УДК 004.75

ІНФОРМАЦІЙНІ РЕСУРСИ ТА ПІДХОДИ ДО ВИМІРЮВАННЯ КІЛЬКОСТІ ІНФОРМАЦІЇ

Маслянко П.П., Ліссов П.М.
Національний технічний університет України «Київський Політехнічний Інститут», Факультет прикладної математики

У статті розглянуте питання побудови інформаційно-комунікаційної системи та інформаційних ресурсів як її компонентів. Запропоновані визначення та класифікація ресурсів. Розглядається проблема вимірювання інформації у інформаційному ресурсі, зокрема алгоритмічний, статистичний та семантичний підхід. Описані можливості застосування даних підходів для різних типів ресурсів.

Ключові слова: *інформаційно-комунікаційна система, інформаційний ресурс, вимірювання інформації.*

Вступ

На сучасному етапі розвитку цивілізації однією з найбільших цінностей стає інформація. Конкурентоспроможність будь-якої організаційної структури, підприємства, фірми прямо залежить від кількості та якості інформації, якою вона володіє.

В [1] показано, що під визначенням «інформатизація організаційних систем» ми розуміємо необхідну і достатню множину правових, організаційних, економічних, наукових та науково-технічних рішень і процесів, спрямованих на створення інформаційно-комунікаційних систем з метою задоволення інформаційних потреб, забезпечення та автоматизацію бізнес-процесів, підтримку прийняття рішень та підвищення ефективності управління організаційною системою із застосуванням інформаційно-комунікаційних технологій. Інформатизація передбачає створення високопродуктивних інформаційних ресурсів і застосувань.

Створення інформаційно-комунікаційної системи (ІКС) передбачає створення інформаційного ресурсу як однієї із її компонент. Для того, щоб створити ресурс, який буде максимально ефективно задовольняти потреби організації, необхідно провести ґрунтовний аналіз як згаданих потреб, так і характеристик та особливостей інформації.

Однією з проблем, яка виникають при цьому є оцінювання інформації і визначення її кількості – вимірювання.

Постановка задачі

Мета статті – дослідження, аналіз та розробка рекомендацій щодо застосування методів вимірювання кількості інформації у різних класах інформаційних ресурсів.

Дані, інформація, знання

Згідно законодавства України інформація – відомості, подані у вигляді сигналів, знаків, звуків, рухомих або нерухомих зображень чи в інший спосіб [2]. Законодавство України визначає дані як інформація у формі, придатній для автоматизованої обробки її засобами обчислювальної техніки [2].

За європейськими стандартами, знання – це комбінація даних та інформації, до яких додається точка зору, навички та досвід експерта, що дає вагомий результат, який може бути використано для прийняття рішень. Знання може бути вичерпним та/або вузьким, індивідуальним та/або колективним. Нажаль, українське законодавство взагалі не визначає поняття «знання».

Згідно інших джерел, “Data is information before it has been given any context, structure and meaning”[3] – Дані це інформація до того, як їм надається контекст, структура та значення.

Найбільш вдалим, на думку авторів, є визначення, за яким дані – це результат простого збору визначених фактів; інформацією вони стають лише при зв’язуванні у щось корисне, комбінацію хто, що, де і як [4]. У свою чергу знання – це розуміння, як і чому щось відбувається [5]. Саме така семантика понять «дані», «інформація» та «знання» будуть використовуватись далі.

Призначенням ІКС є обробка даних, інформації, знань. Дані, інформація та знання являють собою абстрактні поняття. Для використання вони мають бути матеріалізовані у вигляді інформаційних ресурсів [6]. Інформаційний ресурс – сукупність документів у інформаційних системах (бібліотеках,

архівах, банках даних тощо) [7]. Документ – це упорядкована сукупність даних, інформації та знань, яка надає можливості доступу, передачі, обробки, тощо. Прикладом документа може бути паперовий документ, фільм, комп'ютерний файл, тощо [8].

Можна виділити чотири основні типи інформаційних ресурсів: файлові системи, бази даних, інформаційні сховища, інформаційні колектори. Як правило, ресурс є певною комбінацією таких окремих взаємодіючих ресурсів.

Типова модель такого поєднання показана на рис. 1.

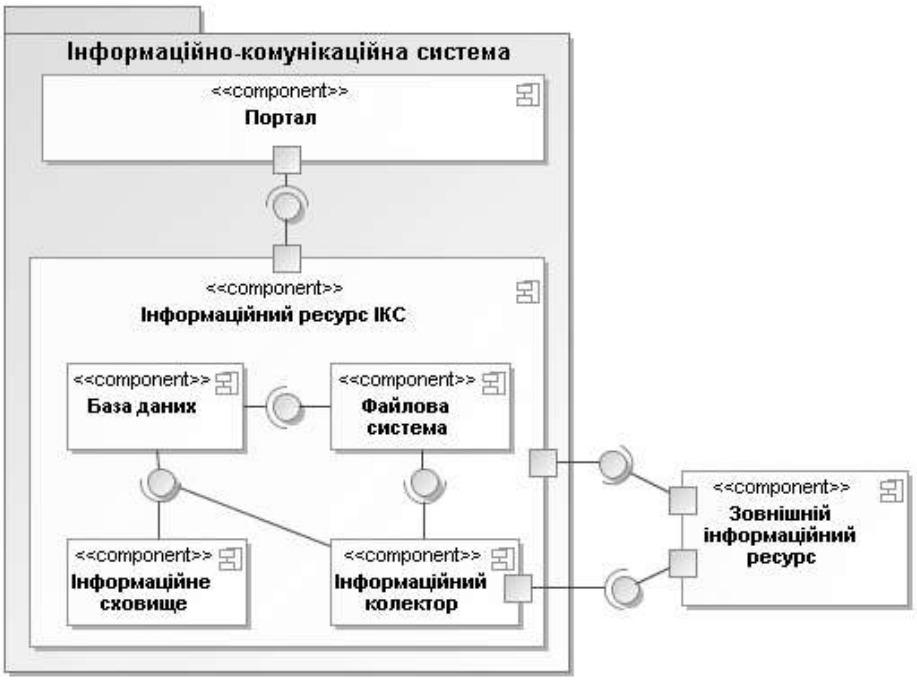

Рисунок 1. Модель організації інформаційного ресурсу ІКС. Діаграма компонентів у нотації UML.

Модель показує взаємозв'язок основних типів ресурсів. Ефективним засобом забезпечення доступу до інформації, якою володіє організаційна структура, є технології веб-доступу, зокрема корпоративні портали [9].

Алфавітний підхід

Алфавітний підхід до вимірювання інформації оцінює розмір інформації як суму структурних елементів повідомлення. Найчастіше такий підхід застосовується при передачі даних. Це пов'язане з тим, що саме така оцінка

дозволяє визначити час передачі повідомлення або необхідні параметри каналу для досягнення визначених його характеристик.

Традиційна і найпростіша міра у рамках такого підходу – геометрична. Одиниця виміру – інформаційний елемент. Міра може бути використана для визначення інформаційної ємності пам'яті комп'ютера. У даному разі в якості інформаційного елементу виступає мінімальна одиниця збереження – біт [10].

Удосконаленням геометричної міри є адитивна міра, запропонована Хартлі у 1928 році (міра Хартлі). Хартлі вперше ввів спеціальне позначення для кількості інформації – I і запропонував наступну логарифмічну залежність між кількістю інформації та потужністю вихідного алфавіту:

$$I = l \log h,$$

де I – кількість інформації, що міститься у повідомленні; l – довжина повідомлення; h – потужність вихідного алфавіту.

При вихідному алфавіті {0,1}; l = 1; h = 2 і основі логарифму 2, маємо:

$$I = 1*\log_2 2 = 1.$$

Дана формула дає аналітичне визначення біту (BIT - BInary digiT) за Хартлі: це кількість інформації, яка міститься в одній бінарній цифрі. Одиницею виміру в адитивній мірі також є біт [11].

Комбінаторна міра оцінює можливість представлення інформації за допомогою різних комбінацій інформаційних елементів в заданому об'ємі. Використовуються типи комбінацій елементів і відповідні математичні співвідношення, які приводяться в комбінаториці.

Комбінаторна міра може бути використана для оцінки інформаційних можливостей деякого автомату, який може генерувати дискретні сигнали (повідомлення) у відповідності до певного правила. Комбінаторна міра використовується для визначення можливостей кодуючих систем, які широко використовуються в інформаційній техніці [12].

Статистичний підхід

Найбільш відомим і широко вживаним на практиці підходом є імовірнісний підхід до вимірювання інформації. На основі даного підходу розроблено великий розділ кількісної теорії інформації, яка за ім'ям її основоположника називається «теорія інформації Шеннона». Головною відмінною особливістю імовірнісного підходу від комбінаторного є той факт, що він базується на імовірнісних припущеннях відносно знаходження певної системи в різних станах. При цьому загальне число елементів (мікростанів, подій) системи не враховується. За кількість інформації тут приймається знята невизначеність вибору із множини можливостей, які мають, в загальному випадку, різну ймовірність [12].

Нехай маємо *n* подій із ймовірностями p_1, p_2, … p_n і необхідно визначити кількість інформації у повідомленні за імовірнісною мірою (позначимо її *H*). Для такої міри *H* Шеннон висуває наступні вимоги:

1. *H* повинна бути неперервною відносно p_i.

2. Якщо всі p_i рівні, то *H* повинна бути монотонно зростаючою від *n*.

3. Якщо вибір розпадається на два послідовних вибора, то первісна *H* повинна бути сумою індивідуальних значень *H* кожного з виборів.

В процесі подальших досліджень Шеннон довів теорему, згідно якої існує єдина функція, що задовольняє всім трьом вимогам, і вона має вигляд:

$$H = -K\sum_{i=1}^{n} p_i \log p_i \ ,$$

де *K* – певна додатна стала [13].

Очевидно, що при рівній імовірності p_i міра Шеннона зводиться до міри Хартлі.

Удосконаленням даного підходу є міра, запропонована у синергетичній теорії інформації. Так Вяткин В.Б. вводить поняття «негентропія», як сутність, протилежна ентропії [14]. Даний підхід є логічним продовженням ідей Бріллюена, згідно якого «Інформація являє собою від'ємний вклад у ентропію»

[15]. Математично величина негентропії визначається за двома різними мірами. Перша міра визначає негентропію як різницю між ентропією до отримання повідомлення і після його отримання. Друга міра визначає негентропію як відношення між такими значеннями [12].

Друга міра очевидно є кращою. Це пов’язано з тим, що будь-яке повідомлення зменшує ентропію саме в певну кількість разів а не на абсолютну величину. Таким чином при появі даного повідомлення першим або другим в системі повідомлень абсолютна величина зміни ентропії буде різною, а відносна – сталою (за умови що повідомлення незалежні).

Відмінний від поглядів Хартлі, Шеннона Вінера і Бріллюена підхід до визначення поняття «кількість інформації» запропонував академік А.Н. Колмогоров, який він назвав алгоритмічним.

Виходячи з того, що найбільш ємним є представлення о кількості інформації «в чомусь» (X) або «про щось» (Y), Колмогоров для оцінки інформації в одному скінченому об’єкті відносно іншого запропонував використовувати теорію алгоритмів. За кількість інформації в одному скінченому об’єкті відносно іншого скінченого об’єкту приймається значення певної функції від складності кожного з об’єктів і довжини програми (алгоритму) перетворення одного об’єкту в інший [16].

Розв’язання задачі визначення кількості інформації в алгоритмічному підході має загальний вигляд і описується наступним чином.

«Відносною складністю» об’єкта Y при заданому X будемо вважати мінімальну довжину "програми" P отримання Y з X. Сформульоване так визначення залежить від "методу програмування". Метод програмування є не що інше, як функція, яка ставить у відповідність програмі P та об’єкту X об’єкт Y" [16].

Так як кожний з об’єктів не може бути нескінченно складним, то можна довести теорему, згідно якої відносній складності об’єкту Y, при заданому методі програмування, може бути поставлена у відповідність інша відносна

складність, яка отримана при іншому методі програмування, така, що виконується нерівність:

$$K_A(Y|X) \leq K_\varphi(Y|X) + C_\varphi,$$

де C_φ - певна стала програмування, яка не залежить від X і Y.

Враховуючи, що при довільних X і Y відносна складність $K_A\ (Y \mid X)$ є скінченою величиною, а $K_A\ (Y) = K_A\ (Y \mid 1)$ можна вважати просто складністю об'єкту Y, А.Н. Колмогоров для оцінки алгоритмічної кількості інформації в об'єкті X відносно об'єкту Y запропонував використовувати формулу:

$$I_A(X:A) = K_A(Y) - K_A(X|Y),$$

при чому $K_A\ (X \mid X) \approx 0$ і, відповідно, $I_A\ (X : X) \approx K_A\ (X)$ [12].

Алгоритмічна інформація найбільш близька до визначення негентропії відображення системних об'єктів в порівнянні з розглянутими раніше інформаційними мірами і навіть, на принциповому рівні суджень, може бути прийнята за її кількісну характеристику. Однак вона не може бути застосована до системних об'єктів. Справа в тому, що коли відображуємий об'єкт є відкритим, ми будемо мати від'ємні значення, негативізм яких може бути доповнений тим, що, як не важко помітити, можливі ситуації, в яких адитивна негентропія відображення сукупності відображуючи об'єктів буде рівна нулю. Таким чином, наприклад, після проведення спостереження і виявлення сукупності об'єктів, які мають безпосередній взаємозв'язок з відображуємим (досліджуємим) об'єктом, в результаті будемо мати, що загальна кількість інформації, яку ми отримали в процесі досліджень, рівна нулю, а це вже нонсенс. Таким чином, ми бачимо, що алгоритмічний підхід, так же як комбінаторний і імовірнісний, не дозволяє отримати розрахункову формулу для негентропії відображення системних об'єктів [14].

Семантичний підхід

Алфавітний та статистичний підхід дуже важко застосувати до вимірювання знань. Обмеження алфавітного підходу в даному випадку очевидні. Статистичний підхід оперує такими поняттями, як імовірність тої чи іншої події, таким чином вимагаючи наявності певних очікуваних результатів досліду. Коли мова йде про знання, дуже часто воно стосується речей, ймовірність яких оцінити апріорно не можна. Тим більше якщо говорити про певні загальні істини та загальні закони. Крім того інколи і для інформації визначення ймовірностей є дуже складним. Спробою запропонувати методи вимірювання таких знань і інформації є семантичний підхід.

Семантичний підхід враховує доцільність та корисність інформації. Застосовується для оцінки ефективності інформації, яка одержується, та її відповідності до реальності.

В рамках даного підходу розглядаються такі міри, як доцільність, корисність, (враховують прагматику інформації) і істинність інформації (враховує семантику інформації).

Кількість інформації I з позицій її доцільності визначається формулою:

$$I = \log \frac{p_1}{p_2},$$

де p_1, p_2 – імовірності досягнення цілі після і до отримання повідомлення, відповідно [10].

Така міра дозволяє, з одного боку, виміряти знання і інформацію. З іншого боку вона, так як і статистичні міри, зводиться до оцінки ймовірностей. Тут стає очевидною характерна особливість семантичного підходу – визначена величина кількості інформації є суб’єктивною. Дійсно, ймовірності досягнення цілі залежать не лише від інформації у даному повідомленні, але в дуже великій мірі від самого суб’єкта, що сприймає повідомлення, від інших його знань, умінь навичок. Це не можна вважати недоліком такої міри, так як дійсно,

інформація, відома всім, навряд-чи має якусь цінність і багаторазове її відтворення не є корисним. Знання є суб'єктивним виходячи з його визначення.

Як вже було вказано, кількість інформації $I_{засв}$, яку «засвоює» користувач, тісно пов'язана із тими знаннями, які користувач має до моменту отримання інформації – з тезаурусом (ТЗ) користувача. На цьому заснована міра корисності інформації. Наприклад, для засвоєння тих знань, які студент отримує у ВНЗ, необхідна середня освіта, інакше студент ці знання не зрозуміє і тому не засвоїть. Тому кожна дисципліна базується на знаннях, які студент повинен отримати на попередніх курсах. Цим пояснюється послідовність учбових дисциплін по рокам навчання.

Залежність інформації, що засвоюється, від тезаурусу, демонструє наступна крива:

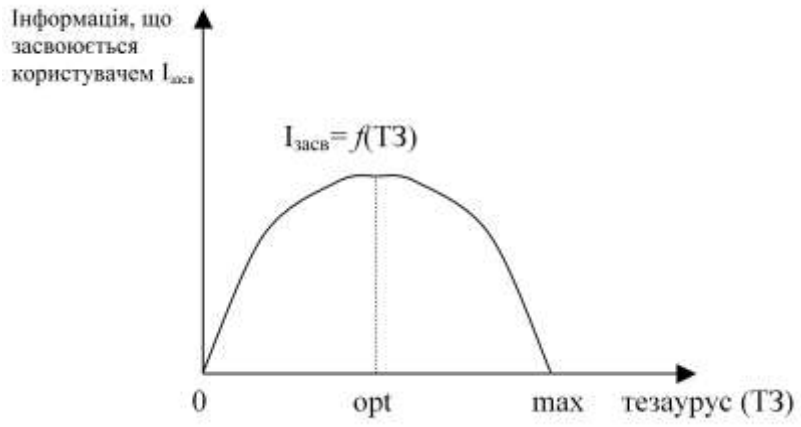

Рисунок 2. Крива засвоєння інформації в залежності від тезаурусу

Очевидно, що при тезаурусі, рівному нулю, інформація не засвоюється, так як є незрозумілою для користувача. З іншого боку при максимальному тезаурусі інформація також не засвоюється, так як вона вже відома користувачу. Максимально засвоюється інформація (тобто, вона максимально корисна) в точці opt, коли користувач має достатній (але не максимально можливий) тезаурус для розуміння інформації, яку він отримує. При значенні тезауруса *i*-го користувача $ТЗ_i$ кількість інформації, яку він засвоює, визначається як $I_{засв} = f(ТЗ_i)$. Сам тезаурус може бути визначений як результат інтелектуального тестування, яке проводиться, наприклад, перед поданням

інформації. При такому тестуванні людині виставляється певний бал, який може розцінюватись як його тезаурус [10].

Міра істинності інформації оцінює інформацію з позицій її відповідності джерелу інформації, тобто реальному світу. Оцінка може бути проведена по шкалі [0, 1]. Тоді для оцінки складного повідомлення (частково істинного) його істинність можна прийняти як функцію від істинності простих повідомлень (які в свою чергу можуть бути істинними 1 чи хибними 0). При цьому функція визначає «вимогливість» щодо істини: середнє арифметичне забезпечить лояльне відношення до неправди, мінімум – визначить що навіть частково хибне повідомлення є все-ж таки хибним [10].

Реалізація підходів

Можливості застосування розглянутих підходів для різних типів ресурсів наведено у табл. 1.

Таблиця 1. Можливість застосування підходів вимірювання інформації.

Тип ресурсу	Призначення	Алфавітний підхід	Статистичний підхід	Семантичний підхід
Файлова система	Персональний ресурс	Автоматично	Не застосовується	Не застосовується
База даних	Зберігання упорядкованих масивів даних	Автоматично	Може бути застосований, можливо автоматизований	Вимагає експертного оцінювання
Інформаційне сховище	Зберігання та аналітична обробка інформації	Автоматично	Може бути застосований, можливо автоматизований	Вимагає експертного оцінювання
Інформаційний колектор	Зберігання знань	Автоматично	Не застосовується	Вимагає експертного оцінювання

Алфавітний підхід використовується для всіх типів інформаційних ресурсів та виконується засобами ресурсу автоматично. Він дозволяє адекватно оцінити апаратні засоби, необхідні для роботи з ресурсом, а також висловити певні вимоги до програмних засобів. Однак така оцінка не може адекватно визначити цінність ресурсу.

Статистичний підхід вимагає визначення параметрів оцінюваної інформації, таких як, наприклад, імовірність тих чи інших подій. Тому автоматичне вимірювання поки що неможливе. Для файлових систем, враховуючи різні типи і завдання даних що зберігаються, така оцінка є неактуальною. Для інформаційних колекторів, які призначені для зберігання знань, статистичні параметри практично не можуть бути визначені, тому така оцінка також не застосовується.

Семантичний підхід може бути використаний для вимірювання інформації та знань. Він не відповідає характеру інформації, яка зберігається у файлових системах. Для баз даних такий підхід може бути використаний із залученням експертів. Наприклад, оцінка істинності даних у базі дозволяє визначити її цінність. Для інформаційних сховищ і колекторів такий підхід також може бути застосований. Необхідність залучення до оцінювання людини викликана, передусім, тим, що комп'ютер (штучний інтелект) поки що не може зрозуміти суть інформації для оцінювання.

Висновки

Кожен з розглянутих підходів може бути застосовано до певних процесів обробки інформації. Так, алфавітний підхід найбільш адекватно може бути використаний при розробці систем збереження і передачі повідомлень. Алфавітний підхід можна найбільш ефективно застосовувати для вимірювання даних. Для вимірювання інформації і знань він практично неприйнятний враховуючи вказані його недоліки.

Статистичний підхід дозволяє частково розв'язати проблеми алгоритмічного підходу. Його можна застосувати для вимірювання інформації, але він ще недостатній для вимірювання знань. Суттєвим недоліком у порівнянні з алгоритмічним підходом є неможливість проведення вимірювання без втручання людини, тому що він базується на ймовірностях, які машина поки-що не може оцінити навіть приблизно.

Семантичний підхід формує основні методи, які можуть бути застосовані для оцінки знань, хоча з успіхом може бути застосований і для оцінки кількості інформації. Однак, на відміну від попередніх підходів, отримана в результаті величина сильно залежить від суб'єктивних особливостей людини, яка дану інформацію (знання) сприймає. Таким чином дана оцінка не може бути використана для визначення абсолютної «об'єктивної» кількості інформації. З іншого боку дана проблема пов'язана саме із суб'єктивним характером знань, що випливає із визначення поняття «знання».

Однак це не означає, що при сучасному рівні продуктивності комп'ютерної техніки слід обмежитись вимірюванням лише даних і інформації як таких, для яких може бути запропонована певна кількісна оцінка.

З іншого боку, всебічне поширення ресурсів, які зберігають саме інформацію і знання (таких як інформаційні сховища, колектори, бібліотеки), існує необхідність у створенні методів визначення принаймні приблизної кількості знань у документі. При цьому важливо створювати саме автоматизовані засоби вимірювання, тому що людина фізично вже не має змоги розібратись в тих об'ємах інформації, які постійно створюються сьогодні.

Таким чином вдосконалення підходів та розробка технологій автоматизації вимірювання інформації в інформаційних ресурсах інформаційно-комунікаційних систем є актуальною проблема розвитку інформаційно-комунікаційних технологій.

ЛІТЕРАТУРА:

1. Маслянко П.П. Основні положення методологій системного проектування інформаційно-комунікаційних систем // Наукові вісті НТУУ „КПІ". 2007, №6 – с 213-219.
2. Постанова Кабінету Міністрів України від 20.01.1997 р. № 40 "Про затвердження Концепції створення Єдиної державної автоматизованої паспортної системи".
3. www.anwers.com
4. Маслянко П.П., Ліссов П.М., Інформаційні ресурси та засоби їх створення // Вісник Східноукраїнського національного університету імені Володимира Даля – №5 (111) – 2007р. – с. 141-145

5. Черненко М., Слепцов С. Принципы классификации управленческих информационных систем // Корпоративные системы – 2004 – №1.
6. Маслянко П.П., Ліссов П.М., Проблеми і технології продукування інформаційних ресурсів // Вісник Східноукраїнського Національного ун-ту імені Володимира Даля - №4(110) (Частина 2) – 2007р – с. 136-141
7. Антопольский А.Б. Вопросы интеграции информационных ресурсов и структура информационного пространства // Техн. информ. общ. – Интернет и совр. общ.: VI Всеросс. объед. конф. СПб, 3 - 6 ноября 2003 г – СПб.: Изд-во Филолог. ф-та СПбГУ, 2003. С. 42-43..
8. П.П., Маслянко. Концепція інформатизації корпоративних структур. Наукові вісті НТУУ „КПІ”. 2003 r., 3, стр. 510-525.
9. Маслянко П.П., Стокоз К.В. Проблеми проектування та застосування порталів // Вісник східноукраїнського національного університету імені Володимира Даля № 5, 2007 – с. 149-157.
10. Информатика. Учебное пособие для студентов Калининградского государственного технического университета, 2003.
11. Хартли Р.В.Л. Передача информации. В сб.: Теория информации и ее приложения. М., 1959.
12. Вяткин В.Б. Задача оценки негэнтропии отражения системных объектов и традиционные подходы к количественному определению информации // матеріалі диссертации «Математические модели информационной оценки признаков рудных объектов», Екатеринбург: УГТУ-УПИ, 2004
13. Шеннон К. Работы по теории информации и кибернетике. М., 1963.
14. Вяткин В.Б. Синергетическая теория информации: общая характеристика и примеры использования // Наука и оборонный комплекс - основной ресурс российской модернизации: Материалы межрегиональной научно-практической конференции. Екатеринбург: УрО РАН, 2002.
15. Бриллюэн Л. Научная неопределенность и информация. М., 1966. С. 34.
16. Колмогоров А.Н. Три подхода к определению понятия "Количество информации". М., 1965.