УДК 519.2

СТЕПЕННОЙ ЗАКОН ЦИПФА И ПРИНЦИП ДИССИММЕТРИИ СИСТЕМЫ

Грушева Д.А.
Научный руководитель - доцент Золожук П.А.

Лесосибирский педагогический институт филиал Сибирского Федерального Университета

Одним из наиболее удивительных примеров степенных законов в гуманитарных науках может служить *закон Ципфа*, определяющий зависимость между *рангом слова* и *частотой слова* для многих натуральных языков. (Под *словом ранга* **r** понимается слово, стоящее на г-м месте в списке слов данного языка, расположенных в порядке убывания частоты их употребления.) Приводимый в литературе вывод закона Ципфа не удовлетворяет нас, по крайней мере по двум причинам.

Во-первых, понятие «энергии» (сложности) и «закона сохранения» вводится ad hoc. Во-вторых, приведенные выше рассуждения не объясняют того замечательного факта, что каждый текст с ципфовским распределением численностей классов обязательно содержит большое количество одноэлементных классов. В приводимых далее рассуждениях нам удалось отказаться от соображений, связанных с использованием внешних по отношению к системе «лимитирующих факторов», «энергии» и проч. Во всяком случае, мы теперь можем не рассматривать эти понятия как исходные. Вместо этого предлагается выбрать в качестве представления системы (текста) множество M не с одним, а с двумя определенным образом согласованными отношениями эквивалентности. Разбиение, связанное с дополнительным отношением эквивалентности, мы назовем коразбиением множества A (относительно разбиения A) .

Определение. Пусть $A = \{X_1, X_2,\dots , X_i, \dots , X_N\}$ - разбиение M. Разбиение $A^* = \{У_1, У_2,\dots, Y_j, \dots , У_p\}$ является коразбиением к разбиению A, если:

1) любое пересечение $X_i \cap Y_j$ содержит не более одного элемента;

2) из $X_i \cap Y_j \neq \emptyset$ следует $X_k \cap Y_j \neq \emptyset$ для всех классов X_k разбиения A таких, что $n_k \geq n_i$.

На рис. 1 разбиение и коразбиение представлено в виде диаграммы: столбцы соответствуют классам разбиения, а строчки - коразбиения. Эту диаграмму можно интерпретировать, конечно, и как гистограмму распределения численностей классов разбиения и коразбиения.

Следующие почти тривиальные утверждения, которые мы приводим без доказательств, проясняют свойства коразбиения и связь его с исходным разбиением.

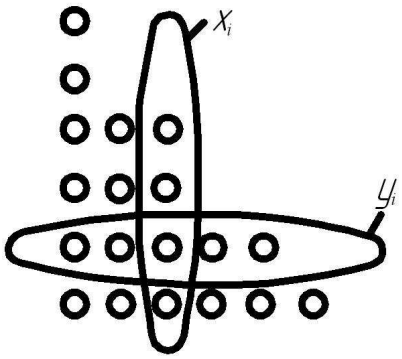

Рисунок 1

Утверждение 1. Если A^*- коразбиение относительно А, то никакое разбиение A_1, содержащее меньшее число классов по сравнению с A^*, не является коразбиением относительно А.

Действительно, каждый класс Y_j из A^* пересекается с самым крупным классом X_1 из А. Следовательно число классов в любом коразбиении A^* не более n_1. С другой стороны число классов в A^* не может быть меньше n_1, так как любые два элемента из X_1 должны содержаться в разных классах A^*.

Утверждение 2. Все коразбиения A^* данного разбиения А изоморфны, и общее число коразбиений составляет:

$$\prod_{i=1}^{n} n_i.$$

Утверждение 3. Если В – коразбиение относительно А, то А – коразбиение относительно В.

Отсюда из утверждения 1 в частности следует, что разбиение $(A^*)^*$ изоморфно разбиению А.

Утверждение 4. Число классов различной мощности в A^* совпадает с числом классов различной мощности в А (хотя общее число классов в А и A^* может не совпадать!).

Утверждение 5. Коразбиение для разбиения на N классов содержит наибольший класс из N элементов.

Рассмотрим теперь некоторое состояние t системы Т, т.е. множество М с разбиением А. Для каждого разбиения А существует коразбиение A^*. Подчеркнем, что все коразбиения определяются самим исходным разбиением и достаточно рассматривать только одно какое-нибудь коразбиение, так как все они изоморфны. Поскольку статус коразбнения A^* в сущности такой же, как и статус разбиения (утверждение 3), для коразбиения естественно выбрать такую же характеристику, как и для исходного разбиения-$H(A^*)$. Сформулируем принцип максимума диссимметрии системы как условие минимальности величины:

$$Ф=H(A)H(A^*). \quad (6)$$

Заметим, что минимум этой величины будет достигаться уже не на таком тривиальном разбиении, как минимум Н (А). При разбиении на одноэлементные множества велик будет член $H(A^*)$, так как все элементы множества М попадут в один класс A^*, и число автоморфизмов A^* составит L!. Собственно принцип максимума диссимметрии системы состоит в том, что наиболее вероятным (или эталонным) состоянием системы считается то, где достигает минимума величина $Ф^*$. Стоит сделать следующее замечание. Минимум величины Н(А) определяет наиболее вероятное состояние системы просто потому, что минимуму Н(А) соответствует максимальный

статистический вес данного состояния в силу равенства $P(A) = \frac{L!}{H(A)}$.

Требование минимальности Ф равносильно требованию максимального статистического веса (вероятности реализации) разбиения А при условии максимального статистического веса, сопряженного с данным разбиением коразбиения A^*. Опять-таки, не будем здесь останавливаться на содержательных мотивировках требования, а перейдем к получению из него нужных следствий. Займемся теперь исследованием разбиения А, для которого Ф минимально. Найти такое разбиение можно чисто комбинаторными методами. Решение этой задачи не сложно, но довольно громоздко. Поэтому мы займемся им позднее, а сначала убедимся, что идем в правильном направлении, решив более простую задачу, которая, однако, должна иметь ответ, близкий к ответу на нашу основную задачу.

Для этого, как и в п. 3, заменим в выражении для H(A) факториалы на более удобную для аналитического исследования функцию n^n. Заметим, что отыскание минимума функционала $\widetilde{\Phi} = \prod_{i=1}^{N} n_i^{n_i} \prod_{i=1}^{P} m_j^{m_j}$ эквивалентно отысканию минимума его логарифма при условии (1):

$$ln\widetilde{\Phi} = ln\widetilde{H}(A) + lnH(A^*) = \sum_{i=1}^{N} n_i ln n_i + \sum_{j=1}^{P} m_j ln m_j. \qquad (7)$$

Выражение (7), в свою очередь, имеет смысл суммы энтропий двух разбиений - А и А. Энтропия является, конечно, вполне естественной мерой диссимметрии разбиения. Однако найти минимум функционала (7) аналитическими методами затруднительно, так как величины n_i и m_j не независимы. Поэтому, прежде всего, мы отыщем связь между n_i и m_j и введем новые, уже независимые переменные. Определим убывающий ряд чисел v_r, пробегающий ровно по одному разу те же самые значения, что и n_i Аналогично определим μ_r - только этот ряд чисел расположим в порядке возрастания. Такая нумерация приведет к более простым расчетным формулам. Обозначим через $P_,$ число классов в разбиении А, содержащих по v_r элементов - кратность v_r. Аналогичную величину для коразбиения A^* обозначим через q_r. С помощью P_r и q_r легко выразить n_i и m_j через v_r и μ_r, а именно:

$$n_i = v_r \; при \sum_{l<r} p_l < i \le \sum_{l \le r} p_l$$

$$m_j = \mu_r \; при \sum_{l<r} q_l < j \le \sum_{l \le r} q_l.$$

При этом мы полагаем p_o и q_o равными нулю. Отсюда
(полагая $\mu_0 = v_{R+1} = 0$) имеем:
$p_1 = \mu_1 - \mu_0$; $p_2 = \mu_2 - \mu_1$; … ; $p_r = \mu_r - \mu_{r-1}$; …
$q_1 = v_1 - v_2$; $q_2 = v_2 - v_3$; … ; $q_r = v_r - v_{r+1}$; …
В новых, уже независимых, переменных (7) записывается следующим образом:

$$ln\widetilde{\Phi} = \sum_{r=1}^{R} (\mu_r - \mu_{r-1}) v_r ln v_r + \sum_{r=1}^{R} (v_r - v_{r+1}) \mu_r ln \mu_r \qquad (8)$$

Представляя H(A) и $H(A^*)$ в этой форме, мы использовали утверждение 4, из которого видно, что число классов разной мощности в А и A^* совпадает.

С помощью новых переменных выразим также объем текста:

$$\sum_{r=1}^{R} (\mu_r - \mu_{r-1}) v_r = \sum_{r=1}^{R} (v_r - v_{r+1}) \mu_r = L \qquad (9)$$

Заметим, что условия (8) и (9) являются единственными ограничениями на переменные v_r и μ_r; величину R можно варьировать в процессе отыскания минимума.

Минимум (8) при условии (9) будем искать методом неопределенных множителей Лагранжа:

$$S = \sum_{r=1}^{R}(\mu_r - \mu_{r-1}) \cdot \upsilon_r \ln\upsilon_r + \sum_{r=1}^{R}(\upsilon_r - \upsilon_{r-1}) \cdot \mu_r \ln\mu_r + \lambda\sum_{r=1}^{R}(\mu_r - \mu_{r-1}) \cdot \upsilon_r + \lambda\sum_{r=1}^{R}(\upsilon_r - \upsilon_{r-1}) \cdot \mu_r.$$

Беря частные производные $\frac{\partial S}{\partial v_r}$ и $\frac{\partial S}{\partial \mu_r}$ приравнивая результаты нулю, получаем два типа равенств:

$$v_r\ln v_r - v_{r+1}\ln v_{r+1} + (v_r - v_{r+1})(\ln\mu_r + 1) + 2\lambda(v_r - v_{r+1}) = 0 \quad (10)$$

$$(\mu_r - \mu_{r+1})(\ln v_r + 1) + \mu_r\ln\mu_r - \mu_{r-1}\ln\mu_{r-1} + 2\lambda(\mu_r - \mu_{r-1}) = 0 \quad (11)$$

Преобразуя (10), имеем:

$$\frac{v_r ln v_r - v_{r+1} ln v_{r+1}}{v_r - v_{r+1}} + ln\mu_r + 1 + 2\lambda = 0,$$

Или, приближенно:

$\ln v_r + 1 + \ln\mu_r + 1 + 2\lambda = 0$,

то есть

$$v_r\mu_r = e^{-2(\lambda+1)} = \text{const}. \quad (12)$$

Уравнение (11) приводит к совершенно таким же результатам.

Соотношение (12) - это и есть закон Ципфа, так как $\mu_r = \sum_{i=1}^{r} p_i$, то есть μ_r равна сумме кратностей всех v_i вплоть до v_r-рангу v_r тогда как v_r - численность класса с данным рангом.