

УДК 004.55, 004.91, 005

**КЛАССИФИКАЦИЯ И СВОЙСТВА ОПЕРАЦИЙ НАД ОБЪЕКТАМИ
ИНФОРМАЦИОННОГО ПРОСТРАНСТВА СРЕДЫ ОБЛАСТИ ЗНАНИЙ¹***Костенко К. И.², Синица С. Г.³***CLASSIFICATION AND PROPERTIES OF OPERATIONS OVER THE OBJECTS OF
INFORMATION SPACE FOR THE SUBJECT AREA DOMAIN**

Kostenko K. I., Sinitsa S. G.

Formalization of shared information space and classification for semantic operations as well as properties based on the results of Web-systems analysis are developed for subject domains information systems. Ways for application of these properties are proposed for realization of such information systems.

Введение

Информационной средой области знаний (СОЗ) называется открытый информационный ресурс, обеспечивающий возможность единого и полного структурно-семантического представления массива предметных и профессиональных знаний и методов их обработки, обеспечивающий обслуживание потребностей образовательной и профессиональной деятельности в предметных и профессиональных знаниях [1]. В работе рассмотрена модель СОЗ, для которой предполагается, что извлекаемые из первоисточников знания структурируются сообществом предметных пользователей в коллективном информационном пространстве (КИП) [2], организация которого отражает существующие представления о структуре системы знаний предметной области. Системы, попадающие под определение СОЗ, находят основное применение в Интернет-сообществах практики и в составе интранет-сетей крупных предприятий, содержащих рассредоточенные группы специалистов, занимающихся совместной исследовательской деятельностью и интеллектуальным трудом. Важной для успешного внедрения и эффективного использования характеристикой СОЗ является продуктивность рабо-

ты пользователей, определяемая средним количеством времени, затрачиваемым на решение комплексных задач. Однородное и формальное описание основных операций процессов формирования коллективной модели ПО и использования КИП позволяет обосновывать состав системы вспомогательных механизмов конкретной СОЗ на этапе ее проектирования, оценить влияние этих механизмов на продуктивность.

**1. Формальная модель
информационного пространства**

Информационное пространство (ИП) СОЗ — это четверка

$$W = (T, M, F, N),$$

где $T = (t_1, t_2, \dots, t_{|T|})$ — последовательность идентификаторов допустимых в ИП типов информационных объектов (ИО);

$M = (m_1, m_2, \dots, m_{|T|})$ — последовательность описаний метаданных ИО соответствующих типов, $m_i = XSD(t_i)$ — XML-схема⁴, описывающая структуру и фиксирующая семантику полей метаданных ИО типа t_i с помощью общепринятых пространств имен DC⁵, OWL⁶;

¹Работа выполнена при поддержке РФФИ (06-07-96618).

²Костенко Константин Иванович, канд. физ.-мат. наук, доцент кафедры информационных технологий, начальник отдела разработки информационных систем Центра интернет Кубанского государственного университета.

³Синица Сергей Геннадьевич, аспирант кафедры высоких технологий прогноза и предупреждения чрезвычайных ситуаций Кубанского государственного университета.

⁴XML Schema. <http://www.w3.org/XML/Schema.html>.

⁵Dublin Core Metadata Initiative. <http://www.dublincore.org/>.

⁶Web Ontology Language (OWL). <http://web4.w3.org/2004/OWL/>.

$F = (f_1, f_2, \dots, f_{|T|})$ — последовательность наименований способов разметки (HTML, BBCode, Wiki-разметка и т. д.) содержимого ИО соответствующих типов;

$N = n_1, n_2, \dots, n_{|N|}$ — множество экземпляров ИО, создаваемых согласованно со структурами T, M, F .

Экземпляр информационного объекта — это четверка $n = (u, t, h, b)$, где $u = URL(n)$ — строка постоянного URL, по которому доступен n ;

t — фиксированный идентификатор типа объекта из T ;

h — XML-документ со значениями полей метаданных, отвечающий XML-схеме m_t ;

b — содержимое n в виде текста, размеченного способом f_t .

Коллективное информационное пространство — это ИП, дополненное XML-схемой описания полей профилей и наборами возможных ролей пользователей, разрешений доступа и изменения ИО всех типов из T для каждой роли и правил их назначения пользователям, профилей зарегистрированных пользователей с назначенными им ролями. Объектной операцией называется всякое действие пользователя, меняющее экземпляр ИО из N .

Семантической операцией назовём комплексную задачу пользователей по изменению КИП, для решения которой необходимо выполнение одной или более объектных операций, изменяющих содержимое b отдельных экземпляров ИО из N в соответствии с задачей и интерпретацией этого содержимого пользователями. Онтологической операцией назовём всякую операцию, изменяющую значения элементов T, M или F .

Далее рассматриваются только объектные и семантические операции.

Связи между объектами задаются значениями соответствующих полей метаданных ИО, предусмотренных XML-схемами M . Иерархические связи (являться — быть частью) между рубриками, узлами таксономий, онтологий и объектами, образующие пространственную навигацию, будем называть вертикальными или сильными. Контекстные гиперссылки, семантические связи между информационными объектами, фрагментами их содержимого, динамическую навигацию, генерируемую на основе учета поведения пользователей, а также элементы адаптивной навигации будем называть горизонтальными или слабыми [3, 4].

Данная формализация структуры КИП позволяет исследовать информационные пространства СОЗ, создаваемых с использова-

нием существующего программного обеспечения для поддержки деятельности сетевых сообществ, аналогичного известным системам Drupal и Wikimedia.

2. Характеристики информационных объектов

Характеристики ИО — это существенные для его использования параметры, которые могут быть извлечены из структур данных, применяемых для хранения и доступа к ИО конкретных СОЗ. Значения характеристик ИО зависят от метаданных и содержимого как самого объекта, так и связанных с ним других объектов КИП.

Определим систему характеристик ИО, объединяющую возможности наиболее распространенных структур данных, используемых для организации и поиска по массивам гипертекстовых документов:

— $URL(n)$ — строка постоянного URL, по которому доступен ИО n .

— Пути $Path(n)$ — множество возможных навигационных траекторий, состоящих из последовательности переходов по гиперссылкам с главной страницы к $URL(n)$. Фиксирует вертикальную навигацию.

— Запрет $Deprecated(n)$ — истинно тогда и только тогда, когда значения полей метаданных n установлены таким образом, что при просмотре ИО n он отображается как запрещенный для использования в пользу другого ИО либо по иной причине.

— Соседи $Adjoin(n)$ — множество объектов, расположенных на расстоянии одного перехода по гиперссылке при просмотре данного объекта. Фиксирует горизонтальную навигацию.

— Ссылочная популярность $Inbound(n)$ — множество ИО, при просмотре которых можно перейти на $URL(n)$ по прямой гиперссылке. Фиксирует горизонтальную навигацию и отражает важность n для пользователей СОЗ.

— Классы $Classes(n)$ — назначенное n подмножество множества рубрик или классов, используемых в среде для тематической организации ИО, определяемое пользователями вручную либо автоматически при создании ИО с помощью одного из известных алгоритмов классификации Web-страниц [5].

— Ключевые слова $Keywords(n)$ — набор наиболее важных слов и словосочетаний из содержимого b ИО n , задаваемых вручную (метки или теги) или выбираемых автоматически при изменении содержимого ИО, например, из его синтаксиса [6] либо из слов, для кото-

рых ИО n является документом «об этом слове» [7]. Передает смысловое содержание ИО.

– Мера похожести $Similar(n, n')$ — вещественное число от 0 до 1, передающее схожесть темы содержимого с ИО n и n' , вычисляемая, например, сопоставлением ключевых слов, тегов или классов [7]. Характеризует тематическую близость документов.

– Мера идентичности $Duplicate(n, n')$ — вещественное число от 0 до 1, передающее схожесть текстов содержимого b ИО n и n' , вычисляемая, например, сравнением выборок контрольных сумм текстов по алгоритму шинглов [8]. Характеристика уникальности содержимого, сопоставление позволяет оценивать степень изменений версий ИО и заимствования, обнаруживать нечеткие дубликаты.

– Видимость в результатах поиска $Search(n)$ — множество поисковых запросов, по которым внутренний поиск СОЗ выводит объект n .

– Фокус содержимого $Focus(n)$ — количество уникальных слов в содержимом n либо доля важных слов, встречающихся в содержимом N , относительно общего количества важных слов во всех текстах ИП, вычисляемая, например, как отношение количества разных слов инвертированного файла индекса поисковой системы [9] СОЗ в содержимом ИО n и общего количества разных слов в таком файле.

– Внимание сообщества $Views(n)$ — доля пользователей СОЗ, просмотревших или скачавших n с момента его добавления.

– Отклик сообщества $Comments(n)$ — доля прокомментировавших членов сообщества.

– Количество информации $Size(n)$ — числовая величина, мера информационной емкости содержимого объекта, например, количество слов в содержимом.

– Эффективность информации $Efficiency(n)$ — числовая величина, определяющая полезность ИО и его релевантность информационным потребностям пользователей СОЗ за определенный период времени. Значение этой величины подсчитывается с учетом внимания и отклика сообщества, средней длительности просмотра пользователями, отношения количеств показов на страницах результатов поиска и переходов с них, количества и эффективности, ссылающихся на n объектов СОЗ [10], результата коллективной фильтрации в СОЗ [11].

3. Базовая система объектных операций

Базовая система объектных операций — это набор объектных операций, достаточный для формирования пользователями произвольного множества экземпляров ИО N и структуры КИП, если заданы фиксированные структуры T , M и X . Базовую систему операций образует набор операций, позволяющий создавать и информационные объекты всех типов, изменять их содержимое в соответствии со способом его разметки и изменять метаданные в соответствии с их XML-схемой. Базовую систему образуют четыре операции: просмотреть, создать новый (с заданием непустого содержимого и метаданных), изменить содержимое, изменить метаданные. Операция удаления не используется при формировании КИП в среде с низким порогом вхождения и заменяется развитым механизмом управления версиями. В таблице далее приведены изменения определенных выше характеристик ИО для этих операций.

4. Целостность и эффективность информационного пространства

ОПРЕДЕЛЕНИЕ. ИП называется целостным, если для него выполняются следующие свойства:

- $\forall n \in N (Size(n) > 0)$;
- $\forall a, b \in N (Duplicate(a, b) \neq 1)$;
- $\forall n \in N (n \notin Adjoin(n))$;
- $\forall n \in N (\neg Deprecated(n) \rightarrow (Classes(n) \neq \emptyset \& Keywords(n) \neq \emptyset \& Path(n) \neq \emptyset))$;
- $Deprecated(n) \Leftrightarrow (Search(n) = \emptyset \& Path(n) = \emptyset)$;
- $\forall a, b \in N (Similar(a, b) = 1 \rightarrow Classes(a) \cap Classes(b) \neq \emptyset)$.

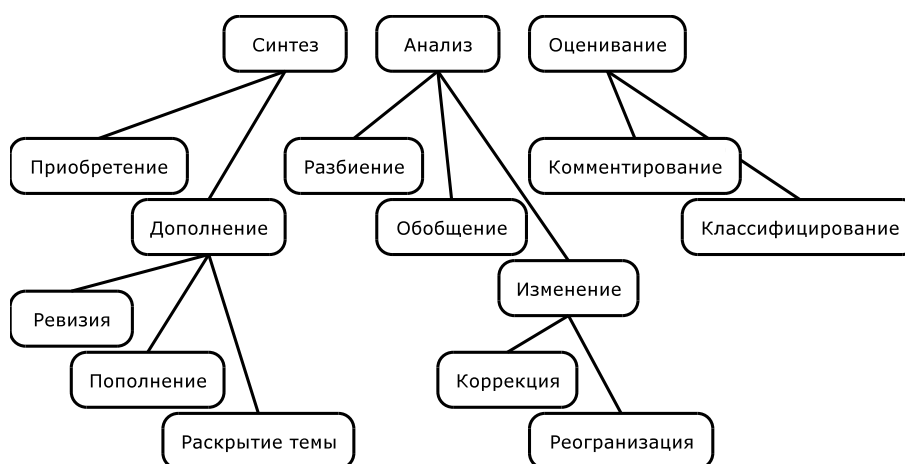
СОЗ с целостным ИП предоставляет удобную для использования навигационную структуру, корректный поиск, актуальные значения метаданных и содержимого ИО. Нарушение целостности ИП в результате выполнения операций пользователями запускает механизмы СОЗ, направленные на восстановление целостности сообществом пользователей и/или администраторами системы.

ОПРЕДЕЛЕНИЕ. ИП называется эффективным, если выполняется

$$\forall n \in N (\neg Deprecated(n) \rightarrow Efficiency(n) > 0).$$

Изменение характеристик информационных объектов базовыми операциями: $\geq$ — в результате операции величина или количество элементов множества характеристики некоторых ИО из КИП может увеличиться; $\leq$ — уменьшиться; $\neq$ — измениться; $=$ — для всех ИО не изменяется.

	<i>просмотреть</i>	<i>создать новый</i>	<i>изменить содержимое</i>	<i>изменить метаданные</i>
<i>URL</i>	$=$	$=$	$=$	$=$
<i>Path</i>	$=$	$=$	$=$	$\neq$
<i>Deprecated</i>	$=$	$=$	$=$	$\neq$
<i>Adjoin</i>	$=$	$=$	$\geq$	$\neq$
<i>Inbound</i>	$=$	$\geq$	$\neq$	$\neq$
<i>Classes</i>	$=$	$=$	$=$	$\neq$
<i>Keywords</i>	$=$	$=$	$=$	$\neq$
<i>Similar</i>	$=$	$\geq$	$\neq$	$\neq$
<i>Duplicate</i>	$=$	$\geq$	$\neq$	$=$
<i>Search</i>	$=$	$\geq$	$\neq$	$\neq$
<i>Focus</i>	$=$	$\leq$	$\neq$	$=$
<i>Views</i>	$\geq$	$=$	$=$	$=$
<i>Comments</i>	$=$	$\geq$	$=$	$=$
<i>Size</i>	$=$	$=$	$\neq$	$=$
<i>Efficiency</i>	$\neq$	$\geq$	$\neq$	$\neq$



Классификация семантических операций в коллективном информационном пространстве

5. Классификация и свойства семантических операций

На рисунке изображена схема классификации семантических операций.

Рассмотрим приведенные на рисунке классы операций и обозначим их свойства, используя соотношения указанных выше характеристик.

5.1. Операции синтеза

К таким операциям относятся те семантические операции, результатом которых является единственный новый объект или новая версия существующего объекта n , где содержимое b включает информацию, полученную

в результате извлечения знаний из эксперта или из внешних по отношению к СОЗ источников. Свойства операции приобретения связаны с требованиями непротиворечивости и максимально возможной интеграции приобретенных знаний в КИП.

Дополнение. Существующий объект n дополняется информацией из внешних источников, образуя новую версию объекта n' . Виды дополнения: ревизия, пополнение, раскрытие темы. Дополнение реализуется с помощью просмотра вертикальной навигации, поиска и базовых операций изменения содержимого и метаданных. Свойства этой операции задаются соотношениями:

$$\bullet \text{ URL}(n) = \text{URL}(n') \& \text{Path}(n) \subseteq \text{Path}(n');$$

- $\forall a \in N(Duplicate(a, n) \geq Duplicate(a, n'))$;
- $Classes(n) \subseteq Classes(n') \& Keywords(n) \subseteq Keywords(n')$;
- $n' \notin Adjoin(n')$;
- Необходимое условие ревизии:
 $Duplicate(n, n') \in (0, 1) \& Size(n') \in (0, Size(n)) \& Adjoin(n) \geq Adjoin(n')$;
- Достаточное условие пополнения:
 $Adjoin(n) \subseteq Adjoin(n') \& Size(n) < Size(n') \& Focus(n) = Focus(n')$;
- Достаточное условие раскрытия темы:
 $Keywords(n) \subseteq Keywords(n') \& Search(n) \subseteq Search(n') \& Size(n) < Size(n') \& Focus(n) < Focus(n')$.

Приобретение. Операция заключается в добавлении нового ИО n , содержимое которого целиком поступает из внешних по отношению к СОЗ источников в результате описания автором. Приобретение выполняется с использованием поиска сходных объектов, просмотра вертикальной навигации с целью определения подходящего раздела/класса для нового объекта и базовой операцией добавления нового ИО. Основные свойства этой операции:

- $Size(n) > 0$;
- $\forall a \in N(Duplicate(a, n) < 1)$;
- $n' \notin Adjoin(n')$;
- $Classes(n) \neq \emptyset \& Keywords(n) \neq \emptyset \& Path(n) \neq \emptyset$.

5.2. Операции анализа

Операции анализа реорганизуют уже существующую в СОЗ информацию с целью повышения эффективности ИП. Свойства операций анализа являются необходимыми условиями целесообразности изменения структуры ИП.

Разбиение. Объект n разбивается на два или более объектов $n_1, \dots, n_r$ и заменяется новой версией n' . Разбиение осуществляется с использованием базовых операций изменения содержимого и метаданных, создания новых ИО. Свойства разбиения:

- $\forall i \in [1..r](n_i \subset Adjoin(n') \& Path(n_i) \subset Path(n))$;
- $Deprecated(n')$;
- $\forall i, j \in [1..r](i \neq j \rightarrow Duplicate(n_i, n_j) = 0)$;
- $Adjoin(n_1) \cup \dots \cup Adjoin(n_r) = Adjoin(n)$;
- $\forall i \in [1..r](Classes(n_i) \subseteq Classes(n))$;
- $Search(n_1) \cup \dots \cup Search(n_r) = Search(n)$;
- $\sum_{i=1}^{|N|} Size(n_i) \geq Size(n)$.

Обобщение. На основе существующих объектов $n_1, \dots, n_r$ создается новый объект n , объ-

единяющий содержимое исходных. Обобщение осуществляется с использованием базовых операций изменения метаданных (исходные объекты помечаются как запрещенные в пользу объединения), создания нового ИО. Свойства обобщения включают:

- $\forall i \in [1..r] \exists k \in N(URL(k) = URL(n_i) \& k \subseteq Inbound(n) \& Deprecated(k))$ (на месте прежних ИО находятся запрещенные к использованию объекты, ссылающиеся на объединенный ИО);
- $Adjoin(n_1) \cup \dots \cup Adjoin(n_r) = Adjoin(n)$;
- $\forall i \in [1..r](Classes(n_i) \subseteq Classes(n))$;
- $Search(n_1) \cup \dots \cup Search(n_r) = Search(n)$;
- $\sum_{i=1}^{|N|} Size(n_i) \leq Size(n)$.

Изменение. Эта операция осуществляет замену объекта n его новой версией n' . При изменении используются базовые операции изменения содержимого и метаданных. Отдельными видами операции являются коррекция, реорганизация. Свойства операции изменения:

- $URL(n) = URL(n') \& Path(n) = Path(n')$;
- $\forall a \in Adjoin(n)(a \in Adjoin(n') \vee Deprecated(a))$;
- $\exists K = 1 + \Delta K(Size(n') = Size(n) * K)$;
- $\exists D = 1 - \Delta D(Duplicate(n, n') = D)$;
- Необходимое условие коррекции:
 $\Delta D = 0 \& \Delta K = 0$.
- Необходимое условие реорганизации:
 $\Delta D < 1 \& Focus(n) = Focus(n')$.

5.3. Операции оценивания

Операциями оценивания являются семантические операции, переносящие мнение пользователя об информационном объекте n в среду в виде значения поля метаданных n или отдельного объекта, связанного с n .

Комментирование. Пользователь создает связанный с n новый ИО n' соответствующего типа, в содержимом которого в форме текста на естественном языке он формулирует свое мнение об n , которое может включать, например, дополнительные факты, доказательство, опровержение, эмоциональную оценку. При комментировании используются базовые операции создания нового ИО и изменения метаданных. Характерные свойства этой операции:

- $Size(n') > 0 \& n' \notin Adjoin(n')$;
- $\forall a \in N(Duplicate(a, n') \neq 1)$;
- $Classes(n') = Classes(n) \& Keywords(n') \subseteq Keywords(n)$;

- $Deprecated(n') = Deprecated(n)$;
- $URL(n) = URL(n') \vee n' \subseteq Adjoin(n)$;

Классифицирование. Пользователь связывает объект n с одним из существующих классов k , либо создает и использует новый класс или добавляет к n метку (тег). При классифицировании используются базовые операции изменения метаданных и создания новых объектов. Свойство классифицирования:

- $\forall k (\exists a \in N (k \in Classes(a)) \rightarrow Classes(k) = \emptyset)$;
- $\forall k' (\forall a \in N (k \in Classes(a) \rightarrow k' \in Classes(a)) \rightarrow k' = k)$;
- $\forall a, b \in N (Similar(a, b) = 1 \rightarrow Classes(a) \cap Classes(b) \neq \emptyset)$;
- $\forall a, b \in N (Similar(a, b) = 1 \& Keywords(a) = Keywords(b) \& Focus(a) = Focus(b) \rightarrow Classes(a) = Classes(b))$.

ТЕОРЕМА. Операции дополнения, приобретения, разбиения, обобщения, изменения, комментирования и классифицирования сохраняют целостность ИП СОЗ.

Доказательство. Следует из определения целостности ИП и свойств семантических операций. $\square$

Заключение

Уточнение семантики операций с помощью подсчета указанных характеристик и проверки выделенных свойств позволяет создавать эффективные правила модерации схем работы пользователей, настраивать параметры ранжирования результатов поиска и параметры профилей пользователей для персонализации, с помощью динамической и адаптивной навигации улучшать целостность и связность КИП, отслеживать появление новой информации в КИП, увеличивать взаимодействие пользователей в СОЗ.

Литература

1. Костенко К.И., Левицкий Б.Е. Системный анализ технологии проектирования, сопровождения и использования информационных сред областей знаний // Изв. вузов. Сев.-Кавказ. регион. (Естеств. науки.) 2005. № 3. С. 11–16.
2. Bannon L., Bodker S. Constructing Common Information Spaces // ECSCW'97: Proceedings of the 5th European CSCW Conference. Dordrecht: Kluwer Academic Publishers. 1997. [<http://www.ul.ie/~idc/library/papersreports/LiamBannon/ECSCW.htm>].
3. Brusilovsky P. Methods and techniques of adaptive hypermedia // User Modeling and User Adapted Interaction. 1996. Vol. 6. P. 87–129.
4. Dourish P., Chalmers M. Running Out of Space: Models of Information Navigation // HCI. 1994. [<http://tinyurl.com/y6fapv>].
5. Максаков А.В. Оценка эффективности масштабируемых алгоритмов классификации текстов // Труды РОМИП'2006 [http://romip.narod.ru/romip2006/09_maksakov.pdf].
6. Тихомиров И.А. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus // Труды РОМИП'2006 [http://romip.narod.ru/romip2006/07_tikhomirov.pdf].
7. Гулин А., Маслов М., Сегалович И. Алгоритм текстового ранжирования Яндекса на РОМИП-2006 // Труды РОМИП'2006 [http://romip.narod.ru/romip2006/03_yandex.pdf].
8. Broder A.Z. Identifying and Filtering Near-Duplicate Documents // Combinatorial Pattern Matching: 11th Annual Symposium. 2000. ISSN 0302–9743 [<http://tinyurl.com/yzwcz7>].
9. Сегалович И.В. Как работают поисковые системы // Конференция «Диалог 2006» [<http://www.dialog-21.ru/trends/?id=15539&f=1>].
10. Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the Web // Stanford University. 1999 [<http://tinyurl.com/3hhqf>].
11. Billsus D., Pazzani M. J. Learning collaborative information filters // In Proceedings of the Fifteenth International Conference on Machine Learning. 1998. P. 46–54 [<http://tinyurl.com/y49onr>].