Влияние качества разметки данных в моделях для предсказания субъективных впечатлений пользователей

М.А. Бакаев, В.А. Хворостов

***Аннотация*— Качество обучающих данных широко признаётся важнейшим условием для успешного создания моделей машинного обучения (ML), однако конкретные составляющие этого понятия могут различаться в зависимости от предметной области и предсказываемых откликов. В сфере человеко-компьютерного взаимодействия (HCI) обычно считается, что трудоемкая разметка людьми графических интерфейсов, т.е. детекция и выделение в них отдельных визуальных элементов, позволяет более точно прогнозировать субъективные впечатления пользователей. В то же время, популярность набирают сервисы для автоматической параметризации интерфейсов, основанные на технологиях компьютерного зрения. В нашей статье описывается экспериментальное исследование с более чем 200 участниками, в котором около 1000 скриншотов веб-интерфейсов были оценены по 3 шкалам субъективного восприятия (визуальная сложность, эстетичность, упорядоченность) с целью изучения влияния типа и качества разметки на точность предсказания оценок. В общей сложности, по результатам ручной разметки и с использованием автоматизированных сервисов было рассчитано 16 метрик (признаков для моделей). Полученные результаты свидетельствуют, что корреляция между качеством разметки и точностью предсказания восприятия по шкалам эстетичности и упорядоченности оказалась отрицательной (-0.768 и -0.644 соответственно), при высоком уровне статистической значимости. Однозначного преимущества метрик, полученных посредством ручной или автоматической разметки выявлено не было, за исключением «ручной» метрики количества элементов. Приводимые нами в заключительной части статьи рекомендации по параметризации интерфейсов и выводы относительно влияющих метрик могут представлять интерес как для исследователей в сфере ML-HCI, так и для практикующих UI/UX дизайнеров.**

***Ключевые слова*— веб-интерфейсы, качество данных, компьютерное зрение, машинное обучение, человеко-компьютерное взаимодействие.**

I. Введение

Одним из слабо формализованных понятий в области машинного обучения является качество данных [1]. В теории, в него включают до 20 аспектов: полнота, непротиворечивость, точность, своевременность, отсутствие дублирований и т.п. В то же время, конкретные составляющие качества обучающих данных зависят от предметной области [2], и наибольшие сложности с их определением возникают при описании поведения такого сложного субъекта как человек, с его когнитивными процессами и эмоциональными реакциями. С практической точки зрения, разработчиков в сфере искусственного интеллекта (ИИ) обычно интересует, приведет ли добавление дополнительных обучающих данных, реальных или синтетических, к повышению показателей качества моделей машинного обучения (ML) [3]. Эта проблема имеет особую актуальность для тех областей, где доступные объемы данных являются ограниченными вследствие значительных трудозатрат, необходимых для их получения.

В сфере человеко-компьютерного взаимодействия (HCI) таковыми являются, прежде всего, данные о субъективных впечатлениях пользователей, которые затруднительно собрать иначе как в ходе индивидуальных опросов. Между тем, эти пользовательские впечатления, составляющие то, что принято называть UX (User Experience), в значительной степени определяют итоговый успех или провал ИТ-продукта [4]. Считается, что субъективную привлекательность графических интерфейсов возможно с приемлемым уровнем качества предсказывать на основе некоторых их количественных характеристик – «метрик». При этом особый эффект на, например, эстетическое впечатление имеют характеристики, отражающие композицию, т.е. гармоничное расположение элементов интерфейса [5].

Расчет таких метрик не является тривиальной задачей, потому что, несмотря на разнообразие применяемых формул, обычно требует данных о размерах всех или большинства визуально «весомых» элементов интерфейса и их координатах на экране. Среднее количество таковых элементов в современных веб-интерфейсах имеет порядок 10^2, что делает их ручную разметку довольно трудоемкой задачей – при том, что автоматическое определение размеров и положения элементов из программного кода (HTML/CSS), без рендеринга в браузере, как правило, является

Статья получена 25 ноября 2024 г.

Статья подготовлена по итогам выступления на конференции «Интернет и современное общество», IMS-2024.

Бакаев Максим Александрович, кандидат технических наук, доцент, Новосибирский государственный технический университет, Россия (bakaev@corp.nstu.ru).

Хворостов Владимир Александрович, старший преподаватель, Новосибирский государственный технический университет, Россия (bakaev@corp.nstu.ru).

непродуктивным. Соответственно, с новым витком развития технологий компьютерного зрения, т.е. примерно 10 лет назад, у научного сообщества и у проектировщиков (дизайнеров) человеко-компьютерных интерфейсов начал расти интерес к автоматизированной детекции элементов; стали появляться соответствующие сервисы для автоматической параметризации интерфейсов [6], [7]. Однако существующая практика и наш собственный опыт показывают, что они пока уступают людям-разметчикам – конечно, при надлежащей мотивации и выстроенном процессе контроля качества, – по показателям как точности (в особенности, при определении типа элементов интерфейсов), так и полноты (детектируя примерно на 25% меньше элементов) [8]. Современные тенденции в этой области – это сочетание метрик, получаемых при участии человека и посредством автоматизированных способов параметризации (см., например, [9]). Однако определение конкретного влияния совмещения различных групп метрик и качества обучающих данных в целом на качество итоговых моделей является открытой и актуальной проблемой в сфере ML-HCI.

Наше данное экспериментальное исследование посвящено оценке эффекта качества ручной разметки интерфейсов на качество моделей, предсказывающих субъективные впечатления пользователей по шкалам визуальной сложности, эстетичности и упорядоченности. В общей сложности мы задействует около 1000 скриншотов веб-страниц и более 200 участников для оценки субъективных впечатлений, разметки интерфейсов и контроля её качества. Для автоматической параметризации скриншотов был использован разработанный нами специализированный онлайн-сервис VA (Visual Analyzer[1]) [6] и сторонний сервис AIM [7]. Анализ полученных нами экспериментальных данных позволяет сделать вывод, что более качественная ручная разметка не даёт гарантированного эффекта для качества предсказывающих моделей, а для некоторых субъективных шкал, неожиданно, имеет даже отрицательный эффект. Излагаемые результаты были ранее представлены нами на научно-исследовательских конференциях INTELS-2022 [10] и PIERE-2024 [11]. В данной статье результаты нескольких наших исследований, касающихся качества обучающих данных для моделей поведения пользователей, обобщаются и представляются на русском языке.

II. Описание экспериментальной части исследования

A. *Используемый материал*

Материалом в нашем исследовании выступали скриншоты главных страниц веб-сайтов, принадлежавших к различным предметным областям: университеты, музеи, аптеки, новостные сайты и т.д. (см. в [12]). Изначально мы собрали более чем 13 тыс. скриншотов, с использованием специально разработанных нами программных инструментов (способных переходить по ссылками из каталогов и сохранять веб-страницы, отрендеренные в браузере, в формате PNG). Затем были вручную отобраны около 1000 скриншотов для последующей разметки, причём критериями отбора являлись следующие:

1) Содержание представлено на английском языке (международные версии веб-сайтов из различных стран мира).
2) Разнообразие предметных областей и дизайнов (с точки зрения цветовой гаммы, расположения элементов и т.п.).
3) Отсутствие видимых технических проблем со скриншотом (например, недозагруженных веб-страниц или всплывающих окон).

Подробнее процесс сбора и отбора скриншотов описаны в наших предыдущих публикациях, например [12].

B. *Участники исследования*

Все участники нашего исследования принимали участие добровольно и выразили информированное согласие. Исследование проводилось в соответствии с принципами Хельсинкской декларации и было одобрено этическим комитетом ФГО НГТУ (номер протокола 7_02_2019).

1) Оценивание субъективных впечатлений

В общей сложности в оценке субъективных впечатлений от скриншотов приняло участие 137 человек (приблизительно равное количество мужчин и женщин), возраст которых находился в диапазоне от 17 до 46 лет. Большинство из них были студентами Новосибирского государственного технического университета (НГТУ), обучающимися на ИТ-специальностях. Участвовали также несколько представителей ИТ-промышленности и наши зарубежные коллеги, ведущие исследования в сфере человеко-компьютерного взаимодействия. Таким образом, большинство оценивающих (89,1%) были из России, но имелись также представители Болгарии, Германии, Южной Африки и ряда других стран.

2) Разметка скриншотов

Разметка скриншотов производилась двумя партиями, с интервалом приблизительно в 1 год. В разметке первой партии участвовало 11 студентов НГТУ (6 мужчин, 5 женщин) в возрасте от 20 до 24 лет. В разметке второй партии изначально принимали участие 40 студентов НГТУ, также в возрасте от 20 до 24 лет, но небольшая их часть была исключена из исследования в связи с низким качеством выполнения задания (в частности, слишком малое количество размеченных элементов интерфейса на скриншотах).

3) Оценка качества разметки скриншотов

В оценке качества разметки скриншотов из первой партии участвовало 20 студентов (10 мужчин, 10 женщин) в возрасте от 20 до 22 лет, никто из которых до этого не был задействован в разметке скриншотов.

C. *Процесс сбора данных*

1) Оценивание субъективных впечатлений

Все задействованные в исследовании скриншоты

[1] http://va.wuikb.pro

были оценены с использованием специально разработанного нами онлайн-опросника ([12]) по трём шкалам субъективного визуального восприятия:

1) Complexity: насколько визуально сложным кажется скриншот интерфейса;
2) Aesthetics: насколько визуально привлекательным кажется скриншот;
3) Orderliness: насколько упорядоченным кажется скриншот.

Оценка по каждой из шкал производилось по порядковой шкале Ликерта из 7 баллов (1 – самая низкая степень выраженности, 7 – самая высокая). Участники были проинструктированы, что они должны давать именно субъективные оценки и что в опроснике нет правильных или ошибочных ответов. Скриншоты были распределены по участникам случайным образом (с повторениями между участниками) и выдавались им в случайном порядке. Базовое количество оцениваемых скриншотов для каждого участника составляло 100, однако некоторые из студентов не завершили оценку всех назначенных им скриншотов, а некоторые вызвались оценить дополнительные.

2) Разметка скриншотов

Для разметки скриншотов участники использовали специальное программное обеспечение, LabelImg (версия 1.8.1), позволяющее помечать элементы интерфейса прямоугольниками и указывать их тип (image, text, button и т.п.). Участникам была выдана инструкция по разметке и назначено примерно равное количество скриншотов каждому (однако, без случайного распределения) с указанием отметить элементы интерфейса как можно более полно. Результаты разметки сохранялись в XML-файлах формата PASCAL VOC и могли использоваться для дальнейшего анализа.

3) Оценка качества разметки скриншотов

Участники, проверяющие качество разметки скриншотов, использовали разработанное нами программное обеспечение, которое позволяло им отметить каждый из ранее размеченных элементов интерфейса как «корректно» или «некорректно». Проверяющим были выданы письменные инструкции с рекомендациями по принятию решения о корректности разметки. Итоговый показатель корректности разметки скриншота (Precision) рассчитывался как доля размеченных элементов, отмеченных как «корректно».

Кроме того, для каждого скриншота проверяющие указывали субъективный показатель полноты разметки (SC – Subjective Completeness), по шкале от 1 (очень малая доля элементов размечена) до 100 (размечены все элементы). Размеченные скриншоты распределялись среди участников примерно поровну, но не случайным образом, а в алфавитном порядке.

4) Расчет значений «визуальных» метрик

Для целей сравнительного анализа, мы осуществили расчет трёх групп метрик для скриншотов:

a) «Ручные» метрики по типу элементов

Для первой группы скриншотов по результатам ручной разметки (из XML-файлов) мы рассчитали 8 метрик, характеризующих основные визуально воспринимаемые элементы на веб-страницах:

1. Общее количество элементов в интерфейсе;
2. Количество текстовых элементов;
3. Доля текстовых элементов в площади скриншота;
4. Количество элементов-изображений;
5. Доля элементов-изображений в площади скриншота;
6. Количество фоновых изображений;
7. Доля фоновых изображений в площади скриншота;
8. Доля пустого пространства на скриншоте (не относящегося ни к одному из размеченных элементов).

b) «Ручные» метрики по расположению элементов

Для второй группы скриншотов на основе результатов ручной разметки (из XML-файлов) мы рассчитали метрики баланса (Balance), равновесия (Equilibrium) и симметрии (Symmetry) в соответствии с формулами, приведенными в [13] (см. также в [11]). Кроме того, было посчитано общее количество элементов (Count of UI elements).

c) «Автоматические» метрики по расположению элементов

Также, для второй группы скриншотов были получены аналогичные метрики (Quadtree Dec balance, Quadtree Dec equilibrium, Quadtree Dec symmetry, No. of UI elements) автоматическим способом – с применением разработанного нами ранее интеграционного онлайн-сервиса VA (Visual Analyzer) [6]. Сервис способен осуществлять параметризацию скриншотов веб-интерфейсов посредством методов компьютерного зрения и в качестве основного компонента включает в себя сервис Aalto Interface Metrics (AIM) [7].

D. План эксперимента и гипотезы

Изначальными независимыми переменными в нашем исследовании являлись три группы метрик и два показателя качества разметки, усредненные для каждого из 11 разметчиков первой партии скриншотов.

Далее, используя метрики в качестве факторов, мы построили модели линейной регрессии для оценок по каждой из трёх шкал субъективного восприятия (Complexity, Aesthetics, Orderliness). Для скриншотов первой группы модели строились для каждого из 11 разметчиков отдельно, с целью изучения эффекта качества разметки [10]. Для скриншотов второй группы модели строились для всех скриншотов сразу, с целью изучения эффекта типа метрик («ручных» и «автоматических») [11]. В качестве показателя качества построенных моделей использовались значения R^2, которые и являлись основными зависимыми переменными в нашем исследовании.

Были сформулированы следующие гипотезы для проверки в ходе статистического анализа (на материале

первой и второй группы скриншотов соответственно):

1. Модели, построенные по данным разметчиков с более высокими показателями качества разметки (Precision и SC), имеют более высокие показатели качества (R^2).
2. Модели, построенные с «ручными» метрикам в качестве факторов, имеют более высокие показатели качества, чем построенные по «автоматическим» метрикам.

III. Результаты анализа данных

A. Описательная статистика

1) Первая группа скриншотов

В общей сложности, для 497 скриншотов веб-интерфейсов, составивших первую группу, было собрано 12705 оценок по субъективным шкалам. Разметкой было охвачено 495 скриншотов, причём общее количество отмеченных элементов интерфейса составило 42716 штук. Вследствие технических проблем еще некоторая часть скриншотов (2,0%) была исключена, так что дальнейший анализ производился для 487 из них. В ходе проверки качества разметки, 37053 элемента были отмечены «корректно» и 4967 – «некорректно». Соответственно, средняя корректность разметки (Precision) в нашем исследовании составила 88,7%, что может быть довольно высоким значением. Описательная статистика с усреднением результатов по разметчикам представлена в Табл. 1 (колонка ID содержит инициалы разметчиков).

Табл. 1. Описательная статистика по разметке (первая группа скриншотов веб-интерфейсов, n=487)

ID	Размечено		Качество разметки	
	Скрин-шотов	Элемен-тов	Precision	SC
AA	54	4802	89,0%	73,0%
GD	44	3520	89,9%	84,3%
KK	44	3927	95,5%	82,5%
MA	44	5349	72,0%	75,1%
NE	44	4994	85,1%	78,3%
PV	43	4544	91,6%	81,7%
PE	42	2569	77,9%	72,0%
SV	43	3737	97,4%	80,4%
ShM	41	1675	89,5%	77,5%
SoM	45	3266	95,9%	56,0%
VY	43	3630	92,8%	95,5%
Total	**487**	**42013**	**88,7%**	**77,8%**

Коэффициент корреляции Пирсона между двумя показателями качества разметки, Precision и SC, не имел статистической значимости (p=0,727). Это свидетельствует о том, что данные два компонента качества различаются и их разделение было целесообразно. Корреляция между SC и средним количеством корректных элементов оказалась значимой (r_{11}=0,622; p=0,041), а между SC и общим количеством элементов – не значимой (r_{11}=0,170; p=0,618). Это свидетельствует как о добросовестности разметчиков при субъективной оценке SC, так и о том, что этот показатель концептуально не тождественен общему количеству элементов в веб-интерфейсе.

2) Вторая группа скриншотов

Во второй группе скриншотов было размечено 495 штук, однако онлайн-сервис не смог обработать часть из них по техническим причинам (в случае AIM они достоверно не известны авторам). Соответственно, такие скриншоты были исключены из дальнейшего анализа, который производился с 368 скриншотами (74,3%), для которых «автоматические» метрики удалось получить полностью. Описательная статистика (средние значения и стандартные отклонения – SD) для двух групп метрик, задействованных в нашем исследовании, представлена в Табл. 2.

Табл. 2. Описательная статистика для метрик расположения элементов (вторая группа скриншотов веб-интерфейсов, n=368)

Группа метрик	Метрика	Ср. знач.	SD
«Ручные» метрики расположения элементов	Count of elements	85,89	37,81
	Balance	0,63	0,18
	Equilibrium	0,99	0,005
	Symmetry	0,85	0,05
«Автоматические» метрики расположения элементов	No. of UI elements	63,93	31,1
	Quadtree Dec balance	0,734	0,19
	Quadtree Dec equilibrium	0,99	0,00007
	Quadtree Dec symmetry	0,54	0,05

В рамках проверки двух групп метрик на мультиколлинеарность были посчитаны парные коэффициенты корреляции Пирсона, однако ни один из них не оказался выше, чем 0,5. Таким образом, включение «ручных» и «автоматических» метрик в единую модель является оправданным со статистической точки зрения.

B. Проверка гипотез

1) Эффект качества разметки

Нами было построено 33 модели линейной регрессии (для каждого из 11 разметчиков и 3 субъективных шкал), в каждой из которой факторами являлись 8 «ручных» метрик по типу элементов, которые мы приводили выше. Полученные для этих моделей значения R^2 представлены в Табл. 3.

Табл. 3. Показатели качества моделей для различных разметчиков и шкал

ID	Качество моделей (R^2)		
	Сложность (Complexity)	Эстетичность (Aesthetics)	Упорядочен-ность (Orderliness)
AA	0,108	0,149	0,114
GD	0,261	0,345	0,222
KK	0,261	0,252	0,152
MA	0,362	0,486	0,295
NE	0,316	0,488	0,416

PV	0,363	0,289	0,199
PE	0,165	0,568	0,611
SV	0,277	0,176	0,213
ShM	0,337	0,324	0,215
SoM	0,304	0,309	0,198
VY	0,204	0,110	0,169
Среднее	**0,269**	**0,318**	**0,255**

Согласно проведённому нами корреляционному анализу, показатель качества разметки SC не имел значимых корреляций (при уровне значимости α=0,05) с R^2 ни для одной из субъективных шкал. Даже для шкалы визуальной сложности (Complexity), которая традиционно связана с количеством элементов в интерфейсе, корреляция составила r_{11}=−0,062 (p=0,856).

Корреляционный анализ для Precision показал наличие значимой **отрицательной** корреляции с R^2 для субъективных шкал эстетичности (r_{11}=−0,768; p=0,006) и упорядоченности (r_{11}=−0,644; p=0,032). При этом для шкалы визуальной сложности корреляция значимой не оказалась (r_{11}=−0,051; p=0,883). Данные результаты представляются довольно неожиданными и заставляют нас отвергнуть гипотезу 1 о положительном влиянии качества разметки на качество моделей.

2) Эффект типа метрик

Прежде всего, мы построили регрессионные модели отдельно для «ручных» метрик расположения элементов (Табл. 4) и для «автоматических» (Табл. 5). В таблицах указаны только значимые факторы (из всего набора рассматриваемых метрик) и значения R^2 для моделей, построенных с этими факторами. Хотя все регрессионные модели были статистически значимы, значения R^2 для них оказались сравнительно невысокими (полужирным шрифтом выделены значения R^2, максимальные из двух групп метрик).

Табл. 4. Модели линейной регрессии для «ручных» метрик расположения

Субъективная шкала	**Значимые факторы (α=0,05)**	**R^2**
Complexity	Count of elements (p<0,001) Equilibrium (p=0,005)	**0,078**
Aesthetics	Count of elements (p=0,04) Equilibrium (p=0,01)	0,025
Orderliness	Count of elements (p=0,002) Equilibrium (p=0,04)	0,034
	Среднее:	0,046

Табл. 5. Модели линейной регрессии для «автоматических» метрик расположения

Субъективная шкала	**Значимые факторы (α=0,05)**	**R^2**
Complexity	No. of UI elements (p=0.008)	0,059
Aesthetics	No. of UI elements (p=0.02) Quadtree Dec balance (p<0.001)	**0,138**
Orderliness	No. of UI elements (p=0.02) Quadtree Dec balance (p<0.001) Quadtree Dec equilibrium (p=0.009)	**0,088**
	Среднее:	**0,095**

Из представленных результатов регрессионного анализа видно, что если среди «ручных» метрик для всех шкал значимыми оказались количество элементов и равновесие, то среди «автоматических» метрик имело место большее разнообразие. При этом, однако, количество элементов интерфейса было значимым для всех шкал и обеих групп метрик. Особо следует отметить, что средний показатель качества моделей для «автоматических» метрик оказался в 2 раза выше.

В рамках дальнейшего изучения предпочтительности метрик, мы объединили метрики из двух групп в единый набор факторов и применили метод обратного (Backwards) их отбора. В Табл. 6 представлены отобранные таким способом метрики с указанием общего их количества, относящегося к каждой группе. Значения R^2 при этом изменились незначительно (увеличение среднего R^2 составило +14%) и в таблице не отражены.

Табл. 6. Отобранные метрики обеих групп как факторы для регрессионных моделей

Субъективная шкала	**Значимые факторы (α=0,05)**	**Метрики**
Complexity	Count of elements (p<0.001) Equilibrium (p=0.01) Quadtree Dec equilibrium (p=0.03)	Ручные: 2 Авто: 1
Aesthetics	Count of elements (p=0.001) Equilibrium (p=0.02) Quadtree Dec balance (p=0.03) Quadtree Dec equilibrium (p<0.001)	Ручные: 2 Авто: 2
Orderliness	Count of elements (p<0.001) Equilibrium (p<0.001) Quadtree Dec balance (p=0.005) Quadtree Dec equilibrium (p=0.008)	Ручные: 2 Авто: 2

Из представленных результатов видно, что во всех случаях «ручные» метрики, отражающие количество элементов и равновесие их расположения, оказались значимыми. Примечательно, что «автоматическая» метрика количества элементов, ранее значимая (см. в Табл. 5), теперь для всех шкал уступила своё место аналогичной «ручной» метрике, что свидетельствует о преимуществе последней. Тем не менее, общее превышение количества отобранных «ручных» метрик (6 штук) над количеством отобранных «автоматических» метрик (5 штук) оказалось незначительным. В совокупности со значениями R^2, представленными в Табл. 4 и Табл. 5, это позволяет сделать вывод, что наша гипотеза 2 о преимуществе метрик ручной разметки не может быть принята.

IV. Выводы и заключение

Итак, в нашей работе мы спланировали и осуществили экспериментальное исследование с целью изучения влияния типа разметки веб-интерфейсов (ручная или автоматическая) её качества на предсказательную точность моделей по субъективным шкалам визуальной сложности, эстетичности и упорядоченности. Мы

предполагали, что качественная ручная разметка во всех случаях позволяет достигнуть более высокого качества моделей, однако наши гипотезы не могут быть приняты по результатам проведённого статистического анализа полученных в эксперименте данных.

A. Эффект качества разметки

В полном противоречии с выдвинутой нами гипотезой 1, мы обнаружили не значимые или статистически значимые отрицательные корреляции между показателями качества разметки и качества моделей, предсказывающих оценки по субъективным шкалам визуального восприятия веб-интерфейсов. В рамках проверки валидности наших результатов, мы проверили некоторые возможные альтернативные объяснения полученного эффекта, однако ни одно из них не оказалось убедительным:

1. Нарушение валидности при распределении оцененных скриншотов. Однако было выяснено, что средние оценки для скриншотов, выданных различным разметчикам, практически не имели статистически значимых различий.
2. Случайности и небрежности при разметке скриншотов. Однако показатели корректности (88,7%) и субъективной полноты (77,8%) разметки являются сравнительно высокими.
3. Случайности и небрежности при оценке качества разметки. Однако показатели Precision и SC имели ожидаемые значения корреляций с другими факторами, в том числе объективными. Так, SC не имело значимой корреляции с общим количеством элементов интерфейса на скриншоте.
4. Нарушения в концептуальной валидности субъективных шкал. Однако оказалось, что корреляции между оценками по шкалам в целом совпадают с теми, что встречаются в источниках. Так, корреляция между шкалами сложности и упорядоченности, как и ожидалось, оказалась отрицательной и значимой (p=0,001).
5. Ошибочность использования коэффициента корреляции Пирсона для R^2. Однако расчет с коэффициентом корреляции Кендалла для порядковых шкал дал аналогичные результаты.

B. Эффект типа метрик

Осуществленная в нашем исследовании проверка гипотезы 2 о том, что ручная разметка скриншотов веб-интерфейсов позволит получить модели для субъективных шкал визуального восприятия с более высоким уровнем качества, чем в случае с распознаванием скриншотов автоматическими сервисами, основанными на технологиях компьютерного зрения, не дала однозначного результата, позволяющего принять гипотезу. Построенные нами модели линейной регрессии оказались значимыми как для «ручных», так и для «автоматических» метрик. При этом средние значения R^2 были на 107% выше именно для «автоматических» метрик, хотя в целом остались сравнительно низкими. Однако при формализованном отборе объединенных метрик обеих как факторов для моделей, общее количество отобранных «ручных» метрик оказалось несколько больше, что свидетельствует о некотором их преимуществе.

При этом объединение метрик обеих групп позволило несколько улучшить качество моделей (на 14%) – предположительно, за счёт более широкого охвата характеристик веб-интерфейсов. Так, и «ручная», и «автоматическая» метрики равновесия для всех шкал оказывались значимыми одновременно – по-видимому, вследствие различия в алгоритмах расчета равновесия в [13] и [7] соответственно. Это еще раз подтверждает важность тщательного выбора признаков для моделей в сфере человеко-комьютерного взаимодействия, в условиях ограниченности объемов данных.

C. Выводы и практические рекомендации

Результаты нашего исследования еще раз подтверждают наличие трудностей при определении, что такое качественные входные данные для моделей, описывающих поведение людей. Обнаруженные нами для двух из трёх субъективных шкал визуального восприятия статистически значимые отрицательные корреляции с качеством разметки скриншотов веб-интерфейсов подчеркивают важность адекватных механизмов контроля качества работы разметчиков. Представляется вполне вероятным, что «ленивые» разметчики пропускали или некорректно отмечали прежде всего именно те элементы, которые меньше влияют и при восприятии веб-интерфейсов пользователями, что в итоге привело к лучшей точности моделей, описывающих субъективное эстетическое впечатление. Действительно, для шкалы визуальной сложности, которая является более объективной, эффекта отрицательного влияния качества разметки обнаружено не было. Можно сказать, что полученные нами результаты находятся в русле новейших идей о том, что модели ИИ для большей реалистичности и успешности работы **должны** содержать когнитивные искажения, свойственные человеку [14].

В свете этого, повышенный интерес вызывает автоматизированная разметка веб-интерфейсов, которая по объективным показателям качества (например, по полноте детекции элементов интерфейса) обычно уступает ручной. Так как в нашем исследовании было обнаружено, что метрики, полученные из автоматической разметки, позволили получить даже более качественные модели, мы рекомендуем исследователям в сфере человеко-комьпютерного взаимодействия и практикующим UI/UX дизайнерам больше экспериментировать с инструментами для автоматической параметризации интерфейсов. Хотя в нашем исследовании затраты времени на разметку одного скриншота веб-интерфейса автоматизированным сервисом и человеком-разметчиком были вполне сопоставимы (20-30 минут), трудозатраты были ниже в первом случае. Следует однако отметить, что автоматизированным способом не

удалось обработать около 25% задействованных скриншотов, что представляется довольно высоким уровнем потерь. Впрочем, и при организации процесса «ручной» разметки могут возникать определенные потери, связанные с контролем её качества и отбраковкой «небрежных» разметчиков [15].

Что касается эффекта рассмотренных метрик на субъективное восприятие, который может представлять практический интерес для веб-дизайнеров, по результатам нашего исследования можно заметить следующее:

1. Количество элементов интерфейса являлось наиболее широко влияющим фактором для всех шкал. При этом вычисление его по результатам «ручной» разметки (Count of elements) давало преимущество по сравнению с автоматическим способом (No. of UI elements).
2. Равновесие, как характеристика пространственного расположения элементов веб-интерфейса, являлось следующим по универсальности влияния фактором. При этом «ручной» и «автоматический» способ вычисления этой метрики способны дополнять друг друга.
3. Метрики баланса и симметрии, вычисленные по результатам «ручной» разметки, не оказывали значимого влияния на рассмотренные шкалы восприятия (в отличие, например, от «автоматической» метрики баланса). Это свидетельствует о важности не только типа и качества разметки, но и выбора конкретных метрик.
4. Расширение набора используемых метрик не всегда дает значительный эффект и должно производиться на основе знаний как в сфере человеко-компьютерного взаимодействия в целом, так и особенностей конкретной предсказываемой шкалы субъективного восприятия. Это подчеркивает важность дальнейших исследований в выбранной нами предметной области.

D. Ограничения и планы дальнейших исследований

Хотя большинство угроз для валидности нашего исследования были рассмотрены выше и признаны несостоятельными, можно отметить еще ряд потенциальных ограничений относительно полученных нами результатов. Прежде всего, это довольно низкие значения R^2 в наших регрессионных моделях. На наш взгляд это ожидаемо и связано как с небольшим объемом данных, задействованных в нашем исследовании (часть из которых, к тому же, в соответствии с планом были невысокого качества), так и с сознательно ограниченным перечнем метрик, выступающих в качестве факторов: 8 метрик для первой партии скриншотов и 8 – для второй. Мы полагаем, что такой подход оправдан, поскольку в соответствии с целями исследования и выдвинутыми гипотезами нас интересовали не абсолютные показатели качества моделей, а их сравнение для различных групп. Тем не менее, мы должны предупредить читателя, что подобные модели вряд ли следует использовать в реальных проектах по проектированию и разработке веб-интерфейсов.

Еще одна потенциальная причина нарушения валидности исследования – не полностью случайные выборки и назначения. В частности, около 25% скриншотов, которые были исключены из второй партии как не прошедшие обработку автоматизированным сервисом, вполне могли иметь значения метрик, отличающиеся от тех, что характеризовали оставшиеся скриншоты, внося, таким образом, искажения в анализ. В целом, причины отказы автоматизированных сервисов и особенности вызывающих их скриншотов заслуживают отдельного изучения.

Планы наших дальнейших исследований включают в себя поиск эффективного «распределения обязанностей» между разметчиками-людьми и автоматизированными инструментами для параметризации человеко-компьютерных интерфейсов. Критерием эффективности при этом может выступать соотношение между достигаемым уровнем качества моделей и затратами, связанными со сбором и обработкой данных. Мы полагаем, что наша текущая статья содержит результаты в этом направлении, представляющие определенный интерес для исследователей в сфере машинного обучения, инженеров по данным и дизайнеров интерфейсов, работающих в ИТ-промышленности.

Благодарности

Авторы хотели бы выразить благодарность всем участникам экспериментального исследования, а также Анне Бортниковой, участвовавшей в обработке данных.

Библиография

[1] Priestley M., O'donnell F., Simperl E. A survey of data quality requirements that matter in ML development pipelines // ACM Journal of Data and Information Quality, 2023. No. 15(2). P. 1-39.

[2] Gudivada V.. Apon A., Ding J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations // Int. J. Adv. Softw. 2017. No. 10. P. 1–20.

[3] Jain A., Montanari A., Sasoglu E. Scaling laws for learning with real and surrogate data // arXiv preprint. 2024. arXiv:2402.04376.

[4] Miniukovich A., Marchese M. Relationship between visual complexity and aesthetics of webpages // Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020. P. 1–13.

[5] Wang X., Tong M., Song Y., Xue C. Utilizing Multiple Regression Analysis and Entropy Method for Automated Aesthetic Evaluation of Interface Layouts // Symmetry. 2024. No. 16(5). 523.

[6] Bakaev M., Heil S., Khvorostov V., Gaedke M. Auto-extraction and integration of metrics for web user interfaces // Journal of Web Engineering. 2018. No. 17(6–7). P. 561-590.

[7] Oulasvirta A., De Pascale S., Koch J., Langerak T., Jokinen J., Todi K., Laine M., Kristhombuge M., Zhu Y., Miniukovich A., et al. Aalto Interface Metrics (AIM): A service and codebase for computational GUI evaluation // Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings, Berlin, Germany, 14–17 October 2018. P. 16–19.

[8] Heil S., Bakaev M., Gaedke M. Assessing completeness in training data for image-based analysis of web user interfaces // CEUR Workshop Proceedings. 2019. Vol. 2500.

[9] Gardey J. C., Grigera J., Rodriguez A., Garrido A. UX-Analyzer: Visualizing the interaction effort for web analytics // Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing. 2024. P. 1774-1780.

[10] Bakaev M., Khvorostov V. Quality of Labeled Data in Machine Learning: Common Sense and the Controversial Effect for User Behavior Models // Engineering Proceedings. 2023. No. 33(1). 3. https://doi.org/10.3390/engproc2023033003.
[11] Bortnikova A., Bakaev M. Who Is to Err? A Limited Effect of Data Labeling in Prediction of Users' Subjective Impressions of Web Designs // IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). Novosibirsk, Russia. 2024. – in print.
[12] Bakaev M., Speicher M., Heil S., Gaedke M. I Don't Have That Much Data! Reusing user behavior models for websites from different domains // International Conference on Web Engineering. P. 146-162. Cham: Springer International Publishing, 2020.
[13] Habay A. A. A Systematic Literature Review of Visual Design Metrics for Graphical User Interfaces // Louvain School of Management, Université catholique de Louvain, 2020. Prom.: Vanderdonckt, Jean. http://hdl.handle.net/2078.1/thesis:25647.
[14] Hagendorff T., Fabi S. Why we need biased AI: How including cognitive biases can enhance AI systems // Journal of Experimental & Theoretical Artificial Intelligence. 2024. No. 36(8). P. 1885-1898.
[15] Saravanos A., Zervoudakis S., Zheng D., Stott N., Hawryluk B., Delfino D. The hidden cost of using Amazon Mechanical Turk for research // International Conference on Human-Computer Interaction; Springer International Publishing: New York, NY, USA, 2021. P. 147–164.

Бакаев Максим Александрович, кандидат технических наук, доцент, Новосибирский государственный технический университет, Россия, ORCID 0000-0002-1889-0692 (bakaev@corp.nstu.ru).
Хворостов Владимир Александрович, старший преподаватель, Новосибирский государственный технический университет, Россия, ORCID 0000-0002-7507-6662 (bakaev@corp.nstu.ru).

How web interface labeling quality affects ML models predicting users' subjective impressions

Maxim Bakaev, Vladimir Khvorostov

***Abstract*— Training data quality is widely recognized as the main pre-requisite for constructing successful Machine Learning (ML) models. However, the concrete aspects of the data quality vary for different domains and outcomes to be predicted by the models. In Human-Computer Interaction (HCI), the common knowledge is that labor-intensive manual labeling of graphical user interfaces (UIs), i.e. identification of visual elements in them, allows to predict users' subjective impressions more accurately. At the same time, computer vision-based services for automated parametrization of UIs gain in popularity. In our paper, we describe an experimental study with over 200 participants and 1000 web UI screenshots, which were assessed on 3 subjective impressions scales: Complexity, Aesthetics and Orderliness. In order to compare the effects of metrics derived from manual vs. automated labeling, as well as of increased data quality in the manual labeling, we calculated in total 16 metrics subsequently used as factors in the predictive ML models. Our results suggest that Pearson correlation of input data quality and the models' quality was highly significant and negative for Aesthetics (-0.768) and Orderliness (-0.644). Neither could we identify consistent advantages of the "manual" metrics over "automated" ones, except for the number of UI elements, in which the automated services were somehow lacking. Our conclusions regarding the labeling of the UIs and the applicability of the considered metrics might be of interest to both ML-HCI researchers and practicing UI/UX designers.**

***Keywords*— web interfaces, data quality, computer vision, machine learning, human-computer interaction.**

REFERENCES

[1] M. Priestley, F. O'donnell, and E. Simperl, "A survey of data quality requirements that matter in ML development pipelines," *ACM Journal of Data and Information Quality*, no. 15(2), pp. 1-39, 2023.

[2] V. Gudivada, A. Apon, J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *Int. J. Adv. Softw*, no. 10, pp. 1–20, 2017.

[3] A. Jain, A. Montanari, and E. Sasoglu, "Scaling laws for learning with real and surrogate data," *arXiv preprint*, 2024. arXiv:2402.04376.

[4] A. Miniukovich, and M. Marchese, "Relationship between visual complexity and aesthetics of webpages," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 25–30 April 2020, pp. 1–13, 2020.

[5] X. Wang, M. Tong, Y. Song, and C. Xue, "Utilizing Multiple Regression Analysis and Entropy Method for Automated Aesthetic Evaluation of Interface Layouts," *Symmetry*, no. 16(5), 523, 2024.

[6] M. Bakaev, S. Heil, V. Khvorostov, and M. Gaedke, "Auto-extraction and integration of metrics for web user interfaces," *Journal of Web Engineering*, no. 17(6–7), pp. 561-590, 2018.

[7] A. Oulasvirta, S. De Pascale, J. Koch, T. Langerak, J. Jokinen, K. Todi, M. Laine, M. Kristhombuge, Y. Zhu, A. Miniukovich, et al, "Aalto Interface Metrics (AIM): A service and codebase for computational GUI evaluation," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*, Berlin, Germany, 14–17 October 2018, pp. 16–19, 2018.

[8] S. Heil, M. Bakaev, and M. Gaedke, "Assessing completeness in training data for image-based analysis of web user interfaces," in *CEUR Workshop Proceedings*, vol. 2500, 2019.

[9] J. C. Gardey, J. Grigera, A. Rodriguez, and A. Garrido, "UX-Analyzer: Visualizing the interaction effort for web analytics," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pp. 1774-1780, 2024.

[10] M. Bakaev, and V. Khvorostov, "Quality of Labeled Data in Machine Learning: Common Sense and the Controversial Effect for User Behavior Models," *Engineering Proceedings*, no. 33(1), 3, 2023, https://doi.org/10.3390/engproc2023033003.

[11] A. Bortnikova, and M. Bakaev, "Who Is to Err? A Limited Effect of Data Labeling in Prediction of Users' Subjective Impressions of Web Designs," in *IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*. Novosibirsk, Russia, 2024, in print.

[12] M. Bakaev, M. Speicher, S. Heil, S., and M. Gaedke, "I Don't Have That Much Data! Reusing user behavior models for websites from different domains," in *International Conference on Web Engineering*, Cham: Springer International Publishing, pp. 146-162, 2020.

[13] A. A. Habay, "A Systematic Literature Review of Visual Design Metrics for Graphical User Interfaces," *Louvain School of Management, Université catholique de Louvain*, 2020. Prom., Vanderdonckt, Jean, 2020, http://hdl.handle.net/2078.1/thesis:25647.

[14] N. Hagendorff, and S. Fabi, "Why we need biased AI: How including cognitive biases can enhance AI systems," Journal of Experimental & Theoretical Artificial Intelligence, no. 36(8), pp. 1885-1898, 2024.

[15] A. Saravanos, S. Zervoudakis, D. Zheng, N. Stott, B. Hawryluk, and D. Delfino, "The hidden cost of using Amazon Mechanical Turk for research," in *International Conference on Human-Computer Interaction*, Springer International Publishing, New York, NY, USA, pp. 147–164, 2021.

Bakaev Maxim Alexandrovich, PhD (Technical Sciences), Associate Professor, Novosibirsk State Technical University, Russia, ORCID 0000-0002-1889-0692 (bakaev@corp.nstu.ru).

Khvorostov Vladimir Alexandrovich, Senior Assistant Professor, Novosibirsk State Technical University, Russia, ORCID 0000-0002-7507-6662 (xvorostov@corp.nstu.ru).