5. ІНФОРМАЦІЙНО-ПОШУКОВІ МОВИ. ДЕСКРИПТОРНИЙ ПОШУК

5.1. Лінгвістичне забезпечення інформаційного пошуку

Лінгвістичне забезпечення – це комплекс засобів, які використовуються для формування, обробки, зберігання та пошуку інформації, а також засобів і методів їх створення, використання та контролю. В бібліотечній діяльності, наприклад, склад та структура лінгвістичного забезпечення залежить від того, в якій сфері воно використовується. Можна виділити 4 функції, покладені на лінгвістичне забезпечення:

1) перетворююча,

2) структуруюча,

3) пошукова,

4) ідентифікаційна.

Виділені функції були класифіковані. Так, перетворююча і структуруюча функції діляться на формальну і семантичну, пошукова - відповідно до типів виконуваних запитів – на адресну, уточнюючу, широко- і вузькотематичну, фактографічну. Ідентифікаційна функція не має різновидів, оскільки її виконання пов'язане з порівнянням ознак документа з елементами бібліографічного опису і кодами, які ідентифікують документ.

Перетворююча функція лінгвістичних засобів – це функція, яка передбачає формальну чи семантичну аналітико-синтетичну обробку документальної та фактографічної інформації за допомогою лінгвістичних засобів.

Пошукова функція лінгвістичних засобів – це функція, яка полягає у здійсненні пошуку документальної або фактографічної інформації за допомогою лінгвістичних засобів , відповідно до типів інформаційних запитів.

Структуруюча функція , або функція структурування лінгвістичних засобів – це функція , яка полягає у застосуванні лінгвістичних засобів для групування документальної або фактографічної інформації з формальних і/або семантичними ознаками .

Ідентифікаційна функція лінгвістичних засобів – це функція, яка передбачає порівняння ознак документа з елементами його бібліографічного опису і/або мовами-ідентифікаторами з метою встановлення повного їх збігу.

Лінгвістичні засоби – це сукупність трьох складових:

1. інформаційно-пошукових мов (ІПМ),
2. правил переводу інформації з природної мови на ІПМ і
3. критерію подання інформації.

Ці складові забезпечують подання змістового наповнення документів, що знаходяться у інформаційних масивах, у вигляді, який є зручним для машинної обробки, а також забезпечують автоматизований пошук інформації.

До складу лінгвістичних засобів входять класифікаційні (ББК – Бібліотечно-бібліографічна класифікація, УДК – Універсальна десяткова класифікація і т. д.) і дескрипторні інформаційно-пошукові мови.

2. Поняття та види інформаційно-пошукових мов

http://www.gpntb.ru/win/book/1/Doc20.HTML

http://www.gpntb.ru/win/book/

http://citforum.ru/internet/services/services0312.shtml

http://www.glossary.ru/cgi-bin/gl_sch2.cgi?RIt%28uwsg.outtu-vuoxqui:l!9n:qo – глосарій

http://vova1001.narod.ru/00009602.htm

5.2.1. Інформаційно-пошукові мови: структура та основні характеристики.

Існує два шляхи інформаційного обслуговування. Один із шляхів інформаційного обслуговування передбачає **компресію первинних документів** у формі вторинних (*рефератів, анотацій, рецензій*), в яких виявлено найбільш важлива (із загальноприйнятої точки зору) інформація, яка на регулярній основі доводиться до користувачів.

Другий шлях полягає у пошуку в документах інформації, яка є важлива в конкретному випадку для задоволення поточної потреби користувача, необхідну

для виконання користувачем своєї функції, незалежно від суспільної значимості інформації.

У такому випадку користувач повинен сформулювати свою інформаційну потребу і надати пошуковій системі відповідний запит. Інформаційно-пошукова система аналізує наявну інформаційну базу, знаходить інформацію, що відповідає запиту, і видає її користувачеві. Ця процедура, яка називається інформаційним пошуком, є найважливішим видом інформаційної діяльності поряд з аналітико-синтетичною обробкою документів.

У цій процедурі виникають дві пов'язані проблеми. Для того щоб викласти свою інформаційну потребу, потрібні засоби, які зрозумілі пошуковій системі. Іншими словами, потрібна особлива мову для формулювання запитів – мова запитів. З іншого боку, документи повинні бути заздалегідь описані на мові опису документів. Потреба в деякій мові опису запитів і документів значно посилюється, коли йдеться про інформаційний пошук в автоматизованому режимі. У зв'язку з цією потребою виникло поняття «інформаційно-пошукова мова».

Інформаційно-пошукова мова (ІПМ) (англ. *information retrieval language, indexing language*) – штучна мова, призначена для формалізованого опису смислового змісту документів, даних, окремих понять або термінів і забезпечення подальшого їх пошуку в інформаційно-пошукових масивах.

ІПМ є основним елементом логіко-семантичного апарата інформаційно-пошукових систем (ІПС).

Необхідність створення штучної мови для виразу смислового змісту інформації з метою її пошуку обумовлена тим, що природна мова, тобто мова, за допомогою якої ми спілкуємось, має такі властивості, котрі заперечують її використання для запису та пошуку інформації.

До таких властивостей головним чином відноситься неоднозначність та багатозначність термінів природної мови. Неоднозначність термінів виражається в синонімії, коли для вираження одного і того ж поняття використовуються різні слова. Неоднозначність та багатозначність термінів

природної мови (омонімія, полісемія) не дозволяють визначити точної відповідності між змістом документа і засобами вираження його змісту.

Наприклад, лексична синонімія – *формування, комплектування* (фондів) і граматична синонімія – *студенти, студентство*. Багатозначність слів проявляється і в омонімії – *атлас* – збірник географічних, історичних та ін. карт, *атлас* – тканина тощо.

Необхідність створення штучної мови визначається і такою властивістю природної мови, як неформалізованість зв'язків слів у ній. *Наприклад*, слова: *дерево, липа* – не мають формальних ознак, що відображають їх родовидовий зв'язок тощо.

Перелічені вище властивості природної мови і обумовили створення штучної ІПМ, *основними вимогами* до якої є:

– однозначність (тобто, кожен запис ІПМ повинен мати тільки одне змістовне значення, пояснення, однакову форму запису засобами ІПМ);

– достатня семантична сила (тобто, здатність відображати з необхідною повнотою та точністю смисловий зміст документів та запитів визначеної предметної області);

– відкритість (тобто, забезпечення можливості корегувати мову).

Отже, формалізація лексики і створення різних ІПМ спричинене необхідністю усунення "надмірності" і "недостатності" природної мови для цілей інформаційного пошуку, а також ліквідації присутній у ній синонімії та омонімії для реалізації "однозначності" інформаційного пошуку.

Структура ІПМ:

– Лексика (словниковий склад). Лексичними одиницями в ІПМ є: лексика природної мови (слова, словосполучення, рубрики), абревіатури, цифрові коди, символи, дати, шифри слів і словосполучень. Сукупність усіх лексичних одиниць мови має назву *словниковий склад.*

– Базисні (аналітичні) відношення виражають зв'язок або структуру мови. Ці відношення ще мають назву *парадигматичні*. Парадигматичними є відношення: рід – вид, ціле – частка, предмет – функція і т. п.

– Граматика ІПМ – сукупність правил формування пошукових образів і пошукових приписів із лексичних одиниць ІПМ. Ці відношення мають назву *синтагматичні*. Синтагматичними є смислові, текстуальні відношення між лексичними одиницями ІПМ, що описують зміст документа або запиту.

– Алфавіт ІПМ представляє собою систему позначень, інде-ксацію лексичних одиниць мови. Як алфавіт ІПМ використовують літерні та цифрові знаки, спеціальні символи.

– Система редагування мови (вилучення застарілих слів, рубрик, зв'язків).

– Правила інтерпретації – правила перекладу з природної мови на ІПМ і навпаки.

Основні характеристики ІПМ:

- *Семантична сила* – здатність ІПМ виражати з необхідною повнотою та точністю смислове значення інформації, що обробляється.
- *Лексична повнота та точність* – характеризує стан слов-ника ІПМ і відповідність його лексиці визначеної предметної області.
- *Парадигматична повнота та точність* – здатність ІПМ достатньо повно та точно відображати інформацію про предмети та поняття, що перебувають в постійних родовидових та інших парадигматичних відношеннях.
- *Синтагматична повнота та точність* – здатність ІПМ повно та точно відображати інформацію про ситуативні відношення, тобто відношення між предметами та поняттями, що виникають у контексті.

5.2.2. Види ІПМ

Нині існують різні класифікації видів та типів ІПМ.

Найпоширенішою є видова класифікація, в основу якої покладені ознаки, які враховують структурні елементи мови. До цих ознак відносяться:

1)спосіб задання лексичних одиниць,

2)спосіб координації (сполучення) лексичних одиниць та

3)спосіб врахування парадигматичних відношень.

Крім того, береться до уваги і тип інформаційного пошуку (традиційний або автоматизований), на який орієнтована та чи інша ІПМ.

1) За способом задання лексичних одиниць: контрольовані і неконтрольовані. **Контрольовані ІПМ** – мови, лексика яких задається заздалегідь за допомогою словників і таблиць (УДК , ББК). **Неконтрольовані ІПМ** – мови, лексика яких не задається словником, а будується на основі вибору необмеженої кількості термінів природної мови з індексованих ресурсів.

2) За координацією лексичних одиниць (способом запису): некоординатні і координатні.

Некоординатні ІПМ – мови, що не допускають координації своїх лексичних одиниць ні в процесі індексування, ні в процесі пошуку.

До **координатних ІПМ** відносяться мови, в яких лексичні одиниці координуються між собою або в процесі індексування, або в процесі пошуку. В залежності від цього розрізняють два види координатних ІПМ: передкоординатні та посткоординатні (або традиційні та нетрадиційні) [73, 103].

Передкоординатні ІПМ – це мови, в яких порядок сполучення лексичних одиниць в ПОД (ПОЗ) встановлюється у процесі індексування за визначеними правилами (наприклад, правилами систематизації, предметизації, бібліографічного опису) та передбачають їх послідовність. До ІПМ такого типу відносяться: ієрархічні та фасетні класифікації, мова предметних рубрик, мова бібліографічного опису. Як правило, передкоординатні ІПМ використовуються при ручному пошуку.

Посткоординатні ІПМ – це мови, в котрих лексичні одиниці ПОД (ПОЗ) задаються переліком в процесі індексування. Запис індексів дозволяє здійснити багатоаспектний пошук інформації. До цих мов відносяться дескрипторні мови, семантичні коди та синтагматичні мови. Це мови, що використовуються при автоматизованому пошуку. В даний час відповідно до стандартів визначають такі види ІПМ, як: *ІПМ предметизації, класифікаційна та дескрипторна ІПМ, вербальні ІПМ* та ін. [17, 21].

Мовою предметних рубрик (ІПМ предметизації) є ІПМ, що призначена для індексування документів (частин документів) та інформаційних запитів засобами предметних рубрик. Суттю предметизації є визначення предметної рубрики, що відповідає змісту, формі та призначенню документа.

В основі **класифікаційної ІПМ** лежить систематична класифікація, що відображає смислові відношення між поняттями. В систематичних класифікаціях лексичні одиниці задаються системою понять, котрі виражаються словами, словосполученнями та реченнями. Класифікаційні мови призначені для формалізації логічних зв'язків слів природної мови.

В основі **дескрипторної ІПМ** лежить алфавітний перелік лексичних одиниць, що виражаються словами або словосполученнями. Ця ІПМ призначена для контрольованого індексування, тобто для контролю за термінологією, що використовується при індексуванні. Під контролем лексики розуміється вилучення неоднозначності слів природної мови.

3) З урахуванням і без урахування парадигматичних відношень: ієрархічні, фасетні та неієрархічні.

Мови ієрархічної структури являють собою ієрархічну класифікацію – систему класів, за якими розподіляються поняття на підставі найбільш істотних ознак, притаманних цим поняттям і, які відрізняють їх один від одного. **Клас** – сукупність об'єктів, що мають один або кілька загальних змістовних ознак. Недолік – неможливість організації позаієрархічних зв'язків (ієрархічна модель даних – неможливість перехресних запитів).

Мови фасетної структури являють собою фасетну класифікацію – сукупність фасетів, що йдуть один за одним у певній послідовності. **Фасет** – одна зі сторін об'єкта, що розглядається, обмежена сукупність однорідних значень за деякою класифікаційною ознакою. Наприклад, фасет колір містить значення: *червоний, білий, зелений, чорний, жовтий*. В основі побудови такої класифікації лежить індуктивний метод, який передбачає дослідження окремих предметів множини, знаходження в них загальних суттєвих ознак і групування на основі останніх термінів, що позначають ці поняття.

Мови неієрархічної структури є продуктом емпіричного підходу до побудови ІПМ. Такий підхід передбачає створення мови на основі індексування реальних текстів. Зі змісту документа вибираються ключові слова, які перетворюються на лексичні одиниці ІПМ. До спеціальних неієрархічних ІПМ з урахуванням сильних і слабких парадигматичних відношень належить дескрипторна мова.

5.2.3. ІПМ предметизації

Предметизація базується на основі предметних класифікацій документів. Предметизація документів використовується при створенні предметних каталогів, картотек, абетково-предметних покажчиків, побудові предметних рубрик.

У вузькоспеціальному значенні предмет розглядають як поняття чи комбінацію понять, що відбивають тему документа.

Окремі теми (предмети) документів відтворюються в предметних рубриках. **Предметна рубрика** – це коротке формулювання теми документа. Суттю предметизації і є визначення предметної рубрики, що відповідає змісту, формі та призначенню документа.

Мова предметних рубрик (ІПМ предметизації) – це ІПМ, що призначена для індексування документів (частин документів) та інформаційних запитів засобами предметних рубрик.

ІПМ предметних рубрик створюється на основі природної мови, і пошукові образи документів представлені у вигляді предметних рубрик. Лексичними одиницями ІПМ виступають *заголовки* та *підзаголовки* предметних рубрик.

Заголовок – це назва головного предмета, документа.

Підзаголовки розкривають аспекти, у яких розглядається предмет, а також форму і призначення документа.

Сукупність лексичних одиниць складає **словник предметних рубрик**, тобто лексику ІПМ предметизації, а правила її побудови – **граматику**. Словник вміщує строго зафіксований на кожний момент часу набір заголовків і підзаголовків, між якими встановлюються певні змістовні й текстуальні зв’язки.

Лексика ІПМ може бути контрольованою (створення списків предметних рубрик конкретних інформаційно-пошукових систем (ІПС)) і неконтрольованою (коли здійснюється вільне формулювання предметних рубрик).

Контрольована лексика властива предметним каталогам і картотекам; неконтрольована – допоміжним предметним покажчикам. У залежності від умов функціонування словник може поповнюватись новими лексичними одиницями, чи скорочуватись, змінювати свою структуру.

Предметна рубрика, розкриваючи зміст документа, виконує такі функції:

- ***Інформаційна функція*** реалізується у тому, що предметна рубрика у згорнутій формі інформує про зміст документа.
- ***Евристична функція*** полягає у тому, що предметна рубрика дає змогу знайти потрібний документ, оскільки пошук значно полегшується завдяки структурі предметних класифікацій, де предметні рубрики розташовані за абеткою.
- ***Комплексна функція*** проявляється у тому, що предметні рубрики створюють комплекси, де збираються документи, присвячені певному предмету, незалежно від галузевого аспекту його розгляду.
- ***Термінологічна функція*** проявляється в тому, що при формуванні предметних рубрик використовують сучасну наукову термінологію і тим самим сприяють опануванню нею користувачів.

Сполучення всіх предметних рубрик, які надаються одному документу, створює ПОД – текст, що термінами ІПМ відтворює зміст документа.

5.2.4. Класифікаційна ІПМ

Класифікаційна ІПМ – це ІПМ, що призначена для індексування документів (частин документів) та інформаційних запитів засобами понять та кодів будь-якої класифікаційної системи.

Аналіз ІПМ класифікаційного типу свідчить, що ці ІПМ характеризуються властивостями, які обумовлюють низьку ефективність їх використання в інформаційному пошуку, особливо з використанням технічних засобів. До таких властивостей відносяться: попередня координація (зв’язок)

слів і словосполучень у рубриці, неможливість оновлення та доповнення, труднощі використання при індексуванні.

Наприклад, Універсальна десяткова класифікація (УДК), бібліотечно-бібліографічна класифікація (ББК) є інформаційно-пошуковими мовами класифікаційного типу.

Системи класифікації в практичному використанні оформлюють у вигляді таблиць класифікації.

Таблиці класифікації – це практичний посібник для систематизаторів, у якому розкриваються принципи побудови та зміст певної класифікаційної системи. Вони призначені для визначення місця конкретного документа в процесі впорядкування документів у логічній послідовності, зв'язку та співпорядкування згідно з ознаками змісту, форми тощо. Перелічувані таблиці класифікації складаються з двох частин: основної класифікаційної таблиці та абетково-предметного покажчика (АПП).

Основна класифікаційна таблиця – це систематичний перелік усіх класів, підкласів та більш дрібних ділень класифікації, які в таблицях називаються *класифікаційними діленнями* [36]. Кожне класифікаційне ділення складається з класифікаційного індексу та словесного формулювання. В разі необхідності до нього можуть належати методичні вказівки, до яких входить і довідковий апарат.

Класифікаційний індекс – це пошуковий образ, побудований засобами класифікаційної ІПМ.

Словесне формулювання дає змогу перекласти інформацію про документ з природної мови на класифікаційну ІПМ, тобто мову індексів і навпаки. Методичні вказівки найчастіше розкривають зміст ділення, якщо зі словесного формулювання його неможливо зрозуміти.

Розглянемо УДК як приклад ІПМ класифікаційного типу.

За структурою УДК – ієрархічна класифікаційна система, яка вміщує основну таблицю, допоміжні таблиці й абетково-предметний покажчик. До її основного ряду входять такі класи:

0	Загальний відділ. Наука та знання. Організація. Інформація. Документація. Бібліотечна справа. Установи. Публікації
1	Філософія. Психологія
2	Релігія. Теологія
3	Суспільні науки. Статистика. Політика. Економіка. Торгівля. Право. Уряд. Військова справа. Фольклор
4	Вільний з 1962 р.
5	Математика та природничі науки
6	Прикладні науки. Медицина. Техніка. Сільське господарство
7	Мистецтво. Архітектура. Ігри. Спорт
8	Мова. Мовознавство. Література
9	Географія. Біографії. Історія

Кожен основний клас підрозділяється на десять відділів, кожен відділ – на десять більш детальних розділів і т. п. Індекси при цьому розглядаються як десяткові дроби.

Як правило, в десятковому ряді УДК цифри 2–8 використовуються для основних підрозділів. Цифра 1 виступає як резерв для нових понять або для відбиття питань загального характеру. Цифра 9 найчастіше означає інші або також резервується для включення нових тем.

Наприклад, дослідженням щодо функціонування веб-форумів як віртуальних спільнот присвоюється такий код УДК: 004.773.2

0	Загальний відділ. Наука та знання. Організація. Інформація. Документація. Бібліотечна справа. Установи. Публікації
00	Загальні питання науки та культури
004	Комп’ютерна наука і технологія. Застосування комп’ютера. Оброблення даних
004.7	Комп’ютерний зв’язок. Комп’ютерні мережі
004.77	Застосування комп’ютерних мереж та послуги. Застосування Інтернету
004.773	Обмін повідомленнями

004.773.2	Дошка оголошень

5.2.5. Дескрипторна ІПМ

До спеціальних ІПМ з урахуванням сильних і слабких парадигматичних відношень відноситься дескрипторна ІПМ (посткоординатна ІПМ), що покладена в основу координатного індексування і призначена для здійснення пошуку в автоматизованій пошуковій системі.

Дескрипторна *ІПМ* – це штучна мова, що була розроблена ще на початку 50-х рр. XX століття. Необхідність її створення обумовлена об'єктивними причинами. В цей час були розроблені перші механізовані системи пошуку інформації і перші спроби використання ЕОМ для вирішення різних інформаційно-пошукових задач. Традиційні мови передкоординатного типу не відповідали вимогам механізованого, а в подальшому і автоматизованого пошуку інформації. Інтенсивний пошук нових семантичних засобів привів до створення принципово нового типу ІПМ – посткоординатного [103]. В основі створення ІПМ посткоординатного типу лежить гіпотеза про те, що основними носіями інформації в будь-якому тексті є ключові слова, під якими слід розуміти всі члени речення, що несуть в собі основне смислове навантаження в тексті. Ключовими словами можуть бути не всі, а тільки деякі частини мови, що означають поняття про об'єкти інформації та дії над ними. Враховуючи те, що не тільки в природній мові, але і в ІПМ окремі ключові слова можуть бути скоординовані одне з іншим в необхідні сполучення для отримання текстів з заданим смисловим змістом, було запропоновано створити такі ІПМ, в яких лексичними одиницями були б слова, а не рубрики. Таким чином були розроблені посткоординатні ІПМ, які в практиці інформаційного пошуку отримали широке розповсюдження.

Так, американський документаліст М. Таубе в 1951 р. розробив ІПМ унітермінів, тобто неформалізованих одиничних термінів, переліком яких можна достатньо повно і точно передати зміст документа або запита. Координація унітермінів відбувається в момент пошуку інформації після

отримання запиту. Але дана мова не відповідала одній із основних вимог – однозначності запиту – і це обумовило необхідність її удосконалення.

Широке розповсюдження отримали ІПМ унітермінів зі штучною граматикою, яка має назву *дескрипторна* ІПМ. Поняття про дескриптор, як термін, що призначений для однозначного запису понять, було введено вченим К. Муерсом [103]. Він запропонував урахувати синонімію в умові унітермінів; здійснювати лексикографічний контроль за ключовими словами. Цей контроль полягав у нормалізації і вилученні неоднозначності та багатозначності термінів. К. Муерс запропонував для використання за контролем лексики спеціальний дескрипторний словник.

Дескрипторна ІПМ стала засобом автоматизованого інформаційного пошуку і є ніби підсумком розвитку предметної мови для комп'ютерних ІПС .

З впровадженням комп'ютерної технології дескрипторні ІПМ стали використовувати в універсальних бібліотеках для монотематичних баз даних (БД) і для електронних каталогів з автоматичним індексуванням документів, що відносяться до однієї галузі.

Таким чином, ***дескрипторна ІПМ*** – це ІПМ, що призначена для координатного індексування документів та інформаційних запитів засобами дескрипторів і / або ключових слів.

В основі дескрипторної ІПМ лежить алфавітний перелік лексичних одиниць, що виражені словами або словосполученнями. Ці ІПМ призначені для контрольованого індексування, тобто для контролю термінології, що використовується в процесі індексування. Під контролем за термінологією слід розуміти вилучення неоднозначності слів природної мови. Для контролю за лексикою використовується інформаційно-пошуковий тезаурус (ІПТ).

3. Інформаційно-пошуковий тезаурус

Термін *тезаурус* у перекладі з грецької мови означає *thesauri* – *скарбничка*, *словник.* Перше видання тезауруса було здійснено в 1852 р. у Британії – доктором П. Роджетом [73. Информатика: Учеб.пособие / Под ред.

К.В.Тараканова. – М.: Книга, 1996. – 304 с.]. В ньому всі терміни та фрази англійської мови були систематизовані не за алфавітом, а згідно з поняттями, котрі вони виражають. Робота по створенню подібного роду тезауруса мала лінгвістичні цілі, і його створення було розпочато ще до появи можливості машинної обробки інформації.

Інформаційно-пошуковий тезаурус (ІПТ) – контрольований словник лексичних одиниць дескрипторної ІПМ, заснований на лексиці однієї природної мови, що відображає семантичні відношення між лексичними одиницями і призначений для організації пошуку інформації шляхом індексування документів і/або запитів. В інформаційній діяльності тезаурус використовується з метою заміни різних слів та словосполучень, що виражають одну і ту ж думку, одним словом або словосполученням – **дескриптором**.

ІПТ призначений для виконання наступних функцій:

1) індексування документів і запитів, тобто перекладу їх змісту з природної мови на дескрипторну;
2) фіксування парадигматичних відношень між дескрипторами, які розширюють можливості інформаційного пошуку.

Як і будь-який словник, ІПТ служить термінологічним посібником в різних галузях знання. Кожний конкретний ІПТ відображає систему понять визначеної галузі знання у вигляді структурованої сукупності термінів.

У залежності від тематичного профілю, котрий визначає склад лексики, розрізняють: багатогалузеві, галузеві та вузькотематичні тезауруси.

Залежно від кількості використання в ІПМ природних мов, тезауруси можуть бути одно- та багатомовними.

Одномовні тезауруси – це ІПТ, що включають лексичні одиниці, які взяті з однієї природної мови.

Багатомовні тезауруси – це ІПТ, що включають лексичні одиниці, які взяті з декількох природних мов і представляють еквівалентні за смисловим значенням поняття кожною із цих мов.

В практиці використовуються також:

– ***макротезауруси*** – це ІПТ, до яких входять лексичні одиниці загального змісту та які охоплюють широку галузь знань;

– ***мікротезауруси*** – це спеціалізовані ІПТ, невеличкі за обсягом, які створені на основі вибору із більш повного ІПТ і додатково включають конкретні вузькі поняття визначеної тематики;

– ***спеціалізовані ІПТ, монотематичні ІПТ*** – побудовані для відображення галузі знання або практичної діяльності;

– ***політематичні ІПТ*** – побудовані для більш ширшої сукупності галузей знань (наприклад, для природничих наук і техніки);

– ***дескрипторний словник*** – це словник декскрипторної ІПМ, в якому наведені в загальному алфавітному ряду дескриптори та їх синоніми без урахування інших відношень між лексичними одиницями (тобто, це спрощений варіант ІПТ, в якому зафіксовані переважно або тільки синонімічні зв'язки).

У відповідності до цією методики за своєю побудовою ІПТ поділяється на два типи:

- тезауруси, що відокремлюють серед своїх лексичних одиниць дескриптори та аскриптори;
- тезауруси, де всі лексичні одиниці є дескрипторами.

Дескриптор – це ключове слово, вибране з групи умовно еквівалентних ключових слів, яке представляє цю групу при індексуванні та пошуку інформації.

Деякий поняття може бути виражене за допомогою двох або більшої кількості термінів, один з яких вибирається в якості основного терміна – дескриптора. Відповідно, дескриптор це лексична одиниця ІПТ, що призначена для використання в ПОД або в ПОЗ. У ролі дескрипторів виступають терміни, що позначають окремі поняття деякої предметної області та задовольняють принципи загальновживаності, поширеності, стислості і термінологічної точності.

Ключове слово (КС) – окреме слово або словосполучення природної мови, що виділяється з тексту інформаційного документа і відображає його основний зміст при індексуванні.

Аскриптор – лексична одиниця в ІПТ, яка не може використовуватись для координатного індексування і підлягає заміні на один або декілька дескрипторів.

Словникова стаття дескриптора (дескрипторна стаття) складається із самого дескриптора (заголовного дескриптора), а також дескрипторів і ключових слів, пов’язаних із заголовним дескриптором за змістом.

У дескрипторній статті лексичні одиниці розподілені в такому порядку:

- заголовний дескриптор;
- ключові слова (умовно синонімічні заголовному дескриптору);
- вищі дескриптори (ширші);
- нижчі дескриптори (вужчі);
- асоціативні дескриптори.

Словникова стаття ключового слова (аскриптора) складається, як і стаття дескриптора, із заголовного слова (в даному випадку аскриптора), відповідного відсилання до дескриптора, який служить його умовним еквівалентом. Більшості аскрипторів однозначно відповідає визначений дескриптор.

3.1. Види відношень в ІПТ

До дескрипторної статті в ІПТ входять лексичні одиниці, які з дескриптором пов’язані такими зв’язками:

1) відношення еквівалентності;
2) ієрархічний зв'язок;
3) асоціативний зв'язок.

Відношення еквівалентності між термінами встановлюється між дескриптором та термінами, що виражають ідентичні або подібні поняття. Відношення еквівалентності між термінами включає три підтипи:

- власне синоніми,
- лексичні варіанти,

- квазісіноніми.

Основними видами синонімів, що включаються в тезауруси, є терміни різного походження, загальновживані слова і наукові терміни, загальновживані терміни та жаргонні або діалектні вирази та ін.

Лексичні варіанти відрізняються від синонімів тим, що вони є деякою модифікацією одного і того ж виразу, наприклад різне написання, абревіатури тощо.

В якості аскрипторів часто можуть використовувати квазісіноніми, тобто такі терміни, значення яких відрізняється, але які розглядаються як еквіваленти для цілей тезауруса; наприклад, як квазісиноніми часто розглядаються антоніми (*ядерна небезпека – ядерна безпека*).

Іншим частим видом квазісінонімов є випадок, коли в якості дескриптора розглядається якийсь узагальнюючий тип, а його підвиди описують як аскриптори до цього дескриптора.

Ієрархічні відношення – виявляють логічне підпорядкування дескрипторів, один з яких визначає поняття роду або цілого (вищий дескриптор), а інший – поняття виду або частки цілого (нижчий дескриптор). Важливо зауважити, що ієрархічні відношення в інформаційно-пошукових тезаурусах встановлюються лише в тому випадку, якщо відношення є істинним незалежно від контексту.

Ієрархічні відношення в тезаурусі можуть використовуватися в трьох логічно різних і взаємно виключних ситуаціях, а саме – для встановлення наступних відношень:

- Родо-видове відношення (відношення вище-нижче). Наприклад: *вибухові речовини – тротил*;
- Відношення частина-ціле. Наприклад: *центральна нервова система – мозок, Львівська область – Львів*;
- Відношення екземпляр-клас. Наприклад, у фразі: *“Виделка входить в стандартний набір столового посуду”* термін *“виделка”* зв’язаний з терміном *“столовий посуд”* відношенням *“екземпляр-клас”*.

Асоціативні відношення – це смислові зв’язки предметів, об’єктів, дій, які виражаються дескрипторами. До асоціативних відносять також і деякі невраховані ієрархічні відношення, якщо їх не оцінюють як сильну парадигматику.

Основним призначенням встановлення асоціативних відношень між дескрипторами інформаційно-пошукового тезауруса є те, що встановлення такого зв'язку може вказати додаткові дескриптори, корисні при індексуванні або пошуку.

Наприклад:

1. наукова дисципліна – об’єкт вивчення або фахівець у цій дисципліні: *математика – математик; неврологія – нервова система;*
2. операції або процеси та їх агент чи інструмент: *контроль температури – термостат; мисливець – полювання*;
3. об'єкти або процеси та їх контрагенти: *рослини – гербіциди*;
4. дії та їх продукти: *ткацтво – тканина; сльозовиділення – сльоза;*
5. дії та їх цілі: *палітурна справа – книга*;
6. об'єкти та речовини і їх властивості (унікальні властивості): *отрути – токсичність; рідина – поверхневий натяг*;
7. поняття, пов'язані причинно-наслідковим зв'язком: *смерть – оплакування;*
8. поняття і одиниці їх вимірювання: *електричний струм – ампер*;

Усі види зв’язку відображаються в словникових статтях дескрипторів і аскрипторів.

Позначкою *С* (синонім) вводяться в дескрипторну статтю упорядковані за алфавітом безліч умовних синонімів поданого заголовного дескриптора. Разом із заголовним дескриптором вони створюють клас умовної еквівалентності.

Дескриптор, що означає або родове поняття, або ціле по відношенню до даного дескриптора, який, у свою чергу, означає частину цього цілого, має назву *вищий (або ширший) дескриптор* і споряджається в заголовній статті позначкою

В (вищий). Так позначається ***родовий дескриптор*** – вищий дескриптор, що означає родове поняття.

Дескриптор, що означає або видове поняття, або частину по відношенню до поняття, яке представлене вищим дескриптором, має назву *нижчий (вузький) дескриптор* і вводиться в статтю з позначкою *Н* (нижчі), тобто ***видовий дескриптор***.

Дескриптор, пов'язаний із головним дескриптором іншими видами семантичних відношень, включений до дескрипторної статті з позначкою *А*, має назву *асоціативний*.

4. Індексування інформаційних ресурсів

Індексування – це процес відображення змісту документа і/або смислового значення інформаційного запиту інформаційно-пошуковою мовою (ІПМ).

Результатом індексування документа є **пошуковий образ документа (ПОД)**, що виражає основний зміст документа, тобто, який представляє собою набір індексів (лексичних одиниць) ІПМ, що відповідають основним предметам і аспектам змісту документа.

Результатом індексування запиту є **пошуковий образ запиту (ПОЗ)**, що виражає зміст інформаційного запиту, тобто, це сукупність понять про головні теми запиту, виражених термінами ІПМ.

Процедура індексування складається з наступних взаємопов'язаних процесів: обстеження змісту документа та встановлення його предмета; ідентифікація основних понять, що складають предмет документа; відбір відповідних термінів індексування для представлення виявлених понять засобами обраної мови індексування і формування ПОД (ПОЗ).

Зауважимо, що в залежності від ІПМ, що прийнята у відповідній ІПС як ПОД, можуть виступати предметні заголовки (рубрики), індекси будь-якої кваліфікації (УДК, МПК, ББК), ключові слова або дескриптори за аналогією, що використовується ІПМ, формується і ПОЗ.

Індексування запиту здійснюється шляхом перекладу його теми або предмета, що цікавить користувача, на прийняту в ІПС ІПМ. Наступними етапами є: формування та розширення ПОЗ (тобто, включення при необхідності в ПОЗ надлишкових дескрипторів) та формування пошукового припису (ПП).

ПП, окрім індексів запиту, включає вказівки (умови), необхідні для здійснення інформаційного пошуку.

У процесі індексування використовується термінологічна база, яка може бути представлена рубрикаторами предметних рубрик, схемами класифікацій, словниками дескрипторів (тезаурусами). В залежності від цього є такі *види індексування:*

- *предметизація* – предметне індексування засобами мови предметних рубрик. Предметне індексування – це індексування предметного змісту документів;
- *систематизація* (класифікаційне індексування) – присвоєння даним або документам класифікаційних індексів відповідно до правил будь-якої класифікаційної ІПМ;
- *координатне індексування* – індексування, що передбачає багатоаспектне викладення основного смислового змісту документа або смислового змісту інформаційного запиту більшістю ключових слів або дескрипторів.

Розглянемо ці види індексування. Аналізуючи види індексування, слід зауважити, що на практиці використовуються одноаспектне та багатоаспектне індексування [103].

Одноаспектне індексування – це індексування документа або запиту за однією ознакою, при цьому пошукові образи можуть створюватись із одної лексичної одиниці ІПМ, наприклад, одного класифікаційного індексу або предметного заголовка.

Багатоаспектне індексування відбувається паралельно за декількома різними ознаками та пов’язане з побудовою складних пошукових образів, що утворюються із декількох лексичних одиниць ІПМ. Розрізняють два режими

багатоаспектного індексування: попередньої координації (передкоординатне індексування) та послідуючої координації (посткоординатне індексування).

Передкоординатне індексування – це процес перекладу змісту документа або запиту з природної мови на ІПМ, при якому формування ПОД відбувається за раніше обумовленими, регламентуючими правилами. Пошуковий образ – це лінійний запис (кортеж), в якому кожна лексична одиниця займає визначене місце. Координація (упорядкування) лексичних одиниць відбувається за рахунок синтагматичних відношень, тобто граматики. Різновидами передкоординатного індексування є предметне та систематичне індексування або *предметизація* та *систематизація* (класифікація).

Пошуковий образ при предметному індексуванні задається правилами (методами) предметизації і представляє собою складну предметну рубрику, що складається із заголовка та ряду підзаголовків, об'єднаних за допомогою таких граматичних засобів, як тире та кома.

Пошуковий образ при систематичному індексуванні задається правилами систематизації та є складним індексом, окремі компоненти якого об'єднані між собою за допомогою знаків (+) та (:). Прикладами пошукових образів даного типу можуть бути класифікаційні індекси УДК, ББК та ін.

Посткоординатне (координатне) індексування – це процес перекладу змісту документа або запиту на ІПМ, при якому пошуковий образ задається списком лексичних одиниць. Послідовність лексичних одиниць може бути будь-якою, кожен індекс має пошукові функції, що забезпечують багатоаспектний пошук інформації. Прикладом пошукового образу в результаті такого індексування може бути пошуковий образ мовою ключових слів або дескрипторів [103].

При формуванні ПОД і ПОЗ часто здійснюється процедура надлишкового індексування, що полягає в доповненні ПОД або ПОЗ індексами, пов'язаними сильними парадигматичними відношеннями з індексами вихідного пошукового образу.

5. Етапи індексування

Процес індексування включає наступні етапи, які здійснюють в зазначеній нижче послідовності:

1. аналіз і визначення змісту документа, як об'єкта індексування;
2. вибір понять, що характеризують зміст документа;
3. вибір термінів індексування для позначення понять;
4. формування пошукового образу документа з термінів індексування.

Перераховані етапи можуть бути об'єднані в складі технологічних процедур при умові належного виконання кожного з етапів.

Пошуковий образ документа (ПОД) формують з вибраних термінів індексування за допомогою граматичних засобів інформаційно-пошукової мови (ІПМ).

1.Аналіз і визначення змісту документа

При аналізі документа індексатору повинна бути надана можливість ознайомитися з документом у повному обсязі. При неможливості вичерпного ознайомлення з документом індексатор повинен вивчити наявні текстові частини документа (основні джерела індексування):

- довідковий апарат документа – заголовок (найменування), анотацію, реферат,
- зміст (зміст), передмова, укладення та ін;
- вступ;
- заголовки частин і глав;
- перші фрази глав і параграфів;
- ілюстрації, схеми, таблиці та підписи до них;
- слова і групи слів, які в тексті підкреслені або виділені поліграфічними засобами.

Індексування тільки по заголовках є неповноцінним. При індексуванні по рефератах і анотаціях слід стежити за адекватністю передачі в них змісту документа.

Термін може характеризувати документ не тільки простою констатацією його присутності-відсутності, але також і важливістю, **інформаційною вагою**

терміна в документі. Ваги термінів можуть обчислюватися за кількістю їх появи в даному документі, а також приписуватися за входженням в заголовок та інші важливі елементи документа. Різні слова можуть також мати власні внутрішні особливості, що впливають на їх можливість бути дескриптором.

Аналіз документів значно полегшують списки **«заборонених» слів**, куди входять слова граматичного або службового характеру, а також слова із надто невизначеним значенням, які очевидно не можуть служити для опису тематики документа і які не входять в число дескрипторів тезауруса. Такий список ще називають **стоп-словник**. Ваги дескрипторів також можуть присвоюватися шляхом експертного аналізу їх смислової важливості для даного документа.

При аналізі документа можна також вказати зв'язок окремих дескрипторів в ньому. Від зв'язку слів дуже сильно залежить зміст документа. *Так, одна справа, коли в документі йдеться, наприклад, про виробництво пива в країнах Балтики, а інша - коли про виробництво пива «Балтика». Розмежування цих документів можна відобразити, вказавши, що в другому випадку слова пиво і Балтика вживаються (майже) завжди поруч.* Можна вказувати, що слова вживаються в документі в одному реченні, в одному абзаці, а також відстань між словами в числі проміжних слів, що в деякій мірі відображає їх смисловий зв'язок. Застосовуються й інші **показники зв'язку**.

При інтелектуальному аналізі документа дескрипторам можна приписувати їх смислову роль в документі, наприклад:

«головний предмет розгляду»,

«характеристика предмета»,

«мета дослідження»

«результат»,

«умови», тощо.

Такі позначки коло дескрипторів мають назву **вказівники ролі**.

2. Вибір понять, що характеризують зміст документа

Число характеристик і понять, відображених в ПОД, визначає його повноту і є найважливішим показником якості індексування.

У ПОД необхідно відобразити всі поняття, які можуть мати цінність для користувачів системи. Якщо у документі виявлено більше ніж одна тема зі сфери інтересів користувачів, то ці теми мають розглядатися окремо.

Тематика, відображена при індексуванні, не повинна обмежуватися вузькими рамками безпосередніх інтересів користувачів ІПС. В ПОД також повинні бути включені поняття, які пов'язані з побічними аспектами документа (наприклад, соціальні та економічні аспекти науково-технічних досліджень).

При виборі понять основним критерієм є потенційна цінність поняття для вираження змісту документа або для його пошуку. При цьому необхідно орієнтуватися на типові запити до ІПС:

- відбирати поняття, найбільш вживані в колективі користувачів ІПС;
- уточнювати склад лексики та граматичні правила ІПМ па основі зворотного зв'язку з користувачами.

Кількість термінів індексування, які приписуються одному документу, визначається кількістю відомостей, що містяться в документі. Обмеження кількості термінів ґрунтується на змістовному відборі найбільш важливих понять.

Повнота індексування, прийнята в кожній ІПС, визначається її функціональним призначенням. Обсяг документа також сильно впливає на повноту індексування.

Необхідно враховувати зазначені фактори і на їх основі виробляти експертний відбір понять з документа, не прагнучи включити до ПОД всі згадані в ньому поняття.

Специфічність ПОД визначається тим, на скільки точне відображення знайшли поняття документа у термінах індексування, і також є одним з параметрів якості індексування. Заміна поняття терміном, що відображає більш широке поняття, призводить до втрати специфічності. Використання ширших термінів допускається в особливих випадках:

- якщо надто специфічний термін незрозумілий користувачам, особливо коли відповідне поняття застосовується тільки в прикордонних областях діяльності;

- якщо в документі поняття розкрито недостатньо повно або є допоміжним для викладу змісту документа.

3. Вибір термінів індексування для позначення понять

У процесі вибору термінів індексування поняття, що характеризують зміст документа, представляють:

- лексичними одиницями, яким надається перевага, (дескрипторами або ключовими словами), обраними за правилами конкретної ІПМ;
- термінами, що відображають нові поняття, перевіривши їх точність і прийнятність за словниками, енциклопедіями, довідниками, інформаційно-пошуковими тезаурусами та іншим джерелам, визнаним авторитетними в даній області.

Вибір термінів індексування здійснюють на основі інформаційно-пошукового тезауруса, який використовують при складанні запитів до ІПС.

При використанні тезауруса допускається скорочувати число термінів, що включаються до ПОД за рахунок виключення загальних понять, які можуть бути залучені на етапі пошуку документа або на етапі складання пошукового розпорядження на підставі посилань у статтях тезауруса.

Поняття, не представлені в словнику індексування, але необхідні для формування ПОД, вказують одним із двох способів:

- новим специфічним терміном, який включають до ПОД і в словник;
- більш загальним терміном, наявним в ІПМ; при цьому специфічний термін направляють до служби ведення ІПМ як кандидата на включення до словника.

Нові поняття представляють найбільш близькими з існуючих в ІПМ лексичних одиниць, а також оцінюють корисність включення нових термінів у словник з точки зору пошуку.

При індексуванні вільними ключовими словами, взятими з тексту документа, вони повинні бути приведені до канонічної формі. Довжину словосполучень переважно обмежується двома-трьома словоформами.

4. Формування пошукового образу документа.

ПОД складається з вибраних термінів індексування, організованих за допомогою граматичних засобів ІПМ даної ІПС.

До складу ПОД можуть бути включені такі категорії даних, передбачені технологією індексування конкретної ІПС:

- ступінь нормалізації термінів індексування і вживаний для цього словник;
- індивідуальні характеристики терміна індексування;
- зв'язок термінів індексування в синтаксичних конструкціях ПОД.

За ступенем нормалізації розрізняють два типи термінів координатного індексування: дескриптори і ключові слова.

Дескриптори можуть представлятися умовними кодами, які зазначені у використовуваному словнику індексування. У цьому випадку ІПС повинна забезпечувати автоматичний пошук орфографічних форм дескрипторів за їх кодами.

Ключові слова з ПОД в багатомовних інформаційних системах повинні бути забезпечені позначками про приналежність до того чи іншого природної мови.

Індивідуальні характеристики термінів індексування є факультативними елементами ПІД і їх використовують для уточнення змісту документа, організації процедур інформаційного пошуку або подальшої аналітико-синтетичної обробки документів в системі.

До **індивідуальних характеристиках** відносять дані про семантичну і морфологічну категорії терміна індексування, його ролі та інформаційну вагу тощо.

Семантична характеристика терміна індексування полягає у віднесенні його до наступних лексикографічних категорій:

1) термін, що виражає науково-технічне поняття;

2) власне ім'я, ідентифікатор;

3) найменування параметра;

4) значення параметра (виражене текстом або іменною величиною);

5) числове вираження;

6) позначення одиниці величини.

Морфологічна характеристика терміна індексування полягає у віднесенні його до таких лексикографічних категорій:

1) похідне слово;

2) складне слово;

3) словосполучення;

4) абревіатура;

5) фрагмент слова.

Морфологічні характеристики використовують в ПОД для реалізації в ІПС смислового аналізу лексичних одиниць на основі їх формальних ознак.

Роль терміна індексування вказують у ПОД для уточнення місця відповідного поняття у змісті документа. Для цього особливими покажчиками ролі, прийнятими в ІПС, відзначають терміни індексування, що відображають такі аспекти документа:

1) об'єкт дослідження, опису;

2) характеристики, властивості, параметри об'єкта;

3) методи та засоби дослідження, технологічну оснастку;

4) складові частини, вузли, деталі об'єкта;

5) сфера застосування об'єкта (галузь господарства, техніки, науки);

6) призначення об'єкта;

7) мета дослідження, розробки, описи;

8) результати дослідження, розробки.

Інформаційна вага терміна індексування у ПОД відображає важливість даного поняття для даного документа. Число градацій інформаційної ваги визначається потребами конкретної ІПС. Слід розрізняти:

1) поняття, що виражають головну тему документа;

2) поняття, що виражають побічні теми документа;

3) поняття, використані в документі як допоміжні для викладу його змісту.

Часом використовується також покажчик від’ємної ваги, яким позначають терміни індексування для вказівки на те, що дане поняття не розглядається в документі.

Терміни індексування в ПІД можуть бути забезпечені **покажчиками зв'язку**, що об'єднують їх у синтаксичні конструкції, які відображають:

1) порядок проходження і взаємне розташування термінів індексування в документі;

2) смислові зв'язки понять у документі;

3) парадигматичні зв'язки дескрипторів в тезаурусі.

Синтаксичні конструкції розглядають як цілісні одиниці ПОД поряд з термінами індексування. Вони можуть бути об'єднані з іншими синтаксичними конструкціями або з окремими термінами індексування в конструкції більш високого порядку.