Исследование «Сами мы не местные»

Александров Юрий Юрьевич

Санкт-Петербургский Академический Университет

Научный руководитель: Кукушкин Владимир
Яндекс, Служба асессоров

22 декабря 2014 г.

Введение

Асессоры — люди, ставящие оценки документам $\langle query, url \rangle \in \mathrm{QU}$.
Группы асессоров для каждого региона:

- «Местные» — асессоры, живущие в этом регионе и имеющие представление о нём.
- «Неместные» — остальные.

Оценки:

VITAL_CANDIDATE

USEFUL

RELEVANT_PLUS

RELEVANT_MINUS

IRRELEVANT

_404

VIRUS

Вопрос: как распределять задания между асессорами, чтобы получать наиболее точные оценки? Дают ли местные асессоры более точные оценки для своего региона, чем неместные?

Метрики качества

$$QU^{\langle 2\rangle} \stackrel{\text{def}}{=} \left\{ ((q,u_1),(q,u_2)) \;\middle|\; (q,u_1),(q,u_2) \in QU, u_1 \neq u_2 \right\}$$

Кликовое предпочтение для пары $(\langle q,u\rangle,\langle q,v\rangle) \in QU^{\langle 2\rangle}$:

$$P_{uv} = \frac{\left(c^{\hat{u}v} + c^{v\hat{u}}\right) - \left(c^{u\hat{v}} + c^{\hat{v}u}\right)}{c^{\hat{u}v} + c^{v\hat{u}} + c^{u\hat{v}} + c^{\hat{v}u} + c^{\hat{u}\hat{v}}}$$

$\mathbf{x}_{(n)} \sim \left(QU^{\langle 2\rangle}\right)^n$ — случайная выборка

Метрики качества оценок $J((q,u))$ асессоров:

- Доля корректных:
  $\frac{1}{n}\left|\left\{(x,y) \in \mathbf{x}_{(n)} \;\middle|\; P_{x,y}\cdot(J(x)-J(y)) > 0 \vee P_{x,y} = J(x)-J(y) = 0\right\}\right|$
- Доля некорректных:
  $\frac{1}{n}\left|\left\{(x,y) \in \mathbf{x}_{(n)} \;\middle|\; P_{x,y}\cdot(J(x)-J(y)) < 0 \vee P_{x,y} = 0 \wedge J(x) \neq J(y)\right\}\right|$
- Доля нечувствительные: $\frac{1}{n}\left|\left\{(x,y) \in \mathbf{x}_{(n)} \;\middle|\; P_{x,y} \neq 0 \wedge J(x) = J(y)\right\}\right|$

Постановка задачи

$\mathcal{C}_i\left(QU^{\langle 2\rangle}\right)$ — значения метрик для контрольной группы (неместные),
$\mathcal{E}_i\left(QU^{\langle 2\rangle}\right)$ — значения метрик для экспериментальной группы (местные).

Требуется проверить три статистических гипотезы при уровне значимости $\alpha = 0.05$:

$$H_0^{(i)}: \mathcal{C}_i\left(QU^{\langle 2\rangle}\right) = \mathcal{E}_i\left(QU^{\langle 2\rangle}\right), \quad i \in \{1, 2, 3\}$$

Подзадачи

- Структуризация данных для удобной и быстрой работы.
- Вычисление оптимального разбиения пар $\langle query, url \rangle$ по количествам кликов и кликовым предпочтениям, когда все корзины с похожими количествами уникальных запросов.
- Случайная выборка локализованных пар $\langle query, url \rangle$ из каждой корзины разбиения для их последующей оценки асессорами.
- Анализ соответствия между оценками асессоров контрольной и экспериментальной групп и кликовыми предпочтениями.

Связанные работы

- G. Kazai, N. Craswell, and et. al. An analysis of systematic judging errors in information retrieval. CIKM '12
  - Корректность судейских оценок по отношению к различным эталонам.
- G. Kazai, N. Craswell, and et. al. User intent and assessor disagreement in web search evaluation. CIKM'13
  - Метрики соответствия судейский оценок и эталонных.

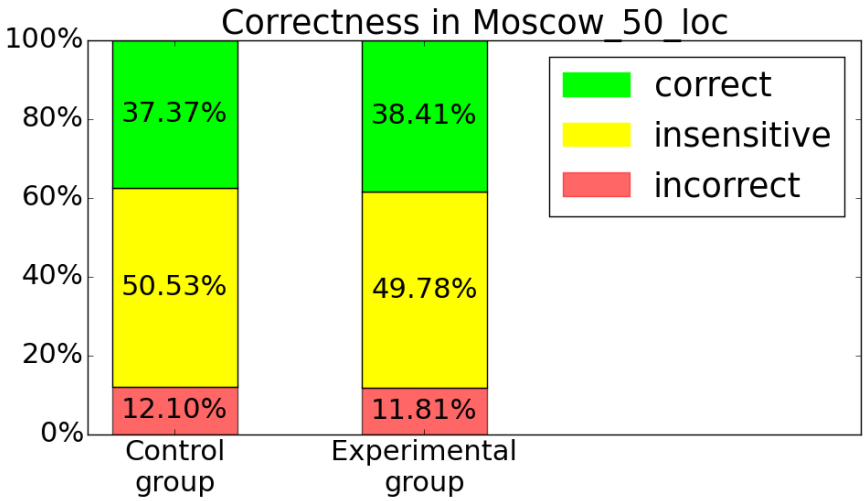
Данные

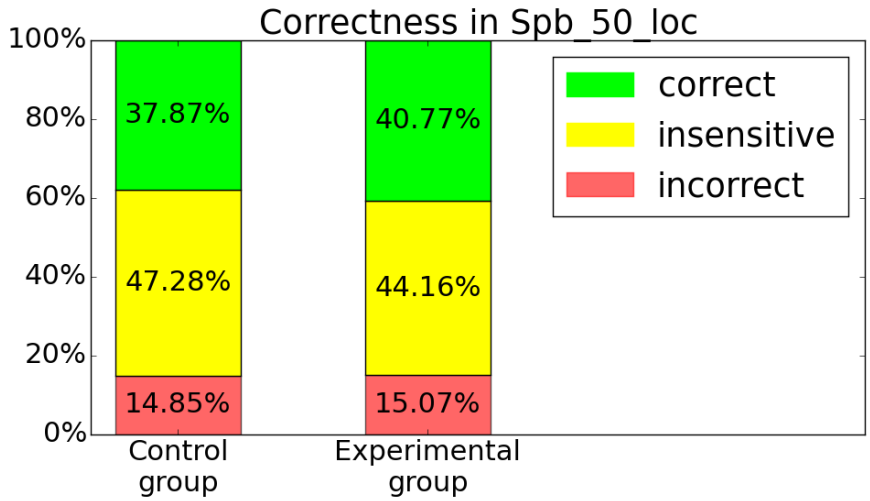
4 города: Москва, Санкт-Петербург, Ижевск, Пермь.

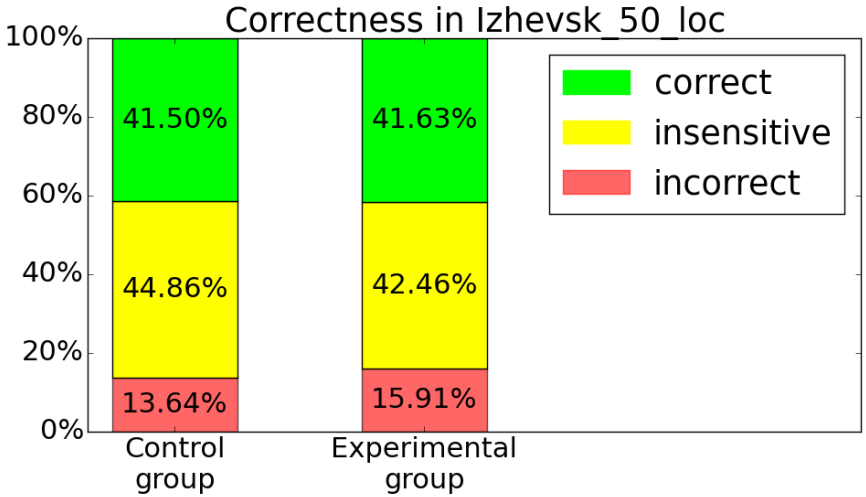
Исходные данные — набор кортежей
<query_id, url1_id, url2_id, click_volume, click_pref>

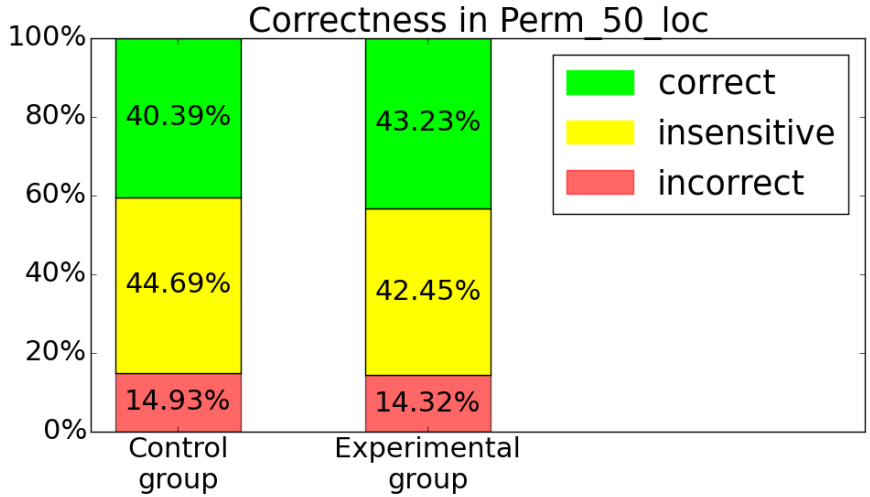
Данные с асессорскими оценками — набор кортежей
<query_id, url_id, label>

Корректность

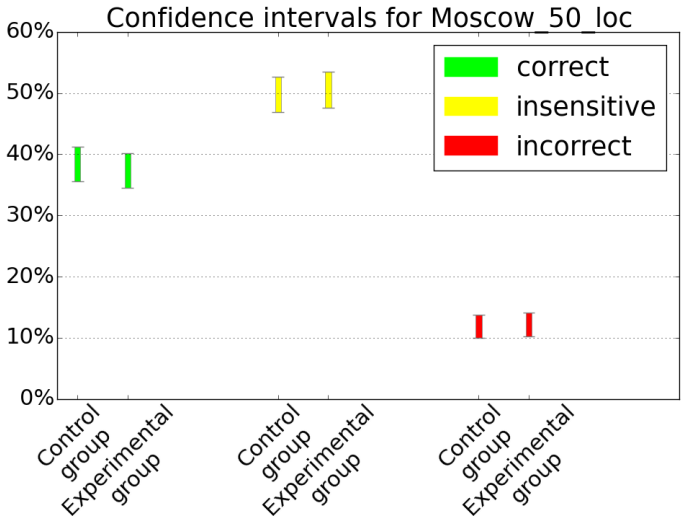


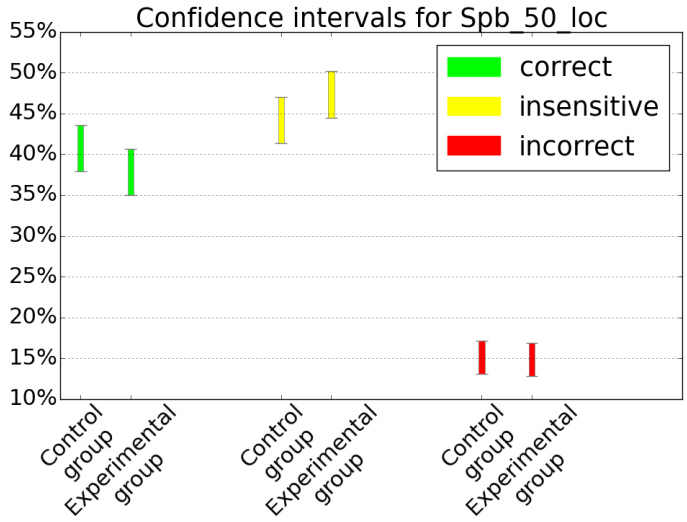


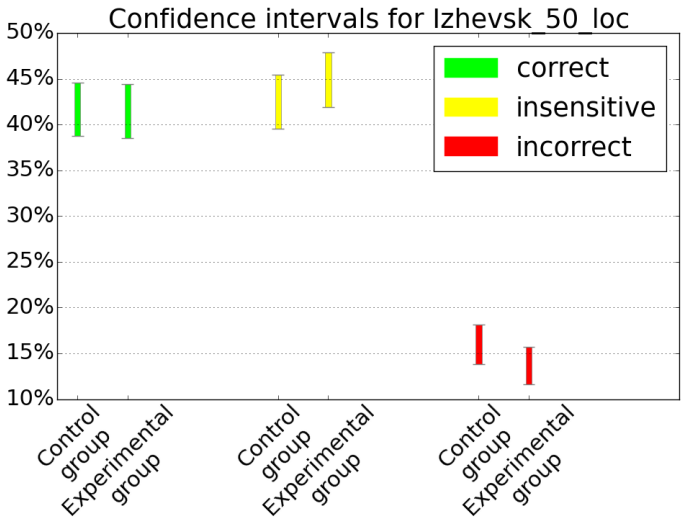


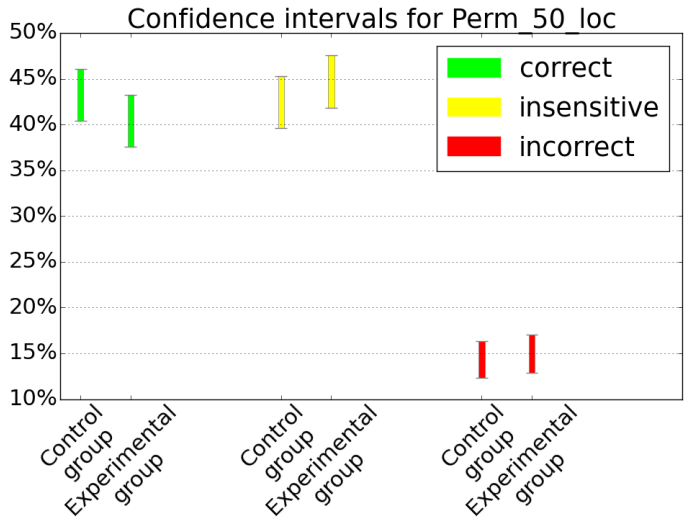


95%-доверительные интервалы корректности









Вывод

Везде 95%-доверительные интервалы пересекаются $\Rightarrow$ оценки контрольной и экспериментальной групп отличаются незначимо.

Использованные средства

https://github.com/alprobit/local_assessors

Что нового для себя

- Работа с различными конструкциями языка Python 3
- Работа с пакетами для вычислений и визуализации
- Знакомство с особенностями асессорских оценок и метрик
- Конкретное применение статистики на практике: например, знакомство с bootstrap-методом