S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports

http://semr.math.nsc.ru

Том 14, стр. 1030–1040 (2017) УДК 519.17

DOI 10.17377/semi.2017.14.087 MSC 68R10

АЛГЕБРА МУЛЬТИРУБРИК НА КОРНЕВЫХ ДЕРЕВЬЯХ ИЕРАРХИЧЕСКИХ ТЕМАТИЧЕСКИХ КЛАССИФИКАТОРОВ

Н.А. ГАЙДАМАКИН, В.А. БАРАНСКИЙ

Abstract. The algebraic formalization of the use of hierarchical thematic classifiers in information retrieval depositories of documents was presented. The definitions of "rubricator ideals" and "multirubrics" on the root trees corresponding to hierarchical thematic classifiers were introduced. A lattice of rubricator ideals was constructed. Relations of thematic dominance, operations of lattice union and intersection of multirubrics were introduced, algorithms for their determination and calculation were presented. Therefore, a lattice of multirubrics was constructed that is isomorphic to the lattice of rubricator ideals. Isomorphism of lattice of multirubrics and boolean lattice of the set of leaf subsets of the root tree of the hierarchical thematic classifier was proved.

Keywords: root tree, hierarchical thematic classifier, ordinal ideal, rubricator ideal, multirubric, lattice of multirubrics, thematic dominance of multirubrics, union of multirubrics, intersection of multirubrics, information retrieval depositories of documents, thematic indexing.

1. Введение

Повсеместно применяемой техникой в организации информационно-поисковых хранилищ документов является использование тематических рубрик (ключевых слов, термов, словоформ) [1]. Каждый документ помечается тематическим индексом, представляющим набор рубрик, информация по которым содержится в соответствующем документе.

Gaydamakin, N.A., Baransky, V.A., Algebra of multirubric on root trees of hierarchical thematic classifiers.

Поступила 1 сентября 2017 г., опубликована 17 октября 2017 г.

Информационные потребности пользователей выражаются так же набором рубрик. Если тематический индекс документа и набор рубрик поискового запроса пересекаются, то документ считается *релевантным* запросу и выдается пользователю. *Пертинентность*, т.е. точность информационного поиска, в простейшем случае определяется мощностью пересечения.

В системах защиты информации рубрики могут использоваться в качестве основы тематического разграничения доступа к документам [2]. Тематические полномочия пользователей выражаются набором разрешенных рубрик. Если набор пользовательских рубрик полностью включается в набор рубрик документа, то доступ к документу разрешается.

При большом количестве рубрик для сокращения индексного описания документов применяют иерархические тематические классификаторы, в которых подмножества рубрик объединяются в рубрики-группы, которые, в свою очередь, объединяются в групповые рубрики более высокого уровня и т.д. Если документ содержит информацию по всем рубрикам некоторой группы, то в его тематический индекс вместо всех рубрик группы, помещается обозначение групповой рубрики. В результате тематические индексы документов образуются наборами узлов (вершин) иерархического тематического классификатора, которые можно трактовать как некие «мультирубрики».

Вместе с тем, насколько известно, алгебраического описания таким образом составленных наборов узлов иерархических тематических классификаторов и операций над ними не представлено.

2. Решетка рубрикаторных идеалов

В формальном плане иерархические тематические классификаторы описываются графом вида «корневое дерево», в котором полустепень исхода любой внутренней вершины больше либо равна 2 (см. рис. 1)[1].

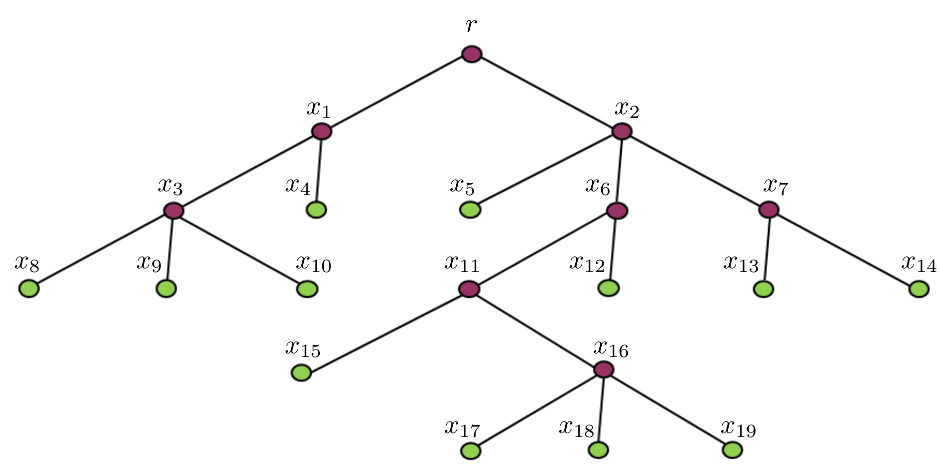

Рис. 1. Пример графа «корневое дерево» иерархического тематического классификатора

[1]Делить рубрику на подрубрики имеет смысл только в том случае, если количество подрубрик не менее двух.

Будем использовать терминологию для графов и решеток в соответствии с [2] и [3]. Под *графом* будем понимать неориентированный граф без петель и кратных ребер, а под *простой цепью* — маршрут в графе, в котором не повторяются вершины и ребра. Хорошо известно (см., например, [3]), что в любом дереве, т.е. связном графе без циклов, любые две его вершины соединены точно одной простой цепью.

Пусть $\boldsymbol{T}$ — корневое дерево с корнем r, т.е. дерево с выделенной вершиной r, которая называется его корнем. Рассмотрим на $\boldsymbol{T}$ естественный строгий частичный порядок $<$, такой, что выполняется $x < y$ в том и только в том случае, когда $x \neq y$ и y принадлежит простой цепи с началом в x и концом в r. Если $x < y$, то будем говорить, что y является *предком* вершины x, а x — *потомком* вершины y. Как обычно, будем писать $x \leq y$, если $x = y$ или $x < y$. Известно, что отношение $\leq$ является отношением частичного порядка на $\boldsymbol{T}$. Если $x < y$ и не существует такого z, что $x < z < y$, то будем говорить, что y является *отцом* вершины x, а x — *сыном* вершины y. Отца вершины x будем обозначать через $father(x)$. Заметим, что корень дерева является единственной вершиной, которая не имеет отца. Как обычно, *антицепью* в дереве $\boldsymbol{T}$ будем называть любое подмножество из $\boldsymbol{T}$, состоящее из попарно несравнимых элементов.

Непустое семейство $\boldsymbol{X} = \{x_1, x_2, \dots, x_t\}$ вершин дерева $\boldsymbol{T}$ называется *кланом*, если оно состоит из всех сыновей некоторой вершины этого дерева, которую будем обозначать через $father(\boldsymbol{X})$ и называть *отцом клана* $\boldsymbol{X}$. Очевидно, каждая вершина дерева, отличная от корня r, принадлежит точно одному клану, т.е. множество вершин дерева, отличных от корня, является объединением попарно непересекающихся кланов. Вершины дерева $\boldsymbol{T}$, не имеющие сыновей, будем называть *листьями*. Множество всех листьев дерева $\boldsymbol{T}$ будем обозначать через $List(\boldsymbol{T})$ или, кратко, через $\boldsymbol{L}$. Вершины, не являющиеся листьями, будем называть *внутренними вершинами* дерева.

Подмножество $\boldsymbol{I}$ дерева $\boldsymbol{T}$ будем называть *порядковым идеалом*, если оно замкнуто относительно взятия потомков вершин, т.е. если некоторая вершина x принадлежит $\boldsymbol{I}$, то $\boldsymbol{I}$ принадлежат и все её потомки (см. рис. 2). Очевидно, множество $\boldsymbol{T}$ и пустое множество $\varnothing$ по определению являются порядковыми идеалами дерева $\boldsymbol{T}$. Порядковый идеал называется *собственным порядковым идеалом*, если он не совпадает с $\boldsymbol{T}$. Легко видеть, что пересечение любого семейства порядковых идеалов дерева $\boldsymbol{T}$ является порядковым идеалом. Для любого подмножества $\boldsymbol{X} = \{x_1, x_2, \dots, x_t\}$ вершин дерева $\boldsymbol{T}$ пересечение всех порядковых идеалов, содержащих $\boldsymbol{X}$, очевидно является наименьшим по включению порядковым идеалом, содержащим $\boldsymbol{X}$. Этот порядковый идеал будем обозначать через $(x_1, x_2, \dots, x_t]$ или, кратко, через $(\boldsymbol{X}]$.

Будем говорить, что множество $\boldsymbol{X}$ порождает порядковый идеал $\boldsymbol{I}$, если $\boldsymbol{I} = (\boldsymbol{X}]$.

Нетрудно заметить, что порядковый идеал $(x_1, x_2, \dots, x_t]$, порожденный множеством вершин $\boldsymbol{X} = \{x_1, x_2, \dots, x_t\}$,

1) состоит из вершин $x_1, x_2, \dots, x_t$ и всех потомков этих вершин;

2) порождается антицепью, составленной из максимальных вершин множества $\boldsymbol{X}$.

Порядковый идеал $(x]$, где x — вершина дерева, будем называть *главным порядковым идеалом*. Ясно, что для любой антицепи $x_1, x_2, \dots, x_t$ порядковый

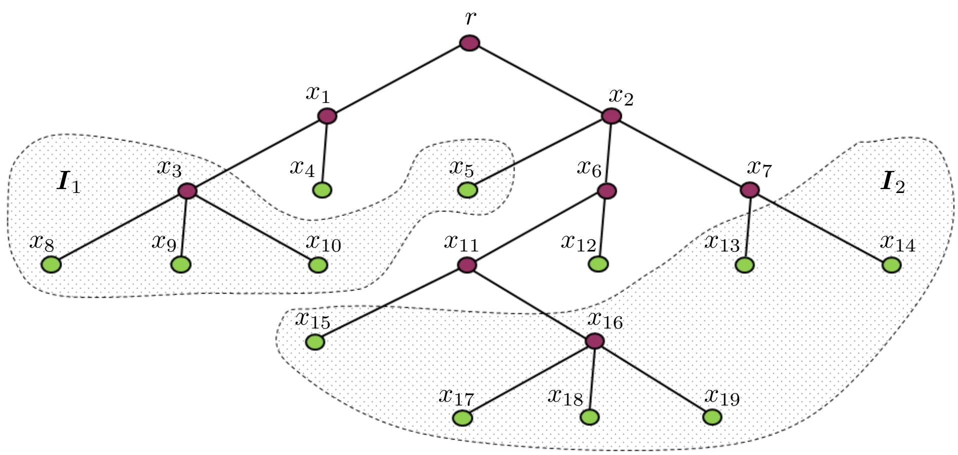

Рис. 2. Примеры порядковых идеалов $\boldsymbol{I}_1 = (x_3, x_5]$, $\boldsymbol{I}_2 = (x_{13}, x_{14}, x_{15}, x_{16}]$

идеал $(x_1, x_2, \dots, x_t]$ является теоретико-множественным объединением попарно непересекающихся главных порядковых идеалов $(x_1], (x_2], \dots, (x_t]$. При этом главный порядковый идеал, порожденный корнем r, совпадает с $\boldsymbol{T}$.

Лемма 1. *Вершина x собственного порядкового идеала $\boldsymbol{I}$ максимальна в $\boldsymbol{I}$ тогда и только, когда ее отец $father(x)$ не принадлежит $\boldsymbol{I}$.*

Доказательство. Покажем сначала, что отец $father(x)$ каждой не максимальной в $\boldsymbol{I}$ вершины x принадлежит $\boldsymbol{I}$. Действительно, если вершина x не максимальна в $\boldsymbol{I}$, то выполняется неравенство $x < y$ для некоторой вершины y из $\boldsymbol{I}$. Поскольку для любой вершины x дерева $\boldsymbol{T}$, отличной от корня, все ее предки образуют цепь (т.е. попарно сравнимы) с наименьшим элементом $father(x)$, неравенство $x < y$ влечет справедливость неравенства $father(x) \le y$, поэтому $father(x)$ принадлежит $\boldsymbol{I}$. С другой стороны, если вершина x максимальна в собственном порядковом идеале $\boldsymbol{I}$, то $x \ne r$, $father(x)$ существует и $father(x)$ не принадлежит этому идеалу. □

Следствие 1. *Любой клан $\boldsymbol{X}$ порядкового идеала $\boldsymbol{I}$ такой, что $father(\boldsymbol{X})$ не принадлежит $\boldsymbol{I}$, состоит из некоторых максимальных вершин порядкового идеала $\boldsymbol{I}$.*

Пусть $\boldsymbol{I}$ — собственный порядковый идеал, $\boldsymbol{A}$ — антицепь его максимальных вершин и $\boldsymbol{X}$ — некоторый клан из $\boldsymbol{I}$ такой, что $father(\boldsymbol{X})$ не принадлежит $\boldsymbol{I}$. Тогда $\boldsymbol{X}$ содержится в $\boldsymbol{A}$ и множество $\boldsymbol{I} \cup \{father(\boldsymbol{X})\}$, где $\cup$ — теоретико-множественное объединение, очевидно является порядковым идеалом, а множество $(\boldsymbol{A} \setminus \boldsymbol{X}) \cup \{father(X)\}$ является антицепью его максимальных вершин. Будем говорить, что антицепь $(\boldsymbol{A} \setminus \boldsymbol{X}) \cup \{father(X)\}$ получена из антицепи $\boldsymbol{A}$ с помощью *сжатия* клана $\boldsymbol{X}$.

Корневое дерево $\boldsymbol{T}$ с корнем r будем называть *рубрикаторным деревом* с корнем r, если любая его внутренняя вершина имеет не менее двух сыновей, т.е. в любом его клане имеется не менее двух вершин. Далее под $\boldsymbol{T}$ мы всюду будем понимать рубрикаторное дерево с корнем r.

Порядковый идеал $\boldsymbol{I}$ дерева $\boldsymbol{T}$ будем называть *рубрикаторным идеалом*, если для любого клана, с одержащегося в $\boldsymbol{I}$, в $\boldsymbol{I}$ содержится и его отец (на рис. 2

порядковый идеал $\boldsymbol{I}_1$ является рубрикаторным идеалом, а — $\boldsymbol{I}_2$ не является). Очевидно, само $\boldsymbol{T}$ и пустое множество $\varnothing$ по определению являются рубрикаторными идеалами. Легко видеть, что пересечение любого семейства рубрикаторных идеалов дерева $\boldsymbol{T}$ является рубрикаторным идеалом. Для любого подмножества $\boldsymbol{X} = \{x_1, x_2, \ldots, x_t\}$ вершин дерева $\boldsymbol{T}$ теоретико-множественное пересечение всех рубрикаторных идеалов, содержащих $\boldsymbol{X}$, очевидно является наименьшим по включению рубрикаторным идеалом, содержащим $\boldsymbol{X}$. Этот рубрикаторный идеал будем обозначать через $(x_1, x_2, \ldots, x_t]_R$ или, кратко, через $(\boldsymbol{X}]_R$.

Будем говорить, что множество $\boldsymbol{X}$ порождает рубрикаторный идеал $\boldsymbol{I}$, если $\boldsymbol{I} = (\boldsymbol{X}]_R$. Множество всех рубрикаторных идеалов дерева $\boldsymbol{T}$ обозначим через $RIdL(\boldsymbol{T})$. Оно является *решеткой* относительно теоретико-множественного включения. Действительно, решеточное пересечение $\boldsymbol{I}_1 \wedge \boldsymbol{I}_2$ двух рубрикаторных идеалов $\boldsymbol{I}_1$ и $\boldsymbol{I}_2$ (или, иными словами, наибольшая нижняя граница для $\boldsymbol{I}_1$ и $\boldsymbol{I}_2$) равно теоретико-множественному пересечению $\boldsymbol{I}_1$ и $\boldsymbol{I}_2$, а решеточное объединение $\boldsymbol{I}_1 \vee \boldsymbol{I}_2$ (или, иными словами, наименьшая верхняя граница для $\boldsymbol{I}_1$ и $\boldsymbol{I}_2$) равно наименьшему рубрикаторному идеалу, содержащему $\boldsymbol{I}_1 \cup \boldsymbol{I}_2$, т.е. $\boldsymbol{I}_1 \vee \boldsymbol{I}_2 = (\boldsymbol{I}_1 \cup \boldsymbol{I}_2]_R$. Будем называть $RIdL(\boldsymbol{T})$ *решеткой рубрикаторных идеалов* дерева $\boldsymbol{T}$.

3. Решетка мультирубрик

Антицепь $\boldsymbol{M}$ дерева $\boldsymbol{T}$ назовем *мультирубрикой*, если она не содержит ни одного клана (на рис. 2 антицепь $\{x_3, x_5\}$ является мультирубрикой, а антицепь $\{x_{13}, x_{14}, x_{15}, x_{16}\}$ не является). Отметим, что по определению пустое множество $\varnothing$ является мультирубрикой. Кроме того, любая вершина x дерева $\boldsymbol{T}$ образует одноэлементную мультирубрику, которую мы будем обозначать так же через x. Очевидно, множество всех максимальных элементов любого рубрикаторного идеала $\boldsymbol{I}$ образует мультирубрику, которую мы будем обозначать через $MR(\boldsymbol{I})$.

Для произвольного множества $\boldsymbol{Y}$ вершин дерева $\boldsymbol{T}$ рубрикаторный идеал $(\boldsymbol{Y}]_R$ и мультирубрика $\boldsymbol{M}$ всех его максимальных элементов могут быть построены с помощью следующего алгоритма.

Алгоритм 1.

1) В качестве начальной антицепи $\boldsymbol{A}_1$ возьмем множество всех максимальных вершин множества $\boldsymbol{Y}$.

2) Если антицепь $\boldsymbol{A}_i$ уже построена и в $\boldsymbol{A}_i$ имеется некоторый клан $\boldsymbol{X}$, то в качестве следующей антицепи $\boldsymbol{A}_{i+1}$ берем антицепь $(\boldsymbol{A}_i \setminus \boldsymbol{X}) \cup \{father(\boldsymbol{X})\}$, т.е. берем антицепь, полученную из $\boldsymbol{A}_i$ сжатием клана $\boldsymbol{X}$.

3) Выполняем шаг 2) до тех пор, пока это возможно, процесс обязательно завершится на некотором шаге t и будет построена антицепь $\boldsymbol{A}_t$, которая является искомой мультирубрикой $\boldsymbol{M}$ и для которой выполняется $(\boldsymbol{Y}]_R = (\boldsymbol{M}]$, т.е. рубрикаторный идеал, порожденный множеством $\boldsymbol{Y}$, состоит из вершин множества $\boldsymbol{M}$ и всех потомков этих вершин.

Доказательство корректности алгоритма. Вместе с антицепями $\boldsymbol{A}_i$ будем рассматривать порядковые идеалы $\boldsymbol{I}_i = (\boldsymbol{A}_i]$, им соответствующие. Ясно, что антицепь максимальных вершин порядкового идеала $(\boldsymbol{Y}]$ совпадает с антицепью вершин, максимальных в множестве $\boldsymbol{Y}$, поэтому $\boldsymbol{I}_1 = (\boldsymbol{A}_1] = (\boldsymbol{Y}]$. Предположим, что антицепь $\boldsymbol{A}_i$ уже построена, порядковый идеал $\boldsymbol{I}_i = (\boldsymbol{A}_i]$, содержит порядковые идеалы $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{i-1}$ и $\boldsymbol{A}_i$ является антицепью всех максимальных вершин в $\boldsymbol{I}_i$. Пусть в $\boldsymbol{A}_i$ имеется некоторый клан $\boldsymbol{X}$. Следующей антицепи $\boldsymbol{A}_{i+1} = (\boldsymbol{A}_i \setminus \boldsymbol{X}) \cup \{father(\boldsymbol{X})\}$ соответствует порядковый идеал $\boldsymbol{I}_{i+1} = (\boldsymbol{A}_{i+1}] = \boldsymbol{I}_i \cup \{father(\boldsymbol{X})\}$, который содержит $\boldsymbol{I}_i$ и, следовательно, содержит все порядковые идеалы $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{i-1}$. Ясно, что антицепь $\boldsymbol{A}_{i+1}$ является антицепью максимальных вершин порядкового идеала $\boldsymbol{I}_{i+1}$. Процесс построения антицепей $\boldsymbol{A}_i$ завершится на некотором шаге t и будет построена антицепь $\boldsymbol{A}_t$, которая не содержит кланов и которой соответствует порядковый идеал $\boldsymbol{I}_t = (\boldsymbol{A}_t]$, содержащий все предыдущие порядковые идеалы вида $\boldsymbol{I}_i$, причем $\boldsymbol{A}_t$ является множеством всех максимальных элементов в $\boldsymbol{I}_t$. Поскольку $\boldsymbol{A}_t$ является мультирубрикой, в силу следствия 1 порядковый идеал $\boldsymbol{I}_t$ является рубрикаторным идеалом и $\boldsymbol{I}_t = (\boldsymbol{A}_t]$. Так как $\boldsymbol{Y}$ содержится в $\boldsymbol{I}_t$, рубрикаторный идеал $(\boldsymbol{Y}]_R$ содержится в $\boldsymbol{I}_t$; верно и обратное включение, поскольку $\boldsymbol{I}_t$ состоит из вершин множества $\boldsymbol{Y}$, их потомков и вершин, получаемых только с помощью сжатия кланов в множествах $\boldsymbol{I}_i$, начиная с множества $(\boldsymbol{Y}]$. Таким образом, $(\boldsymbol{Y}]_R = (\boldsymbol{M}]$, где $\boldsymbol{M} = \boldsymbol{A}_t$. □

Ясно, что любая мультирубрика $\boldsymbol{M}$ является мультирубрикой максимальных вершин порождаемого ей рубрикаторного идеала и $(\boldsymbol{M}]_R = (\boldsymbol{M}]$. Таким образом, имеется естественное *взаимно однозначное отображение* φ *множества* $MRL(\boldsymbol{T})$ всех *мультирубрик* дерева $\boldsymbol{T}$ на *решетку рубрикаторных идеалов* $RIdL(\boldsymbol{T})$ этого дерева, причем для произвольной мультирубрики $\boldsymbol{M}$ выполняется равенство $\varphi(\boldsymbol{M}) = (\boldsymbol{M}]_R = (\boldsymbol{M}]$.

На множестве всех мультирубрик $MRL(\boldsymbol{T})$ определим *отношение частичного порядка* $\leq_M$, полагая $\boldsymbol{M}_1 \leq_M \boldsymbol{M}_2$ для произвольных мультирубрик $\boldsymbol{M}_1$ и $\boldsymbol{M}_2$ в том и только в том случае, когда $\varphi(\boldsymbol{M}_1) \subseteq \varphi(\boldsymbol{M}_2)$. Если $\boldsymbol{M}_1 \leq_M \boldsymbol{M}_2$, то будем говорить, что мультирубрика $\boldsymbol{M}_1$ *тематически доминирует* над мультирубрикой $\boldsymbol{M}_2$ (тематически *меньше* мультирубрики $\boldsymbol{M}_2$ или равна ей).

Нетрудно заметить, что условие $\boldsymbol{M}_1 \leq_M \boldsymbol{M}_2$ для мультирубрик $\boldsymbol{M}_1$ и $\boldsymbol{M}_2$ выполняется тогда и только тогда, когда для любого x из $\boldsymbol{M}_1$ существует y из $\boldsymbol{M}_2$ такой, что $x \leq y$. Отметим, что если хотя бы для одного $x \in \boldsymbol{M}_1$ не найдется $y \in \boldsymbol{M}_2$ такой, что $x \leq y$, то мультирубрики $\boldsymbol{M}_1$ и $\boldsymbol{M}_2$ не сравнимы или $\boldsymbol{M}_2 \subseteq_M \boldsymbol{M}_1$. Несравнимость мультирубрик будем обозначать знаком $\|_M$.

На рис. 3 приведены примеры тематического доминирования и несравнимости мультирубрик.

Отношение $\leq_M$ задает на множестве $MRL(\boldsymbol{T})$ частично упорядоченное множество, изоморфное решетке $RIdL(\boldsymbol{T})$ рубрикаторных идеалов дерева $\boldsymbol{T}$. Следовательно, $MRL(\boldsymbol{T})$ является *решеткой* относительно частичного *порядка* $\leq_M$ и *отображение* φ является *изоморфизмом решетки* $MRL(\boldsymbol{T})$ *на решетку* $RIdL(\boldsymbol{T})$.

Операции *объединения* и *пересечения* в решетке $MRL(\boldsymbol{T})$ обозначим через $\vee_M$ и $\wedge_M$, соответственно. Укажем *алгоритм нахождения объединения* $\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2$ мультирубрик $\boldsymbol{M}_1$ и $\boldsymbol{M}_2$ в решетке $MRL(\boldsymbol{T})$. Поскольку

$$\varphi(\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2) = \varphi(\boldsymbol{M}_1) \vee_R \varphi(\boldsymbol{M}_2) = (\boldsymbol{M}_1] \vee_R (\boldsymbol{M}_2] = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2]_R,$$

мультирубрика $\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2$ является антицепью максимальных вершин в рубрикаторном идеале в виде теоретико-множественного объединения наборов вершин, составляющих мультирубрики $\boldsymbol{M}_1$ и $\boldsymbol{M}_2$ $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2]_R$. Поэтому для получения мультирубрики $\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2$ можно использовать следующий частный случай Алгоритма 1 при $\boldsymbol{Y} = \boldsymbol{M}_1 \cup \boldsymbol{M}_2$.

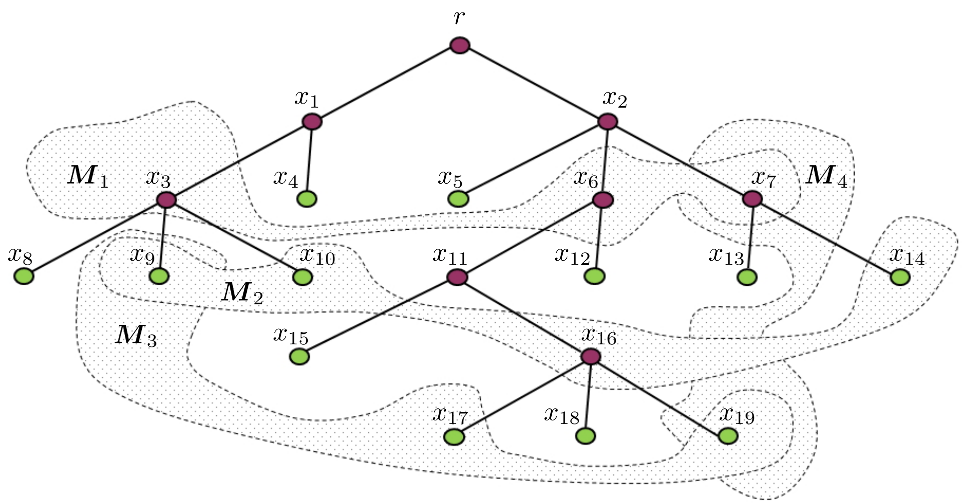

Рис. 3. Примеры тематического доминирования мультирубрик: $\boldsymbol{M}_3=\{x_9,x_{17},x_{19}\} \leq_M \boldsymbol{M}_2=\{x_9,x_{10},x_{14},x_{16}\} \leq_M \boldsymbol{M}_1=\{x_3,x_6,x_7\}$. Примеры несравнимости мультирубрик: $\boldsymbol{M}_4 = \{x_7,x_{19}\} \ \|_M$ $\|_M \ \boldsymbol{M}_3=\{x_9,x_{17},x_{19}\}$, $\boldsymbol{M}_4=\{x_7,x_{19}\} \ \|_M \ \boldsymbol{M}_2=\{x_9,x_{10},x_{14},x_{16}\}$

Алгоритм 2.

1) В качестве начальной антицепи $\boldsymbol{A}_1$ берем антицепь всех максимальных вершин множества $\boldsymbol{M}_1 \cup \boldsymbol{M}_2$ относительно порядка $\leq$ на множестве $\boldsymbol{T}$.

2) Если антицепь $\boldsymbol{A}_i$ уже построена и в $\boldsymbol{A}_i$ имеется некоторой клан $\boldsymbol{X}$ дерева $\boldsymbol{T}$, то, заменяя в $\boldsymbol{A}_i$ этот клан на вершину $father(\boldsymbol{X})$, получаем следующую антицепь $\boldsymbol{A}_{i+1}$.

3) Выполняем шаг 2) до тех пор, пока это возможно, процесс обязательно завершится и будет построена антицепь $\boldsymbol{A}_t$, которая является искомой мультирубрикой $\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2$.

Укажем теперь алгоритм нахождения пересечения $\boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2$ мультирубрик $\boldsymbol{M}_1$ и $\boldsymbol{M}_2$ в решетке $MRL(\boldsymbol{T})$.

Алгоритм 3.

1) В качестве начального множества $\boldsymbol{M}$ берем множество всех вершин x_1 из $\boldsymbol{M}_1$ таких, что для x_1 существует y_2 из $\boldsymbol{M}_2$, удовлетворяющее неравенству $x_1 \leq y_2$.

2) Затем добавляем в $\boldsymbol{M}$ все вершины x_2 из $\boldsymbol{M}_2$ такие, что для x_2 существует y_1 из $\boldsymbol{M}_1$, удовлетворяющее неравенству $x_2 \leq y_1$. Полученное множество вершин $\boldsymbol{M}$ и есть мультирубрика $\boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2$.

Доказательство корректности алгоритма. В начале заметим, что если для трех вершин x, y и z дерева $\boldsymbol{T}$ выполняются два неравенства $x \leq y$ и $x \leq z$, то вершины y и z сравнимы, т.е. верно $y \leq z$ или $z \leq y$. Поскольку

$$\varphi(\boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2) = \varphi(\boldsymbol{M}_1) \wedge_R \varphi(\boldsymbol{M}_2) = (\boldsymbol{M}_1] \wedge (\boldsymbol{M}_2] = (\boldsymbol{M}_1] \cap (\boldsymbol{M}_2],$$

мультирубрика $\boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2$ является множеством максимальных элементов порядкового идеала $(\boldsymbol{M}_1] \cap (\boldsymbol{M}_1]$. Теперь осталось воспользоваться сделанным замечанием. $\square$

Поскольку $MRL(\boldsymbol{T})$ является решеткой, операции пересечения и объединения мультирубрик, вычисляемые по Алгоритмам 2 и 3, удовлетворяют тождествам *идемпотентности*, *коммутативности* и *ассоциативности*.

Также для объединения и пересечения мультирубрик выполняются *законы поглощения*.

На рис. 4 приведены примеры вычислений объединения и пересечения мультирубрик.

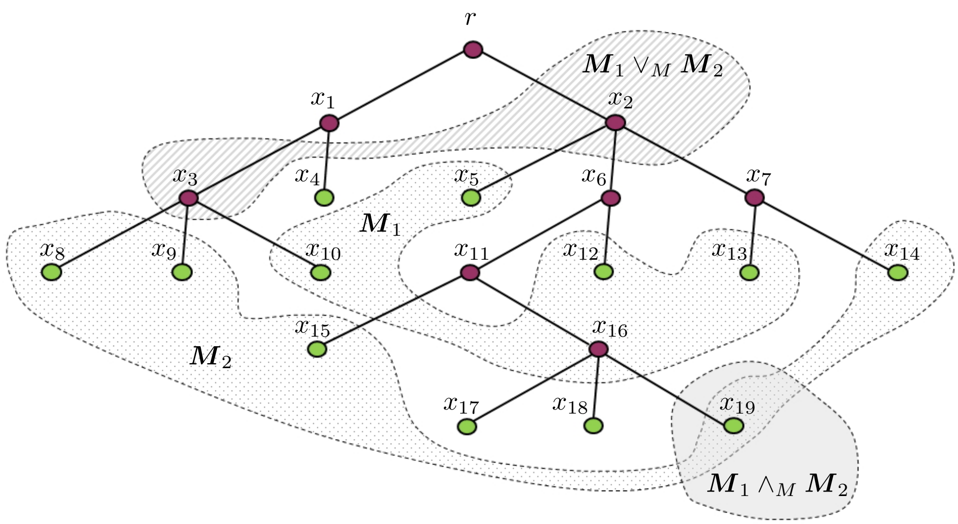

Рис. 4. Примеры вычисления операций объединения и пересечения мультирубрик. Пусть $\boldsymbol{M}_1 = \{x_5, x_{10}, x_{12}, x_{13}, x_{16}\}$ и $\boldsymbol{M}_2 = \{x_8, x_9, x_{14}, x_{15}, x_{19}\}$. Тогда $\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2 = \{x_2, x_3\}$, $\boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2 = \{x_{19}\}$

4. Изоморфизм решетки рубрикаторных идеалов и решетки мультирубрик булевой решетке листовых подмножеств

Далее, рассмотрим отображение $List(x)$, которое каждой вершине x дерева $\boldsymbol{T}$ ставит в соответствие множество всех листьев b таких, что $b \leq x$. Для каждой вершины x дерева $\boldsymbol{T}$ множество $List(x)$ будем называть *листовым множеством вершины* x. Отметим, что листовое множество листовой вершины b состоит из одного элемента — самой вершины b. Количество вершин множества $List(x)$ будем обозначать через $lweight(x)$ и будем называть его *листовым весом вершины* x. Под *высотой* h *вершины* x дерева $\boldsymbol{T}$ будем понимать наибольшую из длин простых цепей от вершины x до листьев. Пусть $\boldsymbol{I}$ — рубрикаторный идеал. Через $List(\boldsymbol{I})$ обозначим множество всех листьев дерева $\boldsymbol{T}$, содержащихся в $\boldsymbol{I}$.

Лемма 2. $(List(\boldsymbol{I}))]_R = \boldsymbol{I}$ *для любого рубрикаторного идеала* $\boldsymbol{I}$.

Доказательство. Очевидно, $(List(\boldsymbol{I}))]_R$ содержится в $\boldsymbol{I}$, поскольку $(List(\boldsymbol{I}))]_R$ является наименьшим рубрикаторным идеалом, содержащим множество $List(\boldsymbol{I})$. Докажем теперь обратное включение. Для любой вершины w из $\boldsymbol{I}$

индукцией по высоте вершины w покажем, что w принадлежит $(List(\boldsymbol{I})]_R$. Листья из $\boldsymbol{I}$ являются вершинами высоты 0 и лежат в $(List(\boldsymbol{I})]_R$. Пусть w имеет высоту $h > 0$. Рассмотрим клан $\boldsymbol{X}$, отцом которого является вершина w. Ясно, что $\boldsymbol{X}$ содержится в $\boldsymbol{I}$ и высота каждой вершины из $\boldsymbol{X}$ не превосходит $h-1$, отсюда в силу предположения индукции клан $\boldsymbol{X}$ содержится в $(List(\boldsymbol{I})]_R$ и, следовательно, $w = father(\boldsymbol{X})$ принадлежит $(List(\boldsymbol{I})]_R$. □

Лемма 3. *Для произвольного подмножества $\boldsymbol{Z}$ множества всех листьев $\boldsymbol{L}$ дерева $\boldsymbol{T}$ выполняется $List((\boldsymbol{Z}]_R) = \boldsymbol{Z}$.*

Доказательство. Применим Алгоритм 1 к множеству $\boldsymbol{Y} = \boldsymbol{Z} = \boldsymbol{A}_1$. Заметим, что $List((\boldsymbol{A}_i]) = \boldsymbol{Z}$ для каждой антицепи $\boldsymbol{A}_i$, поскольку $List((\boldsymbol{A}_1]) = \boldsymbol{Z}$ и всегда выполняется $List((\boldsymbol{A}_i]) = List((\boldsymbol{A}_{i+1}])$. Следовательно, $List((\boldsymbol{Z}]_R) = List((\boldsymbol{A}_t]) = \boldsymbol{Z}$. □

В силу Лемм 2 и 3 отображение $List(\boldsymbol{I})$, которое каждому рубрикаторному идеалу $\boldsymbol{I}$ ставит в соответствие листовое множество всех содержащихся в нём листьев, является *взаимно однозначным отображением* множества $RIdL(T)$ на *булеву решетку* $\mathfrak{B}(\boldsymbol{L})$ всех подмножеств множества $\boldsymbol{L}$ листьев дерева $\boldsymbol{T}$. Поскольку рубрикаторный идеал $\boldsymbol{I}_1$ содержится в рубрикаторном идеале $\boldsymbol{I}_2$ тогда и только тогда, когда множество $List(\boldsymbol{I}_1)$ содержится в множестве $List(\boldsymbol{I}_2)$, отображение $List(\boldsymbol{I})$ является *изоморфизмом* решетки рубрикаторных идеалов $RIdL(\boldsymbol{T})$ дерева $\boldsymbol{T}$ на булеву решетку $\mathfrak{B}(\boldsymbol{L})$ множества $\boldsymbol{L}$ всех листьев этого дерева. Следовательно, $RIdL(\boldsymbol{T})$ является *булевой решеткой.* В силу этого решетка мультирубрик дерева $\boldsymbol{T}$ также является булевой решеткой. Таким образом, справедлива следующая

Теорема 1. *Решетка $MRL(\boldsymbol{T})$ всех мультирубрик и решетка $RIdL(\boldsymbol{T})$ всех рубрикаторных идеалов дерева $\boldsymbol{T}$ являются булевыми решетками, изоморфными булевой решетке $\mathfrak{B}(\boldsymbol{L})$, где $\boldsymbol{L}$ — множество всех листьев этого дерева.*

Рассмотрим отображение ψ решетки мультирубрик $MRL(\boldsymbol{T})$ на булеву решетку $\mathfrak{B}(\boldsymbol{L})$, которое является суперпозицией отображений φ и $List(\boldsymbol{I})$. Ясно, что ψ является изоморфизмом решетки $MRL(\boldsymbol{T})$ на решетку $\mathfrak{B}(\boldsymbol{L})$. Для любой мультирубрики $\boldsymbol{M}$ и произвольного набора $\boldsymbol{X}$ листьев дерева $\boldsymbol{T}$ условие $\psi(\boldsymbol{M}) = \boldsymbol{X}$ очевидно выполняется тогда и только тогда, когда $(\boldsymbol{M}]_R = (\boldsymbol{X}]_R$, т.е. когда $\boldsymbol{M}$ и $\boldsymbol{X}$ порождают один и тот же рубрикаторный идеал.

Нетрудно заметить, что для любой мультирубрики $\boldsymbol{M}$ множество $\psi(\boldsymbol{M})$ совпадает с множеством всех листьев x таких, что $x \le y$ для некоторой вершины y мультирубрики $\boldsymbol{M}$. Поэтому множество $\psi(\boldsymbol{M})$ естественно записывать в виде $List(\boldsymbol{M})$ и называть его *листовым множеством мультирубрики* $\boldsymbol{M}$.

Итак, отображение $List(\boldsymbol{M})$ является *изоморфизмом* решетки мультирубрик $MRL(\boldsymbol{T})$ дерева $\boldsymbol{T}$ на булеву решетку $\mathfrak{B}(\boldsymbol{L})$, где $\boldsymbol{L}$ — множество всех листьев этого дерева. Если для мультирубрики $\boldsymbol{M}$ выполняется $List(\boldsymbol{M}) = \boldsymbol{X}$, то мультирубрика $\boldsymbol{M}$ фактически представляет из себя «*сжатое представление*» заданного подмножества листьев $\boldsymbol{X}$ дерева $\boldsymbol{T}$. Процедура сжатия подмножества листьев X , осуществляемая по *Алгоритму 1*, реализует *отображение заданного подмножества листьев* $\boldsymbol{X}$ в соответствующую *мультирубрику* $\boldsymbol{M}$. Будем обозначать такое отображение через $MR(\boldsymbol{X})$. Таким образом, если $\boldsymbol{X} = List(\boldsymbol{M})$, то $MR(\boldsymbol{X}) = \boldsymbol{M}$.

Вернувшись к примерам на рис. 3 отметим, что

$$List(\boldsymbol{M}_3) = \{x_9, x_{17}, x_{19}\} \subset List(\boldsymbol{M}_2) = \{x_9, x_{10}, x_{14}, x_{17}, x_{18}, x_{19}\} \subset$$
$$\subset List(\boldsymbol{M}_1) = \{x_8, x_9, x_{10}, x_{12}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18}, x_{19}\}.$$

Аналогично для объединения и пересечения на рис. 4 выполняется

$$List(\boldsymbol{M}_1) \cup List(\boldsymbol{M}_2) = \{x_5, x_{10}, x_{12}, x_{13}, x_{17}, x_{18}, x_{19}\} \cup \{x_8, x_9, x_{14}, x_{15}, x_{19}\} =$$
$$= \{x_5, x_8, x_9, x_{10}, x_{12}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18}, x_{19}\} = List(\{x_2, x_3\}) =$$
$$= List(\boldsymbol{M}_1 \vee_M \boldsymbol{M}_2),$$
$$MR(\{x_5, x_8, x_9, x_{10}, x_{12}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18}, x_{19}\}) = \{x_2, x_3\} = \boldsymbol{M}_1 \vee_M \boldsymbol{M}_2,$$
$$List(\boldsymbol{M}_1) \cap List(\boldsymbol{M}_2) = \{x_5, x_{10}, x_{12}, x_{13}, x_{17}, x_{18}, x_{19}\} \cap \{x_8, x_9, x_{14}, x_{15}, x_{19}\} =$$
$$= \{x_{19}\} = List(\boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2),$$
$$MR(\{x_{19}\}) = \{x_{19}\} = \boldsymbol{M}_1 \wedge_M \boldsymbol{M}_2.$$

Поскольку решетка мультирубрик $MRL(\boldsymbol{T})$ изоморфна булевой решетке $\mathfrak{B}(\boldsymbol{L})$ листовых подмножеств, она удовлетворяет тождествам *дистрибутивности* и, следовательно, она *модулярна*.

Кроме того, в силу изоморфизма решетки мультирубрик $MRL(\boldsymbol{T})$ булевой решетке $\mathfrak{B}(\boldsymbol{L})$ легко решается задача определения мощности $\boldsymbol{N_M}$ множества мультирубрик на иерархическом тематическом классификаторе, которая равна количеству подмножеств множества $\boldsymbol{L}$ всех листьев дерева $\boldsymbol{T}$:

$$\boldsymbol{N_M} = 2^{\boldsymbol{N_L}},$$

где $\boldsymbol{N_L}$ — количество листьев корневого дерева иерархического тематического классификатора.

5. Заключение

Как отмечалось во введении, иерархические тематические классификаторы применяются, прежде всего, для сокращения записей тематических индексов в информационно-поисковых хранилищах документов. В этом отношении построенная решетка мультирубрик $MRL(\boldsymbol{T})$ и факт ее изоморфизма булевой решетке $\mathfrak{B}(\boldsymbol{L})$ выражают алгебраическую эквивалентность тематического индексирования документов наборами листовых рубрик и индексирования мультирубриками на корневых деревьях иерархических тематических классификаторов.

Важно отметить, что суть иерархических тематических классификаторов по сравнению с неупорядоченными (несистематизированными) наборами тематических рубрик существенно глубже. Иерархические тематические классификаторы отражают структуру предметной области сведений, содержащихся в информационно-поисковых хранилищах документов. Пользователи, оперируя этой структурой, более «осмысленно» и эффективно формируют свои информационные потребности.

Кроме того, иерархические тематические классификаторы выступают информационной основой для организационной структуры самих хранилищ документов. Документы группируются, размещаются и хранятся в соответствии со структурой классификатора, что обеспечивает эффективный доступ к ним по запросам пользователей.

Представленные соотношения и процедуры определения отношения тематического доминирования $\leq_M$, операций объединения $\vee_M$ и пересечения $\wedge_M$

мультирубрик создают алгоритмическую основу для повышения эффективности функционирования автоматизированных информационных систем, строящихся на принципах информационно-поисковых хранилищ документов. Изоморфизм решеток $MRL(\boldsymbol{T})$ и $\mathfrak{B}(\boldsymbol{L})$ обеспечивает в зависимости от вида представления корневых деревьев (матрицами или списками смежности, двоичными векторами по множеству листьев и т.п.) выбор тех алгоритмов, которые наиболее эффективны для мультирубрик или для листовых подмножеств.

References

[1] N.A. Gaydamakin, *Avtomatizirovannyie informatsionnyie sistemyi, bazyi i banki dannyih*, M.: Gelios ARV, 2002.

[2] N.A. Gaydamakin, *A model of thematic differentiation of access to information for the hierarchical classifier in automatic control systems*, Automation and Remote Control, **3** (2003), 177–189.

[3] M.O. Asanov, V.A. Baransky, V.V. Rasin, *Diskretnaya matematika: grafyi, matroidyi, algoritmyi*, SPb.: Lan, 2010.

[4] V.A. Baransky, V.V. Kabanov, *Obschaya algebra i ee prilozheniya*, Ekaterinburg: UrGU, 2008.

Nikolay Aleksandrovich Gaydamakin
Ural Federal University,
pr. Lenina, 51,
620083, Ekaterinburg, Russia
E-mail address: haid2@bk.ru

Vitaly Anatolievich Baransky
Ural Federal University,
pr. Lenina, 51,
620083, Ekaterinburg, Russia
E-mail address: vitaly.baransky@urfu.ru