КВАНТИТАТИВНАЯ ИСТОРИЯ
QUANTITATIVE HISTORY

МЕТОДЫ ТЕКСТОЛОГИЧЕСКОЙ ГЕНЕАЛОГИЧЕСКОЙ КЛАССИФИКАЦИИ: МАТРИЦА БЛИЗОСТИ VS МАТРИЦА НЕЧЕТКОГО ОТНОШЕНИЯ*

METHODS OF TEXTOLOGICAL GENEALOGICAL CLASSIFICATION: SIMILARITY MATRIX VS FUZZY RELATION MATRIX

Шпирко Сергей Валерьевич,
кандидат физико-математических наук,
старший научный сотрудник факультета управления и прикладной математики (ФУПМ) Московского физико-технического института (МФТИ),
старший преподаватель кафедры исторической информатики исторического факультета МГУ им. М. В. Ломоносова
E-mail: shpirko@yahoo.com

Sergey V. Shpirko

Подавляющее число формализованных моделей текстологической классификации используют в своих построениях т. н. матрицу близости. Данная матрица вычисляется на основе попарного сличения списков средневекового текста и отражает их взаимную схожесть. Проблема заключается в том, что каждый список неоднозначно включается в определенную текстологически близкую группу. Чтобы решить данную проблему, автор настоя-

The most formalized models of textological classification use so called similarity matrix. This matrix is formed as a result of pairwise comparison between copies of some medieval text and reflects their mutual similarity. The problem is that each copy can be ambiguously included in the closest textological group. To solve it the author of this paper develops a model of classification which is based on the application of fuzzy set theory. This model uses so called fuzzy relation ma-

* Статья подготовлена в рамках проекта, поддержанного Российским фондом фундаментальных исследований (РФФИ), грант № 16–06–00365 А «Историко-текстологический анализ средневековых русских текстов на основе применения подходов, алгоритмов и программы нечеткой классификации».

щей работы развивает модель классификации, основывающуюся на применении теории нечетких множеств. Данная модель использует так называемую матрицу нечеткого отношения и позволяет проводить генеалогическую классификацию с заранее заданной степенью уверенности. В настоящей работе рассматриваются методы, использующие матрицу близости и матрицу нечеткого отношения. Результаты их работы сопоставляются на конкретных примерах.

Ключевые слова: текстология, нечеткая классификация, класс эквивалентности, антисимметричное отношение, нечеткий порядок, стеммматология.

trix and enables to carry out genealogical textological classification with the given degree of confidence. This paper discusses methods that use either similarity or fuzzy relation matrix. Their results are compared together with a few examples.

Keywords: textology, fuzzy classification, equivalence class, antisymmetric relation, fuzzy order, stemmatology.

ВВЕДЕНИЕ

Как известно, текст средневекового произведения в процессе своего бытования подвергался многочисленным исправлениям, которые носили как сознательный (стилистические, смысловые), так и бессознательный характер (описки, орфографические исправления, диалектизмы). Подобные исправления могли привести к весьма значительным искажениям исходного текста. Непременным условием для реконструкции максимально близкого к оригиналу текста (архетипа) является классификация сохранившихся списков, выявление текстологически близких групп и установление генеалогопреемственных связей между ними.

Примерно с начала 60-х гг. XX в. для решения задачи формализованной классификации в текстологии стали применяться математические методы. В основе большинства построенных моделей лежит содержательное предположение относительно процесса копирования списков: «Чем ближе генеалогически пара списков, тем меньше различий содержат их тексты». На основе попарного сличения списков составлялась матрица близости (или противоположная ей по смыслу матрица расстояний), для анализа которой применялись различные математические методы, в частности, кластерные, теоретико-графовые.

Возвращаясь к содержательным предположениям модели копирования списков, заметим, что весьма часто переписываемый текст правился дополнительно по одной или нескольким рукописям. Таким образом, можно говорить о принадлежности списка к той или иной группе лишь с определенной долей достоверности. Границы таких групп становятся размытыми, в них выделяют ядро и периферию. По мнению автора настоящей работы, все эти соображения весьма естественным образом вписываются в рамки теории нечетких множеств (fuzzy set theory), появившейся в середине 60-х гг. XX в. и интенсивно и успешно применяющейся в различных отраслях человеческой деятельности (робототехника, медицинская диагностика и т. д.).

В основе предложенной автором нечеткой модели лежат два содержательных предположения: «Чем больше доля унаследованных ошибок (уклонений от нормы) из одного списка в другой, тем достовернее гипотеза о том, что первый список генеалогически предшествует второму» и «Чем меньше доля унаследованных ошибок пары списков, тем более независимы они друг от друга». Построенная на основе данного принципа матрица (матрица нечеткого отношения) уже не будет, в отличие от матрицы расстояний, симметричной. Однако в теории нечетких множеств существуют алгоритмы, позволяющие вычленять из нее как симметричную, так и антисимметричную составляющую. На языке текстологии это как раз и означает выделение текстологически близких групп и выяснение истории их бытования, взаимосвязей. Причем степень детализации получающейся стеммы регулируется специальным внешним параметром — порогом (уверенности, с которой списки объединяются в нечеткие классы и между которыми устанавливаются генеалогопреемственные связи).

В первой части настоящей работы дается обзор наиболее важных методов, использующих в своих построениях матрицу расстояний. Во второй части рассматриваются элементы теории нечетких множеств, необходимые для понимания метода нечеткой генеалогической классификации, излагаемого автором в третьей части. Наконец, в четвертой части проводится сопоставление результатов работы приведенных методов на конкретных примерах.

В приложении автором приводится обоснование важного утверждения, обеспечивающего проведение нечеткой классификации при произвольном значении порога.

I. МЕТОДЫ, ОСНОВАННЫЕ НА ПРИМЕНЕНИИ МАТРИЦЫ РАССТОЯНИЙ

Количественные методы, применяющиеся в текстологии, традиционно делят на две категории — *кладистику* (от др.-греч. κλάδος — ветвь) и *кластерный анализ* (от англ. cluster — пучок)[1]. Первые методы нацелены на выяснение истории бытования письменного текста, установления его архетипа — текста, наиболее близкого к оригиналу. Вторые ставят себе менее амбициозную задачу — определение текстологически близких групп из сохранившихся списков. И первые и вторые методы отталкиваются в своих построениях от так называемой матрицы расстояний.

Для построения данной матрицы производится сличение всех сохранившихся списков изучаемого текста, на основе которого формируется набор узлов разночтений. Узлами разночтений могут быть как отдельные слова, так и целые словосочетания. Данные узлы характеризуют места текста, несовпадающие в разных списках. Далее для каждой пары списков подсчитывается число совпадающих узлов, которое делится на общее число узлов разночтений. Полученное число (коэффициент) в качестве элемента помещается в матрицу расстояний на пересечении соответствующих строки и столбца. Построенная таким образом квадратная и симметричная матрица характеризует взаимную «близость» списков относительно друг друга.

Первый метод кластерного типа был предложен в начале 1960-х гг. американскими текстологами-библеистами Э. Колвеллом и Э. Тьюном[2]. С конца 1980-х гг. данный метод творчески развивается в работах советского текстолога А. А. Алексеева[3]. Общий подход заключается в преобразовании исходной матрицы, перестановке ее строк и столбцов таким образом, чтобы собрать рядом списки с максимально возможным значением коэффициентов. Различие заключается в выборе конкретного алгоритма подобного группирования. Например, можно анализировать все списки в порядке убывания их коэффициентов. Сначала находится пара списков с максимальным коэффициентом, которая и образует первый кластер. Следующая пара (с ближайшим значением коэффициента) образует новый кластер либо включается в первый, если один из списков уже входит в него. Таким образом, все списки будут включены в тот или иной кластер. Переставленные в итоговой матрице списки будут следовать в том же порядке, в каком они объединялись в кластеры. Ниже приведен фрагмент такой матрицы, построенной при анализе греческой новозаветной традиции[4]. Границы выявленных кластеров выделены в матрице цветом.

Итоговая матрица близости, содержащая процент сходства между парами списков (Евангелие от Иоанна, глава 11)

	P^{75}	B	W	ℵ	ς	Ω	C^R	A	Ψ	D	P^{45}	P^{66}	Θ	565
P^{75}		92	72	77	52	52	47	57	65	43	48	57	60	49
B	92		69	75	52	53	48	57	67	49	44	60	54	50
W	72	69		82	57	56	62	64	73	44	52	62	57	55
ℵ	77	75	82		58	60	58	65	67	40	51	58	56	51
ς	52	52	57	58		93	78	76	73	42	56	47	63	65
Ω	52	53	56	60	93		75	77	74	44	54	51	60	67
C^R	47	48	62	58	78	75		78	71	44	64	56	56	69
A	57	57	64	65	76	77	78		76	45	59	59	60	64
Ψ	65	67	73	67	73	74	71	76		46	48	66	59	63
D	43	49	44	40	42	44	44	45	46		68	58	37	30
P^{45}	48	44	52	51	56	54	64	59	48	68		69	52	44
P^{66}	57	60	62	58	47	51	56	59	66	58	69		55	45
Θ	60	54	57	56	63	60	56	60	59	37	52	55		61
565	49	50	55	51	65	67	69	64	63	30	44	45	61	

Как видно из таблицы, Э. Колвелл и Э. Тьюн выделяют 4 кластера (text-type group). Первый (B text-type) состоит из четырех списков: P^{75}, B, W и ℵ, совпадающих друг с другом в диапазоне от 69 до 92%. В то же время совпадения внутри B-кластера выше как минимум на 10% по сравнению с остальными списками (кроме Ψ). Примерно похожая картина наблюдается и для второго кластера (A text-type). Данный кластер состоит из пяти списков: ς, Ω, C^R, A и Ψ. Диапазон совпадений составляет 71÷93%, минимальный «зазор» с большинством остальных списков — 10%. Значительно «хуже» обстоит дело с третьим кластером (Δ text-type). Составляющие его три списка D, P^{45} и P^{66} совпадают друг с другом не более чем на 69%. Причем к P^{66} весьма близки другие три списка (B, W, Ψ), что ставит под сомнение вообще выделение данного кластера. Четвертый кластер (Γ text-type) состоит из двух списков — Θ и 565. Причем Θ совпадает со всеми остальными списками (кроме D) не меньше чем на 52%.

Более того, сам процесс кластеризации является неустойчивым: изменение порядка рассмотрения списков может привести к иным результатам. Так, как показывают Э. Колвелл и Э. Тьюн, если брать Θ в качестве базового списка для сравнения, то наиболее близкими к нему окажутся ς и 565. В то же время, если начать с 565, то ближайшими к нему окажутся не Θ, а C^R, Ω, ς, A и Ψ. Таким образом, приведенный алгоритм кластеризации нуждается в доработке, что и делают авторы. Но все равно остается принципиальная проблема *размытости* границ кластеров. Особенно это справедливо в условиях контаминации, когда два и более текста смешиваются в одном. В этом случае *невозможно однозначно* определить принадлежность списка тому или иному кластеру.

В 1970-е гг. голландский текстолог А. Деес предложил новый метод текстологического исследования, позволяющий представить результаты своей работы в виде неориентированного дерева[5]. Все конечные узлы данного дерева соответствуют сохранившимся спискам, а промежуточные узлы — их гипотетическим предкам (протографам). В 1990-е гг. предложенный метод был творчески доработан голландским математиком Э. Ваттелем[6]. В основе нового подхода лежит предположение, что чем больше коэффициент сходства у пары списков, тем достовернее они окажутся потомками одного и того же гипотетического антиграфа (непосредственного предка).

При использовании данного метода работа начинается с рассмотрения исходной матрицы близости. Все списки анализируются в порядке убывания коэффициентов матрицы. Идея метода состоит в последовательном объединении группируемых узлов: сначала самых близких и в конце — самых отдаленных. Расстояние между группами определяется как среднее арифметическое расстояние всех входящих в него узлов. На очередном шаге находится пара с максимальным коэффициентом, которая объединяется в одну группу. При этом из матрицы вычеркиваются строка и столбец, соответствующие первому узлу, а расстояния новой группы до всех остальных вычисляется по вышеуказанной формуле. Таким образом, в данном случае применяется один из агломеративно-иерархических методов кластерного анализа[7]. Проиллюстрируем работу метода на конкретном примере из шести списков: a, b, c, d, e и f. Исходная матрица приведена на рисунке 1.

	a	b	c	d	e	f
a		90	60	30	30	30
b	90		65	40	40	40
c	60	65		55	20	20
d	30	40	55		70	80
e	30	40	20	70		85
f	30	40	20	80	85	

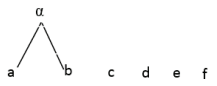


Рис. 1. Матрица расстояний 6X6.
Объединение узлов a, b и возведение их к узлу α

Максимальный элемент исходной матрицы соответствует паре a и b (90% сходства). Объединяем оба узла в одну группу с гипотетическим общим узлом α. Удаляем из матрицы первый столбец и строку, а вторая строка и столбец будут соответствовать новому узлу. Для него коэффициент расстояния от других узлов рассчитывается как среднее арифметическое расстояние для a и b. Новая матрица (размером уже 5X5) представлена на рисунке 2.

	α	c	d	e	f
α		63	35	35	35
c	63		55	20	20
d	35	55		70	80
e	35	20	70		85
f	35	20	80	85	

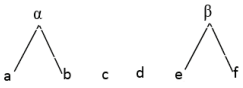


Рис. 2. Матрица расстояний 5X5.
Объединение узлов e, f и возведение их к узлу β

Здесь максимальное значение элемента (85%) соответствует узлам e и f. На место этих узлов в матрицу помещается новый узел β с пересчитанными коэффициентами в строке и столбце. Получаем новую матрицу размером 4X4 (см. рис. 3).

	α	c	d	β
α		63	35	35
c	63		55	20
d	35	55		70
β	35	20	70	

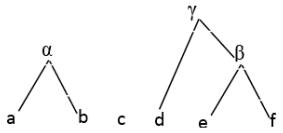


Рис. 3. Матрица расстояний 4X4.
Объединение узлов d, β и возведение их к узлу γ

Продолжая аналогичным образом, приходим к матрице 3X3 и новому узлу *δ*, объединяющему *α* и *c* (рис. 4).

	α	c	γ
α		63	35
c	63		38
γ	35	38	

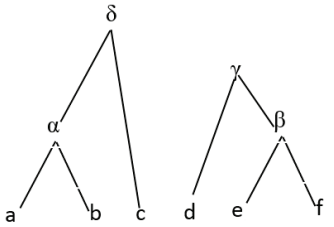


Рис. 4. Матрица расстояний 3X3.
Объединение узлов α, c и возведение их к узлу δ

Наконец, на заключительном этапе объединяются два оставшихся узла δ и γ (рис. 5).

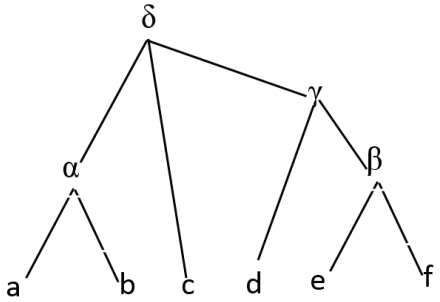


Рис. 5. Построенное дерево списков с архетипом δ

Если представлять группы списков с помощью дерева (одна группа — одна ветвь), то данный метод сопоставим с предыдущим методом (3 кластера: списки ab, c и d, e, f). Но в отличие от метода Колвелла — Алексеева данный метод строит дополнительно промежуточные списки. При этом a priori предполагается, что у любых списков может быть только один антиграф, контаминация не допускается. Также отметим, что данный метод оказывается неустойчивым к начальным данным: если немного «пошевелить» матрицу расстояний, то вид дерева может существенно измениться (антиграф поменяется местами со своим непосредственным потомком).

В некотором смысле кладистические методы являются развитием метода Дееса — Ваттеля. В основе данных методов также лежит предположение о том, что чем больше сходства обнаруживают тексты списков, тем ближе эти списки генеалогически. Используя матрицу расстояний, кладистические методы реконструируют генеалогическое дерево списков. Но в основе подобной реконструкции лежит новый принцип — принцип экономии (парсиномии). Поясним, в чем его суть. Вычислим для каждого дерева длину, слагающуюся из длин его дуг. Длина каждой дуги, соединяющей пару узлов, равна числу чтений, сходных для этих узлов. Из всех деревьев выбираем дерево с максимальной длиной. В этом состоит суть данного принципа — природа эволюционирует из одного состояния в другое за минимальное число шагов. В 1970-е гг. данные подход был применен группой советских ученых во главе с Л. И. Бородкиным и Л. В. Миловым для построения генеалогии списков одного древнерусского юридического памятника[8]. В качестве способа построения генеалогического дерева был выбран алгоритм нахождения покрывающего графа максимальной общей длин[9]. Согласно данному алгоритму все пары списков анализируются в порядке убывания их коэффициентов. На каждом шаге выбирается элемент a_{ij} из матрицы расстояний и соединяются дугой узлы i и j, если это не ведет к появлению цикла. В противном случае переходят к рассмотрению следующего по порядку элемента матрицы. По окончании описанного алгоритма выбирают достаточно «древний» список и «подвязывают» к нему построенное дерево. Ориентированное подобным образом дерево отражает генеалогическую связь между сохранившимися списками произведения.

Автором настоящей работы был предложен новый метод генеалогической классификации[10], опирающийся на применение теории нечетких множеств (fuzzy sets theory)[11]. Суть данного подхода заключается в замене исходной модели объекта (в данном случае текста и истории его бытования) семейством нечетких моделей. Причем выбор конкретной нечеткой модели определяется той степенью точности, с которой необходимо решить поставленную задачу.

II. ЭЛЕМЕНТЫ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ

Нечеткие отношения (fuzzy relations) играют важную роль в теории нечетких множеств. В отличие от классической теории нечеткие отношения предназначены для описания

взаимодействия элементов системы, носящего не дихотомный характер. Другими словами, нечеткие отношения позволяют построить качественную картину процесса или явления, учитывая при этом силу взаимосвязи слагающих его элементов.

Введем ряд необходимых формальных определений. Пусть X — произвольное непустое множество, состоящее из конечного числа элементов.

Определение 1. Нечетким множеством $\tilde{A}$ называется совокупность всех пар вида $\{<\mu_A(x)/x>\}$, где $x \in X$. В качестве аналога такого множества в текстологии может выступать, например, группа близкородственных списков, редакция.

Функцию $\mu_F(x): X \to [0,1]$ называют функцией принадлежности (membership function) нечеткого множества $\tilde{A}$, а X — базовым множеством. Данная функция характеризует степень принадлежности конкретного элемента нечеткому множеству (например, списка к определенной редакции). Чем больше значение функции, тем более достовернее выглядит гипотеза, что соответствующий элемент включен в $\tilde{A}$. При крайних значениях на данной шкале (0 и 1) заключают, что элемент однозначно (четко) не принадлежит/принадлежит нечеткому множеству.

Рассмотрим некоторые основные операции над нечеткими множествами. Пусть заданы нечеткие подмножества $\tilde{A}$ и $\tilde{B}$ множества X. Зафиксируем некоторое значение δ от 0 до 1.

Определение 2. Будем считать, что множество $\tilde{A}$ нечетко включается в $\tilde{B}$, и обозначать $\tilde{A} \subseteq \tilde{B}$, если для любого $x \in X$ выполняются неравенства $(\mu_A(x) \to \mu_B(x)) > \delta$, где

$$\mu_A(x) \to \mu_B(x) \equiv \max\left(\frac{\delta}{1-\delta}(1-\mu_A(x), \mu_B(x)\right).$$

Заметим, что при $\delta = 1/2$ последнее выражение преобразуется в логическую импликацию $\max(1-\mu_A(x), \mu_B(x))$.

Определение 3. Будем считать, что множество $\tilde{A}$ нечетко эквивалентно (равно) множеству $\tilde{B}$ и обозначать $\tilde{A} \approx \tilde{B}$, если справедливы включения $\tilde{A} \subseteq \tilde{B}$ и $\tilde{B} \subseteq \tilde{A}$. Нетрудно показать, что значения функций принадлежности эквивалентных множеств будут одновременно либо меньше порога δ, либо больше (см. приложение 1). Таким образом, в пределе при δ, стремящимся к 0, все нечеткие множества становятся эквивалентными друг другу. И, наоборот, при δ, близком к 1, эквивалентность двух множеств превращается в обычное поэлементное совпадение. Поскольку оба предельных случая с δ малоинтересны для анализа, то обычно выбирают компромиссный вариант между 0,5 и 0,7.

Определение 4. Объединением множеств $\tilde{A}$ и $\tilde{B}$ будем называть нечеткое множество, обозначаемое $\tilde{A} \cup \tilde{B}$ и определяемое как

$$\tilde{A} \cup \tilde{B} = \{<\mu_{A \cup B}(x)/x>\}, x \in X,$$

где $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$.

Определение 5. Пересечением множеств $\tilde{A}$ и $\tilde{B}$ будем называть нечеткое множество, обозначаемое $\tilde{A} \cap \tilde{B}$ и определяемое как

$$\tilde{A} \cap \tilde{B} = \{<\mu_{A \cap B}(x)/x>\}, x \in X,$$

где $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$.

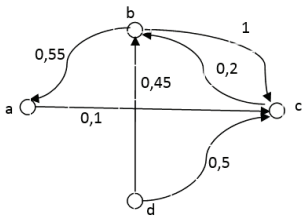
Наряду с нечеткими множествами центральное место в нашем анализе будут играть нечеткие отношения. Нечеткое отношение является частным случаем нечеткого множества в случае, когда базовое множество есть декартово произведение.

Определение 6. Нечетким (нечетким бинарным) отношением на X называется и через $\tilde{\eta} = (X,V)$ обозначается пара множеств X и V, где V является нечетким подмножеством на $X \times X$, т. е.

$$V = \{<\mu_V < x_i, x_j > / < x_i, x_j > \in X \times X\},$$

$$\mu_V < x_i, x_j > \in [0,1].$$

Нечеткое отношение удобно представлять как в виде матрицы, так и в виде ориентированного графа. Ниже приведен пример такого графа, узлами которого являются элементы из X, а ориентированным дугам приписаны соответствующие значения функции μ_V<···> (взвешенные дуги) (рис. 6).



	a	b	c	d
a			0,1	
b	0,55		1	
c		0,2		
d		0,45	0,5	

Рис. 6. Пример графического и матричного представления бинарного нечеткого отношения

Примерами нечетких отношений являются «x приблизительно равно y», «x намного меньше, чем y» на множестве чисел, «x похоже на y», «x не младше, чем y» на множестве людей и т. д. Среди многообразия нечетких отношений выделим в первую очередь два их типа: симметричное и антисимметричное. Первый тип позволяет разбивать исходное базовое множество на семейство экви-

валентных классов (схожих классов), а второй — устанавливать между ними частичный порядок (выделять генеалогопреемственные связи).

Определение 7. Отношение $\tilde{\eta} = (X,V)$ называют нечетко симметричным, если $(\mu_V < x,y > \rightarrow \mu_V < y,x >) \geq \delta$ для любых $x, y \in X$. Граф симметричного отношения обладает следующим свойством: обе дуги каждой пары вершин имеют вес больше или меньше порога. Таким образом, отношение, представленное на рисунке 6, не является симметричным для значения порога 0,5: на графе есть пара вершин a и b, соединенных одной дугой, причем соответствующий вес больше 0,5.

Определение 8. Отношение $\tilde{\eta} = (X,V)$ называется нечетко антисимметричным, если $\neg(\mu_V < x,y > \& \mu_V < y,x >) \geq \delta$ для любых $x, y \in X$. В данном определении используются так называемые логические отрицание $\neg\mu_A(x) = 1 - \mu_A(x)$ и конъюнкция $\mu_A(x) \& \mu_B(x) = \min(\mu_A(x), \mu_B(x))$. Граф антисимметричного отношения обладает следующим свойством: если обе дуги пары имеют вес больше δ, то данные вершины совпадают. Не трудно видеть, что отношение, представленное на рисунке 6, антисимметрично для величины порога 0,5: у каждой пары графа обе дуги имеют веса большие и меньшие 0,5 либо одна из дуг отсутствует.

Определение 9. Отношение $\tilde{\eta} = (X,V)$ называется нечетко связанным, если

$$\max(\mu_A < x,y >, \mu_A < y,x >) \geq \delta$$

для любых $x, y \in X$. Как видно из рисунка 6, данное отношение не является связанным: вершины a и d не соединены, вес дуги, соединяющей *a* с вершиной *c*, ниже порогового значения.

III. НЕЧЕТКАЯ ГЕНЕАЛОГИЧЕСКАЯ КЛАССИФИКАЦИЯ

Как уже указывалось, в общем случае матрица отношений не является ни симметричной, ни антисимметричной. Но, оказывается, существует способ, позволяющий выделить из матрицы как симметричную, так и антисимметричную составляющую[12]. Поскольку соответствующий алгоритм подробно описан в нашей работе[13], прокомментируем его основные этапы:

Алгоритм нечеткой генеалогической классификации

В качестве исходных данных на вход алгоритма поступает матрица нечеткого отношения $\tilde{\eta} = (X,V)$.

1. Транзитивное замыкание

На данном этапе происходит нормализация исходной матрицы, в результате чего отношение $\tilde{\eta} = (X,V)$ становится транзитивным. Транзитивность можно считать аналогией к известному из классической математики неравенству треугольника (сумма длин двух сторон треугольника превосходит длину третьей его стороны). Выполнение требования транзитивности является необходимым для проведения всех последующих операций.

2. Каноническое разбиение на классы эквивалентности

Из отношения $\tilde{\eta} = (X,V)$ выделяется нечеткое отношение $\tilde{\phi} = (X,F)$ следующим образом: $\mu_F < x,y > = \min(\mu_V < x,y >, \mu_V < y,x >)$ для любых $x, y \in X$. Вершины соответствующего графа остаются прежними, а веса дуг становятся одинаковыми для каждой пары. Очевидно, построенное отношение является симметричным, что позволяет разбить его на так называемые классы эквивалентности. А именно, каждому элементу x из базового множества X поставим в соответствие нечеткое множество $\tilde{A}(x) = \{< \mu_{A(x)}(y) / y >\}, y \in X$, где $\mu_{A(x)}(y) = \mu_V < x,y >$. Среди построенных множеств $\tilde{A}(x)$ находим попарно эквивалентные, которые объединяем в одно нечеткое множество A_i, $i = 1, .., m$ (см. определения 3,4). Таким образом, на основе симметричного отношения $\tilde{\phi} = (X,F)$ строится семейство $\Re$ непересекающихся нечетких множеств, на которое распадается базовое множество X: $X = \bigcup_{A_i \in \Re} A_i$ (так называемое каноническое разбиение). Заметим, что таких попарно эквивалентных множеств может и не быть (как в примере, изображенном на рисунке 6). Тогда число m классов эквивалентности в точности совпадает с мощностью базового множества X.

3. Построение нечеткого порядка

Каждому классу A_i поставим в соответствие множество A_i^*, являющееся его δ-сечением:

$$A_i^* = \{< \mu_{A_i^*}(x) / x >\}, \text{где } \mu_{A_i^*}(x) = \begin{cases} 1, & \text{если } \mu_{A_i}(x) \geq \delta \\ 0 & \text{, иначе} \end{cases}.$$

Данное сечение состоит из всех элементов X, на которых соответствующая функция принадлежности принимает значения больше порога. Выберем из четкого множества A_i^* произвольный элемент x_i (эталон). С помощью исходного отношения $\tilde{\phi} = (X,F)$ построим на семействе $\Re$ нечеткое отношение $\tilde{\rho} = (\Re,T)$ следующим образом: для любых двух классов $A_i, A_j \in \Re$ положим $\mu_T < A_i, A_j > = \mu_F < x_i, x_j >$. Оказывается, построенное таким образом отношение $\tilde{\rho} = (\Re,T)$ является антисимметричным (см. определение 8). В терминах теории нечетких множеств данное отношение задает порядок на каноническом разбиении, т. е.

частично упорядочивает классы эквивалентности между собой.

4. Построение нечеткого покрытия множества X из линейно упорядоченных подмножеств

Сначала выделим на X все максимально связные подмножества. Данная операция определяет на каждом из таких подмножеств совершенный порядок (perfect order), что позволяет их линейно упорядочивать. Будем считать, что элемент x нечетко предшествует элементу y, если выполнено соотношение $\mu_F< x, y >\geq \delta$. Интуитивно понятно, что, введя такую меру предшествования, можно упорядочить все элементы базового множества на линейной шкале. Руководствуясь данными соображениями, рассмотрим понятие наименьшего элемента.

Определение 10. Элемент $x \in A$ называется наименьшим, если выполнено неравенство $\mu_F< x, y >\geq \delta$ для любых $y \in A, y \neq x$.

Оказывается, на подмножестве A с совершенным порядком всегда найдется наименьший элемент. Таким образом, определяется следующая процедура линейного упорядочивания. На начальном этапе находим наименьший элемент $x_1 \in A$. Из множества A исключаем элемент x и переходим к рассмотрению множества $X_1 = A\backslash x_1$. Поскольку X_1 содержится в A, то на X_1 также задан совершенный порядок. Отсюда следует, что на X_1 также найдется наименьший элемент $x_2 : \mu_F < x_1, y >\geq \delta$ для любых $y \in X_1, y \neq x_1$. Продолжая аналогичным образом, получаем следующую конечную упорядоченную цепочку: $x = x_0, x_1, x_2, .., x_l$, где l является мощностью множества A. Проводя данную процедуру для всех максимально связных подмножеств, получаем нечеткое покрытие базового множества линейно упорядоченными подмножествами. В графическом представлении данная процедура выражается в выделении ветвей дерева и упорядочивании его узлов (ориентированное дерево). Возвращаясь к примеру на рисунке 6, не трудно видеть, что при величине порога 0,5 множество вершин графа распадается на три связных подмножества: a и b, b и c, d. В результате применения вышеописанной процедуры получаем нечеткое покрытие из трех линейно упорядоченных подмножеств, что отображается в виде следующего дерева (рис. 7).

Заметим, что третье подмножество состоит из одного элемента d, не сравнимого с остальными.

Если искать аналогию предложенному методу в текстологии, то в чем-то его идеи весьма перекликаются с формально-текстологическим подходом, развитым С. А. Бугославским[14] в рамках филологической школы В. Н. Перетца. Идея подхода С. А. Бугославского заключается в восстановлении текста «оригинала» утраченного произведения на основе классификации сохранившихся списков. Причем данная классификация проводится в два этапа.

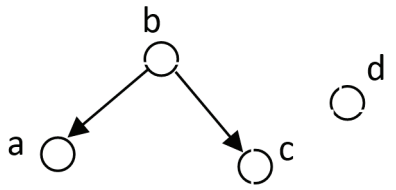


Рис. 7. Дерево нечеткого отношения из рисунка 6

Целью первого этапа классификации является выделение текстологически близких групп списков и установление для каждой из них ее «архетипа». Для этого сначала на основе сравнительного анализа из всех имеющихся списков выбирается самый «типичный». По отношению к типичному чтению все остальные выступают как варианты (разночтения). Собранные варианты и являются материалом для выделения текстологически близких групп. Применяемые при этом приемы исследования базируются на положении о том, что общность ошибок (уклонений от нормальных чтений) свидетельствует об общности происхождения списков (метод «общих ошибок» К. Лахмана)[15]. Архетип каждой родственной группы восстанавливается путем сравнения составляющих ее списков и удаления из каждого списка его индивидуальных чтений, т. е. таких, которые присущи только одному списку.

Целью второго этапа является, как уже было сказано, восстановление текста, максимально близкого к оригиналу произведения. В качестве исходного материала при этом используются «архетипы», полученные на предыдущем этапе. На основе сравнения «архетипов» устанавливается история их взаимоотношения, постепенного накопления в текстах списков изменений, появления редакций. Другими словами, тексты «архетипов» частично упорядочиваются на хронологической шкале.

Сравнивая оба подхода (нечеткий и С. А. Бугославского), отмечаем, что они оба включают в себя два этапы классификации: выделение близких групп (горизонтальное упорядочивание) и установление генеалогопреемственной связи между ними (вертикальное упорядочивание). Также весьма перекликаются роли «эталона» для нечеткого метода и «архетипа» для метода С. А. Бугославского. И тот и другой используются для генеалогического упорядочивания групп. В то же время, в отличие от метода С. А. Бугославского, нечеткий метод учитывает контаминацию списков и позволяет порождать стемму с заданной степенью достоверности.

IV. СРАВНЕНИЕ РЕЗУЛЬТАТОВ РАБОТЫ МЕТОДОВ КОЛВЕЛЛА — ТЬЮНА, ДЕЕСА — ВАТТЕЛЯ И НЕЧЕТКОЙ ГЕНЕАЛОГИЧЕСКОЙ КЛАССИФИКАЦИИ

На основе предложенного алгоритма автором настоящей работы на языке Delphi XE8 (2015) была разработана программа нечеткой генеалогической классификации. Входным параметром данной программы является порог уверенности δ, что позволяет проводить нечеткую классификацию с различной степенью детализации.

Метод Колвелла — Тьюна. В случае с примером Э. Колвелла (см. табл.) интерес представляет диапазон от 65 до 95%. При значении порога меньше 65% все списки (14) включаются в один нечеткий класс. И, наоборот, начиная с 95% все сформировавшиеся классы распадаются на отдельные списки. Первым из одного общего класса выделяется список Θ. За ним из того же класса при величине порога 70÷73% выделяются списки D, P^{45}, P^{66} и 565, что отражено на рисунке 8.

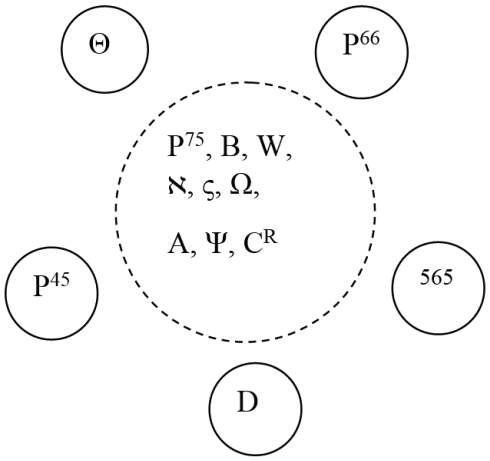


Рис. 8. Нечеткая генеалогическая классификация из примера таблицы (уровень достоверности 73%)

Таким образом, метод нечеткой классификации подтверждает вывод Э. Колвелла о нестабильности третьего и четвертого кластеров (Δ и Γ text-type). При величине порога 74% общий класс разбивается на два класса (B и A-type согласно терминологии Э. Колвелла). Так продолжается до величины порога 76%, когда начинают распадаться оба класса (рис. 9).

На уровне порога 80% процесс дальнейшего выделения новых классов прекращается. Образовавшиеся ранее 3 нечетких класса (списки ς и Ω; W и א; P^{75} и B) сохраняются до величины порога 85%. Далее начинается распад всех классов на отдельные элементы-списки (рис. 10).

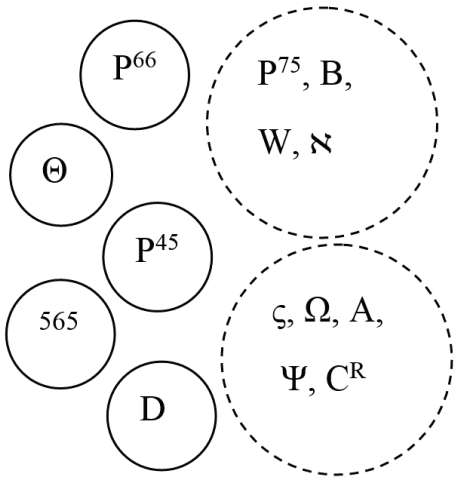


Рис. 9. Нечеткая генеалогическая классификация из примера таблицы (уровень достоверности 74–76%)

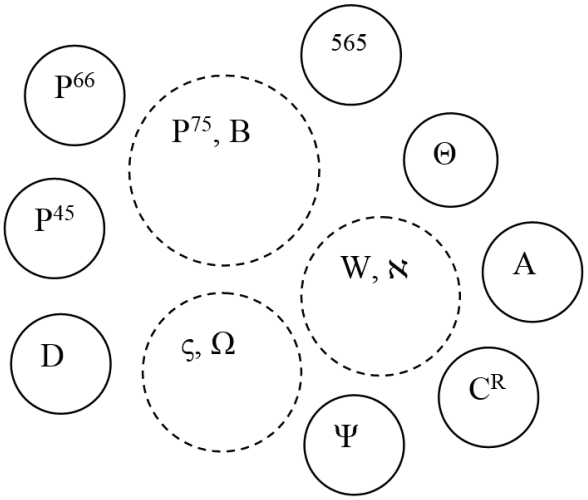


Рис. 10. Нечеткая генеалогическая классификация из примера таблицы (уровень достоверности 80–85%)

Заметим, что в данном примере нечеткая классификация проводится только на «горизонтальном» уровне, списки разбиваются на текстологически близкие группы. Данный факт не вызывает удивления, поскольку исходная матрица является симметричной (см. табл.). Схожий анализ можно провести и для примера, иллюстрирующего метод Дееса — Ваттеля (см. рис. 1).

Метод Дееса — Ваттеля. При величине порога 56% один нечеткий класс, включающий все 6 списков, распадается на два класса. Первый класс включает списки d, e и f, а второй — списки (рис. 11).

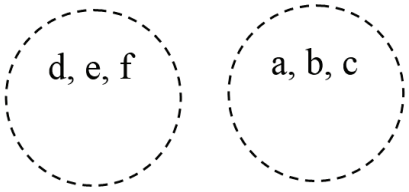


Рис. 11. Нечеткая генеалогическая классификация из примера на рисунке 1 (уровень достоверности 56–65%)

Начиная с уровня 66% из второго класса выделяется список c, а с уровня 81% из первого класса — список d (рис. 12).

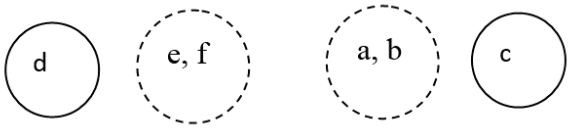


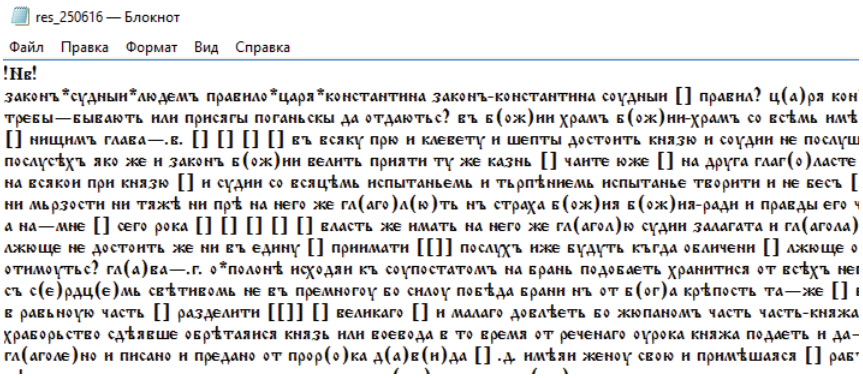
Рис. 12. Нечеткая генеалогическая классификация из примера на рисунке 1 (уровень достоверности 81–85%)

Наиболее стабильным оказывается класс, состоящий из списков a и b. Он сохраняется до уровня 92%, в то время как класс из списков e и f распадается при величине порога 86%. Таким образом, и в данном случае нечеткая классификация дает схожие результаты, что и метод Дееса-Ваттеля.

V. РАБОТА МЕТОДА НЕЧЕТКОЙ КЛАССИФИКАЦИИ НА МАТЕРИАЛЕ «ЗАКОНА СУДНОГО ЛЮДЕМ»

Как известно, Закон Судный людем (ЗСЛ)[16] является древнейшим памятником славянского права и одним из самых древних текстов на славянском языке[17]. Текст оригинала не сохранился. На данный момент автору настоящей работы известно о 57-ми списках краткой редакции ЗСЛ. Проиллюстрируем материале этих списков основные этапы работы метода нечеткой классификации.

В качестве входных данных служит файл с предобработанными текстами списков: каждому чтению соответствует свой узел (разночтения) в тексте. Количество узлов разночтения во всех списках одно и то же (фрагмент данного файла приведен на рисунке 13).


res_250616 — Блокнот

Файл Правка Формат Вид Справка

!Нв!

законъ*судныи*людемъ правило*царя*константина законъ-константина соудныи [] правил? ц(а)ря кон
требы—бываютъ или присягы поганьскы да отдаютьс? въ б(ож)ии храмъ б(ож)ин-храмъ со всѣмь имѣ
[] нищимъ глава—.в. [] [] [] [] въ всяку прю и клеветү и шепты достонть князю и соудни не послуш
послуслѣхъ яко же и законъ б(ож)ии велитъ прияти тү же казнь [] чанте юже [] на друга глаг(о)ласте
на всякои при князю [] и судии со всяцѣмь испытаньемь и тьрпѣниемь испытанье творити и не бесъ [
ни мьрзости ни тяжѣ ни прѣ на него же гл(аго)л(ю)ть нъ страха б(ож)ня б(ож)ня-ради и правды его ч
а на—мне [] сего рока [] [] [] [] [] власть же имать на него же гл(агол)ю судии залагата и гл(агола)
лжюще не достонть же ни въ единү [] приимати [[]] послухъ иже будуть къгда обличени [] лжюще о
отимоуться? гл(а)ва—.г. о*полонѣ исходян къ соупостатомъ на брань подобаеть хранитися от всѣхъ не
съ с(є)рдц(є)мь свѣтивомь не въ премногоү бо силоү побѣда брани нъ от б(ог)а крѣпость та—же [] и
в равьноую часть [] разделити [[]] [] великаго [] и малаго довлѣеть бо жюпаномъ часть часть-княжа
храборьство сдѣявше обрѣтаяися князь или воевода в то время от реченаго оурока княжа подаеть и да-
гл(аголе)но и писано и предано от прор(о)ка д(а)в(и)да [] .д. имѣян женоү свою и примѣшаяся [] раб

Рис. 13. Входной файл программы нечеткой классификации (фрагмент Новгородского списка (Нв))

Вначале программа проводит сличение текстов списков и выделяет все имеющиеся разночтения. Результаты данного этапа выводятся в формате Excel-файла (см. рис. 14).

В первом столбце данной таблицы отражены номера узлов разночтений (всего в тексте около двух тысяч), во втором — все чтения (около 6700). В четвертом столбце приведены собственно чтения.

Далее эксперту необходимо заполнить данную таблицу: в третьем столбце проставить тип разночтения (в данном случае это будет натуральное число от 0 до 14; более подробно о типах ошибочных чтений см.[18]). Самая трудоемкая часть работы эксперта — это заполнение таблицы начиная с пятого столбца: в каждую клетку заносится число из натурального ряда, отражающее информацию о том, является ли данное чтение нормальным для списка (определяемого по столбцу) или нет. В дальнейшем будут учитываться только ошибочные чтения (соответствующее число не равно нулю). Конечно, подобный анализ невозможно провести во всех случаях. Но при этом необходимо помнить, что чем больше будет выделено ошибочных чтений,

тем полнее и актуальнее получится генеалогическая классификация на выходе программы. Фрагмент заполненной таким образом таблицы приведен на рисунке 15.

	A	B	C	D	E	F	G	H	I	J
1					Нв	Ч	ГП	З	Пчт	РМ
2	1	1		законъ*сѹдныи*людемъ						
3	2	2		правило*царя*константина						
4	2	1975		[]						
5	3	3		законъ-константина						
6	3	1976		[[]]						
7	3	5064		константина-законъ						
8	4	4		соудныи						
9	4	1977		сѹдныи						
10	4	4599		соудны						
11	4	6019		сѹд?ныи						
12	5	5		[]						
13	5	3782		всем						
14	6	6		правил?						
15	6	1978		[]						
16	7	7		ц(а)ря						
17	7	1979		[]						

Рис. 14. Таблица разночтений списков
(фрагмент Excel-файла)

	A	B	C	D	E	F	G	H	I
1					Нв	Ч	ГП	З	Пчт
2	1	1	0	законъ*сѹдныи*людемъ	0	0	0	0	0
3	2	2	8	правило*царя*константина	1	1	1	1	1
4	2	1975	8	[]	0	0	0	0	0
5	3	3	9	законъ-константина	1	1	1	1	1
6	3	1976	4	[[]]	0	0	0	0	0
7	3	5064	9	константина-законъ	2	2	2	2	2
8	4	4	11	соудныи	1	0	0	0	0
9	4	1977	11	сѹдныи	0	1	1	1	1
10	4	4599	11	соудны	1	0	0	0	0
11	4	6019	11	сѹд?ныи	0	1	1	1	1
12	5	5	5	[]	0	0	0	0	0
13	5	3782	5	всем	1	1	1	1	1
14	6	6	5	правил?	1	1	1	1	1
15	6	1978	5	[]	0	0	0	0	0
16	7	7	5	ц(а)ря	1	1	1	1	1
17	7	1979	5	[]	0	0	0	0	0
18	7	5065	5	ц(е)с(а)ря	1	1	1	1	1

Рис. 15. Заполненная таблица разночтений списков
(фрагмент Excel-файла)

На следующем этапе задается входной параметр программы — уровень уверенности (надежности классификации) в виде числа от нуля до ста. На рисунке 16 приведен интерфейс программы с уровнем, равным 50.

Далее работает собственно программа нечеткой классификации. Все вычисления отражаются в специальных текстовых файлах, а получающаяся стемма выводится в графическом режиме (см. рис. 17). При этом необходимо по-

мнить, что результат классификации обусловлен также значениям весовых коэффициентов, соответствующих типам разночтений (подробнее см. работу[19]).

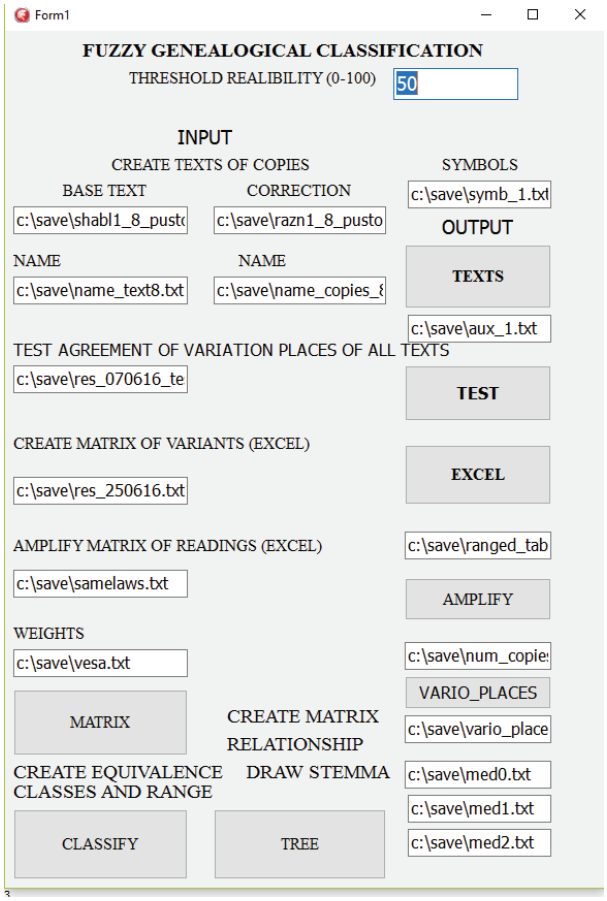


Рис. 16. Интерфейс программы нечеткой классификации
(пороговое значение равно 50)

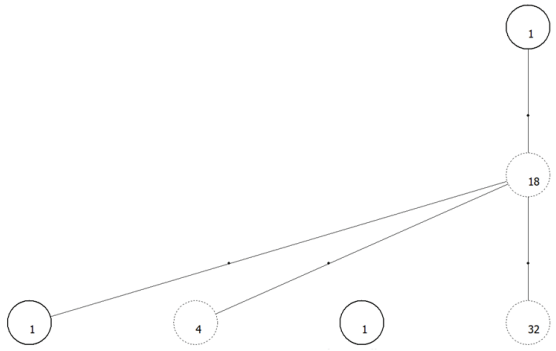


Рис. 17. Стемма списков ЗСЛ
(пороговое значение равно 50)

На данном рисунке образованные нечеткие классы (группы текстологически родственных списков) отражены в виде кружков. Соединяющие кружки дуги отражают генеалогопреемственные связи между классами и их направление. Числа в кружке показывают количество списков, включенных в соответствующий класс.

Заметим, что чем большее значение имеет данный порог, тем детальнее становится полученная стемма (из прежних классов выделяются новые, между ними возникают свои связи и т. д.). Вариант стеммы при пороговом значении, равном 60, приведен на рисунке 18.

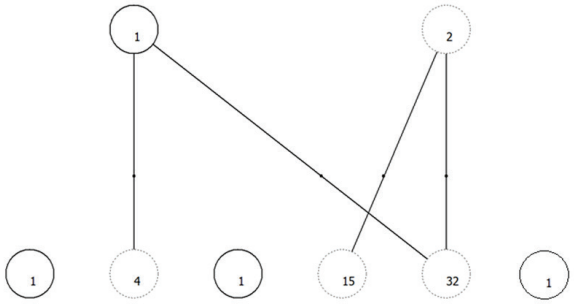


Рис. 18. Стемма списков ЗСЛ
(пороговое значение равно 60)

В заключение автор приносит большую благодарность за внимание и ценные советы Г. С. Баранковой, ведущему научному сотруднику Института русского языка имени В. В. Виноградова и профессору Л. И. Бородкину.

Приложение 1

Утверждение 1. Зафиксируем некоторое пороговое значение $0 \le \delta \le 1$ и рассмотрим два нечетких множества $\tilde{A}$ и $\tilde{B}$. Тогда $\tilde{A} \approx \tilde{B}$ тогда и только тогда, когда выполнены одновременно оба неравенства $\mu_A(x) \le \delta$, $\mu_B(x) \le \delta$ либо $\mu_A(x) \ge \delta$, $\mu_B(x) \ge \delta$ для любого $x \in X$.

Доказательство. Предположим обратное, то есть $\mu_A(x) \le \delta$ и $\mu_B(x) > \delta$. Следовательно, $1-\mu_B(x) < 1-\delta$ или $\frac{\delta}{1-\delta}(1-\mu_B(x)) < \delta$. Два множества $\tilde{A}$ и $\tilde{B}$ являются эквивалентными, если выполняются оба неравенства

$$\max\left(\frac{\delta}{1-\delta}(1-\mu_A(x)), \mu_B(x)\right) > \delta$$

$$\text{и } \max\left(\frac{\delta}{1-\delta}(1-\mu_B(x)), \mu_A(x)\right) > \delta$$

для любого $x \in X$ (см. определение 3). Тогда из второго неравенства следует, что должно выполняться неравенство $\mu_A(x) > \delta$. Полученное противоречие доказывает исходное утверждение.

ПРИМЕЧАНИЯ

1 Прикладная и компьютерная лингвистика / И. С. Николаев, О. В. Митренина, Т. М. Ландо. М., 2016. 320 с.

2 Colwell E. C. and Tune E. W. The quantitative relationship between MS text-types // Biblical and patristic studies in memory of R. P. Casey / Ed. by J. N. Birdsall and P. W. Thomson. Freiburg, 1963. P. 25–32.

3 Алексеев А. А. Текстология славянской библии. СПб., 1999. 190 с.

4 Colwell E. C. and Tune E. W. Ibid.

5 Dees A. Sur une constellation de quatre manuscits // Melanges de linguistique et de literature offertes a Lein Geschiere. Amsterdam, 1975. P. 1–9.

6 Wattel E. Constructing Initial Binary Trees in Stemmatology // Studies in Stemmatology II / Ed. By Reenen P. van, Mulken M. van. Amsterdam, 2004. P. 145–166.

7 Мандель И. Д. Кластерный анализ. М., 1988. 176 с.

8 Бородкин Л. И., Морозова Л. Е. Опыт использования математических моделей и ЭВМ в текстологических исследованиях // Количественные методы в гуманитарных науках. М., 1981. С. 55–66.

9 Басакер Р., Саати Т. Конечные графы и сети. М., 1974. 368 с.

10 Шпирко С. В. Применение теории нечетких множеств к задаче генеалогической классификации в текстологическом исследовании // Историческая информатика: Информационные технологии и математические методы в исторических исследованиях. 2013. № 3. С. 39–51.

11 Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М., 1976. 165 с.

12 Мелихов А. Н., Бернштейн Л. С., Коровин С. Я. Ситуационные советующие системы с нечеткой логикой. М., 1990. 272 с.

13 Шпирко С. В. Указ. соч.

14 Бугославский С. А. Текстология Древней Руси. Т. 2: Древнерусские литературные произведения о Борисе и Глебе. М., 2007. 670 с.

15 Лихачев Д. С. Текстология. М. ; Л., 1983. 640 с.

16 Закон Судный людем краткой редакции / М. Н. Тихомиров, Л. В. Милов. М., 1961. 177 с.

17 Максимович К. А. Закон Судный людем. Источниковедческие и лингвистические аспекты исследования славянского юридического памятника. М., 2004. 240 с.

18 Шпирко С. В. Построение матрицы нечеткого отношения для задачи текстологической генеалогической классификации (на материале «Закона Судного людем») // Исторический журнал: научные исследования. 2017. № 1. С. 22–35.

19 Там же.