I. S. Pechenko, Research Associate, e-mail: ivan.pechenko@intel.com,
O. V. Venger, Senior Researcher, e-mail: oleg.v.venger@intel.com, ZAO "Intel"

Generation of Flow Diagram from Textual Description of the Flow

Nowadays computer systems design is a very complex process, and the specification of components interaction is a serious part of the design process. The quality of the specification is the important factor for system design and test efficiency. Commonly used model of system specification is a set of scenarios (or flows) of components interaction. There is no common standard for flow description, and now 95 % of such specifications are textual scenarios descriptions. They are commonly inaccurate, inconsistent, incomplete and cannot be automatically processed. Here in this work we propose a method of converting textual flow specifications to a form with better readability, usable for automatic processing and more convenient for data sharing between project stakeholders. Special tags are added to the textual flow description to structure it, and then flow diagram is automatically generated from the text using BPMN notation. This method was implemented and successfully used for improving complex systems specifications.

Keywords: *behavior model, behavior scenario, structured text, automatic specification processing, flow diagram, BPMN*

References

1. **Mich L., Franch M., Inverardi P. N.** Market Research for Requirements Analysis Using Linguistic Tools. *Requirement Engineering Journal*. 2004. N. 9 (1). P. 40—56.

2. **Kiyavitskaya N., Zeni N., Mich L., Berry D. M.** Requirements for Tools for Ambiguity Identification and Measurement in Natural Language Requirements Specifications. *Requirement Engineering Journal*. 2008. N. 13. P. 207—240.

3. **Hussain I., Ormandjieva O., Kosseim L.** Automatic Quality Assessment of SRS Text by Means of a Decision-Tree-Based Text Classifier. *In Proceedings of the 7th International Conference on Quality Software*. 2007. P. 209—218.

4. **Tjong S. F., Hallam N., Hartley M.** Improving the Quality of Natural Language Requirements Specifications through Natural Language Requirements Patterns. *In Proceedings of the 6th IEEE International Conference on Computer and Information Technology*. 2006.

5. **Kamalrudin M., Hosking J., Grundy J.** Improving Requirements Quality Using Essential Use Case Interaction Patterns. *In Proceedings of the 33rd International Conference on Sofware Engineering*. 2011.

6. **Denger C., Berry D. M., Kamsties E.** Higher Quality Requirements Specifications through Natural Language Patterns. In *Proceedings of the IEEE International Conference on Sofware — Science, Technology and Engineering — SwSTE'03*. 2003.

7. **Holtmann J., Meyer J., von Detten M.** Automatic Validation and Correction of Formalized, Textual Requirements. *In Proceedings of the IEEE 4th International Conference on Sofware Testing, Verification and Validation Workshops — ICSTW*. 2011.

8. **Gelhausen T., Tichy W. F.** Thematic Role Based Generation of UML Models from Real Word Requirements. *In Proceedings of the 1st IEEE International Conference on Semantic Computing — ICSC*. 2007. P. 282—289.

9. **Sharma V. S., Sarkar S., Verma K., Panayappan A., Kass A.** Extracting High-level Functional Design from Software Requirements. *In Proceedings of the 16th IEEE Asia-Pacific Software Engineering Conference — APSEC*. 2009. P. 35—42.

10. **Deeptimahanti D. K., Sanyal R.** An Innovative Approach for Generating Static UML Models from Natural Language Requirements. *Advances in Software Engineering. Communications in Computer and Information Science*. 2009. V. 30. P. 147—163.

11. **URL:** http://www.bpmn.org/ Дата обращения: 01.01.2014.

12. **URL:** http://www.omg.org/ Дата обращения: 01.01.2014.

13. **Afreen H., Bajwa I. S.** Generating UML Class Models from SBVR Software Requirements Specifications. In 23rd Benelux Conference on Artifical Intelligence — BNAIC. 2011. P. 23—32.

14. **URL:** http://www.omg.org/spec/SBVR/ Data obrasheniya: 01.01.2014.

15. **URL:** http://www.omg.org/spec/ Data obrasheniya: 25.04.2014.

УДК 004.912

М. В. Бочков, д-р техн. наук, проф.,
НОУ ДПО "Центр предпринимательских рисков", г. Санкт-Петербург,
П. Н. Бойков, вед. специалист,
ОАО "НИИ "Рубин", г. Санкт-Петербург, e-mail: boykova@yandex.ru

Инфодинамическая модель поиска пользователя в социальной сети

Исследованы закономерности представления пользователями соцсетей своих данных, на основе которых разработан алгоритм формирования оптимальной стратегии поиска в социальных сетях.

Ключевые слова: *социальная сеть, поиск в социальной сети, алгоритм оптимизации поиска в социальных сетях, инфодинамическая модель*

Социальная сеть как объект исследования

Среди ресурсов в сети Интернет все большую популярность приобретают онлайновые социальные сети (ОСС). К типовым возможностям их участников можно отнести:

- обмен информационными ресурсами с другими участниками ОСС;
- публикация и обсуждение идей;
- выбор и участие в социальных группах (сообществах);
- использование развлекательных и досуговых сервисов ОСС и др.

Очевидной тенденцией в развитии ОСС является рост числа пользователей и развитие их функциональных сервисов [1]. Динамика изменения числа пользователей наиболее популярных ОСС представлена на рис. 1 (см. вторую сторону обложки).

Информационную основу ОСС образуют персональные пользовательские страницы. Как правило, создатели ОСС стремятся получить от пользователя максимум информации. С этой целью регистрационная форма предлагает опубликовать максимум идентификационной и другой персональной информации. На рис. 2 (см. вторую сторону обложки) показаны набор регистрационных данных в ОСС "Вконтакте".

Очевидно, что наиболее полное представление пользователями своих данных повышает точность и полноту результатов запроса и, следовательно, однозначность идентификации участников ОСС. Вместе с тем, среднестатистический пользователь подсознательно стремится представить минимум информации о себе, ограничить круг своего общения, обеспечив себе комфортное общение в ОСС. Таким образом, проявляется конфликт интересов владельцев и пользователей ОСС — одни хотят знать все, а другие хотят обойтись минимумом информации о себе [2].

Постановка задачи. Целью настоящего исследования является выявление закономерностей представления идентифицирующей пользователя информации в ОСС. Знание таких закономерностей и их описание в виде формальной модели позволит сформировать оптимальную стратегию поиска, при которой вероятность точного нахождения требуемого пользователя ОСС за минимальное число итераций поиска будет максимальна.

Модель представления информации пользователями ОСС

Исходные данные. В настоящем исследовании использован модельный фрагмент ООС, сформированный путем обезличивания репрезентативного дампа общедоступных в сети Интернет пользовательских страниц. На основе полученной информации проведен расчет статистик, характеризующих атрибуты регистрационных данных пользователей [3].

Ранжирование пользовательских атрибутов и интерпретация полученных результатов. Для последующих исследований были выделены следующие пользовательские атрибуты:

- идентификатор пользователя в ОСС (a_0);
- фамилия (a_1);
- имя (a_2);
- город проживания (a_3);
- пол пользователя (a_4);
- дата рождения (a_5);
- наименование и год окончания вуза (a_6);
- наименование и год окончания школы (a_7);
- место работы (a_8);
- семейное положение (a_9).

Для исследований были введены следующие упрощения:

- параметр a_0 не использован при анализе, так как является уникальным для каждого пользователя и однозначно идентифицирует пользователя в социальной сети;
- параметры a_4 и a_9 исключены из анализа ввиду малого диапазона принимаемых значений, незначительного влияния на результаты поиска;
- аналогичные значения атрибута a_2 "Иван", "Ваня", "Ванька" считаются эквивалентными поисковому запросу "Иван";
- для атрибута a_3 значения, подобные "СПб", "Санкт-Петербург", "Питер", считались эквивалентными поисковому запросу "Санкт-Петербург".

На рис. 3 (см. вторую сторону обложки) представлены ранжированные значения вероятностей присутствия пользовательских атрибутов, отражающих закономерности отображения информации в ОСС. В дальнейшем данная закономерность используется в качестве модели представления информации пользователями ОСС.

Алгоритм формирования максимальной стратегии

Использование модели представления в свою очередь позволяет решить задачу формирования оптимального алгоритма поиска пользователя на основе математического аппарата инфодинамического моделирования [4], который основывается на анализе распределений значений результатов поиска от используемых атрибутов. Оригинальность предлагаемого подхода состоит в том, что основные формальные соотношения получены в дифференциальной форме (т. е. позволяют оценить влияние конкретных значений атрибутов на конкретные результаты поиска), а средние оценки (энтропия, условная и взаимная информация) представляют усреднение дифференциальных оценок. Такой подход позволяет более детально проанализировать информационные связи в системе отношений между информационными атрибутами и результатами поиска.

Алгоритм формирования максимальной стратегии на основе математического аппарата инфодинамического моделирования состоит из трех этапов.

1. Вначале определяем ряд информационных оценок (собственная информация, условная информация и взаимная информация) поисковой функции и информационных атрибутов [4].

Собственной информацией $I(a_j)$ (или количеством собственной информации) значения информационного атрибута a_j переменной ϑ_j поисковой функции f называется величина

$$I(a_j) = -\log p(a_j),$$

где $p(a_j)$ — вероятность того, что информационный атрибут указан пользователем на странице ОСС.

Условной информацией $I(\beta|a_j)$ в значении β переменной ϑ_j поисковой функции f при заданном значении информационного атрибута a_j называется величина

$$I(\beta|a_j) = I(\beta a_j) - I(a_j).$$

Взаимная информация между значениями β и a_j есть информация в значении информационного атрибута a_j о значении β поисковой функции f (и наоборот, информация в значении β о значении a_j) вычисляется следующим образом:

$$I(\beta;\, a_j) = I(a_j;\, \beta) = I(\beta) - I(\beta|a_j) = I(a_j) - I(a_j|\beta).$$

Взаимная информация между значениями информационного атрибута a_f и поисковой функцией f используется для оценки взаимного влияния атрибута и функции, т. е. данная информационная оценка указывает степень информационной связи присутствия атрибута с конкретным значением поисковой функции.

2. Проведем исследование пользовательских атрибутов для определения влияния атрибутов на результаты поиска с учетом их встречаемости на пользовательских страницах в ОСС. Для этого на основе выборки пользователей, полученной из ОСС, выполним поиск случайных пользователей, изменяя атрибуты поиска и сравнивая полученные результаты. Поиск будем осуществлять на выборке случайных пользователей, на страницах которых содержатся все информационные атрибуты.

Рассчитаем вероятность того, что по заданным атрибутам результаты поиска будут успешными, т. е. пользователь будет найден:

$$P(f(\beta = 1)) = 1 - \frac{n}{N},$$

где n — результат поиска для a_j; N — общее число пользователей.

Результаты расчетов значений условной энтропии поисковой функции и взаимной информации между поисковой функцией и значением каждого атрибута a_j показаны в табл. 1.

Как видно из результатов, приведенных в табл. 1, использование одного пользовательского атрибута в поисковом запросе, даже с учетом того, что пользователь указал на своей странице все атрибуты, не всегда позволяет найти пользователя в ОСС. Таким образом, целесообразно исследовать значения поисковой функции при использовании в поисковом запросе попарно двух атрибутов между собой. При этом в дальнейших расчетах будем использовать среднее значение вероятностей успешного результата поиска, значение которых получены как среднее арифметическое вероятностей для каждого атрибута выбранных пользователей (табл. 2).

В табл. 3, приведены следующие значения: U — номер теста, $P(U_s)$ — вероятность совместного

Таблица 1

Id_user	a_1	a_2	a_3	a_5	a_6	a_7	a_8
36	0,9802	0,9037	0,7861	0,9992	0,9791	0,9996	0,9999
2482	0,9523	0,7504	0,7861	0,9993	0,9842	0,9997	1
227385	0,9792	0,7719	0,9993	0,9998	0,9999	1	1
125812	0,9965	0,9958	0,9937	0,9992	1	0,9998	1
178272	0,9988	0,8178	0,6201	0,9991	0,9977	0,9997	0,9999
93388	1	0,9848	0,9995	0,9994	0,9996	0,9999	0,9998

Таблица 2

Id_user	a_1	a_2	a_3	a_5	a_6	a_7	a_8
Среднее значение	0,9845	0,8707	0,8641	0,9993	0,9934	0,9997	0,9999

Таблица 3

U	$P(U_3)$	$a_i a_j$	f
u_1	0,02	0 0 0 0 0 1 1	1
u_2	0,03	0 0 0 1 0 0 1	1
u_3	0,06	0 0 0 1 0 1 0	1
u_4	0,09	0 0 0 0 1 0 1	1
u_5	0,2	0 0 0 0 1 1 0	1
u_6	0,24	0 0 0 1 1 0 0	1
u_7	0,11	1 0 0 0 0 0 1	1
u_8	0,24	1 0 0 0 0 1 0	1
u_9	0,28	1 0 0 1 0 0 0	1
u_{10}	0,86	1 0 0 0 1 0 0	0
u_{11}	0,11	0 1 0 0 0 0 1	1
u_{12}	0,24	0 1 0 0 0 1 0	0
u_{13}	0,86	0 1 0 0 1 0 0	0
u_{14}	0,03	0 0 1 0 0 0 1	0
u_{15}	0,06	0 0 1 0 0 1 0	0
u_{16}	0,28	0 1 0 1 0 0 0	0
u_{17}	0,08	0 0 1 1 0 0 0	0
u_{18}	0,23	0 0 1 0 1 0 0	0
u_{19}	1	1 1 0 0 0 0 0	0
u_{20}	0,27	1 0 1 0 0 0 0	0
u_{21}	0,27	0 1 1 0 0 0 0	0

присутствия атрибутов $a_i a_j$ на пользовательской странице (данное значение рассчитано для всей полученной выборки) и f — значение решающей функции (результат поиска).

Исходя из результатов, приведенных в табл. 3, вероятности значений решающей функции $p(\beta)$ (где β — значение поисковой функции: 0 — результат поиска неудачен, 1 — поиск успешен), присутствия атрибутов a_j и их комбинаций приведены в табл. 4.

Следующим шагом определяем средние оценки (энтропия, условная и взаимная информация), характеризующие ансамбли значений поисковой функции от значений атрибутов поиска. Для отражения "привлекательности" поискового атрибута выбраны показатели условной энтропии и взаимной энтропии.

Условная энтропия $H(f|\alpha_j)$ функции поиска f при заданном значении атрибута a_j

$$H(f|a_j) = \sum_{\beta=0}^{1} p(\beta a_j) I(\beta|a_j).$$

Взаимная информация $I(f; a_j)$ между поисковой функцией f и атрибутом a_j определяется выражением

$$I(f; a_j) = \sum_{\beta=0}^{1} p(\beta a_j) I(\beta; a_j).$$

Вычисление информационных оценок (для табл. 1) и результаты выбора переменных при построении дерева решений приведены в табл. 5.

3. Дадим содержательную интерпретацию информационных оценок и полученных соотношений с позиции анализа информационных атрибутов процедуры поиска.

Для выработки логически обоснованных критериев выбора состава и порядка атрибута при разработке стратегии поиска необходимо более детально проанализировать содержательную сторону приведенных формальных оценок. Для этого представим связь атрибута и поисковой функции следующим образом:

$$H(f) - H(f|a_j) = I(f; a_j) = H(a_j) - H(a_j|f).$$

Рассмотрим содержательно слагаемые этого выражения.

Энтропия $H(f)$ поисковой функции — среднее количество информации, которое необходимо извлечь для определения значения функции.

Энтропия $H(a_j)$ атрибута — среднее количество информации, которое извлекается при добавлении атрибута поиска.

Взаимная информация $I(f; a_j)$ — среднее количество информации, которое несет атрибут поиска о результатах поиска.

Для решения поставленной задачи интерес представляет взаимная информация как индикатор того, насколько уменьшится диапазон результатов поиска

Таблица 4

Атрибут	Вероятности	Значения a_j, β	Взаимная вероятность $p(\beta a_j)$	Комбинация значений
		a_j 0 1 β 1 1		a_j 0 1 β 1 1
a_1	$p(a_j)$ $p(\beta)$	0,3.....0,7 0,8.....0,2	$p(\beta a_1)$	0,3 0,2
a_2	$p(a_j)$ $p(\beta)$	0,3.....0,7 0,1.....0,9	$p(\beta a_2)$	0,1 0,7
a_3	$p(a_j)$ $p(\beta)$	0,3.....0,7 1.....0	$p(\beta a_3)$	0,3 0
a_5	$p(a_j)$ $p(\beta)$	0,3.....0,7 0,7.....0,3	$p(\beta a_5)$	0,3 0,3
a_6	$p(a_j)$ $p(\beta)$	0,3.....0,7 0,8.....0,2	$p(\beta a_6)$	0,3 0,2
a_7	$p(a_j)$ $p(\beta)$	0,3.....0,7 0,7.....0,3	$p(\beta a_7)$	0,3 0,3
a_8	$p(a_j)$ $p(\beta)$	0,3.....0,7 0,9.....0,1	$p(\beta a_8)$	0,3 0,1

Таблица 5

Атрибут	a_1	a_2	a_5	a_6	a_7	a_8
$H(f\|a_j)$	0,358	0,158	0,363	0,358	0,396	0,279
$I(f;a_j)$	0,802	0,698	0,681	0,802	0, 936	1,049

при наличии того или иного атрибута. Другими словами, из информации $H(a_j)$ оценивается та ее часть, которая позволяет уменьшить энтропию $H(f)$ функции до значения $H(f|a_j)$.

Под стоимостью решения будем понимать время отклика на выполнение поискового запроса и примем его одинаковым для каждого атрибута поиска. Таким образом, критерием оптимизации при выборе следующего теста будет выступать выражение

$$a_j^* = \max_{a_j \in A} I(f, a_j),$$

где a_j^* — следующий атрибут для добавления в запрос.

В общем виде задача формирования оптимальной стратегии поиска в терминах инфодинамического моделирования соответствует задаче конструирования деревьев решений. Процедуру конструирования дерева решений представим в виде алгоритма.

Шаг 1. Задача состоит в выборе атрибута, который целесообразно использовать первым. Результаты вычислений (табл. 6) показывают, что критерию

Таблица 6

Уровень дерева решений	Условие (известные значения переменных)	Атрибуты, из которых осуществляется выбор	Взаимная информация $I(f;a_j)$	Выбор
1	—	a_1 a_2 a_5 a_6 a_7 a_8	0,802 0,698 0,681 0,802 0,936 **1,049**	a_8
2	$a_8 = 1$	a_1 a_2 a_5 a_6 a_7	0,802 0,698 0,681 0,802 **0,936**	a_7
3	$a_7 = 1$	a_1 a_2 a_5 a_6	**0,802** 0,698 0,681 0,802	a_1
4	$a_1 = 1$	a_2 a_5 a_6	0,698 0,681 **0,802**	a_6
5	$a_6 = 1$	a_2 a_5	**0,698** 0,681	a_2

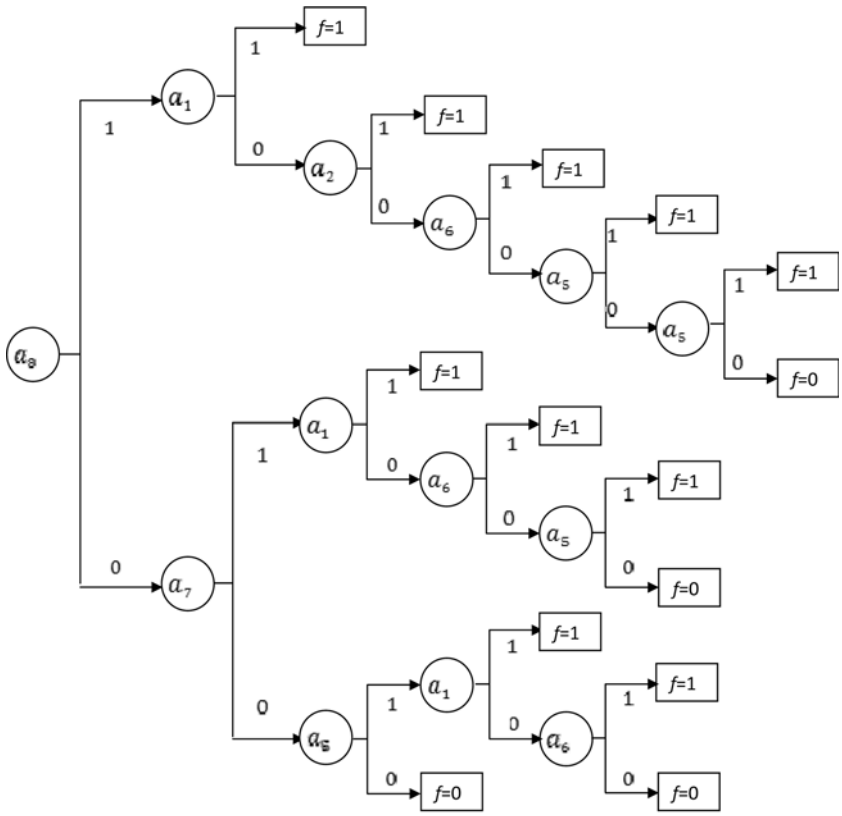

Рис. 4. Дерево решений на основе информационных оценок

оптимизации удовлетворяет атрибут a_8. Построим первый уровень дерева решений при условии, что корневому узлу дерева соответствие этого атрибута не позволит идентифицировать пользователя, поэтому на основе вероятности совместной встречаемости атрибутов целесообразно в качестве второго атрибута использовать a_1. Если атрибут a_1 отсутствует, то в качестве второго атрибута выбирается атрибут, следующий по значению совместной вероятности встречаемости атрибутов с a_8.

Шаг 2. Если атрибут a_8 отсутствует, то выбирается следующий по значимости атрибут, а алгоритм добавления второго атрибута в параметры запроса аналогичен шагу 1.

Шаг 3. На этом шаге выбирается переменная $a_8 = a_7 = 0$. Из таблицы решений видно, что поисковая функция f принимает максимальное значение только в случае, если будут известны следующие пары атрибутов: a_1a_5 и a_5a_6. Таким образом, целесообразно проверить общий для этих пар атрибут a_5. Если он отсутствует, то поиск целесообразно прекратить, так как оставшихся атрибутов недостаточно для идентификации пользователя в социальной сети.

Дерево решений как результат оптимизации представлено на рис. 4.

Экспериментальные исследования полученного алгоритма поиска пользователя социальной сети

На заключительном этапе исследовалась практическая пригодность полученного алгоритма для поиска пользователя в социальной сети, в частности, определялось, возможно ли конструирование множества деревьев решений за приемлемое время. Затем оценивалось преимущество предложенного алгоритма по сравнению с алгоритмом последовательного добавления атрибутов поиска в поисковый запрос.

Для получения априорных сведений о пользователях социальной сети использовались три наиболее популярные социальные сети "ВКонтакте", "Фейсбук" и "Одноклассники". Выборка осуществлялась путем случайного отображения информации о пользователе социальной сети (такая функциональная возможность предоставляется ОСС). В результате эксперимент проводился на общей выборке из 4500 случайно выбранных пользователей социальных сетей.

В ходе эксперимента были получены следующие результаты. Во всех трех социальных сетях алгоритм поиска пользователя сконструировал дере-

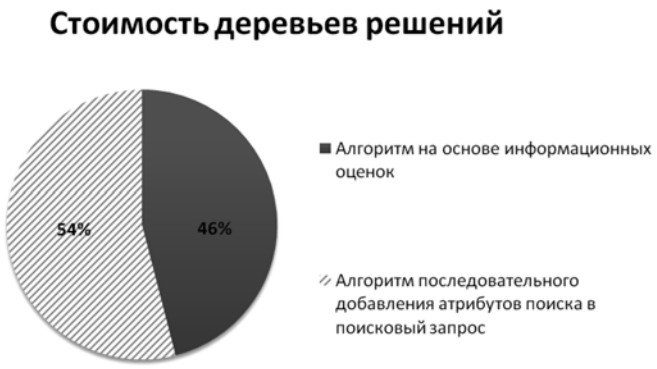

Рис. 5. Сравнение вычислительных затрат двух алгоритмов для успешного поиска пользователя в социальной сети

вья меньшей стоимостью, чем алгоритм последовательного добавления атрибутов поиска в поисковый запрос. В среднем стоимость деревьев решений на 8 % меньше (рис. 5). Также снизились и вычислительные затраты; так, в среднем для успешного поиска пользователя в социальных сетях необходимо использовать три атрибута, а не четыре.

Выводы

- Предложенный алгоритм конструирования дерева решений на основе информационных оценок пользовательских атрибутов позволяет оптимизировать время поиска пользователей в ОСС при достаточно большом числе атрибутов поиска.
- Использование разработанного алгоритма позволяет сократить вычислительные затраты на обработку сервером социальной сети поискового запроса, тем самым давая возможность уменьшить время, затраченное аналитиком на поиск путем последовательного добавления пользовательских атрибутов в поисковый запрос.
- Полученный подход применим к любым базам данных, содержащим большое число объектов учета с множеством идентифицирующих и характеризующих объекты атрибутов, что позволяет говорить о формировании универсальной стратегии поисковой оптимизации.

Список литературы

1. **Губанов Д. А., Новиков Д. А., Чхартишвили А. Г.** Социальные сети: модели информационного влияния, управления и противоборства. МЦНМО, 2010.

2. **Бочков М. В., Бойков П. Н., Яшин А. А.** Социальные сети как основной источник утечки персональных данных // Inside # Защита информации. 2010. № 3.

3. **Бочков М. В., Бойков П. Н.** Способ автоматического рубрицирования неструктурированной информации в сети Интернет // Информационные технологии. 2012. № 2.

4. **Курбацкий А. Н., Чеушев В. А.** Информационный метод анализа и оптимизации в системах поддержки принятия решений. Минск: Ин-т техн. кибернетики НАН Беларуси, 1999.

M. V. Bochkov, Professor, UKC "The Center of enterprise risks", SPb,
P. N. Boykov, Leading Specialist,
Public Corporation of Scientific Research Institute "Rubin", SPbf,
e-mail: boykovpn@yandex.ru

Infodinamichesky Model of Search of the User on a Social Network

Among resources in the Internet online social networks (OSN) are becoming more popular. Information basis OSN is formed personal user pages. Obviously, the most complete of the user's picture of their data increases the accuracy and completeness of query results, and is therefore uniquely of participants' identify OSN. On the other hand, the average user's subconsciously prefer to present a minimum information about themselves, limit their circles of contacts, securing a comfortable communication in OSN. The current article explores the patters of personal data presentation by social media users based on which an algorithm of an optimized social networks search strategy has been developed.

Keywords: *social network, search in social network, the algoritm of an optimized social networks search strategy, infodinamichesky model*

References

1. **Gubanov D. A., Novikov D. A., Chxartishvili A. G.** *Socialnir seti: mo informacionnogo vliyaniya, upravleniya i protivoborstva*. M.: MCNMO, 2010.

2. **Bochkov M. V., Boykov P. N., Yashin A. A.** Socialnie seti kak osnovnoy istochnik utechki personalnich dannich. *Inside. Zashita informacii*, Sant-Petersburg. 2010. N. 3.

3. **Bochkov M. V., Boykov P. N.** Sposob avtomaticheskogo rubricirovaniya nestrukturirovannoy informacii v seti Internet. *Informacionnie technologii.* 2012. N. 2.

4. **Kurbackiy A. N., Cheushev V. A.** *Informacionniy metod analiza i optimizacii v sistemach podderzhki prinyatiya resheniy*. Minsk: Institut tech. kibemetiki NAN Belarus, 1999.