

УДК 004.89

Н.Г. Аксак, С.А. Коргут

Харьковский национальный университет радиоэлектроники, Харьков

## МОДЕЛЬ КАТЕГОРИЗАЦИИ ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ НА ОСНОВЕ ДСМ-МЕТОДА

Рассмотрен подход к построению адаптивной информационно-поисковой системы при помощи категоризации пользователей. Разработана модель категоризации интернет-пользователей на основе метода автоматического порождения гипотез Джона-Стюарта-Милля (ДСМ-метод). Определены основные источники знаний о пользователе информационной системы. Рассмотрен пример отнесения пользователя к категориям при помощи предложенной модели.

**Ключевые слова:** категоризация, интернет-пользователь, ДСМ-метод.

### Введение

**Актуальность.** На современных интернет-ресурсах наряду с общей информацией представлена и узко специализированная, необходимая только определенной категории посетителей сайта. Возникают сложности поиска актуальной информации на ресурсе, которые усложняются также большими объемами и разнородностью данных, содержащихся на веб-сайтах. Одним из вариантов решения данной проблемы предлагается категоризация интернет-пользователя, т.е. отнесение пользователя к той или иной категории и предоставление ему информации, связанной с ней. Под категорией будем понимать виртуальное множество пользователей, обладающих схожими свойствами. Категории формируются в зависимости от тематики ресурса, для посетителей которого проводится категоризация.

Например, одна из категорий состоит из множества пользователей одного пола, другая – из одной возрастной группы; для медицинского веб-ресурса целесообразно разделение пользователей на три категории: врачи, пациенты клиники и посетители, которые не относятся к первым двум категориям. Категоризация позволит получать данные с учетом информационных интересов, что в свою очередь повысит качество поиска информации в сети Интернет. Таким образом, актуальной является разработка модели, позволяющей строить гипотезы о принадлежности интернет-пользователя к определенной категории.

**Постановка задачи.** Дано: информация, передаваемая браузером посетителя интернет-ресурса (доступные данные о пользователе).

**Требуется:** разработать модель категоризации интернет-пользователей, вырабатывающую гипотезы о принадлежности пользователя к категории.

### Решение задачи

Модель категоризации интернет-пользователей представим в виде:

$$D: I \rightarrow Y, \quad (1)$$

где  $D$  – отображающая функция, основанная на ДСМ-методе.

ДСМ-метод представим в виде структуры  $J = \langle C, O, P, V, F, H \rangle$ , где:

- $C$  – непустое конечное множество доступных данных о пользователе;
- $O = \subseteq \alpha(A)$  – непустое конечное множество пользователей, представленных в виде ИПП;
- $P$  – непустое конечное множество свойств пользователя;
- $V = \{+1, -1, 0, \tau\}$  – множество внутренних истинностных значений (+1 – эмпирически истинно, -1 – эмпирически ложно, 0 – эмпирически противоречиво,  $\tau$  – неопределенно);
- $F: O \times P \rightarrow V$ , отображение  $F$  будем называть функцией наличия свойства;
- $H: \alpha(A) \times P \rightarrow V$ , отображение  $H$  будем называть функцией причины свойства;
- $I$  – область определения функции  $D$ , множество данных о пользователе  $I\{i_1, i_2, i_3, i_4\}$ , где
  - $i_1$  – данные о браузере;
  - $i_2$  – данные об операционной системе;
  - $i_3$  – данные о мониторе;
  - $i_4$  – данные о сетевом соединении;
- $Y$  – область значения функции  $D$ , множество категорий  $Y\{y_1, y_2, y_3, y_4\}$ .

ДСМ-метод автоматического порождения гипотез – это один из методов интеллектуального анализа данных, позволяющий с помощью анализа имеющейся базы фактов сделать предположения о причинах наличия или отсутствия определенных свойств у объектов предметной области. ДСМ-метод оперирует сущностями трёх сортов: объекты предметной области, свойства этих объектов, возможные причины свойств. Предполагается, что объекты имеют структуру и причинами свойств объектов являются фрагменты этой структуры.

На вход ДСМ-метод получает некоторое множество изучаемых объектов  $O = \{O_1, O_2, \dots, O_n\}$  (информационный портрет пользователя – ИПП) и сведения об их структуре (параметры ИПП), о наличии или отсутствии у них определенных свойств (социальный статус, пол, возраст), а также, в некоторых случаях, о связи между структурой объектов и их свойств (социальный статус определяется типом процессора, разрешением монитора, скоростью соединения). Кроме того, имеется ряд целевых признаков (социальный статус «высокий»), каждый из которых разбивает исходное множество объектов на четыре непересекающихся подмножества

$$O = O_1 \cup O_2 \cup O_3 \cup O_4 :$$

- объекты  $O_1$ , про которые известно, что они обладают данным признаком (другими словами, пользователи, уровень социального статуса которых известен);
- объекты  $O_2$ , про которые известно, что они не обладают данным признаком;
- объекты  $O_3$ , для которых существуют аргументы как за, так и против того, что они обладают данным признаком;
- объекты  $O_4$ , о которых неизвестно, обладают они этим признаком или нет.

Результатом применения ДСМ-метода являются гипотезы двух типов:

- гипотезы о связи определенных структурных фрагментов (данных ИПП и их комбинаций) изучаемых объектов (ИПП) со свойствами, которыми они обладают;
- гипотезы о наличии или отсутствии целевых признаков у объектов, для которых изначально это было неизвестно, формируемые на основании установленной взаимосвязи между свойствами объектов и их структурными компонентами.

Таким образом, сущностями модели категоризации интернет-пользователей являются:

- $O$  – множество ИПП;
- $P$  – множество свойств объектов (3 уровня социального статуса);
- $C$  – множество возможных причин свойств объектов (операционная система, разрешение монитора, скорость соединения);
- $V$  – множество оценок.

Введём функцию  $F: O \times P \rightarrow V$ , представляющую начальную ситуацию.

- $F(o, p) = +1$  – известно, что объект  $o$  обладает свойством  $p$ ;
- $F(o, p) = -1$  – известно, что объект  $o$  не обладает свойством  $p$ ;
- $F(o, p) = 0$  – есть аргументы как за, так и против того, что объект  $o$  обладает свойством  $p$ ;

- $F(o, p) = \tau$  – неизвестно, обладает ли объект  $o$  свойством  $p$ .

Использование ДСМ-метода позволяет определить, например, какими значениями параметров ИПП должен обладать пользователь, чтобы можно было сделать предположение о социальном статусе «низкий», «средний», «высокий».

Множество объектов:

- $o_1$  – ИПП 1;
- $o_2$  – ИПП 2;
- ...
- $o_{10}$  – ИПП 10.

Таблица 1

Исходные данные для ДСМ-метода

	$O_1$	$O_2$	$O_3$	$O_4$	$O_5$	$O_6$	$O_7$	$O_8$	$O_9$	$O_{10}$
$p_1$	+	+	+	-	-	-	-	-	-	$\tau$
$p_2$	-	-	-	+	+	+	-	-	-	$\tau$
$p_3$	-	-	-	-	-	-	+	+	+	$\tau$
$c_1$	-	+	-	+	-	-	-	-	+	+
$c_2$	+	-	+	-	+	-	-	-	-	-
$c_3$	-	-	-	-	-	+	+	+	-	-
$c_4$	+	+	-	-	-	-	-	-	-	-
$c_5$	-	-	+	-	-	+	+	-	-	-
$c_6$	-	-	-	+	+	-	-	+	+	+
$c_7$	+	-	+	-	-	+	-	-	-	-
$c_8$	-	+	-	-	+	-	-	-	-	-
$c_9$	-	-	-	+	-	-	+	-	-	+
$c_{10}$	-	-	-	-	-	-	-	+	+	-

Для этих объектов выберем следующий набор структурных фрагментов  $C$ :

- $c_1$  – ОС Windows Vista;
- $c_2$  – ОС Windows XP;
- $c_3$  – ОС Unix;
- $c_4$  – разрешение 800×600 и меньше;
- $c_5$  – разрешение 1024×768;
- $c_6$  – разрешение 1200×800 и больше;
- $c_7$  – скорость соединения меньше 56 kbit/s;
- $c_8$  – скорость соединения от 56 kbit/s до 256 kbit/s;
- $c_9$  – скорость соединения от 256 kbit/s до 512 kbit/s;
- $c_{10}$  – скорость соединения выше 512 kbit/s.

Множество целевых признаков:

- $p_1$  – соц. статус «низкий»;
- $p_2$  – соц. статус «средний»;
- $p_3$  – соц. статус «высокий».

Представим каждый из объектов набором структурных компонентов, которыми этот объект обладает:

- $O_1$  { $c_2, c_4, c_7$ };
- $O_2$  { $c_1, c_4, c_8$ };
- $O_3$  { $c_2, c_5, c_7$ };
- $O_4$  { $c_1, c_6, c_9$ };
- $O_5$  { $c_2, c_6, c_8$ };

- $o_6 \{c_3, c_5, c_7\}$ ;
- $o_7 \{c_3, c_5, c_9\}$ ;
- $o_8 \{c_3, c_6, c_{10}\}$ ;
- $o_9 \{c_1, c_6, c_{10}\}$ ;
- $o_{10} \{c_1, c_6, c_9\}$ .

В данном случае положительными примерами для целевого свойства  $p_1$  является объект  $o_1, o_2, o_3$ , отрицательными –  $o_4, o_5, o_6, o_7, o_8, o_9$ , и один (τ)-пример –  $o_{10}$ :

$p_1$ : (+)  $o_1, o_2, o_3$ ; (-)  $o_4, o_5, o_6, o_7, o_8, o_9$ ; (τ)  $o_{10}$ ;

$p_2$ : (+)  $o_4, o_5, o_6$ ; (-)  $o_1, o_2, o_3, o_7, o_8, o_9$ ; (τ)  $o_{10}$ ;

$p_3$ : (+)  $o_7, o_8, o_9$ ; (-)  $o_4, o_5, o_6, o_7, o_8, o_9$ ; (τ)  $o_{10}$ .

Задача состоит в том, чтобы, используя правдоподобные рассуждения, выяснить, обладают ли (τ)-примеры целевым свойством  $p_1, p_2, p_3$ .

*Применение правил первого рода.* В качестве возможных причин одного из свойства  $p_1, p_2, p_3$  у объектов рассмотрим некоторые непустые подмножества множества структурных фрагментов  $C$ .

(+)-условию  $p_1$  удовлетворяют множества:

- $C_1 = \{c_2, c_7\}; \Omega = \{o_1, o_3\}$ ;
- $C_2 = \{c_4\}; \Omega = \{o_1, o_2\}$ .

(-)-условию  $p_1$  удовлетворяют множества:

- $C_3 = \{c_1\}; \Omega = \{o_4, o_9\}$ ;
- $C_4 = \{c_3\}; \Omega = \{o_6, o_7, o_8\}$ ;
- $C_5 = \{c_5\}; \Omega = \{o_6, o_7\}$ ;
- $C_6 = \{c_3, c_5\}; \Omega = \{o_6, o_7\}$ ;
- $C_7 = \{c_6\}; \Omega = \{o_4, o_5, o_8, o_9\}$ ;
- $C_8 = \{c_9\}; \Omega = \{o_4, o_7\}$ ;
- $C_9 = \{c_{10}\}; \Omega = \{o_8, o_9\}$ ;
- $C_{10} = \{c_1, c_6\}; \Omega = \{o_4, o_9\}$ ;
- $C_{11} = \{c_6, c_{10}\}; \Omega = \{o_8, o_9\}$ .

Теперь необходимо выяснить, являются ли найденные множества возможными причинами наличия или отсутствия целевых свойств  $p_1, p_2, p_3$  у объектов, то есть определить функцию  $H$  для данного шага. Как говорилось ранее, правила определения данной функции могут иметь различный вид в зависимости от выбранной стратегии – с запретом (или без запрета) на контр-примеры.

Множество  $C_i \subseteq C$  будем доопределять как

- возможную причину наличия свойства  $p$ , если  $C_i$  удовлетворяет (+)-условию для  $p$ , то есть вкладывается как подмножество в два и более (+)-примера и при этом не вкладывается ни в один (-)-пример; (вкладывается не более чем в один (-)-пример);

- возможную причину отсутствия свойства  $p$ ,  $C_i$  удовлетворяет (-)-условию для  $p$ , то есть вкладывается как подмножество в два и более (-)-примера и при этом не вкладывается ни в один (+)-пример; (вкладывается не более чем в один (+)-пример);

- противоречивую гипотезу, если существуют как (+)-пример, так и (-)-пример, в который вкладывается  $C_i$ .

Возможные причины наличия свойства  $p_1$ :

- $C_1 = \{c_2, c_7\}$ ;

- $C_2 = \{c_4\}$ .

Все множества, удовлетворяющие (-)-условию для  $p_1$ , доопределяются как возможные причины отсутствия свойства  $p_1$ :

- $H(C_1, p_1) = +1$ ;
- $H(C_2, p_1) = +1$ ;
- $H(C_3, p_1) = -1$ ;
- $H(C_4, p_1) = -1$ ;
- $H(C_5, p_1) = -1$ ;
- $H(C_6, p_1) = -1$ ;
- $H(C_7, p_1) = -1$ ;
- $H(C_8, p_1) = -1$ ;
- $H(C_9, p_1) = -1$ ;
- $H(C_{10}, p_1) = -1$ ;
- $H(C_{11}, p_1) = -1$ .

(+)-условию  $p_2$  удовлетворяют множества:

- $C_{12} = \{c_6\}; \Omega = \{o_4, o_5\}$ ;

(-)-условию  $p_2$  удовлетворяют множества:

- $C_9 = \{c_{10}\}; \Omega = \{o_8, o_9\}$ ;
- $C_{11} = \{c_6, c_{10}\}; \Omega = \{o_8, o_9\}$ ;
- $C_{13} = \{c_1\}; \Omega = \{o_2, o_9\}$ ;
- $C_{14} = \{c_2\}; \Omega = \{o_1, o_3\}$ ;
- $C_{15} = \{c_3\}; \Omega = \{o_7, o_8\}$ ;
- $C_{16} = \{c_4\}; \Omega = \{o_1, o_2\}$ ;
- $C_{17} = \{c_5\}; \Omega = \{o_3, o_7\}$ ;
- $C_{18} = \{c_6\}; \Omega = \{o_8, o_9\}$ ;
- $C_{19} = \{c_7\}; \Omega = \{o_1, o_3\}$ ;
- $C_{20} = \{c_2, c_7\}; \Omega = \{o_1, o_3\}$ .

Возможные причины наличия свойства  $p_2$  отсутствуют. Таким образом,

- $H(C_9, p_2) = -1$ ;
- $H(C_{11}, p_2) = -1$ ;
- $H(C_{13}, p_2) = -1$ ;
- $H(C_{14}, p_2) = -1$ ;
- $H(C_{15}, p_2) = -1$ ;
- $H(C_{16}, p_2) = -1$ ;
- $H(C_{17}, p_2) = -1$ ;
- $H(C_{18}, p_2) = -1$ ;
- $H(C_{19}, p_2) = -1$ ;
- $H(C_{20}, p_2) = -1$ .

(+)-условию  $p_3$  удовлетворяют множества:

- $C_{21} = \{c_3\}; \Omega = \{o_7, o_8\}$ ;
- $C_{22} = \{c_6\}; \Omega = \{o_8, o_9\}$ ;
- $C_{23} = \{c_{10}\}; \Omega = \{o_8, o_9\}$ .

(-)-условию  $p_3$  удовлетворяют множества:

- $C_{20} = \{c_2, c_7\}; \Omega = \{o_1, o_3\}$ ;
- $C_{24} = \{c_5, c_7\}; \Omega = \{o_3, o_6\}$ ;
- $C_{25} = \{c_1\}; \Omega = \{o_2, o_4\}$ ;
- $C_{26} = \{c_2\}; \Omega = \{o_1, o_3, o_5\}$ ;
- $C_{27} = \{c_4\}; \Omega = \{o_1, o_2\}$ ;
- $C_{28} = \{c_5\}; \Omega = \{o_3, o_6\}$ ;
- $C_{29} = \{c_6\}; \Omega = \{o_4, o_5\}$ ;
- $C_{30} = \{c_7\}; \Omega = \{o_1, o_3, o_6\}$ ;
- $C_{31} = \{c_8\}; \Omega = \{o_2, o_5\}$ .

Возможные причины наличия свойства  $p_3$ :

- $C_{21} = \{c_3\}$ ;
- $C_{23} = \{c_{10}\}$ .

Получаем:

- $H(C_{21}, p_3) = +1$ ;
- $H(C_{23}, p_3) = +1$ ;
- $H(C_{20}, p_3) = -1$ ;
- $H(C_{24}, p_3) = -1$ ;
- $H(C_{25}, p_3) = -1$ ;
- $H(C_{26}, p_3) = -1$ ;
- $H(C_{27}, p_3) = -1$ ;
- $H(C_{28}, p_3) = -1$ ;
- $H(C_{29}, p_3) = -1$ ;
- $H(C_{30}, p_3) = -1$ ;
- $H(C_{31}, p_3) = -1$ .

*Применение правил второго рода.* Используем полученные на предыдущем шаге (+)- и (-)-гипотезы для определения  $\tau$ -примеров. Такой пример всего один:  $o_{10} \{c_1, c_6, c_9\}$ .

В него не вкладывается ни одна из возможных причин наличия свойств  $p_1, p_2, p_3$  и вкладывается четыре возможных причины отсутствия свойства  $p_1$  ( $C_3, C_7, C_8, C_{10}$ ), две возможных причины отсутствия свойства  $p_2$  ( $C_{13}, C_{18}$ ) и две возможных причины отсутствия свойства  $p_3$  ( $C_{25}, C_{29}$ ). Таким образом,  $\tau$ -пример  $o_{10}$  с большей уверенностью обладает свойством  $p_2$  или  $p_3$ , чем  $p_1$ , т.е. посетитель с информационным портретом  $o_{10}$  скорее имеет «средний» или «высокий» социальный статус, чем «низкий».

*Проверка каузальной полноты.* Проверка каузальной полноты осуществляется рассуждениями типа: условие выполняется, если в каждый исходный (+)-пример вкладывается хотя бы одна возможная причина наличия целевого свойства, а в каждый (-)-пример – хотя бы одна возможная причина его отсутствия. В данном случае каждый исходный положительный и отрицательный пример является объясненным.

Таким образом, получено правдоподобные достаточные условия для того, чтобы пользователю был присвоен определенный социальный статус:

1. Социальный статус «низкий»: ОС Windows XP и скорость соединения меньше 56 kbit/s или разрешение 800×600 и меньше.

2. Социальный статус «высокий»: ОС Unix или скорость соединения выше 512 kbit/s.

Для более точного определения социального статуса посетителя сайта необходимо увеличить количество структурных фрагментов  $C$  (добавить данные обратной связи с пользователем, время посещения, разделить пользователей стационарных компьютеров и ноутбуков и т.д. – факторы, тем или иным образом зависящие от социального статуса) и количество объектов  $O$  (увеличить количество ИПП, участвующих в порождении гипотез).

## Выводы

Предложена модель категоризации интернет-пользователей на основе ДСМ-метода, позволяющая разделить посетителей сайта на категории с целью предоставления им актуальной информации, что в свою очередь позволяет повысить качество поиска информации в сети Интернет.

## Список литературы

1. Обьедков С.А. Алгоритмические аспекты ДСМ-метода и формального анализа понятий / С.А. Обьедков. – М.: РГГУ, 1999.
2. Финн В.К. Базы данных с неполной информацией и новый метод автоматического порождения гипотез / В.К. Финн // Диалоговые и фактографические системы информационного обеспечения. – М., 1981.
3. Виноградов Д.В. Несимметричный ДСМ-метод с учетом контекста / Д.В. Виноградов // Пятая национальная конференция с международным участием. Искусственный интеллект-96. – Казань: 1996. – КИИ-96: Сб. науч. тр.: В 3 т. – Казань: Ассоц. искусств. Интел., 1996.
4. Cooley Robert. Data Preparation for Mining World Wide Web Browsing Patterns / Robert Cooley, Bamshad Mobasher, Jaideep Srivastava // Knowledge and Information Systems. – 1, 1999.

Поступила в редколлегию 17.02.2010

**Рецензент:** д-р техн. наук, проф. Г.Г. Четверяков, Харьковский национальный университет радиоэлектроники, Харьков.

## МОДЕЛЬ КАТЕГОРИЗАЦІЇ ІНТЕРНЕТ-КОРИСТУВАЧІВ НА ОСНОВІ ДСМ-МЕТОДУ

Н.Г. Аксак, С.А. Коргут

*Розглянутий підхід до побудови адаптивної інформаційно-пошукової системи за допомогою категоризації користувачів. Розроблена модель категоризації інтернет-користувачів на основі методу автоматичного породження гіпотез Джона-Стюарта-Мілля (ДСМ-метод). Визначені основні джерела знань про користувача інформаційної системи. Розглянутий приклад віднесення користувача до категорій за допомогою запропонованої моделі.*

**Ключові слова:** категоризація, інтернет-користувач, ДСМ-метод.

## MODEL OF CATEGORIZING OF INTERNET-USERS ON THE BASIS OF DSM-METHOD

N.G. Axak, S.A. Korgut

*Approach is considered to the construction of the adaptive information storage and retrieval system through categorizing of users. The model of categorizing of Internet-users is developed on the basis of method of automatic generation of hypotheses of Dzhoa-Styuart-Mill (DSM-method). The basic sources of knowledges are certain about the user of the informative system. The example of taking of user is considered to the categories through the offered model.*

**Keywords:** categorizing, Internet-users, DSM-method.