УДК 681.3

А.М. Наместников

ИНТЕНСИОНАЛЬНОЕ ПРЕДСТАВЛЕНИЕ КЛАСТЕРОВ ПРОЕКТНЫХ ДОКУМЕНТОВ НА ОСНОВЕ ПРИБЛИЖЕННЫХ МНОЖЕСТВ ПАВЛАКА[1]

Наместников Алексей Михайлович, *кандидат технических наук, доцент, окончил радиотехнический факультет Ульяновского государственного технического университета. Доцент кафедры «Информационные системы» факультета информационных систем и технологий УлГТУ. Имеет статьи и монографию в области интеллектуальных систем хранения и обработки информации. [e-mail: nam@ulstu.ru].*

Аннотация

В статье содержится формальное описание метода интенсионального представления кластеров проектных документов (ПД), сопровождающееся иллюстративным примером. Содержание интеллектуального репозитория есть множество концептуальных индексов, каждый из которых представляет собой отдельный ПД. Для адаптации методики формирования логических правил, которые позволяют определить кластеры, на основе приближенных множеств Павлака применяется понятие лингвистической переменной Л. Заде.

Ключевые слова: интеллектуальная система, кластеризация, лингвистическая переменная, приближенное множество.

Alexey Mikhailovich Namestnikov, *Candidate of Engineering, Associate Professor; graduated from the Faculty of Radio-Engineering of Ulyanovsk State Technical University; Associate Professor at the Chair 'Information Systems' of the Faculty of Information Systems and Technologies; author of articles and a monograph in the field of intelligent systems for storage and processing of information. e-mail: nam@ulstu.ru.*

Abstract

The article contains a formal description for the method of intensional cluster representation of design documents with examples. The content of an intellectual repository is a set of conceptual indexes. Each of them represents a separate design document. To adapt the procedure of the formation of logical rules which let define clusters, on basis of Pawlak rough sets, a notion 'Zadeh linguistic variable' is applied.

Key words: intellectual system, clustering, linguistic variable, rough set.

Введение

Результатом кластеризации некоторого множества объектов является набор кластеров (групп) анализируемых объектов со схожими характеристиками. Часто решается задача не просто группировки объектов, а их интенсионального (обобщенного) представления. В данном контексте интенсиональное описание противопоставляется экстенсиональному – когда информация о группе объектов формируется посредством перечисления всех объектов, принадлежащих этой группе. Недостатком такого описания является его неинтерпретируемость для последующего анализа и громоздкость.

Проблема построения обобщенного описания кластеров информационных ресурсов, в частности текстовых документов, обозначена достаточно давно. Существуют различные способы ее решения, среди которых можно выделить два основных подхода. Используя первый подход, делается попытка описать кластер информационных ресурсов на основе центрального представителя кластера. В качестве такого представителя может пониматься гипотетический (несуществующий) информационный ресурс – центр кластера (например, получаемый на выходе fcm-процедуры кластеризации). Второй подход основывается на формировании некоторого обобщенного описания кластеров. В данной статье предлагается реализация второго подхода содержательной интерпретации кластеров ПД с использованием теории приближенных множеств Павлака [1]. Кроме того, делается дополнительное предположение, что каждый ПД представляется в виде так называемого концептуального индекса, который формально имеет вид нечеткого дерева [2].

1 Содержание интеллектуального проектного репозитория

Кластер ПД можно понимать как неопределенное понятие, которое не может быть охарактеризовано в терминах информации об элементах [1]. В соответствии с подходом приближенных множеств Павлака описание кластера будем представлять в виде нижней и верхней аппроксимаций. Нижняя аппроксимация состоит из всех документов, которые точно соответствуют описанию

1 Работа выполнена при финансовой поддержке РФФИ, проект № 10-07-00064-а.

кластера, а верхняя – из всех документов, которые *возможно* соответствуют тому же описанию. Такое множество, состоящее из нижней и верхней аппроксимаций, называется приближенным множеством. Разница между верхней и нижней аппроксимациями составляет границу области неопределенности кластера.

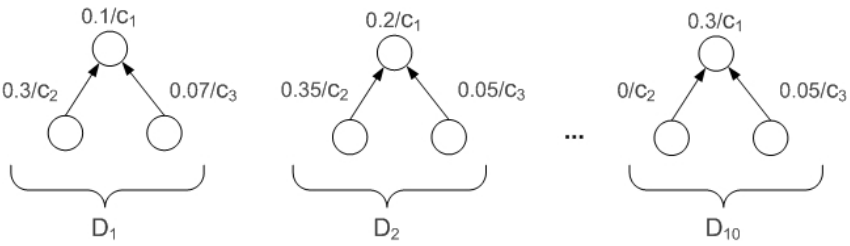

Рис. 1. Иллюстративный пример концептуальных индексов

Каждый ПД, как предполагается в работе [2], будем записывать в виде концептуального индекса – нечеткого вершинного подграфа дерева онтологии интеллектуального проектного репозитория (ИПР) [2, 3]. Концептуальный индекс ПД d формируется как нечеткий граф следующего вида:

$$cI^d = (\tilde{C}, E),$$

где $\tilde{C} = \{< \mu_{C_i} / C_i >\}$;

$$E = \{< C_i, C_k >\}, < C_i, C_k > \in C^2,$$

где C_i, C_k – понятия онтологии ИПР.

Таким образом, ПД в ИПР представляется не в лексическом пространстве терминов, которые удается выделить в документе, а в пространстве понятий предметной области, которые зафиксированы в онтологии ИПР.

Иллюстративный пример содержимого репозитория из 10 концептуальных индексов приведен на рисунке 1.

Для того чтобы иметь возможность определить отношения неразличимости на полном множестве ПД, построим терм-множество лингвистической переменной «Степень выраженности понятия онтологии» так, как показано на рисунке 2, где применяются следующие условные обозначения:

- μ_{C_i} – степень выраженности понятия C_i онтологии ИПР, включенного в концептуальный индекс;
- $\delta(\mu_{C_i})$ – функция принадлежности лингвистической переменной;
- VL (very low), L (low), ML (middle-low), M (middle), MH (middle-high), H (high), VH (very high) – термы лингвистической переменной.

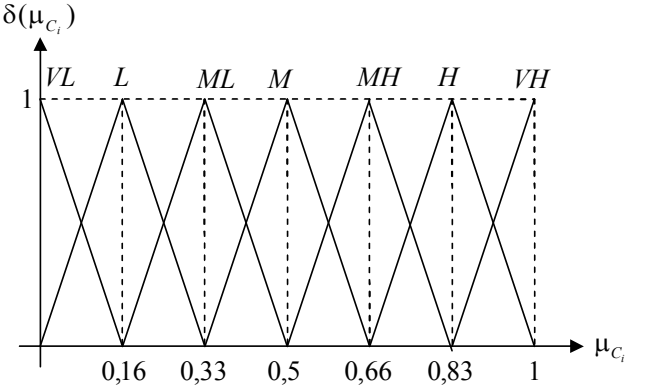

Рис. 2. Лингвистическая переменная «Степень выраженности понятия онтологии»

В качестве примера предположим, что в репозитории находятся разделенные на два класса ($K = \{1, 2\}$) десять ПД, структуры концептуальных индексов которых представлены на рисунке 1. Соответствующие значения степеней выраженности понятий C_1, C_2 и C_3 для указанного множества документов показаны на рисунке 3 а.

На рисунке 3 б представлены значения степеней выраженности понятий для того же множества документов, что и на рисунке 3 а, но не в числовой форме, а в виде значения лингвистической переменной, приведенной на рисунке 2. В алгоритме преобразования из числовой формы в лингвистическую учитывается следующее правило: если значение принадлежности степени выраженности понятия онтологии одинаково для двух соседних термов, то выбор делается в пользу большего терма (например, среди термов VL и L выбран будет терм L).

D	C_1	C_2	C_3	K
D_1	0,1	0,3	0,07	1
D_2	0,2	0,35	0,05	1
D_3	0,33	0,02	0	1
D_4	0	0	0,15	2
D_5	0,4	0,18	0,01	2
D_6	0	0,03	0,1	2
D_7	0,28	0,02	0	1
D_8	0,01	0,15	0,3	2
D_9	0,39	0,19	0	2
D_{10}	0,3	0	0,05	1

а)

D	C_1	C_2	C_3	K
D_1	L	ML	VL	1
D_2	L	ML	VL	1
D_3	ML	VL	VL	1
D_4	VL	VL	L	2
D_5	ML	L	VL	2
D_6	VL	VL	L	2
D_7	ML	VL	VL	1
D_8	VL	L	ML	2
D_9	ML	L	VL	2
D_{10}	ML	VL	VL	1

б)

Рис. 3. Значения степеней выраженности понятий

2 Интенсиональное описание кластеров репозитория

Интенсиональное описание кластеров основывается на множестве правил, которые фактически формируют границы классов. Можно перевести систему знаний из табличной формы (рис. 3 б) в логическую форму, выражая множество документов в дизъюнктивной нормальной форме:

$$\begin{cases}[(C_1 = L) \wedge (C_2 = ML) \wedge (C_3 = VL) \wedge (K = 1)] \vee \\ [(C_1 = L) \wedge (C_2 = ML) \wedge (C_3 = VL) \wedge (K = 1)] \vee \\ [(C_1 = ML) \wedge (C_2 = VL) \wedge (C_3 = VL) \wedge (K = 1)] \vee \\ [(C_1 = VL) \wedge (C_2 = VL) \wedge (C_3 = L) \wedge (K = 2)] \vee \\ [(C_1 = ML) \wedge (C_2 = L) \wedge (C_3 = VL) \wedge (K = 2)] \vee \\ [(C_1 = VL) \wedge (C_2 = VL) \wedge (C_3 = L) \wedge (K = 2)] \vee \\ [(C_1 = ML) \wedge (C_2 = VL) \wedge (C_3 = VL) \wedge (K = 1)] \vee \\ [(C_1 = VL) \wedge (C_2 = L) \wedge (C_3 = ML) \wedge (K = 2)] \vee \\ [(C_1 = ML) \wedge (C_2 = L) \wedge (C_3 = VL) \wedge (K = 2)] \vee \\ [(C_1 = ML) \wedge (C_2 = VL) \wedge (C_3 = VL) \wedge (K = 1)]. \end{cases}$$

Такая форма представления знаний может быть записана более компактно, определяя конъюнкцию как алгебраическое произведение:

$$\begin{aligned} &(C_1^L C_2^{ML} C_3^{VL} K^1) \vee (C_1^L C_2^{ML} C_3^{VL} K^1) \vee \\ &\vee (C_1^{ML} C_2^{VL} C_3^{VL} K^1) \vee (C_1^{VL} C_2^{VL} C_3^L K^2) \vee \\ &\vee (C_1^{ML} C_2^L C_3^{VL} K^2) \vee (C_1^{VL} C_2^{VL} C_3^L K^2) \vee \\ &\vee (C_1^{ML} C_2^{VL} C_3^{VL} K^1) \vee (C_1^{VL} C_2^L C_3^{ML} K^2) \vee \\ &\vee (C_1^{ML} C_2^L C_3^{VL} K^2) \vee (C_1^{ML} C_2^{VL} C_3^{VL} K^1). \end{aligned}$$

Теперь необходимо найти минимальное множество непротиворечивых правил (логических импликаций), которые характеризуют множество ПД, представленных в дизъюнктивной нормальной форме. Для множества условных атрибутов $C = \{C_1, C_2, C_3\}$ и решающего атрибута K такие правила будут иметь форму $C_i^a C_j^b ... C_k^c \to K^d$ или более развернуто:

$$(C_i = a) \wedge (C_j = b) \wedge ... \wedge (C_k = c) \to (K = d),$$

где $\{a, b, c, ...\}$ – допустимые значения из домена, который определяется значениями термов лингвистической переменной «Степень выраженности понятия онтологии».

Метод извлечения правил предполагает формирование так называемой решающей матрицы (decision matrix) для каждого отдельного значения d решающего атрибута K. Решающая матрица для значения d решающего атрибута K представляет список пар «атрибут – значение», которые различны между документами, имеющими $K = d$ и $K \neq d$.

Принимая во внимание таблицу на рисунке 3 б, атрибут K (номер кластера) будет искомой переменной, а $\{C_1, C_2, C_3\}$ – условными переменными. Искомая переменная может принимать два значения: $\{1,2\}$.

Рассмотрим ситуацию, относящуюся к первому кластеру (когда $K = 1$). Все полное множество документов разделим на документы, для которых $K = 1$, и документы, для которых $K \neq 1$. В нашем случае документы, для которых $K = 1$, – $\{D_1, D_2, D_3, D_7, D_{10}\}$, в то время как $K \neq 1$ соблюдается для $\{D_4, D_5, D_6, D_8, D_9\}$. Решающая матрица для $K = 1$ содержит все различия между ПД, для которых $K = 1$, и документами, для которых $K \neq 1$, то есть решающая матрица содержит все различия между $\{D_1, D_2, D_3, D_7, D_{10}\}$ и $\{D_4, D_5, D_6, D_8, D_9\}$. «Положительные» документы ($K = 1$) расположим по строкам, а «отрицательные» ($K \neq 1$) – по столбцам указанной матрицы:

	D_4	D_5	D_6	D_8	D_9
D_1	$C_1^L, C_2^{ML}, C_3^{VL}$	C_1^L, C_2^{ML}	$C_1^L, C_2^{ML}, C_3^{VL}$	$C_1^L, C_2^{ML}, C_3^{VL}$	C_1^L, C_2^{ML}
D_2	$C_1^L, C_2^{ML}, C_3^{VL}$	C_1^L, C_2^{ML}	$C_1^L, C_2^{ML}, C_3^{VL}$	$C_1^L, C_2^{ML}, C_3^{VL}$	C_1^L, C_2^{ML}
D_3	C_1^{ML}, C_3^{VL}	C_2^{VL}	C_1^{ML}, C_3^{VL}	$C_1^{ML}, C_2^{VL}, C_3^{VL}$	C_2^{VL}
D_7	C_1^{ML}, C_3^{VL}	C_2^{VL}	C_1^{ML}, C_3^{VL}	$C_1^{ML}, C_2^{VL}, C_3^{VL}$	C_2^{VL}
D_{10}	C_1^{ML}, C_3^{VL}	C_2^{VL}	C_1^{ML}, C_3^{VL}	$C_1^{ML}, C_2^{VL}, C_3^{VL}$	C_2^{VL}

Это означает, что относительно искомой переменной $K = 1$, например, документ D_3 отличается от документа D_6 атрибутами C_1 и C_3.

Далее, из каждой решающей матрицы формируются соответствующие булевы выражения, по одному выражению для каждой строки матрицы. Так, согласно приведенной выше матрице, получаем пять булевых выражений:

$$\begin{cases}(C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML}) \wedge (C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML}),\\ (C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML}) \wedge (C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML}),\\ (C_1^{ML} \vee C_3^{VL}) \wedge (C_2^{VL}) \wedge (C_1^{ML} \vee C_3^{VL}) \wedge (C_1^{ML} \vee C_2^{VL} \vee C_3^{VL}) \wedge (C_2^{VL}),\\ (C_1^{ML} \vee C_3^{VL}) \wedge (C_2^{VL}) \wedge (C_1^{ML} \vee C_3^{VL}) \wedge (C_1^{ML} \vee C_2^{VL} \vee C_3^{VL}) \wedge (C_2^{VL}),\\ (C_1^{ML} \vee C_3^{VL}) \wedge (C_2^{VL}) \wedge (C_1^{ML} \vee C_3^{VL}) \wedge (C_1^{ML} \vee C_2^{VL} \vee C_3^{VL}) \wedge (C_2^{VL}).\end{cases}$$

В полученном выражении присутствует избыточность. Поэтому следующим шагом будет его упрощение с применением традиционной булевой алгебры. Так утверждение

$$(C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML}) \wedge$$
$$\wedge (C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge$$
$$(C_1^L \vee C_2^{ML} \vee C_3^{VL}) \wedge (C_1^L \vee C_2^{ML}),$$

соответствующее документам $\{D_1, D_2\}$, упрощается до $C_1^L \vee C_2^{ML}$, которое и формирует импликацию:

$$(C_1 = L) \vee (C_2 = ML) \to (K = 1).$$

Аналогичным образом утверждение

$$(C_1^{ML} \vee C_3^{VL}) \wedge (C_2^{VL}) \wedge (C_1^{ML} \vee C_3^{VL}) \wedge$$
$$\wedge (C_1^{ML} \vee C_2^{VL} \vee C_3^{VL}) \wedge (C_2^{VL}),$$

соответствующее документам $\{D_3, D_7, D_{10}\}$, упрощается до $C_1^{ML} C_2^{VL} \vee C_3^{VL} C_2^{VL}$, которое и позволяет получить следующую импликацию:

$$(C_1 = ML \wedge C_2 = VL) \vee$$
$$\vee (C_3 = VL \wedge C_2 = VL) \to (K = 1).$$

В итоге получаем следующее множество правил:

$$\begin{cases}(C_1 = L) \to (K = 1),\\ (C_2 = ML) \to (K = 1),\\ (C_1 = ML) \wedge (C_2 = VL) \to (K = 1),\\ (C_3 = VL) \wedge (C_2 = VL) \to (K = 1).\end{cases}$$

Интерпретировать полученное множество правил можно следующим образом:

Кластер № 1 ИПР включает в себя ПД, у которых:

- степень выраженности понятия C_1 низкая (L) ИЛИ

- степень выраженности понятия C_2 средняя-низкая (ML) ИЛИ

- степень выраженности понятия C_1 средняя-низкая (ML) И степень выраженности понятия C_2 очень низкая (VL) ИЛИ

- степень выраженности понятия C_3 очень низкая (VL) И степень выраженности понятия C_2 очень низкая (VL).

ЗАКЛЮЧЕНИЕ

Решение задачи обобщенного (интенсионального) представления кластеров проектных документов является практически ценным для человеко-машинного взаимодействия. Представление такого решения в виде логических правил (где посылка есть набор высказываний о степени выраженности понятий предметной области, а заключение – номер кластера) является компактным и выразительным по сравнению с традиционным описанием кластеров документов в виде множества наиболее встречающихся слов-терминов. Кроме того, логическая форма является приемлемой для дальнейшей автоматической обработки результатов кластеризации.

СПИСОК ЛИТЕРАТУРЫ

1. Pawlak Z. Rough Sets: Present State and Future Prospects // Intelligent Automation and Soft Computing, 1996, V. 2

2. Наместников А.М., Филиппов А.А. Концептуальная индексация проектных документов // Автоматизация процессов управления. – 2010. – № 2 (20). – С. 34–39.

3. Наместников А.М., Чекина А.В., Корунова Н.В. Интеллектуальный сетевой архив электронных информационных ресурсов // Программные продукты и системы. – 2007. – № 4. – С. 10–13.