УДК 004

Разработка системы типологизации пользователей Web-ресурсов

Кукушкин Илья Михайлович
Волжский политехнический институт (филиал) ВолгГТУ
студент

Лясин Дмитрий Николаевич
Волжский политехнический институт (филиал) ВолгГТУ
к.т.н., доцент «Информатика и технология программирования»

Аннотация
В статье рассматривается реализация системы типологизации пользователей Web-ресурсов. Для иллюстрации работоспособности были приведены скриншоты проведенных экспериментов для системы типологизации пользователей Web-ресурсов.
Ключевые слова: Типологизация, кластеризация, Web.

Development of a system for typology of Web-resources users

Kukushkin Ilia Mihailovich
Volzhskiy Polytechnical Institute, branch of the Volgograd State Technical University
Student

Lysin Dmitry Nikolaevich
Volzhskiy Polytechnical Institute, branch of the Volgograd State Technical University
Ph. D. associate Professor of "Informatics and technology of programming»

Abstract
The article discusses the implementation of the system of typology of users of Web-resources. To illustrate the efficiency, screenshots of the experiments carried out for the system of typology of users of Web resources were presented.
Keywords: Typology, clustering, Web.

При попытке получения знаний из Web мы не можем ориентироваться на строгие структуры и компоненты, так как в Интернете присутствует огромное количество распределённой, гетерогенной, неструктурированной и динамически изменяющейся информации. Несмотря на это, интернет ресурсы научились быть ближе к интернет пользователю, перестали быть изолированными от них. Как только интернет пользователь заходит на интернет ресурс, он сразу оставляет свой след: тем или иным образом

становятся известны его местоположение (география), персональные данные (пол, возраст и т.д.), его история поиска.

Персональные данные пользователей очень ценятся на различных Web ресурсах и у рекламодателей. Данные сведения способствуют продаже рекламы рекламодателей и так же с помощью этих данных можно без проблем разбить пользователей на типы (мужчина - женщина; молодой – пожилой, местный житель – иногородний, семейный человек - холостяк), тем самым увеличив качество рассылки рекламы за счет донесения до пользователя адаптируемой для него информации. Как пример можно привести технологию «Крипта» от Yandex [1]. Проблема типологизации пользователей Web ресурсов заключается в основном в неточности данных из-за не проработанности алгоритма и человеческого фактора, из-за которого могут передаваться неточные данные для обработки. [4] Данная проблема имеет большую актуальность, так как типологизация пользователей активно используется в коммерческих целях, для разбиения пользователей на группы и адаптации информации (таргетирования) для данных групп пользователей. Особый интерес представляет автоматизация процесса составления типологизированного профиля пользователя на основе анализа его поведения, а не предъявленных им при идентификации данных. Разработке алгоритма формирования подобного профиля и его программная реализация для использования на конкретном web-ресурсе была посвящена настоящая работа.

Термин «Типологизации пользователей» характеризует как проведения группировки пользователей по схожим между друг с другом критериями, из которых в последствии строятся группы пользователей с общими характеристиками или интересами [3]. На основе выявленных шаблонов поведения, типичных для тех или иных категорий пользователей, алгоритм типологизации должен давать вероятностную оценку принадлежности текущего пользователя web-ресурса к тому или иному типу, тем самым определяя тот контент, который будет выдаваться пользователю [2,5].

Преимущества и плюсы внедрения алгоритма типологизации пользователей:

1. SEO-продвижение web-ресурса за счет внешних (поведенческих) факторов;
2. Повышение эффективности группировки данных и таргетирования контента пользователям;
3. Уменьшение затрат на продвижения рекламы в места, не относящиеся к тематике сообщества;

Среди теоретических и практических проблем автоматизированной типологизации пользователей особое место занимает проблема сбора и систематизации входных данных для подачи их в алгоритм работы системы. В настоящее время существуют метрики для оценки трудоемкости работ, выполненных программистом, ориентированных на анализ программного кода. Тем не менее, решение проблемы оценки трудоемкости и стоимости

выполненных работ по типологизации пользователей является актуальной задачей.

Наиболее точную оценку трудоемкости и эффективности автоматизации типологизации пользователей можно выявить, сравнив существующие программные продукты, реализующие эту задачу в той или иной мере, на основе следующих критериев:

- форма отображения данных;
- способ ввода информации;
- хранение информации о пользователях;
- формирование групп пользователей;
- качество типологизации.

Для исследования трудоемкости и эффективности автоматизации процесса типологизации пользователей могут быть использованы следующие программные продукты: Majento, KeyAssort, Penguin, TopSite, Rush Analytics. Проведем сравнительный анализ данных программных продуктов с точки зрения их применения для решения задачи получения метрик физической схемы базы данных.

В качестве критериев для сравнительного анализа программных продуктов, выберем следующие:

1. A_1 – форма отображения данных;
2. A_2 – Способ ввода информации;
3. A_3 – Хранение информации о пользователях;
4. A_4 – Формирование групп пользователей;
5. A_5 – Качество типологизации.

Для определения весов критериев воспользуемся аналитической иерархической процедурой Саати. Правила заполнения матрицы парных сравнений представлены в таблице.

Таблица 1. Значения коэффициентов матрицы парных сравнений

X_{ij}	Значение
1	i-ый и j-ый критерий примерно равноценны
3	i-ый критерий немного предпочтительнее j-го
5	i-ый критерий предпочтительнее j-го
7	i-ый критерий значительно предпочтительнее j-го
9	i-ый критерий явно предпочтительнее j-го

Матрица парных сравнений, средние геометрические и веса критериев представлены в таблице 2.

Таблица 2. Матрица парных сравнений, средние геометрические и веса критериев

	A1	A2	A3	A4	A5	Среднее геометрическое	Веса критериев

A_1	1,00	3,00	0,33	0,14	0,14	1,25	0,08
A_2	0,33	1,00	5,00	1,00	1,00	1,38	0,20
A_3	3,00	0,20	1,00	1,00	0,33	1,25	0,13
A_4	7,00	1,00	1,00	1,00	4,00	1,95	0,35
A_5	7,00	1,00	3,00	0,25	1,00	1,84	0,25
Сумма						7,66	1,00

Диаграмма весовых коэффициентов для критериев A_1, A_2, A_3, A_4, A_5 представлена на рис. 1.

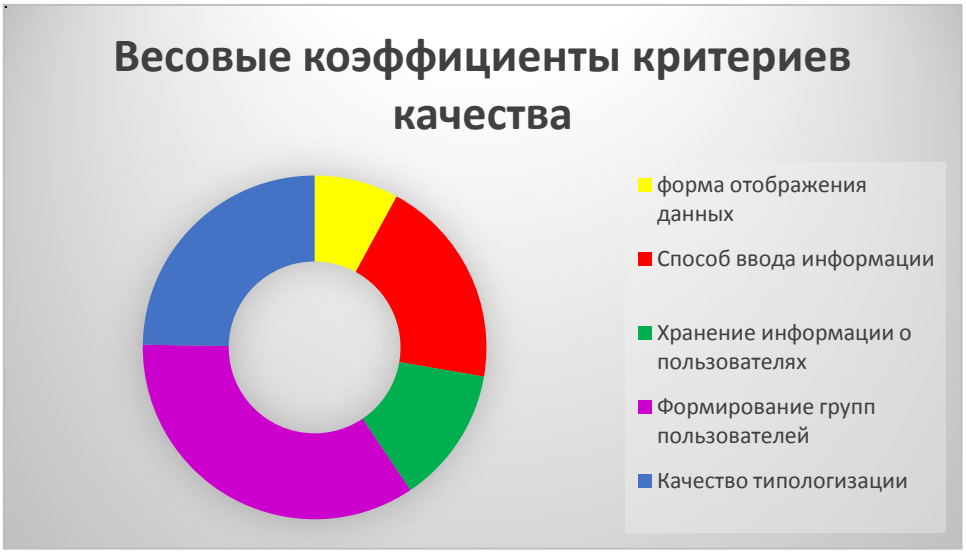

Рисунок 1. Весовые коэффициенты критериев качества

Выполним проверку матрицы попарных сравнений на непротиворечивость.

Суммы столбцов матрицы парных сравнений:

R_1=22.8; R_2=4.9; R_3=9.8; R_4=4.6; R_5=7.8.

Путем суммирования произведений сумм столбцов матрицы на весовые коэффициенты альтернатив рассчитывается вспомогательная величина L = 6.25. Индекс согласованности ИС = (L-N) /(N-1) = (6.25 - 5) /(5-1) = 0.31.

Величина случайной согласованности для размерности матрицы парных сравнений: СлС = 1.6.

Отношение согласованности ОС=ИС/СлС = 0.19. не превышает 0.2, поэтому уточнение матрицы парных сравнений не требуется.

Используя полученные коэффициенты определим интегральный показатель качества для программных продуктов автоматизированной типологизации пользователей.

1. Majento;
2. KeyAssort;
3. Penguin;

4. TopSite;
5. Rush Analytics.

Выберем категориальную шкалу от 0 до 7 (где 0 – качество не удовлетворительно, 7 – предельно достижимый уровень качества на современном этапе) для функциональных возможностей программных продуктов.

Значения весовых коэффициентов a_i соответствующие функциональным возможностям продуктов:

1. Форма отображения данных: $a_1 = 0.08$;
2. Способ ввода информации: $a_2 = 0.2$;
3. Хранение информации о пользователях: $a_3 = 0.13$;
4. Формирование групп пользователей: $a_4 = 0.35$;
5. Качество типологизации: $a_5 = 0.25$; где $\sum a_i = 1$.

Определим (по введенной шкале) количественные значения функциональных возможностей $X_{ij.}$ Вычислим интегральный показатель качества для каждого программного продукта.

Таблица 3. Интегральные показатели качества

Критерии	Весовые коэффициенты	Программные продукты					Базовые значения	Система типологизации пользователей Web-ресурсов
		Key-Assort	Majento	Penguin	Top Site	RA		
форма отображения данных	0,08	2,00	1,00	1,00	4,00	5,00	2,80	6,00
Способ ввода информации	0,20	4,00	3,00	2,00	4,00	6,00	3,80	7,00
Хранение информации о пользователях	0,13	4,00	3,00	1,00	1,00	6,00	2,90	7,00
Формирование групп пользователей	0,35	1,00	1,00	1,00	1,00	5,00	1,80	6,00
Качество типологизации	0,25	4,00	3,00	4,00	1,00	3,00	2,20	5,00
Интегральный показатель качества Q		2,80	1,80	1,38	1,24	4,89	2,36	6,08

Лепестковая диаграмма значений характеристик качества функциональных возможностей (критериев) представлена на рисунке 2.

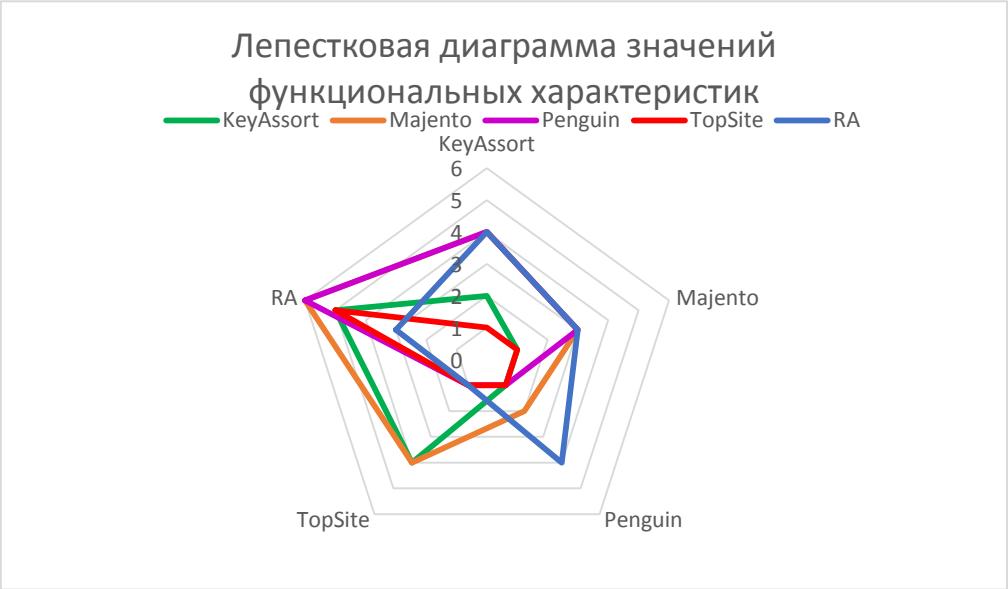

Рисунок 2. Лепестковая диаграмма значений функциональных характеристик

Сравнительный анализ программных продуктов для автоматизированной типологизации пользователей показал, что только два из пяти рассмотренных программных средств имеют значения интегрального показателя качества, превышающего базовое значение, что и доказывает важность и актуальность данной работы.

Анализ пользовательского интерфейса системы проводился с помощью средства CogTool. Были предусмотрены задачи, которую необходимо было выполнить. С помощью средства CogTool было спрогнозировано время, затрачиваемое на выполнение определенной задачи.

В таблице 4 приведены задачи, поставленные пользователям, и прогнозируемое время выполнения задачи.

Таблица 4. Общий результат эксперимента

№ эксперимента	**Задача**	**Время выполнения задачи, сек.**
1	Группировка и сохранение данных	9,1
2	Группировка и очистка данных, после идет сохранение данных	9,4
3	Сохранение данных в исходной таблице и группировка данных	8,4

При помощи средства CogTool, было установлено, что

спрогнозированное время, затрачиваемое на выполнения необходимых задач невелико. Следовательно, интерфейс системы типологизации пользователей Web-ресурсов не вызывает затруднений для пользователя, что свидетельствует о легком и удобном интерфейсе для пользователей.

Оценка эффективности и трудоемкости процесса сбора и систематизации входных данных для типологизации пользователей может быть определена на следующих этапах проектирования:

1. При сборе и использовании входных данных для типологизации;
2. При определении критериев для выполнения группировки пользователей;
3. При редактировании ранее собранных данных с сайта.

Анализ ресурсозатратности системы проводился с помощью средства Visual Studio Analyzer. Были предусмотрены задачи, которую необходимо было выполнить. С помощью средства Visual Studio Analyzer было спрогнозировано ресурсов процессора, затрачиваемое на выполнение определенной задачи.

В таблице 5 приведены задачи, поставленные пользователям, и прогнозируемое общее процессорное время выполнения задачи (в эксперименте используется 30 записей).

Таблица 5. Общий результат эксперимента

№ эксперимента	**Задача**	**Общее процессорное время выполнения задачи, мсек.**
1	Группировка и сохранение данных	4506
2	Группировка и очистка данных	3278
3	Сохранение данных в исходной таблице и группировка данных	4222

При помощи средства Visual Studio Analyzer, было установлено, что количество процессорного времени, затрачиваемое на выполнения необходимых задач невелико. Следовательно, процесс системы типологизации пользователей Web-ресурсов не требует слишком мощных мощностей.

Программное обеспечение, необходимое для функционирования программы

Для работы программы необходимо:

1. .NET Framework 4.0;
2. ОС Windows 7, 8, 10.
3. Собственный хостинг или доступ к внешнему хостеру.
4. Браузер с подддержкой HTML5 и CSS3.
5. SQL - DbFordge

Программа написана на языке программирования C#. Скрипт по сборке данных с пользователя написан на языке PHP.

Преимущества разработанной программной системы:

1. ПО легко модифицируется в случае изменения требований;
2. Система имеет дружественный пользователю интерфейс;
3. Система предоставляет отчет в виде текстового файла;
4. Система выдает информацию о работе (сохранения, удаления).
5. Предоставляемая информация понятна пользователя;
6. в системе предусмотрен ввод и получение информации;
7. Сбои в работе ПО, а также аппаратных средств не замедляют работу системы.

Функционал программной системы типологизации пользователей web-ресурсов по их поведению описан в виде диаграммы вариантов использования (рис.3).

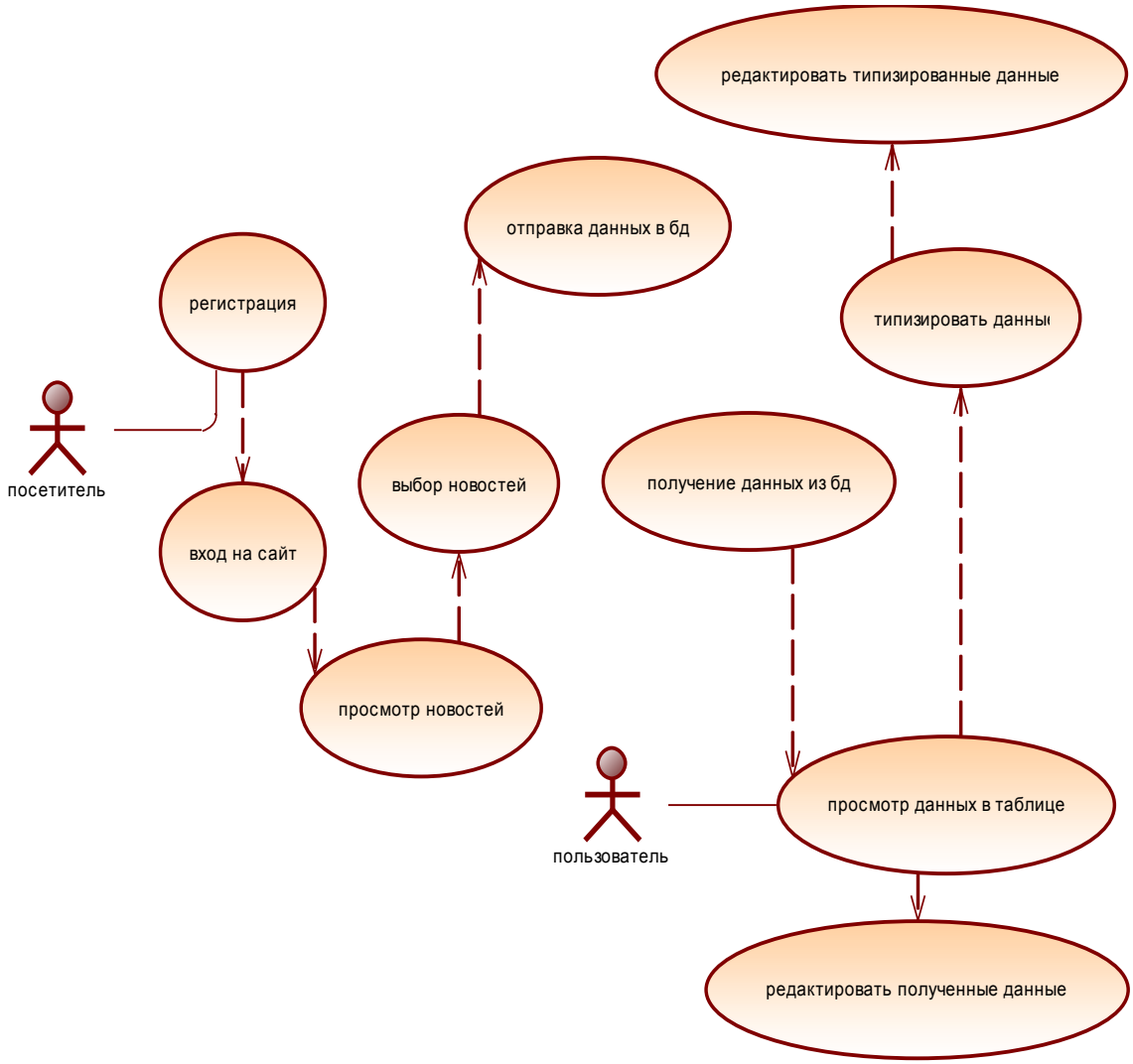

Рисунок 3. Общая диаграмма вариантов использования системы

Исходными данными для типологизации являются информационные блоки, получаемые в заголовках HTTP-запросов пользователя, а также данные о периодичности этих запросов:

- Информация о ip адресе;
- Информация о ссылке, откуда пришел посетитель;
- Информация о браузере, через который заходил посетитель;
- Время, и дата, когда заходил посетитель;
- Электронная почта зарегистрированного пользователя (у зарегистрированного пользователя);
- Имя пользователя (у зарегистрированного пользователя);
- Новости, выбранные пользователем;
- Пол пользователя (у зарегистрированного пользователя);
- Возраст пользователя (у зарегистрированного пользователя).

На основе анализа предметной области были выявлены входные данные в виде сущностей и их атрибутов, непосредственно влияющие на процесс типологизации. Эти данные позволяют сформировать обобщенный профиль пользователя, который сопоставляется с типовыми профилями и дается вероятностная оценка принадлежности пользователя к тому или иному типу.

Особенности функционирования разработанной системы:

- Программа позволяет добавлять данные как вручную, так и через файл или автоматически через базу данных.
- Программа предоставляет возможность пользователю сохранить таблицу с данными в xml файл.
- Программа позволяет давать редактировать данные в таблице с данными (в редактирование входит добавление и удаление строк).
- Программа позволяет полученные данные от пользователя разбивать на группы (типизировать) и выводить в таблицу, в которой они распределены по группе с помощью вложенности.
- Программа позволяет получать данные от сборщика – php-модуля, работающего в контексте наблюдаемого web-ресурса.

Вывод

Основной целью работы, описываемой в данной статье, было повышение эффективности типологизации пользователей Web ресурсов для адаптации предоставляемого им контента.

В ходе работ получены следующие теоретические и практические результаты:

1. С использованием современных технологий проектирования информационных систем разработано описание системы типологизации пользователей web-ресурсов.
2. На основе разработанной математической модели разработаны алгоритмы и программная реализация системы типологизации пользователей web-ресурсов на базе модельного сайта.
3. Выполнена оценка эффективности предложенных методов и алгоритмов. На основании оценки сделан вывод, что разработанная система

позволяет более эффективно осуществлять типологизацию пользователей web-ресурсов.

Библиографический список

1. 1Мороховец Ю.Е. Исследование и разработка методов автоматической кластеризации интернет-пользователей и интернет- ресурсов. М., 2014. 306 с.
2. Филиппов С. А., Ковалев Д. Ю. Кластеризация профилей пользователей в рекомендательных системах поддержки жизнеобеспечения на основе реальных неявных данных // Труды XVIII Международной конференции. DAMDID/RCDL’2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016. 2016. С.98-103.
3. Алфимцев А.Н., Девятков В.В., Сакулин С.А. Персонализация в гипертекстовых сетях на основе распознавания действий пользователей и нечеткого агрегирования // Вестник МГТУ им.Баумана, Сер. «Приборостроение». 2014. №3. 121 с.
4. Киселев М. В. Оптимизация процедуры автоматического пополнения веб-каталога // Megaputer Intelligence, 2014. С. 22.
5. Прохорова А.М. Роль методов анализа и прогнозирования поведения пользователей на образовательном портале // Вестник Научных Конференций, 2017. №89. С.3