ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

ПРОГРЕССИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

PROGRESSIVE INFORMATION TECHNOLOGIES

УДК 004.55

Аксак Н. Г.

Канд. техн. наук, доцент, профессор кафедры электронных вычислительных машин Харьковского национального университета радиоэлектроники, Харьков, Украина

РАЗРАБОТКА СИСТЕМЫ ПЕРСОНАЛИЗАЦИИ СПЕЦИАЛИЗИРОВАННОГО ВЕБ-ПОРТАЛА

Актуальность. Решена актуальная задача персонификации веб-портала, предоставляющего бизнес сервисы (телемедицина, консультации, удаленный мониторинг, дистанционное образование и т. д.).

Цель работы – разработка системы персонификации веб-портала, предоставляющего специализированные услуги, что позволяет учитывать предпочтения пользователей с целью улучшения качества обслуживания, ускорения информационного поиска, исключения неинтересных страниц, а также удержания клиентов.

Метод. Предложена обобщенная модель процесса персонализации Интернет-сервиса, в которой на основе сочетания агентских и нейросетевых технологий предложен метод адаптации веб-ресурса, автоматически генерирующий контент для определенных категорий Интернет-пользователей. Также предложена объектная модель документов сайта в виде графа для поиска актуальной информации, что позволило осуществить персонализацию. Использование мультиагентной структуры позволило осуществить взаимодействие компонентов разработанной модели. Метод включает совокупность следующий действий: автоматическая выработка гипотез, что дает возможность определить наличие или отсутствие целевых свойств пользователя; анализ поведения пользователя по его серфингу в Интернете, что позволяет выдавать более релевантные результаты; построение информационного портрета для сбора статистически значимой совокупности информационных характеристик с целью планирования дальнейших действий; параллельная кластеризация пользователей с использованием самоорганизующихся карт Кохонена с целью ускорения обработки больших данных. Для ускорения вычислений самоорганизующиеся карты Кохонена адаптированы под симметричные мультипроцессоры системы. Показано, что для уменьшения времени вычислений необходимо выбирать конфигурацию вычислительной системы кратную размерности входных данных.

Результаты. Разработаны программное обеспечение и веб-интерфейс, реализующие предложенные модели и метод, используемые при проведении вычислительных экспериментов по верификации модели, оценки адекватности и исследованию свойств модели и метода.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенных моделей и методов. Применение совокупности методов и средств может быть использовано на практике для продвижения товаров и услуг в сети, для предоставления различных сервисов или отдельных его составных частей, для развития бизнеса.

Ключевые слова: персонализация, ДСМ-метод, нейросетевая кластеризация, мультиагентная система, информационный портрет пользователя.

НОМЕНКЛАТУРА

DBSCAN – Density Based Spatial Clustering of Applications with Noise;

БД – база данных;

БЗ – база знаний;

ДСМ – метод автоматического порождения гипотез Джона Стюарта Милля;

НС – нейронная сеть;

A – множество агентов;

A_{data} – агент обработки информации;

A_{expert} – агент эксперта;

A_{user} – агент пользователя;

$B_{behavior}$ – модель анализа поведения пользователя;

B_H – модель предоставления сервиса;

B_{JSM} – модель порождения гипотез;

B_R – модель процесса персонализации Web-портала;

B_{SOM} – модель категоризации пользователей;

C – множество данных о пользователе;

C^+ – множество целевых положительных свойств пользователя;

C^- – множество целевых отрицательных свойств пользователя;

DOI 10.15588/1607-3274-2018-1-11

$c^k_n()$ – значение k-го входного признака $C^k()$, характеризующее n-ый экземпляр;

$D_{behavior}$ – задача анализа поведения пользователя;

D_H – множество заданий для предоставления сервиса;

D_{JSM} – задача порождения гипотез о наличии или отсутствии определенных свойств пользователя $U(i)$;

D_R – множество заданий для адаптации Web-портала;

D_{SOM} – задача ускоренной кластеризации пользователей с помощью сети Кохонена;

E – агентская середа;

G – граф веб-серфинга пользователя $U()$;

IPU множество информационных портретов пользователей;

K – количество веб-страниц в графе переходов G;

L – количество кластеров веб-пользователей;

M – общее число посетителей Web-портала;

MAS – мультиагентная система;

N – количество целевых свойств Интернет-пользователей;

O – множество целевых свойств пользователей $U()$ Web-портала;

Op_1 – количество операций последовательного алгоритма обучения нейронной сети;

Op_p – количество операций параллельного алгоритма обучения нейронной сети на p вычислителях;

Arc – множество дуг графа *Res*;

p – количество вычислителей;

R – множество категорий пользователей Web-портала;

Res – объектная модель документов Web-портала в виде графа;

RS – возврат в результаты поиска;

$rs_a()$ – показатель возврата в результаты поиска;

S – множество гипотез;

S^- – гипотезы, являющиеся причиной отсутствия целевого свойства;

S^+ – гипотезы, являющиеся причиной наличия целевого свойства;

SOM – self-organizing map;

T – время посещения страниц Web-портала;

$t_a()$ – время посещения a-ой страницы;

TR – глубина просмотра Web-портала;

$tr_a()$ – количество переходов;

U – множество Интернет-пользователей;

V – множество вершин графа G;

τ – время генерации страниц;

$v_a()$ – a-ая вершина графа G;

W – множество весов графа G;

ϖ – степень релевантности отображенной информации;

$w_{ab}()$ – вес дуги, соединяющей a-ую вершину с b-ой;

X – множество вершин графа *Res;*

X^t – страница, которую пользователь посетил во время t;

$\bar{X}_k$ – адаптированная страница;

$\Gamma^{-n}(\bar{X}_k)$ – обратное соответствие, показывающее для каких вершин графа *Res* вершина $\bar{X}_k$ является конечной;

Λ – множество интерфейсов Web-портала;

λ_j – интерфейс j-го кластера;

P – отображение, являющееся решением задачи D_R;

δ – вероятность сброса небезопасных страниц;

π_i – стационарная вероятность распределения процессов;

Υ – отображение, являющееся решением задачи D_{SOM};

Δ – количество вершин графа *Res;*

Φ – отображение, являющееся решением задачи D_{JSM};

Ω – количество образов веб-интерфейсов;

Ψ – отображение, являющееся решением задачи $D_{behavior}$;

$\Re$ – показатель важности страницы;

$\mathbb{X}$ – матричное представление входных данных;

$\mathbb{Y}$ – выходное множество;

$\mathbb{Z}$ – полное множество универсум;

ВВЕДЕНИЕ

С ростом Интернета, использованием социальных сетей, мобильных устройств, подключенных и сообщающихся объектов, информация увеличивается экспоненциально. Обрабатывать такую лавину информации, а также осуществлять поиск с каждым днем становится все сложнее. С одной стороны, в этом гигантском хранилище информации для нахождения необходимого ресурса обычно приходится осуществлять длительный серфинг по Интернету, с другой – очень много уделять внимания для удержания посетителей на сайте.

При разработке Web-ресурса одной из самых важных задач является сделать его максимально привлекательным для потенциальных пользователей и придать ему индивидуальность. В современном мире в условиях выросшей конкуренции веб-сайт, как представительство фирмы, является целевой рекламой, обеспечивает информационную поддержку клиентов и сотрудников фирмы. Независимо от того создается сайт для представления какой-либо фирмы или для заработка на рекламе, основной его функцией является привлечение как можно большей аудитории.

Целью данной работы являлась разработка системы персонализации сервис-ориентированного веб-портала, позволяющей учитывать предпочтения пользователей для улучшения качества обслуживания, ускорения информационного поиска, исключения неинтересных страниц, а также удержания клиентов.

1 ПОСТАНОВКА ЗАДАЧИ

Пусть имеется информация о пользователе $U(i)$ ($i = \overline{1, M}$) (имя браузера, номер версии, язык, платформа, встроенные расширения, адрес предыдущей страницы, часовой пояс, время посещения страницы, информация о мониторе и т.п.) $C(i) = [C^1(i), C^2(i), \ldots, C^k(i)]$.

Тогда задача персонализации веб-ресурса *Res* при ограниченном количестве образов веб-интерфейсов Ω будет заключаться в разработке:

– модели персонализации B_R, ее составных компонентов и их взаимодействия;

– объектной модели документов Интернет-ресурса в виде графа *Res;*

– метода персонализации веб-портала, позволяющего для каждой $R(n)$-ой категории пользователей ($n = \overline{1, L}$) адаптировать информационное наполнение страниц $\overline{X}_k \in \overline{Res}$ и $\overline{Res} \subset Res$ ($k < \Delta$).

Критерием эффективности разработки является удовлетворение требованиям $k : \forall(D_i \in D_R)[(\tau < \tau^{\max}) \& (\varpi > \varpi^{\min})] \Rightarrow B_R$, где D_i – подзадача общей задачи D_R.

2 ОБЗОР ЛИТЕРАТУРЫ

Одной из наиболее актуальных проблем обработки больших данных является кластеризация веб-пользователей на основе их общих свойств. В статье [1] представлен способ определения сходства интересов Интернет-пользователей. Веб-журналы доступа пользователей обеспечивают точную и объективную информацию о посетителях. Записи журнала содержат IP-адрес веб-пользователя, дату и время запроса, URL-адрес запрашиваемой страницы, протокол запроса, код возврата сервера с указанием статуса обработки запроса и при успешном запросе размер страницы. Из журнала веб-сервера извлекается пользовательский шаблон, состоящий из страниц, которые пользователь посетил и потраченного на это времени. Проведенные эксперименты показали, что предложенный метод кластеризации группирует веб-пользователей со схожими интересами.

В работе [2] предложено объединение веб-пользователей на основе эволюции посещения веб-страниц. Обнаруженные закономерности изменений информационных потребностей веб-пользователей используются для их группировки. Сгенерированные на основе исторических веб-сессий кластеры Web-пользователей, могут быть использованы для персонализированных веб-приложений: веб-рекламы и веб-кэширования.

Чтобы получить информацию об интересах пользователей на веб-страницах в работе [3] исследуется поведение клиента посредством изучения записей веб-журнала. Время, проведенное на веб-странице, и типы совершенных операций показывают степень заинтересованности веб-пользователя. Исследуемые данные представляют собой журналы пользователей, собранные за шесть месяцев. В работе предложена модификация алгоритма кластеризации *K*-means для группировки пользователей путем вычисления начальных центроидов на основе выбранного веб-контента.

Любой Интернет-портал может постоянно совершенствоваться, опираясь на информационную потребность пользователя. Для сбора и анализа данных о пользователях в работе [4] используется метод роевого интеллекта, благодаря которому выявляются «путешествия» веб-пользователей с одинаковыми интересами. Результаты кластеризации сравниваются с методами DBSCAN и K-means.

Благодаря самоорганизации, простоте и быстродействию упрощенная модель нейронной сети Кохонена предлагается использоваться в информационно-поисковой системе [5]. Для ее успешного применения необходимо решить задачи формирования содержательного образа документа и идентификацию кластера.

Проблемы обработки больших данных, связанные с их многообразием, со сложностями сбора, хранения, управления и анализа, объемом памяти и скоростью вычислений рассмотрены в [6, 7]. Описаны методики и алгоритмы, используемые для управления большими наборами данных. Показана целесообразность применения самоорганизующихся карт для анализа данных большой размерности.

3 МАТЕРИАЛЫ И МЕТОДЫ

В общем виде процесс персонализации Интернет-ресурса, предоставляющего сервисные услуги, описывается моделью (1), которая включает следующие компоненты [8]: Web–интерфейс, блок интеллектуальных методов, агентский блок, а также блок накопления и анализа опыта (рис. 1).

Агент пользователя» A_{user} собирает информацию о пользователе, на основе которой решается задача порождения гипотез D_{JSM} и анализируется поведение пользователя $D_{behavior.}$ По выходным данным решенных задач «Агент обработки информации» A_{data}, формирует информационный портрет пользователя *IPU*(*i*), и передает/принимает информацию блоку накопления и анализа информации. «Агент эксперта» A_{expert} на основе полученной информации от «Агента пользователя» A_{user} и «Агента

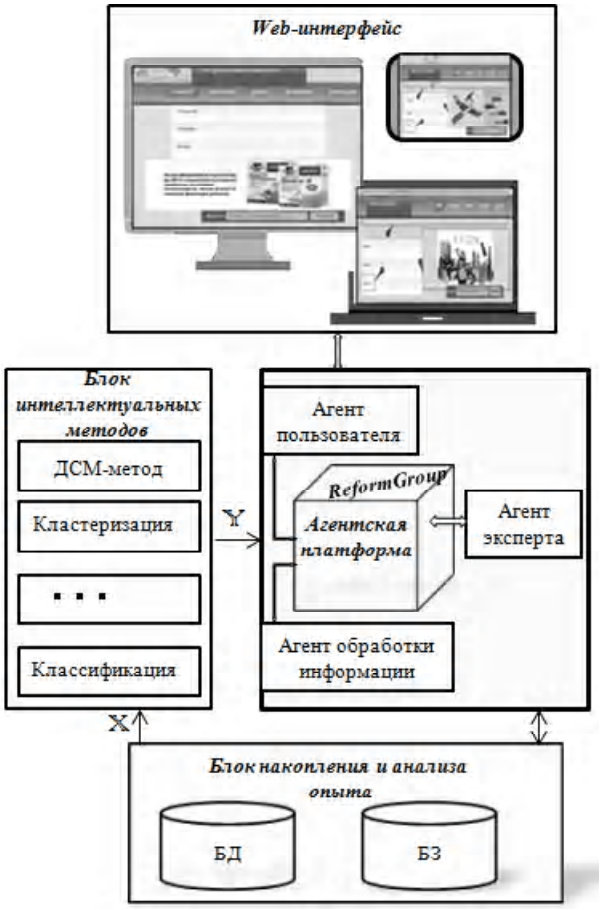


Рисунок 1 – Обобщенная модель персонализации специализированного Веб-портала: $\mathbb{X}$ – матричное представление входных данных; $\mathbb{Y}$ – выходное множество; БД – база данных, БЗ – база знаний

обработки данных» A_{data} обращается за предоставлением услуг D_H. Под услугами будем понимать, например, анализ медицинских изображений (меланомы, опухолей, маммографии, ишемической болезни сердца и т. п.), удаленный мониторинг состояния здоровья человека и т.д. Для решения задачи категоризации пользователей D_{SOM} используется множество информационных портретов пользователя *IPU*. Для каждой категории веб-пользователей «Агент пользователя» A_{user} вырабатывает оптимальную стратегию, позволяющую учитывать личные предпочтения, и настраивает соответствующий контент.

Модель процесса персонализации Интернет-ресурса *Res* [9] выражается как преобразование входных значений R в выходные величины Λ:

$$B_R \subset R \times \Lambda. \quad (1)$$

Полное множество (универсум) $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$ включает в себя $B_R \subset (R \times \Lambda)$, это означает, что существует такое подмножество $R = \{R(1), R(2), ..., R(L)\}$, $P \subset \mathbb{X}$ и отношений между ними Λ, на которых строится модель B_R.

Таким образом, интерфейс Интернет-ресурса, предоставляющего сервисы для пользователей, адаптируется на основе кластеров $R = \{R(1), R(2), ..., R(L)\}$ благодаря модели B_R при конечном числе образов веб Ω. Для выходных величин Λ построено множество заданий, решение которых принадлежит множеству $D_R = D_{JSM} \cup D_{behavior} \cup D_{SOM}$.

Отображение $\Pi : R \to \Lambda$ позволяет для каждой категории $R(n)$ ($n = \overline{1, L}$) получить такое $\lambda_j \in \Lambda$ ($j = \overline{1, \Omega}$), которое является решением задачи D_R, полученное в виде графовой модели (2) иерархического представления адаптированного Интернет-ресурса

$$\overline{Res} = < \bar{X}, \Gamma^{-n}, \lambda_j >, \quad (2)$$

где $\bar{X} = \{\bar{X}_k\}$ – множество адаптированных страниц ($k < \Delta$), сформированных в результате отображения $\bar{X}_k = \Gamma^{-k}(x_m)$ множества интересного для пользователя информационного наполнения $x_m \subset x$.

С помощью метода автоматического порождения гипотез Джона Стюарта Милля (ДСМ-метод) делаются предположения о причинах наличия или отсутствия определенных свойств Интернет-пользователя $U(i)$, ($i = 1 \div M$) во время его посещения Интернет-ресурса по собранным данным $C(i) = [C^1(i), C^2(i), ..., C^k(i)]$.

Элементы множеств $C = [C^1, C^2, ..., C^k]$ будем называть примитивными элементами (атомами) [10]. Поведение пользователя $U(i)$ будем называть (целевыми) свойствами $O(i) \in O$. Исследуемый объект (Интернет-пользователь) представляется в виде конечного множества примитивных элементов $C(i) = [C^1(i), C^2(i), ..., C^k(i)] \in C$. Такое подмножество представляется фрагментом. Пользователь может обладать (или не обладать) некоторым множеством целевых свойств. Предполагается, что как у наличия, так и у отсутствия набора целевых свойств может быть причина (не обязательно единственная), эта причина является фрагментом структуры объекта. Множество C включает объекты с известными целевыми свойствами, как положительными C^+, так и отрицательными примерами C^-, а также объекты с неопределенными примерами. Объекты с известными целевыми свойствами образуют обучающую выборку, объекты с неопределенными примерами – тестовую. Множество O содержит все интересующие целевые свойства пользователей Интернет-ресурса, предоставляющего сервисные услуги.

Реализация ДСМ-метода представляет собой итеративную процедуру и состоит из трех основных этапов: индукции, аналогии и абдукции. Модель причинно-следственных связей, используемая ДСМ-методом для этапов индукции и аналогии, может быть представлена графом, изображенном на рис. 2.

Автоматическое порождение гипотез о возможных причинах наличия или отсутствия целевых свойств у веб-пользователей происходит на этапе индукции. Гипотезы формируются отдельно для каждого целевого свойства на основе поиска общих фрагментов у пользователей, обладающих (или не обладающих) данным свойством, и представляют собой пару – фрагмент и свойство. Для каждого целевого свойства генерируются два множества гипотез: S^+ – гипотезы, являющиеся причиной наличия целевого свойства, и S^- – гипотезы, являющиеся причиной его отсутствия.

Результатом этапа индукции являются $2N$ множеств гипотез: $\Phi : C \to S$. Отображающая функция Φ для конечного множества $C(i)$ формирует $(S_n^+ \cup S_n^-) \in S$ ($n = \overline{1, N}$), которое является решением задачи D_{JSM}, полученное в виде модели порождения гипотез ДСМ-методом (3)

$$B_{JSM} \subset C \times S. \quad (3)$$

На этапе аналогии порожденные гипотезы применяются с целью определения наличия или отсутствия целевых свойств у неопределенных примеров. Классифицированные неопределенные примеры пополняют множества положительных и отрицательных объектов. Этапы индукции и аналогии повторяются до тех пор, пока множества порождаемых гипотез не перестанут изменяться. После этого выполняется этап абдукции, на котором проверяется условие каузальной полноты – объясняют ли сформированные гипотезы исходные обучающие данные [11].

В результате определяются значения параметров, описывающих пользователя, по которым можно сделать предположение, например, о социальном статусе («низкий», «средний», «высокий»). Располагая данными о часовом поясе и IP-адресе, а также сеткой распределения IP-адресов между Интернет-провайдерами, можно выдвинуть гипотезу о географическом расположении

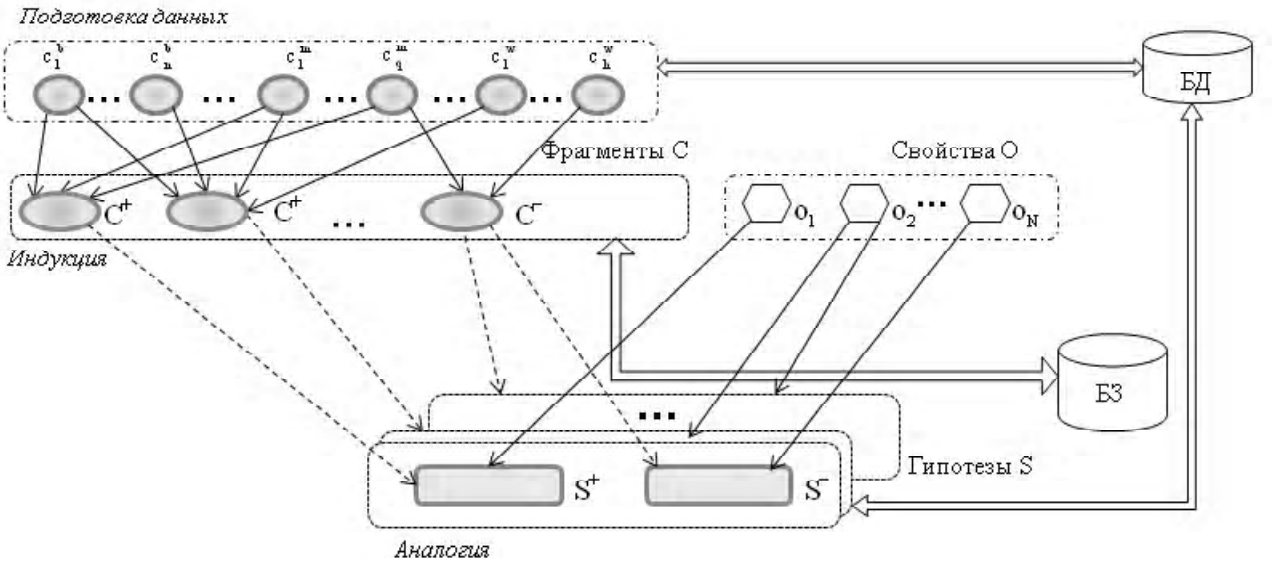


Рисунок 2 – Модель порождения гипотез ДСМ-методом

пользователя (город, область, страна). По времени соединения можно вычислить место соединения (домашнее или служебное помещение, Интернет-кафе и пр.). Анализируя статистику за продолжительный период времени, можно отделить пользователей стационарных компьютеров от мобильных, которые подключаются через разных провайдеров и из разных типов подсетей.

Также строится поведенческий граф с целью вычисления важности посещенных страниц. Каждое направленное ребро представляет переход между двумя вершинами, отражающими количество переходов в качестве веса страницы. Другими словами, граф переходов пользователей (4) является взвешенным графом с ребрами, содержащими вес веб-страницы и вершинами, в виде URL адреса страниц

$$G(i) =< V(i), W(i), T(i), TR(i), RS(i), \delta(i) >, \qquad (4)$$

где $V(i) = \{v_a(i)\}$ – множество вершин; $W(i) = \{w_{ab}(i)\}$ – множество весов; $T(i) = \{t_a(i)\}$ – время посещения; $TR(i) = \{tr_a(i)\}$ – глубина просмотра; $RS(i) = \{rs_a(i)\}$ – возврат в результаты поиска, принимает значение 0, если возврат не произошел, 1 – если возврат имел место быть; $\delta(i)$ – вероятность сброса небезопасных страниц; $(a, b = 1,..., K)$ – количество вершин в графе переходов.

Отображение $\Psi : G \to \Re$ позволяет для $\forall$ U(i) получить такое $\pi_i \in \Re$ (π_i – стационарная вероятность распределения процессов $X = \{X^t,\ t \ge 0\}$, X^t – страница, которую пользователь посетил во время t, $(t > 0)$), которое является решением задачи $D_{behavior}$, полученное в виде модели определения важности страниц: $B_{behavior} \subset G \times \Re$.

Иными словами, по времени посещения страницы, типу переходов и информации о поведении пользователя на выходе будет получено значение показателя важности страницы.

Таким образом, получаем информационный портрет пользователя $IPU(i) = [S(i), \Re(i)]$.

Отображение $\Upsilon : IPU \to R$ позволяет для множества IPU получить такое $r_i \in R$, которое является решением задачи D_{SOM}, полученное в виде модели категоризации пользователей (5):

$$B_{SOM} \subset IPU \times R. \qquad (5)$$

Агентский блок представлен следующими множествами *MAS={A,E,Res}*. Множество адаптивных агентов *A* может быть представлено в виде $A=\{A_{user}, A_{expert} A_{data}\}$, среда *E* представляет собой программную платформу для выполнения агентов и предоставляет функциональность для создания/уничтожения агентов, для приема/передачи сообщений; *Res* – объектная модель документов в виде графа (6)

$$Res =< X, Arc, \Lambda >, \qquad (6)$$

где $X = \{X_a\}$ – множество вершин графа, представляющее страницы сайта, каждая страница $X_a = \{x\}$ представлена множеством информационных элементов: текстовые блоки, меню, ссылки, графические элементы и т. д. ($x = \{1,...,|x|\}$ –.), Arc – множество дуг графа, при этом дуга $arc = (a, b)$ принадлежит графу только, если направление предполагается заданным от вершины *a* к вершине *b* ($a, b = 1,..., \Delta$ – количество вершин графа *Res*), $\Lambda = \{\lambda_c\}$ – множество интерфейсов веб-ресурсов *Res*, $(c = 1,..., \Omega)$ – количество образов веб-интерфейсов.

Блок накопления и анализа информации включает базу данных (БД), хранящую множества $C = [C^1, C^2, ..., C^k]$, множество Интернет-пользователей *U*, целевые свойства *O*, множество информационных портретов пользователей *IPU*, а также базу знаний (БЗ), содержащую общие закономерности о возможных при-

чинах наличия (отсутствия) целевых свойств и информацию, являющуюся результатом накопленного опыта.

4 ЭКСПЕРИМЕНТЫ

Поведенные эксперименты хорошо согласуются с предлагаемой моделью процесса персонализации. Распараллеливание осуществлено с помощью технологий OpenMP и MS MPI на языке программирования C++ в операционной системе Microsoft Windows Compute Cluster Server 2003 на четырехядерных вычислителях Intel Core 2 Quad CPU Q8200 @2.33GHz. Веб-интерфейс реализован с помощью языка гипертекстовой разметки HTML, каскадных таблиц стилей CSS и скриптового языка JavaScript. Программное обеспечение для мобильных агентов реализовано на платформе .NET Framework с использованием языка C# для операционной системы Windows и Windows Mobile. Язык JavaScript позволил получить максимально полную информацию о стране пользователя, которая применялась для автоматической локализации страницы, а также размеры дисплея, что позволило автоматически перестраивать ресурс в зависимости от них.

Кластеризация пользователей является одной из самых трудоемких подзадач при веб-персонализации, поскольку требуется обработка большого объема данных. Поэтому для ускорения вычислений самоорганизующиеся карты Кохонена были адаптированы под SMP системы. Для проведения вычислительных экспериментов использована база данных компании Маркет Репорт [12]. Было подано множество образцов $IPU = \{IPU^1,...,IPU^M\}$, где $IPU^j = \left(ipu_1^j,...,ipu_m^j\right)^T$, $j = \overline{1,M}$. Сеть состоит из одного слоя, имеет m входных нейронов, соответствующих координатам рассматриваемых образцов, и k^2 выходных нейронов, представляющих собой квадратную решетку размером $k \times k$.

Параллельная реализация на системах с общей памятью, включающих p вычислителей, основана на одновременной работе максимально возможного количества нейронов в одной группе $Group_{\max}(p,k)$. Количество операций последовательного алгоритма обучения нейронной сети выражается следующим соотношением

$$Op_1 = 6 + T\left(8 + M\left(2 + 21k^2 m + 3\sum_{i=2}^{k^2}\frac{1}{i}\right)\right),$$

параллельного алгоритма соответственно выражением

$$Op_p = 6 + T\left(8 + M\left(2 + 21Group_{\max}(p,k)m + 3\sum_{i=2}^{Group_{\max}(p,k)}\frac{1}{i} + 3\sum_{i=2}^{p}\frac{1}{i} + (3m+2)p\right)\right).$$

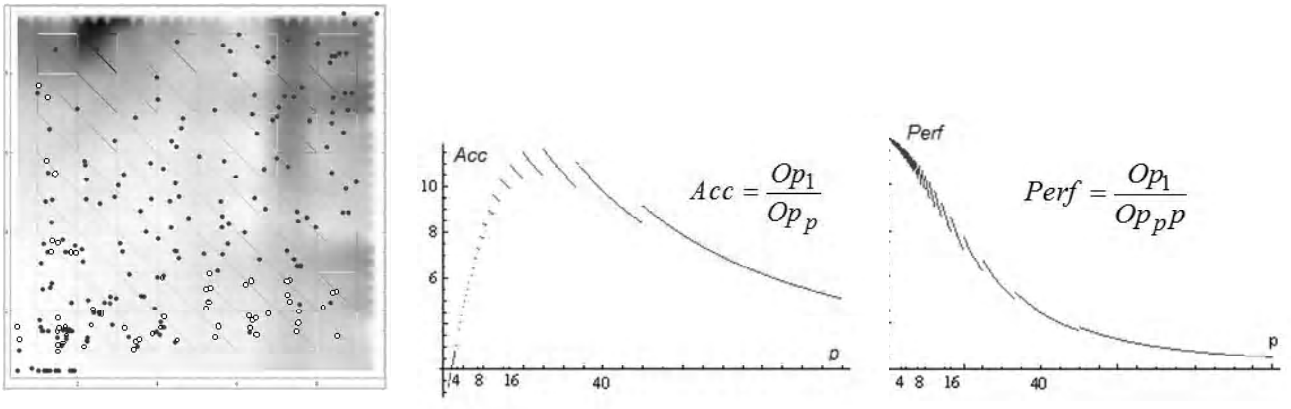
5 РЕЗУЛЬТАТЫ

Результаты моделирования параллельной реализации сети Кохонена для 250 посетителей веб-ресурса приведены на рис. 3.

На рис. 3а темным цветом отмечены украинские посетители, светлым цветом – зарубежные пользователи. Из рисунка видно, что все пользователи разбиты на 4 кластера, причем большинство посетителей сосредоточено в одном кластере. Экспериментальный анализ показал, что для параллельной реализации кластеризации пользователей наиболее целесообразно выбирать количество ускорителей p кратное величине k^2. Таким образом, наибольший выигрыш по времени получается в том случае, когда в зависимости от размерности входных данных выбрана соответствующая конфигурация вычислительной системы (рис. 3б, 3в).

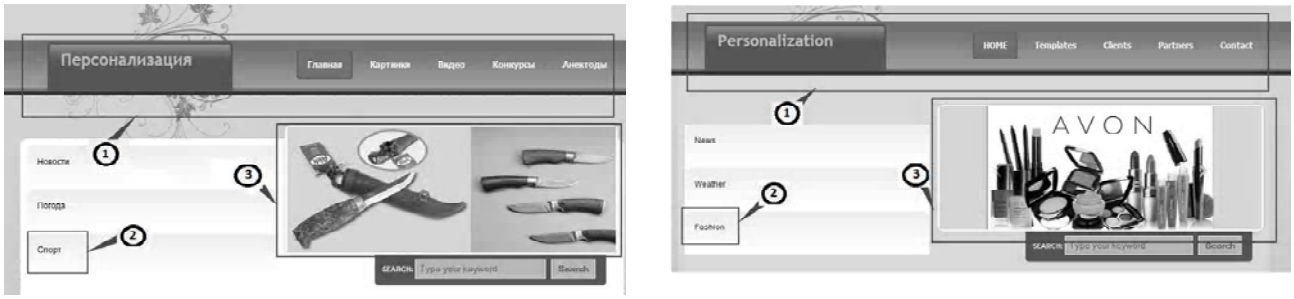
Для разработанного веб-ресурса на основе выявленной категории пользователей с помощью самоорганизующейся сети Кохонена осуществлена адаптация информационного наполнения сайта (рис. 4).

Для пользователей, первый раз посетивших веб-ресурс, отображается стартовая страница. В зависимости от полученных данных о пользователе (страна, размер дисплея, пол, возраст и т. д.) на главную страницу выносятся различные разделы. Так, на рис. 4 а для посетителя мужского пола можно заметить, что в области, обозначенной схематически цифрой ②, есть раздел «Спорт».



а б в

Рисунок 3 – Результаты кластеризации: а – визуализация данных на основе прямоугольных кусочно-гладких карт Кохонена; б – график ускорения; в – график эффективности; для ускорение используется величина *Acc*; для эффективности *Perf*



а б

Рисунок 4 – Персонализация: а – для молодых мужчин, местожительство которых Украина, предлагается реклама ножей; б – для девушек, проживающих за рубежом, предлагается реклама косметики

В то же время, как показано на рис. 4б, для женской аудитории скрыт этот раздел и добавлен раздел «Мода», от выбранного пользователем цвета, подстроен фон сайта.

6 ОБСУЖДЕНИЕ

Предложенный на основе обобщенной модели метод позволяет решать комплексную задачу адаптации веб-портала совместно с предоставлением специализированных услуг в отличие от подходов, в которых Интернет-портал постоянно совершенствуется, опираясь на информационную потребность пользователя. Предложенный метод отличается тем, что кластеризация пользователей выполнена благодаря построенным информационным портретам на основе выдвинутых гипотез о наличии или отсутствии определенных свойств пользователя и с учетом посещения им интересных страниц, в отличие от подходов, в которых кластеризация пользователей осуществляется на основании сходства сеансов по времени посещения веб-страниц, обнаружения шаблонов поведения по историческим веб-сессиям и с использованием метода роевого интеллекта [1–5].

Эффективность использования предложенного метода при увеличении количества посетителей Интернет-портала повысится, если в конфигурации вычислительной системы увеличить количество вычислителей до величины, кратной размерности карты Кохонена. При этом целесообразно увеличить количество персональных агентов, которые смогут адаптироваться под инцидент персонализированного общения с клиентом для более длительного удержания его на сайте.

ВЫВОДЫ

В работе решена актуальная задача персонификации веб-портала, предоставляющего бизнес услуги.

Научная новизна полученных результатов состоит в том, что получила дальнейшее развитие обобщенная модель адаптации специализированного веб-портала, которая в отличие от существующих, описывает взаимодействие веб-интерфейса, блока интеллектуальных методов, агентского блока, блока накопления и анализа опыта, благодаря сочетанию интеллектуального анализа данных для решения задач предоставления сервиса, адаптации самоорганизующихся карт Кохонена под симметричные мультипроцессоры для параллельной категоризации большого количества Интернет-пользователей, динамического распределения ролей между агентами для управления различными инцидентами в зависимости от предоставляемого сервиса, что позволяет улучшить качество обслуживания, ускорить информационный поиск, исключить неинтересные страницы, а также дольше удерживать клиентов за счет учета предпочтений пользователей. Впервые предложена объектная модель документов Веб-портала в виде графа, в которой в качестве узлов выступают страницы и их составные элементы, каждый из которых является потомком и/или предком другого элемента, а дуги – переходы между ними, благодаря чему можно определить путь до элементов, содержащих интересующую пользователя информацию, что позволило персонализировать контент сайта. Также в работе впервые предложен метод персонализации Интернет-сервиса, в котором выявление причинно-следственных закономерностей обусловило возможность выработки гипотез о наличии или отсутствии определенных свойств пользователя; на основе собранной о нем информации, по его серфингу в Интернете построен информационный портрет пользователя, благодаря которому осуществлена ускоренная кластеризация пользователей, что позволило настраивать контент, осуществлять консультации в режиме реального времени, предоставлять актуальную информацию по различным аспектам предлагаемого сервиса.

Практическая значимость полученных результатов заключается в том, что разработано программное обеспечение и веб-интерфейс, реализующие предложенные модели и метод, используемые при проведении вычислительных экспериментов по верификации модели, оценки адекватности и исследованию свойств моделей и метода. Разработанный веб-ресурс позволяет настраивать информационное наполнение (навигацию, поисковую поддержку, персонализацию веб-приложений, веб-рекламу и т.д.) в зависимости от состояния внешней среды и от индивидуальных предпочтений пользователя, а также позволяет рекомендовать предложенные модели и метод для построения систем, направленных на продвижение товаров и услуг в сети, предоставление различных сервисов или отдельных его составных частей, для развития бизнеса. В результате собранной и проанализированной информации о пользовательских предпочтениях сформированы кластеры пользователей со схожими свойствами, благодаря чему осуществлена персонализация Web-пространства.

Перспективы дальнейших исследований заключаются в том, чтобы ускорить обработку персональных данных пользователей большого объема с использованием распределённой файловой системы Hadoop, а также осуществить персонализацию веб-ресурса для каждого пользователя с учетом его предпочтений.

БЛАГОДАРНОСТИ

Полученные результаты соответствуют проблематике госбюджетной темы «Эволюционные гибридные системы вычислительного интеллекта с изменяемой структурой для интеллектуального анализа данных», раздел «Эволюционные гибридные методы и модели интеллектуальной обработки информации с изменяемой структурой в условиях неопределенности» (№ ДР 011U000458).

СПИСОК ЛИТЕРАТУРЫ

1. Xiao J. Clustering of web users using session-based similarity measures/ J. Xiao, Y. Zhang // Computer Networks and Mobile Computing, 2001. Proceedings. 2001 International Conference on. – IEEE, 2001. – P. 223–228. DOI: 10.1109/ICCNMC.2001.962600
2. Chen L. COWES: Web user clustering based on evolutionary web sessions / L. Chen, S. S. Bhowmick, W. Nejdl // Data & Knowledge Engineering. – 2009. – Vol. 68, No. 10. – P. 867–885. DOI: 10.1016/j.datak.2009.05.002 ·
3. Selvakumar K. Enhanced K-Means Clustering Algorithm for Evolving User Groups / K. Selvakumar, L. S. Ramesh, A. Kannan //Indian Journal of Science and Technology. – 2015. – Vol. 8, No. 24. – P. 1. DOI: 10.17485/ijst/2015/v8i24/80192
4. Ganesan S. Evolving interest based user groups using PSO algorithm / S. Ganesan, A. I. U. Sivaneri, S. K. Selvaraju // Recent Trends in Information Technology (ICRTIT), 2014 International Conference on. – IEEE, 2014. – P. 1–6. DOI: 10.1109/ICRTIT.2014.6996196
5. Андреева К. А. Применение нейронной сети Кохонена для классификации web-страниц информационно-поисковой системой сайтов / К. А. Андреева, Р. С. Шайдуров, Е. П. Моргунов // Актуальные проблемы авиации и космонавтики. – 2015. – Т. 1, № 11– С. 380–381.
6. Zerhari B. 'Big data clustering: Algorithms and challenge' / B. Zerhari, A. A. Lahcen, S. Mouline // Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15). – 2015.
7. Kurasova O. Strategies for big data clustering / O. Kurasova et al. // 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. – IEEE, 2014. – C. 740–747. DOI: 10.1109/ICTAI.2014.115
8. Axak N. Development of multi-agent system of neural network diagnostics and remote monitoring of patient / N. Axak // Eastern-European Journal of Enterprise Technologies. – 2016. – 4/9 (82) – P. 4–11. DOI: http://dx.doi.org/10.15587/1729-4061.2016.75690
9. Axak N. Decision support system for intelligent site / N. Axak, S. Korgut, P. Komoda //Elektronika (LIV). – 2013. – No. 8. – P. 52–59.
10. Аншаков О. М. ДСМ-метод: теоретико-множественное объяснение / О. М. Аншаков // НТИ. Сер. 2. – 2012. – № 9.
11. Финн В. К. Индуктивные методы Д. С. Милля в системах искусственного интеллекта. Часть I / В. К. Финн // Искусственный интеллект и принятие решений. – 2010. – № 3. – С. 3–21.
12. Shklovets A. V. Visualization of high-dimensional data using two-dimensional self-organizing piecewise-smooth Kohonen maps / A. V. Shklovets , N. G. Axak // Optical Memory and Neural Networks. – 2012. – Vol. 21, No. 4. – P. 227–232. DOI: 10.3103/S1060992X12040066.

Статья поступила в редакцию 25.04.2017.
После доработки 20.06.2017.

Аксак Н. Г.
Канд. техн. наук, доцент, професор кафедри електронних обчислювальних машин Харківського національного університету радіоелектроніки, Харків, Україна

РОЗРОБКА СИСТЕМИ ПЕРСОНАЛІЗАЦІЇ СПЕЦІАЛІЗОВАНОГО ВЕБ-ПОРТАЛУ

Актуальність. Вирішено актуальне завдання персоніфікації веб-порталу, який надає бізнес сервіси (телемедицина, консультації, віддалений моніторинг, дистанційна освіта і т.д.).

Мета роботи – розробка системи персоніфікації веб-порталу, який надає спеціалізовані послуги, що дозволяє враховувати переваги користувачів з метою поліпшення якості обслуговування, прискорення інформаційного пошуку, виключення нецікавих сторінок, а також утримання клієнтів.

Метод. Запропоновано узагальнену модель процесу персоналізації Інтернет-сервісу, в якій на основі поєднання агентських і нейромережевих технологій запропонований метод адаптації веб-ресурсу, що автоматично генерує контент для певних категорій Інтернет-користувачів. Також запропонована об'єктна модель документів сайту у вигляді графа для пошуку актуальної інформації, що дозволило здійснити персоналізацію. Використання мультиагентної структури дозволило здійснити взаємодію компонентів розробленої моделі. Метод включає сукупність наступний дій: автоматичне вироблення гіпотез, що дає можливість визначити наявність або відсутність цільових властивостей користувача; аналіз поведінки користувача на його серфінгу в Інтернеті, що дозволяє видавати більш релевантні результати; побудова інформаційного портрета для збору статистично значущою сукупності інформаційних характеристик з метою планування подальших дій; паралельна кластеризація користувачів з використанням самоорганізуючих карт Кохонена з метою прискорення обробки великих даних. Для прискорення обчислень самоорганізуючі карти Кохонена адаптовані під симетричні мультипроцесори системи. Показано, що для зменшення часу обчислень необхідно вибирати конфігурацію обчислювальної системи кратну розмірності вхідних даних.

Результати. Розроблено програмне забезпечення та веб-інтерфейс, що реалізують запропоновані моделі і метод, що використовуються при проведенні обчислювальних експериментів по верифікації моделі, оцінки адекватності, дослідженню властивостей моделі та методу.

Висновки. Проведені експерименти підтвердили працездатність запропонованих моделей та методів. Застосування сукупності методів і засобів може бути використано на практиці для просування товарів і послуг в мережі, для надання різних сервісів або окремих його складових частин, для розвитку бізнесу.

Ключові слова: персоналізація, ДСМ-метод, нейромережева кластеризація, мультиагентна система, інформаційний портрет користувача.

Axak N. G.

PhD, Associate professor, Professor of Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

DEVELOPMENT OF PERSONALIZATION SYSTEM OF SPECIALIZED WEB PORTAL

Context. The actual task of personifying a Web portal providing business services (telemedicine, consultations, remote monitoring, distance education, etc.) has been solved.

Objective - development of a personalization system for a web portal that provides specialized services, which allows to take into account preferences of users for the improvement of quality of service, an acceleration of information search, an exception of uninteresting pages, and a customer retention.

Method. The generalized process personalization model of Internet service is offered. The method of adaptation of the Web-resource based on the combination of agent and neural network technologies is proposed in a model which automatically generates content for certain categories of Internet users. The document object model of site in a graph form to search of relevant information was proposed that allows the site personalization. The use of multi-agent structure allowed to realize interaction of the components of the developed model. The method includes the following actions: automatic generation of hypotheses, which determines the presence or absence of target properties of the user; analysis of the user's behavior on his surfing the Internet that allows to give more relevant results; construction of information portrait for collection statistically significant set of information characteristics for the purpose of planning of further actions; parallel clustering of users with use of the self-organizing Kohonen maps for the purpose of an acceleration of processing big data. The self-organizing Kohonen maps are adapted to symmetric multiprocessing system for accelerating computations. Thus, the configuration of the computing system shall be a multiple of the dimension of the input data for reduction of computation time.

Results. For the proposed models and method, software and a web interface are developed. They are used to realization computing experiments to verification of the models, valuation of the adequacy and study the properties of the model and method.

Conclusions. The conducted experiments have confirmed the proposed models and methods. The use of a set of methods and tools can be used in practice to promote goods and services in the network, to provide various services or individual parts of it, for business development.

Keywords: personalization, JSM-method, neural network clustering, multi-agent system, informative portrait of user.

REFERENCES

1. Xiao J., Zhang Y. Clustering of web users using session-based similarity measures, *Computer Networks and Mobile Computing, 2001. Proceedings. 2001 International Conference on. IEEE,* 2001, pp. 223-228. DOI: 10.1109/ICCNMC.2001.962600.
2. Chen L., Bhowmick S. S., Nejdl W. COWES: Web user clustering based on evolutionary web sessions, *Data & Knowledge Engineering*, 2009, Vol. 68, No. 10, pp. 867–885. DOI: 10.1016/j.datak.2009.05.002.
3. Selvakumar K., Ramesh L. S., Kannan A. Enhanced K-Means Clustering Algorithm for Evolving User Groups, *Indian Journal of Science and Technology*, 2015, Vol. 8, No. 24, P. 1. DOI: 10.17485/ijst/2015/v8i24/80192.
4. Ganesan S., Sivaneri A. I. U., Selvaraju S. K. Evolving interest based user groups using PSO algorithm, Recent *Trends in Information Technology (ICRTIT), 2014 International Conference on, IEEE,* 2014, pp. 1–6. DOI: 10.1109/ICRTIT.2014.6996196.
5. Andreeva K. A., Shajdurov R. S., Morgunov E. P. Primenenie nejronnoj seti Kohonena dlja klassifikacii web-stranic informacionno-poiskovoj sistemoj sajtov, *Aktual'nye problemy aviacii i kosmonavtiki,* 2015, Vol. 1, No. 11, pp. 380–381.
6. Zerhari B., Lahcen A. A., Mouline S. Big data clustering: Algorithms and challenge, *Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15),* 2015.
7. Kurasova O. et al. Strategies for big data clustering, *2014 IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE*, 2014, pp. 740–747.
8. Axak N. Development of multi-agent system of neural network diagnostics and remote monitoring of patient, *Eastern-European Journal of Enterprise Technologies*, 2016, 4/9 (82), pp. 4–11.
9. Axak N., Korgut S., Komoda P. Decision support system for intelligent site, *Elektronika (LIV)*, No. 8/2013, pp. 52–59.
10. Anshakov O. M. DSM-metod: teoretiko-mnozhestvennoe ob#jasnenie, *NTI. Ser. 2.* 2012, № 9.
11. Finn V. K. Induktivnye metody D.S. Millja v sistemah iskusstvennogo intellekta. Chast' I, *Iskusstvennyj intellekt i prinjatie reshenij*, 2010, No. 3, pp. 3–21.
12. Shklovets A. V., Axak N. G. Visualization of high-dimensional data using two-dimensional self-organizing piecewise-smooth Kohonen maps, *Optical Memory and Neural Networks*, 2012, Vol. 21, No. 4, pp. 227–232. DOI: 10.3103/S1060992X12040066.