НОВЫЕ ТЕХНОЛОГИИ СЕМАНТИЧЕСКОГО ПОИСКА ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ

Смирнов Иван Валентинович
к.ф.-м.н., доцент
Заведующий лабораторией
ИСА ФИЦ ИУ РАН
Москва

2016

Методы обнаружения текстовых заимствований

- Метод «шинглов»
  - Текст разбивается на перекрывающиеся «шинглы» - последовательности из слов
  - Заимствования ищутся на основе совпадения шинглов текстов
  - Эффективен для «copy&paste» заимствований
  - Неэффективен для заимствований с сильным перефразированием

Методы обнаружения текстовых заимствований

- Метод «мешка слов»
  - Для фрагмента текста строится вектор взвешенных слов (терминов)
  - Заимствования ищутся путем вычисления косинусообразной меры близости векторов фрагментов разных текстов
  - Малоэффективен при значительных перефразированиях и поиске в больших коллекциях текстов
  - Промышленные реализации отсутствуют

Другие методы

- Поиск подстрок
- На основе совпадения последовательности цитирований (для научных текстов)
- Учет авторского стиля
- Гибридные методы

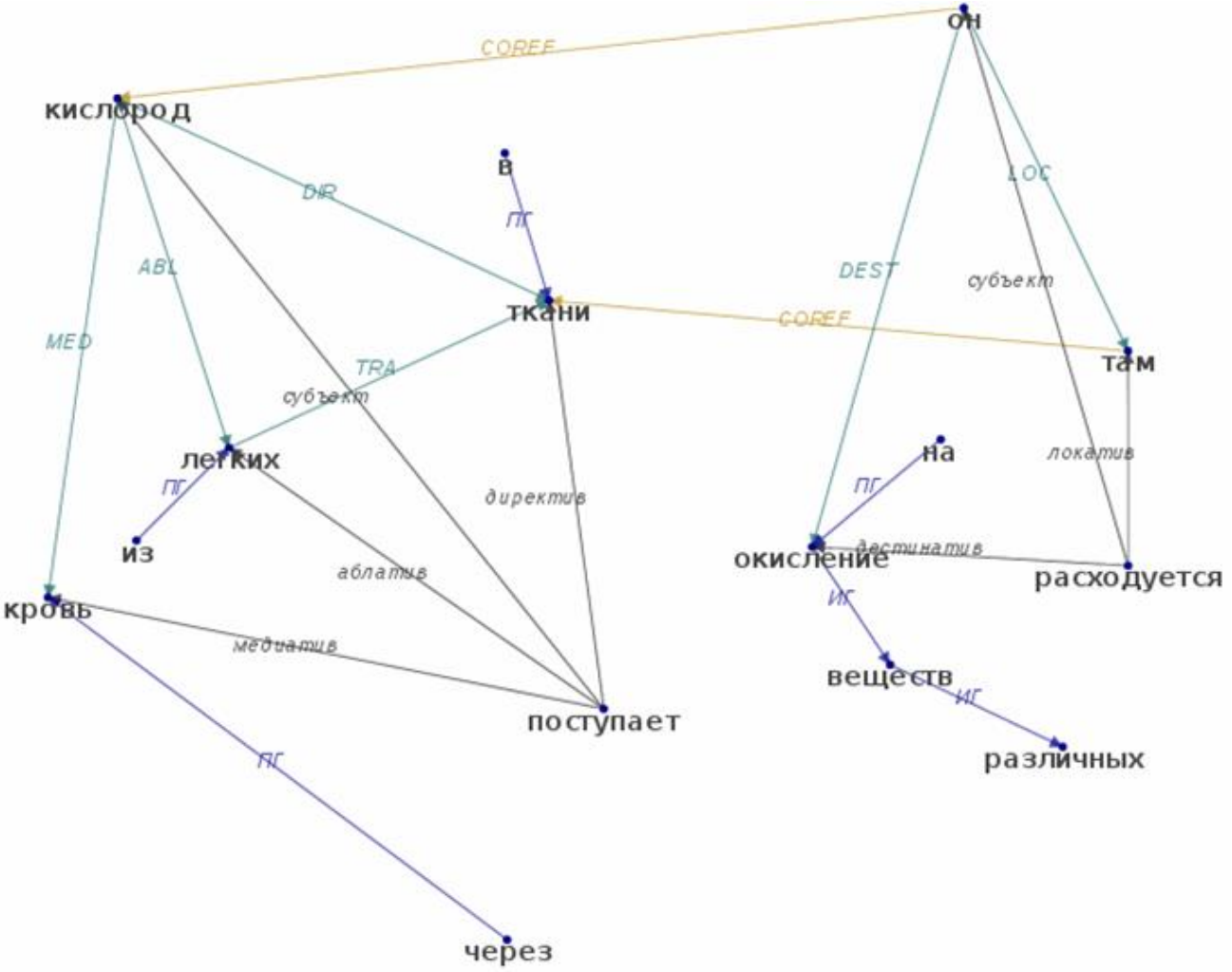
Дополняют основные методы, немного улучшая их качество

Семантический поиск заимствований

- Учитываются не последовательности слов, а семантика высказываний, представленная в виде семантической сети
- Учитывается не просто наличие слов, а значения слов и смысловые связи между ними
- Вычисляется смысловая близость текстов путем сравнения их семантических сетей

Семантическая сеть текста

Кислород поступает в ткани из легких через кровь.
Там он расходуется на окисление различных веществ.



Отличительные особенности семантического поиска заимствований

- Полный лингвистический анализ текстов, включая морфологический и синтактико-семантический анализ
- Выявление смысловых заимствований
- Нечувствительность к сильному перефразированию:
  - перестановке слов и предложений местами
  - замене слов и словосочетаний синонимами
  - разбиению и объединению предложений
- Поддерживаемые языки:
  - русский
  - английский
  - татарский (в разработке)

Примеры обнаруживаемых заимствований

Проверяемый текст

Текст докладной записки делится на две части: Констатирующая (описательная), где излагаются имевшие место факты или описывается ситуация, вторая, где излагаются предложения, просьбы.

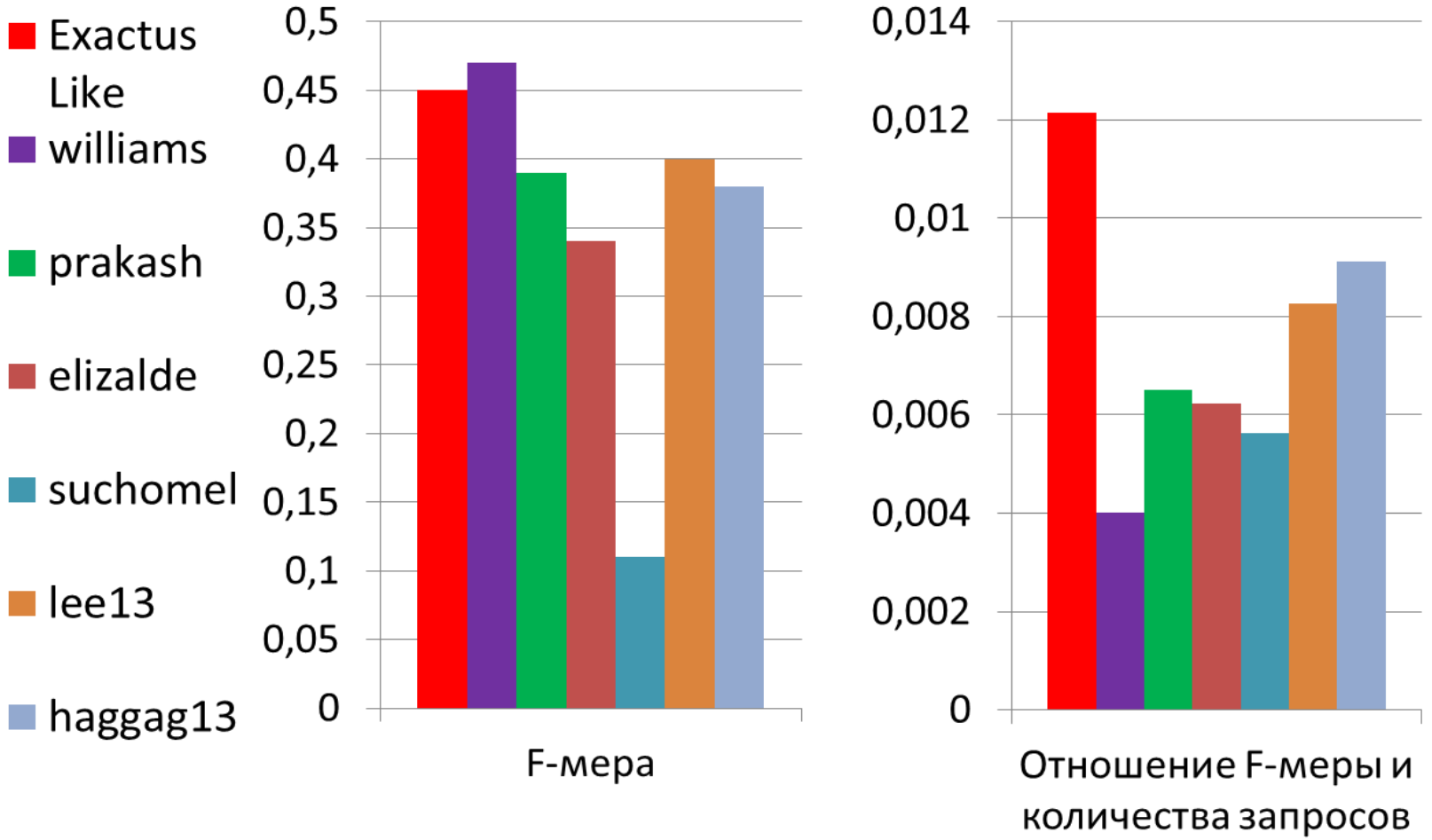
Сам метод заключается в следующем: на каждом шаге мы выбираем один из исходных элементов и вставляем его на нужную позицию в уже отсортированном списке, до тех пор, пока набор исходных данных не будет исчерпан.

Текст источника

Докладная записка обычно состоит из **двух частей**: в первой **описывается** сложившаяся **ситуация**, во **второй излагаются предложения**, **просьбы**, делаются конкретные выводы.

На **каждом шаге** алгоритма **мы выбираем один** из **элементов** входных **данных** и восстанавливаем **его** на **нужную позицию** в **уже отсортированном списке**, до **тех пор пока набор** входных **данных не будет исчерпан**.

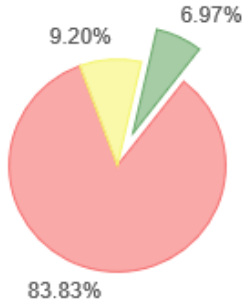
Оценка метода семантического поиска заимствований на международных соревнованиях CLEF'2014



Метод показывает 2-е место по F-мере, 1-е место по соотношению F-мера/количество проверяемых фрагментов

Оценка оригинальности документа - 6.97%

Процент некорректных заимствований - 83.83%



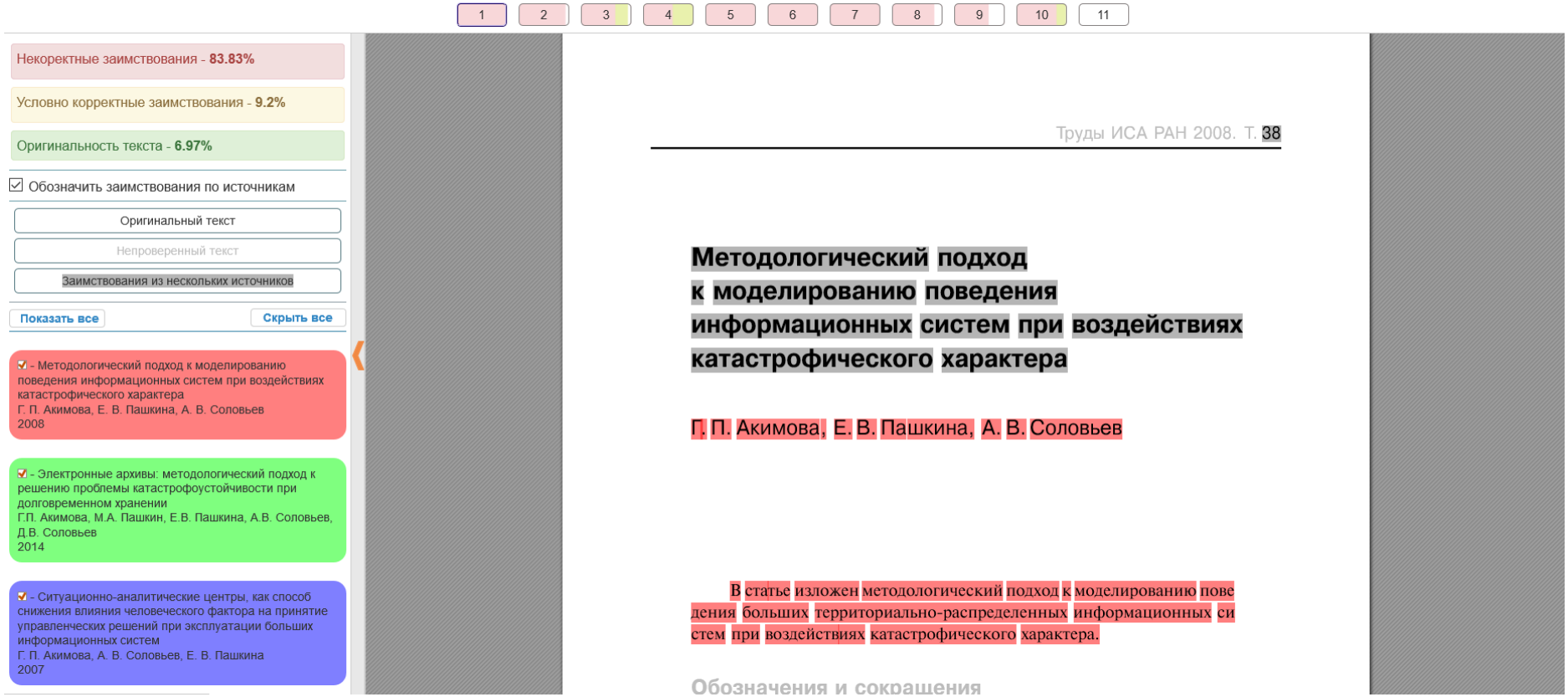
Просмотр заимствований в документе

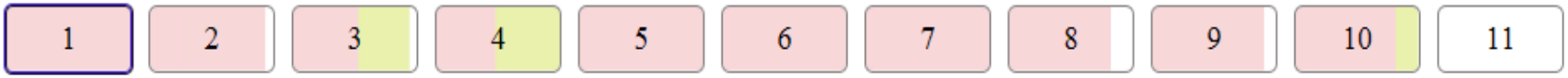
Время выполнения: 16 с.

Документы из базы

Источники ▲

Источники заимствования	В списке литературы	Заимствования
1. Методологический подход к моделированию поведения информационных систем при воздействиях катастрофического характера Авторы: Г. П. Акимова, Е. В. Пашкина, А. В. Соловьев. Год публикации: 2008. Тип публикации: статья научного журнала. http://www.isa.ru/proceedings/images/documents/2008-38/62-72.pdf Показать заимствования (84)	–	95.5%
2. Электронные архивы: методологический подход к решению проблемы катастрофоустойчивости при долговременном хранении Авторы: Г.П. Акимова, М.А. Пашкин, Е.В. Пашкина, А.В. Соловьев, Д.В. Соловьев. Год публикации: 2014. Тип публикации: статья научного журнала. http://www.isa.ru/proceedings/images/documents/2014-64-3/t-14-3_91-98.pdf Показать заимствования (61)	–	69.3%
3. Ситуационно-аналитические центры, как способ снижения влияния человеческого фактора на принятие управленческих решений при эксплуатации больших информационных систем Авторы: Г. П. Акимова, А. В. Соловьев, Е. В. Пашкина. Год публикации: 2007. Тип публикации: статья научного журнала. http://www.isa.ru/proceedings/images/documents/2007-29/113-122.pdf Показать заимствования (9)	✓	10.2%


1
2
3
4
5
6
7
8
9
10
11
Некоректные заимствования - 83.83%
Условно корректные заимствования - 9.2%
Оригинальность текста - 6.97%
Обозначить заимствования по источникам
Оригинальный текст
Непроверенный текст
Заимствования из нескольких источников
Показать все
Скрыть все
- Методологический подход к моделированию поведения информационных систем при воздействиях катастрофического характера
Г. П. Акимова, Е. В. Пашкина, А. В. Соловьев
2008
- Электронные архивы: методологический подход к решению проблемы катастрофоустойчивости при долговременном хранении
Г.П. Акимова, М.А. Пашкин, Е.В. Пашкина, А.В. Соловьев, Д.В. Соловьев
2014
- Ситуационно-аналитические центры, как способ снижения влияния человеческого фактора на принятие управленческих решений при эксплуатации больших информационных систем
Г. П. Акимова, А. В. Соловьев, Е. В. Пашкина
2007
Труды ИСА РАН 2008. Т. 38
Методологический подход к моделированию поведения информационных систем при воздействиях катастрофического характера
Г. П. Акимова, Е. В. Пашкина, А. В. Соловьев
В статье изложен методологический подход к моделированию поведения больших территориально-распределенных информационных систем при воздействиях катастрофического характера.
Обозначения и сокращения



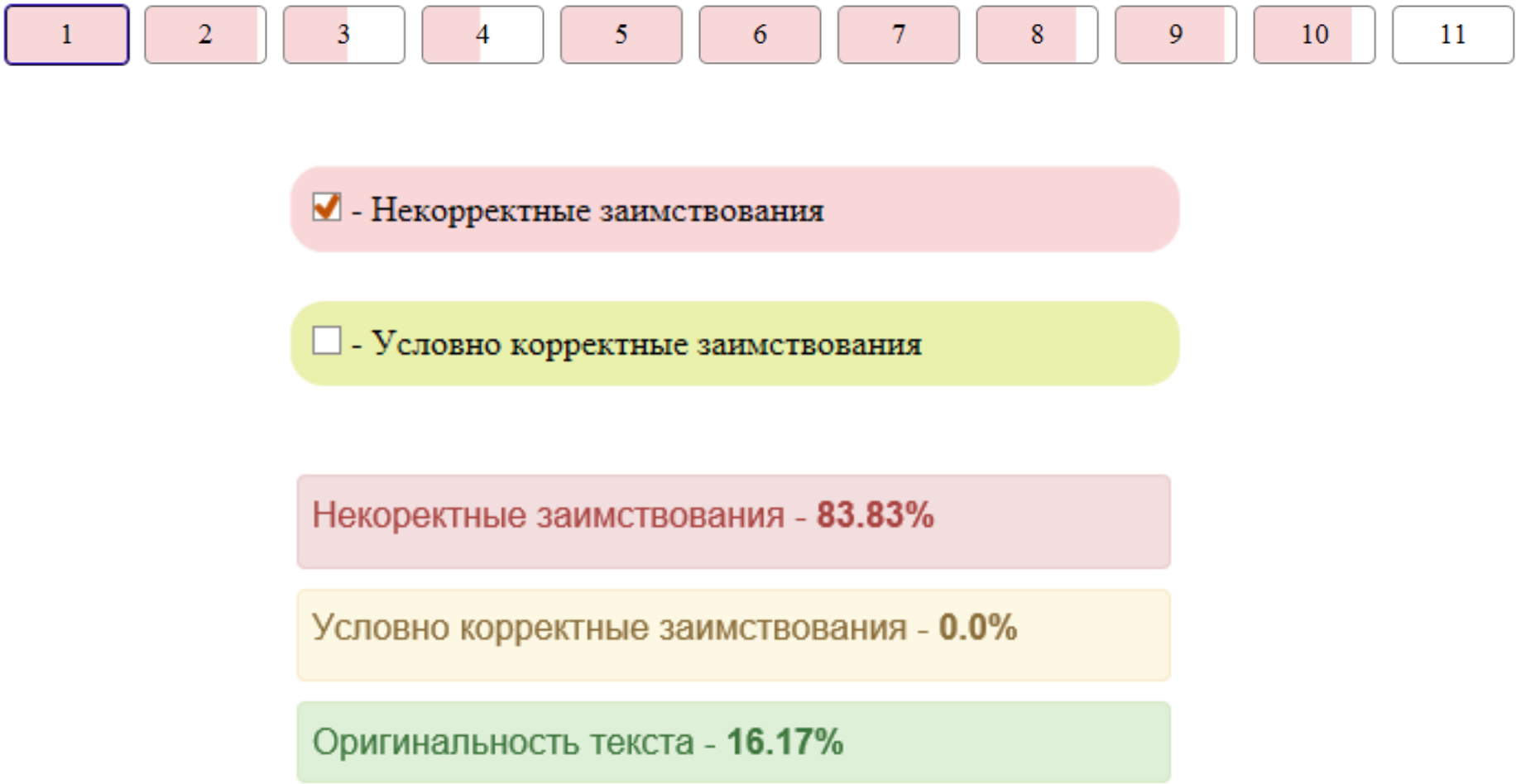
☑ - Некорректные заимствования

☑ - Условно корректные заимствования

Некоректные заимствования - **83.83%**

Условно корректные заимствования - **9.2%**

Оригинальность текста - **6.97%**


1
2
3
4
5
6
7
8
9
10
11
- Некорректные заимствования
- Условно корректные заимствования
Некоректные заимствования - 83.83%
Условно корректные заимствования - 0.0%
Оригинальность текста - 16.17%

8. На основании составленных частных моделей рисков, противодействия им, функционирования информационной системы создается общая модель катастрофоустойчивости. На основании данной модели проводятся расчеты и определяются стратегии повышения защищенности элементов ИС.

Идеальным вариантом проведения полноценного моделирования катастрофоустойчивости информационной системы является создание ситуационно-аналитических центров, реализующих модели функционирования ИС (см. [15]). Их использование значительно повышает эффективность функционирования и развития информационной системы в целом.

Заключение

Большие информационные системы все чаще становятся неотъемлемой частью производственного процесса на промышленных предприятиях, коммерческих организациях и в государственных структурах. Чем крупнее организация, тем большая по масштабам информационная система требуется для охвата и управления всем производственным и/или технологическим циклом, но и тем больше риск потери критически важной информации.

Сгруппировать по фрагментам Сгруппировано по документам ×

- **Ситуационно-аналитические центры, как способ снижения влияния человеческого фактора на принятие управленческих решений при эксплуатации больших информационных систем**

Авторы: Г. П. Акимова, А. В. Соловьев, Е. В. Пашкина
Год публикации: 2007. http://www.isa.ru/proceedings/images/documents/2007-29/113-122.pdf
Показать заимствования Скрыть заимствования

1. Введение **Большие информационные системы все чаще становятся неотъемлемой частью производственного процесса** на **промышленных предприятиях**, **коммерческих организациях** и в **государственных структурах**. <...>

ются стратегии повышения защищенности

ведения полноценного моделирования ка
ационной системы является создание си
гров, реализующих модели функциониро
зование значительно повышает эффектив
звития информационной системы в целом.

Заключение

Большие информационные системы все чаще становятся неотъем лемой частью производственного процесса на промышленных предпри ятиях, коммерческих организациях и в государственных структурах. Чем крупнее организация, тем большая по масштабам информационная си стема требуется для охвата и управления всем производственным и/или технологическим циклом, но и тем больше риск потери критически важ ной информации.

Сгруппировать по фрагментам Сгруппировано по документам

- **Ситуационно-аналитические центры, как способ снижения влияния человеческого фактора на принятие управленческих решений при эксплуатации больших информационных систем**

Авторы: Г. П. Акимова, А. В. Соловьев, Е. В. Пашкина

Год публикации: 2007. http://www.isa.ru/proceedings/images/documents/2007-29/113-122.pdf

Показать заимствования Скрыть заимствования

1. Введение **Большие информационные системы все чаще становятся неотъемлемой частью производственного процесса** на **промышленных предприятиях, коммерческих организациях** и в **государственных структурах**. <...>

Заключение

Большие информационные

Реализация технологии

TextAppliance

- ✓ Отказоустойчивость
- ✓ Масштабируемость
- ✓ Простота интеграции (JSON/XML-RPC)
- ✓ Поддержка Big Data

- ✓ Полнотекстовая индексация
- ✓ Извлечение и индексация метаданных
- ✓ Поддержка распространённых форматов документов
- ✓ Распознавание текста

Система поиска заимствований РУКОНТекст

РУКОНТ

СИСТЕМА ПОИСКА ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ

СКОЛКОВО РУКОНТекст

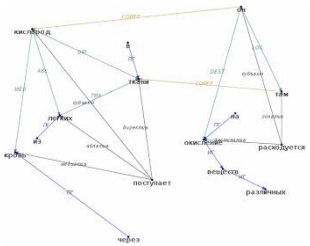
Программно-аппаратный комплекс TextAppliance

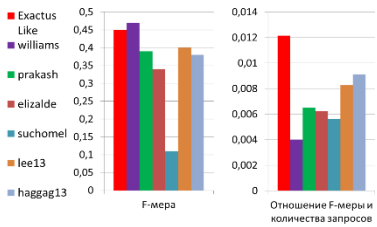
ТСА

TextAppliance

Метод семантического поиска заимствований

РАН





Exactus Like

like.exactus.ru

textapp.ru

ИСА ФИЦ ИУ РАН

ООО «ТСА»

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

Телефон/факс: +7 (499) 135-04-63

e-mail: ivs@isa.ru

Основные публикации

1. Sochenkov, Ilya, Denis Zubarev, Ilya Tikhomirov, Ivan Smirnov, Artem Shelmanov, Roman Suvorov, and Gennady Osipov. "Exactus Like: Plagiarism Detection in Scientific Texts." In *Advances in Information Retrieval*, pp. 837-840. Springer International Publishing, 2016.
2. Zubarev, D., Sochenkov, I.: Using Sentence Similarity Measure for Plagiarism Source Retrieval — Notebook for PAN at CLEF 2014. In: CEUR Workshop Proceedings, CEUR-WS.org, Eds. L. Cappellato, N. Ferro, M. Halvey and W. Kraaij. 2014. P.p. 1027–1034.
3. Осипов, Г. С., И. В. Смирнов, И. А. Тихомиров, И. В. Соченков, Д. В. Зубарев. "Exactus Like – система выявления заимствований в научных текстах“ // Труды международной конференции КРЫМ-2015
4. Зубарев Д. В. Поиск потенциального плагиата на основе метода многокритериальной оценки сходства текстов // Труды III Всероссийской научной конференции молодых ученых с международным участием ТПСА. 2014. Том 2, С. 148-157
5. И.В. Соченков. Метод сравнения текстов для решения поисково-аналитических задач // Искусственный интеллект и принятие решений. М.: ИСА РАН, 2013, №2, с.95-106.