4. ЕФЕКТИВНІСТЬ ІНФОРМАЦІЙНОГО ПОШУКУ

4.1. Поняття про критерій відповідності. Види релевантності.

Функція інформаційної системи полягає у виділенні з пошукового масиву тих документів, які містять інформацію, що задовольняє інформаційну потребу користувача. Однак, інформаційна потреба виражається в інформаційному запиті, формулювання якого може лише більш-менш приблизно висловлювати дійсну інформаційну потребу.

Інформаційний запит подається пошуковій системі у вигляді ***пошукового образу запиту (ПОЗ)***, тобто формалізованого переліку термінів. Крім того задається формальний ***критерій відповідності (КВ)*** документа запиту.

Критерій відповідності (match criterion) – ознака або сукупність ознак, за якими визначається ступінь відповідності між запитом або пошуковим приписом і пошуковим образом документа, самим документом або записом його частини для прийняття рішення про видачу або не видачу конкретного документа на інформаційний запит, що обробляється системою.

Пошуковий образ запиту разом з критерієм відповідності складають ***пошуковий припис***: ПП = ПОЗ + КВ.

Інформаційна система у відповідь на запит, виконуючи пошуковий припис, видає деяку сукупність документів.

Однак не всі документи у видачі задовольняють інформаційну потребу. Як правило, вони лише формально відповідають пошуковому припису. Документи, які дійсно є відповідними до потреби користувача, називаються ***пертинентними***.

Інформаційна потреба є досить складним психічним явищем, і підвищення ступеня пертинентності видачі є не тільки непростою метою для досягнення, але її навіть важко чітко поставити як практичну задачу.

Задача визначення відповідності результатів видачі задокументованому запиту є простішою. Документи, що відповідають запиту, називаються ***релевантними***.

Релевантність – це ступінь відповідності змісту документа, який знайдений в результаті інформаційного пошуку, змісту інформаційного запиту.

Оцінка рівня релевантності буде залежати від того, хто здійснюватиме цю оцінку. Якщо оцінку здійснюватиме автор запиту, то він буде оцінювати не стільки релевантність, скільки пертинентність, в тій мірі, в якій йому вдасться ознайомитися з документом. Якщо ж релевантність оцінюватиме працівник системи, то він зможе об'єктивно враховувати тільки формальне входження елементів пошукового припису в документ, не задаючись питанням про відповідність запиту потребі користувача («Який запит, така і відповідь»).

Але саме остання характеристика визначає ефективність самої системи. Таким чином, треба розрізняти ***формальну або технічну релевантність*** і ***смислову або семантичну релевантність.***

Формальна релевантність (formal match criterion)– наявність у документі контекстних ситуацій, які задекларовані у користувацькому запиті.

Семантична або смислова релевантність (relevancy criterion)– відповідність змісту документа інформаційній потребі користувача.

Індивідуально-прагматична релевантність або ***пертинентність*** (***pertinence)*** – ступінь задоволення інформаційної потреби користувача, яка сформульована у вигляді інформаційного запиту змістом документа, який знайдений в результаті інформаційного пошуку.

Для організації видачі документів система повинна вміти оцінювати релевантність апріорі, до видачі, щоб видати саме релевантні документи. А для визначення якості роботи системи оцінку релевантності виданих документів виробляють апостеріорі, після видачі. Апостеріорна релевантність дуже залежить від апріорної, але ці характеристики різні за своєю природою. Так, при автоматичному пошуку система має лише пошуковий припис та пошукові образи документів. Це означає, що система може встановлювати відповідність лише цих об'єктів, що призводить до введення ще одного параметра – ***системної чи алгоритмічної релевантності***.

Системна (алгоритмічна) релевантність – це відповідність пошукового припису пошуковому образу документа.

Людину цікавить семантична релевантність, але ІПС видає результати з системною релевантністю.

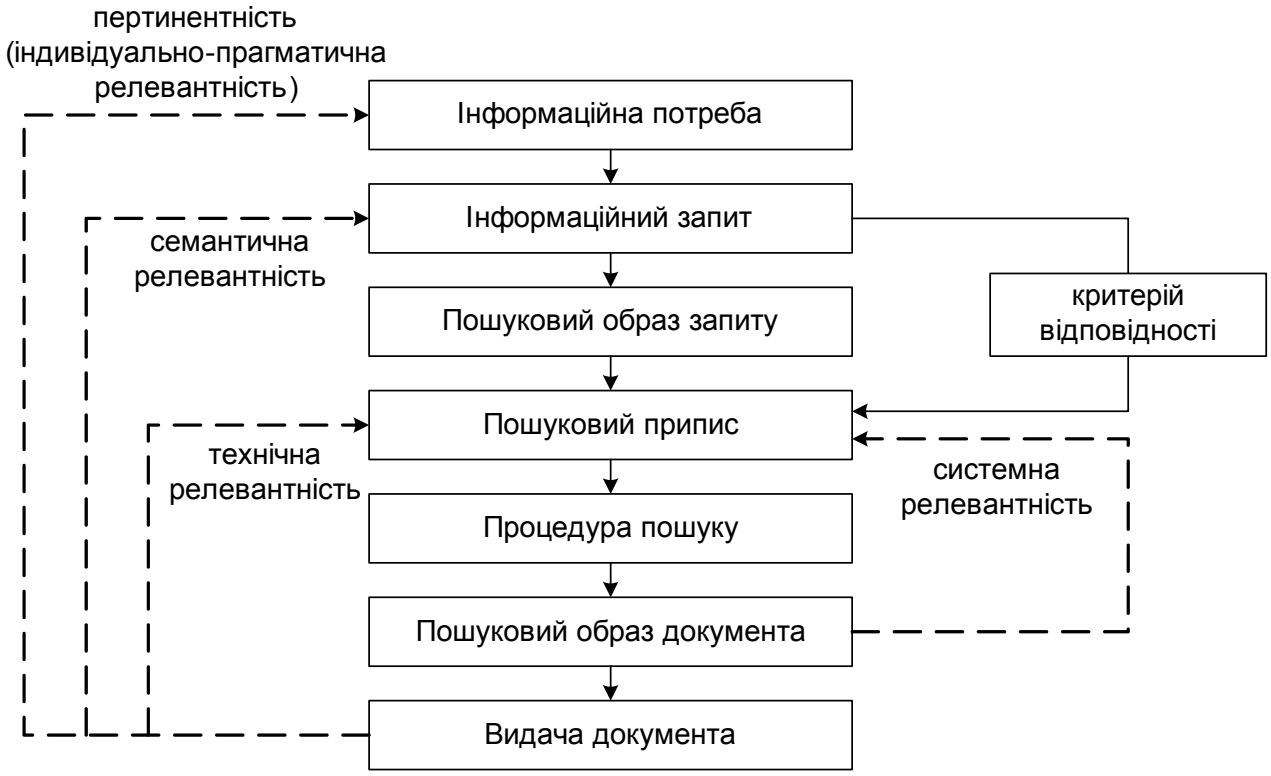


Рис.1. Організація видачі документів

Найпростіший і найпоширеніший критерій релевантності полягає у вимозі повного збігу пошукового образу документа з пошуковим приписом. Але такий критерій застосовується лише до обмеженого виду запитів, наприклад до пошуку за повним бібліографічним описом, або до пошуку всіх документів в деякому тематичному класі за прийнятою класифікацією знань.

4.2. Визначення ступеня релевантності результатів

В реальних пошукових системах при всебічному координатному індексуванні ймовірність повного збігу припису з пошуковим образом документа дуже низька. Тому необхідно якось оцінювати не абсолютну, а відносну релевантність – ***ступінь релевантності*** – на основі часткового збігу пошукового припису з пошуковим образом документа. При цьому система повинна видавати документ, якщо ступінь його релевантності запиту перевищив певний достатньо високий поріг.

Існує багато методів обчислення ступеня системної релевантності. Кожен з існуючих методів розроблений з метою змоделювати людське сприйняття подібності та відмінності змісту текстів за формальними ознаками. Розглянемо деякі з них.

У найпростішому випадку **ступінь релевантності** можна оцінювати за відношенням кількості дескрипторів у запиті – A ,знайдених в документі, до загальної кількості N дескрипторів в запиті:

$$R1 = A / N$$

Вимога повного збігу запиту з документом відповідає R1 = 1 і A = N = M, де M – загальна кількість дескрипторів в пошуковому образі документа.

У практичних пошукових системах поріг релевантності встановлюють певною допустимою різницею (d) між загальною кількістю дескрипторів в запиті N і кількістю тих, що знайдені в документі. Значення ***R1 = (N-d) / N*** при цьому менше 1.

Відсутність у документі деяких дескрипторів із запиту означає, що в цьому документі надається неповна інформація на запит, але є корисні відомості.

Якщо пошук на повний збіг дає незадовільний результат, тоді проводять пошук на збіг всіх, крім одного дескриптора запиту (d = 1), крім двох (d = 2) і т. д. Якщо ж запит складається лише з одного терміна, то пошук здійснюється лише на повний збіг.

Наявність у документі дескрипторів, які відсутні у запиті, зазвичай не враховується при пошуку, але це може означати, що не весь документ важливий для користувача, і це знижує ступінь його фактичної релевантності.

Це враховується у складнішому випадку визначення ступеня релевантності, коли за критерій релевантності приймається величина **R2 = A / M** – відношення кількості знайдених дескрипторів в документі до кількості всіх дескрипторів в пошуковому образі документа.

Вимога повного збігу документа із запитом тут також відповідає R2 = 1, а при частковому збігу значення R2 знаходиться в межах від 1 до 0. Як показує практика, для систем з таким критерієм релевантності задовільне видавання спостерігається при встановленні порогу видавання в інтервалі від R2 = 0,25 до

R2 = 0,4. Очевидно, що R2 залежить від прийнятої глибини та різнобічності індексування документів, від середньої кількості M дескрипторів в пошуковому образі документа.

При багатослівному пошуковому образі документа запит також повинен бути багатослівним. Якщо M = 10, то пошук за одному поняттю ніколи не дасть R2> 0,1 і система нічого не видасть. У запит доведеться додавати нові терміни, таким чином "пояснюючи" системі свою потребу. Якщо у першому випадку R1 для збільшення кількості виданих документів потрібно було видаляти дескриптори із запиту, то для критерію R2 навпаки слід розширювати запит.

4.3. Оцінювання ефективності пошуку з точки зору користувача

До цього часу ми розглядали, як система оцінює корисність того чи іншого документа для користувача. Тепер зупинимося на тому, як користувач може оцінити корисність системи для себе.

Взагалі ефективність системи для замовника визначається як її технічною якістю, так і економічною, тобто вартістю в обернено пропорційній залежності. Ми будемо розглядати виключно технічну ефективність.

Показник технічної ефективності може бути визначений порівнянням реальної діючої системи з ідеальною моделлю. Ідеальна модель може бути визначена (як це було зроблено основоположником науково-технічної інформатики К. Муерсом) так:

"*ідеальна інформаційно-пошукова система* – це система, яка з документального фонду видає лише ті і всі ті документи, які б відібрав сам користувач, якби він міг уважно прочитати кожен з них".

Звідси випливає, що про ефективність пошукової системи можна говорити лише після здійснення пошуку та оцінки його результатів, оскільки відбір документів залежить від конкретної справи, і заздалегідь визначеним бути не може. Цей факт значно знижує цінність визначення ефективності систем, яке потрібно здійснити наперед, коли йдеться про придбання, впровадження або розробку системи. Наперед можна визначити тільки ефективність стосовно

технічної релевантності, а «в ході справи» користувач судить про систему за її реальною пертинентністю, яка є завідома нижчою.

Співвідношення множини реально виданих документів D_r з множиною ідеальної видачі D_i характеризується такими підмножинами:

A – документи, які реально видані системою і входять в бажану видачу;

B – документи, які видані системою, але не входять в бажану видачу;

C – документи, які не видані системою, але входять в бажану видачу;

D – документи, які не входять ні в реальну, ні в бажану видачу.

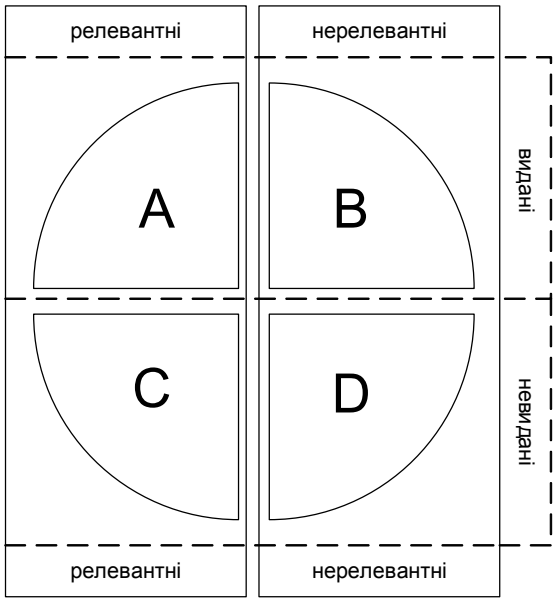
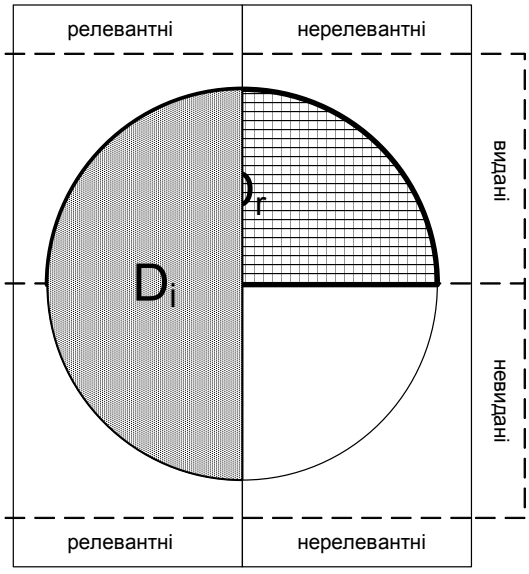


Рис.2. Підмножини документів

В ідеальному випадку:

$$D_r = D_i = A,\ B = C = D = 0$$



В реальних ситуаціях розглядається співвідношення кількості документів у цих множинах. n_a - кількість документів у множині A, n_b - кількість документів у B, n_c - кількість документів у C, n_d - кількість документів у D.

Найчастіше розглядають два відношення:

Коефіцієнт точності пошуку (або точність пошуку) (*Precision*)– це відношення кількості релевантних результатів до загальної кількості документів, посилання на які містяться у відповіді пошукової системи.

Коефіцієнт точності:

$$P = n_a / (n_a + n_b)$$

Коефіцієнтом повноти пошуку (або повнотою пошуку) (*Recall*) називають відношення кількості отриманих релевантних результатів до загальної кількості документів, що існують в пошуковому масиві, і є релевантними даному пошуковому запиту.

- Коефіцієнт повноти:

$$R = n_a / (n_a + n_c)$$

Множина B, що містить документи видачі, які не відповідають запиту, називається шумом (інформаційний шум).

Коефіцієнт інформаційного шуму (*noise ratio)* – це кількісна характеристика інформаційного пошуку, яка визначається шляхом ділення кількості виданих в результаті виконання пошуку нерелевантних документів на загальну кількість виданих документів (релевантних і не релевантних).

Відносна кількість шумових документів у видачі:

$$N = n_b / (n_a + n_b)$$

Сума коефіцієнту шуму та коефіцієнту точності становить 1.

$$N + P = 1.$$

Множина C, яка містить релевантні документи, які не видані користувачу, називається ***втратами***. Відношення числа «загублених» документів до загальної кількості релевантних документів в масиві називається ***коефіцієнтом втрат***, або ***коефіцієнтом мовчання*** (*silence*)

$$S = n_c / (n_a + n_c).$$

Коефіцієнти втрат і шуму не є самостійними показниками ефективності пошуку. Вони однозначно пов'язані з коефіцієнтами повноти і точності:

$$N = 1 – P, S = 1 – R$$

Очевидно, що чим вищими є коефіцієнти повноти і точності, тим вищою є ефективність пошуку. При роботі з різними інформаційними системами значення цих показників буде різним. Якщо у видачі всі документи пошукового масиву виявляться релевантними, і не буде жодного шумового документа, тоді повнота і точність матимуть найвищого значення - 1. У протилежному випадку, якщо у видачі не буде жодного релевантного документа, а видані документи виявляться шумовими, коефіцієнти R і P дорівнюватимуть 0. В інших випадках значення коефіцієнтів повноти і точності знаходяться в діапазоні від 0 до 1.

Конкретні величини R і P в кожній спробі пошуку можуть бути різними, тому про ефективність системи в цілому не можна судити після здійснення одного пошуку. Для цього обчислюють середнє арифметичне показників R і P для великої кількості пошуків за типовими запитами. Такі усереднені коефіцієнти R і P коливатимуться між 0 і 1, ніколи не досягаючи своїх граничних значень. Коефіцієнт повноти R характеризує ймовірність того, що деякий релевантний документ в масиві буде виданий у відповідь на запит. Коефіцієнт точності P характеризує ймовірність того, що деякий документ у видачі виявиться релевантним. Часто величину коефіцієнтів виражають у відсотках, помноживши їх розрахункові відносні величини на 100.

Слід зауважити, що повнота і точність пошуку залежить не тільки від системи, але і від типу запитів. За одним типам запитів система може проводити пошук краще, а за іншими – гірше. Також ефективність може залежати від уявлення про реальну потребу в отриманні тієї чи іншої інформації. Тому при вказівці характеристик системі слід зазначати умови проведення випробувань і характер запитів, на підставі яких ці характеристики були обчислені.

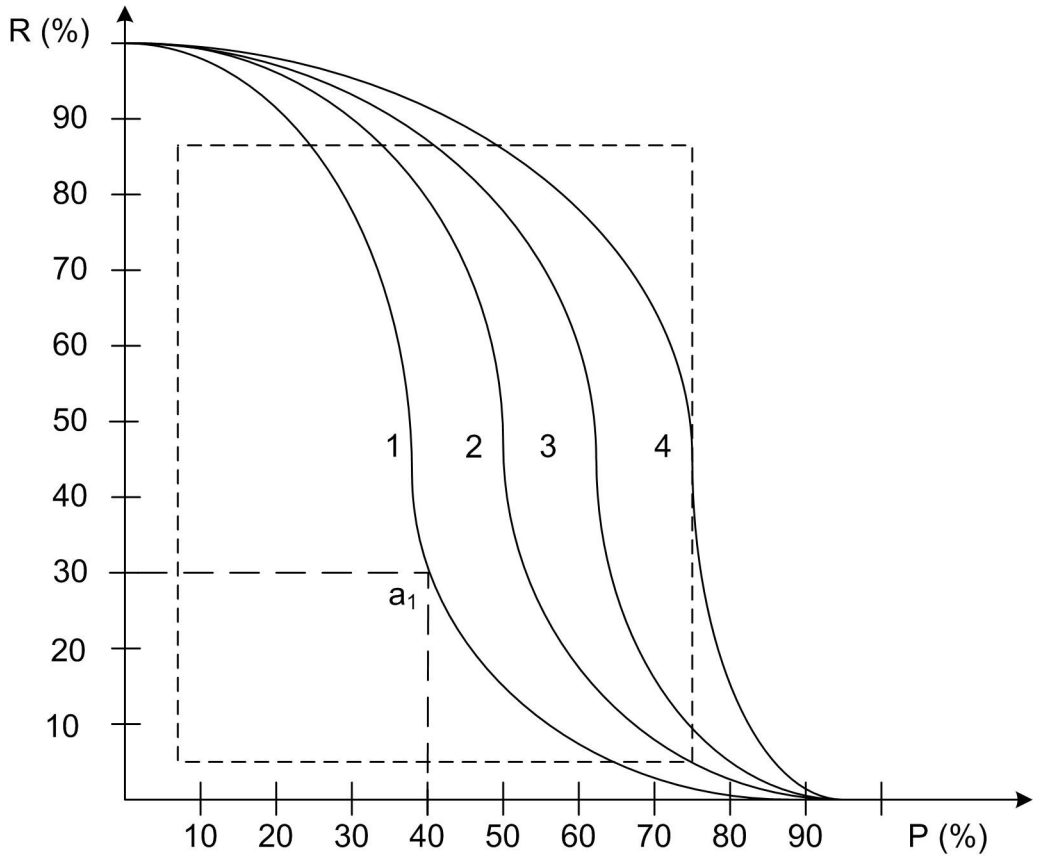
Повнота і точність є незалежними характеристиками інформаційної системи. Неможливо знайти метод розрахунку коефіцієнта повноти за заданим коефіцієнтом точності і навпаки.

Проте існує емпірично виявлені обмеження на ці показники у реальних системах. Якщо система погано сконструйована, то її повнота і точність можуть бути як завгодно малими, тут ніякого обмеження немає. Якщо штучно сформувати таку систему, яка буде містити тільки документи, які релевантні всім запитам певного типу, і яка буде видавати весь свій масив у відповідь на кожний запит, то ми виявимо в цьому випадку стовідсоткову повноту і точність: П = Т = 1 . Але в практичних випадках таке твердження неможливе. Ані повнота, ані точність реальної інформаційної системи ніколи не досягають 100%.

Окрім того, якщо в наявній інформаційній системі намагатися шляхом зміни умов роботи або критерію видачі підвищити один з цих показників, то інший неминуче падає.

4.4. Співвідношення повноти і точності пошуку. Показники ефективності (необов'язкове).

Співвідношення повноти і точності характеризується обереженою залежністю, що можна зобразити на графіку:



На координатних осях відкладені значення коефіцієнтів повноти і точності P і R в відсотках, а точки чотирьох кривих показують значення коефіцієнтів P і R , властиві чотирьом різним інформаційним системам. Так точка a1 показує, що система «1» в режимі, що забезпечує 40% точності видає тільки 30% релевантних документів. Це вважається не дуже хорошим показником. Система «2» при тій же точності видає близько 80%, а система «3» - близько 90% релевантних документів, наявних у пошуковому масиві. Чим вище проходить графік показників системи, тим вища її ефективність. Але крайні точки цих графіків практично недосяжні, реальні системи не виходять за межі, приблизно зазначені на рисунку пунктирним квадратом.

Як повнота, так і точність пошуку показують ступінь задоволеності користувача роботою системи. Але ці коефіцієнти не можуть відповісти на питання, яка система краща – та, що забезпечує більшу повноту, чи та, що забезпечує більшу точність. Для інтегральної оцінки системи використовується показник ефективності (E), який дорівнює сумі коефіцієнтів повноти і точності.

$$E1 = P + R$$

Вважається, що система забезпечує непогані результати, якщо цей показник наближається до 1, а якщо він перевершує 1, то система – дуже ефективна. При налаштуванні інформаційної системи на максимальну повноту або на максимальну точність величина E1 визначається в основному тим коефіцієнтом, на максимум якого система налаштовується. На підставі показника E1 можна зробити висновок, що система «1» не дуже хороша, оскільки в середній частині діапазону роботи він приймає значення близько 70%, в той час як для інших систем «2», «3», «4» досягає значень близько 100% ... 115% ... 130% відповідно.

Має сенс також показник ефективності E2, який дорівнює добутку коефіцієнтів повноти і точності.

$$E2 = P * R$$

Цей показник дозволяє оцінити дію системи в центрі діапазону роботи, де для ефективної системи він повинен приймати значення близько 0,5. У нашому

прикладі це виконується для системи «4». Для систем «1», «2», «3» показник E дорівнює приблизно 0,1…0,2…0,3 відповідно.

Показники повноти і точності інформаційної системи не враховують одну важливу обставину – обсяг пошукового масиву. Очевидно, що відшукати необхідні документи серед великого числа непотрібних набагато важче, ніж у тому випадку, коли їх у масиві більшість.

4.5. Коефіцієнт селективності. Коефіцієнт уточнення.

Для оцінки здатності системи відсіювати непотрібні документи вводять ***коефіцієнт селективності*** (специфічності) (specificity) S, що дорівнює відношенню числа невиданих нерелевантних документів до загального числа нерелевантних документів в масиві.

$$S = n_d / (n_d + n_b),$$

Коефіцієнт S приймає максимальне значення S = 1 при відсутності у видачі шумових документів, тобто коли P = 1, а N = 0. У цьому він схожий з коефіцієнтом точності, але його величина залежить також від характеру пошукового масиву. При одній і тій же точності показник S є тим вищим, що менше релевантних документів в масиві. Очевидно, що коефіцієнт селективності найбільш об'єктивно оцінює роботу самого механізму інформаційної системи.

Недоліком коефіцієнта селективності, а також коефіцієнта повноти є те, що в практичних випадках ми не можемо їх визначити з достатньою впевненістю, оскільки без повного перегляду величезних масивів ми не знаємо кількості невиданих релевантних документів n_c. З неминучою похибкою ми можемо дізнатися лише відносний вміст релевантних документів в масиві $\mathbf{t} = (n_a+n_c)/N = (n_a+n_c)/(n_a+n_b+n_c+n_d)$, який можна шляхом дослідів визначити як ймовірність отримання релевантного документа при виборі навмання документів з масиву. Тоді величина n_c може бути визначена через ймовірність t:

$$n_c = N\,t - n_a,$$

де величина пошукового масиву N часто буває відома, і тоді коефіцієнти приймають вигляд:

$$S = (N\,(1–t)– n_a – n_b) / (N(1–t) – n_a),$$
$$R = n_a / N\,t.$$

На жаль, і цей спосіб обчислення показників ефективності не може бути використаний у всіх випадках, оскільки не можливо оцінити обсяг пошукового масиву Інтернету. Тому для пошукових машин в мережі Інтернет показники повноти і селективності пошуку залишаються невизначеними.

Що стосується показника точності, то він також вимагає співвіднесення із загальним характером пошукового масиву. Роботу інформаційної системи характеризує не стільки точність, скільки поліпшення точності в порівнянні з випадковим вибором документів. Для оцінки цього можна скористатися «***коефіцієнтом уточнення***» (*adjustment coefficient*), що дорівнює відношенню коефіцієнта точності до ймовірності випадкового документа в масиві виявитися релевантним.

$$Adj = T/t$$

Іншими словами, ***Adj*** показує наскільки концентрація релевантних документів у видачі перевершує концентрацію їх у вихідному масиві.

Застосовуються й інші коефіцієнти, що враховують відносну важливість видачі релевантних і нерелевантних документів, але вони мають суто суб'єктивний характер і застосовуються лише до конкретних систем у заздалегідь обумовлених ситуаціях.