Exactus

Всегда то, что искали.

ИНТЕЛЛЕКТУАЛЬНАЯ МЕТАПОИСКОВАЯ СИСТЕМА EXACTUS

www.exactus.ru

Особенности поискового алгоритма и архитектуры Exactus.

Тихомиров Илья Александрович
к.т.н.

Российский семинар по Оценке Методов Информационного Поиска РОМИП-2007.
(Переславль-Залесский, 18 октября)

www.isa.ru
РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ СИСТЕМНОГО АНАЛИЗА
117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Exactus

Состояние дел в области поисковых алгоритмов

- Статистические алгоритмы поиска (с учетом морфологии).
- Лингвистические алгоритмы поиска (без учета статистики).

Цель разработчиков Exactus объединение статистических и лингвистических методов поиска.

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Exactus

Особенности алгоритма поиска Exactus (1)

Учет статистических характеристик текста:

- TF*IDF веса термов (с некоторой модификацией).
- Значимость фрагментов текстов (заголовки, разметка, удаленность от начала документа и т.д.)

- Примечательный момент – в алгоритме не используется PageRank.

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Exactus

Особенности алгоритма поиска Exactus (2)

Учет лингвистических характеристик единиц текста:

- Значения синтаксем (наиболее близкое понятие – семантические валентности).
- Сочетаемость синтаксем в конкретном предложении.

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Exactus

Особенности алгоритма поиска Exactus (3)

Пример: «Митрофанушка не знал, что говорит прозой».

«Митрофанушка» – субъект (значение синтаксемы определено через глагол и признаки: личное существительное именительного падежа, принадлежащее к классу имен собственных).

«Прозой» - медиатив (значение синтаксемы определено через глагол и признаки: существительное в творительном падеже, принадлежащее к классу признаковые).

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Exactus

Особенности алгоритма поиска Exactus (4)

1. Поиск возможен только в проиндексированной коллекции.
2. На этапе индексации производится преобразование документов к внутреннему формату Exactus, обсчет TF*IDF.
3. Производится синтаксический и семантический анализ текстов (выявление подчинения синтаксем и их значений).
4. Полученные в результате анализа данные укладываются в линейные упорядоченные списки.
5. Поиск представляет собой слияние линейных упорядоченных писков.

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Особенности архитектуры Exactus

- Модули расположены на узлах кластерной установки.
- Управление задачами осуществляется посредством PVM-машины (Parallel Virtual Machine).
- Модули разделены на два типа: основные (лингвистические процессоры, индексаторы и т.д.) и вспомогательные (агрегаторы, синхронизаторы и т.д.).
- Система является кросс-платформенной, код написан на C и C++. Поиск и индексация для РОМИП'2007 производились под управлением Linux Debian 4.0.
- Экспериментальная установка состоит из 8-и узлов кластера пиковой производительностью 100 Gigaflops.
- В качестве вычислительных узлов используются обычные персональные компьютеры, объединенные в стойку.
- Для взаимодействия узлов используется Gigabit Ethernet.

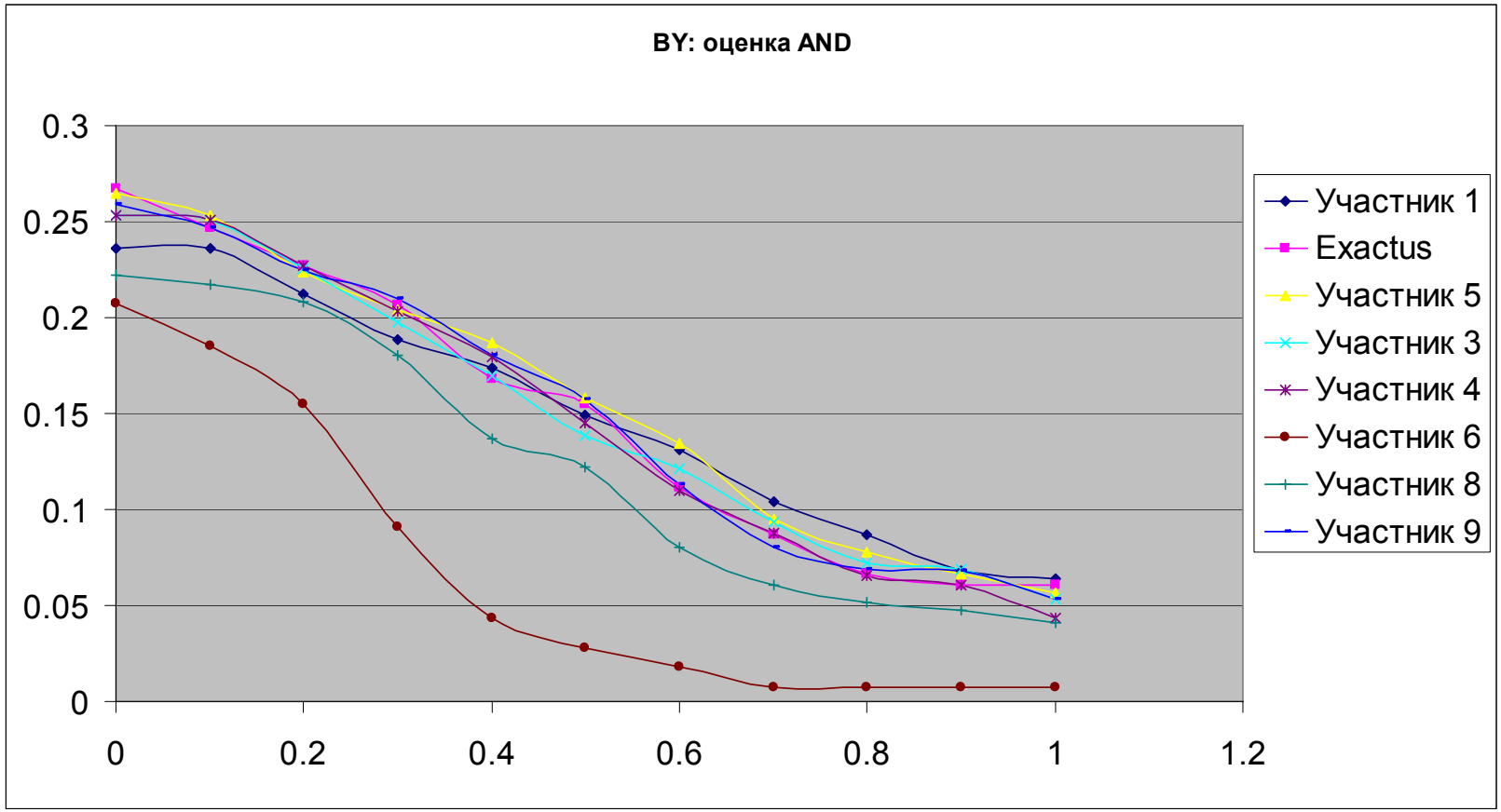
117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Краткий анализ результатов участия в РОМИП-2007 (1)

- Exactus принимал участие в поиску по коллекциям BY и LEGAL.
- Наилучшие результаты достигнуты системой в AND-оценке по точности.
- Хорошие оценки достигнуты по другим показателям.
- Отсутствие Page Rank в алгоритме Exactus не привело к отставанию от других систем в точности и полноте поиска.

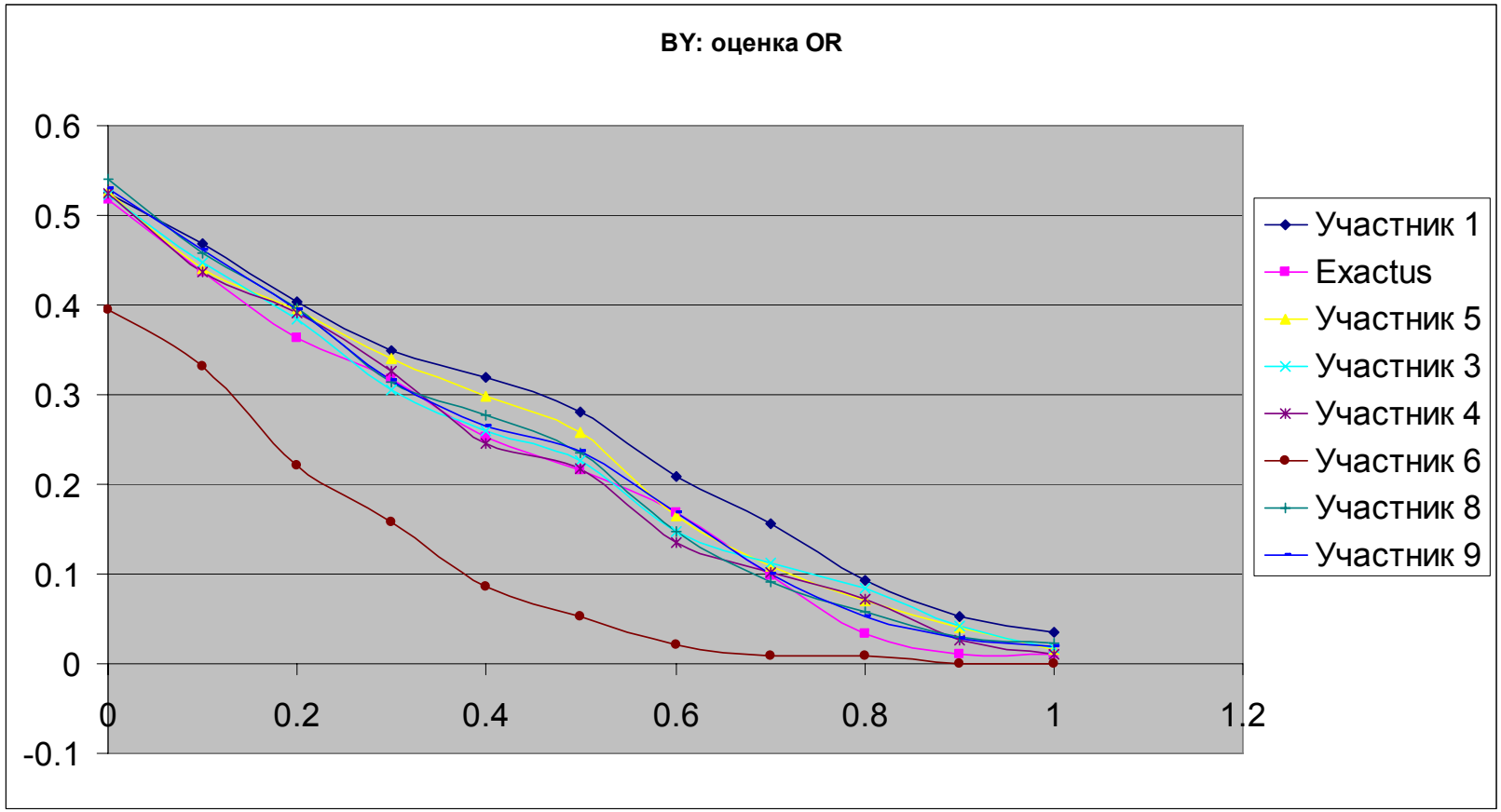
117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Краткий анализ результатов участия в РОМИП-2007 (2)



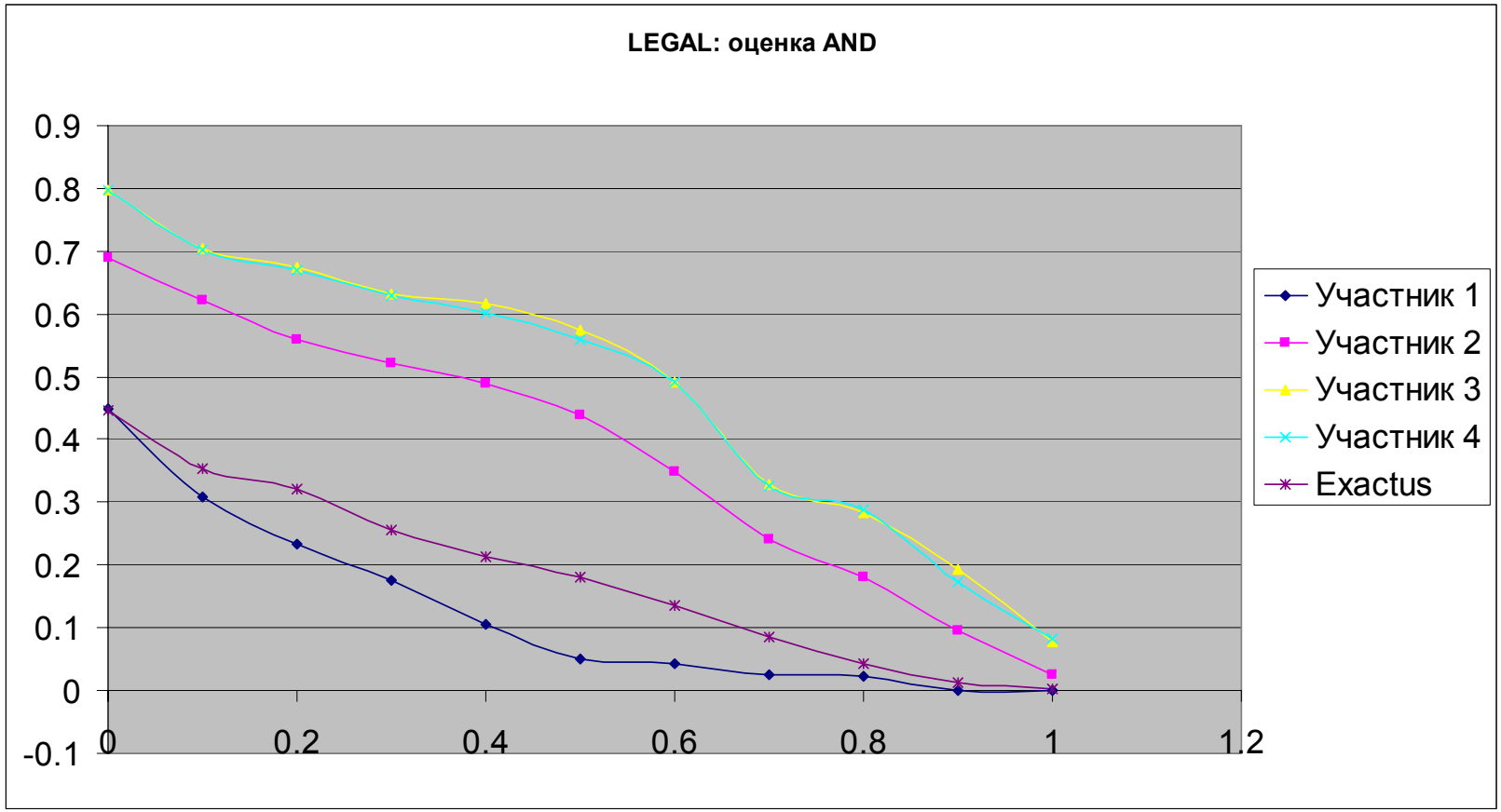
117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Краткий анализ результатов участия в РОМИП-2007 (3)



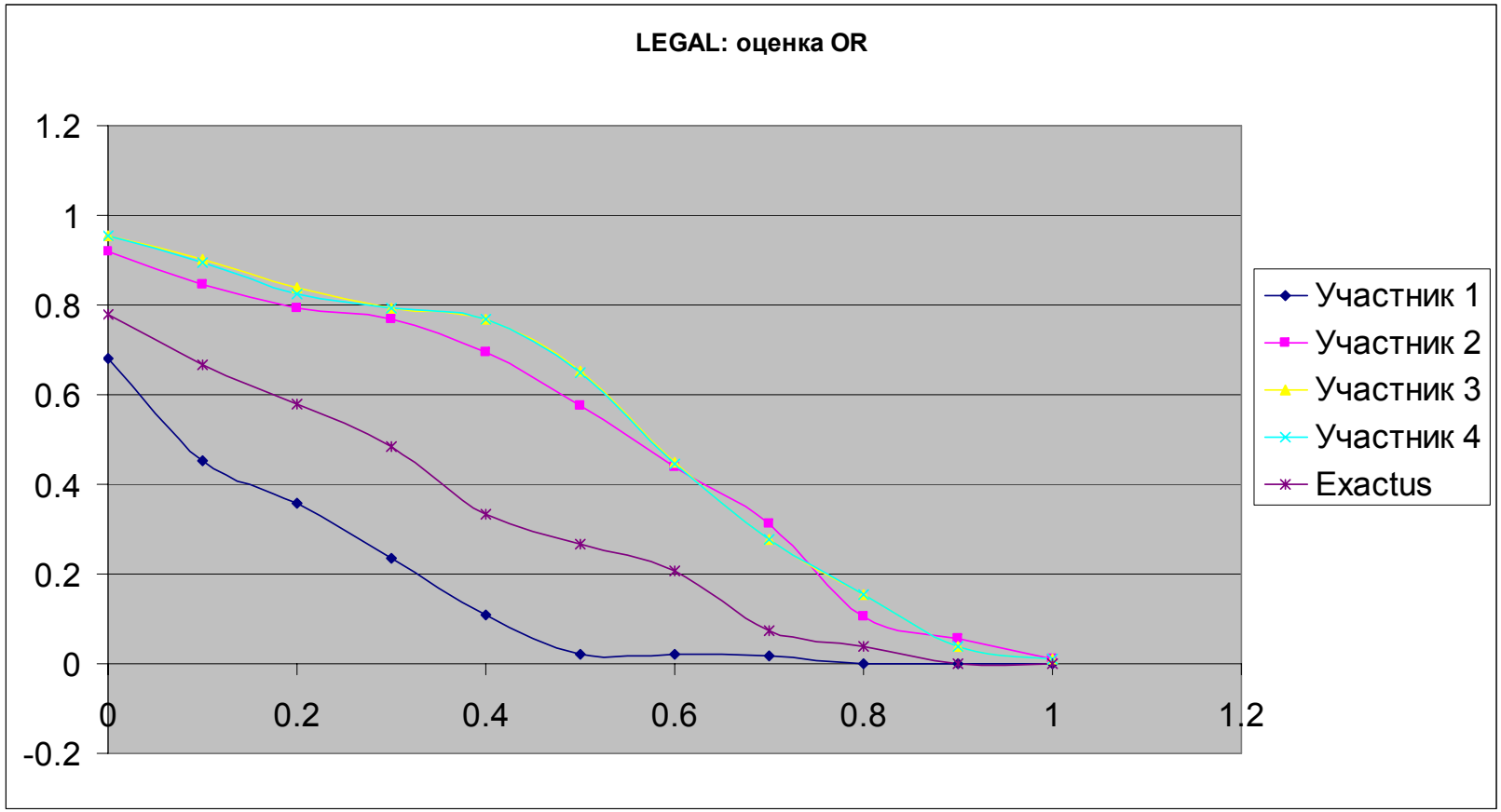
117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Краткий анализ результатов участия в РОМИП-2007 (4)



117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Краткий анализ результатов участия в РОМИП-2007 (5)



117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Краткий анализ результатов участия в РОМИП-2007 (6)

- Полученные на РОМИП результаты показывают перспективность симбиоза лингвистических и статистических алгоритмов поиска и возможность их применения в реальных условиях.
- Скорость поиска Exactus сравнима с современными поисковыми машинами на больших объемах данных (не более 2х секунд на любой запрос по коллекции РОМИП).
- Индексация и лингвистический анализ, по-прежнему, остаются узким местом Exactus. Единственный путь преодоления барьеров скорости анализа - использование современных вычислительных систем и параллельных вычислений.

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Направления развития Exactus

- Включение разновидности Page Rank в алгоритм поиска Exactus.
- Использование каталога ресурсов для улучшения результатов индексации (повышение рэйтингов).
- Совершенствование метода трансформации семантической сети текста в линейные упорядоченные списки для последующего использования при поиске.

- В перспективе, команда Exactus планирует расширить свое участие в дорожках РОМИП и проверить свои новые алгоритмы каталогизации и контекстно-зависимого аннотирования.

117312, Москва, В-312,
проспект 60-летия Октября, 9, ИСА РАН
тел./факс (495) 1350463, E-mail: matandra@isa.ru

Exactus

СПАСИБО ЗА ТЕРПЕНИЕ!

Вопросы принимаются по адресу:

Институт системного анализа РАН
117312, Москва, пр-т 60-летия Октября, 9
Тел./факс: (495) 135-04-63
e-mail: matandra@isa.ru
Тихомиров Илья Александрович