Методи відбору документів

Ольга Косовська

Кафедра соціальних комунікацій та інформаційної діяльності, Національний університет "Львівська політехніка", УКРАЇНА, м.Львів, вул.С.Бандери, 12,
E-mail: olgakosovska@gmail.com

In the paper brief review of Zipf's and Bradford's laws, which can be used for selecting and arranging documents, is presented.

Ключові слова – документ, закон.

Термін «документ» походить від латинського слова documentum, яке, в свою чергу, походить від docere, що означає навчати. Звідси випливає що термін «документ» у минулому мав більш точне, ніж сьогодні, значення: «те, що служить для навчання». Наявність декількох визначень не заважає документам брати найактивнішу участь в житті суспільства, переносячи інформацію крізь час і простір.

Відмічено, що зростання документів носить експоненціальний характер[2]. При цьому щорічний приріст потоків науково-технічної інформації складає 7-10%.

Швидкість зростання документів досить велика, виникає питання: як знайти потрібний документ за короткий час, як вибрати серед усіх документів найбільш важливий і значущий? Постає задача у розробці не лише методів відбору матеріалів, але і у визначенні методів рангування (сортування) документів у певному порядку. Існують закони, які допомагають у всьому цьому розібратись.

Найбільш відомий гіперболічний закон, який відноситься до статистичної обробки текстів, сформульований Ціпфом [1]. Він стосується розподілу слів в достатньо великих вибірках тексту. Точніше, Дж. Ціпф, зібравши величезний статистичний матеріал, спробував показати, що розподіл слів природної мови підпорядковується одному простому закону, який можна сформулювати наступним чином. Якщо до якогось досить великого тексту скласти список усіх в ньому слів, що повторюються, потім розмістити ці слова в порядку спадання частоти їх повторюваності в даному тексті і пронумерувати в порядку від 1 (порядковий номер найбільш часто вживаного слова) до *R*, то для будь-якого слова добуток його порядкового номера (рангу) (в такому списку) та частоти його повторюваності в тексті буде величиною постійною, що має приблизно однакове значення для будь-якого слова з цього списку. Аналітично закон Ціпфа може бути виражений у вигляді

$$fr = const, \qquad (1)$$

де *f* - частота зустрічальності слова в тексті; *r*- ранг (порядковий номер) слова в списку; *const*- емпірична постійна величина.

Отримана залежність графічно виражається гіперболою.

Найважливішим для розглянутої нами проблеми є той факт, що і документи всередині будь-якої галузі знань можуть розподілятися відповідно

до цього закону. Окремим випадком його є закон Бредфорда, безпосередньо пов'язаний вже не з розподілом слів в тексті, а з розподілом документів всередині якої-небудь тематичної області.

Досліджуючи явище розкиданості статей тої чи іншої тематики в наукових журналах, Бредфорд виявив цікаве відношення між кількістю журналів та кількістю опублікованих них статей на ту чи іншу тему. На підставі встановленого факту С. Бредфорд сформулював закономірність розподілу публікацій за виданнями.

Основний сенс закономірності полягає в наступному: якщо наукові журнали розташувати в порядку спадання числа статей з потрібної тематики, то журнали в отриманому списку можна розбити на три зони таким чином, щоб кількість статей в кожній зоні по потрібній тематиці була однаковою. При цьому в першу зону, так звану зону ядра, входять профільні журнали, безпосередньо присвячені даній тематиці. Кількість профільних журналів в зоні ядра невелика. Другу зону утворюють журнали, частково присвячені заданій області, причому число їх суттєво зростає в порівнянні з числом журналів в ядрі. Третя зона, найбільша за кількістю видань, об'єднує журнали, тематика яких досить далека від розглянутого предмета.

Таким чином, при рівному числі публікацій з певної тематики в кожній зоні число найменувань журналів різко зростає при переході від однієї зони до іншої. С. Бредфорд встановив, що кількість журналів в третій зоні буде приблизно в стільки разів більше, ніж у другій зоні, у скільки разів число найменувань у другій зоні більше, ніж у першій. Позначимо P_1 - число журналів в 1-й зоні, P_2- в 2-й, P_3 - число журналів в 3-й зоні.

Якщо a - відношення кількості журналів 2-ї зони до числа журналів 1-ї зони, то закономірність може бути записана так:

$$P_1 : P_2 : P_3 = 1 : a : a^2 \quad \text{або} \quad P_3 : P_2 = P_2 : P_1 = a \qquad (2)$$

Цю залежність називають законом Бредфорда.

Висновок

Множина документів набуває обрисів системи, в якій елементи взаємопов'язані, а закономірності, що керують цими зв'язками, можуть бути вивчені.

Література

1. Солтон Дж.. Динамические библиотечно-информационные системы. – Москва, Мир, 1979.
2. Чурсин Николай. Популярная информатика. – Киев, Техника, 1982.