Обща постановка

Процесът на извличане на информация започва с въвеждането от потребителя на заявка към системата. Заявките са формални описания на информационната потребност, например низ въведен в полето на търсачката. При извличането на информация с една заявка не се идентифицира по уникален начин един-единствен обект от съвкупността. Напротив, обикновено на заявката отговарят повече от един обекта, вероятно с различни степени на релевантност. Под „обект“ се разбира запис, който съхранява определен обем от информация в базата данни, като в зависимост от приложението, обектът може да е текстов, графичен, аудио- или видео-документ.

Повечето системи за извличане на информация изчисляват числов коефициент на релевантност на всеки от документите в базата по отношение на изпратената от потребителя заявка, и ранжират (подреждат в намаляващ ред) така оценените документи според техния коефициент. Най-високо ранжираните обекти са тези, които се връщат като резултат на потребителя. Процесът може да претърпи и повече от една итерация, ако потребителят не е удовлетворен от резултата и желае да прецизира заявката си.

Оценки на резултата

Съществуват различни техники за измерване и оценка на резултата от работата на системите за извличане на информация. Всяка от тях изисква съвкупност от документи и потребителска заявка.

Важни показатели за оценка и управление на качеството са:

Наличност / Достъпност на данните (Availability)

Пълнота на данните (Completeness) — параметър, измерващ съществуването или отсъствието на данни.

Съгласуваност на данни (Consistency) — Съгласувани данни са тези, при които при възможно наличие на дублиране на данни, те са с еднакво и налично съдържание.

Релевантност / Съответност на данни (Relevance) — Този показател изисква стойностите на данните да попадат в приемлив обсег или да са от определена типизирана съвкупност.

Навременност / Свежест на данни (Timeliness/Freshness) — Този параметър използва времето за записване на данните и времето, когато данните се смятат актуални. Разликата между тези времена би показала дали данните са свежи, или са стари.