Описание на документи/ресурси чрез ключови думи.
Търсене по ключови думи,
избрани от предварително дефинирана структура

Описание чрез неподредено множество от ключови думи

Метод за описание: явен

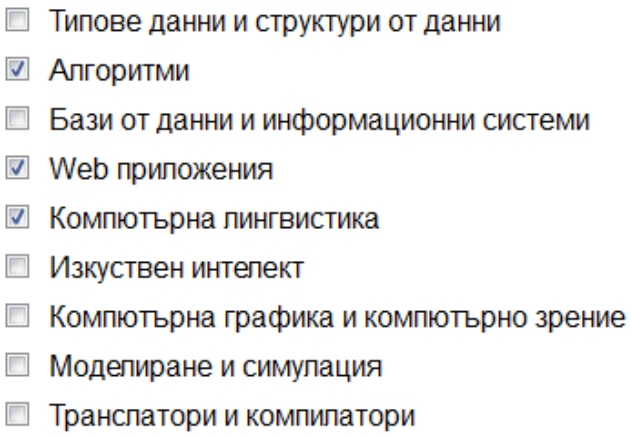
Модел за представяне на документите: и трите са възможни, но обикновено се използват моделите, базирани на теория на множествата.

Обектите се описват отделно и независимо едни от други с помощта на ***неподредено множество от ключови думи***. В някои литературни източници се нарича още *списък от ключови думи*, тъй като понятието списък е по-познато и близо до потребителите. От техническа гледна точка обаче това не е коректно, защото списъчната структура по същество е подредена.

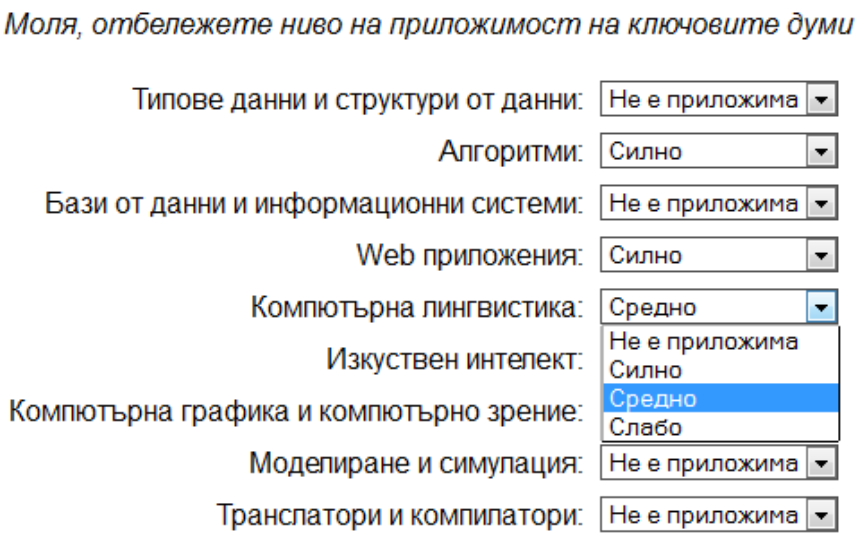
В началото, преди да започне регистрацията на документите, администраторът или който друг има права за това, дефинира неподредено множество от определен брой (колкото са необходими, но не прекалено много) ключови думи. При регистрацията на нов документ се избират само тези ключови думи, които най-точно го описват.

Методът може да се реализира по следните два начина:

1. Чрез избор на ключовите думи посредством HTML полета за маркиране (checkboxes) (фигура 1). В този случай ключовите думи имат „бинарно" поведение – всяка една от тях или е избрана или не е; или присъства в описанието на обекта или не.
2. Чрез избор на ключовите думи с помощта на HTML падащи менюта (select boxes). За всяка ключова дума съществува по едно падащо меню (фигура 2), което позволява на потребителя да посочи не само дали тя е приложима за описание на съответния ресурс, но и да отбележи колко точно е приложима. Така се въвежда *степен на приложимост / значимост* на дадената ключова дума към описанието на конкретния обект.



Фигура 1. Описание на обект чрез избор на ключови думи от неподредено множество, посредством HTML полета за маркиране



Фигура 2. Описание на обект чрез претеглен избор на ключови думи от неподредено множество, посредством HTML падащи менюта

Ако ключовите думи се избират от списък от checkbox-ове, има няколко лесни варианта да се изчислят коефициентите на подобие между всеки два документа (или между заявката и всеки от документите):

1. **Просто съвпадение (simple match)**

$$Sim(d_i, d_j) = |KW_i \cap KW_j| \quad (1)$$

където:

$Sim(d_i, d_j)$ – коефициент на подобие между документите d_i и d_j.

KW_i – множество от ключовите думи, описващи *i*-тия документ.

KW_j – множество от ключовите думи, описващи *j*-тия документ.

Т.е. уравнение (1) връща *броя на общите ключови думи*.
Ако някой не помни, символът ∩ означава сечение, а правите черти, които заграждат всичко, връщат броя.

Проблемът на простото съвпадение е, че получената стойност не е нормализирана в определен интервал, примерно [0, 1]. Ако имате два обекта, описани с по 2 напълно съвпадащи ключови думи, и други два обекта, описани с по 3 напълно съвпадащи ключови думи, то се оказва, че втората двойка обекти, всъщност е по-подобна от първата двойка. А това не е вярно, защото и при двете двойки, подобието е 100%.

2. **Коефициент на Jaccard**

$$Sim_{Jaccard}(d_i, d_j) = \frac{|KW_i \cap KW_j|}{|KW_i \cup KW_j|} \quad (2)$$

Тук всички означение са вече известни.

Коефициентът на Jaccard изчислява степента на подобие като отношение между *броя на общите ключови думи* и *броя на всички уникални* (за двете множества) ключови думи.

Стойността му е нормализирана в интервала [0, 1], което е предимство спрямо простото съвпадение.

3. Коефициент на Dice

$$Sim_{Dice}(d_i, d_j) = \frac{2 \times |KW_i \cap KW_j|}{|KW_i| + |KW_j|} \quad (3)$$

Според коефициента на Dice, подобието се намира като *2 пъти броя на общите се раздели на броя на всички (само за двете множества) ключови думи*. Тук в знаменателя дублиранията не се филтрират както е при Jaccard!

Ако в първото множество има **5** ключови думи, а във второто **3**, и **2** от тях съвпадат между множествата, то знаменателят при Dice ще бъде 8 (5+3), а при Jaccard 6 (5+1). При Jaccard ще бъде 5+1, защото 2 от ключовите думи в KW_j вече ги има в KW_i.

Т.е. при описания случай подобието според Jaccard ще бъде 2/6 = 0.33, а според Dice 4/8 = 0.5.

Очевидно и коефициентът на Dice е нормализиран в интервала [0, 1].

Saif Mohammad и Graeme Hirst научно са доказали, *че отношението между два коефициента на подобие, изчислени по формулата на Jaccard се запазва и в случая, когато коефициентите се преизчислят по формулата на Dice*. Да, има разлика в абсолютните стойности, както видяхме и на примера по-горе, но отношенията между отделните коефициенти се запазват. Затова, ако не ви интересуват абсолютните стойности, а само отношенията между коефициентите, тогава *няма значение* дали ползвате мярката за сходство на Jaccard или на Dice. При търсенето и ранкирането на резултатите, по принцип не ни интересуват особено абсолютните стойности на подобията, а само отношенията между тях. Резултатите се подреждат на база отношения. Най-големите коефициенти, най-отгоре. Без значение каква точно е стойността.

В литературата коефициентът на Dice се среща още и като индекс на Sørensen. Двамата учени са го предложили независимо един от друг, но Dice го е публикувал първи. Ей затова като измислите нещо хубаво, публикувайте го!

Ако ключовите думи са претеглени и за тях са посочени ***нива на приложимост*** (както е на фигура 2), е задължително мярката за сходство да отчете не само броя на съвпадащите ключови думи, но и съответните им нива. Как може да стане това? Ами – когато 2 ключови думи (по една от множество) съвпаднат, но нивата им на приложимост не са максимални, *няма да броите съвпадението като 1 пълно съвпадение, а ще го броите за 0.x съвпадения*. Като ще намалявате единицата обратно-пропорционално на нивото на приложимост, т.е. колкото по-голяма е приложимостта на думата, толкова по-малко ще намалявате.

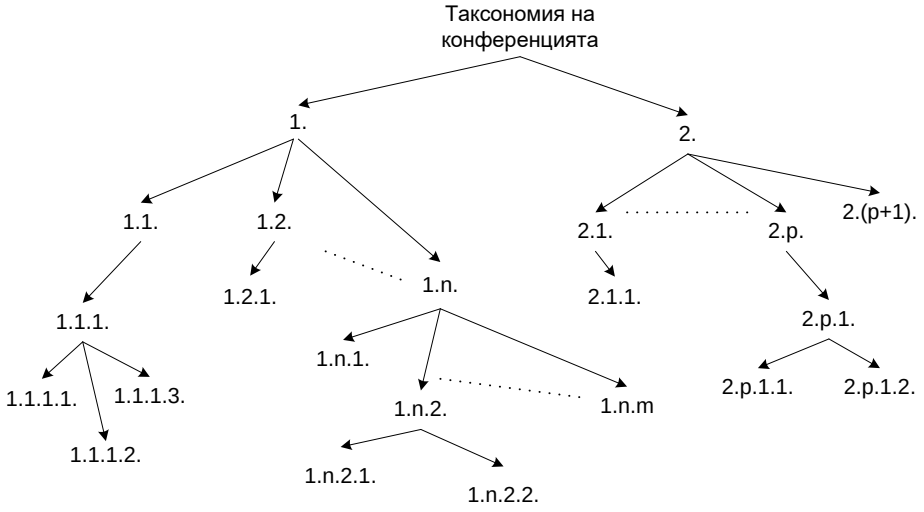
$$|KW_i \cap KW_j| => \sum_{k_m \in KW_i, k_n \in KW_j} (1 - (1 - w_m) - (1 - w_n))$$

Описание на документите/ресурсите чрез таксономия от ключови думи

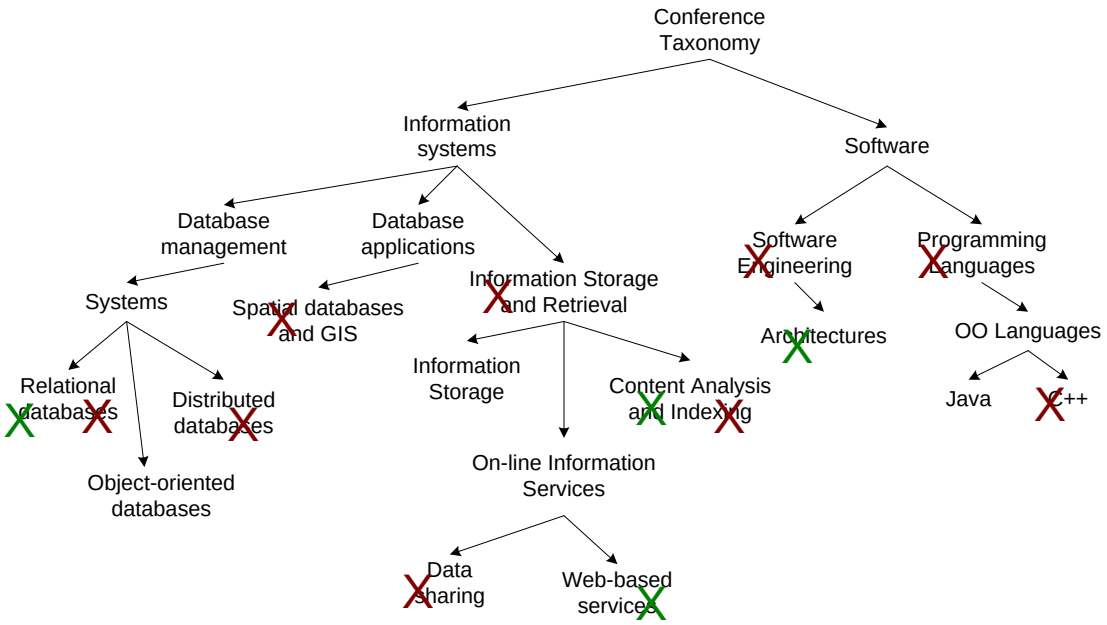
Изборът на ключови думи от предварително дефинирано, *неподредено множество* описва индивидуално и независимо всички документи. Но неподреденият характер на множеството изисква *размерът му да бъде ограничен до разумен* (не голям) *брой* семантично непрепокриващи се елементи - например 20 до 30. В противен случай, потребителският интерфейс ще бъде много неудобен за работа. Ако множеството от възможни *ключови думи съдържа стотици елементи* и те не са структурирани по никакъв начин, на потребителя ще му бъде *необходимо изключително много време*, за да ги прочете и да избере тези, които най-добре описват съответния ресурс. Това би могло силно да го демотивира и той да се откаже, или да направи повърхностен избор. От друга страна, *малкият брой ключови думи води до липса на конкретика* в тях или невъзможност напълно да се обхванат тематичните области на документите.

Посоченият проблем може до голяма степен да бъде решен, ако изборът на ключови думи се прави не от неподредено множество, а от ***таксономия***. Предимството на този метод е пряко следствие от *йерархичната структура* на таксономията. Тя предоставя допълнителна и много важна информация – семантичните връзки между отделните ключови думи, което позволява:

1. Мерките за сходство да отчитат ***не само броя на точно съвпадащите*** ключови думи, но и ***семантичната близост между несъвпадащите***.
2. Да се изчисли ***ненулев коефициент на подобие*** между два документа, дори когато ***те не споделят нито една обща ключова дума***.
3. В таксономията да участват *много повече ключови думи* - стотици, дори хиляди. По-големият им брой осигурява ***по-детайлно и точно описание*** на обектите, без това да доведе до неудобство при работа с потребителския интерфейс. Елементите са групирани в обобщени разклонения в дървото и потребителят "отваря" само тези клони, които го интересуват.



Фигура 3. Обща структура на таксономията



Фигура 4. Примерна таксономия с избраните ключови думи, които описват единия документ (маркирани със зелено) и и другия документ (маркирани с тъмно червено).

Съгласно фигурата, ключовите думи, които описват двата документа са:

KW_i = {Relational databases, Content analysis, Web-based services, Architectures}

KW_j = {Relational databases, Distributed databases, Spatial DB & GIS, Information storage and retrieval, Content analysis, Data sharing, Software Engineering, Programming languages, C++}

Въпреки, че те се съхраняват в неподредени множества, семантичните връзки между тях не се губят и могат по всяко време да бъдат извлечени от таксономията.

Документите обикновено ще се описват не с по една, а с по множество ключови думи, избрани от предварително дефинираната таксономия. Затова, за да се изчисли точно степента на подобие между тях е необходимо да се използва мярка за сходство, която намира семантичната близост между две множества от възли (концепти) в обща таксономия. Тази долу (4) аз съм я предложил. Тя се базира на коефициента на Dice и се извежда от него. Само, че *вместо броя на точно съвпадащите ключови думи, отчита и семантичната близост между несъвпадащите*. Защо това е възможно? Ами защото въпросните ключови думи са йерархично свързани в таксономията, а не са независими. И примерно в зависимост от дължината на пътя между тях, може да се намери семантичната близост помежду им.

$$Sim(d_i, d_j) = \frac{\sum_{k_m \in KW_i} \max_{k_n \in KW_j} (Sim(k_m, k_n)) + \sum_{k_n \in KW_j} \max_{k_m \in KW_i} (Sim(k_n, k_m))}{|KW_i| + |KW_j|} \quad (4)$$

Където:

k_m – m-тата ключова дума, описваща i-тия документ

k_n – n-тата ключова дума, описващата j-тия документ

KW_i – множество от ключовите думи, описващи i-тия документ.

KW_j – множество от ключовите думи, описващи *j*-тия документ.

$Sim(k_m, k_n)$ – семантична близост между *m*-тата ключова дума, описваща *i*-тия документ и *n*-тата ключова дума от *j*-тия документ.

$\max_{k_n \in KW_j}(Sim(k_m, k_n))$ – семантична близост между *m*-тата ключова дума, описваща *i*-тия документ и семантично най-близката ѝ ключова дума, описваща *j*-тия документ.

Ако таксономията се преобразува в неподредено множество, чрез игнориране на семантичните връзки между отделните елементи, формула (4) винаги ще дава абсолютно същия резултат като коефициента на Dice (3). Но за разлика от него, (4) може да бъде използвана и при множества, чиито елементи са семантично свързани.

За да се изчисли $Sim(d_i, d_j)$ трябва преди това да се изчислят всички подобия $Sim(k_m, k_n)$ между всички ключови думи, описващи единия документ и всички ключови думи, описващи другия документ. Това може да стане по един от следните два начина:

- на база на структурните характеристики на таксономията - разстояние, дълбочина, плътност и др.
- на базата на информационното съдържание на възлите.

Една от широко използваните мерки за определяне на семантичната близост между два възела в таксономия е формулираната от *Zhibiao Wu u Martha Palmer*.

$$Sim_{Wu\ \&\ Palmer}(k_m, k_n) = \frac{2 \times N_0}{2 \times N_0 + N_1 + N_2} \quad (5)$$

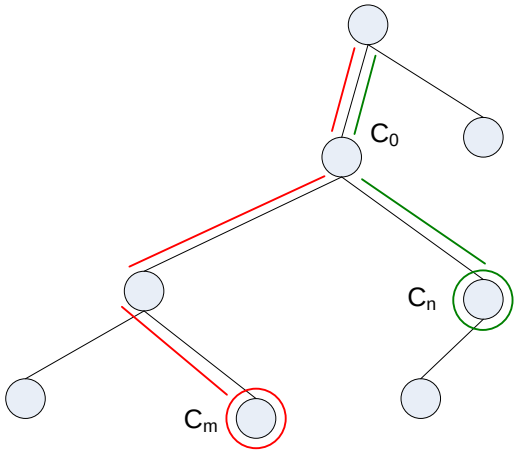
където:

N_0 - разстоянието (в брой ребра/дъги) между корена и най-близкия общ родител C_0 на двата възела (C_m) и (C_n), между които се търси семантичната близост (фиг. 5).

N_1 - разстоянието от C_0 до единия от възлите, примерно C_m. C_m представлява *m*-тата ключова дума от *i*-тото множество.

N_2 - разстоянието от C_0 до другия възел – C_n. C_n представлява *n*-тата ключова дума от *j*-тото множество.

Тъй като мярката за сходство на Wu и Palmer е симетрична, няма значение дали C_m принадлежи на *i*-тото множество, а C_n на *j*-тото или обратното. При по-внимателно вглеждане в (5) се вижда, че тя всъщност представлява коефициент на Dice приложен върху множествата от ребра, изграждащи пътищата от корена до двата възела, между които се търси семантична близост.



Фигура 5. Визуално представяне на мярката за сходство между два възела в таксономия на Wu и Palmer

Dekang Lin също предлага мярка за сходство, базираща се на информационно съдържание на възлите. В случая таксономията се допълва с функция $p:C \rightarrow [0,1]$, такава че за всеки възел (концепт) $c \in C$, $p(c)$ представлява вероятността в таксономията да се срещне възелът c или някой негов наследник. Т.е. ако c_1 е в отношение "IS-A" с c_2, тогава $p(c_1) <= p(c_2)$. От тук следва, че вероятността на корена (ако има такъв) е 1, защото всеки възел е негов наследник. Тъй като по-ниската вероятност означава по-високо информационно съдържание, възлите по-дълбоко в йерархията са по-информативни от тези, намиращи се по-плитко. Мярката за сходство на Lin (6) е подобна на тази на Wu и Palmer, но вместо разстояния, отчита информационното съдържание на възлите.

$$Sim_{Lin}(k_m, k_n) = \frac{2 \times \log P(C_0)}{\log P(C_m) + \log P(C_n)} \quad (6)$$

където

$P(C_0)$ - вероятността в таксономията да се срещне най-близкият общ родител C_0 (на възлите, чиято семантична близост се изчислява) или някой негов наследник.

$P(C_m)$ - вероятността да се срещне възелът (или негов наследник), представляващ *m*-тата ключова дума от едното множество.

$P(C_n)$ - вероятността да се срещне възелът (или негов наследник), представляващ *n*-тата ключова дума от другото множество.

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