

І н ф о р м а т и з а ц і я т а і н ф о р м а ц і й н і т е х н о л о г і ї

УДК 681.3: 519.8

ЛАНДЕ Д.В., доктор технічних наук, старший науковий співробітник,
завідувач відділом Інституту проблем реєстрації
інформації НАН України

ФУРАШЕВ В.М., кандидат технічних наук, старший науковий співробітник,
заступник директора Науково-дослідного інституту
інформатики і права НАПрН України

**ІНДЕКСИ МЕДІАПРИСУТНОСТІ І ПОРІВНЯННЯ
В ІНФОРМАЦІЙНОМУ ПРОСТОРІ**

Анотація. Про застосування індексів медіаприсутності і порівняння при дослідженнях в інформаційному просторі.

Ключові слова: індекси порівняння, індекси медіаприсутності, інформаційний простір, контент-моніторинг, монополізм.

Аннотация. О применении индексов медиаприсутствия и сравнения при исследованиях в информационном пространстве

Ключевые слова: индексы сравнения, индексы медиаприсутствия, информационное пространство, контент-мониторинг, монополизм.

Summary. About application of indexes of media presence and comparison at researches in information space.

Keywords: indexes of comparison, indexes of media presence, information space, content-monitoring, monopolism.

Постановка проблеми. Під час проведення аналітичних досліджень велику увагу приділяють присутності в інформаційному просторі організацій, брендів, політичних партій, інших об'єктів. Для незалежної цифрової оцінки рівня присутності об'єктів в інформаційному просторі пропонується застосовувати підходи, що застосовуються в економіці при визначенні економічного монополізму, які базуються на дослідженнях Лоренця, Джині, Гувера, Гейла тощо.

Метою роботи є дослідження методів оцінки рівня присутності об'єктів в інформаційному просторі.

Виклад основних положень. Зупинимось на змісті зазначених індексів більш детально, маючи на увазі той факт, що там, де в “чистій” економіці застосовують слово “прибутки”, будемо вважати доцільним використання терміну “рівень медіаприсутності”.

Крива Лоренця [1] – це графічне зображення функції кумулятивного розподілу. Вона була запропонована американським економістом Максом Отто Лоренцом в 1905 році як показник нерівності в доходах населення. Крива Лоренця – це представлення функції кумулятивного розподілу, в якому акумулюються долі медіаприсутності. У прямокутній системі координат крива Лоренця є опуклою вниз і проходить під діагоналлю одиничного квадрата, розташованого в I координатних чвертях.

У разі рівного розподілу медіаприсутності кожна група об'єктів зустрічається в інформаційному просторі кількість разів, пропорційну своїй чисельності. Такий випадок описується кривою рівності (line of perfect equality), прямої, що сполучає початок

координат і точку (1;1). У разі повної нерівності (коли лише один об’єкт представлений в інформаційному просторі) крива (line of perfect inequality) спочатку “прилипає” до осі абсцис, а потім з точки (1;0) “злітає” до точки (1;1). Крива Лоренця розташована між кривими рівності і нерівності.

З кривої Лоренця можна вивести кількісні показники нерівності, наприклад коефіцієнт Джіні та індекс Гувера (який ще називають індексом Робін Гуда).

Коефіцієнт Джіні [2] – це статистичний показник, що свідчить про ступінь розшарування об’єктів по відношенню до якої-небудь ознаки, що вивчається (наприклад, по рівню представлення в інформаційному просторі).

Ця статистична модель була запропонована і розроблена італійським статистиком і демографом Коррадо Джіні і опублікована в 1912 році в його праці “Варіативність і мінливість ознаки”.

Коефіцієнт Джіні розраховується як відношення площі фігури, утвореної кривою Лоренця і кривою рівності, до площі трикутника, утвореного кривими рівності і нерівності. У разі повної рівності коефіцієнт буде рівний 0; у разі повної нерівності він буде рівний 1. Коефіцієнт можна розрахувати по формулі:

$$G = \frac{\sum_{i=1}^N \sum_{j=1}^N |y_i - y_j|}{2N^2 \bar{y}},$$

де: G коефіцієнт Джіні N – число об’єктів;

y_k – частка ознак об’єкту k ;

\bar{y} – середнє арифметичне часткою значень ознаки.

Застосування коефіцієнта Джіні надає можливості:

- порівнювати розподіл ознаки в сукупностях з різним числом одиниць;
- може використовуватися для порівняння розподілу ознаки між різними сукупностями об’єктів;
- дозволяє відстежувати динаміку нерівномірності розподілу ознаки в сукупності на різних етапах динаміки розвитку.

До недоліку коефіцієнта Джіні слід віднести його деяку неінваріантність щодо обсягів сукупності, що досліджується. Так, чим на більшу кількість груп поділена одна і та ж сукупність, тим вище для неї значення коефіцієнта Джіні.

Індекс Робін Гуда (Robin Hood index), також відомий як індекс Гувера (Hoover index) [3], – це ще один показник нерівності по параметрах, що має зв’язок з кривою Лоренця. Він рівний тій частці параметра в сукупності об’єктів, яку необхідно перерозподілити для досягнення рівності. Графічно він може бути представлений як щонайдовший вертикальний відрізок, що сполучає фактичну криву Лоренця з лінією рівності (бісектрисою I координатних чвертей).

Індекс Гувера також належить напіввідкритому інтервалу [0;1]. Якщо ж рівень присутності в інформаційному просторі неподільний до нескінченності, то говорять про частку, перерозподіл якої максимально наближає дану сукупність об’єктів до рівності.

Індекси Тейла [4], як і коефіцієнт Джіні, можуть приймати значення від 0 (повна рівність) до 1 (повна нерівність). Чим вищий індекс, тим більше нерівномірним є розподіл витрат (або доходу).

Індекси Тейла відносяться до більш загального класу “показників загальної ентропії”.

Індекс обчислюється таким чином:

$$T = \frac{1}{N} \sum_{i=1}^N y_i \ln(y_i),$$

де: y_i – рівень доходу індивіда i .

Загальна формула для “індексів загальної ентропії” (T при $\alpha \rightarrow 1$):

$$GE(\alpha) = \frac{1}{\alpha^2 - \alpha} \left\{ \frac{1}{N} \sum_{i=1}^N (y_i)^\alpha - 1 \right\}$$

Показник чутливий до змін в нижній частині розподілу, якщо має значення, близьке до нуля; однаково чутливий до змін у всьому діапазоні розподілу, якщо $\alpha \rightarrow 1$ (індекс Тейла); і чутливий до змін у верхній частині розподілу, якщо α приймає більше значення.

Індекс надає більшу вагу групам в нижній частині розподілу, ніж коефіцієнт Джіні, але на відміну від коефіцієнта Джіні, індекси Тейла важко інтерпретувати.

На базі системи контент-моніторингу інформаційних ресурсів мережі Інтернет InfoStream [5] було реалізовано програмний комплекс оцінки присутності об’єктів в інформаційному просторі. На Рис. 1 та 2 наведено криві Лоренця та значення коефіцієнтів Джіні та індексів Робін Гуда за наведеними запитами для таких об’єктів, як компанії (нафтотрейдер та банки, відповідно).

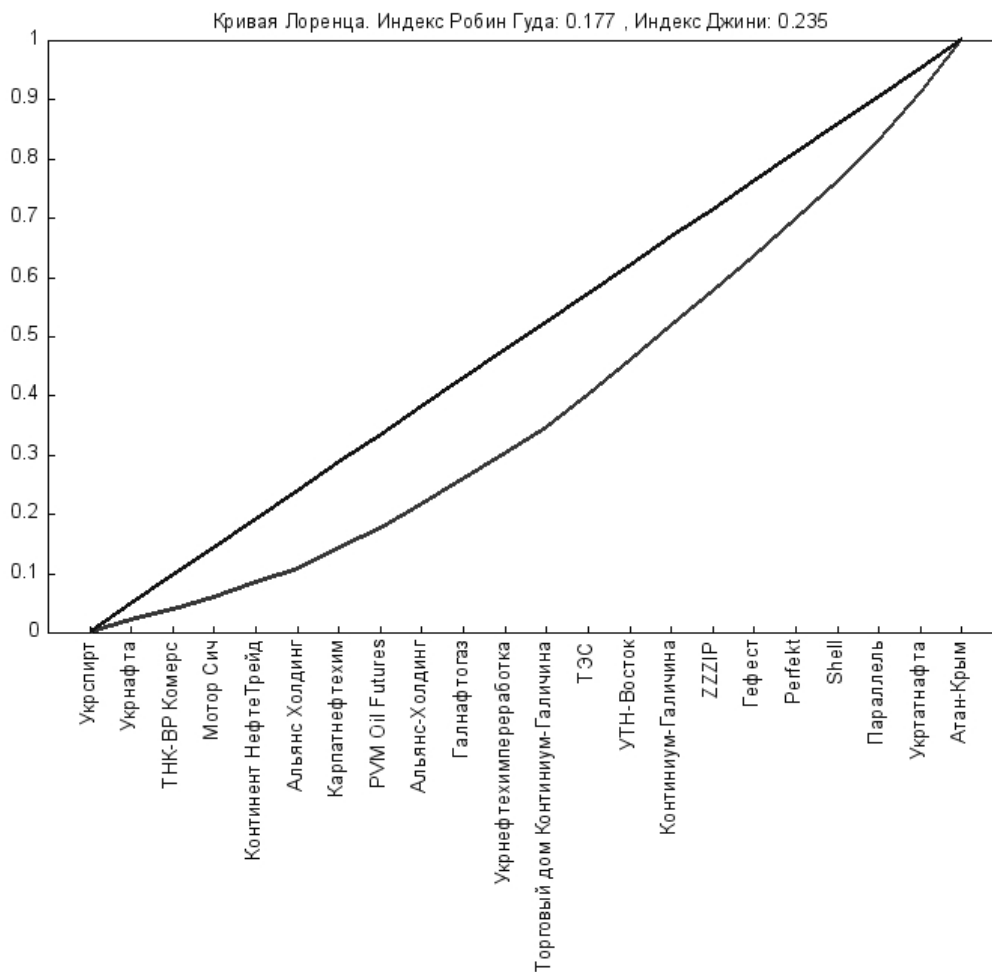


Рис. 1. Медіаприсутність компаній за запитом “Нафтотрейдер Україна”

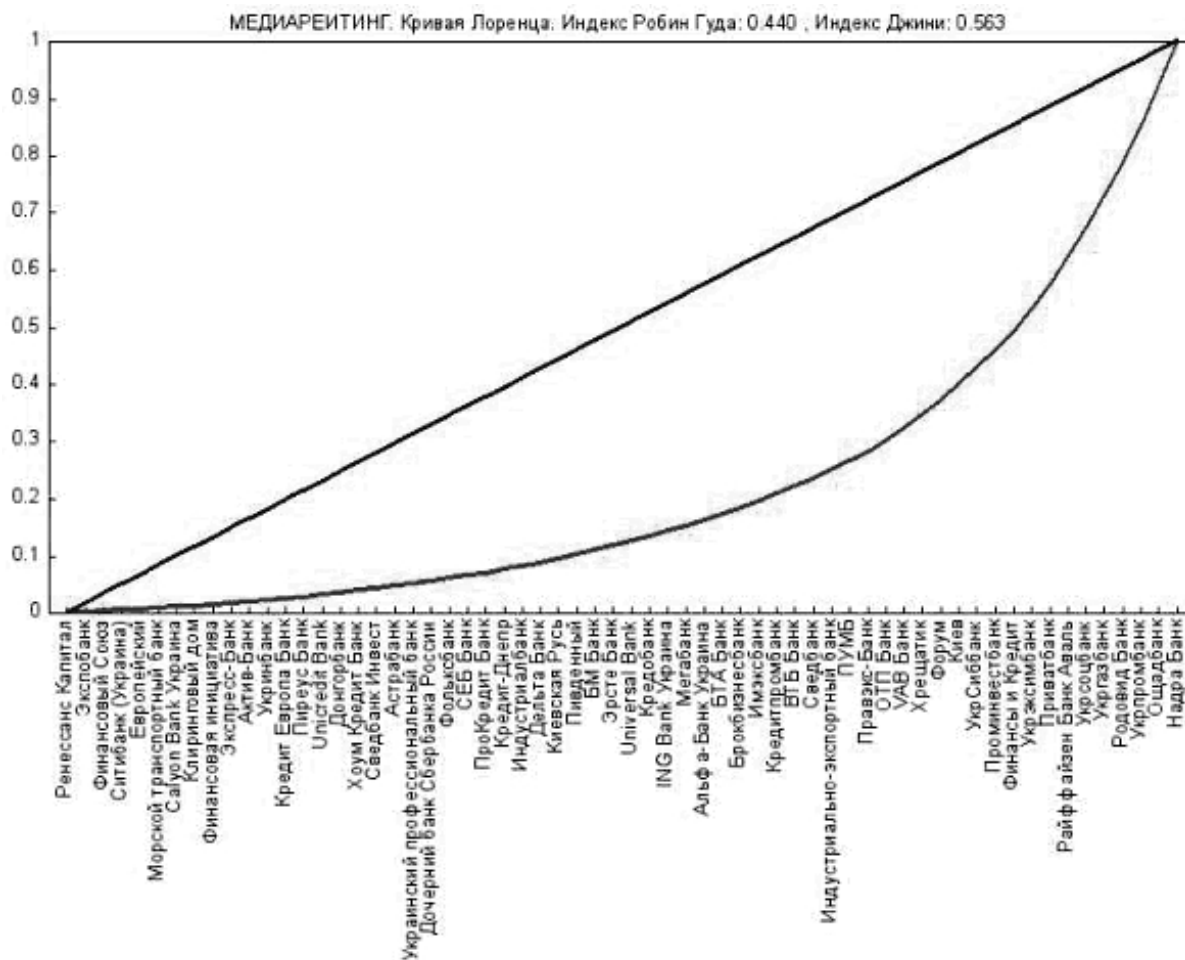


Рис. 2. Медіаприсутність банків (2010 р.)

Індекси порівняння. При проведенні аналітичних досліджень на базі застосування засобів контент-моніторингу інформаційного простору важливе значення має оцінка адекватності застосування саме такого показника, як присутність в інформаційному просторі (або медіаприсутність). Для отримання об’єктивних цифрових оцінок застосовуються індекси, що знайшли на цей час широке застосування у політології та соціології [6].

Нехай досліджується n об’єктів (наприклад, компаній), які впорядковані та пронумеровані. Нехай (V_1, V_2, \dots, V_n) - кількість публікацій в інформаційному просторі, що відносяться до цих компаній, (R_1, R_2, \dots, R_n) – долі цих компаній на визначеному ринку. Нехай загальні суми визначаються таким чином:

$$\sum_{i=1}^n V_i = V, \quad \sum_{i=1}^n R_i = R.$$

Задача складається в знаходженні розподілу кількості публікацій в інформаційному просторі між компаніями у відповідності з долями на ринку.

Уведемо позначення. Нехай $v_i = \frac{V_i}{V}$, $r_i = \frac{R_i}{R}$ – долі медіаприсутності та долі на ринку, відповідно. При $\frac{R_i}{V_i} < \frac{R}{V}$ компанія і недостатньо представлена в інформаційному

просторі, при $\frac{R_i}{V_i} > \frac{R}{V}$ можна стверджувати, що компанія має надмірне представлення в медіапросторі.

У ідеальному випадку компанія має частку медіаприсутності рівну долі на ринку:

$$v_i = r_i, \quad i = \overline{1, n}.$$

У реальності компанії на ринках показують неможливість досягнення рівності між долями медіаприсутності та ринковою.

Множину різних підходів до вимірювання присутності в медіапросторі можна розділити на декілька груп.

Індекси абсолютних відхилень. Перша група індексів характеризує показність за допомогою абсолютних відхилень, тобто різниць між долями на ринку і в інформаційному просторі. Ідеальна показність досягається при $v_i = r_i$, що відповідає нульовому значенню індексів. Можливі два варіанти обліку відхилень: знаходження максимального відхилення і використання деякого усереднювання.

Максимальне відхилення:

$$MD = \max_{i=1, n} |r_i - v_i|.$$

Найпростіший з можливих індексів. Він показує величину спотворення для самої неточно представлені компанії.

Індекс Рє. Цей індекс є середнім арифметичним абсолютних відхилень:

$$I_{Rae} = \frac{1}{n} \sum_{i=1}^n |r_i - v_i|.$$

Індекс має ясну інтерпретацію: наскільки в середньому кожна компанія не відповідає своєму точному представленню в інформаційному просторі. Але індекс має значний недолік: його значення залежить від числа компаній. Коли число компаній, представлених в інформаційному просторі, індекс приймає дуже низькі значення, що може зовсім не означати хорошу показність.

Квадратичні індекси. Попередня група індексів заснована на середньому арифметичному в різних варіантах. Унаслідок їх лінійності по відхиленнях індекси можуть не відбивати зміну показності при зміні розподілу долей, оскільки однаково враховують великі і малі відхилення.

Квадратичні індекси дозволяють моделювати різне відношення до структури відхилень. Невеликі відхилення в загальному випадку усунути не можна. Якщо в результаті розподілу деякі компанії мають значно вищі абсолютні відхилення від точної частки, ніж інші компанії, то дана ситуація повинна характеризуватися гіршою показністю в інформаційному просторі, ніж рівніший розподіл відхилень. Для віддзеркалення цієї ідеї запропонований відповідний індекс.

Індекс Галлахера. У літературі цей індекс часто називається індексом найменших квадратів (least squares index):

$$Lsq = \sqrt{\frac{1}{2} \sum_{i=1}^n (r_i - v_i)^2}.$$

Зведення в квадрат значно збільшує відмінність між великими і малими відхиленнями в порівнянні із звичайним підсумовуванням. Малі різниці слабкіше впливають на індекс, ніж більші, які сильно збільшують індекс.

У соціально-економічній статистиці розглядаються завдання вимірювання структурних відмінностей. Прикладом може служити порівняння галузевих структур економік різних регіонів, порівняння структури фактичного випуску з планованим. У цій області були розроблені ряд індексів. Виявляється, що ці індекси можна використовувати в завданні вимірювання представлення компаній в інформаційному просторі.

Індекс Гатєва. Індекс розраховується по наступній формулі:

$$I_{Gatev} = \sqrt{\frac{\sum_{i=1}^n (r_i - v_i)^2}{\sum_{i=1}^n (r_i^2 + v_i^2)}}.$$

Індекс приймає вищі значення, коли компанії мають приблизно рівне значення. При цьому, чим кількість компаній більша, а їх значення менше, тим значення індексу вище. Таким чином, індекс чутливіший до малих компаній, ніж індекс Галлахера.

Індекс Салаї. Він був введений при дослідженні відмінностей в структурі використання бюджету часу у різних груп населення і розраховується по формулі, яка може бути використана в даній наочній області:

$$I_{Szalai} = \sqrt{\frac{\sum_{i=1}^n \left(\frac{r_i - v_i}{r_i + v_i} \right)^2}{n}}.$$

Цей індекс відрізняється від всіх розглянутих вище індексів з цієї групи. Чим значніше компанія, тим більше значення прийматиме $(r_i + v_i)^2$ що приводить до зменшення її внеску в загальній сумі. Це збільшує значущість малих компаній. Індекс Салаї приймає близькі до 1 значення, коли велика кількість компаній не представлена в інформаційному просторі. Таким чином, індекс дуже чутливий до некоректного представництва малих компаній, що помітно відрізняє від всіх інших.

Індекс Алєськерова-Платонова. Індєкси абсолютних відхилень і квадратичні індєкси вимірюють показність через значення відхилень, але рівні перевищення частки в інформаційному просторі над часткою на ринку приводять до різних ефектів з погляду пропорційності. Причиною тому є різна значущість відхилення для великих і малих компаній.

Індекс Алєськерова-Платонова розраховується тільки по компаніях, представлених в інформаційному просторі:

$$R = \frac{1}{k} \sum_{i=1}^k \frac{r_i}{v_i}.$$

Коли частина компаній не присутня в інформаційному просторі (що відповідає реальній ситуації на більшості ринків), то компанії, представлені в ньому в середньому на кожен відсоток частки на ринку, отримують більш за один відсоток представництва в інформаційному просторі. Індекс показує середнє перевищення частки в інформаційному просторі над часткою на ринку для k компаній з ненульовою медіаприсутністю. Якнайкраще значення індексу рівне одиниці, що відповідає відсутності компаній, які не отримали представництва в інформаційному просторі, і повній пропорційності розподілу.

Висновки.

Нерівномірність представлення в інформаційному просторі різних об'єктів тієї самої природи може свідчити про інформаційний монополізм, рекламне засилля тощо.

Важливо знати, як насправді співвідноситься, наприклад, медіаприсутність компанії на визначеному ринку та її реальна доля на цьому ринку. Якщо ці величини корелюють, має сенс займатися вивченням медіаприсутності, якщо ні – необхідно застосовувати інші методи. Визначення областей, де кореляція має місце, з одного боку звужує область застосування методів контент-моніторингу, а з іншого – забезпечує отримання надійних результатів подальших досліджень у цих областях.

Використана література

1. Lorenz M.O. Methods of measuring the concentration of wealth // Publications of the American Statistical Association, 1905 (Publications of the American Statistical Association, Vol. 9, №. 70, 9 (70): 209 – 219.
2. Gastwirth J.L. The Estimation of the Lorenz Curve and Gini Index // The Review of Economics and Statistics, 1972 (The Review of Economics and Statistics, Vol. 54, № 3, 54 (3): 306 – 316.
3. Hoover E.M. jr. The Measurement of Industrial Localization // Review of Economics and Statistics, 1936, 18, №. 162-171.
4. Theil H. Economics and Information Theory. Chicago: Rand McNally and Company, Amsterdam: North-Holland, 1967. 488 p.
5. Ландэ Д.В. Основы моделирования и оценки электронных информационных потоков : монографія / Д.В. Ландэ, В.Н. Фурашев, С.М. Брайчевский, А.Н. Григорьев. – К. : Инжиниринг, 2006. – 176 с.
6. Карпов А.В. Измерение представительности парламента в пропорциональных избирательных системах // Моделирование в социально-политической сфере. – 2008. – № 1(2). – С. 10 – 21.

~~~~~ \* \* \* ~~~~~