УДК 004.912

О.В. БАРМАК, О.В. МАЗУРЕЦЬ
Хмельницький національний університет

МЕТОДИ АВТОМАТИЗАЦІЇ ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ У НАВЧАЛЬНИХ МАТЕРІАЛАХ

В статті наведено аналіз методів для автоматичного визначення семантичних термінів у навчальних матеріалах. Для аналізу взята вибірка з 30 навчальних матеріалів. Для виявлення семантичних термінів застосовані методи частотної оцінки, оцінки TFIDF та дисперсної оцінки. Отриманні результати порівнювались зі списками термінів, що запропонували автори навчальних матеріалів. В результаті проведеного аналізу встановлено, що найбільша ефективність у розв'язанні задачі виявлення у навчальних матеріалах семантичних термінів досягається в ході використання метода дисперсної оцінки. Виявлено фактори, що перешкоджають ефективному аналізу навчальних матеріалів.

Ключові слова: навчальні матеріали, аналіз, ключові терміни, дисперсна оцінка.

O. BARMAK, O. MAZURETS
Khmelnytsky National University

METHODS OF AUTOMATION OF DEFINITION OF SEMANTIC TERMS IN EDUCATIONAL MATERIALS

The article presents the analysis of methods of automatic detection of the semantic terms in educational materials. 30 educational materials are taken for the analysis as the samples. The methods of term frequency, TFIDF and disperse evaluation are used to identify semantic terms. The obtained results were compared with lists of terms offered by the authors of the educational materials. The analysis shows, that the maximum efficiency in the task of identifying of semantic terms in educational materials reached by using the method of disperse evaluation. The factors that hinder effective analysis of educational materials have been found.

Keywords: educational materials, analysis, key terms, disperse evaluation.

Постановка проблеми в загальному вигляді

Поширення інформаційних технологій та розвиток глобальної мережі й телекомунікацій привели до значних змін у вищій освіті [1]. Одним із проявів цих змін стало виникнення нової форми освіти – дистанційної. На сучасному етапі велика кількість навчальних закладів пропонують послуги цієї форми навчання – як в світі, так і в Україні [2].

Характерною рисою дистанційної освіти є повний перехід на інформаційні технології, що визначає необхідність суттєвої формалізації та стандартизації навчального процесу [3]. Зокрема, є загальноприйнятим підхід [4] застосування як інструменту навчання матеріалів у вигляді цифрових документів визначеної структури, і тестів як інструмента контролю рівня отриманих знань [5]. Для розробки та використання курсів навчальних дисциплін за наведеним принципом використовуються спеціалізовані середовища, найбільш розповсюдженим із яких є Moodle [6].

За таких умов, потенційна якість отриманих освітніх послуг безпосередньо визначається відповідністю навчальних матеріалів курсу вимогам стандартів освіти (робочим планам, структурі навчального плану тощо), і тестів – навчальним матеріалам [7]. Необхідність автоматизації процесу створення такого контенту та оцінки його якості поширюється на всі форми освіти (а не тільки дистанційну). Якщо структурна відповідність може бути оцінена шляхом порівняння зі стандартами (базуючись на структурі цифрового документу, то задача оцінки семантичної відповідності залишається актуальною [8].

Аналіз останніх досліджень

Одним із способів вирішення задачі оцінки семантичної відповідності є аналіз термінологічної бази навчальних матеріалів. З навчальної точки зору, ключовою властивістю контенту є його семантика, яку формалізовано відображають у вигляді семантичної мережі [9], вузлами якої є терміни, що несуть семантичне навантаження, а дуги відображають характер зв’язку між вузлами [10].

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 актуальність його аналізу. Тому переважно аналіз саме термінів, що використовуються у навчальних матеріалах, дозволяє визначити якість цих навчальних матеріалів та їх відповідність вимогам. Крім цього, визначені семантичні терміни у навчальних матеріалах, спроможні допомогти при створенні тестів, адже тести як засіб перевірки якості засвоєння сенсу навчальних матеріалів відповідним чином ставлять на меті задачу перевірки якості засвоєння термінів як складових семантичних одиниць навчальних матеріалів. Отже, задача автоматизації визначення семантичних термінів у навчальних матеріалах є актуальною задачею сучасної освіти.

Термінами можуть бути як ключові слова, так і ключові словосполучення. Ключові словосполучення можуть містити довільну кількість слів, і бути семантичними мережами малої ємності. В рамках аналізу їх склад доцільно спрощувати до двох слів. При цьому в процесі пошуку як мінімум одне із цих слів можна розглядати як термін в межах навчальних матеріалів. Тож питання автоматизації пошуку ключових слів у контенті навчальних матеріалів є першочерговою задачею в процесі вирішення

розглядуваної проблеми.

Постановка задачі. Дослідження сучасних відомих методів аналізу текстів для оцінки їх ефективності й придатності до використання у задачі автоматизації пошуку ключових семантичних термінів у контенті навчальних матеріалів.

Викладення основних матеріалів дослідження

Застосування різноманітних методів аналізу текстів дозволяє зіставити окремим словам або словосполученням тексту деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті. Ці методи розрізняються за алгоритмами обрахунку вказаних вагових значень [11]. Найбільш розповсюдженими методами аналізу текстів є частотна оцінка, оцінка TFIDF та дисперсна оцінка.

Частотна оцінка Tf (term frequency) є частотою згадувань певного слова і у тексті, що розглядається, й обчислюється наступним чином [12]:

$$Tf_i = \frac{n(i)}{\sum_k n_{ik}}, \tag{1}$$

де $n(i)$ – кількість згадувань слова i у тексті, $\sum_k n_{ik}$ – загальна кількість слів у тексті.

Оцінка TFIDF є добутком частоти згадувань слова у тексті *Tf* (term frequency) та зворотної документарної частоти слова *Idf* (inverse document frequency) [13]:

$$TfIdf = Tf * Idf, \tag{2}$$

де

$$Tf_i = \frac{n(i)}{\sum_k n_{ik}},$$

$$Idf_i = \log\frac{D}{d_i},$$

де D – кількість фрагментів, на які розбивається текст при аналізі; d_i – кількість фрагментів, у яких дане слово присутнє.

Дисперсна оцінка за змістом близька до оцінки TFIDF, та є оцінкою дискримінантної сили слів. Вона дозволяє відділити із загального переліку широковживаних у тексті слів слова, що розташовані рівномірно. Якщо деяке слово A в тексті, що складається з N слів, позначене як A_k^n, де індекс k – номер появи даного слова в тесті, а n – позиція даного слова в тексті, то інтервал між послідовними появами слова при таких позначеннях буде величина $\Delta A_k^m = A_{k+1}^m - A_k^n = m - n$, де на m-ій і n-ій позиціях в тесті знаходиться слово A, яке зустрілось k+1-й і k-й рази. Тоді дисперсійна оцінка розраховується наступним чином [14]:

$$\sigma = \frac{\sqrt{(\Delta A^2) - (\Delta A)^2}}{(\Delta A)} \tag{3}$$

де (ΔA) – середнє значення послідовності $\Delta A_1, \Delta A_2, \Delta A_k$; (ΔA^2) – послідовності A_1^2, A_2^2, A_k^2; K – кількість появи слова A в тексті.

Запропоновано проаналізувати вибірку навчальних матеріалів шляхом використання запропонованих методів для визначення ключових семантичних термінів за схемою, наведеною на рис. 1.

Для проведення експериментів за наведеною вище схемою було розроблено тестове програмне забезпечення, що реалізує обробку контенту навчальних матеріалів трьома розглянутими методами (частотний аналіз, аналіз TFIDF та дисперсний аналіз) з відповідними ваговими параметрами.

В процесі обробки контенту переліки ключових слів, отримані за відповідними методами, обмежуються за кількісним порогом й формують множини B_1, B_2, B_3. В подальшому ці множини порівнюються із множиною B_A, утвореною переліком ключових термінів, який сформовано автором навчального матеріалу. Перетин цих множин $B_k \bigcap B_A$ визначає ефективність відповідного методу k.

Максимальна область перетину авторського переліку зі сформованими за стосунком переліками $B_k \bigcap B_A \to \max$ визначає найбільш ефективний метод автоматизації пошуку ключових семантичних термінів у контенті навчальних матеріалів.

Ефективність наведених методів пропонується визначати за наступною формулою:

$$E_k = \frac{N_{Ak}}{N_A} \cdot 100\%, \tag{4}$$

де N_{Ak} – кількість термінів у авторському (B_A) та сформованому за k-м методом (B_k) переліками термінів, що співпали ($B_k \bigcap B_A$); N_A – кількість термінів у переліку термінів B_k, сформованому експертом (автором).

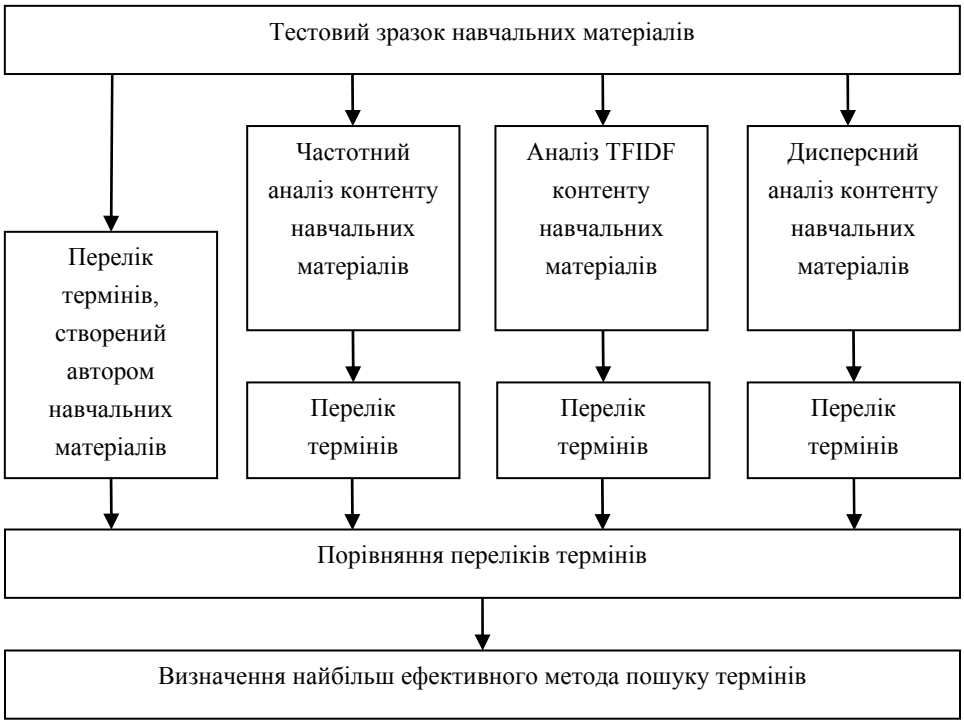

Рис. 1. Схема оцінки ефективності методів обробки текстів

Результати дослідження

В результаті тестування (на прикладі лекційного матеріалу «Введення у реляційну модель даних» навчального курсу «Вступ до реляційних баз даних» [15]) розробленим програмним забезпеченням отримуються три переліки ключових термінів за відповідними методами аналізу та проводиться їх порівняння у сукупності з авторським переліком. Деякі результати порівняння наведено у табл. 1.

На основі наведених даних дослідження, за формулою (4) побудовано діаграму ефективності розглянутих методів формування переліку ключових термінів у порівнянні з авторським переліком (рис. 2). Ефективність методу частотної оцінки склала 33,3%, методу оцінки TFIDF – 30,3%, методу дисперсної оцінки – 84,8%.

Таблиця 1

Фрагмент порівняльної таблиці аналізу термінів

№ п/п	Термін	Визначено автором	Частотний аналіз	Аналіз TFIDF	Аналіз дисперсії
1	реляційна база даних	+	+		+
2	тип даних	+	+	+	+
3	домен	+		+	+
4	реляційна модель даних	+			+
5	обмеження цілісності	+		+	+
6	заголовок відношення	+			+
7	значення відношення	+	+		+
8	перша нормальна форма	+			+
9	модель даних	+	+		+
10	СКБД	+	+		+
11	реляційне числення	+			
12	цілісність сутності	+		+	+
13	булевий тип	+			+
14	зовнішній ключ	+		+	+
15	SQL	+			
16	заголовок відношення				+
17	унікальність значень			+	

Аналогічним чином було досліджено 30 лекцій із різних навчальних курсів й обраховано середню ефективність кожного із методів. Середня ефективність методу частотної оцінки склала 27,1%, методу оцінки TFIDF – 45,5% та методу дисперсної оцінки – 88,3% (рис. 3).

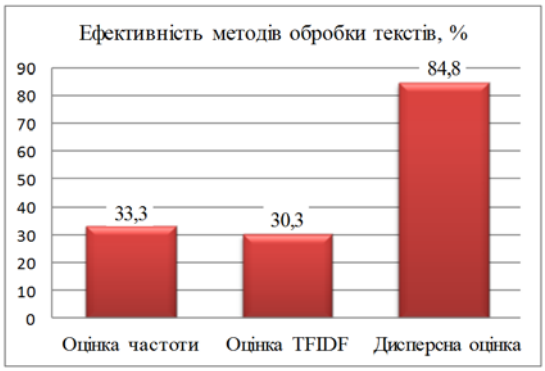

Рис. 2. Діаграма ефективності методів обробки текстів

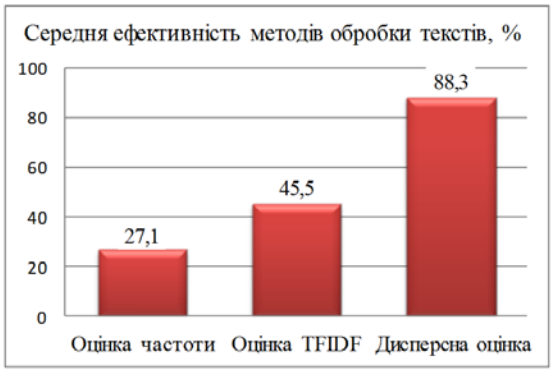

Рис. 3. Діаграма середньої ефективності методів обробки текстів

Таким чином, метод дисперсної оцінки продемонстрував найвищу ефективність серед досліджуваних методів, показавши при цьому мінімальну ефективність 67,7%, максимальну – 100%.

Дискусія

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.

Отриманий результат аналізу контенту лекції методом дисперсного оцінювання дозволив визначити перелік слів, найбільш близький до переліку, сформованого експертом (автором курсу).

Слід зауважити, що одним з факторів, які зменшують ефективність обробки, є недостатня якість вхідних даних. У процесі аналізу було встановлено групи факторів, що перешкоджають ефективному аналізу навчальних матеріалів:

- некоректна побудова переліку ключових термінів авторами курсів;
- некоректне формування контенту навчальних матеріалів.

До некоректної побудови списків термінів авторами призводять:

- використання одночасно декількох мов (українська та англійська);
- включення в перелік термінів одночасно абревіатур і повних назв при використанні в тексті лише одного з варіантів;
- включення в перелік термінів тегів, що визначають рубрикацію й тематичну приналежність текстів, але не розкриваються в рамках їх контенту;
- включення в перелік термінів таких, що використовуються в тексті одноразово.

Причинами некоректного формування контенту навчальних матеріалів авторами є:

- некоректне використання абревіатур (наприклад, «СКБД», «Система керування БД», «Система керування базами даних);
- використання протягом викладення матеріалу термінів одночасно кількома мовами;
- непослідовність та диспропорції при викладенні матеріалу.

З метою підвищення ефективності застосування методу дисперсної оцінки контенту навчальних матеріалів визначено подальші напрямки досліджень:

- для підвищення ефективності формування переліків ключових термінів навчальних матеріалів із використанням методу дисперсної оцінки є доречним використання додаткових фільтрів для відсіювання категорій слів та попередньої обробки контенту навчальних матеріалів.
- з метою наближення вигляду результатів пошуку ключових термінів до авторських зразків, необхідно визначити ефективні методи об'єднання знайдених слів у словосполучення та колокації.

Висновки

В результаті дослідження методів аналізу текстів було встановлено, що найбільшу ефективність у вирішенні задачі автоматизації пошуку ключових слів у контенті навчальних матеріалів досягнуто методом дисперсної оцінки. Виявлені фактори, що перешкоджають ефективному аналізу навчальних матеріалів.

Подальші дослідження направлені на вдосконалення методики використання дисперсної оцінки при обробці контенту навчальних матеріалів та пошук ефективних методів об'єднання знайдених слів.

Література

1. Нові інформаційні технології в освіті [Електронний ресурс]. – 2015. – Режим доступу : http://it-

tehnolog.com/statti/novi-informatsiyni-tehnologiyi-navchannya/.

2. University of the People [Електронний ресурс]. – 2015. – Режим доступу : http://www.uopeople.org/.

3. Концепція якості освіти [Електронний ресурс]. – 2015. – Режим доступу : http://osvita.ua/.

4. Факультет заочно-дистанційного навчання, післядипломної освіти та довузівської підготовки ХНУ [Електронний ресурс]. – 2015. – Режим доступу : http://dn.tup.km.ua/.

5. Аванесов В.С. Композиция тестовых заданий / Аванесов В.С. – М. : Центр тестирования, 2002.

6. Moodle – Open-source learning platform [Електронний ресурс]. – 2015. – Режим доступу : https://moodle.org/

7. Снитюк В.Е. Интеллектуальное управление оцениванием знаний / В.Е. Снитюк, К.Н. Юрченко. − Черкассы, 2013. − 262 с.

8. Кірей К.О. До проблеми стандартизації термінології освітніх інформаційно-телекомунікаційних технологій / К.О. Кірей, Л.О. Кірей // Вісник Черкаського університету / Черкас. нац. ун-т ім. Богдана Хмельницького. Сер.: Педагогічні науки. – Черкаси, 2009. – Вип. 146. – С. 27–29.

9. IDEF5 – Ontology Description Capture Method [Електронний ресурс]. – 2015. – Режим доступу : http://www.idef.com/IDEF5.htm.

10. Титенко С.В. Побудова дидактичної онтології на основі аналізу елементів понятійно-тезисної моделі С.В. Титенко // Наукові вісті НТУУ "КПІ". – 2010. – № 1. – С. 82–87.

11. Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учебное пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В. и др. – М. : МИЭМ, 2011. – 272 с.

12. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA. Europhys. Lett, 2002. 57(5). P. 759-764.

13. 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.

14. Ландэ Д.В. Компактифицированный горизонтальный граф видимости для сети слов / Д.В. Ландэ, А.А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения». – Киев : КПИ, 2013. – С. 158–164.

15. Введення у реляційну модель даних – НОУ «ИНТУИТ» [Електронний ресурс]. – 2015. – Режим доступу : http://www.intuit.ru/studies/professional_skill_improvements/1426/courses/74/lecture/2218.

References

1. New Information Technologies in Education. URL: http://it-tehnolog.com/statti/novi-informatsiyni-tehnologiyi-navchannya/.

2. University of the People. URL: http://www.uopeople.org/.

3. Koncepcia Yacosti Osvityu. URL: http://osvita.ua/.

4. Facultet Zaochno-Distanciynogo Navchannya, Pislyadiplomnoi Osvity ta Dovuzivskoi Pidgotovki KhNU. URL: http://dn.tup.km.ua/.

5. Avanesov V.S. Kompozicia Testovih Zadaniy. M., Centr Testirovanya, 2002.

6. Moodle – Open-source learning platform. URL: https://moodle.org/.

7. Snituk V.E., Yurchenko K.N. Intelektualnoe Upravlenie Ocenivaniem Znaiy. Cherkassy, 2013. − 262 s.

8. Do Problemi Standartizacii Terminologii Osvitnih Informaciyno-Telekomunikaciynih Tehnologiy / K. O. Kirey ., L. O. Kirey // Visnik Cherkaskogo Universitetu / Cherkaskiy Nacionalnogo Universitet im. Bogdana Khmeinitskogo. – Cherkasy, 2009. Ser.: Pedagogichni Nauki, Vip. 146. – C. 27-29.

9. IDEF5 – Ontology Description Capture Method. URL: http://www.idef.com/IDEF5.htm.

10. Titenko S.V. Pobudova Didaktichnoi Ontologii na Osnovi Analizu Elementiv Onyatiyno-Tezisnoi Modeli // Naukovi visti KhTUU "KPI". – 2010. – № 1. – S. 82–87.

11. Avtomaticheskaya Obrabotka Tekstov na Yestestvennom Yazike i Kompiuternaya Lingvistika: Uchebnoe Posobie / Bolshakova E.I., Klishinskiy E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. – M.: MIEM, 2011. – 272c.

12. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett, 2002. – 57(5). – P. 759-764.

13. 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.

14. Lande D.V., Snarskiy A.A. Kompaktificirovanniy Gorizontalniy Graf Vidimosti dlya Seti Slov // Trudi Mejdunarodnoy Nauchnoy Konferencii «Intellektualniy Analiz Informacii IAI-2013. Znania I Rassujdenia» – KPI, Kiev: 2013. – c. 158-164.

15. Vvedennya u Relyaciynu Model Danih – NOU «INTUIT». URL: http://www.intuit.ru/studies/professional_skill_improvements/1426/courses/74/lecture/2218.

Рецензія/Peer review : 2.3.2015 р. Надрукована/Printed :15.4.2015 р.
Рецензент: д.т.н., проф. Сорокатий Р.В.