Title: Indeksowanie treści - repozytoria cyfrowe : warsztaty (Warszawa, 27 marca 2014 r.)

Author: Anna Seweryn

Citation style: Seweryn Anna. (2014). Indeksowanie treści - repozytoria cyfrowe : warsztaty (Warszawa, 27 marca 2014 r.). "Nowa Biblioteka" Nr 2 (2014), s. 159-163

SPRAWOZDANIA

Nowa Biblioteka
nr 2 (15), 2014

Anna Seweryn
Zakład Zarządzania Informacją
Instytut Bibliotekoznawstwa i Informacji Naukowej
Uniwersytet Śląski w Katowicach
e-mail: anna.seweryn@us.edu.pl

INDEKSOWANIE TREŚCI – REPOZYTORIA CYFROWE (WARSZAWA, 27 MARCA 2014 R.)

Problematyka efektywnej organizacji dostępu do informacji na potrzeby usprawnienia procesów jej wyszukiwania stanowi od zawsze jedno z kluczowych zagadnień nauki o informacji, istotne zarówno w wymiarze teoretycznym, jak i – zwłaszcza – w kontekście praktycznej działalności informacyjnej. W ostatnich latach, w związku z intensywnym rozwojem technologii informacyjno-komunikacyjnych, szczególnej wagi nabrały kwestie związane z zarządzaniem informacją cyfrową, w tym z automatycznym indeksowaniem treści dokumentów elektronicznych gromadzonych, przechowywanych i udostępnianych w systemach komputerowych, takich jak biblioteki cyfrowe czy repozytoria sieciowe. Dostrzegając rynkową potrzebę pogłębiania specjalistycznej wiedzy i doskonalenia umiejętności zawodowych w tym zakresie, warszawska firma szkoleniowa Nova Skills Sp. z o.o. zamieściła w swojej ofercie – wśród rozmaitych kursów (m.in. z zakresu zarządzania, finansów, marketingu, prawa czy psychologii) – warsztaty *Indeksowanie treści – repozytoria cyfrowe*. Propozycja skierowana została przede wszystkim do administratorów serwisów internetowych i instytucjonalnych repozytoriów dokumentów cyfrowych oraz bibliotekarzy (w szczególności kadr działów IT bibliotek tradycyjnych i cyfrowych).

Szkolenie odbyło się w Warszawie 27 marca 2014 r. i miało na celu prezentację metod, technik i narzędzi przetwarzania elektronicznych dokumentów tekstowych na potrzeby automatycznego indeksowania oraz wyszukiwania informacji. Do poprowadzenia warsztatów zaangażowano jako eksperta Piotra Malaka z Instytutu Informacji Naukowej i Bibliologii Uniwersytetu Mikołaja Kopernika w Toruniu. Prowadzący niewątpliwie

posiada rozległą wiedzę w tym zakresie – w 2011 r. na podstawie rozprawy doktorskiej na temat *Porównanie skuteczności metod automatycznych i kognitywnych w tworzeniu charakterystyk wyszukiwawczych dokumentów, ze szczególnym uwzględnieniem słów kluczowych* uzyskał stopień doktora nauk humanistycznych w zakresie bibliologii; jest autorem szeregu publikacji z dziedziny informacji naukowej, dotyczących przede wszystkim problematyki zarządzania informacją i przetwarzania języka naturalnego na potrzeby systemów informacyjno-wyszukiwawczych[1].

Program szkolenia zakładał realizację dwóch modułów – teoretycznego i praktycznego. Zagadnienia teoretyczne prezentowane były w formie wykładu. Prowadzący rozpoczął prelekcję od scharakteryzowania podstawowych pojęć dotyczących indeksowania i wyszukiwania pełnotekstowego, porównania metod automatycznych z tradycyjnym, manualnym opracowaniem rzeczowym oraz wyjaśnienia różnic pomiędzy indeksowaniem sekwencyjnym (polegającym na indeksowaniu treści dokumentów *ad hoc* w poszukiwaniu wyrażeń sformułowanych w zapytaniu) i podejściem pro-aktywnym (gdy indeksowanie odbywa się niezależnie od zapytań kierowanych do systemu informacyjno-wyszukiwawczego). Następnie omówione zostały procesy związane z przygotowaniem tekstów do automatycznego indeksowania (tzw. *preprocessing*). Zwrócono uwagę na takie kwestie, jak: konieczność ujednolicenia strony kodowej plików w repozytorium wyszukiwawczym (obecnie zalecany jest standard UTF-8), zasadność oczyszczania plików ze znaczników (tzw. *parsing*), problemy związane z normalizacją tekstów (m.in. ujednolicanie wielkości liter czy sposobu zapisu liczebników, rozpoznawanie nazw własnych) w kontekście precyzji i kompletności wyszukiwania, specyfika słownictwa należącego do różnych stref wyodrębnionych ze względu na frekwencję (słownictwo częste, charakterystyczne, gramatyczne, rzadkie) oraz przydatność i zasady tworzenia wykazów słów małoznaczących / nieistotnych w procesie wyszukiwania (tzw. stoplist). Przedstawione zostały dwie najczęściej stosowane metody ujednolicania zapisu i znaczenia wyrazów: *stemming*, polegający na sprowadzaniu wyrazów do wspólnego rdzenia graficznego, oraz lematyzacja, której istotą jest ustalanie ich podstawowej formy gramatycznej. Wykładowca podkreślił zalety i wskazał ograniczenia obydwu tych strategii, podkreślając ich wpływ na efektywność wyszukiwania informacji – najważniejszym atutem *stemmingu* jest szybkość przetwarzania tą metodą tekstów zgromadzonych w repozytorium (kosztem mniejszej dokładności indeksowania), zaś lematyzacja zapewnia większą precyzję, ale wymaga większego nakładu czasu. Na przykładach zaczerpniętych z języka polskiego zilustrowano trud-

[1] Na podstawie rozprawy doktorskiej P. Malak opracował książkę *Indeksowanie treści. Porównanie skuteczności metod tradycyjnych i automatycznych* [4], wysoko ocenioną przez środowisko naukowe – zob. np. recenzję J. Woźniak-Kasperek [6].

ności natury formalnej i semantycznej, jakie wiążą się z optymalizacją dokumentów tekstowych na potrzeby ich automatycznego przetwarzania w elektronicznych systemach informacyjnych. Zasygnalizowana została ponadto dostępność narzędzi informatycznych opracowanych dla języka polskiego, służących poprawie efektywności indeksowania pełnotekstowego (projekty takie jak STEMPEL czy Morfologik, a także taksonomie, tezaurusy, ontologie).

Drugi blok zagadnień teoretycznych dotyczył różnych aspektów tworzenia reprezentacji treści dokumentów elektronicznych. Przedstawione zostały ogólne zasady wyszukiwania pełnotekstowego oraz trudności związane z realizacją tego procesu w systemach opartych na różnych metodach przechowywania i identyfikowania tekstów: pojedynczych lub zbiorczych plikach tekstowych oraz bazach danych. Zwrócono uwagę, że reprezentacje treści, tworzone na potrzeby automatycznego dopasowania dokumentu i zapytania, nie zawsze pozwalają odzyskać oryginalny dokument. Omawiając popularne sposoby indeksowania, wyróżniono reprezentacje: unigramowe (np. słowozbiór, lista frekwencyjna), wektorowe (macierz VSM) oraz n-gramowe (wielozbiór). Podkreślono fakt, iż wyszukiwanie polega na dopasowaniu terminów z zapytania do zapisów indeksowych (a nie bezpośrednio do treści dokumentów), toteż istotną kwestią jest sposób organizacji indeksów haseł, zapisywanych w odrębnych plikach o strukturze inwersyjnej. Najbardziej pożądanym sposobem prezentacji wyników wyszukiwania są listy rankingowe, w których kolejność odpowiedzi może być ustalana z wykorzystaniem algorytmów ważenia wyrazów wykorzystujących metody: statystyczne (związane z częstością wystąpień danego słowa w dokumencie i w całej kolekcji) lub probabilistyczne (np. Okapi BM25, wykorzystujące oprócz cech statystycznych również funkcje prawdopodobieństwa wystąpienia danego słowa). Część teoretyczna szkolenia zakończyła się omówieniem najczęściej stosowanych miar relewancji, stanowiących podstawę oceny efektywności systemów informacyjno-wyszukiwawczych (P@n – trafność wyników wyszukiwania dla n pierwszych wyników, R@n – kompletność wyników wyszukiwania dla n pierwszych wyników, MAP – współczynnik średniej trafności odpowiedzi na zapytania kierowane do systemu). Podczas wykładu P. Malak przedstawił ponadto kilka praktycznych wskazówek dotyczących sposobów usprawniania procesów indeksowania i wyszukiwania informacji w dużych repozytoriach cyfrowych.

Moduł określony jako praktyczny, zapowiadany w ofercie Nova Skills jako zasadnicza część proponowanego szkolenia, niestety z praktyką miał wspólny jedynie… temat. Początkowo program tego modułu miał obejmować szeroki wachlarz zagadnień: 1) Pakiet Lucene – najpopularniejsze otwarte oprogramowanie do indeksowania i wyszukiwania zasobów; 2) Pakiet SMART; 3) Biblioteki cyfrowe – indeksowanie pli-

ków DjVu; 4) Pakiet NLP Toolkit [3]. Kilka dni przed datą szkolenia program został jednak zmodyfikowany – treści ograniczono do zastosowania pakietu SOLR jako jednej z najczęściej wykorzystywanych platform wyszukiwania pełnotekstowego, opartej na bibliotece Lucene i rozwijanej przez Apache Foundation jako oprogramowanie typu *open source*. Zakres szkolenia, zgodnie z informacją zamieszczoną w drukowanych materiałach szkoleniowych i na stronie WWW organizatora [2], zawierał następujące zagadnienia: „możliwości, przeznaczenie i zastosowanie SOLR; instalacja i konfiguracja pakietu SOLR i Lucene; konfiguracja modułów dla języka polskiego; przygotowanie tekstów do indeksowania: indeksowanie różnych formatów plików, implementacja stoplist, *stemming* za pomocą pakietu Morfologik; ćwiczenia – indeksowanie i wyszukiwanie informacji: rozpoznawanie języka dokumentu i podejmowanie odpowiedniej akcji, sortowanie i filtrowanie list wyników, definiowanie parametrów ustalania wagi wyrazów, grupowanie wyników". Ku zaskoczeniu i rozczarowaniu większości uczestników, ta część szkolenia, deklarowana jako „praktyczna", odbyła się *de facto* w formie krótkiego i dość pobieżnego pokazu wybranych funkcjonalności oprogramowania, bez aktywnego zaangażowania kursantów do wykonania jakichkolwiek ćwiczeń. Jedyną osobą w wynajętej sali, która miała do dyspozycji komputer z zainstalowanym specjalistycznym oprogramowaniem, był prowadzący szkolenie; kursanci mogli jedynie obserwować wykonywane przez niego działania na ekranie projekcyjnym.

Konsternacja uczestników szkolenia była tym bardziej uzasadniona, że na etapie naboru kursantów podkreślano warsztatowy charakter proponowanego szkolenia (pierwotnie było ono zatytułowane „Indeksowanie treści w teorii i praktyce – warsztaty") i zapewniano, że istotnym jego elementem będą „ćwiczenia instalacji i konfiguracji serwera indeksującego wraz z wyszukiwarką oraz dostosowania zainstalowanego systemu do pracy z polskojęzycznymi dokumentami" [2]. Precyzując korzyści z udziału w proponowanych warsztatach, organizatorzy przekonywali, że uczestnicy m.in. „samodzielnie zainstalują system indeksująco wyszukiwawczy i skonfigurują go do pracy z tekstami w języku polskim" oraz „będą mieli okazję przetestować różne ustawienia systemu indeksującego oraz sprawdzić ich wpływ na efektywność procesu wyszukiwania informacji" [2]. Obietnic tych jednak nie dotrzymano – szkolenie w przeważającej części miało charakter poprawnego, ale tradycyjnego akademickiego wykładu, wspomaganego prostą prezentacją w PowerPoincie i drukowanymi materiałami pomocniczymi.

Przykład tego szkolenia pokazuje, że do sukcesu edukacyjnego nie wystarczy atrakcyjna, tzn. interesująca i aktualna tematyka ani kompetentny trener – uchybienia organizacyjne (a za takie można uznać niezapewnienie uczestnikom warsztatów odpowiednio skonfigurowanego sprzętu

komputerowego) mogą przesądzić o niepowodzeniu przedsięwzięcia i zrazić kursantów do udziału w kolejnych inicjatywach proponowanych przez daną firmę szkoleniową. Problematykę podjętą podczas marcowych warsztatów z pewnością warto kontynuować w szerszym zakresie, zwłaszcza w aspekcie praktycznym, podczas kolejnych szkoleń. Firma NovaSkills przewiduje w październiku br. organizację dwu spotkań poświęconych tym zagadnieniom – 14 października ma odbyć się konferencja naukowa *I Forum Indeksowania Treści – rozwiązania praktyczne* [1], natomiast dzień później, 15 października, komplementarne warsztaty: *Indeksowanie treści – rozwiązania praktyczne* [5]. Zaproszeni prelegenci i zapowiadane tematy wystąpień rokują wysoki poziom merytoryczny; należy oczekiwać, że tym razem szkolenia zostaną przygotowane z większą dbałością o aspekty organizacyjne.

Bibliografia:

[1] *I Forum Indeksowania Treści – rozwiązania praktyczne*. W: *NovaSkills. Wiedzą inspirujemy przyszłość* [online]. Wrzesień 2014. [Data dostępu: 25.09.2014]. Dostępny w World Wide Web: http://www.nskills.pl/Kalendarz,wydarzen/I_Forum_Indeksowania_Tresci_-_rozwiazania_praktyczne/153,,0,11,1.

[2] *Indeksowanie treści – repozytoria cyfrowe*. W: *NovaSkills. Wiedzą inspirujemy przyszłość* [online]. Luty 2014. [Data dostępu: 20.05.2014]. Dostępny w World Wide Web: http://www.novaskills.pl/Kalendarz,wydarzen/Indeksowanie_tresci_-_repozytoria_cyfrowe/58,,0,21,1.

[3] *Indeksowanie treści w teorii i praktyce – warsztaty* [online]. Marzec 2014. [Data dostępu: 20.05.2014]. Dostępny w World Wide Web: http://www.novaskills.pl/docs/Indeksowanie_tresci.pdf.

[4] Malak P.: *Indeksowanie treści. Porównanie skuteczności metod tradycyjnych i automatycznych*. Warszawa 2012. ISBN 978-83-61464-42-6.

[5] *Warsztaty: „Indeksowanie treści – rozwiązania praktyczne*. W: *NovaSkills. Wiedzą inspirujemy przyszłość* [online]. Wrzesień 2014. [Data dostępu: 25.09.2014]. Dostępny w World Wide Web: http://www.nskills.pl/Kalendarz,wydarzen/Warsztaty_Indeksowanie_tresci_-_rozwiazania_praktyczne/154,,0,11,1.

[6] Woźniak-Kasperek J.: *Automatyczne i tradycyjne indeksowanie treści*. „Zagadnienia Informacji Naukowej" 2012 nr 1, s. 95-98. ISSN 0324-8194.