**3. Побудова тезаурусу предметної області видавництва та поліграфії**

**Мета складової**: отримати практичні навички з аналізу інформаційних документів і побудови тезарусу.

**Загальні відомості**

Інформація, що циркулює в системі видавничо-поліграфічного виробництва, має бути певним чином структурована та систематизована з метою ефективного пошуку даних, формування масивів релевантних відомостей, описання змістовних особливостей інформаційних об’єктів відповідної предметної області. На вхід системи інформаційного забезпечення видавництва найчастіше поступають «сирі» дані, які характеризуються хаотичністю, відсутністю адекватного розподілу для вирішення локальних завдань видавничої справи. Систематизація інформаційних потоків стає можливою завдяки методиці семантичного аналізу документів на основі онтології предметної області.

Практика семантичного аналізу документів передбачає формування відповідного набору термінів предметної області,пов’язаних один з одним змістовними зв’язками; створення адекватного контенту інформаційних об’єктів і їхню сегментацію; здійснення автоматичного реферирування документів. Вказані аспекти дозволяють структурувати та оптимізовувати потоки інформації в системі інформаційного забезпечення видавничої діяльності.

Для опису якої-небудь предметної області завжди використовується певний набір термінів, кожен з яких позначає або описує яке-небудь поняття або концепцію з даної предметної області.

Сукупність термінів, що описують дану предметну область, з вказівкою семантичних стосунків (зв'язків) між ними називаються ***тезаурусом***.

Такі стосунки в тезаурусі завжди указують на наявність смислового (семантичного) зв'язку між термінами.

Основним відношенням (зв'язком) між термінами в тезаурусі є зв'язок між *ширшими* (виразнішими) і *вужчими* (більш спеціалізованими) поняттями. Виділяють два підвиди цього відношення:

* Один термін позначає поняття, що є частиною поняття, що позначається іншим терміном (наприклад, «видавництво» і «друкарня»)
* Один термін позначає елемент класу, що позначається іншим терміном («спеціальні види друку» і «райдужний друк»).

Це відношення на множині термінів є відношенням часткового порядку, тобто множини термінів з такими зв'язками утворює ациклічний граф, або поліієрархічну структуру.

Існують також і інші зв'язки між термінами. Наприклад, одне поняття або концепція може бути позначене декількома термінами, синонімами. Деякі терміни можуть бути антонімами для інших. Часто серед термінів, що відносяться до одного поняття, виділяють єдиний (для кожної мови тезауруса) *найбільш переважний* (*найбільш відповідний*) *термін*, який найдобріше характеризує, або позначає дане поняття. Решта термінів є *менш переважними* (*менш відповідними*).

Окрім вищеописаних, між термінами можуть існувати також та інші, *асоціативні зв'язки*, якщо поняття, що позначаються цими термінами, як-небудь зв'язані між собою по своєму сенсу, за винятком описаних вище ієрархічних зв'язків.

У багатомовних тезаурусах існують також *зв'язки еквівалентності* між термінами на різних мовах. Виділяють повну (строгу) еквівалентність, і декілька видів часткової (нестрогою) смислової еквівалентності термінів на різних мовах.

Тезаурус часто містить *коментарі* до термінів, що розкривають для користувача сенс терміну, а також що пояснюють, як слід його використовувати.

В процесі формування контенту видання результатом семантичного аналізу інформаційних документів найчастіше виступає автоматичне реферирування.

Автоматичним реферуванням є процес складання анотацій, або коротких викладів матеріалу, іншими словами, витягання найбільш важливих або характерних фрагментів з одного або багатьох джерел інформації [5].

Процес реферирування може бути здійснений на основі використання певних методів [5].

***Метод складання витримок***, припускає акцент на виділення характерних фрагментів (як правило, пропозицій). Для цього методом зіставлення фразових шаблонів, виділяються блоки найбільшою лексичною і статистичною релевантності. Створення підсумкового документа в даному випадку — просто з'єднання вибраних фрагментів.

У більшості методів застосовується модель лінійних вагових коефіцієнтів. Основу аналітичного етапу в цій моделі складає процедура призначення вагових коефіцієнтів для кожного блоку тексту відповідно до таких характеристик, як розташування цього блоку в оригіналі, частота появи в тексті, частота використання в ключових пропозиціях, а також показники статистичної значущості. Сума індивідуальних ваг, що, як правило, визначена після додаткової модифікації відповідно до спеціальних параметрів налаштування, пов'язаних з кожною вагою, дає загальну вагу всього блоку тексту.

***Метод формування короткого викладу*** вимагає використання могутніх обчислювальних ресурсів для систем обробки природних мов (NLP — natural language processing), зокрема граматики і словники для синтаксичного розбору і генерації природно-мовних конструкцій. Крім того, для реалізації цього методу потрібні якісь онтологічні довідники, що відображають міркування здорового глузду і поняття, орієнтовані на предметну область, для ухвалення рішень під час аналізу і визначення найбільш важливої інформації.

**Завдання**

1. Сформуйте тезаурус предметної області вашого магістерського дипломного проекту. Для цього виділіть систему термінів, які описують процеси мультимедійного видавництва; опишіть семантичні зв’язки між термінами з вказанням більш ширшого та вужчого поняття, найбільш переважних і менш переважних термінів. Проаналізуйте зв’язки еквівалентності та асоціативні зв’язки між поняттями тезаурусу.

2. Наведіть види гіперпосилань, які використовуються в практичному результаті вашого магістерського дипломного проекту. Охарактеризуйте ці види гіперпосилань та порівняйте їх з посиланнями аналогічних мультимедійних продуктів.

3. Сформуйте формальну структуру тексту розділу 1 вашого магістерського дипломного проект. В якості маркеру використайте лексичні об’єкти, задані конкретним ідентифікатором або назвою.

4. Складіть інформативний реферат обраної статті на тематику вашого магістерського дипломного проекту методом складання витримок.

**Контрольні запитання**

1. Що являє собою тезаурус та яке його функціональне призначення?

2. Як виконується автоматичне реферирування?

3. Яким чином відбувається аналіз текстового документу

4. Проаналізуйте сфери використання автоматичного реферування у видавничій діяльності.