Информационная технология КЛЮЧИ К ТЕКСТАМ®: функции, возможности, приложения

Крейнес М.Г.,
сотрудник ООО «БАЗИСНЫЕ ТЕХНОЛОГИИ»

Информационная технология КЛЮЧИ К ТЕКСТАМ® предназначена для решения задач поиска и анализа неструктурированной текстовой информации, зафиксированной на естественных языках (русском и английском).

Введение

Информационная технология КЛЮЧИ К ТЕКСТАМ® основана на вычислительном формировании моделей семантики отдельных документов в форме списков «ключевых» слов с весами. На базе оригинального решения этой классической задачи удалось реализовать широкий набор разнообразных вычислительных сервисов по смысловому поиску и анализу текстовой информации. Технология КЛЮЧИ К ТЕКСТАМ® защищена тремя действующими патентами на полезные модели. Программное ядро технологии КЛЮЧИ К ТЕКСТАМ® составляют десять официально зарегистрированных программ для ЭВМ, реализующих содержательную обработку текстовой информации. Априорная информация о смысле, значении и сочетаемости слов в работе информационной технологии КЛЮЧИ К ТЕКСТАМ® не используется. Программная реализация технологии включает в себя более ста программных модулей, функционирует в операционной среде Linux и ориентирована на параллельную вычислительную реализацию в распределённых вычислительных средах.

Основные функции информационной технологии КЛЮЧИ К ТЕКСТАМ®

Информационная технология **КЛЮЧИ К ТЕКСТАМ®** обеспечивает следующие функции по обработке и анализу текстовой информации на естественных языках:

- вычислительное формирование системы наукоёмких вторичных информационных ресурсов, характеризующих тематику и содержание текстовых документов —

словарей лемм — идентификаторов грамматически нейтральных форм слов документов (словарные лингвистические модели текстов) и списков ключевых слов с весами (словарные модели семантики текстов), характеризующих тематику и содержание документов;
- контекстный поиск документов по заданному числу слов из сформированного пользователем списка слов. Возможен поиск по трём категориям слов — словам, входящим в текст, ключевым словам текста, «главным» ключевым словам текста;
- вычислительное построение модели семантики текста в форме аннотации, сформированной из фрагментов анализируемого документа;
- поиск текстов, содержательно и тематически сходных с выбранным пользователем образцом;
- вычислительное формирование моделей семантики текстовых коллекций (групповых — вычислительное определение наличия и состава тематически однородных групп документов в текстовой коллекции при отсутствии априорной информации о наличии и составе таких групп, так называемая кластеризация текстовых коллекций, и словарных семантических моделей в форме списков слов, тематически связанных в рамках конкретной анализируемой коллекции);
- вычислительная категоризация документов на основании моделей семантики текстовых коллекций и моделей семантики отдельных текстов;
- семантическая навигация в рамках текстовых коллекций на основании моделей семантики коллекций.

Анализ и поиск текстовой информации в технологии КЛЮЧИ К ТЕКСТАМ® основаны на использовании наукоёмких вторичных информационных ресурсов — моделей семантики отдельных документов, вычислительно сформированных в форме списков «ключевых» слов по первичным информационным ресурсам — текстам на естественных языках.

Как это делается

Лемматизация документов

Информационная технология КЛЮЧИ К ТЕКСТАМ® использует в качестве исходного «материала» для построения семантических моделей текстовых документов лемматизированный словарный состав документа. Лемма слова — это идентификатор грамматически нейтральной формы слова. В наших проектах используется оригинальный алгоритмический лемматизатор, основанный на формальном описании знаний о морфологии языка в определённом стандарте представления соответствующих знаний. Поэтому процесс «подключения» новых европейских языков требует только создания необходимого описания соответствующего языка.

Вычислительное построение ключевых слов текстов на естественных языках

Эффективная организация поиска и анализа информации в рамках информационной технологии КЛЮЧИ К ТЕКСТАМ® основана на использовании наборов

ключевых слов, являющихся носителями основной тематики и содержания текстов — результатах автоматического смыслового индексирования — наукоёмкого вычислительного анализа текстов на естественных языках. Наборы ключевых слов документов являются основным типом наукоёмких вторичных информационных ресурсов, автоматически формируемых в ходе загрузки текстовых документов в хранилище данных, используемое информационной технологией КЛЮЧИ К ТЕКСТАМ®.

Задача вычислительного определения ключевых слов текстов на естественных языках с пятидесятых годов XX века привлекала математиков, программистов, специалистов в области анализа данных, лингвистов и психологов. К сожалению, господствующие сегодня подходы к решению проблемы извлечения смысла текста, сформировавшиеся в эпоху кибернетического романтизма, не дают достаточно надёжного и универсального предметно независимого решения. Информационная технология КЛЮЧИ К ТЕКСТАМ® основана на оригинальном вычислительном решении этой классической задачи: на формировании моделей семантики отдельных документов в форме списков «ключевых» слов с весами. Слова, входящие в такие списки слов, в нашей технологии формально определяются как наиболее сильно связанные в конкретном документе в некотором комбинаторном смысле. Оказалось, что вычисляемые множества слов — уникальные, устойчивые и воспроизводимые характеристики документа, а разумный читатель воспринимает их в качестве носителей основной тематики и содержания текста. Последнее свойство наборов «ключевых» слов, формируемых в результате формальных вычислений, доказывает их содержательность (так называемое свойство интерпретируемости) и обосновывает возможность именовать их ключевыми словами без кавычек.

Основные свойства моделей семантики отдельных документов в форме списков «ключевых» слов:

- *уникальность* — содержательно близким текстам соответствуют близкие списки ключевых слов с весами;
- *устойчивость* — при незначительных изменениях текста списки ключевых слов с весами меняются незначительно;
- *воспроизводимость* — для произвольного текста вычисленный при неизменных условиях список ключевых слов с весами всегда неизменен.

Для вычислительного построения моделей семантики отдельных документов в форме списков «ключевых» слов необходимы:

- текст документа;
- репрезентативная для языка, на котором написан документ, коллекция документов;
- знания о морфологии языка документа, представленные в определённом формате.

При построении моделей семантики отдельных документов в форме списков «ключевых» слов при желании пользователя могут быть учтены:

- словарь слов, воспринимаемых пользователем в качестве синонимов;
- словарь устойчивых словосочетаний, воспринимаемых пользователем в качестве единого означающего для определённых объектов.

Поиск содержательно похожих документов

В информационной технологии КЛЮЧИ К ТЕКСТАМ® поиск содержательно похожих документов использует анализ наукоёмких вторичных информационных ресурсов документов — сравнение списков ключевых слов с весами. Результатом такого сравнения является количественная оценка смысловой близости пары документов. При этом

значению единица (100%) соответствует полное смысловое совпадение, а значению ноль — отсутствие смысловых совпадений. Поиск содержательно похожих документов может выполняться по желанию пользователя или в рамках определённых фиксированных сценариев работы, например, для автоматического построения списков подозрительно схожих документов, так называемых «тревожных списков». В обоих случаях задаётся документ — образец, для которого отыскиваются содержательно схожие документы. Оценивание смысловой близости — достаточно трудоёмкая вычислительная процедура. Поэтому для ограничения потребностей системы в вычислительных ресурсах выполняется предварительный поиск «подозрительных» документов, содержащих ключевые слова документа — образца, на основании которых формируется задание и запрос на поиск. По результатам вычислений попарных оценок смыслового сходства образца и найденных документов формируется не возрастающий по величине оценки содержательного сходства с образцом отчёт о результатах поиска.

Контекстный поиск по заданному числу слов из сформированного пользователем списка слов с использованием моделей семантики отдельных документов в форме списков ключевых слов

Основная для традиционных поисковых систем функция — контекстный поиск (поиск документов, в которых встречаются слова запроса) — реализована и в информационной технологии КЛЮЧИ К ТЕКСТАМ®. Возможности нашей технологии позволили сделать эту стандартную функцию более удобной и комфортной для пользователя по сравнению с реализацией контекстного поиска в популярных поисковых системах. Пользователю предоставлен выбор: осуществлять контекстный поиск среди всего словарного состава документов, среди рассчитанных ключевых слов, которые являются носителями основного содержания и тематики документа, или среди «главных» (самых значимых) ключевых слов. Пользователь имеет возможность задать в качестве запроса набор слов в произвольной грамматической форме (поиск выполняется по леммам слов) и указать пороговое число слов из запроса. При наличии в документе числа слов из запроса не меньше порогового идентификатор документа включается в отчёт о результатах поиска. Отчёт упорядочен по невозрастанию числа слов из запроса в документах (первым в отчёте приводится ссылка на документ, использующий максимальное среди других найденных документов число слов из запроса). Это позволяет пользователю обойтись без написания запроса объёмом в страницу или больше в виде логических формул. При равенстве числа слов в запросе и пороговой величины для числа слов из запроса реализуется стандартный вариант контекстного поиска.

Построение моделей семантики документа в форме аннотаций

Формируемые при использовании информационной технологии КЛЮЧИ К ТЕКСТАМ® в результате формальных вычислений списки ключевых слов текстовых документов достаточно легко интерпретируются разумным человеком. Модель семантики документа в форме аннотации позволяет достаточно точно, полно и ярко представить тематику и содержание документа

в виде небольшого текста на естественном языке. В информационной технологии КЛЮЧИ К ТЕКСТАМ® в качестве аннотации формируется ограниченный по объёму (параметр, управляемый пользователем) набор предложений текстового документа. Предложения, включаемые в аннотацию, выбираются в ходе вычислительной процедуры таким образом, чтобы суммарный вес включённых в отобранные предложения ключевых слов документа был максимален при алгоритмическом обеспечении разнообразия ключевых слов. Для каждого документа реализуется одновременно два режима аннотирования: в контексте запроса (в аннотацию включаются «тяжёлые» предложения документа со словами из запроса) и в контексте самого найденного документа (в аннотацию включаются «тяжёлые» предложения без дополнительных ограничений). В результате пользователь получает, соответственно, модельное представление о содержании в найденном документе информации, связанной с непосредственным запросом, и об «информационной начинке» документа, как он есть. Аннотация в контексте запроса и аннотация в контексте самого документа могут существенно различаться.

Смысловая навигация по текстовой коллекции с использованием адаптивного диалогового тезауруса (АДТ)

Адаптивный диалоговый тезаурус (АДТ) — управляемый пользователем диалоговый инструмент смысловой навигации по текстовым коллекциям, являющийся оригинальным компонентом информационной технологии КЛЮЧИ К ТЕКСТАМ®. АДТ формируется вычислительно для произвольной выбранной пользователем коллекции документов по наукоёмким вторичным информационным ресурсам, характеризующим коллекцию: словарным лингвистическим моделям текстов и словарным моделям семантики текстов. Средствами формирования коллекции могут быть поиск (контекстный или по сходству с документом — образцом) и/или отбор документов по доступной метаинформации. АДТ — это список всех слов, используемых в документах отобранной коллекции, упорядоченный по невозрастанию суммарного веса слов в документах коллекции или числа документов, использующих слово. Именно этим обусловлено название нашего инструмента смысловой навигации. Действительно, АДТ предоставляет информацию о тематике и содержании конкретной коллекции, начиная с наиболее значимых слов (с максимальным суммарным весом) или с наиболее распространённых слов (с максимальным числом использующих их документов). Пользователь имеет возможность выбрать любое слово из представляющего АДТ списка слов и тем самым определить подмножество исходной коллекции, документы которого используют отобранное слово. АДТ автоматически перестраивается для новой коллекции, предоставляя пользователю модель её тематики и содержания.

Результатом использования АДТ является организация быстрого полностью управляемого пользователем в диалоговом режиме адаптированного к семантическим и словарным характеристикам конкретной коллекции поиска текстовой информации. Существенными положительными особенностями такого поиска являются доступность для пользователя априорной информации об уменьшении числа документов при выборе очередного шага диалоговой поисковой процедуры и принципиальная невозможность для пользователя сформировать запрос, по которому не удаётся найти релевантных документов. Таким образом, мы решаем две практические проблемы поиска текстовой информации: именно отсутствие априорных данных о возможности сужения множества документов в ходе поиска и формирование запроса, дающего «не пустой», но обозримый результат поиска, являются основными трудностями при применении стандартных методов поиска текстовой информации для решения нетривиальных задач поиска.

Кластеризация текстовых коллекций (групповая модель семантики текстовой коллекции)

Использование в качестве модели семантики отдельных документов наукоёмких вторичных информационных ресурсов, характеризующих их тематику и содержание, позволило решить классическую проблему вычислительного выделения в произвольной текстовой коллекции тематически однородных групп при отсутствии априорной информации о признаках и составе таких групп (задача кластеризации текстовых коллекций). Результатом решения данной задачи для текстовой коллекции является перечень тематически однородных групп документов коллекции, множества ключевых слов, характеризующих тематику каждой выявленной группы, и аннотации групп, составленные из предложений документов, включённых в группу. Модели семантики отдельных документов позволяют вычислительно находить содержательно близкие тексты и формировать оценки содержательной близости отдельного документа к группе документов. Это обеспечивает в коллекциях объёмом до 10–15 тысяч текстов вычислительный итеративный поиск тематически однородных групп документов в отсутствие априорной информации о наличии и составе таких групп, и вычислительное формирование ключевых слов и аннотаций групп текстов. Результаты кластеризации представляют групповую семантическую модель коллекции и одновременно являются мощным навигационным средством. Ключевые слова, аннотации и состав тематически однородных групп предоставляют пользователю необходимую информацию для содержательного анализа текстовой коллекции, реализации смысловой навигации в рамках коллекции и необременительного поиска в коллекции конкретной информации.

Семантическая модель текстовой коллекции (словарная модель семантики)

Адаптивный диалоговый тезаурус (АДТ), формируемый при использовании информационной технологии КЛЮЧИ К ТЕКСТАМ®, — простейшая, но весьма объёмная словарная модель семантики. Для масштабных текстовых коллекций АДТ может включать десятки и даже сотни тысяч слов. Поэтому для практически неограниченных по объёму коллекций текстов на естественном языке сделан следующий шаг — АДТ преобразуется в вычислительно и технологически эффективную ограниченную по объёму модель семантики коллекции, насчитывающую примерно две сотни слов. Модель семантики коллекции строится непосредственно на основании моделей семантики отдельных документов. Для этого словарный состав последних моделей объединяется и упорядочивается по невозрастанию суммарного веса слов в документах коллекции. Общий объём списка слов, включённых в модель семантики коллекции, ограничивается наличием хотя бы одного слова из списка почти в каждом документе коллекции.

Для организации процессов поиска информации используется семантическая модель и результаты её анализа. Во-первых, множества слов, связанных в рамках коллекции по комбинаторному критерию, формально аналогичному комбинаторному критерию для выявления ключевых слов отдельных документов. Во-вторых, множества слов, для которых характерен общий контекст в рамках слов, включённых в семантическую модель коллекции (формальный аналог решения задачи кластеризации текстовой коллекции для множеств

слов, использование которых связано по упомянутому комбинаторному критерию). Множества первой категории показывают наличие смысловых связей слов. Близкий контекст использования определяет группы слов, характерные для одной предметной области (тематики). В семантической модели коллекции удаётся формально выделить слова, маркирующие отдельные особо значимые темы (они выделяются из общих контекстов в рамках семантической модели), и слова, определяющие стилистические особенности всей коллекции (неэффективные для использования при поиске). Различия смысловых связей и контекстов использования слов существенны для идентификации важной информации.

Возможности информационной технологии КЛЮЧИ К ТЕКСТАМ®

Практическая реализация информационной технологии КЛЮЧИ К ТЕКСТАМ® — это современная наукоёмкая программно-техническая инфраструктура, обеспечивающая высокотехнологическое выполнение рассмотренных выше базовых функций содержательного анализа текстовой информации и целого набора служебных функций, позволяющих конечному пользователю комфортно использовать базовые функции для решения практических задач поиска и анализа текстовой информации.

К числу наиболее существенных служебных функций относятся:

- преобразование в текстовый формат вновь появляющихся первичных текстовых информационных ресурсов в форматах doc, docx, rtf, ps, pdf, автоматический вычислительный анализ вновь появляющихся первичных текстовых информационных ресурсов и соответствующее пополнение хранилища наукоёмких вторичных информационных ресурсов, используемого информационной технологией КЛЮЧИ К ТЕКСТАМ®;
- интеграция с внешними базами данных и информационными системами на уровне обмена XML-файлами для получения из них метаинформации о документах и для передачи внешним системам результатов поиска и анализа текстовой информации;
- вычислительное распознавание формализованных во внешней «партнёрской» информационной системе документов и использование результатов распознавания в качестве дополнительной характеристики документов (метаинформация о типе документа);
- формирование набора параметров для формирования пользователем зоны поиска по метаинформации по структуре XML-файла, описывающего метаинформацию о документах, и перечню типов формализованных документов;
- обеспечение возможности самостоятельного формирования пользователем зоны поиска в рамках хранилища наукоёмких вторичных информационных ресурсов на основании метаинформации о вычислительно семантически проиндексированных документах;
- вычислительное выделение содержательной неформализованной части в формализованных документах;
- личная и ролевая идентификация пользователей, ограничение прав доступа пользователей в соответствии с идентификацией, возможность практически неограниченного по времени сохранения индивидуальных настроек и конкретного сеанса работы пользователя.

Практическое применение технологии КЛЮЧИ К ТЕКСТАМ® или возможные сервисы для пользователей

В настоящее время информационная технология КЛЮЧИ К ТЕКСТАМ® используется в государственном учреждении «Дирекция целевой научно-технической программы» Министерства образования и науки РФ для технологической реализации набора разнообразных сервисов

смыслового поиска и анализа текстов. Основные предоставляемые служебные и базовые сервисы:

- автоматическое пополнение хранилища вторичных наукоёмких ресурсов, характеризующих тематику и содержание документов, по результатам вычислительного анализа документов, поступающих в систему экспертиз Дирекции;
- интеграция с базой данных системы экспертиз на уровне обмена XML-файлами для получения метаинформации о документах и для передачи системе экспертиз результатов поиска и анализа текстовой информации в рамках определённых сценариев совместного функционирования;
- вычислительное распознавание в результате поиска по шаблонам типовых документов системы экспертиз для их включения/исключения (по желанию пользователя) в процедуры смыслового анализа и поиска текстовой информации;
- формирование «тревожного списка» для заявочных и отчётных документов, поступающих в систему экспертиз Дирекции, по результатам вычислительного поиска содержательно подобных документов. Документы из «тревожного списка» могут быть проанализированы специалистами для оценки наличия прямых содержательных заимствований и вычислительно для выявления наличия прямых текстуальных заимствований (плагиата);
- формирование по результатам вычислительного поиска списков документов, которые ввиду содержательного и тематического сходства с заявочными или отчётными материалами целесообразно использовать при проведении экспертизы конкретных документов;
- предоставление пользователю агрегированного описания тематики и содержания отдельных документов (списков ключевых слов и аннотаций) или сформированных пользователем коллекций;
- технологическое обеспечение решения разнообразных аналитических задач.

Формирование «тревожного списка» — мощное технологическое средство борьбы с текстуальными и содержательными заимствованиями.

В задачах выявления плагиата и скрытого цитирования технология КЛЮЧИ К ТЕКСТАМ® позволяет:

- находить документы, содержательно (не текстуально) похожие на заданный в качестве образца документ;
- выявлять в документе наличие прямых (текстуальных) заимствований из других документов;
- идентифицировать фрагменты, заимствованные из других документов, и источники заимствований.

Эти возможности реализуются для анализа документа в целом или для отдельных фрагментов документа (глав, разделов, страниц и т.п. по выбору пользователя или в рамках стандартных сценариев). Допустимый объём прямого текстуального заимствования — параметр, управляемый пользователем.

Уникальными характеристиками технологии КЛЮЧИ К ТЕКСТАМ® для задач выявления плагиата являются:

- возможность выявления содержательных (не текстуальных) заимствований;
- нечувствительность выявления текстуальных заимствований к традиционным методам маскировки плагиата (замена слов их синонимами, изменение порядка слов, предложений, абзацев в заимствованном тексте).

Технология КЛЮЧИ К ТЕКСТАМ® также может быть эффективно использована для выявления плагиата и его источников в диссертационных работах, в научных публикациях, в методических материалах и в студенческих работах.

Основные типы аналитических задач, эффективно решаемых с помощью информационной технологии КЛЮЧИ К ТЕКСТАМ®:

- выявление основных тематических направлений исследований и динамики их изменения в результате вычислительного выявления наличия и состава тематически однородных групп в коллекциях документов, например, в отчётной документации, соответствующей определённому приоритетному направлению или критической технологии;
- формирование системы терминов для конкретных тематических направлений, приоритетных направлений и/или критических технологий в результате вычислительного формирования и анализа модели семантики коллекций документов;
- получение статистических характеристик проводимых исследований в результате комбинированного использования отбора документов по метаинформации и смыслового поиска/анализа текстовых документов и смысловой навигации по коллекциям документов.

Для задач управления научной деятельностью значительный практический интерес представляет сопоставление решения аналитических задач на массивах заявочной и отчётной документации с решениями аналогичных задач на массивах научных публикаций в соответствующих предметных областях. Это позволяет не только на технологическом уровне решить задачи выявления заимствований научных результатов (чужих и собственных), но и предоставляет объективные характеристики предметных областей и тенденций их развития в научных публикациях и в финансируемых НИР-ОКР на базе семантических моделей соответствующих текстовых коллекций.

Программное ядро технологии КЛЮЧИ К ТЕКСТАМ®

1. Программа формирования семантических информационных ресурсов, соответствующих коллекции текстовых документов (свидетельство об официальной регистрации программ для ЭВМ № 2007613580), правообладатели и авторы Крейнес М.Г., Афонин А.А.

2. Программа формирования лингвистических информационных ресурсов, соответствующих коллекции текстовых документов (свидетельство об официальной регистрации программ для ЭВМ № 2007613582), правообладатели и авторы Крейнес М.Г., Афонин А.А.

3. Программа формирования семантических информационных ресурсов, соответствующих одному текстовому документу (свидетельство об официальной регистрации программ для ЭВМ № 2007613581), правообладатели и авторы Крейнес М.Г., Афонин А.А.

4. Программа лемматизации текстовых документов (свидетельство об официальной регистрации программ для ЭВМ № 2007613675), правообладатели и авторы Крейнес М.Г., Афонин А.А.

5. Программа поиска документов в коллекции, для которой сформированы лингвистические и семантические информационные ресурсы (свидетельство об официальной регистрации программ для ЭВМ № 2007613583), правообладатели и авторы Крейнес М.Г., Афонин А.А.

6. Программа анализа семантического и тематического соответствия документов в коллекции, для которой сформированы лингвистические и семантические информационные ресурсы (свидетельство об официальной регистрации программ для ЭВМ № 2007613673), правообладатели и авторы Крейнес М.Г., Афонин А.А.

7. Программа кластеризации коллекции документов, для которой сформированы лингвистические и семантические информационные ресурсы (свидетельство об официальной регистрации программ для ЭВМ № 2007613677), правообладатели и авторы Крейнес М.Г., Афонин А.А.

8. Программа аннотирования документа или группы документов (свидетельство об официальной регистрации программ для ЭВМ № 2007613674), правообладатели и авторы Крейнес М.Г., Афонин А.А.

9. Программа формирования списка ключевых слов тематически однородной группы документов (свидетельство об официальной регистрации программ для ЭВМ № 2007613584), правообладатели и авторы Крейнес М.Г., Афонин А.А.

10. Программа семантической и тематической навигации по коллекции документов (свидетельство об официальной регистрации программ для ЭВМ № 2007613676, правообладатели и авторы Крейнес М.Г., Афонин А.А.)

Патенты, защищающие информационную технологию КЛЮЧИ К ТЕКСТАМ®:

1. Патент на полезную модель № 60751 «Система формирования лингвистических данных для поиска и анализа текстовых документов». Патентообладатели и авторы Крейнес М.Г., Афонин А.А.

2. Патент на полезную модель № 62263 «Система формирования семантических данных для поиска и анализа текстовых документов». Патентообладатели и авторы Крейнес М.Г., Афонин А.А.

3. Патент на полезную модель № 80597 «Система построения агрегированного интегрального представления знаний о тематике и содержании коллекции текстовых документов». Патентообладатели и авторы — Крейнес М.Г., Афонин А.А.