Институт системного анализа
Российской академии наук

Exactus Expert

Реляционно-ситуационные методы в задачах анализа научно-технических текстов

д.ф.-м.н., проф. Осипов Г.С.
к.т.н., доцент Тихомиров И. А.

+7 (499) 135 04 63

117312, Москва,
пр-т 60-летия
Октября, 9

Положение дел

Научно-технический текст это написанная по определенным правилам публикация, отражающая результаты научно-технической деятельности. Типичные представители научных текстов – статьи в научных журналах, трудах конференций, научно-технические отчеты, авторефераты и диссертации и тд.

Разработки в области анализа научных текстов:

- E-library
- Google Scholar
- Scopus
- Exactus Expert
- …

Задачи анализа научно-технических текстов

Задачи можно сгруппировать следующим образом:

- Формирование коллекций и баз данных научно-технических текстов.
- Индексация коллекций.
- Предоставление пользователям доступа к коллекциям и функциям аналитической обработки текстов:
  - Поиск документов по запросу (семантический, по ключевым словам, по образцу, по полям, фразовый поиск и тд).
  - Анализ связанности (индексы цитирования, Хирш и тд.).
  - Анализ качества текстов.
  - Определение перспективных и бесерспективных направлений исследований.
  - Выявление возможного дублирования, плагиата, частичных заимствований.
  - Классификация и кластеризация документов.
  - Выявление коллективов исследователей и научных школ.
  - Формирование пользовательских коллекций документов.
  - …

Применяемые методы

Фактически, все современные подходы к поиску и анализу научных текстов учитывают статистические закономерности, такие как:

- TF*IDF веса термов,
- Page Rank, и др.

Зачастую декларируется возможность **семантического поиска и анализа текстов**, хотя на самом деле применяются простейшие методы морфологического анализа, статистика, попарная встречаемость слов в текстах или надстройки в виде тезаурусов и онтологий!

Цель разработчиков систем анализа научных текстов

Первоочередная цель – разработка новых методов поиска, извлечения и анализа текстовой информации, для построения развитых аналитических инструментов, решающих важные прикладные задачи при большом количестве данных(big data) с высоким уровнем качества.

Реляционно-ситуационная модель текста

- Реляционно-ситуационная модель текста опирается на развитые научные теории и формальные системы:
  - Коммуникативная грамматика(Золотова Г.А.).
  - Неоднородные семантические сети (Осипов Г.С.).
- Реляционно-ситуационные методы объединеняют статистические и лингвистические методы обработки текста.
- При анализе текстов используются словари, тезаурусы, онтологии и лингвистические базы знаний.
- Решения ИСА РАН ориентированы на семантический поиск и анализ текстов, а также визуализацию результатов аналитической обработки.

Вышеперечисленное обеспечивает высокое качество и эффективность решения поисково-аналитических задач.

Коммуникативная грамматика русского языка

- Коммуникативная грамматика русского языка [Золотова и др., 2004] основана на связи синтаксиса и семантики.
- Синтаксис имеет дело с осмысленными единицами, несущими обобщенный категориальный смысл в конструкциях разной степени сложности. Эти единицы характеризуются взаимодействием морфологических, семантических и функциональных признаков. Они получили название ***синтаксем***.

Коммуникативная грамматика: именные синтаксемы

- Исходная точка движения (*выйти из комнаты*) .
- Производитель действия (*утвержден президентом*) .
- Владелец отчуждаемого объекта (*лишить субъект РФ права на что-л.)*.
- Местонахождение (*войска сосредоточены в районе Багдада*).
- Размер, исчисляемая мера величины (*измерять информацию в байтах*).
- … (описано 85 значений)

Типы семантических отношений: примеры

Один компонент называет местонахождение другого компонента *(В Париже с успехом прошли гастроли Большого театра)*.
Один компонент обозначает причину проявления другого компонента спустя какое-то время. *(Появление регулировщика на дороге приводит к затруднению движения)*.
Один компонент обозначает сопровождающее другой компонент действие, сопутствующий предмет, сопровождающее лицо (*Президент встретился с коллегой в своей загородной резиденции*).
Один компонент выражает отношение владения другим компонентом (*Абрамовичу принадлежит ф/клуб «Челси»)*.

Exactus Expert

От значения слов к значению высказываний

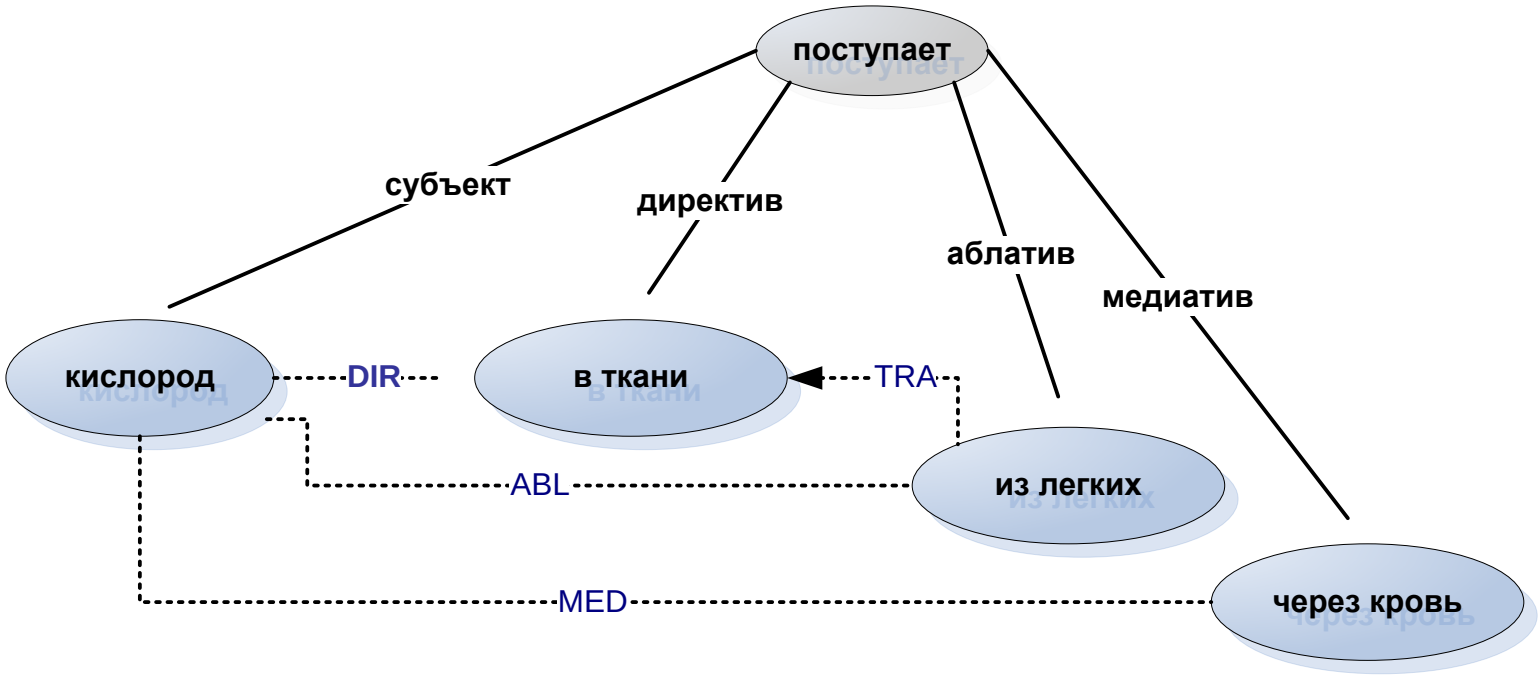
Значение высказывания определяется множеством значений входящих в него минимальных семантико-синтаксических единиц и семейством отношений на них.

Реляционно-ситуационная модель текста

- $M = < S, T_s, R , I_s >$
- S – множество синтаксем, $S = \{s_1, s_2, \ldots, s_n\}$, s_i – синтаксема
- R – семейство отношений на множестве синтаксем, $R \subseteq S \times S$
- T_s – типы синтаксем
- $I_s : S \rightarrow T_s$
- $s = < W, P, \tau >$
- $\tau \in T_s$, $T_s = \{‘p’, ’n’\}$
- W – слово
- P – свойства синтаксемы, включая ее категориальный класс, предлоги и иные морфологические свойства
- τ – тип синтаксемы (‘*p*’ -предикатное слово; ‘*n*’ -именная синтаксема)
- $R = \{(s^1, s^2)\}$ семейство бинарных отношений, состоит из трех типов:
  - R_p – отношение между предикатным словом и именной синтаксемой
  - R_n – отношение между именными синтаксемами
  - R_c – отношение для представления кореференции

Пример сети высказывания

Пример: кислород поступает в ткани из легких через кровь



Возникающая в результате конструкция является алгебраической системой с множеством значений в качестве основного множества, семейством отношений на основном множестве и семейством правил в качестве функций.

Exactus Expert

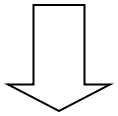
Уровни анализа текстов

- Графематический анализ.
- Морфологический анализ.
- Синтаксический анализ.
- Семантический анализ.

Графематический анализ

Основная задача графематического анализа – разбиение текста на слова.

The mother brings her son to school.

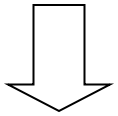


the	mother	brings	her	son	to	school

Морфологический анализ

Основная задача – определить характеристики слов (род, число, падеж, форма, спряжение и т.д.).

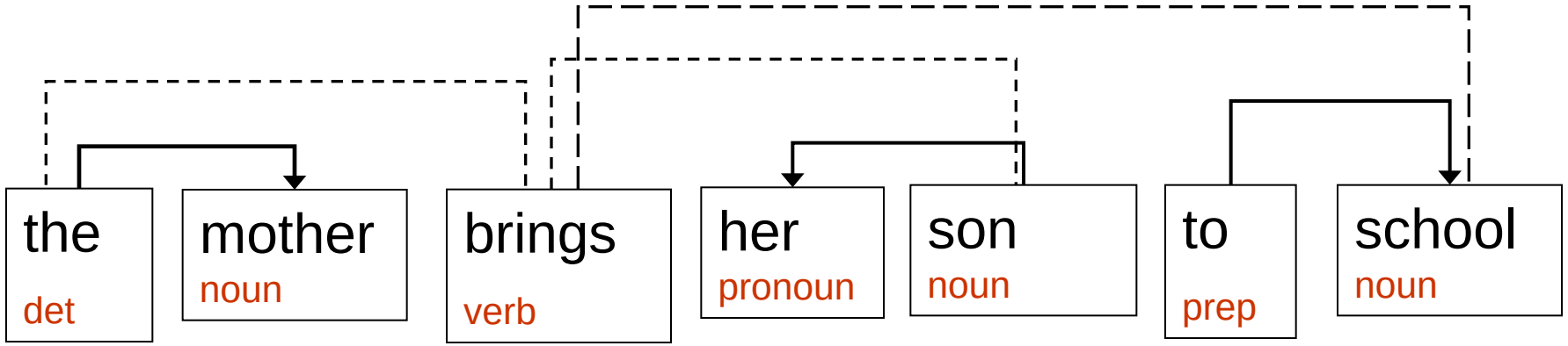
the	mother	brings	her	son	to	school



the	mother	brings	her	son	to	school
det	noun	verb	pronoun	noun	prep	noun

Синтаксический анализ

Основная задача – установить синтаксические связи между словами.



Семантический анализ

Семантический анализ это переход от языкозависимых конструкций к инвариантным к языку значениям и конструкциям.

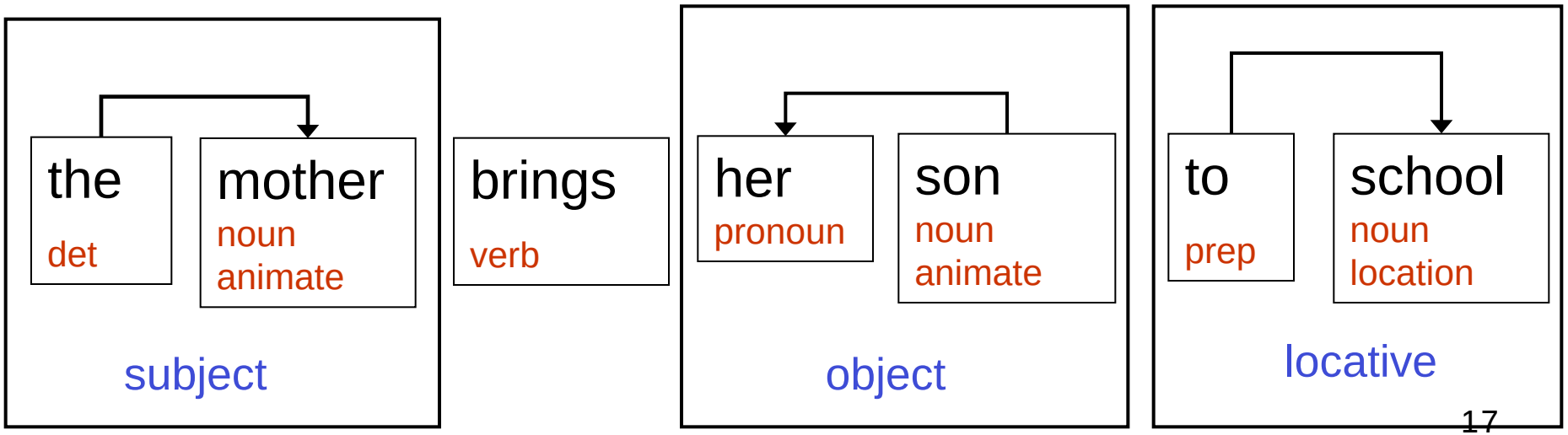
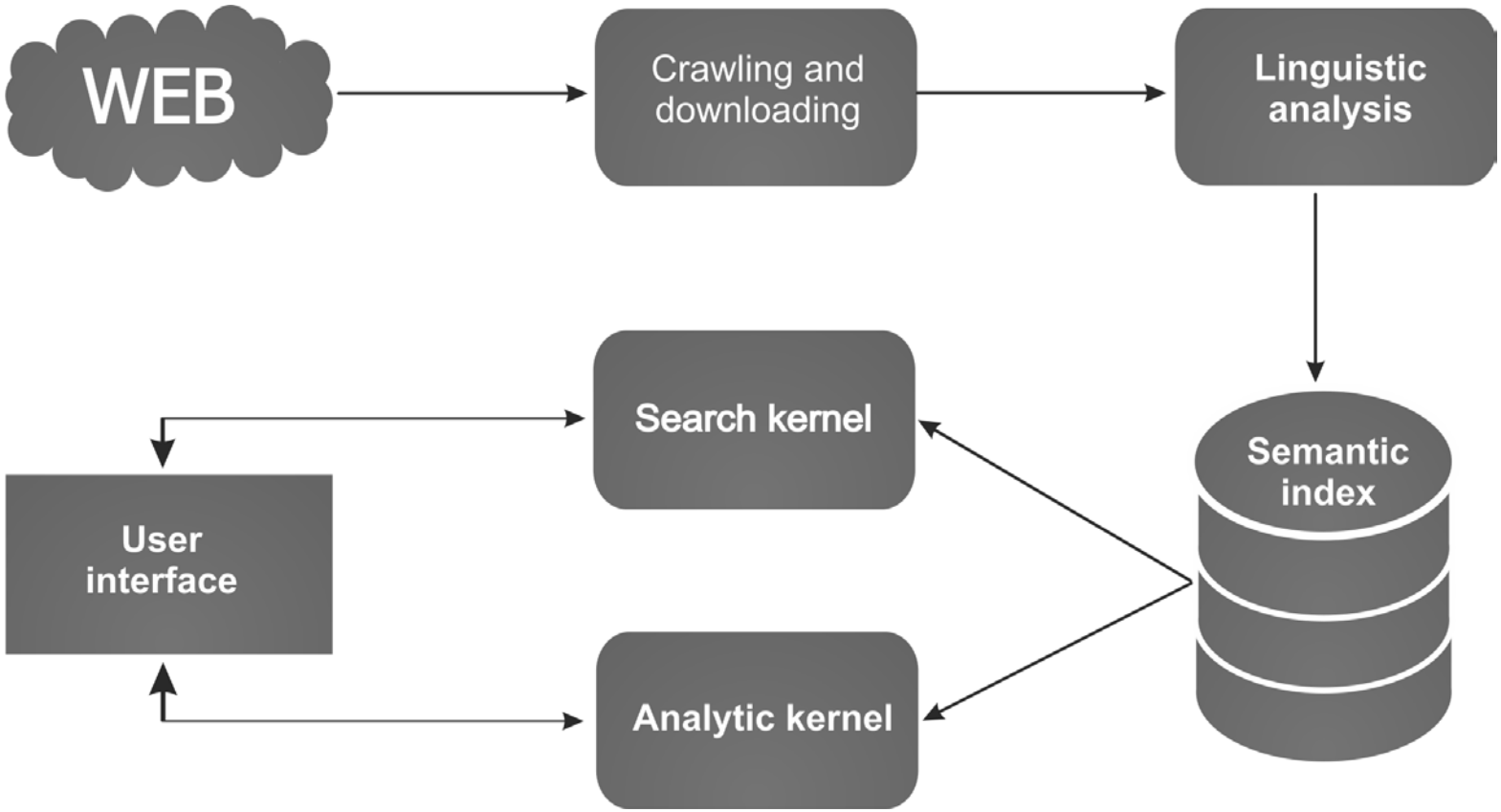


Схема работы поисково-аналитической машины Exactus Expert

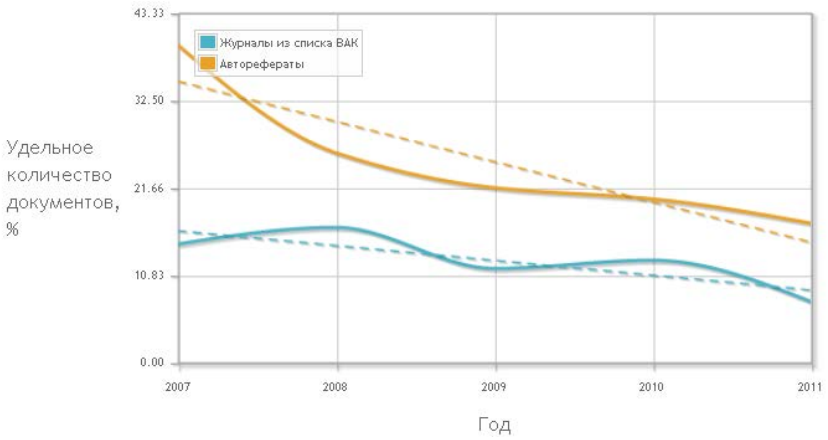


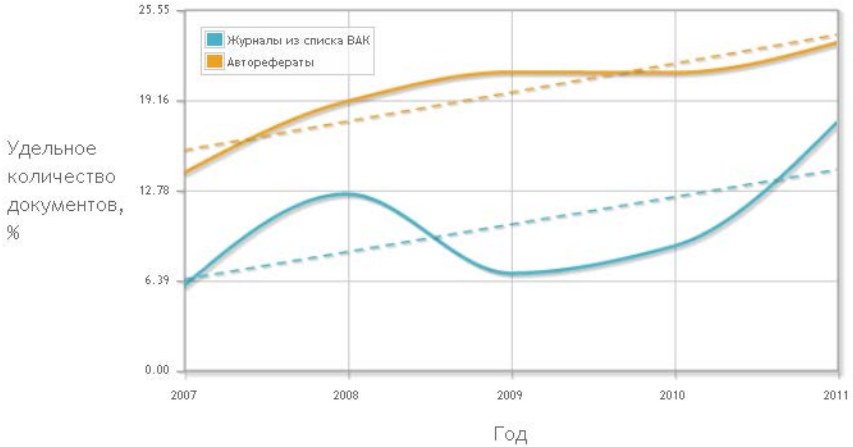
Реализация поисково- аналитической машины Exactus Expert

- Реализация на вычислительном кластере под управлением Linux Debian.
- Распределённые вычисления обеспечивают стабильность работы при высокой нагрузке и масштабируемость.

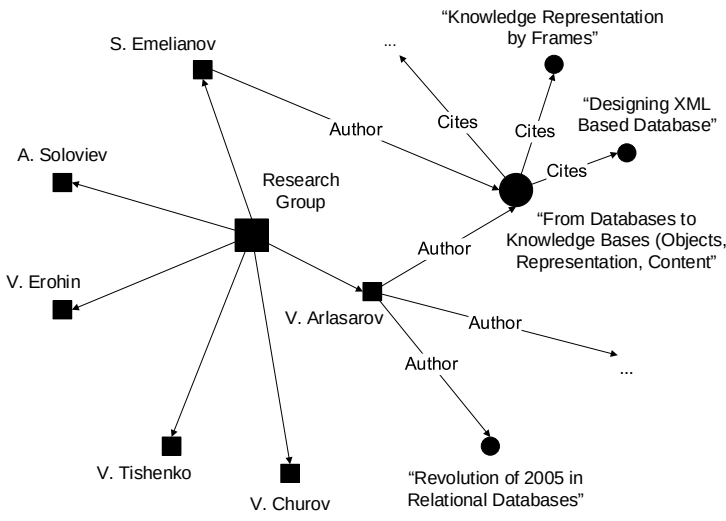
Примеры анализа научных текстов

Анализ публикационной активности по теме «экспертные системы»:





Граф связанности публикаций



Пример оценки качества научной публикации:

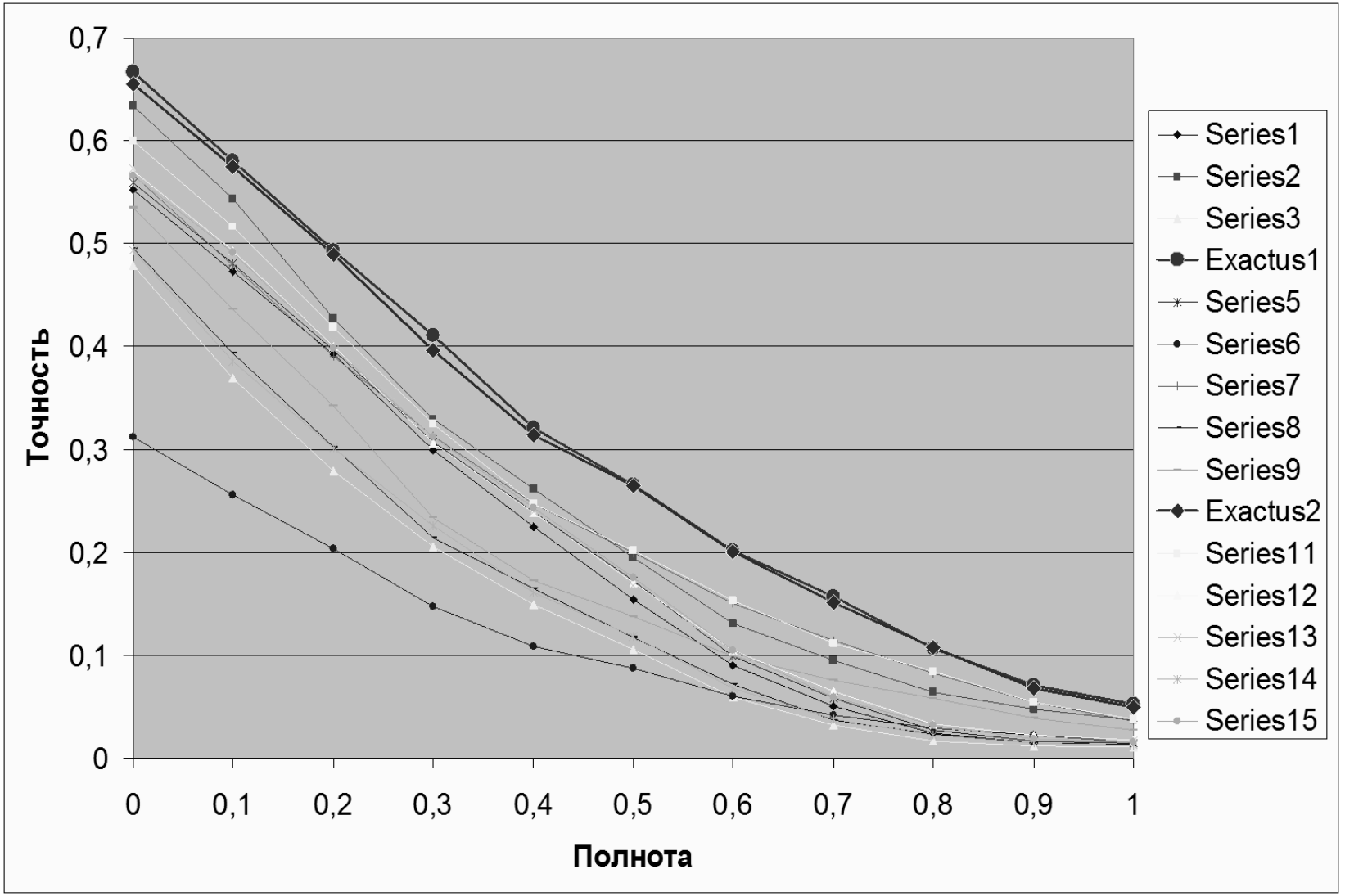
Индекс соответствия текста формальным требованиям: 5 (0..5).

- Текст содержит 38% общенаучной лексики.
- Текст содержит 0% ненаучной лексики.
- Список литературы присутствует.
- Постановка проблемы присутствует. Коэффициент достоверности: 0.85 (0..1).
- Описание методов присутствует. Коэффициент достоверности: 0.86 (0..1).
- Выводы присутствуют. Коэффициент достоверности: 0.87 (0..1).

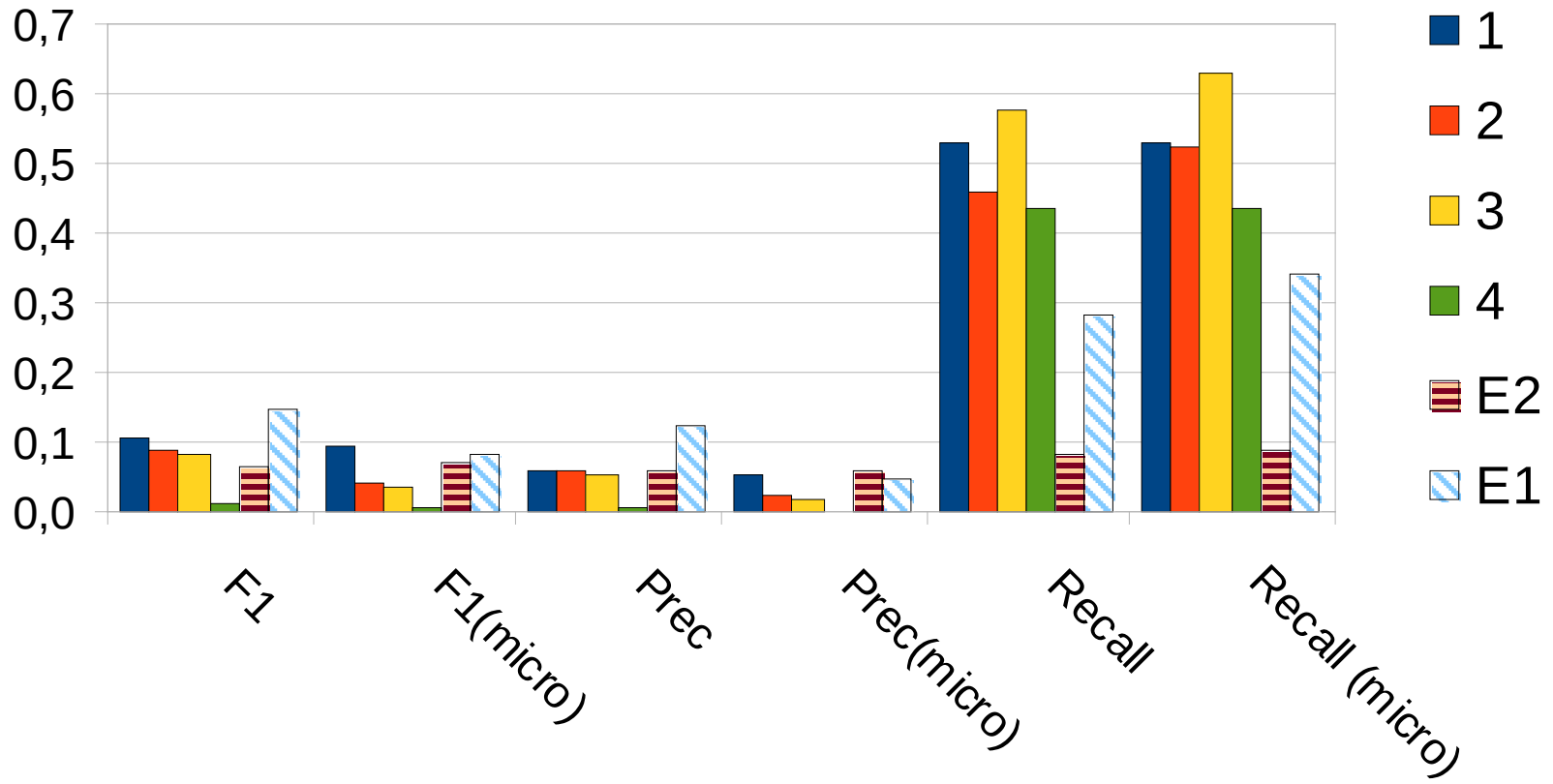
Степень логической и семантической правильности (отсутствие дефектов): 4 (0..5).

- Количество нарушений падежного согласования: 0 (0..5).
- Низкая степень нарушения синтаксической связности.
- Количество нарушений согласования однородных существительных и управляющего слова: 2.
- Содержание плеоназмов - умеренное.

Оценка поисковых машин на РОМИП'2008 – первое место у Exactus

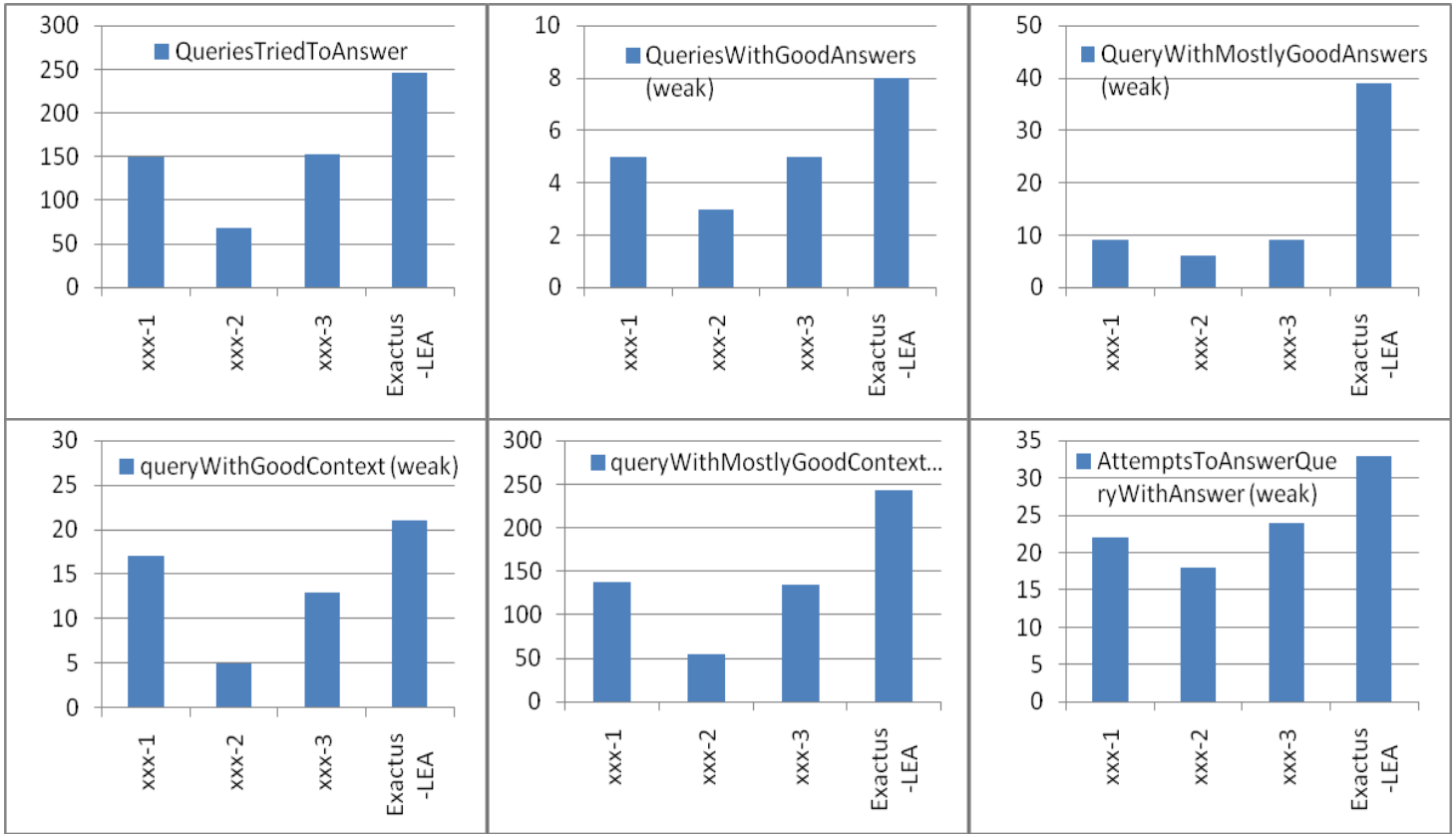


Оценка алгоритмов классификации Web-страниц на РОМИП'2009



Алгоритм EXACTUS (E1 и E2) показал лучшие результаты по точности и F-мере

Оценка алгоритмов вопросно-ответного поиска на РОМИП'2010



Алгоритм EXACTUS показал лучшие результаты по всем параметрам

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ПОЛЕЗНУЮ МОДЕЛЬ

№ 62719

СИСТЕМА СЕМАНТИЧЕСКОГО МЕТАПОИСКА, АНАЛИЗА И ИНДЕКСАЦИИ ИНФОРМАЦИИ

Патентообладатель(ли): ***ИНСТИТУТ СИСТЕМНОГО АНАЛИЗА РОССИЙСКОЙ АКАДЕМИИ НАУК (RU)***

Автор(ы): ***Осипов Геннадий Семенович (RU), Тихомиров Илья Александрович (RU), Смирнов Иван Валентинович (RU)***

Заявка № **2006135491**

Приоритет полезной модели **09 октября 2006 г.**

Зарегистрировано в Государственном реестре полезных моделей Российской Федерации ***27 апреля 2007 г.***

Срок действия патента истекает **09 октября 2011 г.**

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Б.П. Симонов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ИЗОБРЕТЕНИЕ

№ 2473119

СПОСОБ И СИСТЕМА СЕМАНТИЧЕСКОГО ПОИСКА ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Патентообладатель(ли): ***Учреждение Российской академии наук Институт Системного Анализа РАН (ИСА РАН) (RU)***

Автор(ы): ***Осипов Геннадий Семенович (RU), Тихомиров Илья Александрович (RU), Соченков Илья Владимирович (RU), Смирнов Иван Валентинович (RU)***

Заявка № **2011132873**

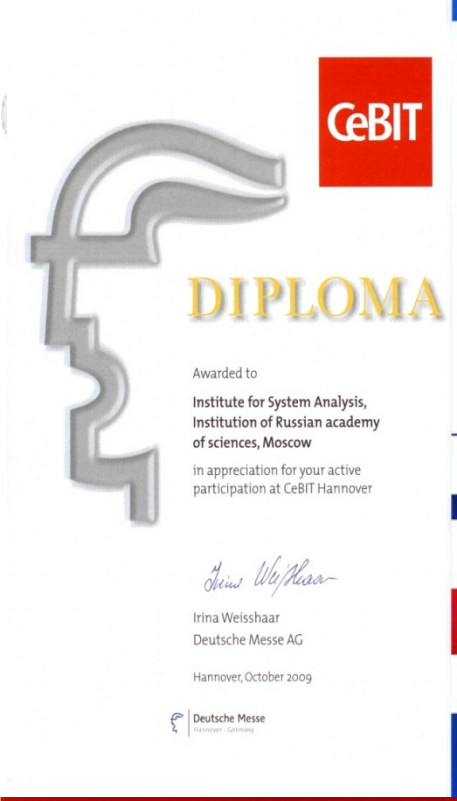
Приоритет изобретения **05 августа 2011 г.**

Зарегистрировано в Государственном реестре изобретений Российской Федерации ***20 января 2013 г.***

Срок действия патента истекает **05 августа 2031 г.**

Руководитель Федеральной службы по интеллектуальной собственности

Б.П. Симонов


CeBIT
DIPLOMA
Awarded to
Institute for System Analysis, Institution of Russian academy of sciences, Moscow
in appreciation for your active participation at CeBIT Hannover
Irina Weisshaar
Deutsche Messe AG
Hannover, October 2009
Deutsche Messe


Российская академия наук
Министерство связи и массовых коммуникаций Российской Федерации
Российский фонд фундаментальных исследований
Правительство Москвы
ДИПЛОМ
участника
Двадцать первая ежегодная выставка информационных и коммуникационных технологий
SofTool
www.softool.ru


VI Московский международный салон инноваций и инвестиций
СВИДЕТЕЛЬСТВО
участника


ДИПЛОМ
Награждается
Учреждение Российской академии наук
Институт системного анализа РАН
за активное участие
в работе экспозиции Минобрнауки России на международной выставке информационных технологий, телекоммуникаций и оргтехники CeBIT в 2011 году
Заместитель Министра образования и науки Российской Федерации
С.В.Иванец
Президент Международного Союза приборостроителей и специалистов по информационным и телекоммуникационным технологиям, академик РАН
Н.А.Кузнецов


SofTool 2011
XXII ежегодная выставка информационных и коммуникационных технологий SOFTOOL
Российская национальная премия
Конкурс лучших решений в области информационных технологий «ПРОДУКТ ГОДА»
НАГРАЖДАЕТСЯ
2 МЕСТО
ПРОДУКТ ГОДА SOFTOOL
Поисково-аналитическая машина Exactus
компании
Учреждение российской академии наук
Институт системного анализа РАН
в номинации
Интернет технологии
Академик-секретарь Отделения нанотехнологий и информационных технологий РАН
Е.П. Велихов
Председатель Организационного комитета выставки Softool, член Президиума РАН
Ю.В. Гуляев
Заместитель министра связи и массовых коммуникаций РФ
И.И. Массух


ВЫСОКИЕ ТЕХНОЛОГИИ XXI ВЕКА
HIGH TECHNOLOGY OF
ДИ
10-й ЮБИЛЕЙ
«ВЫСОК
ОРГКОМИТЕТОМ НАГРАЖДАЕТСЯ
Институт системного анализа (ИСА РАН)
ЗА УЧАСТИЕ В ВЫСТАВКЕ «ВТ-XXI 2009»
И ДОСТИЖЕНИЯ
В ОБЛАСТИ ВЫСОКИХ ТЕХНОЛОГИЙ
МИНИСТР ПРАВИТЕЛЬСТВА МОСКВЫ,
ЗАМЕСТИТЕЛЬ ПРЕДСЕДАТЕЛЯ ОРГКОМИТЕТА
Е.А. ПАНТЕЛЕЕВ
21-24 АПРЕЛЯ 2009 г.
МОСКВА, ЦВК «ЭКСПОЦЕНТР»


международный форум
МЕГАПОЛИС XXI ВЕК


Дипломъ
Институт Системного Анализа Российской Академии Наук
За участие в подготовке и проведении Международного Форума "Мегаполис: XXI век"
3-4 апреля 2008 г.

Внедрения

Решения ИСА РАН внедрены:

- Сервисы для Центрального коллектора электронных библиотек БИБКОМ (2013 год).
- Система интеллектуального поиска и анализа научных публикаций “Exactus Expert” – expert.exactus.ru (2011 год).
- Медицинский центр Банка России – медицинская электронная библиотека (внедрение 2008-2009 годы).
- Московская медицинская академия им. Сеченова - медицинская электронная библиотека (внедрение 2007 год).
- Поиск по сайту ИСА РАН - www.isa.ru (2009 год).
- Общее использование – www.exactus.ru (2008 год).
- Выполнено более 20 НИР и НИОКР для различных заказчиков – Бюджет союзного государства Россия-Беларусь, РосНаука, РосОбразование, РФФИ, программы Президиума РАН, ОНИТ РАН и др.

Медицинская электронная библиотека

Общий поиск | Журналы | Клинические руководства | Авторефераты | Метапоиск

Введите данные о статье, и/или задайте слова, содержащиеся в тексте статьи

Обновление коллекции от 10.01.2010

Выбор источников поиска

Авторы

Заглавие

Журнал: Все

Год выпуска с: 1997 по 2009

Сортировать по: релевантности

Поисковый запрос

Поиск | Оставить замечание

Вас приветствует Медицинская электронная библиотека!

На этой вкладке Вы можете выполнять поиск в коллекции медицинских журналов. Вы можете указать авторов публикации, слова, содержащиеся в названии публик и\или задать ключевые слова в строке "Поисковый запрос". Вы можете указать временной период публикации, а также тип сортировки результатов поиска: по релевантности или по дате.

Для поиска в отдельной коллекции перейдите на соответствующую вкладку.

Если вы хотите, чтобы слово обязательно присутствовало в результатах поиска, поставьте перед ним знак +, например "+астма". Если наоборот, слово не должно содержаться в результатах поиска, перед ним необходимо поставить знак -.

Используйте всплывающие подсказки по заданию параметров поиска, щелкая мы по значкам.

Контакты | Информационные сервисы

ПОИСК ПО АТРИБУТАМ

Национальный цифровой ресурс

Сервис интеллектуального поиска по атрибутам

КОНТЕКСТУМ

Консорциум

Информационная технология сбора цифрового контента

Журналы ВАК | БИБКОМ | Посмотреть тематики

Авторы

Заглавие

Годы публикации с 1970 по 2013

Поисковый запрос

+Термины | Поиск

О сайте | Контакты | Дизайн — AlterEGO | Разработка — ООО "Технологии системного анализа"

© КНИГА СЕРВИС | бибком

matandra | Зарегистрироваться

Пароль | Забыли пароль?

Войти | Запомнить меня

Рус Eng

Добавить в закладки

Exactus | Искать

РАН

УЧРЕЖДЕНИЕ РОССИЙСКОЙ АКАДЕМИИ НАУК ИНСТИТУТ СИСТЕМНОГО АНАЛИЗА РАН

On-line ресурсы института | Интернет-конференции | Виртуальная приемная директора

Поиск

Тихомиров | Найти!

НОВОСТИ | ОБ ИНСТИТУТЕ | СТРУКТУРА ИНСТИТУТА | УЧЕНЫЙ СОВЕТ | СОВЕТ МОЛОДЫХ УЧЕНЫХ | ИССЛЕДОВАНИЯ И РАЗРАБОТКИ | ПЕРИОДИЧЕСКИЕ ИЗДАНИЯ | ОБРАЗОВАНИЕ | ПАРТНЕРЫ

Искать по **Тихомиров**

Всего найдено 6 записей.

Количество строк: 20

1. Сайт ИСА РАН - Контакты СМУ
   ... *Тихомиров* Илья Александрович, к. т. н. лаборатория 0 2. тел. ...
2. Сайт ИСА РАН - Совет молодых ученых
   .... *Тихомиров* Илья Александрович, к. т. н., лаб. ...
3. Сайт ИСА РАН - Лаб. 0-2. Динамические интеллектуальные системы
   ... служебный городской: 8 499 135 04 63. *Тихомиров* Илья Александрович Ст. науч. сотр., к. т. н., комната Г 1521... ...

Exactus Expert

Интеллектуальный поиск и анализ научных публикаций

Добро пожаловать Уважаемый Гость

сменить пароль | выйти

Организация: не определена

Задачи

Мои коллекции(0)

Мой профиль

О системе

Министерство образования и науки Российской Федерации

Статистика системы

Всего пользователей 175

Подключено организаций 131

Документов в базе 1021547

Ежедневное количество поисковых запросов 41

Поиск по запросу | Тематический анализ | Оценка качества научных текстов

справка | настройки поиска | расширенный поиск

Авторы

Заглавие +Термины

Годы публикации с 2007 по 2011

Поисковый запрос

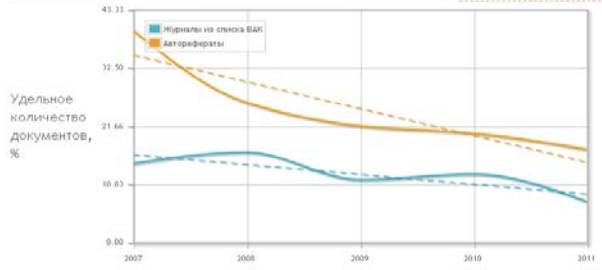
экспертные системы +Термины | Поиск

Перевести запрос

Отобразить тренды

Найдено документов: 963

Статистика по техническим наукам



Основные разработчики

Dr.Sc, prof.
Gennady Osipov

PhD
Ilya Tikhomirov

PhD
Ivan Smirnov

PhD
Olga Vybornova

Dr.Sc, prof.
Sergey Krylov

PhD
Olga Zavjalova

Researcher
Ilya Sochenkov

PhD-student
Alexander Shvets

PhD-student
Dmitry Devyatkin

PhD-student
Artem Shelmanov

PhD-student
Roman Suvorov

Researcher
Alexander Leshkin

Контакты

www.exactus.ru
expert.exactus.ru

Институт системного анализа
Российской академии наук
лаборатория 0-2
«Динамические интеллектуальные системы»
117312, Москва,
пр. 60-летия Октября, 9
тел.: +7 (499) 135-04-63
e-mail: tih@isa.ru
Тихомиров Илья Александрович