ИНФОРМАЦИОННО ТЪРСЕНЕ. МЕТОДИ И АЛГОРИТМИ

1. Същност на информационно търсене. Основни понятия. Приложение.

Информационното търсене (ИТ) е определена последователност от операции, които се изпълняват с цел намиране на документи, съдържащи определена информация или за получаване на фактически данни, представляващи отговори на дадени въпроси.

В по-широк смисъл под информационно търсене се разбира област на обработка на данните, включваща съвкупност от аритметични, логически и входно-изходни операции, крайната цел на които е да се извърши подбор по предварително зададени признаци на всички данни (във вид на документи, фактически справки и т.н). Тези данни трябва да съдържат търсената информация и да представляват отговори на въведената информация. Такива данни (документи) обикновено се наричат релевантни.

При информационното търсене се решават две основни задачи: най-напред е необходимо да се установи за какви обекти (документи) или явления става дума в текста на запитването, а след това да се изясни какво е съотношението между тях. По този начин автоматизацията на информационното търсене фактически се свежда до моделиране на разбирането на текстовете и до тяхното съдържателно съпоставяне.

При информационното търсене се решават две основни задачи:

- да се установи за какви обекти (документи) или явления става дума в текста на запитването;
- да се изясни какво е съотношението между тях.

По този начин автоматизацията на информационното търсене фактически се свежда до моделиране на разбирането на текстовете и до тяхното съдържателно съпоставяне.

Информационното търсене е свързано с откриване на факти и документи в информационните системи (ИС), и основно в информационно-търсещите системи (ИТС) и системите за управление на бази от данни (СУБД).

Под **факти** и **числова информация** се разбират структурирани данни, представени в определен формат. Този формат отразява тяхното логическо и физическо представяне, както и от избрания модел за организация и съхранение на данните. Той зависи от физическата памет и от формата на файловете, в който са записани.

Под термина **текст** се разбира група неструктурирани, безформатни данни. А под термина **документ** – текстове на естествен език.

При СУБД потребителят формулира заявка на специализаран (процедурен или непроцедурен) език. При информационните системи той открива набор от документи, свързани с интересуващата го тема или имащи отношение към нея. Критерият на търсене се основава на отношение, наречено сходство. Понятието **сходство** се въвежда за определяне на идентичността на два обекта (факти, документи и др.). Може да се представи като разтояние между два обекта. Колкото повече те си приличат, толкова са по-близко. Друга геометричната интерпретация на понятието сходство е ъгълът между два вектора, представящи обектите във векторното пространство. Обектите се разглеждат като вектори. Сравняват се не самите обекти, а техните представяния. Сравнението може да бъде между заявки и документи, както и между документи.

Други понятия, имащи пряко отношение към разглежданата тема са:

- ✓ Информационен език;
- ✓ Критерий за съответствие;
- ✓ Познавателен образ на документа (ПОД);

- ✓ Индексиране.

Информационен език (ИЕ) се нарича определена семантична система, предназначена за изразяване на основното съдържание на документите и информационните запитвания с цел в масива на документите да се намерят такива документи, които съдържат необходимата информация.

Правилата за превод от естествения език на информационен (и обратно) се задават във вид на двуезичен речник и се реализират по съответен алгоритъм.

Критерий за съответствие (КС) е съвкупност от правила, по които се определя степента на смисловата близост между познавателния образ на документите и съдържанието на информационното запитване.

Особеното на критерия за съответствие като част от процеса за търсене на информация се състои в необходимостта от определяне на последователността от правила за установяване на съответствието. Проблем възниква, тъй като същността на информационното търсене се състои в подбор на "съответни" данни (документи, обекти) от информационния масив. Трудностите произтичат както от лошото определяне на смисъла на съответствието, така и поради неопределеността или дори фактическата грешка в описанията, както в запитването, така и в познавателния образ на документа.

Познавателен образ на документа (ПОД) е основното смислово съдържание на документа, изразено с термините на информационния език (но не и цялата информация, съдържаща се в него), което е поставено в еднозначно съответствие с дадения документ и е предназначено за намиране в масива на документите. Познавателните образи на документите формират информационния масив, в който се извършва процесът на търсенето. Познавателният образ фактически се явява представител на дадения документ и съдържа информация, която пряко или косвено показва мястото на библиографското описание на документа.

Индексиране се нарича процедурата, която осъществява формирането на познавателния образ, т.е. изразява основното смислово съдържание на документа с термините на информационния език. С тези термини се изразява и запитването (въпроса, заявката) на клиента.

Информационното търсене може да се разглежда като област от обработката на данните, която включва съвкупност от операции за проверка на всички данни, съдържащи търсената информация.

Стратегията за търсене в автоматизиранит информационни системи (АИС) включва следните елементи:

- ✓ логико- семантични отношения – отношения на думите и фразите на езика;
- ✓ правила за сравнение – определят процедура по съпоставянето на елементите на запитването и на познавателния образ. Зависят от избрания език на системата;
- ✓ методика за използване на логически семантични елементи и на правилата за сравнение. Чрез нея се определя последователността на работа с елементите.

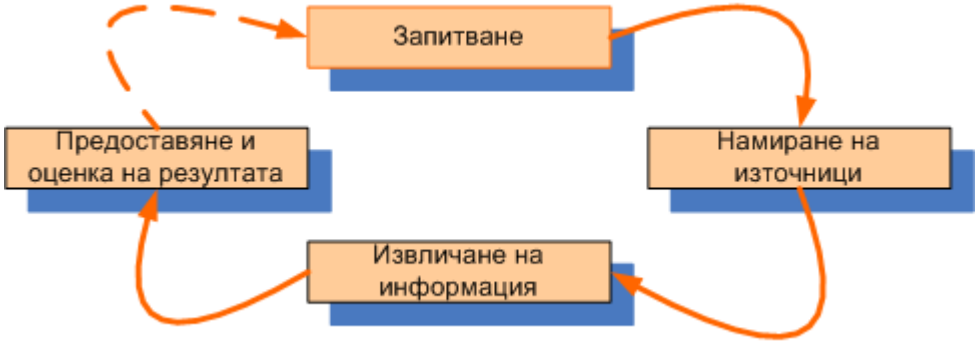
Информационното търсене се основава на последователното прилагане на правилата за сравнение.

Предмет на информационното търсене в БД, ИС и Интернет е:

- ✓ Търсене на документи;
- ✓ Търсене на информация в документи;
- ✓ Извличане на метаданни от документи;
- ✓ Търсене на текст, изображения, звук и видео.

2. **Етапи на информационно търсене**

Информационно търсене се състои от четири основни етапа (фиг.1): запитване; определяне на източниците на информация; извличане на информация и предоставяне на заявителя.



Фиг. 1. Етапи на инфрмационно търсене

Запитването е формализиран начин за изразяване на информационни потребности от потребителите. За изразяване на информационните потребности се използва език на запитване (информационен език). Обикновно всички търсещи системи позволяват въвеждането на запитване да се извършва на естествен език.

При търсенето и **намирането на източници** на информация се използва критерии за съответствие. Особеното на критерия за съответствие, като част от процеса за търсене на информация, се състои в необходимостта от определяне на последователността от правила за установяване на съответствието. Под това се разбира последователно усложняващ се анализ на обекта с постепенен обхват на все по-голям брой от неговите признаци.

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.

След намирането на източниците информация следва извличането на информация от откритите обекти на запитването. Повечето системи изчисляват числов коефициент на релевантност на всеки от обектите на запитване и аранжират (подреждат в намаляващ ред) оценените обекти според техния коефициент.

Най-високо аранжираните обекти по време на извличане на информация са тези, които се връщат, като резултат на потребителя. Съществуват различни техники за измерване и оценка на тези резултати. Точността на оценяването е отношението на броя извлечени документи, които са релевантни, към общия брой извлечени документи.

Процесът на предоставяне и оценка може да има повече от една итерация. Например, ако потребителят не е удовлетворен от резултата от информационното търсене и желае да прецизира заявката си.

Като показатели за оценка на информационното търсене се използват:

- ✓ Наличност / Достъпност на данните (Availability);
- ✓ Пълнота на данните (Completeness) - параметър, измерващ съществуването или отсъствието на данни;
- ✓ Съгласуваност на данни (Consistency) - съгласувани данни са тези, които при възможно наличие на дублиране на данни, 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 за актуални. Разликата между тези времена показва дали данните са свежи.

3. Методи и алгоритми

В инфорнационните системи и базите от данни документите се представят пряко или косвено. При прякото представяне документите се съхраняват в обикновена форма, а при косвеното – чрез различни начини за индексиране. По индекса се определя адресът или идентификаторът на документа.

В съответствие с представянето на документите има два вида информационно търсене:

- ✓ **Пряко търсене** или търсене по ред;
- ✓ **Косвено търсене** или класификация на документите.

При прякото търсене се използват различни схеми за сравнение, като карайни автомати, компаратори. Прилагат се и ключови думи, както и комбинации от тях. Може да се зададе и относителното „разстояние" между ключовите думи.

При косвеното търсене се използват различни **ключови думи** или термини, избрани в съответствие със схемата на индексиране. Ключовите думи се обединяват в групи, наречени **клъстери**. Клъстерите се изпозват за съставяне на **тезауруси** т.е. на речници за ключови думи или синоними. В големите бази от данни се изпозват и клъстери от документи.

За изпълнение на потребителските заяваки за търсене се използват следните подходи:

- ✓ Търсене на точно съвпадение на термини;
- ✓ Търсене с използване на мерки за сходство (несходство)

Търсенето на точно съвпадение на термини използва следните методи:

- Търсене с индексиране на текста;
- Търсене по абзац в пълен текст;
- Търсене с използване на крайни автомати.

При **търсене с индексиране на текст** се въвежда и използва терминът индекс, кото има определена структура. Най-простият индекс се представя като наредена двойка параметри *I (a, s)*, където *a* е стойност на атрибута, а *s* е списък на адресите на елементите (записи, документи и др.), съхраняващи тази стойност.

Друг алтернативен начин на търсене е чрез сигнатури. Текстовете се представят като редове от битове, получени при преобразуване на текста с функция на разложение. За целта се изполват два подхода. При единия всяка дума от текста се преставя чрез функция на разложение на битове. Тази функция се нарича сигнатура. Тя се може да се съхранява в два или повече байта. Така текстът на документа се представя като конкатенация на сигнатури на отделни негови думи (без повторения) и се съгранява във файл. При информационно търсене в теста реално се извършва търсене в сигнатурния файл.

При другия подход тестът се разделя на блокове с фиксирана дължина. За определяне на сигнатурите на даден блок всяка нова дума се представя като припокриващи се тройки символи, като например думата STUDENT: STU-UDE-ENT.

След това всяка тройка се изобразява в номер от бита чрез функцията на разложение. Търсенето в текста се реализира по същия начин, както е при първия подход.

Индексирането позволява да се намали необходимата памет за съхранение и обработка, както и времето за търсене на информацията. То е удачно да се използва когато:

- Системите имат ограничен брой заявки;
- Не е необходимо пълно инвертиране на тестовете;
- Размерът на БД не е голям;
- Документите, които се съграняват и обработват са на един език;
- Актуализацията на БД не трябва да се изпълнява в реално време.

Търсене по абзац в пълен текст се извършва със специални алгоритми. Един такъв алгоритъм е следния: Първо образецът се разполага в началото а текста. Сравнението започва от края на образеца. Последният символ на образеца се сравнява със съответния символ от текста. Ако има съвпадение сравнението продължава от дясно на ляво. Ако няма съвпадение образецът се мести надясно и търсенето продължава. С цел по-бърза обработка на текста могат да се използват различни видове премествания.

Последните се избират по следните критерии:

Ако част от образеца съвпада с текста, то се прави опит да се намери място на този фрагмент от образеца в лявата част на текста. Ако такова място се намери, образецът се премества в дясно, така че да се изравнят съвпадащите части.

Ако няма съпадение, се прави опит да се намери сравняваният символ от текста в останалата (разположена вляво) част от образеца. При успешен опит образецът се премества вляво с толкова, че да се изравнят съвпадащите части . При нуспешен опит образецът се премества надясно, така че неговият символ да се намери веднага след разгледания символ в текста.

Основните методи за информационно търсене се различават по начина на търсене. Те използват:

- ✓ Търсене по съвпадение на термини;
- ✓ Логически изрази;
- ✓ Сравнение с числови стойности и интервали;
- ✓ Контекстно търсене;
- ✓ Търсене с изполване на критерии за сходство (несходство).

При методите, основани на **търсене по съвпадение на термини** се задава предварително пълно и/или частично съвпадение на ключови думи (термини). За изпълнение на заявките за търсене се могат да се използват:

Точно съвпадение – изисква се да се намерят всички документи, които съдържат зададената ключова дума, а не част от нея. Например, STUDENT.

Частично съвпадение – кагато се използват несъществени символи, като [*], [$], [?].и др - STUDENT?, STUDENT??, STUD* и др.

Някой методи на търсене използват логическите операции за да генерират заявки за търсене като логически изрази. Чрез основните оператори на булевата логика AND, OR, NOT могат да се съставят разнообразни логически изрази за точно формулиране на дадено информационно запитване.

- Оператор AND (И) – произведение, пресичане или конюнкция. При наличност на две множества с него се определя тяхната пресечна обща част.

- Оператор OR (ИЛИ) – логическо събиране, обединение или дизюнкция. При наличност на две множества с него се определя ново множество, съдържащо двете изходни множества.

- Оператор NOT (НЕ) – отрицание или допълнение. Ако се приеме, че множеството А дефинира всички описания на документи, съдържащи термина X, и че някои от тях могат да съдържат и термина Y, а множеството В – всички описания на документи, съдържащи термина Y, с оператора НЕ се дефинира подмножество, съдържащо описания на документи с термина X, но без нито едно описание на документ с термина Y.

Примери за заявки с логически изрази:

MEN **AND** WOMEN

MEN **OR** WOMEN

STUDENT **NOT** TEACHER

Методите, основани на **сравнение с числови стойности и интервали** се използват за търсене на документи, публикувани в определено сремe или определен период между две зададени дати.

При медодите за **контекстно търсене** се дефинира съдържанието, както и разстоянието между търсените термини в търсения документ.

Чрез операторите на контекстуалната граматика е възможно да се търсят описания на документи, съдържащи термини, за които се изисква: да бъдат съседни (ADJ), да отстоят един от друг (NEAR), да са синоними (SYN), да се намират в едно изречение (WITH) или в един абзац (SAME). Търси се и по отделни елементи (полета на запис) като заглавие, име на домейни и пр.

Търсене по сходство (Similarity) се основава на термина сходство (виж т.1). При тези методи документите се представят като вектори D_j от ключови думи (термини).

$$D_j = (t_{i1}, t_{i2, \ldots\ldots, } t_{in)}, 1 \le i \le m \text{ и } 1 \le j \le n, \quad (1)$$

където: *m* е общия брой на документите, а *n* – броя на ключови думи (термини).

Всеки от термините t_{ij} може да бъде:

- 1 или 0 - 1 означава наличие на j-тия термин в i-тия документ, 0 – отсъствие;
- число (константа), което показва теглото (колко пъти се среща термина в документа) j-тия термин в i-тия документ.

Документите се представят като индексна матрица (2) на която, редовете са самите документи представени като вектори, както е показно в (1), а стълбовете – вектори на термините. Последните показват наличието на този термин във всеки документ.

Всяка заявка също се представя като вектор Q_j

$$Q_j = (t_{i1}, t_{i2, \ldots\ldots, } t_{in)}, 1 \le i \le m \text{ и } 1 \le j \le n, \quad (3)$$

където t_{ij} означава наличието или от отсъствието (или теглото) на термина в *j*-тата заявка.

При това матрично представяне може да се определи **мярка на сходство** между използваните термини в даден документ и в заяваката. Сходните термини (термините с еднакви тегла) се групират и се съставят тезауруси (виж т.1). Изчислява се сходството между реда, представящ заявката и всеки от редовете на матрицата на документите (2). Така могат да се изберът тези документи, за които стойността на мярка на сходство ≥ от зададената от потребителя.

Ако броя на документите е голям е по-удачно да се изчислява сходство на заявката не с вектрите на документите, а с представители на клъстерите.Наричат ги **центроиди**. Векторът на центроида G_j се представя в следния вид:

$$G_j = (g_{i1}, g_{i2, \ldots\ldots}, g_{in}) , \quad (4)$$

$$g_{i1} = 1/m.\sum t_{ij} \quad 1 \le i \le m \quad (5)$$

където G_j е центроид на j-тия клъстер от документи за n термини , $1 \le j \le n$, а g_{ij} – средно тегло за m документа.

Съществуват различни мерки сходство. Наричат ги функции на сходство или фунции на съвпадение. Такива се коефициент на Дайс (Dice similarity coefficient), коефициент на Жакар (Jaccard Index) , коефициент на препокриване/съвпадение (Rand Index), коефициент на косинуса и др.

Освен понятието сходство се използва и понятието **несходство.** Понякога е по-удачно да се търсят различните, а не еднаквите, например когато те са по-малко на брой. Математически фукцията на несходство може да се преобразува във функция на сходство чрез уравнението (6).

$$\text{сходство} = (1 + \text{несходство})^{-1} \quad (6)$$

Различните алгоритми и стратегии за информационно търсене са подробно описани и представени в Help функциите на конкретни реализации (интернет търсачки, портали и др.) и са придружени с много примери.