Metric of Hierarchical Choices and Possibilities of its Application in Computational Linguistics

Bērziņš A. A.
Riga Technical University
Riga, Latvia
ansis@latnet.lv

Abstract

We are defining a new distance – the distance of hierarchical choices, and proofing that it's a metric. It is defined on a space of results of hierarchical clustering, i.e., binary trees with labelled leafs and unlabelled inner nodes, so-called dendrograms. This new metric can be used in many domains of computational linguistics, for example, for expert evaluation of idiomometric results.

Keywords: metric; distance; clustering; categorisation; hierarchy; dendrogram; tree; graph; binary; expert evaluation

DOI: 10.28995/2075-7182-2021-20-1031-1040

Метрика иерархического выбора и возможности её применения в компьютерной лингвистике

Берзинь А. У.
Рижский технический университет
Рига, Латвия
ansis@latnet.lv

Аннотация

В докладе рассказывается о новой, предложенной нами метрике иерархического выбора, которая может применяться для оценки расстояния между результатами иерархической категоризации, т.е., бинарными деревьями с мечеными листьями и немечеными внутренними узлами, так называемыми дендрограммами. Эта новая метрика может применяться при решении многих задач компьютерной лингвистики, например, при экспертной оценке идиомометричеких результатов.

Ключевые слова: метрика; расстояние; категоризация; иерархия; дендрограмма; дерево; граф; бинарный; экспертная оценка

1 Категоризация

Иерархическая категоризация (или кластеризация) является классическим способом упорядочения, который достаточно широко использовался уже в «докомпьютерном веку»[1], с применением в различных областях.

Например, при решении определённых филологических задач в области идиомометрии (лингвометрии, диалектометрии) [20] [21], абсолютные величины расстояний не суть их

[1] *The classical theory comes with two general principles of organization for categories: hierarchical categorization and cross-categorization.*

HIERARCHICAL CATEGORIZATION: A partition of a category into subcategories such that all members are in one, and only one, subcategory.

Biological taxonomies are common examples. For instance, we view animals as being elephants, raccoons, tigers, or the like. These group together into larger categories (mammals, etc.) and can be split up into smaller categories (e.g., various kinds of elephants). [8]

характеристики – более воспринимаемыми являются результаты проведённой по ним категоризации. Кроме того, иерархическая категоризация применяется к данным атмосферных наук, в частности – климатических, метеорологических и загрязнения воздуха [5][2]; измерениям качества мощности [7] и генерированию распределения энергии [9][3]; региональным описательным данным [6][4] и даже аэрофотосъёмке [24][5]; текстовым данным и документам [10]; биометрическим данным [18][6]; данным микрочипов в биостатистике [3][7]; данным социального поведения, например, производительности студентов [15]; и т.д.

Ознакомимся с алгоритмом агломеративной иерархической категоризации:

1. Пусть у нас имеются n объектов, которые мы хотим распределить по категориям, а также задано расстояние d между объектами.
2. Создадим n начальных одноэлементных категорий – для каждого объекта по одной, и определим расстояния между ними как расстояния между составляющими их объектами.
3. Найдём две наиближайшие (с наименьшим расстоянием) категории и объединим их в одну так, что новая будет содержать все элементы обеих. Таким образом количество категорий уменьшится на одну.
4. Рассчитаем расстояния между новой и остальными категориями по формуле:
 а) в случае метода наименьшего расстояния:
$$\min\{ d(x,y): x \in A, y \in B\}$$
 б) в случае метода наибольшего расстояния:
$$\max\{ d(x,y): x \in A, y \in B\}$$
 в) в случае метода среднего расстояния:
$$\frac{1}{\mathrm{car}\, d(A)\cdot \mathrm{car}\, d(B)} \sum_{x\in A}\sum_{y\in B} d(x,y)$$

По сути, результат такой категоризации – это помеченное двоичное дерево, при чём помеченными являются лишь все его листья, остальные узлы и все дуги являются немеченными.

---

2 *Over the past 50 years, clustering has been widely applied to atmospheric science data in particular, climate and meteorological data. Since the 1980's, air pollution studies began employing clustering techniques, and has since been successful .. In particular, two well known and commonly used clustering methods i.e. k-means and hierarchical agglomerative, that have been applied in air pollution studies have been reviewed.* [5]

3 *Distribution Generation is expected to become more important in future generation system. In distribution system, due to load uncertainties the load exceeds the generating capacity which leads to power loss and unreliable operation of the system. To overcome this problem DG units are incorporated into the distribution system to meet the excess demand which results in power loss minimization, improvement of voltage profile, power quality improvement, reliable operation, etc., clustering techniques are proposed .. for optimal placement of DG.* [9]

4 *Numerical agglomerative hierarchical classification is fundamentally an unsupervised method of grouping individuals on which there are multivariate data so as to identify natural groups in them and perhaps in the populations from which they are drawn and where no prior classification exists or is assumed. We have used the technique to make a tectonic regionalization of the Zagros region and to see whether it can increase our understanding of the regional tectonics. We first identified 137 sub-areas as units for each of which we had recorded 18 quantitative variables; these formed our data ..* [6]

5 *.. метод позволяет осуществлять адекватную поддержку технологии распознавания реалистических картографических изображений в геоинформационных системах с применением категоризации. .. для каждой формируемой категории учитывается комбинированная база прототипов, расширяемая с использованием простых критериев (цвета, текстуры, контуров). Введение этапа категоризации в видеосистеме позволяет упростить задачу распознавания объектов.* [22]

6 *In data analysis, the hierarchical clustering algorithms are powerful tools allowing to identify natural clusters, often without any priori information of the data structure, and are quite often used because provide a graphical representation of the resulting partitions, a hierarchy or dendrogram, revealing more information than non-hierarchical algorithms that returns a unique partition. Moreover, it is not necessary specify the number of clusters à priori. .. we do a comparative study using a set of experiments using two-dimensional synthetics and real-world data sets, based on the biometrics of the hands.* [17]

7 *Agglomerative (bottom-up) hierarchical clustering algorithms are important analysis tools for microarray data. They are useful for organizing the genes and samples from a set of microarray experiments so as to reveal biologically interesting patterns. Hierarchical methods are especially useful, because they enable analysts to simultaneously examine groupings of the data into many small clusters (e.g. repeat samples from the same patient) and a few large clusters (e.g. different prognosis groups).* [3]

Рисунок 1: Пример результатов агломеративной иерархической категоризации идиомов.

Алгоритм дивизивной иерархической категоризации приводит к результатам того же типа – меченым двоичным деревьям, называемаем дендрограммами. В свою очередь, неиерархическая категоризация приводит к результатам другого вида.

2 Экспертная оценка

В некоторых областях компьютерной лингвистики имеется т.н. «золотой стандарт», см., напр., [16]. Анализ диалектальных речевых фонограмм к таковым не относится. Да и разработать такой стандарт было бы практически невозможно и с точки зрения объёма труда, и с точки зрения ответственности. Посему для оценки диалектометрических результатов данного вида требуется идти по другому пути – относительно-сравнительного анализа.

Рассматриваемые нами методы разрабатывались для расчёта расстояний между идиомами, и эти числовые расстояния являются выходными данными этих методов. Однако оценивать результаты данного типа слишком сложно: чтобы сравнить их между собой, нужно было бы их каким-то (неизвестным нам) способом нормировать, потому что эти расстояния относительны, т.е., абсолютные числовые значения различны для каждого метода, а – для сравнения с экспертной оценкой – эксперты должны бы были взять на себя числовое определение соответствующих расстояний, что вряд ли возможно.

Значит, единственный выход – это просить экпертов (вместо числовой оценки расстояний) вручную составлять деревья подобные тем, которые образуются в итоге категоризации, т. е., в соответствии со своими знаниями и чутьём на каждом поочерёдном шагу выбирать ближайшую пару.

Сперва мы решили рассмотреть известные методы экспертной оценки (МЭО) и оценить их применяемость к нашим данным. В разных источниках их число колеблется от 4 до 7[8], но все они

[8] *Наиболее распространенными методами упорядочения альтернатив (факторов) являются: 1) ранжирование; 2) непосредственная оценка; 3) последовательнои сравнение и 4) парное сравнение.* [22]

Окончательная количественная оценка определяется с помощью четырех основных методов экспертных оценок и множества их разновидностей:

1) метод простой ранжировки (или метод предпочтения);

2) метод задания весовых коэффициентов;

3) метод парных сравнений;

4) метод последовательных сравнений.

Метод простой ранжировки заключается в том, что каждого эксперта просят расположить признаки в порядке предпочтения. Цифрой один обозначается наиболее важный признак, цифрой два – следующий за ним по важности и т.д. полученные данные сводятся в следующую таблицу. ...

Метод задания весовых коэффициентов заключается в присвоении всем признакам весовых коэффициентов. Весовые коэффициенты могут быть проставлены двумя способами:

ориентированы на упорядочение элементов на одной оси, попарное сравнение, ранжирование или оценку в количественных единицах. Наши входные данные, т.е., результат нашей иерархической категоризации – двоичные деревья, при чём с метками на всех листьях, но без таковых на внутренних узлах, не соответствуют ни одному из указанных типов входных данных, поэтому ни один из описанных методов к ним не применим. Это привело нас к выводу, что метод экспертной оценки придётся разрабатывать самим.

3 Метрика

Очевидно, дабы таковой МЭО создать, нам необходим какой-то критерий оценки: т.е., расстояние, в лучшем случае – метрика, заданная на пространстве наших обьектов – деревьев, созданных

---

1) всем признакам назначают весовые коэффициенты так, чтобы суммы коэффициентов была равна какому-то фиксированному числу (например, единице, десяти или ста);

2) наиболее важному из всех признаков придают весовой коэффициент, равный какому-то фиксированному числу, а всем остальным – коэффициенты, равные долям этого числа. ...

Метод последовательных сравнений заключается в следующем:

1) эксперт упорядочивает все признаки в порядке уменьшения их значимости: $A_1>A_2>...>A_n$;

2) присваивает первому признаку значение, равное единице: $A_1=1$, остальным же признакам назначает весовые коэффициенты в долях единицы;

3) сравнивает значение первого признака с суммой всех последующих.

Возможны три варианта: $A_1 > A_2 + A_3 + \ldots + A_n$

$A_1 = A_2 + A_3 + \ldots + A_n$

$A_1 < A_2 + A_3 + \ldots + A_n$

Эксперт выбирает наиболее соответствующий, по его мнению, вариант и приводит в соответствие с ним оценку первого события;

1) сравнивает значение первого признака с суммой всех последующих за вычетом самого последнего признака.

Приводит оценку первого признака в соответствие с выбранным из трех вариантов неравенством:

$A_1 > A_2 + A_3 + \ldots + A_{n-1}$

$A_1 = A_2 + A_3 + \ldots + A_{n-1}$

$A_1 < A_2 + A_3 + \ldots + A_{n-1}$

2) процедура повторяется до сравнения A_1 *с* $A_2 + A_3$.

После того как эксперт уточнил оценку первого признака в соответствии с выбранным им неравенством из трех возможных:

$A_1 > A_2 + A_3$

$A_1 = A_2 + A_3$

$A_1 < A_2 + A_3$

он переходит к уточнению оценки второго признака A_2 *по той же схеме, что и в случае первого, т.е. сравнивается оценка второго признака с суммой последующих. ...*

Метод парных сравнений. Согласно ему все признаки попарно сравниваются между собой. На основании парных сравнений путем дальнейшей обработки находятся затем оценки каждого признака. [23]

Vērtēšanas uzdevumos apakšmērķi var būt:

1) labākā objekta noskaidrošana (šāda nostāja derīga arī lemšanas uzdevumiem);

2) dažu labāko objektu noskaidrošana, salīdzinot ar citiem;

3) objektu ranžēšana, resp., sakārtošana rindā pēc kāda izvēlēta kritērija. Parasti labākais objekts dabū vietu ar mazāku kārtas numuru, piem., pirmo vietu, otro vietu (attiecīgi rangu 1, rangu 2, u.t.t.);

4) objektu novērtēšana nosacītās vienībās (ballēs) jeb citiem vārdiem, kvantitatīvu rādītāju piešķiršana objektiem. [11]

Известны следующие методы экспертных оценок:

Метод ассоциаций. Основан на изучении схожего по свойствам объекта с другим объектом.

Метод парных (бинарных) сравнений. Основан на сопоставлении экспертом альтернативных вариантов, из которых надо выбрать наиболее предпочтительные.

Метод векторов предпочтений. Эксперт анализирует весь набор альтернативных вариантов и выбирает наиболее предпочтительные.

Метод фокальных объектов. Основан на перенесении признаков случайно отобранных аналогов на исследуемый объект.

Индивидуальный экспертный опрос. Опрос в форме интервью или в виде анализа экспертных оценок. Означает беседу заказчика с экспертом, в ходе которой заказчик ставит перед экспертом вопросы, ответы на которые значимы для достижения программных целей. Анализ экспертных оценок предполагает индивидуальное заполнение экспертом разработанного заказчиком формуляра, по результатам которого производится всесторонний анализ проблемной ситуации и выявляются возможные пути её решения. Свои соображения эксперт выносит в виде отдельного документа.

Метод средней точки. Формулируются два альтернативных варианта решения, один из которых менее предпочтителен. После этого эксперту необходимо подобрать третий альтернативный вариант, оценка которого расположена между значений первой и второй альтернативы. [27]

путём иерархической категоризации. Следует отметить, что вычисление расстояний между объектами данного типа не является слишком распространённой задачей, однако отдельные попытки имели место быть, посему в мировом контексте имеются некоторое предложение соответствующих метрик.

Наиболее распространённым является т.н. расстояние вращения – оно задаётся как наименьшее количество вращений (ротаций), необходимых для получения из одного дерева другого дерева[9]. Существуют также разработанные предложения алгоритмов расчёта данного расстояния[10]. И хотя в основном описана работа с деревьями без меток, утверждается, что переход на деревья с метками сути не изменит. Однако способ определения самих этих вращений противоречит природе объектов, которые описываем мы: например, получается, что одна операция вращения может создать (в нашем понимании, т. е., исходя из свойств наших объектов) принципиально другое дерево, которое не соответствует такому близкому расстоянию (в понимании расстояния вращений) (см. рисунок 2).

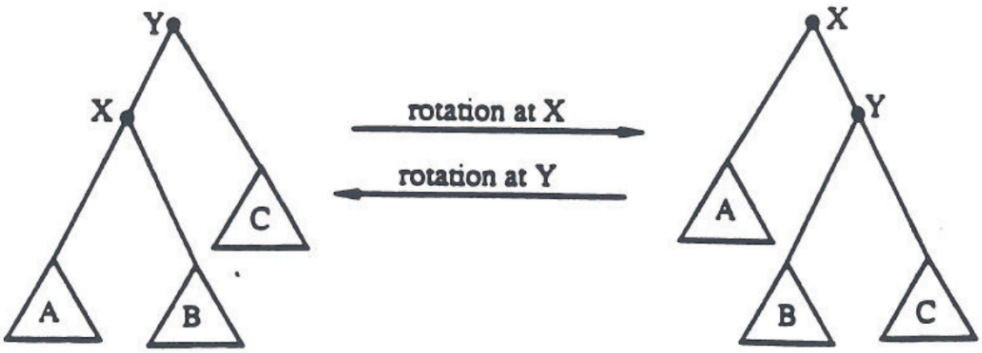

Рисунок 2: Схематическое определение расстояния вращений. Треугольниками обозначены поддеревья. Отображённое дерево может являться частью большего дерева. [17]

Мы нашли ещё одно, менее известное расстояние, которое называется расстоянием АКМ[11]. В нём нас не устроил сам подход, что в нижних ветвях различия имеют меньший вес, потому что в нашем случае важны все шаги выбора, независимо от того, когда они выполняются.

---

[9] *A rotation is an operation that changes one binary tree into another. In a tree of size n there are n — 1 possible rotations, one corresponding to each nonroot internal node. Figure 1 shows the general rotation rule and the effect of a particular rotation on a particular tree. The rotation corresponding to a node changes the structure of the tree near that node, but leaves the structure elsewhere intact. A rotation maintains the symmetric order of the nodes, but changes the depths of some of them. Rotations are the primitives used by most schemes that maintain "balance" in binary trees.*

A rotation is an invertible operation; that is, if tree T can be changed into T' by a rotation, then T' can be changed back into T by a rotation. The rotation graph for trees of size n (denoted RG(n)) is an undirected graph with one vertex for each tree of size n and an edge between vertices T and T' if there is a rotation that changes T into T'.

Any binary tree of size n can be converted into any other by performing an appropriate sequence of rotations. Therefore the rotation graph is connected. We can define the rotation distance between two trees as the length of the shortest path in the rotation graph between the two trees, i.e. the minimum number of rotations required to convert one tree into the other. [17]

[10] *Tree rotations (left and right) are basic local deformations allowing to transform between two unlabeled binary trees of the same size. Hence, there is a natural problem of practically finding such transformation path with low number of rotations, the optimal minimal number is called the rotation distance. Such distance could be used for instance to quantify similarity between two trees for various machine learning problems, for example to compare hierarchical clusterings or arbitrarily chosen spanning trees of two graphs, like in SMILES notation popular for describing chemical molecules.*

There will be presented inexpensive practical greedy algorithm for finding a short rotation path, optimality of which has still to be determined. It uses introduced partial order for binary trees of the same size: t1 ≤ t2 if t2 can be obtained from t1 by a sequence of only right rotations. [4]

There are a number of ways of measuring the difference in shape between two rooted binary trees with the same number of leaves. Pallo (Computer Journal, 9, 171–175, 1986) introduced a left weight sequence, which is a sequence of positive integers, to characterize the structure of a binary tree. By applying the AVL tree transformation on binary trees, we develop an algorithm for the efficient transformation of the left weight sequences between two binary trees. [2]

[11] *ρ i(A) is the i'th ancestor of A, such that ρ0(A) is A itself, p1(A) is the parent of A, and ρ2(A) is the grandparent of A, etc. We use ρ(A) as shorthand for ρ1(A).*

ch(A) is the set of children of A.

Таким образом мы пришли к выводу, что ни одна из найденных нами метрик не соответствует нашей задаче. Поэтому надо искать подход с другой стороны – попытаться самим задать метрику, которая бы численно характеризовала суть взаимоотношений наших объектов.

Что же является тем отличием, которое характерно нашим оценкам – как автоматизированным, так и мануальным? В принципе, это выбор, который делается на каждом шаге путём нахождения ближайшей пары. Итак, если дан набор из *n* языков, то созданные деревья можно описать также как *n*-элементные наборы, состоящие из вновь созданных (или выбранных) пар, которые уже в следующем шаге являются элементами выбора. Скажем, набор из 5 языков (*Auleja, Baļtinova, Dundag, Rudzātys, Vileks*) мы можем иерархически категоризировать, например, так:

1-й шаг: выбираем ближайшую пару (*Baļtinova, Vileks*).
2-й шаг: выбираем ближайшую пару (*Rudzātys, Auleja*).
3-й шаг: выбираем ближайшую пару ((*Baļtinova, Vileks*), (*Rudzātys, Auleja*)).
4-й шаг: выбираем ближайшую пару (((*Baļtinova, Vileks*), (*Rudzātys, Auleja*)), *Dundag*).

...или эдак:
1-й шаг: выбираем ближайшую пару (*Baļtinova, Vileks*).
2-й шаг: выбираем ближайшую пару ((*Baļtinova, Vileks*), (*Rudzātys*)).
3-й шаг: выбираем ближайшую пару (*Dundag, Auleja*).
4-й шаг: выбираем ближайшую пару (((*Baļtinova, Vileks*), *Rudzātys), (Auleja*, *Dundag)*).

В 1-м случае, технически стерев скобки и пробелы, а также добавив пустое множество, потому что до 1-го шага множество наших выборов было пустым, мы получаем такое множество подмножеств нашего набора:

{Ø, {*Baļtinova, Vileks*}, {*Rudzātys, Auleja*}, {*Baļtinova, Vileks, Rudzātys, Auleja*}, {*Baļtinova, Vileks, Rudzātys, Auleja, Dundag*}}

В свою очередь, во 2-м случае:

{Ø, {*Baļtinova, Vileks*}, {*Baļtinova, Vileks*, *Rudzātys*}, {*Dundag, Auleja*}, {*Baļtinova, Vileks, Rudzātys, Auleja, Dundag*}}

Так как первое и последнее подмножества тривиальны (соответственно, пустое и полное), то они не зависят от выбора, поэтому их сократим:

{{*Baļtinova, Vileks*}, {*Rudzātys, Auleja*}, {*Baļtinova, Vileks, Rudzātys, Auleja*}}

...и:

{{*Baļtinova, Vileks*}, {*Baļtinova, Vileks*, *Rudzātys*}, {*Dundag, Auleja*}}

...

π(A, B) is the set of edges on the path between A and B. Notice that for any edge, we always write the parent node first. For example, for the tree in Figure 1, we have π(U, W) = {(T,U),(S,T),(S,W)}.

...

In the following we propose a new distance measure, the AKM distance, that satisfies the seven properties mentioned in Section 4.2.

We define the AKM distance as a distance measure between nodes in a label tree:

$$d_{AKM}(A,B) = \sum_{(X,Y)\in\pi(A,B)} w(X,Y)$$

where

$$w(X,Y) = \begin{cases} \dfrac{1}{\log|ch(X)|+1} & \text{if\_X\_is\_root} \\ \dfrac{1}{\tau}\cdot\dfrac{w(\rho(X),X)}{\log|ch(X)|+1} & \text{otherwise} \end{cases}$$

This implies that dissimilarities at lower levels in the tree are deemed less significant for values of τ greater than 1. Also, due to the term log|ch(X)|+1, the distance between two siblings decreases logarithmically as more siblings are added. This prevents nodes with many children having a deciding impact on the weight between two nodes. [1]

Итак мы получаем *n-2 = 3*-элементные множества подмножеств, которые характеризуют пошаговый выбор иерархической категоризации, значит отношение несоответствий к *n-2* может достаточно хорошо характеризовать то, что желаем.

Определение 1. Расстоянием иерархического выбора *назовём деление объёма разницы между множествами множеств элементов всех нетривиальных выборов иерархической категоризации на количество нетривиальных выборов.*

Из этого определения также следует, что *n>2*, так как только в таком случае нетривиальные выборы существуют. При *n=2* тривиальны оба выбора, при *n=1* тривиален единственный выбор – пустое множество, при *n=0* выборов вообще нет (*n*, конечно, неотрицательно).

Так как по определению множества порядок элементов множества не имеет значения, то расстояние иерархического выбора от первого примера до второго будет делением

|{{Baļtinova, Vileks}, {Rudzātys, Auleja}, {Baļtinova, Vileks, Rudzātys, Auleja}} \ {{Baļtinova, Vileks}, {Baļtinova, Vileks, Rudzātys}, {Dundag, Auleja }}| = |{Rudzātys, Auleja}, {Baļtinova, Vileks, Rudzātys, Auleja}}| = 2 на объём разницы аргументов, итого 2/3.

Обозначим расстояние иерархического выбора *hi(X, Y)*, где *X* и *Y* – множества, на которых оно определено.

Лемма 1. *Расстояние иерархического выбора симметрично.*

Доказательство. Пусть *A* и *B* – любые пространства, для которых расстояние иерархического выбора определено. В обоих множествах – аргументах расстояния – одинаковое количество элементов: *n-2*. Обозначим *m* число совпадающих элементов. Тогда, отняв от первого множества второе, объём результата будет равен объёму первого множества |*A*| минус число совпадающих элементов: *n-2-m.* И расстояние иерархического выбора будет его деление на *n-2*: $hi(A,B) = \frac{n-2-m}{n-2}$. В свою очередь, от второго множества отняв первое, объём результата будет равен объёму второго множества |*B*| минус число совпадающих элементов: *n-2-m*. И расстояние: $hi(B,A) = \frac{n-2-m}{n-2}$. Значит расстояние в обоих направлениях совпадает, т.е., является симметричным.

Лемма доказана. ■

Лемма 2. *Расстояние иерархического выбора треугольно (соответствует аксиоме треугольника).*

Доказательство. Мы должны доказать, что $hi(A,B) + hi(B,C) \geq hi(A,C)$ для любых множеств A, B и C, для которых оно определено.

Так как расстояние иерархического выбора может быть определено только для множеств одинакового объёма, то и знаменатель деления *n-2* у всех будет одинаков. Так как *n-2* всегда положительно, то доказуемое выражение эквивалентно выражению $|A\backslash B| + |B\backslash C| \geq |A\backslash C|$.

Для большей наглядности обозначим d_{XY} число несовпадающих элементов во множествах X и Y. Тогда, фактически, $d_{XY} := |X\backslash Y| = |Y\backslash X|$

В случае трёх множеств одинакового объёма (в нашем случае – объём *k:=n-1*), максимальное возможное *d* третьей пары множеств является функция от *d* двух остальных пар множеств:

$d^{max}_{AC} = min(k, d_{AB} + d_{BC})$

Если у «среднего» множества *B* с одним множеством *A* не совпадают d_{AB} элементов, а с другим множеством *C* – d_{BC} элементов, тогда у множеств *A* и *C* не могут совпасть больше чем $d_{AB}+d_{BC}$ элемента, ибо, если бы совпадало большее количество, тогда из этого бы автоматически следовало, что и у множества *B* с каким-то из остальных множеств совпадает больше элементов. В свою очередь, если $d_{AB}+d_{BC}$ превышает максимальное количество элементов, то, конечно, число *k* ограничивает максимальное несовпадение раньше, чем сумма.

Из этого следует, что $|A\backslash B| + |B\backslash C| = d_{AB} + d_{BC} \geq min(k, d_{AB} + d_{BC}) = d^{max}_{AC} \geq |A\backslash C|$

Лемма доказана. ■

Лемма 3. *Расстояние иерархического выбора недегенеративно.*

Доказательство. *hi(A,A)* будет *0* $\forall A$, ибо, ежели все элементы совпадают, тогда количество несовпадающих равно нулю.

В свою очередь, если *hi(A,B)=0*, тогда все элементы *A* и *B* совпадают, из чего следует, что совпадают и сами множества *A* и *B*.

Лемма доказана. ■

Теорема 1. *Расстояние иерархического выбора является метрикой.*

Доказательство. Метрикой называют такое неотрицательное расстояние, которое соответствует всем трём аксиомам метрики[12].

В соответствии с определением 1, расстояние иерархического выбора неотрицательно, так как оно задано как дробь, числитель которой неотрицателен, а знаменатель – положителен.

В соответствии с леммой 1 для расстояния иерархического выбора в силе аксиома симметрии.

В соответствии с леммой 2 для расстояния иерархического выбора в силе аксиома треугольника.

В соответствии с леммой 3 для расстояния иерархического выбора в силе аксиома недегенеративности (идентичности).

Теорема доказана. ■

4 Применение в компьютерной лингвистике

В начале статьи мы перечислили несколько областей применения иерархической категоризации – и лингвистических, и нелингвистических. Нас самих к данному вопросу привели **задачи идиометрические**, т.е., по заданию и вычислению расстояний между человеческими идиомами. Как правило, подобные оценки расстояний не абсолютны, а относительны, т.е., между собой могут сравниваться не числовые значения расстояний, а только некие структуры отношений, которые эти расстояния задают. Вот тут и приходит на выручку иерархическая категоризация, и получается, что структуры на выходе – дендрограммы. В таком случае при экспертной оценке эксперты также рисуют дендрограммы на заданных комплектах идиомов, и после этого надо численно оценить, т.е., «измерить» расстояния между дендрограммами экспертными и дендрограммами, полученными путём автоматизированного расчёта расстояний между идиомами.

Во-вторых, иерархическая категоризация используется в **статистических парсерах**, так как достаточно качественно описать синтактическую и семантическую взаимозаменяемость более простыми структурами не представляется возможным[13]. В данном случае проведение экспертной оценки всего прокатегоризированного лексикона вряд ли возможно ввиду его объёма, однако она может быть проведена на отдельных выборках, что скорее всего создаст представление о качестве построенных деревьев взаимозаменяемости.

---

[12] *МЕТРИКА, расстояние на множестве X, – определенная на декартовом произведении X×X функция ρ, с неотрицательными действительными значениями, удовлетворяющая при любых x, y* $\in X$ *условиям:*

1) ρ(x,y)=0 тогда и только тогда, когда x=y (аксиома тождества);

2) ρ(x,y)+ρ(y,z)≥ ρ(x,z) (аксиома треугольника);

3) ρ(x,y)=ρ(y,x) (аксиома симметрии).

Множество X, на к-ром может быть введена М., наз. метризуемым. Множество X, наделенное некоторой М., наз. метрическим пространством. [25]

[13] *An ideal type of clusters for NLP is the one which guarantees mutual substitutability, in terms of both syntactic and semantic soundness, among words in the same class.*

... In this paper we adopt the merging approach and propose an improved method of constructing hierarchical clustering.

... This algorithm produces a balanced binary tree representation of words in which those words which are close in meaning or syntactic feature come close in position. [19]

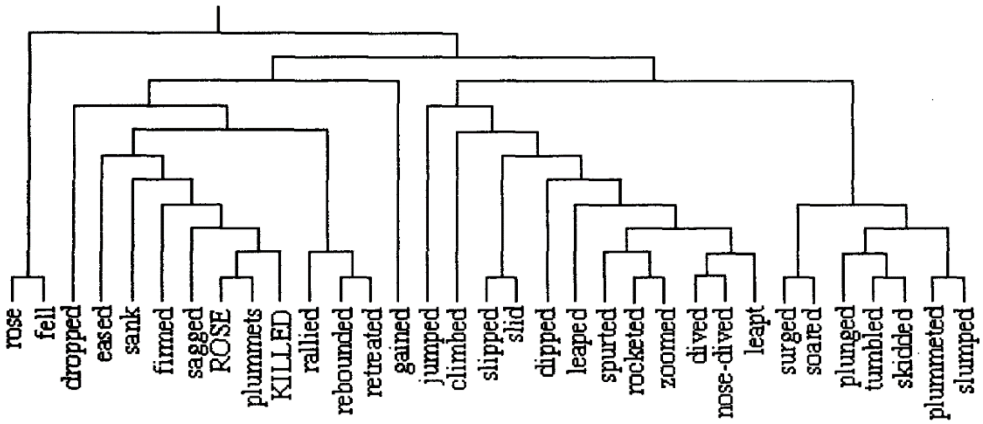

Рисунок 3: Пример результатов иерархической категоризации лексикона. [19]

Иерархическая категоризация применяется и в **семантической типологии**, т.е., области, занимающейся описанием сходств и различий семантической организации разных единиц: идиомов (языков), лингвистических свойств, стимулирующих воздействий, информантов и т. п. Матрицы сходства путём аггломеративной иерархической категоризации проецируются в дендрограммы[14], к которым уже применима наша метрика.

5 Выводы

Нам удалось задать метрику на пространстве результатов иерархической категоризации. Данная метрика по своей конструкции соответствует интуитивному пониманию о близости или отдалённости рассматриваемых нами обьектов, поэтому экпериментально численно проверять её соответствие мы, строго говоря, не должны. Достаточно того, что мы математически доказали, что она является метрикой. Воспользовавшись ею, мы, во-первых, сможем оценить уровень компетентности экспертов: есть несколько способов[15], как это делается, но в нашем случае реально применим только один – оценка компетентности, основываясь на данных самого опроса, именуемый также определением степени единодушия или консенсуса[16].

Во-вторых, мы ею сможем оценить соответствие автоматизированных оценок экспертной оценке, скорее всего – средней (тех экспертов, которые будут признаны компетентными). [17]

Итак, мы создали новый метод экспертной оценки для результатов иерархической категоризации, имеющих довольно широкое применение и в компьютерной лингвистике, и в других областях науки.

---

[14] *Figure 3 shows a dendrogram of the similarity matrix underlying Figure 2 created by a Hierarchical Cluster Analysis using an agglomerative clustering method. Such a dendrogram can be interpreted as representing the results of a series of successive splits (when read top-down) or lumps (bottom-up) of categories. Information reduction in this case is the result of forcing the similarity matrix, not into a small number of dimensions, but into a series of binary choices.* [13]

[15] *Методы определения степени компетентности экспертов принято делить на такие группы ... :*

- *оценка компетентности экспертов в зависимости от их оценки объектов;*
- *взаимооценка;*
- *самооценка;*
- *оценка по объективным документальным данным об эксперте.*

Кратко рассмотрим некоторые методы, относящиеся к указанным группам. Компетентность эксперта, например, определяют в зависимости от того, насколько его оценки согласованы с оценками большинства. Однако в том случае, когда относительную значимость некоторого множества альтернатив оценивают, к примеру, т экспертов, из которых т-1 экспертов совершенно некомпетентны в рассматриваемом вопросе, а один является высококвалифицированным специалистом, то их оценки с высокой степенью вероятности могут сильно отличаться друг от друга. [26]

[16] *Literatūrā ir sastopamas metodes, kas tīri formāli ļauj uzlabot ekspertu vienprātības pakāpi. Viena no tām ... paredz, ka no kopējās ranžējumu tabulas ... pēc kārtas izslēdz kāda eksperta doto ranžējumu.* [11]

[17] На момент подачи публикации мы указанные эксперименты уже провели (на данных латышских говоров) и получили неплохие результаты, которое здесь не описываем, ибо это выходит за рамки заданной темы.

Литература

[1] *Caspersen K.M., Madsen M.B., Eriksen A.B., Thiesson B.* A Hierarchical Tree Distance Measure for Classification. // Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods. Porto: "SCITEPRESS", 2017.

[2] *Chen Y.J., Chang J.M., Wang Y.L.* An efficient algorithm for estimating rotation distance between two binary trees. // International Journal of Computer Mathematics. Vol. 82, No. 9, September 2005, "Taylor & Francis", 2005.

[3] *Chipman H., Tibshirani R.* Hybrid hierarchical clustering with applications to microarray data // Biostatistics. Volume 7, Issue 2. Oxford: Oxford University Press, 2006.

[4] *Duda J.* Practical estimation of rotation distance and induced partial order for binary trees. arXiv:1610.06023 [ocs.DS]. Cornell University, 2016.

[5] *Govender P., Sivakumar V.* Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019) // Atmospheric Pollution Research. Volume 11, Issue 1. İstanbul: Hava kirlenmesi araştirmalari ve denetimi Türk milli komitesi, 2020.

[6] *Hashemi S.N., Mehdizadeh R.* Application of hierarchical clustering technique for numerical tectonic regionalization of the Zagros region (Iran) // Earth Science Informatics. Volume 8. Springer, 2015.

[7] *Jasiński M., Sikorski T., Leonowicz Z., Borkowski K., Jasińska E.* The Application of Hierarchical Clustering to Power Quality Measurements in an Electrical Power Network with Distributed Generation // Signal Analysis in Power Systems. Energie, 13(9):2407. Basel: MDPI, 2020.

[8] *Lakoff G.* Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: The University of Chicago Press, 1987.

[9] *Karunakar Jureedi N.V.V., Rosalina K.M., Prema Kumar N.* Clustering Analysis and its Application in Electrical Distribution System // International Journal of Recent Advances in Engineering & Technology (IJRAET). Bhubaneswar: IRD India, 2020.

[10] *Kuang D., Park H.* Fast rank-2 nonnegative matrix factorization for hierarchical document clustering // Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Association for Computing Machinery, 2013.

[11] *Markovičs Z.* Ekspertu novērtējuma metodes. Rīga: RTU izdevniecība, 2009.

[12] *Mayer-Schönberger V., Cukier K.* Big Data: A Revolution that Will Transform how We Live, Work, and Think. New York: An Eamon Dolan Book | Houghton Mifflin Harcourt, 2013.

[13] *Moore R., Donelson K., Eggleston A., Bohnemeyer J.* Semantic typology: New approaches to crosslinguistic variation in language and cognition. De Gruyter, April 13, 2015. *https://doi.org/10.1515/lingvan-2015-1004*

[14] *O'Neil C., Schutt R.* Doing Data Science. Sebastopol: O'Reilly Media, 2019.

[15] *Rana S., Garg R.* Application of Hierarchical Clustering Algorithm to Evaluate Students Performance of an Institute // Second International Conference on Computational Intelligence & Communication Technology (CICT). Ghaziabad: ABES Eng. Coll., 2016.

[16] Similarity (State of the art). // ACL Wiki for Computational Linguistics. The Association for Computational Linguistics. *https://aclweb.org/aclwiki/Similarity_(State_of_the_art)* Число доступа: 6.XI.2019.

[17] *Sleator D.D., Tarjan R.E., Thurston W.P.* Rotation Distance, Triangulations, and Hyperbolic Geometry. // Journal Of The American Mathematical Society. Volume 1. Number 3. July 1988.

[18] *Sousa L., Gama J.* The Application of Hierarchical Clustering Algorithms for Recognition Using Biometrics of the Hand // International Journal of Advanced Engineering Research and Science (IJAERS). Volume 1, Issue 7. Jaipur: AI Publications, 2014.

[19] *Ushioda A.* Hierarchical Clustering of Words and Application to NLP Tasks // Proceedings of the Fourth Workshop on Very Large Corpora, 1996.

[20] *Берзинь А.У.* Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». М.: Издательство РГГУ, 2006.

[21] *Берзинь А.У.* Применение распознавателей фонем для автоматического определения уровня близости языков // Труды международной конференции «Диалог 2016», 2016.

[22] *Бешелев С.Д., Гурвич Ф.Г.* Экспертные оценки. М.: «Наука», 1973.

[23] *Громова Н.М., Громова Н.И.* Основы экономического прогнозирования. Учебное пособие. Старая Русса: Старорусский политехнический колледж, 2007.

[24] *Дудинова О.Б.* Метод категорийной классификации объектов при компьютерном анализе аэроснимков // Системи обробки інформації. Випуск 7 (144). Харків: Харківський університет Повітряних Сил імені Івана Кожедуба, 2016.

[25] Метрика // Математическая энциклопедия. – Т. 3 – М.: «Советская энциклопедия», 1982.

[26] *Полегенько А.Ф., Князский О.В.* Оценка относительной компетентности экспертов в экспертной группе с использованием матриц парных сравнений // Озброєння та військова техніка. № 3. Київ: Центр. НДІ озброєння та військ. техніки ЗС України, 2014.

[27] Экспертное оценивание. *http://ru.wikipedia.org/wiki/Экспертное_оценивание* Число доступа: 31.X.2019.