

# Экспериментальное исследование методов выявления нечетких дубликатов научных публикаций

*Дербенёв Н. В., Козлюк Д. А., Никитин В. В., Толчеев В. О.  
Кафедра Управления и информатики НИУ «МЭИ»*

# Нечёткие дубликаты. Чем это плохо

- Показателем эффективности научной работы заявлено количество публикаций.
- Результаты публикуются повторно.
  - ✓ Популяризация полезна.
  - ✗ Усложняется поиск материалов для новых исследователей.
  - ✗ Тратится время ученых.
  - ✗ Искажается картина ценности идей и результатов.

# Причина возникновения нечетких дубликатов научных публикаций

Показателем качества деятельности ученого заявлено количество публикаций



Необходимость написания большого количества статей



Уменьшение времени на оригинальные научные исследования



Дублирование одних и тех же результатов в разных работ

# Особенность поиска нечетких дубликатов среди научных публикаций

- Затруднен доступ к полнотекстовым документам
- В свободном доступе библиографические описания
- Пытаемся выявлять дубли-ПО, анализируя БО.

# Исследуемые методы выявления нечетких дубликатов

- Меры близости:
  - Коэффициент ассоциативности Джаккарда
  - ОКА (обобщенный коэффициент ассоциативности)
- Шинглы и их модификации:
  - Метод шинглов
  - Winnowing
- Специализированные расстояния:
  - Коэффициент Жаро
  - Коэффициент Жаро-Винклера
  - Расстояние Левенштейна

# Обобщенный коэффициент ассоциативности (ОКА)

Дербенев Н. В., Толчеев В. О., ИОИ-2012 (Будва)



# Варианты целевого критерия

## Варианты целевых критериев

|  |  |
|--|--|
| Максимальное значение одного из показателей ( <i>полнота</i> или <i>точность</i> ) при заданных ограничениях на другой | $\exists \theta_* \in [0; 1]: P(\theta_*)_* = \max_{\theta \in [0; 1]} P(\theta), R(\theta_*) \geq C$ <p style="text-align: center;">или</p> $\exists \theta_* \in [0; 1]: R(\theta_*)_* = \max_{\theta \in [0; 1]} P(\theta), R(\theta_*) \geq C$ |
| Максимальное суммарное значение <i>полноты</i> и <i>точности</i>   | $\exists \theta_* \in [0; 1] : \forall \theta [0; 1] (R(\theta) + P(\theta)) \leq (R(\theta_*) + P(\theta_*))$   |
| Баланс между <i>полнотой</i> и <i>точностью</i> (близость значений этих параметров)                                    | $\exists \theta_* \in [0; 1] : \forall \theta \in [0; 1] R(\theta_*) \approx P(\theta_*)$  |

# Выбор целевого критерия

- Ценные ресурсы:
  - Время экспертов
  - Полные тексты (они часто недоступны)
- Требования к полноте:
  - Обеспечение полноты не ниже заданного высокого уровня, чтобы не пропускать нечеткие дубликаты
- Требования к точности:
  - Максимизация точности, дабы отсеять оригинальные документы



Целевой критерий: максимизация **точности**, при заданном уровне **полноты** ( $\geq 90\%$ )

# Особенности выборки

- 150 пар библиографических описаний с eLibrary.ru, около 50 полнотекстовых документов
- Экспертные оценки:
  - «дубликат» — «не дубликат»,
  - вспомогательная подсветка,
  - 3 эксперта,
  - согласованная окончательная оценка.

# Результаты стандартных методов

| Метод                                 | Пороговое значение | Полнота | Точность    |
|---------------------------------------|--------------------|---------|-------------|
| Коэффициент ассоциативности Джаккарда | 0,46               | 0,91    | <b>0,71</b> |
| Шинглы (длина = 2 слова)              | 0,34               | 0,934   | <b>0,71</b> |
| Жаро                                  | 0,68               | 0,91    | 0,689       |
| Жаро-Винклер                          | 0,68               | 0,9     | 0,692       |
| Расстояние Левенштейна                | 0,46               | 0,92    | 0,68        |
| Winnowing k=2 t=3                     | 0,31               | 0,92    | <b>0,71</b> |
| ОКА                                   | 0,45               | 0,9     | <b>0,73</b> |

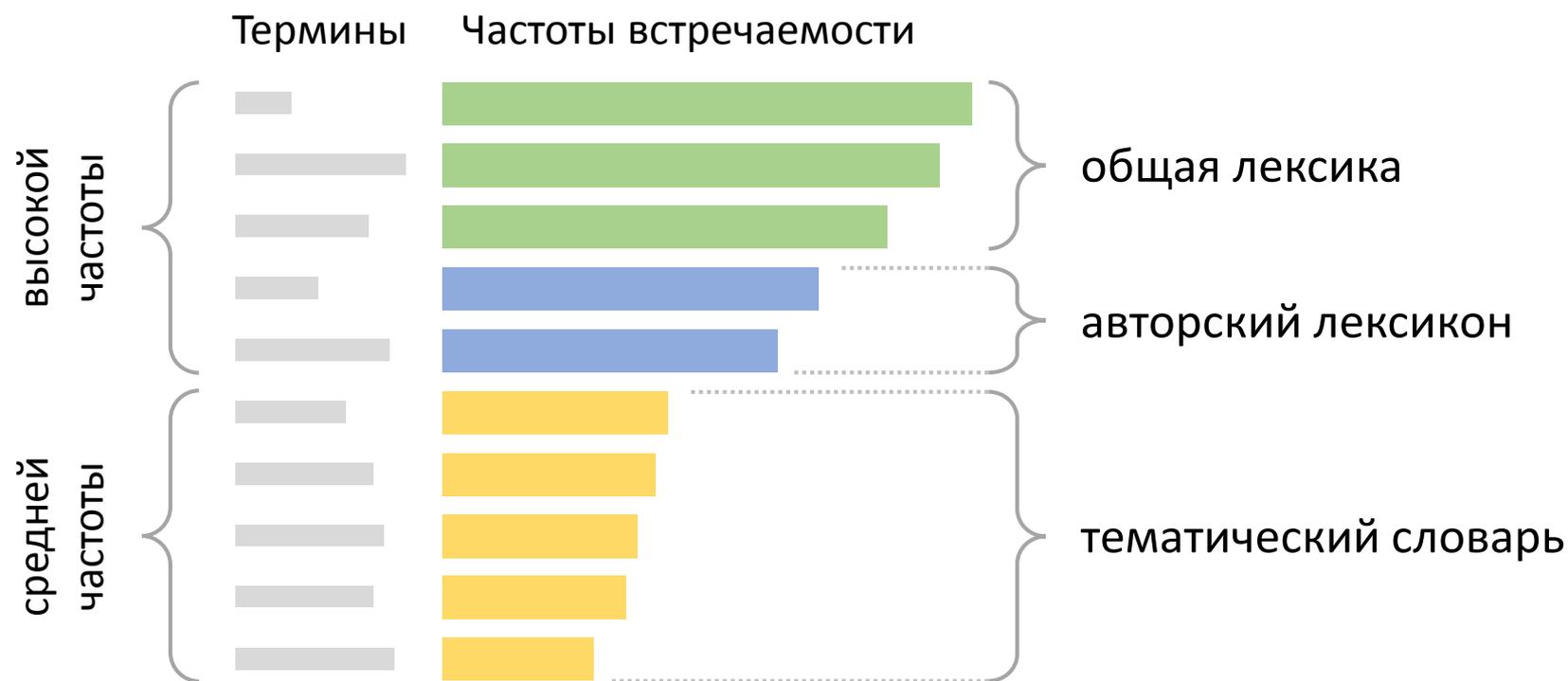
# Результаты стандартных методов (со стеммингом)

| Метод                                 | Пороговое значение | Полнота | Точность     |
|---------------------------------------|--------------------|---------|--------------|
| Коэффициент ассоциативности Джаккарда | 0,54               | 0,908   | <b>0,712</b> |
| Шинглы (длина = 2 слова)              | 0,38               | 0,908   | <b>0,718</b> |
| Жаро                                  | 0,74               | 0,908   | 0,701        |
| Жаро-Винклер                          | 0,74               | 0,908   | 0,701        |
| Расстояние Левенштейна                | 0,52               | 0,871   | 0,701        |
| Winnowing k=2 t=4                     | 0,35               | 0,917   | <b>0,720</b> |
| ОКА                                   | 0,57               | 0,908   | <b>0,724</b> |

# Возможные варианты увеличения точности:

- Рассматриваемые варианты, которые не дали результата:
  - Анализ стоп-слов, встречающихся в двух сравниваемых документах с одинаковой частотой (наличие одних и тех же служебных слов, союзов, предлогов и т.п.)
  - Анализ последовательностей служебных символов, включая знаки препинания и сокращения
- Эффективные варианты увеличения полноты-точности:
  - Использование авторского словаря

# Учёт авторского словаря

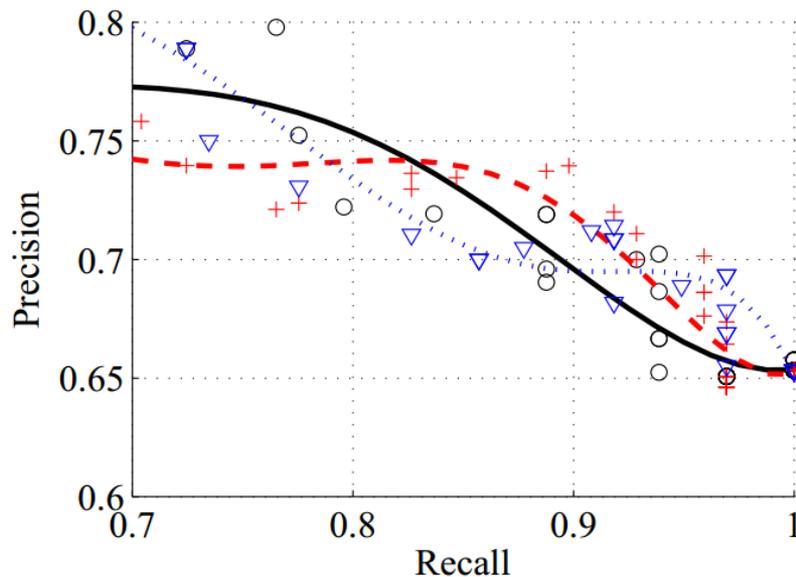


# Результаты методов с использованием словаря автора

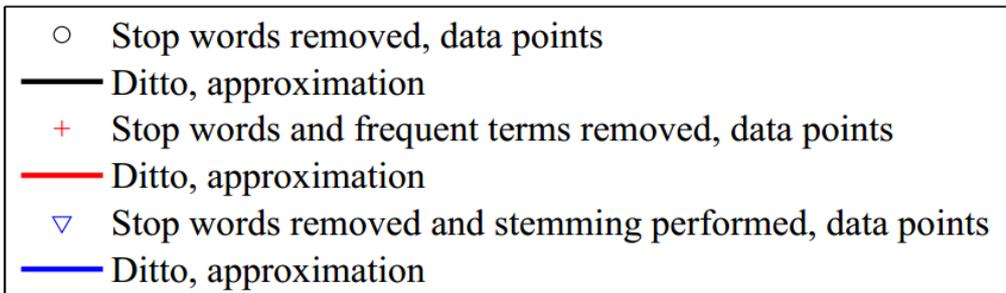
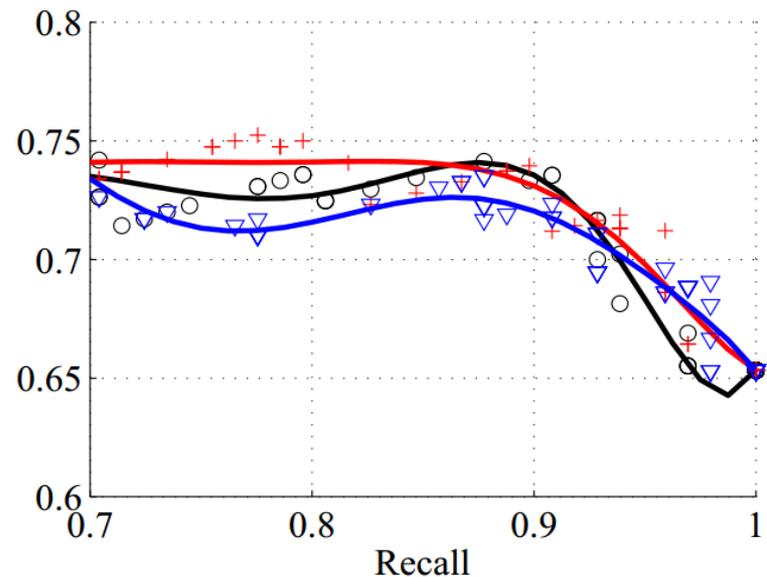
| Метод                                 | Пороговое значение | Полнота | Точность     |
|---------------------------------------|--------------------|---------|--------------|
| Коэффициент ассоциативности Джаккарда | 0,47               | 0,92    | <b>0,74</b>  |
| Шинглы (длина = 2 слова)              | 0,33               | 0,92    | <b>0,71</b>  |
| Жаро                                  | 0,71               | 0,91    | 0,69         |
| Жаро-Винклер                          | 0,71               | 0,92    | 0,69         |
| Расстояние Левенштейна                | 0,46               | 0,91    | 0,72         |
| Winnowing k=2 t=4                     | 0,35               | 0,917   | <b>0,72</b>  |
| ОКА                                   | 0,57               | 0,908   | <b>0,745</b> |

# Результаты

## Джаккард



## ОКА



# Можно ли по сходству описаний судить о сходстве документов?

- Из 300 документов 150 исследуемых пар в открытом доступе — 31 (10%)
  - ...к вопросу о ценных ресурсах.
- Среди 31 документа — 6 полных дубликатов.
  - Это 3 пары с выявленными схожими БО.
- «Самоплагиат» встречается редко.

Спасибо за внимание!