Търсене и извличане на информация
Information Retrieval (IR)

1. Въведение

Ако до преди едно-две десетилетия, голяма част от данните бяха структурирани или полу-структурирани и се съхраняваха главно в релационни бази от данни, в момента не е така. Особено след масовото разпространение на системите за управление на съдържание и на социалните мрежи, които отреждат на потребителя ролята на основен генератор на съдържание. Милиони статуси и изображения се публикуват всеки ден в социалните медии. Множество документи се добавят в различни интернет сайтове, електронни библиотеки, системи за обучение, сайтове на научни конференции, популярни онлайн списания и др. Но за да бъде полезно някому, всичкото това съдържание трябва да бъде достъпно и най-важното лесно откриваемо. Затова, *всяка система за управление на документи*, или дори по-общо казано на произволен тип ресурси, *трябва да предоставя необходимите функционални възможности за търсене*, както *по допълнителни външни описатели* (ключови думи и текстови анотации), така *и по съдържание*.

Обикновено когато се говори за документи, хората разбират текстови документи, но това понятие всъщност е доста по-широко и може да включва всякакъв тип *обекти*, в това число изображения, аудио, видео, физически предмети, дори и хора. При изображенията, аудио и видео файловете, сравнително лесно може да се реализира търсене по основното им мултимедийно съдържание. Например при изображенията, може да се търси по цветово съдържание и неговото пространствено разпределение. Извън това, всички нетекстови обекти (например хора, автомобили, имоти и др.) могат лесно да се опишат с текстови анотации или ключови думи, избрани от предварително дефинирани структури, и в последствие да се реализира тяхното бързо и ефективно търсене по предоставените им описателни характеристики.

Търсенето по съдържание е актуален проблем с важно практическо значение, който намира приложение в различни сфери на дейност, например за:

- *Търсене на подобни документи*, дори в колекции от *неструктурирани и неиндексирани* такива. Всяка система за управление на документи трябва да предоставя възможности за тяхното бързо и ефективно търсене, както и да *подрежда резултатите от търсенето по степен на подобие със заявката*.
- Изграждане на глобални *търсещи машини в Интернет*. Търсещите машини са незаменимо средство за достъп до информация в необятното уеб пространство, състоящо се към момента от почти два милиарда уебсайта.
- *Групиране на документи в клъстери* на база на техни общи признаци. Например автоматизирано групиране на потребители по интереси; групиране на музика или книги по жанрове и тематични направления; и т.н.

- *Двойкосъчетание на обекти*. Автоматизирано решаване на задачата за оптимално разпределение на ресурсите (на английски assignment problem). Позволява оптимално да се разпределят ресурси по консуматори; служители по задачи; финансови средства по проекти; да се идентифицират и назначават подходящи оценители и рецензенти; да се свързват подходящи мъже и жени в сайтовете за запознанства; и много други.
- *Изграждане на препоръчващи системи.* Препоръчващите системи са много популярни напоследък в Интернет. Например, ако потребителят чете определена публикация, този тип системи автоматично ще му препоръчат други подобни статии, на същата или близка тема, които също биха му били интересни. На този принцип електронните магазини откриват и препоръчват подобни или свързани продукти, а платформите за видео споделяне - подобни или свързани клипове. Разбира се няма как да се подминат и спорните алгоритми за подбор на съдържанието на социалните мрежи, които често са критикувани, и то с право, че предлагат само „още от същото“, което не позволява на потребителя да се докосне до алтернативни мнения и гледни точки.

Търсенето на подобни документи е фундаментална задача, без решение на която огромното количество данни и документи, публикувани в Интернет, биха били неоткриваеми, а от там и безполезни. Търсенето обикновено се извършва чрез подадена от потребителя заявка, а резултатът от него е върнато, или извлечено, множество от документи, свързани по някакъв начин с нея. Важно е обаче да се определи не само кои документи са свързани със заявката, но и колко точно са свързани с нея, като резултатите от търсенето се подредят в низходящ ред по степен на подобие със заявката.

2. Оценка на резултатите от търсенето

Търсенето се извършва чрез подадена от потребителя *заявка*, като върнатите резултати обикновено се ранкират, т.е. подреждат по степен на подобие със заявката. За да се оцени работата и точността на дадена система за търсене, и респективно методите и алгоритмите, които използва тя, е необходимо да се прецени доколко върнатите от търсенето резултати са адекватни. Кой обаче може да дава оценки дали даден резултат е адекватен или не? Разбира се хората (потребителите). Целта на автоматизираното търсене е да моделира субективната човешка представа за подобие и семантична свързаност между документите, и да открие и извлече само тези документи, които според потребителите наистина са свързани със заявката за търсене.

Преди да се разгледат най-често използваните показатели за оценка на резултатите от търсенето, трябва да се уточнят някои базови класификационни понятия. За по-добра съгласуваност със специализираната литература в тази област, те ще бъдат представени с оригиналните си термини на английски, а описанието им разбира се ще бъде на български.

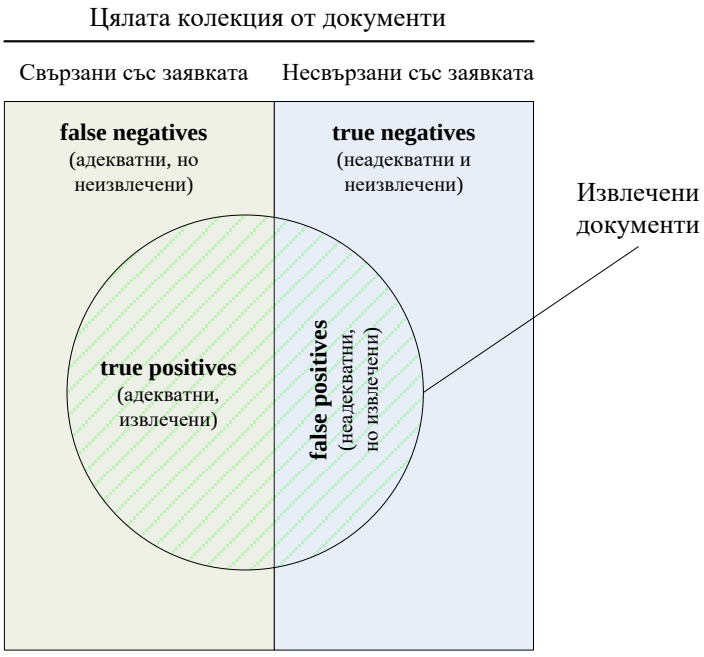
True positives (**TP**) – броят на върнатите документи, които системата за търсене *вярно* е класифицирала като адекватни.

False positives (**FP**) – броят на върнатите документи, които според системата са адекватни, но в действителност не са. Това са фалшиво положителни резултати, т.е. системата ги е класифицирала грешно.

True negatives (**TN**) – броят на документите, които системата *вярно* класифицира като неадекватни. Т.е. и според системата, и в действителност, документите нямат общо със заявката за търсене.

False negatives (**FN**) – броят на документите, които според системата са неадекватни, но в действителност са адекватни и са свързани със заявката за търсене. И тук системата е класифицирала въпросните документи погрешно.

По принцип се счита за нереалистично, при търсене системата да върне всички свързани със заявката документи и всички върнати действително да бъдат адекватни. Тъй като търсенето включва анализ на съдържанието, което може да бъде двусмислено и подвеждащо, или на допълнителни външни описатели, които могат да бъдат неточни, поради неправилен избор, то е практически невъзможно върнатите резултати да не съдържат известно количество грешки и шум. На фигура 1 визуално е представена взаимовръзката между разгледаните основни четири понятия за класифициране на резултатите като адекватни или не. Почти всички показатели за оценка на адекватността на резултатите се дефинират като комбинация от тези четири понятия.



Фигура 1. Взаимовръзка между действително позитивните резултати (true positives) – адекватни и правилно извлечени; фалшиво позитивните резултати (false positives) – неадекватни, но въпреки това извлечени; фалшиво негативните резултати (false negatives) – адекватни, но пропуснати/неизвлечени; и действително негативните резултати (true negatives) – неадекватни и правилно неизвлечени.

Прецизност (Precision)

Показва каква част от извлечените резултати са адекватни.

$$precision = \frac{\text{брой адекватни извлечени}}{\text{брой на всички извлечени}} = \frac{TP}{TP + FP} \quad (1)$$

Например, ако търсенето върне 20 резултата и от тях 15 наистина са свързани със заявката, а останалите 5 не са, тогава прецизността е 15/20, т.е. 0.75.

Прецизността (precision) се оценява много лесно – просто се преценява каква част от върнатите резултати са адекватни.

Откриваемост (Recall)

Показва каква част от адекватните, т.е. свързаните със заявката, документи за извлечени. Този показател се нарича още *чувствителност (sensitivity)*, а в някои източници може да се срещне и като *точност на връщане*.

$$recall = \frac{\text{брой адекватни извлечени}}{\text{брой на всички адекватни}} = \frac{TP}{TP + FN} \quad (2)$$

Връщайки се към горния пример, ако в базата от данни има например 50 адекватни, то тогава откриваемостта ще 15/50, т.е. 0.3.

За разлика от прецизността, откриваемостта се оценява доста по-трудно. Ако базата от данни (колекцията от документи) е статична и предварително се знае колко документи съдържа, свързани с конкретната заявка – тогава е лесно. Но обикновено БД е динамична и то често променяща се. На практика е почти невъзможно да се знае във всеки един момент колко документи съдържа, свързани с конкретна потребителска заявка. Разбира се има и изключения – при предварително подготвени тестови данни и тестови заявки.

Прецизността и откриваемостта са взаимосвързани, като често прецизността се измерва при предварително зададени нива на откриваемост. Обикновено при ниско ниво на откриваемост, прецизността е по-висока.

Точност (Accuracy)

Традиционно, точността се дефинира като отношението на всички вярно класифицирани документи към броя всички документи в базата от данни.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Точността често е критикувана и обвинявана за подвеждаща мярка. Защо? Нека, например, в базата от данни да има 100 документа. 5 от тях са свързани с дадената заявка за търсене, т.е. са адекватни, и 95 не са. Тогава, ако системата не върне нито един резултат, точността ще бъде 0.95 (или 95%). Формално погледнато това е вярно, но за потребителя е странно и неочаквано да получи 95% точност, след като системата не е намерила и върнала нито един свързан документ, въпреки че в действителност има такива. Освен това, тъй като точността отчита TN в числител, то изчислените ѝ стойности винаги ще бъдат много, и неправомерно, високи. Причината е, че при огромна колекция от разнородни документи, броят на несвързаните с дадена заявка документи, винаги ще бъде многократно по-голям от броя на свързаните с нея.

Друг проблем с точността, който важи и за всички останалите мерки (без прецизността) е, че в реална ситуация много трудно могат да се намерят стойностите на TN и FN, освен ако не става въпрос за предварително подготвени тестови данни и тестови заявки.

Top-N точност

В машинното обучение, особено при класификация на обекти, много често се използва *Top-N* точността. Класификаторът, определяйки какъв е обекта, връща не едно единствено предположение, а няколко, сортирани в низходящ ред по степен на вероятност. Според Top-1 точността (обичайната дефиниция за точност), класификаторът правилно е определил типа на обекта, ако предположението с най-висока вероятност отговаря на действителния обект. Top-N точността дава възможност за по-толерантно и не толкова строго тълкуване на точността. Например, при Top-5, обектът се счита за правилно класифициран, ако присъства сред първите 5 предположения на класификатора, дори и предположението с най-висока вероятност да е съвсем различен обект от действителния. В този случай, при Top-1 точността, класификацията е <u>не</u>правилна, защото предположението с най-голяма вероятност изобщо не е действителният обект. Но според Top-5 точността, класификацията е правилна, защото действителният обект присъства сред първите 5 предположения, макар и да не е най-вероятният. В редица ситуации, в които не е необходимо с голяма точност да бъде определен типа на даден обект, подобно свободно тълкуване на точността е оправдано, дори се насърчава. Например в препоръчващите системи, които търсят други подобни обекти – други песни в същия или в свързани жанрове в платформите за видео споделяне; други подобни, но не съвсем същите, продукти в електронни магазини, в онлайн книжарници и др. Именно, чрез подобно по-толерантно тълкуване на точността, потребителят може да открие и други интересни предложения, освен това, което е търсил.

3. Оценка на подредбата на резултатите

Разгледаните по-горе мерки оценяват само доколко намерените резултати са адекватни, но не и дали са подредени правилно. Подредбата зависи от мярката за сходство, която се използва, за да се изчисли степента на подобие между заявката и всеки един от документите.

За да се оцени дали подредбата на резултатите е правилна, може да се определи *степента на корелация* между автоматичната подредба и тази, предоставена от реални хора. В случая абсолютната стойност на коефициентите на подобие, между заявката и всеки един от върнатите резултати, няма значение. Важни са отношенията между подобията, защото именно те определят подредбата. Освен това, получаването от страна на потребителите на точна абсолютна стойност (например 81% или 64,5% и т.н.) за подобията между заявката и резултатите е изключително трудно. Един е склонен да дава по-високи стойности, друг по-ниски. Но ако им се каже „Подредете резултатите по степен на подобие“, тогава вероятността различните потребители да ги подредят по един и същ начин е значително по-висока, отколкото да дадат съизмерими абсолютни стойности за коефициентите на подобие.

За да се определи степента на корелация между автоматичната подредба и референтната, предоставена от потребителите, обикновено се използва линейният корелационен коефициент на Пиърсън (*Pearson*), като се „сравняват“ позиционно *отношенията между всички двойки резултати*. Отношението между два резултата i и j може да бъде „по-голямо“, „по-малко“ или „равно“. Когато i > j, това означава, че i се намира в списъка с резултати преди j. Ако коефициентите на подобие между заявката и двата резултата i и j са еднакви, то отношението между i и j е „равно“ и резултатите следва да се изведат с еднаква тежест (позиция), дори ако се налага да бъдат визуално подредени един под друг. За да се изчисли корелационният коефициент на Пиърсън (4), тези отношения трябва да се кодират с числа. Например: 1 за по-малко; 2 за равно; и 3 за по-голямо.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (4)$$

където:

r_{xy} – корелационен коефициент на Пиърсън. Показва степента на линейна корелация между множеството от отношения X и множеството от отношения Y.

x_i – кодът на *i*-тото отношение от множеството X, *i*-тата стойност на X.

y_i – кодът на *i*-тото отношение от множеството Y.

$\overline{x}$ – средно-аритметичната стойност на всички елементи от X. Т.е. т.нар. извадкова средна на X.

$\overline{y}$ – извадковата средна на Y.

n – броят на елементите (отношенията) в множествата X или Y.

Отношението между всички двойки резултати означава да се отчетат отношенията между всеки един резултат и всеки друг резултат, а не само между съседните резултати. Например при 4 резултата: 1-ви с 2-ри, 1-ви с 3-ти, 1-ви с 4-ти, 2-ри с 3-ти, 2-ри с 4-ти, и 3-ти с 4-ти. По принцип биха могли да се отчетат отношенията и само между съседните резултати (например: 1-ви с 2-ри, 2-ри с 3-ти и 3-ти с 4-ти), но отчитането „всяко с всяко“ дава значително по-точна оценка на подредбата, особено в случаите когато разместването е с повече от една позиция. Отчитането на отношенията само на съседните резултати води до два проблема, които лесно могат да бъдат илюстрирани с примери. Нека е дадено следното отношение между подобията на съседните резултати 2, 3 и 4 със заявката:

sim(резултат_2, заявка) > sim(резултат_3, заявка) < sim(резултат_4, заявка)

От тук е ясно, че степента на подобие на резултати 2 и 4 със заявката е по-висока от тази на резултат 3. Т.е. резултати 2 и 4 трябва да се подредят в списъка преди резултат 3. Но кой от двата резултата, 2 или 4, трябва да бъде първи и кой втори в подредбата? Ако се отчитат отношенията *само* между съседните резултати, тогава отговор на този въпрос не може да бъде даден изобщо. Не може, защото не е известно какво е отношението между резултати 2 и 4. Ако обаче се отчитат отношенията „всеки със всеки“, тогава ще се знае и какво е правилното отношение между резултати 2 и 4, и това ще позволи много по-точно да се сравни автоматичната подредба на резултатите с тази, предоставена от потребителите.

Другият проблем при отчитането на отношенията само между съседни резултатите е, че не позволява да се разграничи замяната на два съседни резултата (което се счита за малка грешка) от замяната на два доста по-отдалечени резултата (което е значително по-голяма грешка). При отчитането на отношенията между всички двойки резултати, такова разграничение е възможно. Нека, например, системата да връща 200 подобни документи. Това означава 39800 отношения между отделните резултати. Ако са разменени два съседни резултата, примерно 5-тия и 6-тия, грешката ще бъде 2/39800. Ако обаче са разменени 5-ти и 80-ти, което се счита за значително по-сериозна грешка, тогава отношенията с всички 75 междинни резултати ще бъдат различни при автоматичната подредба, спрямо референтната. Така грешката ще бъде (2*75)/39800, т.е. 75 пъти по-голяма, което в действителност е правилно. Правилно е защото размяната на два съседни резултата не е особен проблем, но изместването на адекватни в края на списъка и на неадекватни в началото, е проблем.

4. Модели за представяне и методи за описание на документите

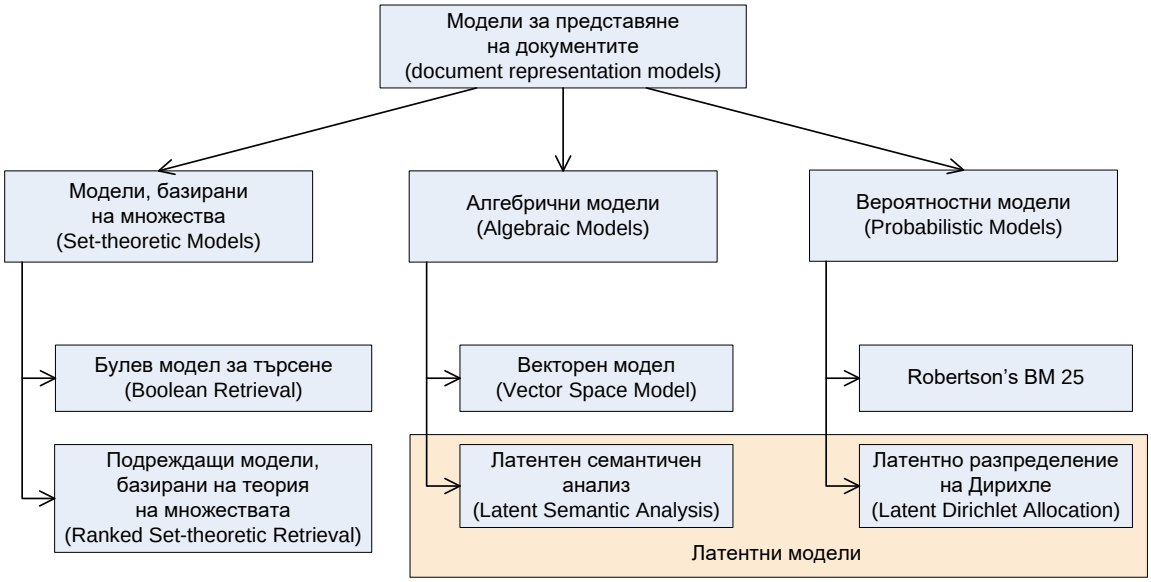
За да се изчисли степента на подобие между заявката и отделните документи е необходимо да се намери някакъв тип *семантично представяне* и на двете – и на заявката, и на документите. Избраният модел на представяне трябва да пресъздава *семантиката, или смисъла,* на документа и същевременно с това да позволява автоматизираното му обработване с помощта на математически средства. Целта е в крайна сметка с помощта на една или множество математически формули да се достигне до едно число, което показва доколко свързан е един документ с друг документ. Този показател може да се нарече *коефициент на подобие* или *семантична близост* между документите. Неговото/нейното изчисляване е пряко следствие от избрания модел за представяне на документите.

Моделите за представяне на документи (заявката също е документ) могат да се разделят на 3 големи групи:

- ***Модели, базирани на теория на множествата***. Документите се представят чрез *множество от краен брой характеристики* - най-често думи. Това множество може да бъде неподредено, но обикновено се използват *характеристични вектори*. *Семантичната близост* между документите се изчислява с помощта на метрики, познати от теорията на множествата.
- ***Алгебрични модели***. Документите се представят чрез вектори, матрици или кортежи, съдържащи реални числа. Тези числа в общия случай представляват *теглата на думите (term weights)* в документа. Теглата отразяват както локални, така и глобални характеристики - т.е. доколко дадената дума е важна за описанието на документа, но и какво е значението ѝ в рамките на цялата колекция от документи. *Семантичната близост* между документите обикновено се изчислява като косинуса на ъгъла между векторите, но може и по други начини.
- ***Вероятностни модели***. Процесът на търсене и извличане се разглежда като вероятностен извод - оценява се вероятността даден документ да бъде свързан с потребителската заявка. И тук документите се представят чрез вектори или матрици, но числата в тях представляват вероятността съответните думи да се срещат в

документите. В този смисъл, вероятностните и алгебричните модели много си приличат. Единствената разлика между тях е начинът на изчисляване на елементите на векторите. Но в крайна сметка вероятностите също се определят на база на статистически характеристики като честота на повява и др.

На фигура 2 е представена класификация на моделите, заедно с някои техни конкретни представители. По-голямата част от тях ще бъдат подробно разгледани, реализирани и експериментално изследвани.



Фигура 2. Класификация на моделите за представяне на документи

На фигурата прави впечатление, че латентния семантичен анализ и латентното разпределение на Дирихле са групирани заедно, като групата е наречена „Латентни модели“. При останалите модели, всички думи (или характеристики), описващи семантиката на документа се третират като независими. Например думите „шапка“ и „капела“ ще се обработят като съвсем отделни и несвързани, макар и в реалността да имат синонимно значение. При латентните модели не е така. Те позволяват да се откриват взаимовръзки, не само между документите, но и между отделните думи. Така думите, които имат подобно значение се групират в по-общи латентни теми (latent topics). Темите се наричат латентни или скрити, тъй като не са явно посочени в документите, но се извличат от думите, имащи подобен смисъл. Например думите „ракета“, „совалка“, „космос“, „сонда“, „планета“ и т.н. образуват космическа тема. Приема се, че даден документ е свързан с космическата тема, ако съдържа която и да е от тези думи (една или повече). Така два документа могат да се определят като семантично-свързани, дори и да не съдържат едни и същи думи, а само взаимосвързани такива. По този начин латентните модели успешно се справят със синонимите и отчасти с полисемията – различните значения на една и съща дума.

Пряко свързани с моделите за представяне на документи са и *методите за тяхното описание*. Най-общо те биват:

- *Явни*. При тях, *от потребителите се изисква* явно да предоставят допълнителна информация, която да опише съответните документи. Тези методи са задължителни,

в случая когато не е възможно от ресурсите автоматизирано да се извлекат подходящи описатели/характеристики. Но могат да се използват и в много други случаи. Описанието може да се извърши чрез предварително дефинирани *ключови думи* или анотация в *свободен текст*.

- *Неявни*. Не е необходимо потребителите да правят каквото и да било, а описанието на ресурсите се получава чрез *анализ на тяхното съдържание*. Подходящи са за текст, изображения, до голяма степен аудио и видео съдържание, но не и за описание например на хора или предмети.

Част от ресурсите, например текстови документи, изображения, аудио и видео файлове, могат да се опишат както явно, така и неявно. За други обаче явното описание е единственото възможно. Например, ако е необходимо да се опишат интересите на хора, за които липсва информация и няма от къде да се получи автоматично. Тогава всеки човек трябва явно да предостави тази информация за себе си. Може да го направи като избере интересите си от предварително фиксиран списък с ключови думи; или от таксономия от ключови думи; или сам да въведе съответните ключови думи; или пък да се опише в свободен текст. Всички тези варианти са възможни, но избраният метод за описание на обектите до голяма степен ще предопредели и модела за представянето им, както и начина за изчисляване на семантичната близост между тях.

При подготовката на този файл са използвани материали от:

Калмуков, Й. Методи и алгоритми за търсене и извличане на документи. Издателство Primax Русе, 2022 г., ISBN 978-619-7242-93-5