ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ ТЕРМІНІВ У НАВЧАЛЬНИХ МАТЕРІАЛАХ

Крак Ю.В[1]., Бармак О.В.[2], Мазурець О.В.[2]
[1]*Інститут кібернетики НАН України, м.Київ*
[2]*Хмельницький національний університет, м.Хмельницький*

Сучасна дистанційна освіта характеризується повним переходом на інформаційні технології, що визначає необхідність суттєвої формалізації та стандартизації навчального процесу. Так, загальноприйнятим є підхід застосування навчальних матеріалів у вигляді цифрових документів визначеної структури як інструменту навчання, й тестів як інструмента контролю рівня отриманих знань.

Для розробки й використання курсів навчальних дисциплін за наведеним принципом використовуються спеціалізовані віртуальні навчаючі середовища, напрклад Moodle [1]. При використанні таких середовищ, потенційна якість отриманих освітніх послуг безпосередньо визначається відповідністю навчальних матеріалів курсу вимогам стандартів освіти (робочим планам, структурі навчального плану тощо), й тестів – навчальним матеріалам. Слід зауважити, що необхідність автоматизації процесу створення такого контенту та оцінки його якості поширюється на всі форми освіти, а не тільки дистанційну [2].

Структурна відповідність навчальних матеріалів вимогам стандартів може бути оцінена шляхом аналізу структури відповідних цифрових документів. Задача ж оцінки семантичної відповідності залишається актуальною.

Зі змістовної точки зору, ключовою властивістю контенту є його семантика, яку формалізовано відображають у вигляді семантичної мережі, вузлами якої є терміни, що несуть семантичне навантаження, а дуги відображають характер зв’язку між вузлами. Зв'язок між термінами навчальних матеріалів залежить від багатьох факторів (галузь знань, тип лекції, літературні здібності автора тощо) й може змінюватися у широких межах без втрати якості викладання, що знижує можливість формалізації аналізу. Тому саме аналіз термінів, що використовуються у навчальних матеріалах, дозволяє визначити якість цих навчальних матеріалів та їх відповідність вимогам.

Оскільки тести є засобом перевірки якості засвоєння сенсу навчальних матеріалів й ставлять на меті задачу перевірки якості засвоєння термінів як складових семантичних одиниць навчальних матеріалів, то виділення семантичних термінів з навчальних матеріалів може забезпечити допомогу та контроль при розробці наборів тестових завдань. Тому задача автоматизації визначення семантичних термінів у навчальних матеріалах є актуальною задачею інформаційних технологій у сучасній освіті.

Метою роботи є розробка інформаційної технології автоматизованого визначення семантичних термінів у контенті навчальних матеріалів.

Для автоматизації пошуку ключових слів доцільно використовувати різноманітні методи аналізу текстів, таких як частотна оцінка, оцінка TFIDF й дисперсна оцінка [3]. Ці методи дозволяють зпівставити окремим словам або словосполученням тексту деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті.

Попередніми дослідженнями було визначено найбільш ефективним методом аналізу текстів метод дисперсної оцінки [4]. Результат застосування частотного аналізу свідчить, що цей метод надає велику вагу не тільки ключовим словам, а й словам із максимальною частотою – сполучникам, прийменникам і часткам, що відіграють велику роль для зв’язності тексту, проте не несуть навантаження з точки зору семантичної структури. Метод TFIDF дозволяє дещо відсіяти слова, що використовуються для зв’язування тексту, через їх велике значення розповсюдженості у контенті, але значна вага надається словам, важливість яких є обмеженою в рамках локальних елементів контенту, зокрема, наприклад, важливі слова із практичних прикладів та важливі оператори й змінні у лістингах програмного коду; тому даний метод використовується переважно для аналізу масивів незв’язних текстів, й продемонстрував низьку ефективність при аналізі контенту навчальних матеріалів. Результат аналізу контенту лекції методом дисперсного оцінювання дозволив визначити перелік слів, найбільш близький до переліку, сформованого експертом, що й обумовило його подальше використання у задачі визначення термінів у навчальних матеріалах.

У навчальних матеріалах присутні терміни трьох видів (слова, словосполучення та абревіатури), кожен з яких вимагає окремого підходу до пошуку. Так слова потребують для пошуку спеціальний алгоритм на базі дисперсної оцінки. Абревіатури є стійкими звязними сукупностями літер, тому можуть розпізнаватись як слова. Словосполучення технічно є сталими сукупностями ключових слів (колокаціями) й потребують для пошуку спеціального алгоритму поза дисперсним оцінюванням.

Незважаючи на високу ефективність методу дисперсної оцінки, ряд факторів унеможливлюють його монопольне застосування для вирішення розглядуваної задачі. Це, зокрема, необхідність

попередньої перевірки тексту на відповідність нормам ведення наукової літератури; потреба в виявленні як ключових слів, так і ключових словосполучень; удосконалення алгоритмів пошуку кючових слів і словосполучень з використанням методу дисперсної оцінки, та інші. Тому виникла потреба в розробці нової інформаційної технології, яка із використанням методу дисперсної оцінки дозволяє ефективно й автоматизовано визначати семантичні терміни в навчальних матеріалах.

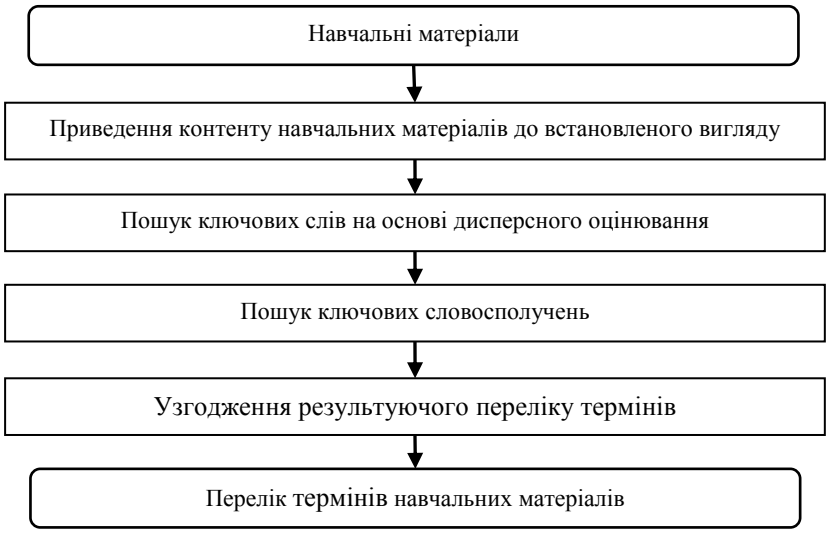

Рисунок 1 – Загальна схема інформаційної технології автоматизованого визначення термінів у навчальних матеріалах

Ефективному застосуванню методу дисперсної оцінки значно шкодить недотримання норм формування та ведення наукової літератури. Оскільки низька якість контенту навчальних матеріалів знижує ефективність його сприйняття також і рецепієнтом, то вимога приведення контенту навчальних матеріалів до встановленого вигляду розглядається як загальна. Основною проблемою, що вирішується в процесі приведення контенту навчальних матеріалів до коректного вигляду, є усунення неоднозначної іменованості термінів. Хоча є допустимим використання абревіатур, часткових скорочень та повних назв (наприклад, «СКБД», «Система керування БД», «Система керування базами даних»), в тому числі кількома мовами (наприклад, українською та англійською), для використання в рамках окремого матеріалу обирається лише один варіант, оскільки як для машинної, так і для розумової ідентифікації термінів бажаною (а в науковій літературі – обовязковою) є уніфікація ідентифікатора. Навіть у випадку введення нового скорочення, на позиції введення скорочення присутня також і повна назва. Відповідно, усунення неоднозначної іменованості термінів дозволить уникнути випадків, коли при автоматичному аналізі контенту навчальних матеріалів одне поняття буде розглядатись як кілька окремих термінів.

Приведений до встановленого вигляду контент навчальних матеріалів використовується для його аналізу методом дисперсної оцінки. Отриманий перелік лематизується із виключенням повторів й фільтрується за частиною мови, після чого обмежується за обсягом відповідно до вимог.

Для виявленні ключових словосполучень у контенті навчальних матеріалів проводиться пошук неперервних скупчень ключових слів протягом тексту без врахування службових частин мови. Після чого проводиться визначення порядку й форм слів у словосполученнях за частотним аналізом. Після чого проводиться рейтингове обмеження обсягу й формується результуючий перелік словосполучень.

В результаті послідовного виконання алгоритмів пошуку ключових слів та словосполучень утворюються два відповідних переліки термінів. На їх основі формується узагальнений перелік термінів, до якого входять всі словосполучення й ті слова, значення дисперсної оцінки яких перевищує значення дисперсної оцінки звязаних із цим словом колокацій. Узагальнений перелік термінів ранжується за значеннями їх дисперсної оцінки.

Для дослідження ефективності запропонованої технології було розроблено відповідне тестове програмне забезпечення. Згенеровані програмою переліки термінів порівнювались із переліками, сформованими експертами (авторами) для відповідних навчальних матеріалів. Аналіз результатів по тестовій виборці (30 лекцій із різних навчальних курсів) показав середню ефективність 94,3% (мінімальний збіг переліків 85,8%, максимальний – 100%).

Висока ефективність запропонованої технології надає підставу до її ефективного застосування у вирішенні актуальних задач, таких як оцінка відповідності навчальних матеріалів змістовим вимогам, оцінка відповідності наборів тестових завдань навчальним матеріалам, семантична допомога при створенні тестів, автоматизована генерація переліків ключових слів та анотацій.

Література

1. Moodle – Open-source learning platform. – [Електронний ресурс]. – 2015. – https://moodle.org/
2. Снитюк В.Е., Юрченко К.Н. Интеллектуальное управление оцениванием знаний. − Черкассы, 2013. − 262с.
3. Ventura, J. & Silva, J. (2007). New Techniques for Relevant Word Ranking and Extraction. In Proceedings of 13th Portuguese Conference on Artificial Intelligence, Springer-Verlag, pp. 691-702.
4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272с.