

МОДЕЛЬ ИНТЕГРАЦИИ ДАННЫХ В ЕДИНОЕ ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО ПРЕДПРИЯТИЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДА N-ГРАММ

© 2012 А.Н. Подобрый

ФНПЦ ОАО “НПО “Марс”, г. Ульяновск

Поступила в редакцию 02.11.2012

В статье рассматривается подход к реализации сопоставления данных с помощью модифицированного метода n-грамм. Выводится модель интегрированной информационной системы и доступа к сопоставленным данным в рамках корпоративной информационной сети. Статья предназначена для специалистов, занимающихся внедрением автоматизированных систем управления предприятием, а также для специалистов занимающихся вопросами интеграции данных.

Ключевые слова: интеграция данных, n-грамм, матрица доступа, предприятие, модель интеграции.

Интеграция данных – обеспечение единого согласованного представления данных для ряда информационных ресурсов, объединенных общим смысловым содержанием, и/или на основе общего представления – частных представлений.

Интеграция данных в информационных системах понимается как обеспечение единого унифицированного интерфейса для доступа к некоторой совокупности, неоднородных независимых источников данных. Таким образом, для пользователя информационные ресурсы всей совокупности интегрируемых источников представляются как новый единый источник. Система, обеспечивающая пользователю такие возможности, называется *системой интеграции данных* [1, 2].

Система интеграции данных освобождает пользователей от необходимости знания, данные из каких источников, кроме интегрированного, они используют, каковы свойства этих источников и как осуществить доступ к ним. Доступ к данным многих источников обеспечивается через единый интерфейс, который представляет совокупность данных из множества независимых источников в терминах единой модели данных на основании прав доступа.

Существуют три основных типа информационного поиска: булевый, поиск по релевантности и поиск по сходству. Для сопоставления неоднозначных наименований объектов предлагается использовать метод нечеткого поиска основанный на релевантности – количественного критерия схожести. В основе данного метода лежит модифицированный метод n-грамм.

Целями статьи являются построение математической модели интегрированной информационной системы на основании интеграции данных из разных информационных систем с помощью сопоставления данных, используя метод

Подобрый Александр Николаевич, ведущий инженер-программист. E-mail: mars@mv.ru

n-грамм, и матрицы прав доступа сотрудников предприятия.

МОДЕЛЬ ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Основным понятием описываемой модели является понятие информационного объекта, являющимся сущностью информационной системы [3].

Набором информационных объектов назовем множество упорядоченных пар вида:

$$x = \{ \langle a_1, b_1 \rangle, \langle a_2, b_2 \rangle, \dots, \langle a_n, b_n \rangle \},$$

$$a_i \triangleleft a_j \text{ для } i \triangleleft j, i, j \in \{1..n\},$$

где a – имя атрибута (идентификатор),
 b – характеристика (свойство) объекта.

Набор свойств b_i можно представить как кортеж полей

$$b_i = \{e_{i1}, e_{i2}, \dots, e_{in}\},$$

каждое поле которого состоит из набора:

$$e_{ij} = (v_{ij}, s_{ij}),$$

где v_{ij} – значение атрибута,

s_{ij} – тип атрибута

Таким образом, набор свойств информационного объекта заданных множеством характеристик ставят в соответствие каждому объекту некоторое свойство $b_i \in B = \{b_1, b_2, \dots, b_n\}$.

Через O^x обозначим множество атрибутов элемента x , через $o_i(x)$ – характеристика атрибута o_i объекта x .

Информационную систему рассмотрим как структуру информационной схемы, описывающей характеристики входящих в эту систему информационных объектов.

Информационная схема включает в себя следующие характеристики:

- множество характеристик информационных объектов;
- множество атрибутов для информационных объектов каждой характеристики;
- множество используемых атрибутов;

- множество связей между информационными объектами;

- множество иерархий (вложенностей) объектов.

Связи между информационными объектами зададим элементами множества $L = \{l_1, l_2, \dots, l_k\}$, где каждый элемент является тройкой:

$$l_j = \{o_{j1}, o_{j2}, r_j\},$$

где o_{j1}, o_{j2} – два связанных между собой объекта, r_j – наименование (вид) роли, по которой связываются объекты.

Иерархия (дерево вложенности) объектов опишем множеством $T = \{t_1, t_2, \dots, t_n\}$, где каждый элемент – это пара $t_i = (o_i, o_j)$, где $i \in \{1..n\}$, в которой первый элемент соответствует объекту с тем же индексом, а второй указывает на объект, который является родительским по отношению к данному в иерархии.

Информационной схемой назовем набор: $M = \langle O, B, L, T \rangle$,

где $O = \{o_1, o_2, \dots, o_n\}$ – множество атрибутов информационных объектов;

$B = \{b_1, b_2, \dots, b_n\}$ – множество возможных характеристик атрибутов;

$L = \{l_1, l_2, \dots, l_k\}$ – множество связей между информационными объектами;

$T = \{t_1, t_2, \dots, t_p\}$ – множество иерархий (вложенностей) объектов

Информационной системой построенной по схеме M назовем набор:

$$S^M = \langle M, S, \beta \rangle,$$

где $M = \langle O, B, L, T \rangle$ – информационная схема;

$S = \{x_1, x_2, \dots, x_n\}$ – множество информационных объектов;

$\beta: S \rightarrow B$ – отображение, ставящее каждому объекту его характеристику.

Причем для каждого информационного объекта $x \in S$ выполнено следующее условие:

Все атрибуты x имеют значение и тип, т.е. для любой пары $\langle a, b \rangle \in x$ существует $b \in \beta(a)$.

Пусть $\bar{M} = \{M_1, M_2, \dots, M_n\}$ – множество схем информационных систем. Используя $S^{\bar{M}} = \{S_1^{M_1}, S_2^{M_2}, \dots, S_n^{M_n}\}$, обозначим множество информационных систем, каждая из которых имеет соответствующую схему из $\bar{M}$.

Для того, чтобы множества информационных систем можно было рассматривать как единую информационную систему, необходимо, чтобы разные информационные системы, входящие в $S^{\bar{M}}$ могли иметь пересекающиеся множества атрибутов объектов и соответствующие им характеристики.

Таким образом, для построения интегрированной информационной системы, необходимо, чтобы отображения – B , ставящие каждому объекту его свойство на разных информационных системах совпадали.

Множество информационных систем $S^{\bar{M}}$ непротиворечиво если существует $S^M = \{S_1^{M_1}, S_2^{M_2}, \dots, S_n^{M_n}\}$ – множество информационных систем, где $S_i^M = \langle M_i, S_i, \beta_i \rangle$ и $M_i = \langle O_i, B_i, L_i, T_i \rangle$, для которых:

$$\bar{S} = \bigcup_{1 \leq i \leq N} S_i, \bar{O} = \bigcup_{1 \leq i \leq N} O_i, \bar{B} = \bigcup_{1 \leq i \leq N} B_i, \bar{L} = \bigcup_{1 \leq i \leq N} L_i, \bar{T} = \bigcup_{1 \leq i \leq N} T_i.$$

При условии, что существует отображение $\beta: S \rightarrow B$, являющееся расширением каждого отображения β_i для $1 \leq i \leq n$.

Информационная система S' называется интегрированной на множестве информационных систем $S^{\bar{M}}$, если $S^M = \{S_1^{M_1}, S_2^{M_2}, \dots, S_n^{M_n}\}$ непротиворечиво.

Таким образом, будем считать информационную систему $S' = \langle \bar{M}, \bar{S}, \bar{\beta} \rangle$, где $\bar{M} = \langle \bar{O}, \bar{B}, \bar{L}, \bar{T} \rangle$ интегрированной на множестве $S^{\bar{M}}$.

ОПИСАНИЕ МОДЕЛИ СРАВНЕНИЯ ХАРАКТЕРИСТИК ИНФОРМАЦИОННЫХ ОБЪЕКТОВ

N-граммой на алфавите E некоторого языка $L(E)$ будем представлять набор символов длиной n строки Y [5],

где $E = \{v, e_1, e_2, \dots, e_l\}$ – алфавит;

$L(E)$ – некоторый язык на алфавите E ;

Y – строка символов.

N-грамма может совпадать с какой-либо строкой, быть его подстрокой или не входить в язык $L(E)$:

- если алфавит $E = \{v, e_1, e_2, \dots, e_l\}$ и строка $Y = \{y_1, y_2, \dots, y_n, \$\}$, где $\forall i, y_i \in A$, то n -грамма – это последовательность из n символов, принадлежащая одному слову;

- если строки – это тексты, то n -грамма – это последовательность из N слов одного текста.

Число вхождений строки Y опишем множеством: $C(w) = C(w_1, w_2, \dots, w_n)$,

где $W = \{w_1, w_2, \dots, w_n\}$ есть совокупность всех слов рассматриваемого языка $L(E)$

Вероятность $P(W)$ появления n -граммы $W = \{w_1, w_2, \dots, w_n\}$ вычисляется по формуле:

$$P(w_i) = \frac{C(w_i)}{\sum_{w_j} C(w_j)},$$

где w_i – n – грамма;

$C(w_i)$ – количество вхождений w_i ;

$\sum C(w_j)$ – общее число возможных n – грамм.

Если вероятность появления символов в любой позиции строки Y имеют одну и ту же вероятность, то формулу вероятности можно представить:

$$P(w_1 \dots w_n) = \prod_{i=1}^n p(w_i)$$

Таким образом, любые перестановки символов строки Y имеют одну и ту же вероятность.

Релевантность есть степень (коэффициент) соответствия поискового шаблона $P = \{p_1, p_2, \dots, p_n\}$ и просматриваемого текста. Данный коэффициент можно воспринимать как процент вхождения поисковой строки к общему объему текста.

Формулу релевантности можно представить как:

$$R = \frac{\sum_{i=1}^N r(i)}{N};$$

$$r(i) = \frac{\text{sovp}(str1, str2, i) + \text{sovp}(str2, str1, i)}{C(str1, i) + C(str2, i)},$$

где $\text{sovp}(Y_1, Y_2, i)$ – сумма совпадений всех подстрок длиной i из строки Y_1 в строке Y_2 ;

$C(str1, i)$ – общее число возможных n – грамм длиной i ;

N – фиксированная длина максимальной подстроки.

Данный метод позволяет получить схожие тексты информации с высокой долей вероятности. За счет увеличения фиксированной длины подстроки N , и в случае, когда N будет равно длине строки, при отсутствии точных дублей строк будет получено полностью равномерное распределение.

Недостатком данного метода можно считать низкую степень релевантности при сравнении строк с одинаковым набором слов, но с разным порядком следования.

Пример 1:

Вычислим коэффициент релевантности наименований двух атрибутов: Y_1 = (“Иванов Сергей”) и Y_2 = (“Сергей Иванов”)

При $N=3$ $R=0.72$,

$N=4$ $R=0.6$

Для сравнения данных атрибутов разобьем строки Y_1 и Y_2 на слова с помощью пробельного символа x ,

где $G^1 = \{g_1, g_2, \dots, g_m\} \in Y_1$ – набор слов первой строки;

$G^2 = \{g_1, g_2, \dots, g_k\} \in Y_2$ – набор слов второй строки;

$|G^1| = m, |G^2| = k$ – количества слов в строках.

Таким образом, вероятность совпадения слова g_j строки Y_1 и строки Y_2 при длине n -грамма равной i получаем:

$$r(i) = \frac{\text{sovp}(g_j, str2, i)}{C(g_j, i)},$$

где $\text{sovp}(g_j, str2, i)$ – сумма совпадений всех подстрок длиной i из набора слов G в строке Y_2 ;

$C(g_j, i)$ – общее число возможных n – грамм длиной i в слове G

Формула релевантности соответствия g_i – отдельно взятого слова строки Y_1 со строкой Y_2 будет выражаться:

$$R(g_j) = \frac{\sum_{i=1}^N r(i)}{N}.$$

Коэффициент релевантности строки Y_1 со строкой Y_2 :

$$R(G^1) = \frac{\sum_{j=1}^m R(g_j)}{m},$$

где $R(g_j)$ – коэффициент релевантности g_i слова строки Y_1 ;

m – количество слов в строке Y_1 .

Аналогично выводится коэффициент релевантности для строки Y_2 .

Общий коэффициент соответствия двух строк Y_1 и Y_2 можно представить как:

$$R_{main} = \frac{R(G^1) + R(G^2)}{2},$$

где $R(G^1)$ – коэффициент релевантности строки Y_1 со строкой Y_2 ;

$R(G^2)$ – коэффициент релевантности строки Y_2 со строкой Y_1 .

Пример 2:

Вычислим коэффициент релевантности наименований двух атрибутов: Y_1 = (“Бензопила белая”) и Y_2 = (“Бензопила”)

При $N=3$,

$R(G^1) = 0.66, R(G^2) = 1, R_{main} = 0.83$

Таким образом, на основании полученных трех коэффициентов релевантности строк, фиксированная длина максимальной подстроки можно с высокой долей вероятности определить однозначные атрибуты. Для определения однозначного соответствия двух строк достаточно сделать длину подстроки плавающей в зависимости от максимальной длины g_j .

Результат сопоставления можно увидеть на рис. 1, где данные берутся из разных корпоративных информационных систем и представляются в едином виде.

ПОСТРОЕНИЕ ЕДИНОГО ПОЛЬЗОВАТЕЛЬСКОГО ИНТЕРФЕЙСА

Информационной основой автоматизированной интегрированной информационной си-

Единая точка входа

Форма для поиска приходников и требований накладных:

Поставщик: ООО "ТК "ДЛЯ ВАС"

сокращенный поиск

Склад:

Период с: 01.01.2012 по: 24.10.2012

Номер приходного ордера:

№	Руков.	Клад.	Дата	№ прих орд	Склад	Поставщик
1	v	v	13.01.2012	9558	98001	ООО "ТК "ДЛЯ ВАС"
2	v	v	19.01.2012	184	98001	ООО "ТК "ДЛЯ ВАС"
3	v	v	27.01.2012	105	98001	ООО "ТК "ДЛЯ ВАС"
4	v	v	07.02.2012	498	98001	ООО "ТК "ДЛЯ ВАС"
5	v	v	02.03.2012	1044, 1097, 1316	98001	ООО "ТК "ДЛЯ ВАС"
6	v	v	23.04.2012	2757	98001	ООО "ТК "ДЛЯ ВАС"

Internet Explorer предоставлен: КИСП

НДС 18% Приходный ордер №3095

Дата составления		Склад	Поставщик		Код
10.05.2012		98001	ООО "ТК "ДЛЯ ВАС"		5287

№	Наименование	Ед. изм.	Колл по докум	Колл принято	цена	сумма без НДС	сумма НДС	Всего с НДС	Пор.номер по скл.карточке
1	Лист г/к 2,0 мм ст.3	т	0.1020	0.1020	27139.80	2768.26	498.29	3266.55	21678
2	Лист х/к 1,2 мм ст.08пс	т	0.0310	0.0310	31055.16	962.71	173.29	1136.00	19488
3	Труба 15х2,8 ГОСТ 3262-75	т	0.0320	0.0320	31589.06	1010.85	181.95	1192.80	18525
4	Труба 40х40х2,0 ГОСТ 8639-82	т	0.0420	0.0420	33101.67	1390.27	250.25	1640.52	17859
5	Уголок 32х32х4,0	т	0.0130	0.0130	28920.00	375.96	67.67	443.63	19416
6	Уголок 45х45х4,0	т	0.0680	0.0680	27389.85	1862.51	335.25	2197.76	18523
7	Уголок 63х63х5,0	т	0.0590	0.0590	27440.68	1619.00	291.42	1910.42	25576
8	Швеллер 14	т	0.0710	0.0710	29118.59	2067.42	372.14	2439.56	18544
9	Труба 108х5 ГОСТ 8732-78	т	0.1350	0.1350	45203.41	6102.46	1098.44	7200.90	23561

Данные КИС «Марс 1С»
Данные КИС «Флагман»

Рис. 1. Окно программы просмотра приходных ордеров из КИС «Марс 1С» и КИС «Флагман»

стемы управления предприятия должна являться единая распределенная база данных, в которой каждый пользователь, обладающий соответствующими правами доступа, может своевременно получить информацию из разных информационных систем [7].

Объект информационного ресурса интегрированной информационной системы можно представить как набор полей: $h = \{ \langle a_1, b_1 \rangle, \langle a_2, b_2 \rangle, \dots, \langle a_n, b_n \rangle \}$,

$$a_i \diamond a_j \text{ для } i \diamond j, \\ i, j \in \{1..n\}, a_i \in \bar{O}, b_i \in \bar{B},$$

где a – идентификатор объекта,

b – значение объекта.

Информационным ресурсом интегрированной информационной системы есть набор:

$$U = \{H, V, D\},$$

где $H = \{h_1, h_2, \dots, h_n\}$ – множество объектов информационного ресурса;

$V = \{v_1, v_2, \dots, v_n\}$ – множество представлений;

$D = \{d_1, d_2, \dots, d_n\}$ – множество прав доступа пользователей;

Представлением V информационного ресурса интегрированной информационной системы называется именованная функция или процедура представления пользователю набора информационных сущностей одного типа, определенных в виде реляционной таблицы. Представление, в свою очередь, включает в себя следующие множества:

- набор входных параметров $P = \{p_1, p_2, \dots, p_n\}$

– на основании их значений и формируется конкретный экземпляр представления;

- набор выходных параметров $K = \{k_1, k_2, \dots, k_n\}$

– поля, обеспечивающие уникальность каждой записи.

Множество всех информационных ресурсов обозначим $U = \{u_1, u_2, \dots, u_l\}$,

где l – количество всех ресурсов.

Матрица прав доступа сотрудников предприятия к интегрированным объектам информационных ресурсов представляется:

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1k} \\ S_{12} & S_{22} & \dots & S_{2k} \\ \dots & \dots & \dots & \dots \\ S_{l1} & S_{l2} & \dots & S_{lk} \end{pmatrix},$$

где S_{ij} – множество прав доступа на просмотр и редактирование;

S_{i1} – информационный ресурс из множества U ;

S_{ik} – информационный объект из множества $\bar{O}$.

Множество прав доступа к информационным ресурсам S_{ij} состоит из двоек вида:

$$(s_p, s_c), \text{ причем } s_p, s_c \in \{0;1\},$$

где s_p – правила доступа на просмотр;

s_c – правила доступа на изменение.

Таким образом матрицу прав доступа можно представить:

$$S = \begin{pmatrix} (s_{p11}, s_{c11}) & (s_{p12}, s_{c12}) & \dots & (s_{p1k}, s_{c1k}) \\ (s_{p21}, s_{c21}) & (s_{p22}, s_{c22}) & \dots & (s_{p2k}, s_{c2k}) \\ \dots & \dots & \dots & \dots \\ (s_{pl1}, s_{cl1}) & (s_{pl2}, s_{cl2}) & \dots & (s_{plk}, s_{clk}) \end{pmatrix}$$

	S ₁	S ₂	...	S _m
S ₁	1	0		1
S ₂	0	1		0
...				
S _m	1	0		1

Рис. 2. Пример матрицы прав доступа отдельно взятого сотрудника

Тогда совокупность элементов прав доступа к информационным ресурсам интегрированной информационной системы можно показать:

$$E = \{O, U, S\},$$

где O – множество информационных объектов из множества $\bar{O}$;

U – множество информационных ресурсов;
 S – правила доступа.

Основным объектом идентификации корпоративной информационной сети защищенной инфраструктуры предприятия является учетная запись. Данные которой хранятся в доменной службе, представляющей из себя распределенную базу данных сведений о сетевых ресурсах [8].

Доменная учетная запись сотрудника предприятия будем называть множеством пар вида: $s = \{ \langle a_1, b_1 \rangle, \langle a_2, b_2 \rangle, \dots, \langle a_n, b_n \rangle \}$, где a – наименование атрибута;

b – значение атрибута.

Через A^s обозначим множество атрибутов элемента s , через $a_i(s)$ – характеристика атрибута a_i объекта s .

Информационной системой доменной службы называется кортеж:

$$D = \langle A, B, L, l \rangle,$$

где $A = \{a_1, a_2, \dots, a_n\}$ – множество атрибутов учетных записей;

$B = \{b_1, b_2, \dots, b_n\}$ – множество возможных значений возможных атрибутов;

$L = \{l_1, l_2, \dots, l_n\}$ – множество иерархий (вложенностей) объектов;

$l: A \rightarrow B$ – отображение, ставящее каждому атрибуту его значение.

Множество иерархий вложенности объектов $L = \{l_1, l_2, \dots, l_n\}$ – это набор из (o_i, o_j) , где $i \in \{1..n\}$, в котором первый элемент соответствует объекту с тем же индексом, а второй указывает на объект, который является родительским по отношению к данному в иерархии.

Таким образом, интегрированная информационная система S' и доменная система D образуют набор данных о сотрудниках предприятия и учетных записей корпоративной информационной сети предприятия. Тогда функция соответствия F – это отображение двух множеств $S' \xrightarrow{F} D$ (рис. 3).

В общем случае каждому объекту из S' может соответствовать произвольное количество объектов из D , и наоборот: каждому объекту из D может соответствовать произвольное количество объектов из S' . Это связано с тем, что в соответствие ставятся не сами объекты, а их свойства (рис. 3).

Одним из примеров реализации описанной модели является использование “портальных технологий”, обеспечивающих единую точку входа, знакомый для пользователей и удобный в работе веб-интерфейс (рис. 4).

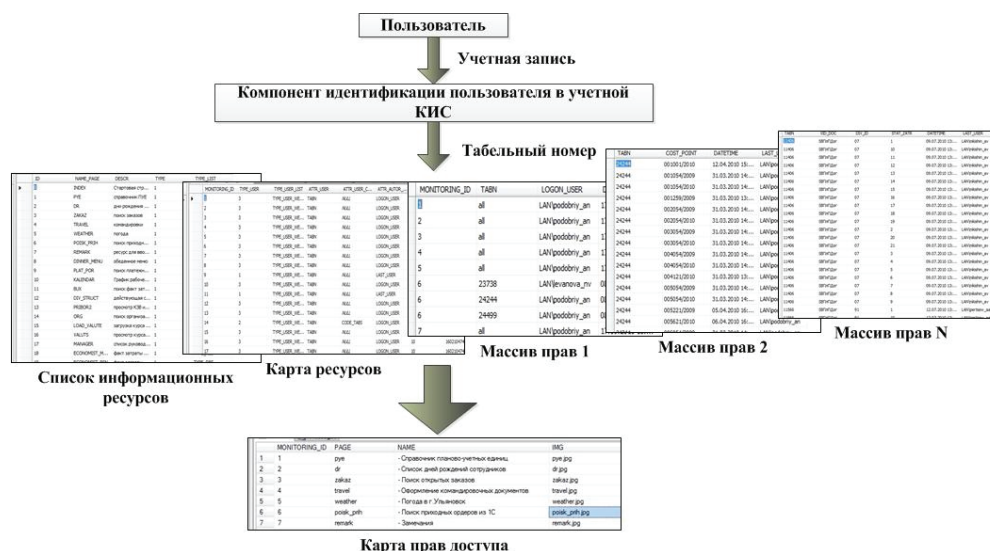


Рис. 3. Карта прав доступа пользователя к информационным ресурсам

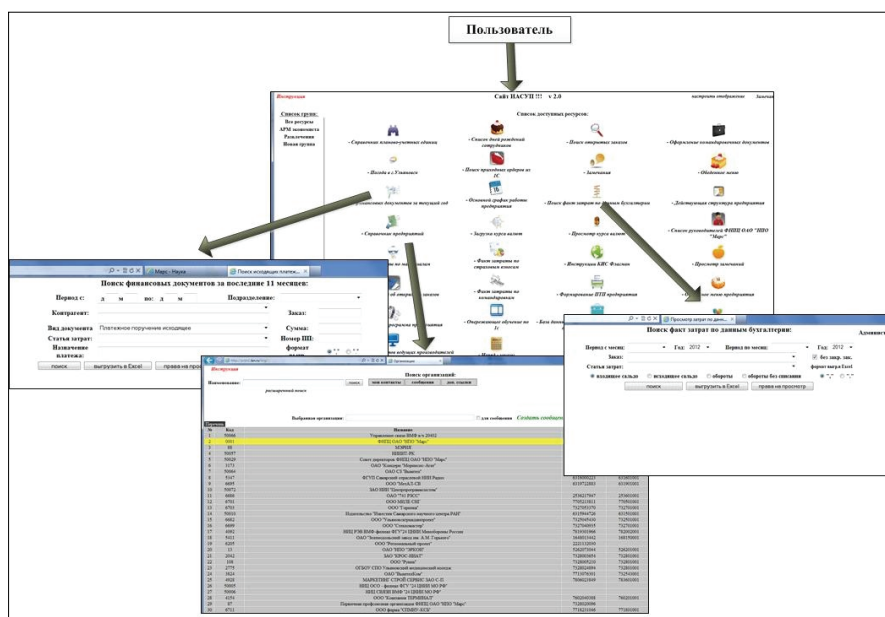


Рис. 4. Единая точка входа пользователей

ЗАКЛЮЧЕНИЕ

В статье представлена математическая модель интегрированной информационной системы основанной на интеграции данных из разных информационных систем. Отталкиваясь от данной системы, построена модель сопоставления данных, используя модифицированный метод n-грамм. Выведена формула среднего коэффициента релевантности сравнения двух наименований и предложено решение для определения однозначного соответствия двух строк.

В качестве примера приведены сравнения двух наименований и вычисления вероятностей их вхождения.

Приведена модель построения единого пользовательского интерфейса на основании матрицы прав доступа сотрудников.

СПИСОК ЛИТЕРАТУРЫ

1. Levy A.Y. Logic-Based Techniques in Data Integration. Logic-based Techniques in Data Integration. In: Logic Based Artificial Intelligence. Edited by J. Minker. Kluwer Publishers, 2000.
2. Manolescu I., Florescu D., Kossman D. Answering XML Queries over Heterogeneous Data Sources. Proc. Of the 27th VLDB Conference, Roma, Italy, 2001.
3. Петров В.Н. Информационные системы. СПб.: Питер, 2002. 430 с.
4. Сирли Э. Корпоративные хранилища данных. Планирование, разработка и реализация. Т.1. М.: Вильямс, 2001. 230 с.
5. Ахатов А.Р. Алгоритмы программной системы контроля текстовой информации на основе n-граммной языковой модели // Актуальные проблемы современной науки. 2009. № 3. С. 156-161.
6. Герасименко В.А., Малюк А.А. Основы защиты информации. М.: Инкомбук, 1997.
7. Подобрий А.Н. Информационная модель безопасности веб-портала на предприятиях. Инноватика-2011: Материалы международной конференции. Махачкала (март 2011).
8. Подобрий А.Н. Модель доступа к веб-порталу на современном предприятии // Известия Самарского научного центра РАН. 2011. Т.13. № 4(2). С. 475-478.

DATA INTEGRATION MODEL IN A COMMON INFORMATION SPACE COMPANIES USING METHOD N-GRAM

© 2012 A.N. Podobriy

Federal Research-and-Production Center
'Research-and-Production Association 'Mars', Ulyanovsk

In the article considers an approach to the implementation of mapping data with the modified method of n-gram and model of an integrated information system, and access to associated data within a corporate network information. Article is intended for professionals involved in the introduction of automated business management systems, as well as for all professionals involved in data integration.

Key words: data integration, n-grams, access matrix, enterprise, integration model