактную информацию для каждой из них, такую как: адрес, телефон, факс, e-mail, после чего программа самостоятельно оформляет полученные данные в виде таблицы.
2. После чего с помощью функции **“Objects”**, позволяющей собирать информацию о физическом или юридическом лице по набору параметров, таких как: название, контактные данные, ИНН и другие, собирается информация об организациях по информации, полученной с помощью функции “Контакты”.
3. После чего на основе полученных об организациях данных применяется **процедурный поиск в интернете** для получения недостающей информации.

На рис. 11 представлен процесс сбора информации с использованием ресурсов SiteSputnik при условии, что у компании в наличии нет веб-сайта и всю информаци необходимо собирать из иных источников.

4.7. Пример реализации фрагмента технологии ИО ИА

Необходимо отметить, что качество результата тестируемой технологии на данный момент существенно ограничены возможностями демонстрационной версии программы SiteSputnik как с точки зрения доступных дополнительных функций, так и с точки зрения глубины простого поиска.

В качестве тестовых сайтов для проверки работоспособности поиска были выбраны две торговые сети, отличающиеся друг от друга информационным наполнением официальных сайтов:

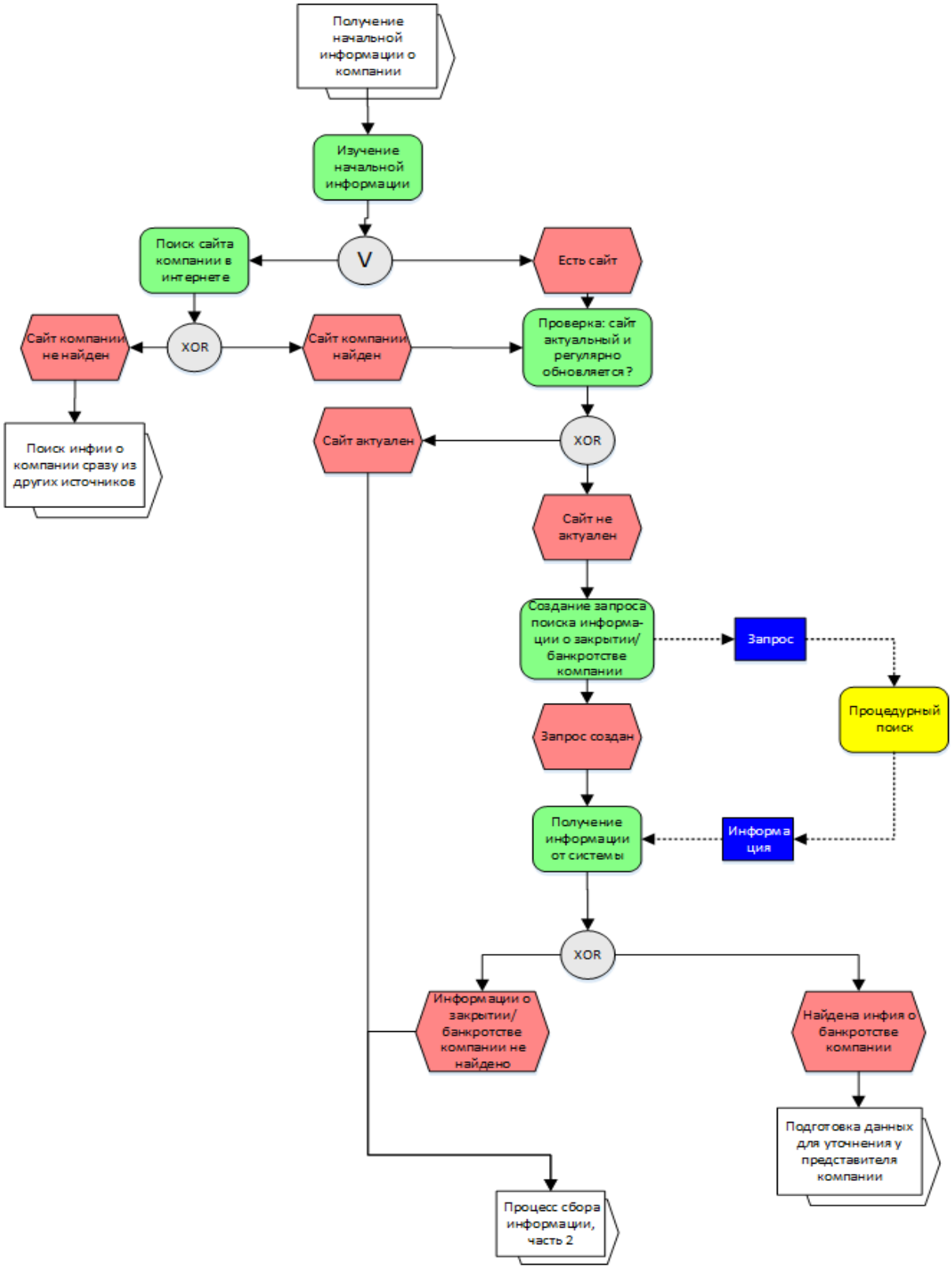
- ГК Дикси, публичная компания, входящая в топ-5 ритейлеров FMCG и регулярно выкладывающая на своем веб-сайте свежие пресс-релизы, финансовые и операционные показатели, актуальную документацию, отчетность для инвесторов и т.д.
- PRISMA, крупная сеть супер- и гипермаркетов, имеющая актуальный информативный веб-сайт с довольно небольшим количеством необходимой информации.

4.7.1. Выгрузка информации с веб-сайта организации.

Синтаксис:

(1) host:(2) || (3)=(4), где:
(1) - текст поискового запроса
(2) - сайт организации, по которому ведется поиск
(3) - поисковая система, в которой ведется поиск
(4) - глубина поиска (количество страниц)

поисковой системы был выбран Яндекс, поскольку при поиске информации в русскоязычном интернете о российских компаниях, он, исходя из практического опыта, выдает более релевантные результаты по сравнению с конкурентами. При этом для более



исунок 9 - Процесс сбора информации с использованием ресурсов SiteSputnik при условии, что у компании в наличии есть веб-сайт

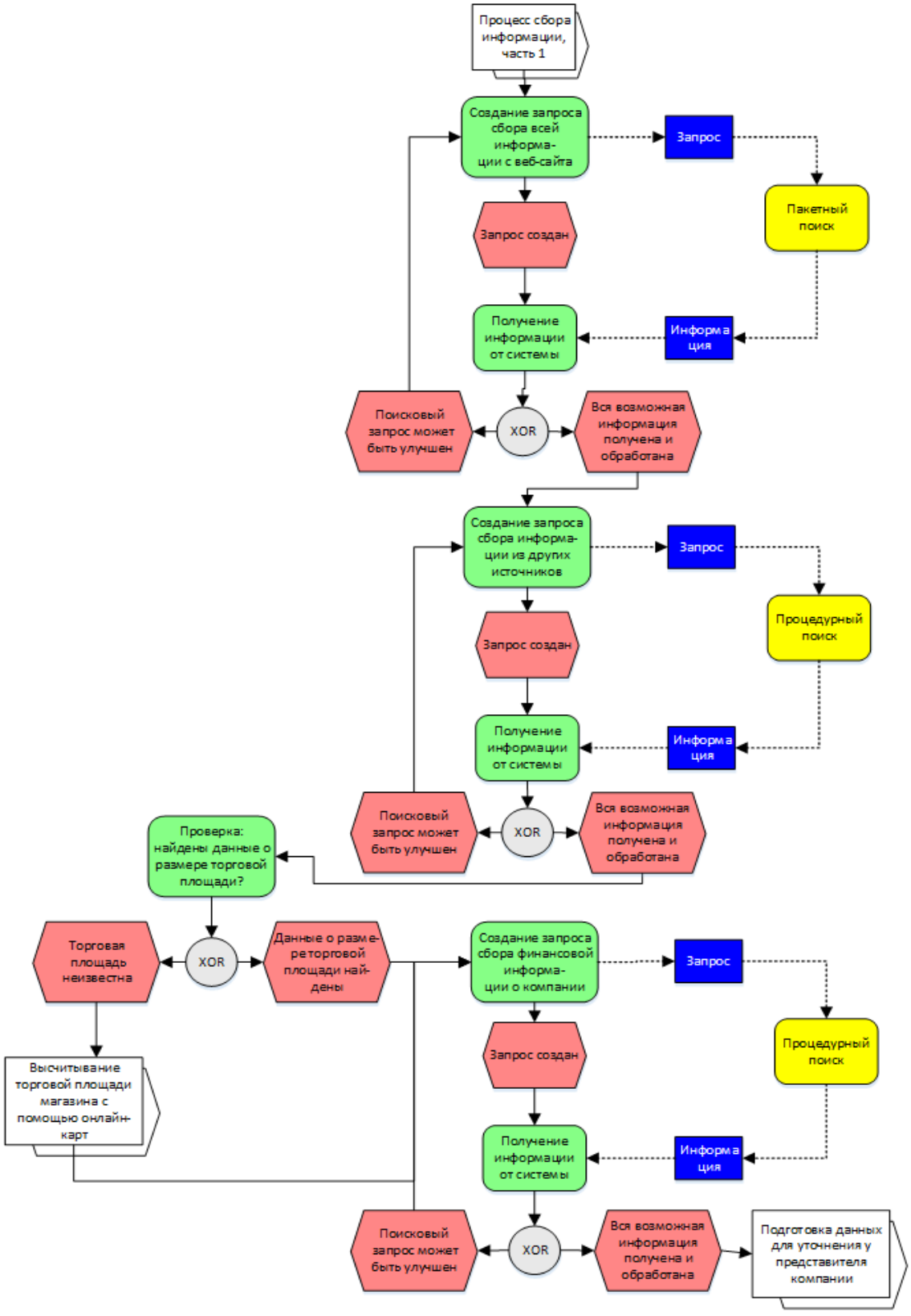


Рисунок 10 - Процесс сбора информации с использованием ресурсов SiteSputnik при условии, что у компании в наличии есть веб-сайт - продолжение

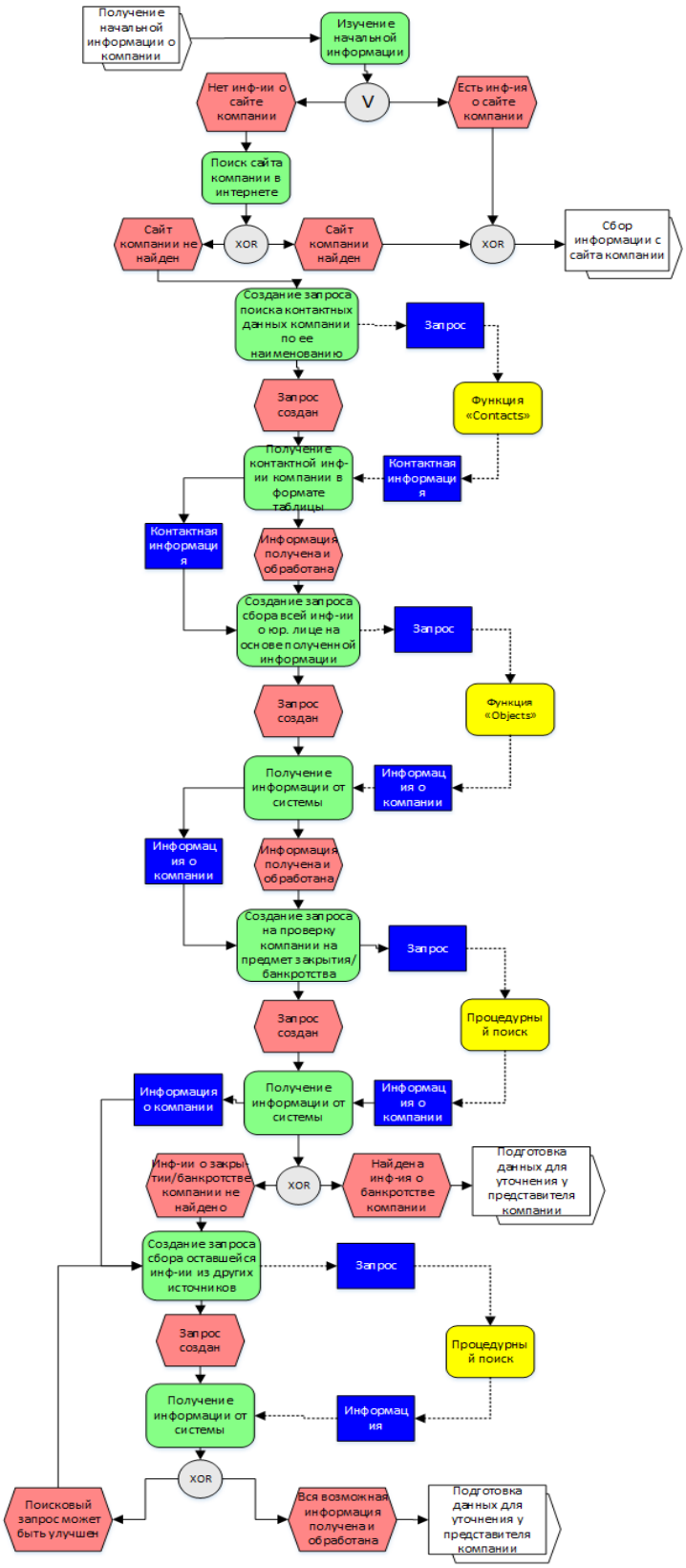


Рисунок 11 - Процесс сбора информации с использованием ресурсов SiteSputnik при условии, что у компании нет веб-сайта

сложного поиска по другим источникам, таким как, например, средства массовой информации, как уже было упомянуто выше, необходимо использовать как минимум 2-3 поисковые системы.

Для постоянно изменяющейся во времени информации, такой как количество магазинов и объемы выручки, используется сортировка результатов поиска не по релевантности страницы, а по последней дате ее изменения (YANDEX->Дата)

- **Юридическое название**

(((юридическое / юр) /2 (название | лицо)) | ооо | оао | зао | пао | нао | ао | ип) host:www.dixygroup.ru || Yandex=1

- **Менеджмент сети**

(директор | глава | руководитель | менеджмент | начальник | акционер | хозяин | учредитель) host:www.dixygroup.ru || Yandex=1

- **Фактический адрес**

((центральный | наш | головной | главный | компании) /3 (офис | адрес) | (нас /3 найти) | (мы /3 находимся)) host:www.dixygroup.ru || Yandex=1

- **Телефон/Факс/e-mail**

(телефон | тел | факс | (связаться /3 нами) | (написать /2 нам)) host:www.dixygroup.ru ||Yandex=1

- **Интернет-магазин**

((интернет | онлайн) /2 (магазин | каталог)) host:www.dixygroup.ru || Yandex=1

- **Общее количество магазинов сети на 1.1.2016**

((количество | число | открыто | существует | работает | “у нас” | новый | еще) / 3 (магазин | минимаркет |супермаркет | гипермаркет | отдел | точка)) host:www.dixygroup.ru || YANDEX->Дата=1

- **Общая торговая площадь магазинов сети на 1.1. 2016**

(общий | средний | суммарный | минимальный | максимальный | стандартный | обычный) /2 ((торговая +площадь) | размер) /2 (магазин | минимаркет |супермаркет | гипермаркет | отдел | точка) host:www.dixygroup.ru YANDEX->Дата=2

- **Чистая выручка (без учета НДС) торговой сети в 2014-2015 гг., млрд. руб.**

(выручка | прибыль | продажи) host:www.dixygroup.ru YANDEX->Дата=1

- **Количество РЦ/складов на 1.1. 2016**

"(рц | ((распределительный | логистический) / 1 центр) | (складское /1 помещение) | склад) host:www.dixygroup.ru || YANDEX->Дата=1

- **Площадь складов**

+площадь /3 ((распределительный | логистический) / 1 центр) | (складское /1 помещение) | склад) host:www.dixygroup.ru || YANDEX->Дата=1 ^^Площадь складов

- Пример текстового отображения в интерфейсе SiteSputnik:

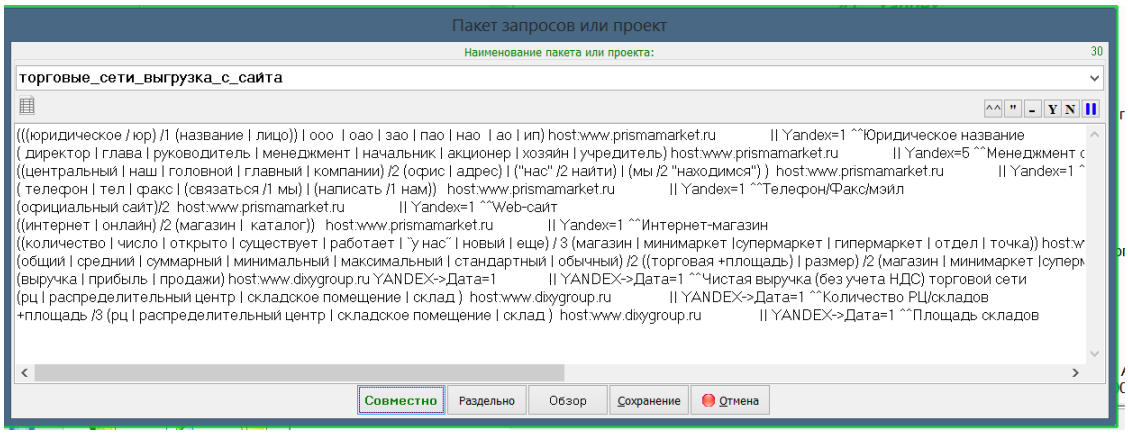
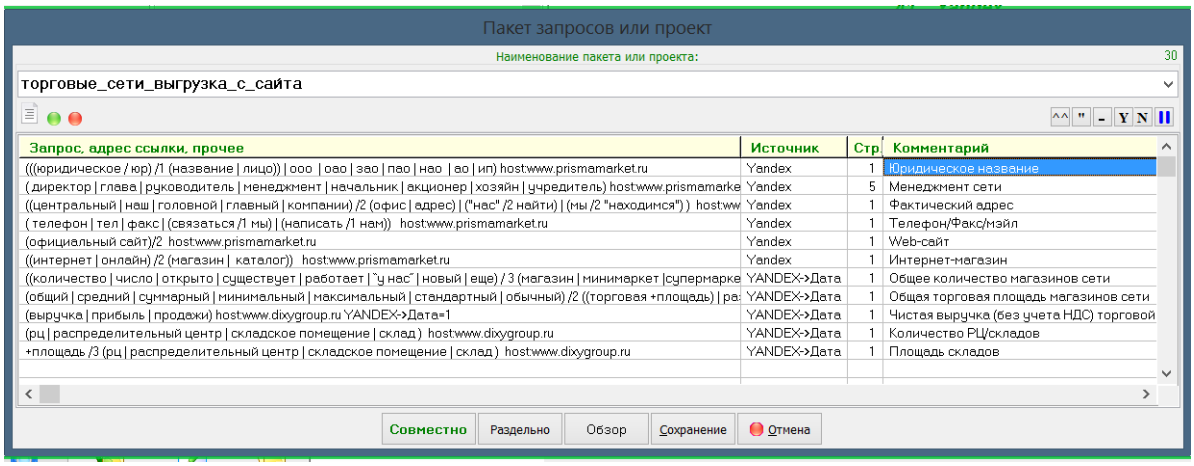


Рисунок 12 - Пример текстового отображения пакета запросов в интерфейсе SiteSputnik

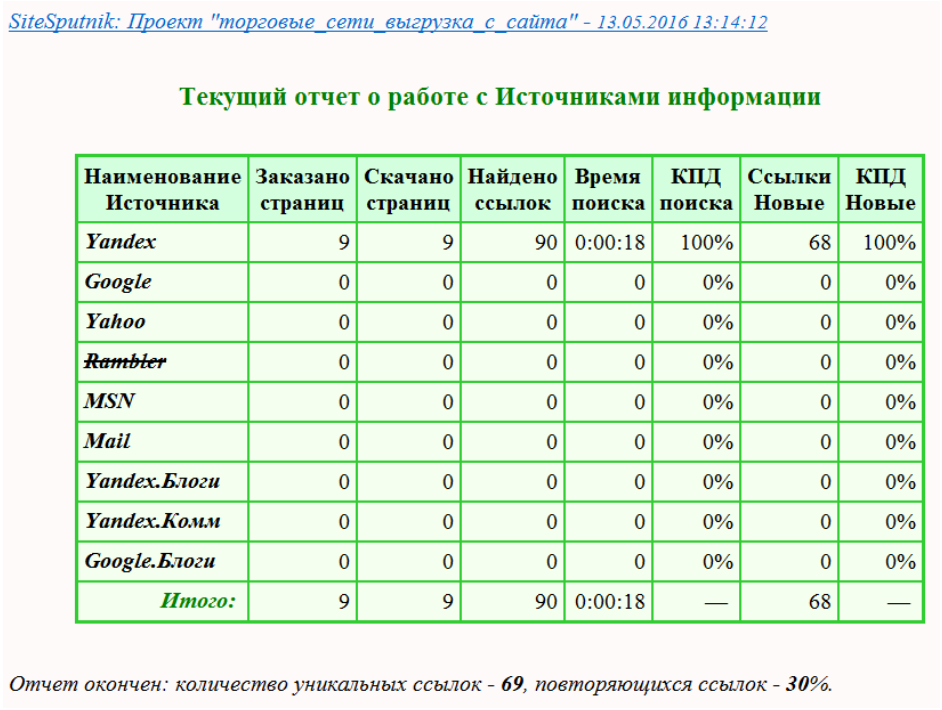
- Пример табличного отображения:



Запрос, адрес ссылки, прочее	Источник	Стр	Комментарий
(((юридическое / юр) /1 (название | лицо)) | ооо | оао | зао | пао | нао | ао | ип) host:www.prismamarket.ru	Yandex	1	Юридическое название
(директор | глава | руководитель | менеджмент | начальник | акционер | хозяйн | учредитель) host:www.prismamarke	Yandex	5	Менеджмент сети
((центральный | наш | головной | главный | компании) /2 (офис | адрес) | ("нас" /2 найти) | (мы /2 "находимся")) host:ww	Yandex	1	Фактический адрес
(телефон | тел | факс | (связаться /1 мы) | (написать /1 нам)) host:www.prismamarket.ru	Yandex	1	Телефон/Факс/мэйл
(официальный сайт)/2 host:www.prismamarket.ru	Yandex	1	Web-сайт
((интернет | онлайн) /2 (магазин | каталог)) host:www.prismamarket.ru	Yandex	1	Интернет-магазин
((количество | число | открыто | существует | работает | ~у нас~ | новый | еще) / 3 (магазин | минимаркет |супермарке	YANDEX->Дата	1	Общее количество магазинов сети
(общий | средний | суммарный | минимальный | максимальный | стандартный | обычный) /2 ((торговая +площадь) | ра	YANDEX->Дата	1	Общая торговая площадь магазинов сети
(выручка | прибыль | продажи) host:www.dixygroup.ru YANDEX->Дата=1	YANDEX->Дата	1	Чистая выручка (без учета НДС) торговой
(рц | распределительный центр | складское помещение | склад) host:www.dixygroup.ru	YANDEX->Дата	1	Количество РЦ/складов
+площадь /3 (рц | распределительный центр | складское помещение | склад) host:www.dixygroup.ru	YANDEX->Дата	1	Площадь складов

Рисунок 13 - Пример табличного отображения пакета запросов в интерфейсе SiteSputnik

4.7.2. Пример результирующей поисковой выдачи:



SiteSputnik: Проект "торговые_сети_выгрузка_с_сайта" - 13.05.2016 13:14:12

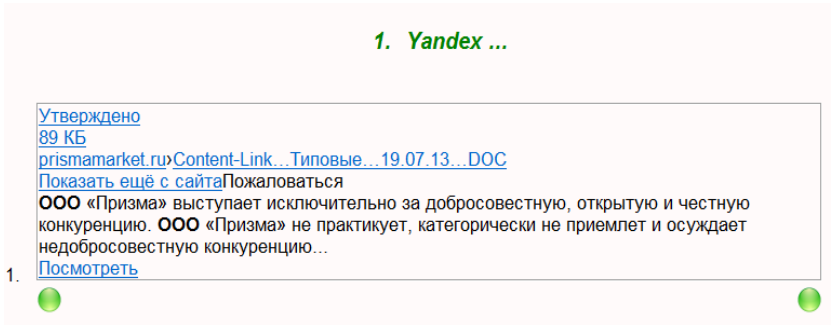
Текущий отчет о работе с Источниками информации

Наименование Источника	Заказано страниц	Скачано страниц	Найдено ссылок	Время поиска	КПД поиска	Ссылки Новые	КПД Новые
Yandex	9	9	90	0:00:18	100%	68	100%
Google	0	0	0	0	0%	0	0%
Yahoo	0	0	0	0	0%	0	0%
~~Rambler~~	0	0	0	0	0%	0	0%
MSN	0	0	0	0	0%	0	0%
Mail	0	0	0	0	0%	0	0%
Yandex.Блоги	0	0	0	0	0%	0	0%
Yandex.Комм	0	0	0	0	0%	0	0%
Google.Блоги	0	0	0	0	0%	0	0%
Итого:	9	9	90	0:00:18	—	68	—

*Отчет окончен: количество уникальных ссылок - **69**, повторяющихся ссылок - **30%**.*

Рисунок 14 - Пример отчета о работе с источниками информации

- **Юридическое название:**

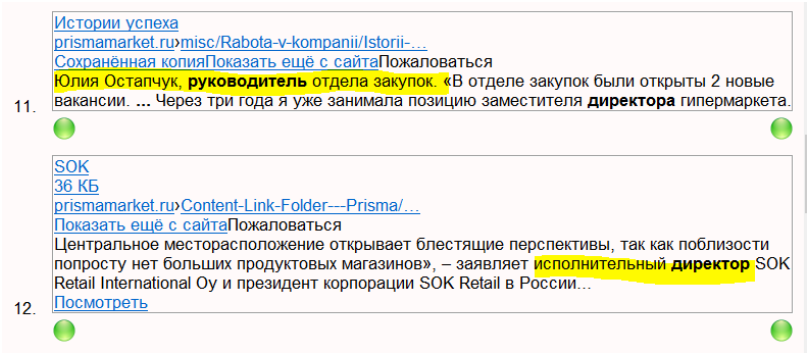


1. Yandex ...

1. Утверждено
89 КБ
prismamarket.ru›Content-Link…Типовые…19.07.13…DOC
Показать ещё с сайтаПожаловаться
ООО «Призма» выступает исключительно за добросовестную, открытую и честную конкуренцию. **ООО** «Призма» не практикует, категорически не приемлет и осуждает недобросовестную конкуренцию…
Посмотреть

Рисунок 15 – Пример поисковой выдачи по запросу о юридическом названии

- **Менеджмент:**



11. Истории успеха
prismamarket.ru›misc/Rabota-v-kompanii/Istorii-…
Сохранённая копияПоказать ещё с сайтаПожаловаться
Юлия Остапчук, **руководитель** отдела закупок. «В отделе закупок были открыты 2 новые вакансии. **...** Через три года я уже занимала позицию заместителя **директора** гипермаркета.

12. SOK
36 КБ
prismamarket.ru›Content-Link-Folder---Prisma/…
Показать ещё с сайтаПожаловаться
Центральное месторасположение открывает блестящие перспективы, так как поблизости попросту нет больших продуктовых магазинов», – заявляет исполнительный **директор** SOK Retail International Oy и президент корпорации SOK Retail в России…
Посмотреть

Рисунок 16 – Пример поисковой выдачи по запросу о менеджменте компании

- **Фактический адрес:**

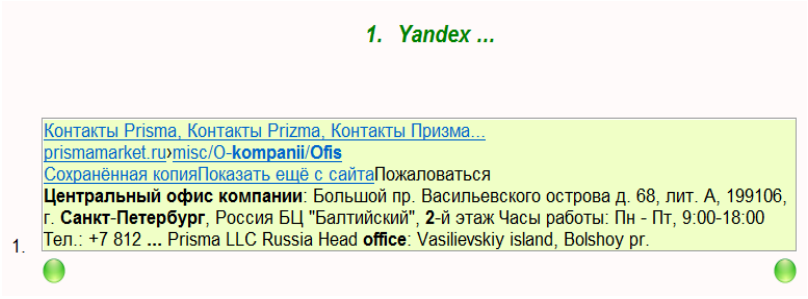


Рисунок 17 – Пример поисковой выдачи по запросу о фактическом адресе компании

- **Телефон/факс/email:**

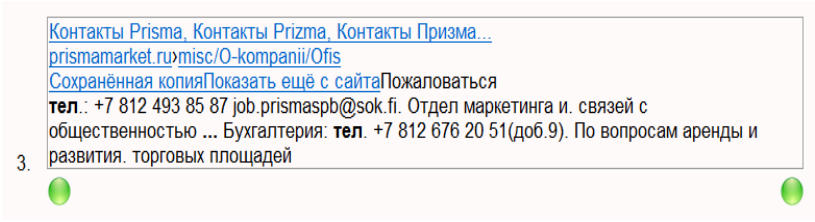


Рисунок 17 – Пример поисковой выдачи по запросу о контактных данных компании

- **Интернет-магазин:**

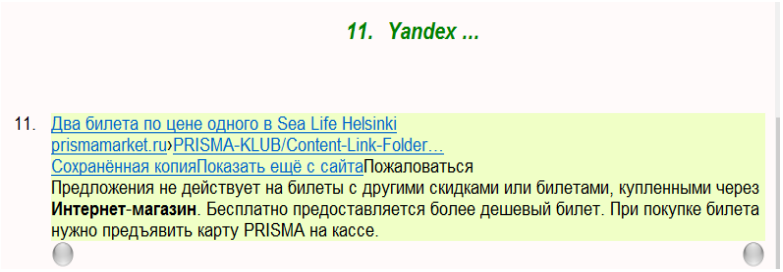


Рисунок 18 – Пример поисковой выдачи по запросу об адресе интернет-магазина компании

Все полученные по запросу результаты нерелевантны, из чего можно сделать вывод, что у торговой сети PRISMA нет интернет-магазина.

- **Общее количество магазинов сети на 1.1.2015:**

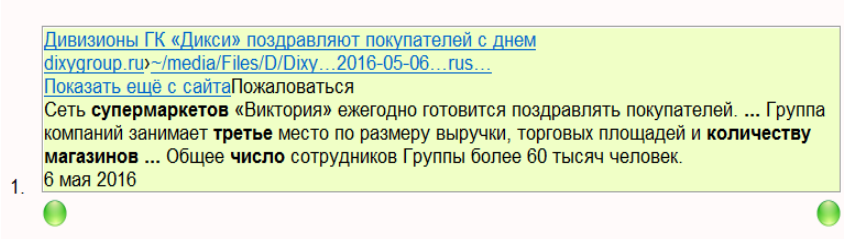


Рисунок 19 – Пример поисковой выдачи по запросу об общем количестве магазинов

Первым результатом запроса является документ в формате pdf, отрывок из содержания которого можно видеть на рис. 20. Документ содержит не только информацию о количестве магазинов, но и об их площади.

Группа компаний «ДИКСИ» (ММВБ: DIXY) - одна из лидирующих российских компаний в сфере розничной торговли продуктами питания и товарами повседневного спроса.

Открыв первый магазин «ДИКСИ» в 1999 году в Москве, после периода интенсивного органического развития и приобретения в июне 2011 года Группы Компаний «Виктория», по состоянию на 31 марта 2016 года Группа управляла 2 744 магазинами, включая: 2 595 магазинов «у дома» «ДИКСИ», 110 магазинов «Виктория», 1 магазин Cash и 38 компактных гипермаркетов «Мегамарт» и «Минимарт».

География деятельности Группы распространяется на четыре федеральных округа России: Центральный, Северо-Западный, Приволжский и Уральский, а также на Калининград и Калининградскую область.

Торговая площадь ГК «ДИКСИ» по состоянию на 31 марта 2016 года составляла 925 914 кв.м.

В 2015 году общая выручка Группы компаний «ДИКСИ» достигла 272 млрд рублей (4,5 млрд долларов США).

Рисунок 19 – Пример найденного по запросу документа.

- **Торговая площадь:**

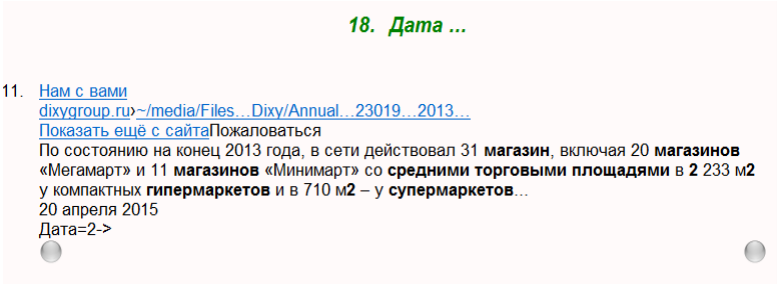


Рисунок 21 – Пример поисковой выдачи по запросу об общей торговой площади магазинов.

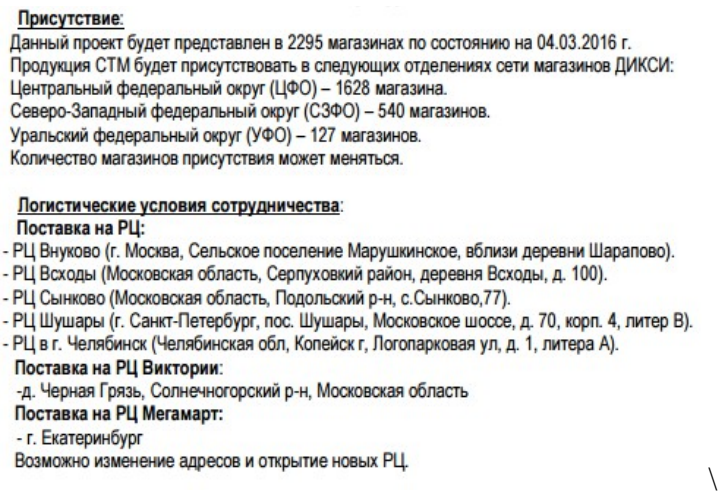
Полученная информация датируется 2013 годом и потому является устаревшей. Это объясняется тем, что, поскольку в своих отчетах компания публикует информацию как о количестве магазинов, так и об их площади, повторяющиеся ссылки были скрыты, новой актуальной информации найдено не было.

- **РЦ и склады:**

1.

Рисунок 21 – Пример поисковой выдачи по запросу о наличии у компании РЦ или склада

Содержимое файла в формате pdf:



Присутствие:
Данный проект будет представлен в 2295 магазинах по состоянию на 04.03.2016 г.
Продукция СТМ будет присутствовать в следующих отделениях сети магазинов ДИКСИ:
Центральный федеральный округ (ЦФО) – 1628 магазина.
Северо-Западный федеральный округ (СЗФО) – 540 магазинов.
Уральский федеральный округ (УФО) – 127 магазинов.
Количество магазинов присутствия может меняться.

Логистические условия сотрудничества:
Поставка на РЦ:
- РЦ Внуково (г. Москва, Сельское поселение Марушкинское, вблизи деревни Шарапово).
- РЦ Всходы (Московская область, Серпуховский район, деревня Всходы, д. 100).
- РЦ Сынково (Московская область, Подольский р-н, с.Сынково,77).
- РЦ Шушары (г. Санкт-Петербург, пос. Шушары, Московское шоссе, д. 70, корп. 4, литер В).
- РЦ в г. Челябинск (Челябинская обл, Копейск г, Логопарковая ул, д. 1, литера А).

Поставка на РЦ Виктории:
-д. Черная Грязь, Солнечногорский р-н, Московская область
Поставка на РЦ Мегамарт:
- г. Екатеринбург
Возможно изменение адресов и открытие новых РЦ.

Рисунок 22 – Пример найденного по запросу документа.

4.7.3. Качественная оценка результата поиска

Информация, полученная с помощью пакетного поиска, отличается высокой релевантностью - нужная информация, при условии наличия ее на сайте компании, была показана в 1-2 строчке поисковой выдачи, в случае нахождения информации, отвечающей нескольким поисковым запросам, на одной странице, ее дубликаты в выдачах последующих запросов были удалены. Оригинал веб-страницы или документа открывается одним кликом по ссылке, при этом в большей части случаев переходить на оригинальную страницу не было необходимости, поскольку искомая информация была вынесена в аннотацию.

Время, затраченное на поиск всей информации, примерно равно времени, необходимому для массовой замены шаблонного адреса веб-сайта в текстовом представлении пакетного запроса на реальный адрес веб-сайта рассматриваемой компании и значительно меньше, чем время, затрачиваемое на самостоятельный просмотр всех страниц веб-сайта в поисках необходимых данных.

ЗАКЛЮЧЕНИЕ

В результате работы был предложен метод автоматизации процесса сбора первичной информации в информационном агентстве, который позволяет сократить временные и стоимостные затраты в ходе проведения аналитического исследования путем повышения производительности труда и высвобождения человеческих ресурсов. Также предложенный способ позволит повысить качество собираемой информации, ее полноту, достоверность, оперативность, а следовательно, ценность выпускаемых информационных продуктов и даже, возможно, положение ИА на конкурентном рынке.

В ходе работы была изучена специфика деятельности ИА и проводимых им аналитических исследований, описан обобщённый процесс проведения аналитического исследования. После этого была обследована универсальная технология сбора первичной информации, применяемая в настоящее время в ИА, описаны основные информационные источники и обобщенный процесс сбора первичной информации.

После выявления существующих недостатков процесса сбора первичной информации, были выделены специфические требования к информационной технологии, предназначенной для автоматизации этого процесса, после чего был проведен обзор и последующее сравнение технологий, существующих в данный момент на рынках систем автоматизации маркетинговых исследований и конкурентной разведки. В результате сравнительного анализа была выбрана технология, наиболее полно соответствующая предъявленным требованиям.

Выбор технологии был осложнен тем фактом, что большая часть корпоративных систем поиска и анализа информации предназначены для комплексной поддержки бизнес-процессов маркетинга и конкурентной разведки предприятий и, таким образом, слабо соответствуют специфике процессов, протекающих в рассматриваемом ИА и имеют большое количество избыточного функционала.

В работе дано описание инструментария выбранной технологии и его адаптации к специфике аналитических исследований, проводимых ИА. На примере типичного для ИА аналитического исследования, представлен обновлённый процесс сбора первичной информации и применение в нем инструментария выбранной технологии, формализованы и описаны новые алгоритмы действий.

Приведен пример практической реализации технологии для решения задач поиска и обработки первичной информации, демонстрирующий преимущество использования выбранной технологии перед выполнением операций вручную, результатам функционирования технологии дана качественная оценка. Практическая реализация осложнялась наличием в свободном доступе только демонстрационной версии программы, которая хоть и дает возможность сделать выводы о возможностях полной версии, однако

тем не менее существенно ограничена и не позволяет проводить сколько-нибудь масштабных исследований.

Дальнейшим шагом в работе по автоматизации процесса сбора первичной информации является тестирование технологии на реальных данных в масштабах всего ИА, что в настоящий момент в рамках данной работы представляется невозможным.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Александров, А. Аналитика по-русски / А. Александров // Открытые системы. – 2007. - №8
2. Беляевский И.К. Маркетинговое исследование: информация, анализ, прогноз. Учебное пособие / И.К. Беляевский- М.: Финансы и статистика, 2001. – 320 с.
3. Вирник, Ю.П. Обзор информационных технологий, применяемых в аналитической работ / Ю.П. Вирник // Аналитический вестник Совета Федерации ФС РФ. – 2010. - №9(395)
4. Вороной А. Сравнительный анализ информационно-аналитических систем для обработки открытых источников информации / А. Вороной, П. Манько // Маркетинг и маркетинговые исследования. - 2007. - №3 (69)
5. Доронин А.И. Бизнес-разведка. 2-е изд., перераб. и доп / А.И. Доронин — М.: Ось-89, 2003. — 384 с.
6. Завьялов П.С. Маркетинг в схемах, рисунках, таблицах / П.С. Завьялов. – М.: Издательский Дом «ИНФРА-М», 2007http:///h
7. Иванов Л.А. Исследование рынка собственными силами. Мастер-класс / Л.А. Иванов. - СПб.: Питер, 2006 .- 144 с.
8. Колик А. Альтернатива: мы или конкуренты / А. Колик // М.: ИП Стрельбицкий, 2010. – 210 с.http:///h
9. Ландэ Д.В. Поиск знаний в Internet. / Ландэ Д.В. - М.:Диалектика, 2005. - 272 с.
10. Левкин И.М., Микадзе С.Ю. Добывание и обработка информации в деловой разведки. / И.М Левкин, С.Ю. Микадзе - СПб: Университет ИТМО, 2015. - 460 с. Нуралиев С.У. Маркетинг: Учебник для бакалавров / С.У. Нуралиев, Д.С. Нуралиева. — М.: Издательско-торговая корпорация «Дашков и К°», 2013. http:///h
11. Токарев Б.Е. Методы сбора и использования маркетинговой информации / Б.Е. Токарев. - М.: Юрист, 2001http:///h
12. Ющук Е. Интернет-разведка. Руководство к действию / Е. Ющук. - М.: Вершина, 2007
13. Анализ информации — превращение данных в аналитические выводы // ИКФ "АЛЬТ", Global Intelligence Alliance - [Электронный ресурс] URL: http://www.marketing.spb.ru/lib-research/Intelligence_Process.htm (дата обращения: 16.05.2016)
14. Аналитическая база “700 торговых сетей FMCG России. Демонстрационная версия // Информационное агентство «ИНФОЛайн» [Электронный ресурс]. URL: http://infoline.spb.ru/upload/iblock/be7/be78839a6d00e2327ba8c034fab15fd5.pdf (дата обращения: 16.05.2016)
15. Аренков И.А. Бенчмаркинг и маркетинговые решения. Монография. // Энциклопедия маркетинга. [Электронный ресурс]. URL: http://www.marketing.spb.ru/read/m12/4.htm (дата обращения: 16.05.2016)

16. Деревяшко, В.В. Влияние фактора старения информации на ее ценность для организации [Текст] / В.В. Деревяшко // Экономические науки. - 2010. - №1. - С. 425-427

17. Мыльников А.Б. Программа SiteSputnik (СайтСпутник). Сравнительный анализ поисковиков // Персональный сайт Алексей Борисовича Мельникова. – 2008. - [Электронный ресурс]. URL: http://sitesputnik.livejournal.com/804.html (дата обращения: 16.05.2016)

18. Нежданов И.Ю. Технологии конкурентной разведки // электронная книга – 2009. [Электронный ресурс]. URL: http://www.ci2b.info/wp-content/uploads/2013/01/Технологии-КР-Нежданов-ИЮ-20130102.pdf (дата обращения: 16.05.2016)

19. Основные технологические и рыночные тренды // Компания «Ай-Теко». [Электронный ресурс]. URL: http://www.i-teco.ru/solutions/business_intelligence_products/technological_and_market_trends/ (дата обращения: 16.05.2016)

20. Проблемы современных методов анализа текста // Компания «Ай-Теко». [Электронный ресурс]. URL: http://www.i-teco.ru/solutions/business_intelligence_products/modern_methods_of_text_analysis/ (дата обращения: 16.05.2016)

21. Schwartz B. Google’s Matt Cutts: 25-30% Of The Web’s Content Is Duplicate Content & That’s Okay // Search Engine Land. – 2013. [Электронный ресурс]. URL: http://searchengineland.com/googles-matt-cutts-25-30-of-the-webs-content-is-duplicate-content-thats-okay-180063 (дата обращения: 16.05.2016)