Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК_____

Прашкович Артем Юрьевич

Задача профилирования интернет-пользователя

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1 40-80-02 Системный анализ, управление и обработка информации

_____
(подпись магистранта)

Научный руководитель
Герман Олег Витольдович
кандидат технических наук, доцент

_____
(подпись научного руководителя)

Минск 2016

ВВЕДЕНИЕ

Ежедневно пользователи всемирной паутины получают массу информации. Иногда она бывает полезной, однако большая ее часть не представляет интереса для человека, который ее просматривает. Это связано с тем, что каждый человек – это личность, он уникален. То, что представляется занимательным одному, другому может показаться бессмысленным. Поэтому у каждой статьи найдется благодарный читатель, однако вероятность того, что им будет именно тот, кому она в данный момент предъявлена, очень невелика.

Для того чтобы изменить такое положение дел, необходимо обдуманно подойти к тому, какую информацию показывать конкретному пользователю, вместо того, чтобы действовать наугад или полагаться исключительно на предпочтения редактора, как это часто происходит. Чтобы принять взвешенное решение о выборе данных для отображения, необходима информация о потребителе, которая позволила бы сделать вывод о его предпочтениях и интересах, либо помогла по другим признакам отсечь информацию, которая заведомо не представляет ценности для данного Интернет- пользователя.

Профилирование пользователя – достаточно известная и распространенная задача, решаемая в настоящее время различными способами [1, 2]. Суть профилирования заключается в том, что пользователю произвольного информационного Интернет-ресурса предоставляют не весь контент, а в первую очередь то, в чем, предположительно, он может быть заинтересован. Предположение обычно строится на основе многих факторов: документов, которые пользователь смотрел в прошлом, его географического положения, приватной информации из личного профиля пользователя и т.д. В данной работе для решения задачи профилирования определяются тематики, интересные пользователю, и оценивается близость того или иного документа к необходимым тематикам.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования.

Целью настоящей диссертационной работы является применение математических методов в задачах классификации.

Достижение данной цели потребовало решения следующих задач:

1. Изучить существующие стандарты и подходы для профилирования пользователей.
2. Изучить синтаксис и семантику существующих алгоритмов и программных продуктов для профилирования пользователей информационных ресурсов.
3. Разработать алгоритм, базирующийся на теории нечетких нейронных сетей, позволяющий решать задачу классификации, основываясь на обучающих таблицах.
4. Спроектировать и разработать библиотеку для кластеризации пользователей на основе страниц, которые он посещал.

Объектом исследования являются методы и алгоритмы профилирования пользователей. Предметом исследования являются методы и алгоритмы классификации.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническими заданиями и планами работ кафедры «Информационных технологий автоматизированных систем» по теме «Задача профилирования интернет-пользователя».

Научная новизна

Разработан оригинальный метод классификации многомерных объектов на основе составления системы линейных алгебраических неравенств и получения приближенного нечеткого решения на базе техники Монте-Карло. Данный метод позволяет построить нечеткий классификатор на основе четкого классификатора.

Практическая значимость

Разработано программное средство для классификации пользователей интернет-сайта использующее созданный алгоритм.

Личный вклад магистранта

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя О. В. Герман, заключается в формулировке целей и задач исследования.

Опубликованность результатов диссертации

1. Прашкович А.Ю.. Нечеткие нейронные сети в решении задачи профилирования пользователя. //Информационные технологии и управление: материалы 52-й научной конференции аспирантов, магистрантов и студентов. Минск, 23 – 25 апреля 2016 г. – Минск, БГУИР, 2016. – стр. 69

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Профилирование или персонализация пользователя – это разумное ограничение предъявляемой пользователю информации с целью выделения более важного содержания для данного индивидуума. Задачей профилирования является правильный отбор пар «пользователь – набор отображаемых данных» путем отсеивания неинтересной пользователю информации. Решение этой задачи позволит потребителям услуг тратить меньше времени на просмотр информации и больше – на ее практическое применение.

Существует масса подходов к персонализации, но можно выделить два основных: персонализация пользователя путем изменения формы отображения данных и персонализация содержания – собственно, предъявляемых данных [1]. К первому подходу можно отнести, например, индивидуальную разметку страницы – сюда входит как расположение элементов друг относительно друга, так и цветовая гамма, темы оформления и так далее. Ко второму подходу относятся все способы выделения конкретных данных – более значимых для пользователя – по сравнению с остальными. В данной работе представлен именно метод персонализации Интернет-пользователей путем изменения набора отображаемых данных.

Один из подходов к профилированию пользователей базируется на применении нейронной сети с обучением на основе обучающих таблиц. Предположим, что все пользователи у нас разделены на группы (они же классы) обозначим из как D. Каждая группа характеризуется набором ключевых фраз (слов) K, каждая из которых имеет свою частоту для каждой из групп и обозначим её как X. Суммируя вышесказанное, получаем таблицу исходных данных в общем виде (таблица 1):

Таблица 1 – Таблица исходных данных нейронной сети в общем виде без выходного сигнала

	k_1	k_2	…	k_n
D_1	X_{11}	X_{21}	…	X_{n1}
D_2	X_{12}	X_{22}	…	X_{n2}
…	…	…	…	…
D_m	X_{1m}	X_{2m}	…	X_{nm}

Выходная функция нейронной сети (распознающая функция) имеет следующий общий вид:

$$Y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \quad (3)$$

где $a =< a_0, a_1, \dots, a_n >$ – вектор весов ключевых фраз (слов);

x_i – частота встречаемости ключевой фразы (ключевого слова) i.

При условии, что Y больше или равно нулю, считаем входной объект относится к распознаваемой области, в противном случае – нет. Например, имеется группа экономических текстов. У экономических текстов свои ключевые слова со своими частотами встречаемости в каждом документе. Распознающая функция (3) вычисляется на входном документе. Если ее значение больше либо равно 0, то документ относится к категории экономических.

Исходными данными для определения весовых коэффициентов $< a_0, a_1, \dots, a_n >$ являются данные таблицы 2

Таблица 2 – Обучающая таблица нейронной сети в общем виде с предопределенным выходным сигналом

	X_1	X_2	…	X_n	Y
D_1	X_{11}	X_{21}	…	X_{n1}	$\geq$ **0** ИЛИ $<$ **0**
D_2	X_{21}	X_{22}	…	X_{n2}	$\geq$ **0** ИЛИ $<$ **0**
…	…	…	…	…	…
D_m	X_{1m}	X_{2m}	…	X_{nm}	$\geq$ **0** ИЛИ $<$ **0**

Для нахождения весов ключевых слов решаем систему линейных алгебраических неравенств (1):

$$\begin{cases} a_0 + a_1x_{11} + a_2x_{21} + \cdots + a_nx_{n1} \geq 0 \ (< 0) \\ a_0 + a_1x_{12} + a_2x_{22} + \cdots + a_nx_{n2} \geq 0 \ (< 0) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_0 + a_1x_{1\mathrm{m}} + a_2x_{2\mathrm{m}} + \cdots + a_nx_{n\mathrm{m}} \geq 0 \ (< 0) \end{cases} \quad (1)$$

Если система имеет решение, то получаем функцию распознавания в форме линейного многочлена, используемую нейронной сетью. В общем случае система может не иметь решения. В таком случае можно поставить задачу построения нечеткой нейронной сети с линейной распознающей

функцией. Математически эта задача записывается в виде системы (5). Введены дополнительные неотрицательные переменные Δ, сумма которых стремится к минимуму. Введение переменных Δобусловлено необходимостью «смягчить» условия выполнения системы неравенств (4). Назовём Δ коэффициентами сдвига.

$$\begin{cases} \Delta_1^2 + \Delta_2^2 + \cdots + \Delta_m^2 \rightarrow min \\ a_0 + a_1 x_{11} + a_2 x_{21} + \cdots + a_n x_{n1} + (-\Delta_1) \geq 0 \text{ или } < 0 \\ a_0 + a_1 x_{12} + a_2 x_{22} + \cdots + a_n x_{n2} + (-\Delta_2) \geq 0 \ (< 0) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 + a_1 x_{1\mathrm{m}} + a_2 x_{2\mathrm{m}} + \cdots + a_n x_{n\mathrm{m}} + (-\Delta_m) \geq 0 \ (< 0) \\ \Delta_1, \Delta_2, \ldots, \Delta_m \geq 0 \end{cases} \quad (2)$$

Полученная система (5) всегда будет иметь решение. Но решение системы (5) не всегда будет удовлетворять обучающей таблице 3.2. Возникает необходимость оценить погрешность решения, связанную с введением коэффициентов сдвига.

Для оценки погрешности, воспользуемся методом Монте-Карло. Суть применения метода заключается в следующем: случайным образом генерируем набор параметров. Для оценки погрешности нам так же необходимо ввести два счетчика: N и $\widetilde{N}$.

Каждый случайно сгенерированный набор параметров подставляем в систему неравенств (5). Если набор является решением системы, то увеличиваем счетчик $\widetilde{N}$ на единицу. Затем проверяем, сколько неравенств из системы (4) выполняется на этом наборе данных; выражаем это количество в процентом отношении и добавляем к счетчику N.

Стоит отметить, что чем больше итераций будет произведено, тем более точные результаты будут получены. Рекомендуемо число итераций не меньше двухсот.

Ответом на поставленную задачу определения того, относится ли входной документ к данному профилю (кластеру), является отношение счетчиков N и $\widetilde{N}$:

$$Q = \frac{N}{\widetilde{N}}, \quad (3)$$

где N – счётчик положительных решений системы неравенств (5), сгенерированных методом Монте-Карло;

$\widetilde{N}-$ счётчик положительных решений уравнений из системы неравенств (4), выраженный в процентном отношении;

$Q-$ отношение величин счетчиков.

Произведя подсчеты и определив значения счетчиков, подставим их в формулу (6), мы получаем отношение величин счетчиков. Если отношение превышает 88%, то условно можем считать, что входное множество сигналов относятся к данному профилю (относится к данному классу документов).

В противном случае для определения профиля пользователя воспользуемся алгоритмом CLOPE.

Рассмотрим алгоритм CLOPE применительно к задаче кластеризации поисковых профилей.

Пусть:

$QP = \cup \{qp_m\}$ – множество поисковых профилей;

$qp_m = \cup \{q_{mf}\}$ – множество f-ых поисковых запросов m-ого поискового профиля;

$C_{qp} = \cup \{cqp_k\}$ – множество кластеров, разбивающее множество поисковых профилей QP так, что $cqp_q \dots cqp_k = \{qp_1 \dots qp_n\}$ и $Cqp_i \neq \emptyset \wedge Cqp_i \cap Cqp_y = \emptyset, i \geq 1, y \leq k.$

Каждый кластер Cqp_i описывается следующими характеристиками:

$D(Cqp_i)$ – множество уникальных поисковых запросов;

$Occ(q, Cqp_i)$ – частота вхождений поискового запроса q в кластер.

Задача кластеризации сводится к нахождению такого разбиения множества поисковых профилей на кластеры, при котором глобальная функция стоимости имеет максимальное значение:

$$profit\ (Cqp_i, r) \rightarrow max, \quad (4)$$

где $profit\ (Cqp_i, r) = \frac{\sum_{i=1}^{k} \frac{S(Cqp_i)}{W(Cqp_i)^r} \times |Cqp_i|}{\sum_{i=1}^{k} |Cqp_i|}$ – глобальная функция стоимости;

$S(Cqp_i) = \sum_{q \in D(Cqp_i)} Occ(q, Cqp_i) = \sum_{qp_m \in Cqp_i} |qp_m|$ – площадь, занимаемая гистограммой кластера;

$W(Cqp_i) = |D(Cqp_i)|$ – ширина гистограммы кластера Cqp_i;

r – коэффициент отталкивания, положительное натуральное число, r=2.

С помощью параметра r, названного авторами CLOPE коэффициентом отталкивания (repulsion), регулируется уровень сходства транзакций внутри

кластера, и, как следствие, финальное количество кластеров. Этот коэффициент подбирается пользователем. Чем больше r, тем ниже уровень сходства и тем больше кластеров будет сгенерировано.

В результате кластеризации поисковый профиль конечного пользователя qp окажется в определенном кластере.

При посещении сайта, который просматривается пользователем в связи с постоянными информационными потребностями, система управления контентом, использующая персонализацию, должна предъявить пользователю список рекомендуемых для просмотра страниц, включающий две группы ссылок:

1) на уже ранее просмотренные страницы для постоянных потребностей;

2) на новые страницы, которые не просмотрены, но могут содержать нужную информацию для постоянных потребностей.

Рассмотрим алгоритм формирования списка.

Формирование первой группы ссылок.

Шаг 1. Для предъявления пользователю не просмотренных страниц, соответствующих постоянной информационной потребности, производится расширение его поискового профиля. Для этого поисковые запросы, входящие в состав кластера Cqp^*, ранжируются по частоте их вхождения в кластер. В расширенный поисковый профиль $QPsim$ выбирается некоторое количество (lim) поисковых запросов с наибольшей частотой вхождения:

$QPsim = \cup \{q_c, fq_c\}$, причем $|QPsim| \leq lim$, (5)

– частотность поискового запроса в кластере.

Шаг 2. Далее формируется множество ссылок L_u на страницы, которые были просмотрены другими пользователями в сессиях с поисковыми запросами из кластера Cqp^*. Ранжирование ссылок осуществляется в соответствии с частотностью поискового запроса fq_c и релевантностью страницы:

$$W_g = fq_c * rel_g,$$

Где W_g – вес ссылки на страницу rv_g;

fq_c – частотность поискового запроса q_c;

rel_g – релевантность g-ой страницы.

Релевантность g-ой страницы рассчитывается как отношение суммы скалярных оценок, полученных в a-ой сессии к количеству сессий, в которых было зафиксировано обращение к g-ой странице:

$$rel_g = \frac{\sum_{a=1}^{u} tsq_{ga}}{u}$$

где u – количество сессий, в которых имел место просмотр g-ой страницы;

tsq_{ga} – оценка релевантности g-ой страницы для a-ой сессии.

Оценка релевантности основана на модели предпочтений, представленной множеством индикаторов вида

$$rel_g = (tt_g, np_g, nvf_g, depth_g),$$

где $tt_g = \sum t_g$ – суммарное время пребывания на странице;

np_g – количество обращений к странице;

nvf_g – количество просмотренных пользователем фрагментов g-ой страницы;

$depth_g$ – глубина просмотра g-ой страницы;

Формирование второй группы ссылок.

Шаг 1. Расчет принадлежности ранее просмотренных страниц постоянным интересам пользователя. Для этого рассчитывается коэффициент линейной корреляции между порядковым номером сеансов, имевших место у пользователя, и количеством сеансов, в которых была просмотрена анализируемая страница. Страница принадлежит интересам пользователя, если $(kor_a > 0{,}7)$.

Шаг 2. Ранжирование ссылок на страницы в соответствии с величиной корреляционной связи.

ЗАКЛЮЧЕНИЕ

Необходимость и актуальность алгоритма отбора релевантных документов, позволяющего производить правильный отбор пар «пользователь – набор отображаемых данных» путем отсеивания неинтересной и ненужной пользователю информации, очевидна.

Решение задачи профилирования пользователя в сети Интернет позволяет ему тратить меньше времени на просмотр информации и больше на практическое её применение.

Наряду с этим, задача профилирования позволяет повысить производительность средств распространения рекламных информационных материалов в Интернет и эффективность рекламного и информационного воздействия на пользователей.

В рамках диссертации рассмотрено два подхода к профилированию интернет пользователей: на основе нечетких нейронных сетей и с применением алгоритма кластеризации CLOPE.

Мной разработан подход, основанный на нечеткой нейронной сети, позволяющий определить профиль пользователя, в случае, кода четкая нейронная сеть не может дать однозначный ответ. Суть метода заключается в нахождении отношения площадей фигур, образуемых системами линейных неравенств четкой и нечеткой нейронных сетей. Если отношений площадей больше 12%, то такой результат мы не можем считать действительным. В этом случае, для определения профиля пользователя, используем алгоритм CLOPE.

В завершение подчеркнем преимущества алгоритма CLOPE:

- Высокие масштабируемость и скорость работы, а так же качество кластеризации, что достигается использованием глобального критерия оптимизации на основе максимизации градиента высоты гистограммы кластера. Он легко рассчитывается и интерпретируется. Во время работы алгоритм хранит в RAM небольшое количество информации по каждому кластеру и требует минимальное число сканирований набора данных. Это позволяет применять его для кластеризации огромных объемов категорийных данных (large categorical data sets);
- CLOPE автоматически подбирает количество кластеров, причем это регулируется одним единственным параметром – коэффициентом отталкивания.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Скуратов А. К ., Ефремов С. В. Персонификация и персонализация как основа современных порталов // Телематика '2003: Тр. X Всероссийской на -учно- методической конф. – 2003. – Т . 1. – С. 183–185

[2] Некрестьянов И . С. Тематико-ориентированные методы информационного поиска Дис.. канд . физ.- мат. наук : 05.13.11. – СПб ., 2000

СПИСОК СОБСТВЕННЫХ ПУБЛИКАЦИЙ

1. Прашкович А.Ю.. Нечеткие нейронные сети в решении задачи профилирования пользователя. //Информационные технологии и управление: материалы 52-й научной конференции аспирантов, магистрантов и студентов. Минск, 23 – 25 апреля 2016 г. – Минск, БГУИР, 2016. – стр. 69