AUTOMATYCZNE I TRADYCYJNE INDEKSOWANIE TREŚCI

Książka Piotra Malaka *Indeksowanie treści*[1] jest zmodyfikowaną wersją rozprawy doktorskiej[2]. Autor jest absolwentem kierunku Informacja naukowa i bibliotekoznawstwo na Uniwersytecie Mikołaja Kopernika w Toruniu. Od roku 2000 pracuje w tamtejszym Instytucie Informacji Naukowej i Bibliologii. Jego zainteresowania naukowe obejmują, jak pisze na stronie http://www.home.umk.pl/~piomk/?page_id=75, „efektywne zarządzanie czasem, zadaniami, kontaktami, informacją i wiedzą; data mining, information retrieval and extraction; corporate knowledge – wiedza korporacyjna; inżynieria lingwistyczna; technologie informacyjne wspierające pracę osób niepełnosprawnych"[3]. Czytelnicy „Zagadnień Informacji Naukowej" znają Piotra Malaka z artykułu *Rozwój badań nad przetwarzaniem języka naturalnego*[4]. Książka *Indeksowanie treści* jest moim zdaniem udaną próbą ponownego włączenia do dyskursu naukowego z zakresu bibliologii i informatologii problematyki indeksowania automatycznego.

Zagadnienia metod kwantytatywnych, lingwistyki komputerowej i ich zastosowań w bibliotekoznawstwie i w nauce o informacji, zwłaszcza w tzw. automatycznym indeksowaniu i/lub abstraktowaniu (streszczaniu), mają stosunkowo długą historię. Tworzą ją prace wielu badaczy i praktyków, od Hansa Petera Luhna i opracowanego przez niego pod koniec lat 50. XX w. systemu automatycznego tworzenia indeksów na podstawie tytułów dokumentów KWIC (KeyWords in the Context), poprzez Phyllis Baxendale, Geralda Saltona, Karen Spärck Jones, Sephena Robertsona, po Dereka Austina, żeby wymienić tylko niektórych spośród tych, którzy zainicjowali zainteresowanie tą problematyką czy (jak np. Salton) stworzyli podstawy teoretyczne do wielu późniejszych udanych prac. W Polsce większość podobnych prac lokowała się poza ścisłym kręgiem świata biblioteczno-informacyjnego, choć trzeba przyznać, że były one dość skrupulatnie śledzone i analizowane, przede wszystkim przez reprezentantów dyscypliny, która wówczas nosiła nazwę bibliotekoznawstwa i informacji naukowo-technicznej. Na pamięć z pewnością zasługują publikacje Mirosława

[1] P. Malak: *Indeksowanie treści: porównanie skuteczności metod tradycyjnych i automatycznych*. Warszawa 2012, 196 ss. (Nauka – Dydaktyka – Praktyka; 133).

[2] P. Malak: *Porównanie skuteczności metod automatycznych i kognitywnych w tworzeniu charakterystyk wyszukiwawczych dokumentów, ze szczególnym uwzględnieniem słów kluczowych*. Praca doktorska, promotor dr hab. prof. Uniwersytetu Wrocławskiego Adam Pawłowski. Wrocław, Uniwersytet Wrocławski, 2011. Recenzenci: prof. dr hab. Irena Kamińska-Szmaj, prof. dr hab. Wiesław Babik.

[3] Pisownia oryginalna.

[4] P. Malak: *Rozwój badań nad przetwarzaniem języka naturalnego.* „Zagadnienia Informacji Naukowej" 2010, nr 2(96), s. 21-30.

Dąbrowskiego (także napisana wspólnie z Krystyną Laus-Mączyńską książka *Metody wyszukiwania i klasyfikacji informacji*[5]), Heleny Dryzek, Janusza Bienia, Leonarda Bolca czy Czesława Daniłowicza. Ponad trzydzieści lat temu na łamach „Zagadnień Informacji Naukowej"[6] ukazał się do dziś aktualny artykuł Wojciecha Seroki napisany na podstawie pracy magisterskiej *Metody nadawania wag deskryptor*om (1980 r., promotor prof. dr hab. Michał Tempczyk). Wymienione osoby i publikacje są tylko okruchami mojej pamięci o ludziach i ich dokonaniach, nie jakimś uporządkowanym wyborem. Problematyka automatycznego indeksowania od dziesiątek lat do dziś z różnymi efektami jest podejmowana przez teoretyków i praktyków. Zmieniają się narzędzia, możliwości i moc sprzętu, kontekst kulturowy i scjentystyczny, ale problem nadal pozostaje aktualny i nie w pełni rozwiązany. Ostatnie lata, przynajmniej na gruncie polskiego bibliotekoznawstwa i nauki o informacji, to czas swoistego regresu w tym obszarze. Trudno byłoby znaleźć w piśmiennictwie polskim (ale już nie zagranicznym) z ostatnich kilkunastu lat poważne i rzetelne publikacje poświęcone tej problematyce oparte na solidnym fundamencie badania. Tymczasem potrzeba i waga tych zagadnień wcale nie maleje. Indeksowania automatycznego nie unikniemy i nie ma też powodów, żeby próbować go uniknąć. Sądząc po dotychczasowych efektach, przynajmniej w najbliższym czasie nie będzie możliwe wdrożenie go zamiast indeksowania manualnego (wykonywanego przez ludzi), ale będzie to narzędzie komplementarne, wspierające i przyspieszające procesy kognitywne. Książka Piotra Malaka w jakimś stopniu zapełnia lukę istniejącą i w polskiej bibliologii, i informatologii, i w praktyce biblioteczno-informacyjnej. Jest pierwszą od lat poważniejszą próbą zmierzenia się z pewnymi aspektami zagadnienia indeksowania automatycznego i przybliżenia choćby niektórych aspektów tego złożonego problemu.

Książka wyraźnie dzieli się na dwie, względnie niezależne części. Nie było to zapewne intencją Autora, również ja do tego nie zachęcam, ale w pewnych sytuacjach można każdą z tych części czytać jako dwa odrębne teksty. Obie składają się z dwóch rozdziałów. Pierwszy rozdział *Związki NLP z informacją naukową* jest wprowadzeniem do teorii przetwarzania języka naturalnego. Autor dokonał w nim pewnych ustaleń terminologicznych związanych z nazwą badań nad tekstami języka naturalnego, zaprezentował ważniejsze kierunki badawcze, ich cele i genezę. W kolejnym rozdziale zostały scharakteryzowane wybrane metody komputerowego przetwarzania i reprezentowania języka naturalnego, w tym analiza kwantytatywna tekstów, niektóre metody reprezentacji treści (wielozbiór, listy frekwencyjne, reprezentacje wektorowe), sposoby nadawania wag wyrazom i optymalizacji treści lingwistycznej.

Rozdziały trzeci i czwarty, tworzące drugą część publikacji, są prezentacją warunków, przebiegu i rezultatów badania przeprowadzonego przez Autora. W mojej ocenie to najcenniejsze elementy książki. Zostały w nich scharakteryzowane zasady, zgodnie z którymi było prowadzone badanie, system stworzony na jego potrzeby, cele i przedmiot badań, hipotezy i rezultaty. Badanie

[5] M. Dąbrowski, K. Laus-Mączyńska: *Metody wyszukiwania i klasyfikacji informacji*. Warszawa 1978.

[6] W. Seroka: *Niektóre zagadnienia deskryptorów ważonych*. „Zagadnienia Informacji Naukowej" 1981, nr 2 (39), s. 61-81.

polegało, najogólniej mówiąc, na porównaniu skuteczności wyszukiwawczej charakterystyk wyrażonych w postaci słów kluczowych, najpierw wygenerowanych automatycznie, a następnie wskazanych przez osoby indeksujące. Korpus badanych tekstów, związanych tematycznie z bibliologią i informatologią, tworzyły artykuły opublikowane w trzech rocznikach (2005-2007) „Przeglądu Bibliotecznego" i „Zagadnień Informacji Naukowej", a także artykuły z wybranych materiałów konferencyjnych. W sumie korpus tekstów poddanych badaniu zawierał ok. 850 tys. tokenów pochodzących ze 183 tekstów. Po sprowadzeniu do postaci podstawowej korpus liczył niecałe 39 tys. leksemów o średniej częstości wystąpień wynoszącej 22. „Oprócz tekstów poświęconych informacji naukowej badaniu poddano również artykuły z zakresu nauk ekonomicznych i zarządzania. Ten zbiór został utworzony głównie w celu weryfikacji ustaleń analiz przeprowadzonych na tekstach z zakresu informacji naukowej i bibliologii. Teksty te stanowiły mniejszy zestaw, łączna objętość wynosi ok. 195 300 tokenów pochodzących z 54 artykułów"[7]. Jako ciekawy przyczynek można przytoczyć, że wśród 20 leksemów o najwyższych frekwencjach dla obu czasopism 12 z nich pokrywa się. Są to: biblioteka (na pierwszym miejscu w obu czasopismach), dane (drugie miejsce w ZIN-ie i piąte w PB), informacja (trzecie miejsce w obu czasopismach), praca, naukowy, książka (na siódmym miejscu w ZIN-ie i dziewiątym w PB), język, bibliografia, użytkownik, badać, system i bibliograficzny.

Najważniejsze rezultaty badania można streścić następująco. Autorzy, dołączając słowa kluczowe do własnych tekstów, używają przeciętnie dwa razy mniej jednostek leksykalnych w porównaniu z zestawami tworzonymi przez osoby indeksujące. Ponad dwukrotna przewaga liczebna słownictwa użytego przez osoby indeksujące nie doprowadziła jednak do pełnej zgodności leksyki obu podzbiorów. „Średni stopień zgodności leksykalnej pomiędzy zestawami utworzonymi w wyniku procesów kognitywnych kształtował się na poziomie ok. 75%. Poziom ten wydaje się być górną granicą zgodności indeksowania tradycyjnego przeprowadzanego na tych samych dokumentach przez różne osoby"[8]. Oceniając możliwości automatycznego generowania słów kluczowych na podstawie rezultatów przeprowadzonego badania, Autor doszedł również do konkluzji, że „uzyskane poziomy zgodności zbiorów słownictwa, otrzymanego automatycznie i wskazanego w wyniku procesów indeksowania tradycyjnego, są zbyt niskie, żeby uznać proces automatyczny za wystarczający i równoważny z opracowaniem tekstu przez człowieka. Jednakże listy leksemów otrzymane automatycznie (…), mogą z powodzeniem wesprzeć proces opracowania rzeczowego dokumentów przez człowieka"[9]. Na podstawie doświadczenia zdobytego w efekcie przeprowadzonego badania Autor sformułował kilka postulatów technicznych dotyczących podobnych badań w przyszłości, propozycje dalszych badań oraz możliwości praktycznego wykorzystania uzyskanych rezultatów. Jego zdaniem „wyniki badań dotyczących możliwości automatycznego generowania słów kluczowych charakteryzujących treść dokumentu mogą okazać się przydatne we wszelkiego rodzaju repozytoriach" i bibliotekach cyfrowych.

[7] P. Malak: *Indeksowanie treści…*, op. cit., s. 122.
[8] P. Malak: *Indeksowanie treści…,* op. cit., s. 166.
[9] P. Malak: *Indeksowanie treści…,* op. cit., s. 167.

Dodam od siebie, że wygenerowane automatycznie listy leksemów mogłyby również być wykorzystane w co najmniej dwóch innych celach: przez twórców i administratorów języków informacyjno-wyszukiwawczych do określania i weryfikowania poziomu szczegółowości jednostek leksykalnych; do zwizualizowania realnego obrazu naszej dyscypliny, który wyłania się z analizy tematycznej publikacji. Dobrze by się stało, gdyby eksperyment opisany w recenzowanej publikacji, przeprowadzony przez jedną osobę w dość ograniczonym czasie wyznaczonym reżimem pracy nad rozprawą doktorską zainicjował szerszy projekt tego typu na dużo liczniejszym i bardziej zróżnicowanym materiale badawczym.

Książka Piotra Malaka może również być wykorzystana do zadań zapewne nie branych przez Autora pod uwagę, a mianowicie do analiz terminologicznych. Umiejscowienie problematyki indeksowania w kontekście automatyzacji pozwala na przykład lepiej zrozumieć, dlaczego jest to indeksowanie treści, a nie dokumentu (w „klasycznym" rozumieniu *dokumentu* w bibliologii i informatologii), dlaczego treść (*content*) nie jest tym samym co obiekt będący jej nośnikiem itd. Dobrze byłoby, gdyby również udało się przy okazji lektury doprowadzić do rozdzielenia w naszej specjalistycznej terminologii na poziomie leksykalnym znaczeń angielskich terminów *information retrieval* i *information searching,* które zwykle tłumaczone są na polskie *wyszukiwanie informacji,* choć nie zawsze jest to wybór optymalny.

Podsumowując, książka Piotra Malaka, choć nie pozbawiona pewnych mankamentów – pisanie o nich nie wydaje mi się konieczne, bo nie zmieniają mojej pozytywnej opinii, a mogą wyprowadzić uwagę potencjalnych czytelników poza to, co jest istotą tej publikacji – pod wieloma względami jest wyjątkowa. Po pierwsze, Autor podjął w niej problematykę zarzuconą jakiś czas temu przez innych badaczy i w pewnym sensie w Polsce niepopularną. Po drugie, wyraziście i praktycznie wykazał, w czym się wyraża interdyscyplinarność bibliologii i informatologii zarówno na poziomie analiz naukowych, praktyki badawczej, jak i realnych zastosowań. Wreszcie ostatni argument, lecz nie najmniej ważny, nie ograniczył swojej aktywności tylko do zreferowania cudzych badań, ale przeprowadził własne, prawidłowo zaplanowane, przygotowane, przeprowadzone i skomentowane badanie. Bardzo zachęcam do zapoznania się z książką Piotra Malaka *Indeksowanie treści***.**

Jadwiga Woźniak-Kasperek