

## Технологии извлечения знаний из текста

*Николай Ильин, Сергей Киселев, Владислав Рябышкин, Сергей Танков*



Основную часть знаний аналитики получают в результате сравнения, анализа и синтеза информации из разрозненных фактов, размещенных в текстах. При работе с большими потоками документов процесс автоматического структурирования текстовой информации заменяет экспертный процесс выделения фактов и объектов, выполняемый вручную. В статье рассматриваются примеры использования новых технологий извлечения знаний из текстов на русском языке, ориентированных на работу с большими хранилищами данных.

До 85% новых знаний аналитики до сих пор получают, изучая тексты. В ближайшем будущем наиболее востребованными станут системы с максимально автоматизированными ETL-процессами структурирования контента (extract, transfer, load — «извлечение, преобразование, загрузка»). Важной чертой таких систем будет функция оперативного анализа информации, полученной по запросу для выбора дальнейшего направления исследования документов (автопилотирование направления исследования), выполняемой с помощью методов интеллектуального анализа текста.

К наиболее актуальным средствам интеллектуального анализа текстов относятся технологии выделения фактографической информации об объектах с учетом анафорических ссылок на них (ссылочные местоимения на объект, поименованный в тексте); нечеткий поиск; тематическое и тональное (точное и полное) рубрицирование; кластерный анализ хранилищ и подборок документов; выделение ключевых тем; построение аннотаций; построение многомерных частотных распределений документов и их исследование с помощью OLAP-технологий; использование методов интеллектуального анализа текста для определения направления исследования больших подборок документов и извлечения новых знаний.

В современных системах используется двухфазная технология аналитической обработки. В первой фазе (ETL) производится автоматизированный анализ отдельных документов, структуризация их контента и формирование хранилищ исходной и аналитической информации. Во второй фазе (OLAP, Text Mining, Data Mining) — извлечение в оперативном режиме знаний из хранилища или из полученной по запросу подборки документов. На наш взгляд, к наиболее интересным системам аналитической обработки относятся ClearForest, Convera RetrievalWare, Hummingbird KM, IBM Text Miner, инструменты компании IQMen, Inxight Smart Discovery Extraction Server, Ontos Miner, Oracle Text, ODB-Text, TextAnalyst, инструменты компании Smartware, XANALYS Link Explorer, X-Files, инструменты компании «Гарант-Парк-Интернет» и «Медialogия». Попробуем проанализировать современное состояние дел в области аналитических технологий на примерах конкретных систем.

### Особенности аналитической обработки

Первичная аналитическая обработка в фазе ETL требует значительных вычислительных ресурсов. Наш опыт эксплуатации систем с объемом фондов 5-10 млн. единиц хранения показывает, что если объем входных документов и время построения индекса принять за единицу, а запросы дополнительной памяти на диске и времени, требуемые на каждом из следующих видов обработки, как dV и dT соответственно, то получается следующая картина:

- выполнение индексирования: dV = 0,3-2, dT = 1;
- построение семантической сети: dV = 0,2-0,4, dT = 2-3;
- построение рубрик: dV = 0,001, dT = 0,1;
- создание аннотации и ключевых тем: dV = 0,1, dT = 1-2;
- терминологические векторы документов: dV = 0,1, dT = 0,02;
- хранилище аналитических данных: dV = 0,3, dT = 0,5;
- база данных фактографической информации, объединенной в досье: dV = 0,3, dT = 3.

Объем вторичных данных может быть в 3-4 раза больше объема документов, а время, необходимое на извлечение новых знаний, больше времени индексирования в семь-девять раз.

В ходе аналитической обработки происходит выделение текста фактографической информации об объекте, причем с учетом всех ссылок. Для этого сначала выделяются все предложения с упоминаниями об объекте (создается дайджест), в которых могут встречаться названия объекта («Иванов»), ссылки на него (анафория: «он», «который»...), а также обобщающие определения (корреференты: «воин», «семьянин»...). Нахождение и разрешение корреферентов и анафор дает увеличение объема дайджеста на 15-30%, а значит, и объема фактографической информации.

В начале исследования аналитики в первую очередь стремятся к полноте запроса, а не к его точности, поэтому объем релевантной подборки документов составляет сотни или тысячи единиц. Дальнейшее исследование проблемы производится уже после получения подборки документов с помощью кластерных, семантических карт или других методов. Такая технология работы аналитика сегодня типична как для работы в Internet, так и при работе со специализированными системами. Русский язык плохо поддается описанию формализмами различных уровней: морфологией, синтаксисом, семантикой. Например, для идентификации морфологических признаков лексики на русском языке необходимо выполнить также предсинтаксический анализ предложения для снятия омонимии. В любом случае реализации этих формализмов используют нечеткую модель анализа текста.

К наиболее актуальным направлениям извлечения знаний из текста на сегодняшний день относятся:

- аналитическая обработка фактов; ведение досье;
- извлечение и структурирование фактографической информации;
- поиск информации по запросам на естественном языке с использованием тезаурусов;
- направления поиска информации, объектов в хранилище документов, в подборке документов;
- аннотирование документов, построение дайджестов по объектам;
- проведение тематического анализа документов (кластеризация и рубрицирование);
- построение и динамический анализ семантической структуры текстов;
- выделение ключевых тем и информационных объектов;
- определение общей и объектной тональности сообщений;
- исследование частотных характеристик текстов.

## Поиск

Исторически первой и присутствующей сегодня во всех системах является *векторная модель* поиска, изобретенная Дж. Сэлтоном в 60-х годах. Большинство машин работают по принципу наличия в релевантном документе всех терминов запроса, учета их встречаемости в документах и их средней языковой частотности. Эта модель используется при обработке запросов на естественном языке, особенно на поисковых страницах сайтов; она же применяется для поиска похожих документов.

Продолжает активно использоваться *булева модель* поиска, которая позволяет вводить в запрос логические операторы, контекстные ограничения на расстояние между словами, строить разветвленные мощные запросы, использовать стоп-словарь и лексические шаблоны аналогично регулярным выражениям в скриптовых языках. Профессиональные системы, в дополнение к перечисленным базовым моделям, предоставляют поиск с использованием нечеткой булевой модели поиска, позволяющей поисковой машине доставлять документы, которые она считает релевантными, даже если некоторые «слабые» элементы запроса в них не встречаются.

Для семантического поиска широко используются *тезаурусы*, за счет которых происходит расширение запроса. Например, при поиске документов по автотранспортным происшествиям, запрос «ДТП» имеет фактор расширения 1:150, т. е. из одной лексики системой фактически генерируется 150 лексем для сервера поиска (см. рис. 1, правый фрейм). Активное использование тезаурусов русского языка сдерживается сегодня отсутствием актуальных словарей синонимов.

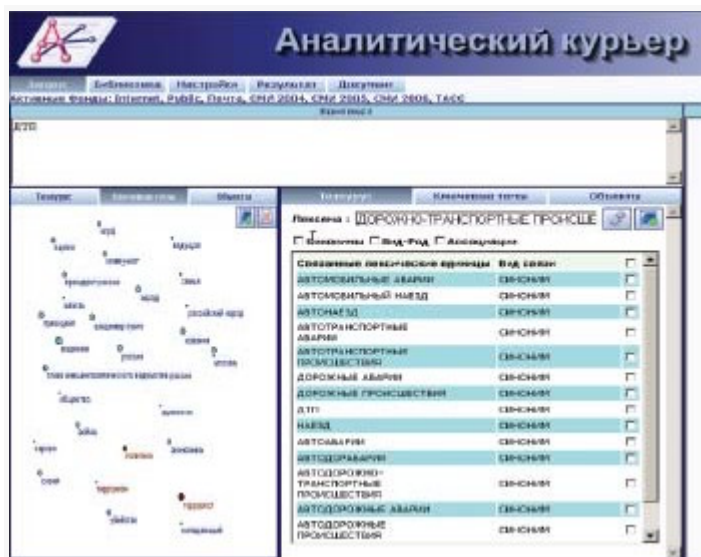


Рис. 1. Пример расширения запроса с использованием тезауруса в системе «Аналитический курьер» компании «Ай-Тек». На левом фрейме — карта наиболее важных ключевых тем подборки документов

Многие специалисты скептически относятся к идее осмысленного *диалога аналитика и системы* на формальном языке, поэтому имеет смысл максимально подстраивать язык запросов к мышлению и лексике аналитика. Проблемы в том виде, в котором с ними сталкивается сотрудник, зачастую трудно сразу сформулировать с помощью поисковых запросов. Возможность исполнять запросы на естественном языке с последующим использованием технологии навигации в полученной подборке документов может дать новые результаты, поскольку исследование направляется полученной информацией, а не только знаниями эксперта.

## Направления поиска

В одном из интервью Гари Флэйк, руководитель исследовательской лаборатории Yahoo!, сказал: «Если бы Web-поиск был совершенен, он бы выдавал ответ на каждый запрос, и это происходило бы так, будто на вопрос отвечает умнейший человек в мире, у которого есть под рукой вся справочная информация, и все это выполняется меньше, чем за мгновение». Пока же современные системы предоставляют визуальный интерфейс для анализа «препарированной» ими подборки документов, предоставляя аналитику выбор направления для дальнейшего анализа несколькими способами.

На рис. 1 (левый фрейм) представлен пример карты ключевых тем, полученной подборки, темы которой наилучшим образом (в математическом смысле — выделение в полученной подборке тем, имеющих максимальную дисперсию при определенном математическом ожидании) будут уточнять запрос, перемещая нужные темы в поле контекстного поиска.

Альтернативным способом поиска является поиск объектов и их взаимосвязей, выделенных автоматически из текста документов в фазе ETL-процесса. Этот способ позволяет исследовать связи объектов из документов без указания контекстного критерия на фильтрацию документов. Например, можно произвести поиск взаимосвязей объекта «Чейни» с другими объектами (рис. 2). Это можно использовать для навигации к нужным объектам, для получения и анализа документов о связях этих объектов. Дальнейшее развитие методов анализа связей объектов связано с решением задач типизации связей между объектами. В свою очередь, их решение ограничено качеством синтаксических анализаторов русского языка и тезаурусов.

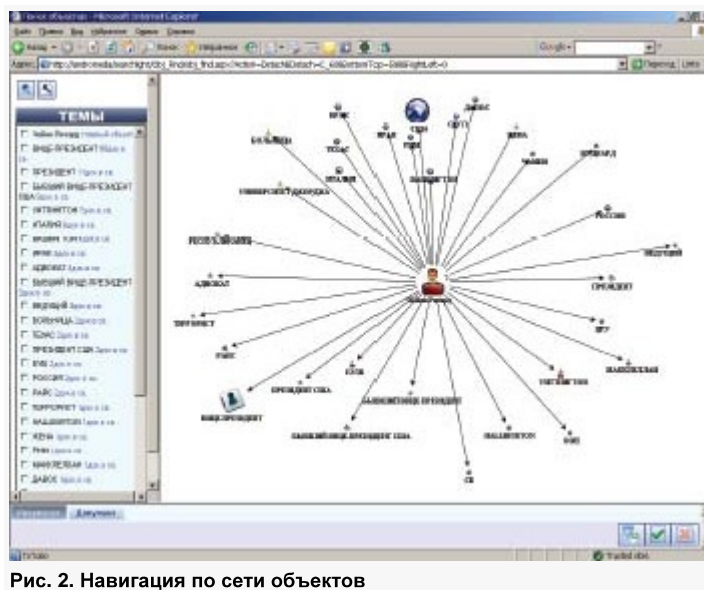


Рис. 2. Навигация по сети объектов

Очень полезен метод навигации в подборке документов с использованием OLAP-технологии. Система «на лету» строит многомерное представление полученной подборки документов с измерениями из полей карточки: рубрики, авторы, дата публикации, источники и др. Аналитик может погружаться в элементы разных измерений (например, в регионы федерального округа), просматривать документы в ячейках с нужными значениями частот и др. Дополнительно могут использоваться общие методы анализа и прогноза данных. На рис. 3 показана схема получения списка публикаций из ячейки двумерного распределения публикаций по регионам и подрубрикам рубрики «Политика». Этот метод используется при анализе динамики публикаций и факторов, ее определяющих.

## Автоматическое аннотирование

Открытые источники информации делают доступными огромное количество публикаций и тем самым ставят проблему эффективной работы с большими объемами документов. Предоставление сжатого смысла первоисточников в виде аннотаций в несколько раз повышает скорость анализа. Однако, наш опыт показывает, что аннотации — статичный результат, он используется при анализе «бумажных» документов, а при анализе коллекций электронных документов более наглядное и структурированное представление содержания одного или коллекции электронных документов дает интерактивная семантическая карта взаимосвязей тем документов. Современные системы аналитической обработки текстовой информации обладают средствами автоматического составления аннотаций. При этом существует два подхода к решению этой задачи [4].

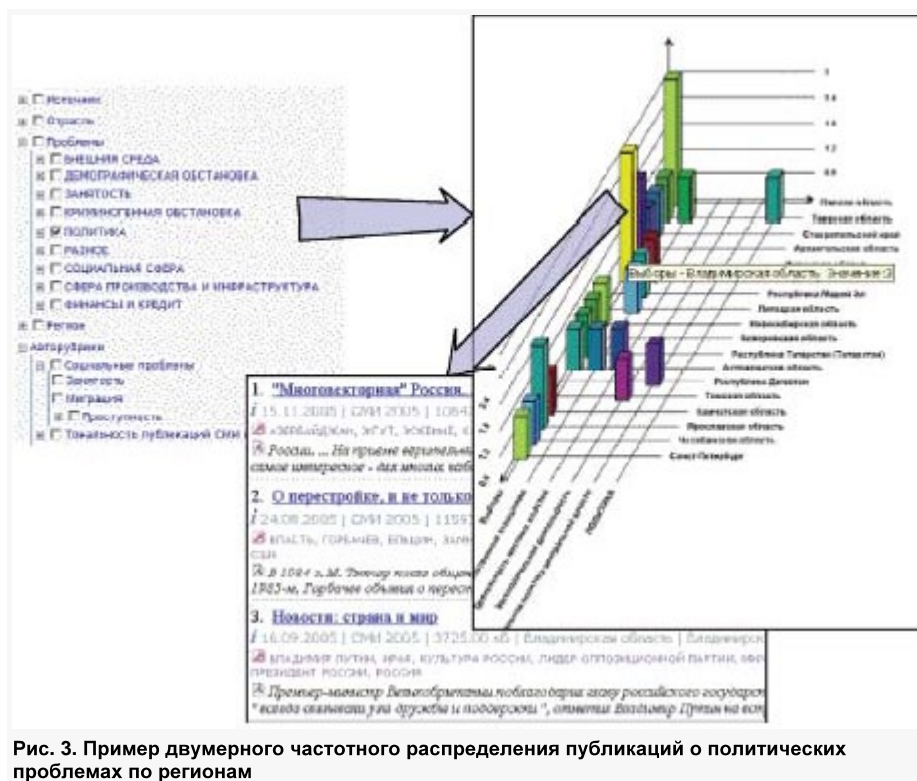


Рис. 3. Пример двумерного частотного распределения публикаций о политических проблемах по регионам

В первом подходе программа-аннотатор извлекает из первоисточника небольшое количество фрагментов, в которых наиболее полно представлено содержание документа. Это могут быть предложения, содержащие термы запроса; фрагменты предложений с окружением термов несколькими словами и др. В более развитых системах выделяются предложения, прямо содержащие ключевые темы документа (но не кореферентные ссылки на них).

При втором подходе аннотация представляет собой синтезированный документ в виде краткого содержания. Аннотация, сформированная в соответствии с первым подходом, качественно уступает получаемой при синтезе. Для повышения качества аннотирования необходимо решить проблему обработки кореферентных ссылок в русском языке. Еще одной проблемой, возникающей при синтезе аннотаций, является отсутствие средств семантического анализа и синтеза текста на русском языке, поэтому сервисы аннотирования ориентированы либо на узкую предметную область, либо требуют участия человека.

Большинство программ-аннотаторов построены по принципу выделения фрагментов текста. Так, исследовательская система eXtragon [1] ориентирована на аннотирование Web-документов. Для каждого предложения документа вычисляется вес на основе информации о ключевых словах, значимых словосочетаниях, их месте в тексте и присутствии в запросе, после чего предложения ранжируются, и из нескольких фраз с максимальным весом составляется реферат. В системе «Аналитический курьер» аннотация документа автоматически формируется из его фрагментов, а ее объем зависит от главных тем документа и настроек. В аннотацию по объектам или проблемам могут включаться анафорические предложения документа. Кроме этого, имеется компонент создания общей аннотации на основе взаимосвязей тем в семантической сети этой подборки документов.

## Тематическое рубрицирование и тональность

Технология автоматического рубрицирования используется при наличии сложившейся иерархии понятий в прикладной области. Технология основана на использовании метода распознавания образов применительно к текстам. Направления развития модели тематического рубрицирования связаны как с методами классификации, так и с методами выделения характерной лексики в корпусе обучающих рубрикатор документов для ее последующей классификации. Так, в системах «Аналитический курьер» и в модуле рубрицирования компании «Гарант-Парк-Интернет» каждый рубрикатор представлен в виде вероятностной нейросети. Эксперт предварительно создает типичные для рубрики коллекции документов, затем рубрикатор «обучается» на этих примерах и ставится на поток документов. Для русского языка потенциальная точность рубрицирования зависит от многих факторов: комплексности проблем, представленных в тексте (информационные

сообщения хорошо рубрицируются, поскольку они монотемны), от модели и максимальной размерности нейросети, репрезентативности лексики в тематике рубрики. В наибольшей степени точность зависит от качества лингвистического анализа, используемого для выделения словаря рубрик, в том числе от наличия средств разрешения анафории. Для текстов на русском языке качество рубрицирования («точность» × «полнота») может достигать 85%, что уступает качеству рубрицирования, выполняемому экспертами вручную. Во многих системах под рубрицированием понимается фильтрация документов по заранее сохраненным критериям запросов, что дает еще более слабые результаты, поскольку не учитываются факторы значимости одной и той же лексики для различных рубрик.

Еще одна задача классификации текста — рубрицирование тональности публикаций. Система должна определять эмоциональную окраску сообщений, как общую, так и по отношению к объектам документа. Нейросетевая модель, применяемая обычно при тематическом рубрицировании, здесь не работает. Каким бы хорошим словарем ни обладала система, главные проблемы классификации состоят в наличии инверсии смысла (тональности) и наличии анафорических ссылок на целевой объект, с которыми связана тональная лексика (например, во фразе «неэффективно борется с уличной преступностью» присутствует кратная инверсия тональности «борется с» но «неэффективно»). Специальный семантический анализ должен выделять те семантические роли слов, которые имеют отношение к эмоциональной окраске нужного объекта. Полнота определения тональности определяется качеством идентификации объектов в предложении. Правильное разрешение кореферентных ссылок на объект анализа повышает количество выделяемых упоминаний объекта и фактов, а значит, полноту анализа, на 30-80% в зависимости от содержания фактов. На рынке сегодня почти нет систем, которые выполняли бы функцию тонального рубрицирования.

## Динамический анализ тематической структуры публикаций

В отличие от авторубрицирования, выполняемого в фоновом режиме, анализ тематической структуры полученной подборки документов производится оперативно. Этот метод, кластерный анализ, используется при анализе новых проблем или событий, в которых тематическая структура динамична и еще неустойчива. При большом числе публикаций по проблеме важно выделить основные, репрезентативные группы тем — кластеры. Так, в новостном потоке «Яндекс.Новости» сообщения автоматически группируются в кластеры, соответствующие событиям [3]. Нужно помнить о том, что в обработке страниц поисковыми сайтами участвует малая часть всего текста сообщения, что приводит к существенному шуму в аналитической обработке. Однако, в отличие от новостных сайтов, цель которых — краткое изложение новостей дня, в информационно-аналитических системах пользователю необходимо разобраться в архиве, собираемом зачастую в течение нескольких лет. К примеру, в программе «Аналитический курьер» при объединении документов в кластер учитывается общность лексики и значений полей карточки. Кластеры могут пересекаться, что указывает на взаимосвязь их тем, можно погружаться в список документов любого кластера и в отдельные документы.

## Семантические карты подборки документов

Кластеризация позволяет разделять подборку документов на статистические смысловые группы, однако зачастую аналитику нужен более тонкий инструмент для обнаружения редких, но важных связей между темами подборки. В этом случае объектом анализа является семантическая карта взаимосвязей тем документов, а не сами документы. Карта представляет собой ориентированный граф, размеры узлов и толщина линий связи на котором соответствуют относительному весу тем и связей в подборке. Связи могут быть либо *типизированными* (определен семантический тип связи), либо *логическими* (установлен факт их наличия). Направление стрелки связи показывает причинно-следственную связь между темами — на более частную тему указывает стрелка. Толщина стрелки между темами отражает ее важность. В вершинах и связях находятся гиперссылки, ведущие к связанному набору документов. Выбрав узел на карте аналитик погружается в темы, непосредственно связанные с темой узла, как бы увеличивая масштаб карты и центрируя карту на теме. При этом состав тем карты изменится, появятся темы, наиболее тесно связанные с выбранной. Этот метод анализа часто используется также для совместного анализа нескольких карт, поиска похожих ситуаций или семантических шаблонов в различных картах и другие задачи. На рис. 4 представлен пример семантической карты.



Рис. 4. Пример семантической карты верхнего уровня

## Извлечение и структурирование фактографической информации

Для выделения объектов и их свойств (адреса, поездки, встречи, бизнес и т. п.) используются компоненты управления фактографической информацией и ведения досье. Например, в терминах системы Xfiles [2] факт об объекте является структурированным представлением фрагментов текста документа в виде значения факта: его суть, время и место совершения, его участники. Факты выделяются из предложений, содержащих упоминания объектов или ссылки на них. Технология выделения фактов основана на использовании специальных семантико-лингвистических методов, которые дают возможность получить точность и полноту фактов, сравнимую с экспертными.

Зачастую факты содержат информацию о взаимосвязях объектов и классифицируются как *прямые* (имеется факт о связи двух объектов); *нечеткие* (нет фактов); *общего места и времени* (для пары различных фактов различных объектов); *косвенные*, или *транзитивные* (через общий третий объект-связь у пары фактов различных объектов); *рефлексивные* (между парой атрибутов досье, связанных семантически). Если в одном из них появляется факт с определенным объектом-связью, то в симметричном атрибуте для объекта-связи также появляется этот факт. Скажем, атрибут «продажа акций» имеет симметричный атрибут «покупка акций». Симметричные атрибуты «срабатывают» по прямым связям. Свойство симметричности задается при создании атрибутов независимо от того, в какие досье они входят. При включении атрибута в другое досье свойство симметричности сохраняется.

Все эти свойства необходимы в системах аналитической разведки, немислимых без следующих сервисов: автоматическое выявление прямых и косвенных (т. е. через третье лицо) связей объекта; автоматическое выявление связей объектов по месту и времени (когда события произошли с разными объектами в одном месте или в близкое время); типизация связей, представленных различной лексикой; формирование групп объектов, связанных между собой общностью фактов (например, место, время, содержание факта); построение карты связей объектов для различных типов связей, визуализация и фильтрация связей; поиск оптимальных (обычно кратчайших) связей между заданными объектами; построение многомерных частотных распределений фактов. Сегодня системы извлечения фактов являются наиболее эффективным инструментом выделения нужной для принятия решений информации, заменяя ее поиск.

## Добыча данных

Широкое применение методов искусственного интеллекта позволяет порождать гипотезы — предложения по дальнейшему исследованию. Типичная технология анализа взаимосвязей проблем содержит следующие фазы:

- получение подборки документов по запросу;
- получение ее семантической карты;
- просмотр документов о связи выделенной пары тем;

- кластерный анализ этих документов;
- анализ документов нужных кластеров;
- резюме о структуре связи тем.

Так, и типичная технология анализа динамики развития проблемы в регионе (стране) включает следующие фазы:

- получение подборки документов по запросу;
- получение двумерного частотного распределения рубрик-проблем по регионам;
- выделение значимой проблемы в исследуемом регионе;
- получение частотного распределения рубрики-проблемы в регионе по времени;
- анализ документов в пиковые периоды времени;
- кластерный анализ этих документов;
- предложения по нормализации проблемы.

К примеру, многие ежедневно ездят на работу по Москве, но эти факты еще не свидетельствуют о наличии связи между ними, однако если два дипломата работали в одно время в небольшой стране, то с большой вероятностью следует, что они могли быть знакомы. Система должна уметь предлагать аналитику такого типа гипотезы.

## Заключение

Необходимо отметить обостряющуюся проблему с качеством базовых средств лингвистического анализа текста на русском языке. Другой важной проблемой является разрешение анафорических ссылок, требующая создания общего тезауруса русского языка. Эти проблемы сдерживают развитие методов интеллектуального анализа русскоязычных текстов, а их решение возможно скорее в рамках академических исследований (тем более, что это соответствует заявленным государственным приоритетам).

К наиболее актуальным методам сегодня можно отнести: семантические сети тем и объектов текста документов, выделение фактографической информации с учетом анафорических ссылок, возможность параллельной обработки распределенных архивов документов, различные стратегии нечеткого поиска, тематическое и тональное рубрицирование, кластеризация документов, аннотирование, анализ многомерных частотных распределений документов.

## Литература

1. П.Браславский, И.Колычев, Автоматическое реферирование веб-документов с учетом запроса. Грант ООО «Яндекс» № 102707, [company.yandex.ru/grant/2005/11\\_Braslavski\\_102707.pdf](http://company.yandex.ru/grant/2005/11_Braslavski_102707.pdf)
2. Сергей Киселев, Модель информационной системы бизнес-разведки. [«Открытые системы», 2005, № 5-6.](#)
3. Илья Сегалович, Михаил Маслов, Денис Нагорнов, «Как работают новые Яндекс.Новости». [company.yandex.ru/articles/smi-mirror.html](http://company.yandex.ru/articles/smi-mirror.html)
4. Удо Хан, Индерджиет Мани, Системы автоматического реферирования. [«Открытые системы», 2000, № 12.](#)

*Николай Ильин — начальник управления информационных систем Спецсвязи ФСО России, Сергей Киселев ([kiselev@i-teco.ru](mailto:kiselev@i-teco.ru)) — заместитель директора департамента разработки информационно-аналитических систем, Владислав Рябышкин ([ryabyshkin@i-teco.ru](mailto:ryabyshkin@i-teco.ru)) — системный аналитик компании «Ай-Техо», Сергей Танков — заместитель начальника управления информационных систем Спецсвязи ФСО России.*