ANNA SZCZEPAŃSKA
Uniwersytet Warszawski
Instytut Informacji Naukowej i Studiów Bibliologicznych
e-mail: aszczepanska@uw.edu.pl

PODSTAWOWE STRATEGIE WYSZUKIWANIA INFORMACJI I ICH WYKORZYSTANIE W PRAKTYCE

ABSTRAKT: W artykule dokonano przeglądu strategii wyszukiwania stosowanych we współczesnych systemach informacyjno-wyszukiwawczych. Zaprezentowano sześć powszechnie używanych strategii: wyszukiwanie proste, strategię formowania klas, kolejnych klas, podwójnych klas, pomnażanie odwołań oraz indeksów cytowań. Przedstawiono właściwe dla każdej z tych strategii zasady prowadzenia wyszukiwań zilustrowane przykładami kwerend przeprowadzonych w bazach danych dostępnych w Bibliotece Uniwersyteckiej w Warszawie. Podjęto próbę odpowiedzi na pytanie, w odniesieniu do jakich problemów wyszukiwawczych użycie danej strategii powinno przynieść najlepsze efekty oraz z jakich mankamentów należy zadawać sobie sprawę przy ich stosowaniu.

WPROWADZENIE

Celem niniejszego artykułu jest przedstawienie strategii wyszukiwania, które mogą być efektywnie wykorzystywane we współczesnych systemach informacyjno-wyszukiwawczych dostępnych online. Wybrano sześć najpowszechniej stosowanych strategii, które zilustrowane zostały przykładami wyszukiwania[1]. Dokonana analiza ma służyć wskazaniu, w jakich sytuacjach dana strategia może być użyteczna oraz ewentualnych mankamentów jej stosowania. Jednak przed omówieniem konkretnych strategii należy rozstrzygnąć kilka kwestii terminologicznych. Używany w artykule termin „strategia wyszukiwania" rozumiany jest jako przemyślany plan działań prowadzący do takiego sposobu zapisania problemu wyszukiwawczego, który pozwoli zidentyfikować maksymalną liczbę

[1] Wyszukiwania prezentowane w przykładach przeprowadzone zostały w oparciu o bazy danych, do których dostęp wykupuje Biblioteka Uniwersytecka w Warszawie.

relewantnych dokumentów przy minimalnej liczbie operacji przeszukiwania systemu informacyjnego. Z kolei „taktyka wyszukiwania" zdefiniowana jest jako działanie przedsięwzięte w procesie wyszukiwania informacji, często mające charakter heurystyczny, którego celem jest uzyskanie zbioru relewantnych wyników [2].

W świetle tych rozróżnień terminologicznych może nasunąć się pytanie czy słuszne jest używanie takich sformułowań jak „strategia formowania klas", czy „strategia indeksów cytowań". Skoro strategia rozumiana jest jako całość działań podejmowanych od momentu zaistnienia potrzeby informacyjnej do jej zaspokojenia, właściwiej byłoby mówić o „taktyce formowania klas", czy „taktyce indeksów cytowań". Wynika to z faktu, że działania, które zostaną opisane w dalszej części stanowią tylko niewielki fragment zespołu operacji składających się na wyszukiwanie informacji, np. dotyczą etapu konstruowania instrukcji wyszukiwawczej. Takie rozwiązanie byłoby słuszne z logicznego punktu widzenia, lecz w praktyce istnieje bardzo mocno zakorzeniony zwyczaj, aby używać takich sformułowań jak właśnie „strategia formowania klas" (mowa tu zwłaszcza o literaturze anglojęzycznej, zob. np. Harter, 1986; Meadow 1992). Nazywanie rodzaju strategii według rodzaju taktyk użytych na etapie interakcji z systemem jest swoistym skrótem myślowym, który ma wskazać najbardziej charakterystyczne elementy obranej metody postępowania. W niniejszym artykule zdecydowano się używać nazw zgodnie z powszechnie przyjętą praktyką, mając jednak świadomość wykazanej powyżej nieprecyzyjności takiego rozwiązania.

Kolejna kwestia wymagająca komentarza dotyczy tłumaczenia angielskich nazw analizowanych strategii. W piśmiennictwie polskim nie pojawiły się do tej pory powszechnie przyjęte odpowiedniki tych nazw. Praktyka wskazuje, że osoby zajmujące się wyszukiwaniem informacji często posługują się angielskimi nazwami strategii jako zrozumiałymi wśród specjalistów. Przyjęte w artykule tłumaczenia

[2] Więcej na temat definiowania i rozumienia tych pojęć zob. Szczepańska, 2006.

nazw strategii to jedynie propozycje, wraz z rozwojem badań w tej dziedzinie okaże się zapewne, jakie polskie odpowiedniki tych terminów przyjmą się w praktyce.

WYSZUKIWANIE PROSTE

Można zastanawiać się, czy wyszukiwanie proste (ang. *briefsearch)* kwalifikuje się do miana odrębnej strategii wyszukiwania. W wyszukiwaniu prostym instrukcja sformułowana jest w najprostszy możliwy sposób. Zawiera jeden lub kilka terminów, połączonych operatorami Boole'a. Jednak rzadko zdarza się, aby problematyka zapytania dała się odzwierciedlić za pomocą pojedynczego terminu wyszukiwawczego. Z tego względu, zależnie od tego jak szczegółowego lub ogólnego użyje się terminu wyszukiwawczego, wyszukiwanie proste charakteryzuje się zazwyczaj niskim współczynnikiem kompletności bądź dokładności.

Zdarzają się jednak zapytania informacyjne, w przypadku których uzasadnione jest wykorzystanie tego sposobu formułowania instrukcji. Znajduje ono zastosowanie w sytuacjach, kiedy zapytanie informacyjne dotyczy ściśle zdefiniowanego problemu, który ma unikalną reprezentację językową. Wyszukiwanie proste szczególnie użyteczne jest w przypadkach, gdy można użyć określonej, nazwy własnej, ponieważ zazwyczaj eliminuje to problem występowania terminów synonimicznych. Oczywiście dane zagadnienie może zostać uwzględnione w artykule dotyczącym ogólniejszych zagadnień - w takim wypadku artykuł zostanie zaindeksowany zapewne terminem szerszym. Np., w przypadku poszukiwania artykułów o Witoldzie Gombrowiczu, wzmianki o nim mogą pojawić się w artykułach opisanych terminem „pisarze polscy". Jednak w przypadku wyszukiwania artykułów, które opisane są terminem szerszym niż poszukiwany, założyć można, że zazwyczaj zawierają jedynie niezbyt obszerne, ogólne wzmianki o interesujących zagadnieniach. Czy takie

wzmianki są istotne dla prowadzonego wyszukiwania zależy już od specyfiki zapytania informacyjnego.

Najbardziej powszechne zastosowanie wyszukiwania prostego to sytuacja, w której wyszukiwanie ma na celu znalezienie konkretnego rekordu, który można jednoznacznie zidentyfikować na podstawie posiadanych wcześniej danych – np. wyszukiwanie opisu książki na podstawie nazwiska autora i słów z tytułu, w celu sprawdzenia nazwy wydawcy.

Nieco innym rodzajem wyszukiwania prostego, jest tzw. wielokrotne wyszukiwanie proste (ang. *multiple briefsearch*). Różni się ono sposobem przeprowadzenia, lecz nie budową instrukcji. Używane jest w sytuacji, gdy temat wyszukiwania można wyrazić za pomocą pojedynczych terminów, lecz jego specyfika powoduje, że zbiór wyników z jednej bazy najprawdopodobniej będzie niewystarczający. W takim wypadku taką samą instrukcję wyszukiwania wprowadza się do możliwie wielu baz danych. Poniżej wyszukiwanie proste zilustrowane zostanie takim właśnie przykładem zrealizowanym w bazach danych dostępnych w Bibliotece Uniwersyteckiej w Warszawie.

Przykład 1

Wyszukane mają być informacje na temat francuskiego pisarza Sébastien Roch Nicolas (1741-1794), który znany był pod pseudonimem Chamfort (jest to nazwa miejsca urodzenia Nicolasa). W encyklopediach ogólnych biogram tego pisarza można znaleźć pod hasłem „Chamfort" (zob. *Encyclopædia Britannica Online*, 2006; *Nowa encyklopedia…*, 1998, t. 1, s. 679). Dlatego też w wyszukiwaniu zdecydowano się nie używać jego prawdziwego nazwiska a jedynie pseudonimu.

Chamfort był moralistą i aforystą, przez pewien czas pełnił funkcje dyrektora Bibliothèque Nationale w Paryżu, lecz dziś jest raczej mało znaną postacią. Nie można spodziewać się dużego zbioru rekordów w wyniku planowanego wyszukiwania, toteż zdecydowano się przeprowadzić wyszukiwanie symultanicznie w wielu bazach.

Przeszukane zostały bazy konsorcjum *ProQuest*, gdyż spośród wszystkich serwisów, do których dostęp wykupiła BUW, pozwala ono na

prowadzenie wyszukiwania w największej liczbie baz jednocześnie. Dla analizowanego przypadku są to następujące bazy: *ABI/INFORM Dateline, ABI/INFORM Global, ABI/INFORM Trade & Industry, Career and Technical Education, Dissertations and Theses; ProQuest Dissertations and Theses - A&I, ProQuest Agriculture Journals, ProQuest Computing, ProQuest Education Journals, ProQuest Nursing Journals, ProQuest Science Journals, ProQuest Social Science Journals, ProQuest Telecommunications* (14 baz). Mając na względzie prawo Bradforda, zgodnie z którym aż 1/3 publikacji na dany temat znajduje się w piśmiennictwie nie należącym do danej dziedziny, nie wykluczono z wyszukiwania nawet baz, w których rejestracja prac o Chamforcie wydaje się mało prawdopodobna. Instrukcja wyszukiwania zawiera jedynie termin „Chamfort", nie został wybrany żaden indeks, w związku z tym system domyślnie przeszukuje wszystkie pola opisu bibliograficznego oraz abstraktu tekstu dokumentu, jeśli jest on dostępny.

Wynik wyszukiwania to 11 rekordów, z których tylko pierwszy jest nierelewantny (dotyczy muzyka, który używa podobnego nazwiska). Większość rekordów to opisy dysertacji, w których zajmowano się osobą Chamforta. Są wśród nich zarówno prace francuskie, jak i napisane na uniwersytetach innych państw, tak więc wyszukujący może ocenić, które prace mogą być dla niego dostępne (dodatkowo *ProQuest* oferuje możliwość płatnego zamówienia kopii takiej pracy). Temat wyszukiwania nie sugerował tego, że wyszukanie w bazie dysertacji może okazać się tak użyteczne. Gdyby selekcjonowano bazy odpowiednie dla tej tematyki wybór tej konkretnej bazy nie byłby wcale oczywisty.

Trzy artykuły pochodzą z prasy codziennej i są recenzjami książki zawierającej biografię Chamforta. Dla użytkownika, który nie wiedział o istnieniu takiej publikacji może być to cenna informacja.

W celu dokonania oceny użyteczności zastosowania omawianej strategii dokonano dodatkowego wyszukiwania w bazie firmy *JSTOR*, która m.in. bogatą kolekcję czasopism z zakresu nauk humanistycznych. Instrukcja pozostała taka sama, przeszukiwane są pola tytułu, abstraktu oraz nagłówków artykułów (*caption*).

Wynik wyszukiwania to dwa artykuły z czasopism naukowych, które dotyczą właśnie osoby Chamforta. Baza oferuje dostęp do pełnych tekstów, informacje w nich zawarte mają większą wartość naukową niż informacje, które uzyskano w poprzednim wyszukiwaniu.

Jednak materiały pochodzące jedynie z dwóch źródeł mogą nie być dla użytkownika wystarczające. Właśnie w takim wypadku należy użyć wyszukiwania w wielu bazach – dostarczone materiały mogą być niższej jakości (nie jest to jednak regułą), lecz alternatywą jest niewystarczający zbiór rekordów.

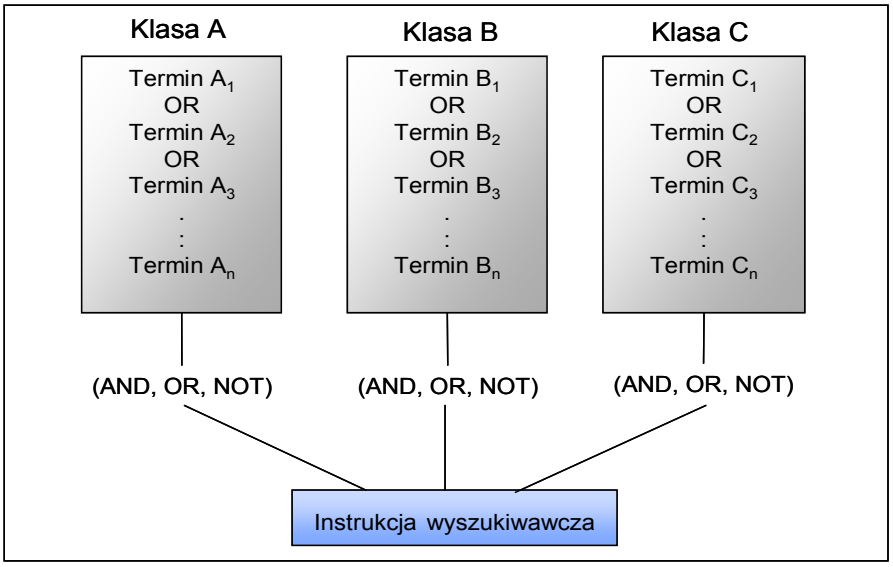
Najczęściej jednak wyszukiwanie proste nie stanowi samodzielnej strategii wyszukiwania. Specjaliści informacji używają go zwykle, aby sprawdzić zawartość bazy danych. Wprowadzenie takiej instrukcji dostarcza wtedy informacji, czy dany termin uwzględniany jest w bazie danych, które rekordy zostały zaindeksowane przy jego pomocy (szczególnie użyteczne przy terminach niejednoznacznych), oraz jak wiele rekordów w bazie dotyczy wyszukiwanego tematu, czy też jednego z jego aspektów. Wyszukiwanie proste wykorzystywane w taki sposób stanowi często punkt wyjścia lub komponent innych strategii wyszukiwawczych.

STRATEGIA FORMOWANIA KLAS

Strategia formowania klas (ang. *building blocks strategy*) jest najczęściej stosowną strategią. Jedną z przyczyn tego jest zapewne fakt, iż strategia ta jest bardzo użyteczna do przeszukiwania niekontrolowanego środowiska, jakim jest Internet oraz przy wyszukiwaniu pełnotekstowym.

Chcąc wyszukiwać informacje za jej pomocą, należy najpierw zidentyfikować główne terminy, którymi można odzwierciedlić treść zapytania informacyjnego. Należy również ustalić, jakie relacje zachodzą między tymi terminami tak, aby można je było wyrazić przy pomocy operatorów Boole'a. Następnie dla każdego terminu należy znaleźć inne wyrażenia, które reprezentują daną klasę zagadnień. Mogą to być synonimy, terminy węższe, szersze lub jedynie powiązane, wyrażone w języku naturalnym oraz wszystkich stosowanych w danej bazie językach informacyjno-wyszukiwawczych. Należy również podjąć decyzję, jakie indeksy będą przeszukiwane. Terminy i wyrażenia

reprezentujące dany aspekt zapytania informacyjnego, połączone za pomocą operatora OR tworzą klasy. Liczba rekordów, które będą relewantne dla takiej klasy zazwyczaj jest stosunkowo duża. Jednak ostateczna instrukcja wyszukiwawcza, zawierająca wszystkie uformowane klasy, połączone odpowiednimi operatorami, zwykle skutkuje już nie tak licznym zbiorem relewantnych rekordów. Sposób tworzenia takiej strategii najłatwiej wytłumaczyć przedstawiając ten proces za pomocą poniższego schematu.



Tablica 1: Strategia formowania klas.

Przykład 2

Zapytanie informacyjne dotyczy informacji na temat pielgrzymki do Egiptu, którą w 2000 roku odbył Jan Paweł II. Poszukiwane są reportaże, analizy, komentarze publikowane na ten temat przez media na całym świecie, choć język publikacji ma być ograniczony do angielskiego.

Wyszukiwanie przeprowadzone zostanie w bazie *Factiva.com*. Według informacji zamieszczonych na stronie serwisu, baza oferuje dostęp do wiadomości z 120 agencji prasowych na całym świecie. Są to nie tylko publikacje prasowe, lecz również transkrypcje programów radiowych i telewizyjnych wiodących stacji informacyjnych. Przeprowadzenie przeszukiwania w pełnych tekstach artykułów w tak dużej bazie skutkowałoby bardzo długą listą wyników. Zapewne duża część z nich byłaby nierelewantna. Dlatego w celu zwiększenia dokładności wyszukiwania zdecydowano się przeszukiwać jedynie nagłówki i pierwsze akapity wiadomości (ang. *headlines and lead paragraph*; w bazie pole to oznaczone jest HLP). Ponadto, aby wyeliminować krótkie wzmianki prasowe, w instrukcji zamieszczono komendę pomijania wszystkich publikacji krótszych niż pięćset słów. W tym celu do instrukcji

za pomocą operatora AND dodano komendę „WC>500” Kolejne wprowadzone ograniczenie dotyczy daty. W prasie często publikowane bywają zapowiedzi oraz spekulacje czy przypuszczenia na temat papieskich pielgrzymek. Ponieważ nie wchodzą one w obręb wyszukiwania, ograniczono zasięg czasu publikacji do przedziału: od 24 lutego 2000 r. (pierwszy dzień pielgrzymki) do 1 kwietnia 2006 r. (dzień przeprowadzenia wyszukiwań). W bazie *Factiva.com* to ograniczenie wprowadza się w odrębnym polu, dlatego zapis tej komendy nie zostanie odwzorowany w prezentowanej poniżej instrukcji wyszukiwawczej. W instrukcji zawarto następujące terminy:

Klasa 1 Papież	Klasa 2 Pielgrzymki	Klasa 3 Egipt
“John Paul” pope Wojty?a	pilgrimage* trip* visit* travel*	Egypt Cairo “Mount Sinai”
Instrukcja wyszukiwawcza: Papież AND Pielgrzymki AND Egipt HLP=[("John Paul" OR pope OR Wojty?a) AND (pilgrimage* OR trip* OR visit* OR travel*) AND (Egypt OR Cairo OR “Mount Sinai")] AND WC>500		

Możliwości, którymi dysponują bazy pozwalają ograniczyć liczbę terminów zawartych w instrukcji. Zamiast wpisywać termin „Wojtyła” oraz „Wojtyla” (czasami pojawiający się w prasie zagranicznej), wystarczy użyć znaku „?”, maskującego jedną literę. Podobnie nie trzeba wpisywać takich terminów, jak np. „trip” w liczbie pojedynczej i mnogiej, czy odmieniać czasowników – „visit”, visited”, „visiting”, itd. Frazy należy mieścić w cudzysłowie. W klasie pierwszej nie wpisano frazy „John Paul II” oraz „Karol Wojtyła”, aby nie ograniczać zbytnio wyszukiwania. Istnieje bardzo mała szansa, aby artykuły zawierające termin „Wojtyła” dotyczyły innej osoby niż papież. W ostatniej klasie obok terminu „Egypt” uwzględniono „Cairo” oraz „Mount Sinai” (Góra Synaj). Pobieżny przegląd wydawnictw informacyjnych pozwolił stwierdzić, że wizyta w Egipcie trwała tylko dwa dni, w trakcie których Jan Paweł II odwiedził te dwa miejsca.

W wyniku wyszukiwania otrzymano 38 rekordów. Wszystkie są relewantne. Zupełnie odrębnym zagadnieniem jest kwestia ich pertynencji, ponieważ wiele z artykułów zawiera bardzo podobną treść.

Wskaźnik dokładności osiągnął w przeprowadzonym wyszukiwaniu rzadko uzyskiwaną wartość równą 1 (100%). Jednak zapewne zupełnie

inaczej przedstawia się wskaźnik kompletności, którego nie można dokładnie określić. Autorzy publikacji prasowych często posługują się metaforami, czy innymi sformułowaniami, które wykorzystane w tytułach artykułów nie wskazują precyzyjnie tematyki publikacji. Można przypuszczać, że szczególnie liczne były artykuły relewantne, w których nagłówkach nie zawarto sformułowań z klasy trzeciej. Potwierdza to dodatkowe wyszukiwanie, dla którego instrukcja została tak sformułowana, aby wyłonić relewantne rekordy, które umknęły w poprzednim wyszukiwaniu: HLP=[("John Paul" OR pope OR Wojty?a) AND (pilgrimage* OR trip* OR visit* OR travel*) **NOT** (Egypt OR Cairo OR "Mount Sinai")] AND WC>500. Ponieważ taka instrukcja dotyczyłaby wszystkich papieskich pielgrzymek, poza tą do Egiptu, ograniczono daty publikacji artykułów do czasu trwania pielgrzymki 24-26.02.2000.

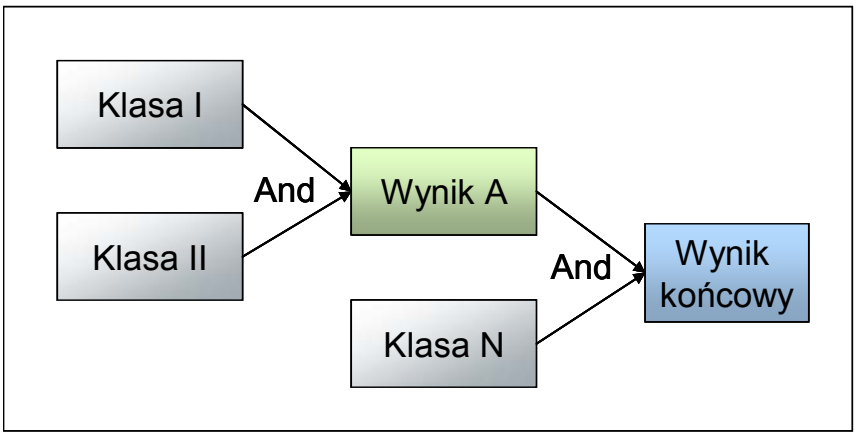
Wśród trzynastu odnalezionych rekordów dziesięć jest relewantnych. Tak więc można założyć, iż współczynnik kompletności pierwszego wyszukiwania jest raczej niski. Jeśli osoba formułująca zapytanie informacyjne kładłaby nacisk na kompletność wyszukiwania, należałoby użyć różnych taktyk wyszukiwawczych, aby poprawić ten współczynnik.

Istnieje wiele problemów i sytuacji wyszukiwawczych, dla których uzasadnione jest posługiwanie się tą strategią. Szczególnie użyteczna jest w sytuacjach przeszukiwania materiałów o niekontrolowanym zasobie słownictwa, jak również w sytuacjach, gdy przeszukuje się symultanicznie wiele baz, w których mogą funkcjonować różnie sformułowane hasła. Zaletą tej strategii jest również to, że stosunkowo łatwo daje się wprowadzać niewielkie modyfikacje instrukcji wyszukiwania, w celu osiągnięcia określonego poziomu współczynników kompletności (ścisłe określenie tego współczynnika jest oczywiście niemożliwe) i dokładności. Współczynnik kompletności będzie ulegał zmianom w zależności od tego jak wiele terminów uwzględnimy w poszczególnych klasach. Z kolei dokładność wyszukiwania jest zależna od liczby klas połączonych operatorami AND lub NOT (w sytuacji, gdy dołączylibyśmy klasę za pomocą operatora OR malałaby precyzja a rosła kompletność). Strategia ta

możliwa jest do stosowania w większości obecnie istniejących baz, jakkolwiek nie we wszystkich.

STRATEGIA KOLEJNYCH KLAS

Strategia kolejnych klas *(*ang. *succesive facet strategy)* również wykorzystuje metodę budowania klas odzwierciedlających różne aspekty poszukiwanego zagadnienia. Różni się ona od strategii formowania klas sposobem wprowadzania wybranych grup terminów do systemu. Stosowana jest najczęściej w sytuacji, gdy wyszukujący obawia się, że wprowadzenie do systemu wszystkich wyróżnionych faset jednocześnie będzie skutkowało bardzo małym zbiorem relewantnych rekordów. W takim wypadku należy zacząć od stworzenia tylko jednej fasety, wprowadzić ją do systemu i w zależności od uzyskanych rezultatów podjąć decyzję, jaką kolejną klasę należy wprowadzić i za pomocą jakiego operatora. Może pojawić się sytuacja, kiedy już pierwsze wyszukiwanie wykaże, że dodatkowe fasety są niepotrzebne, lecz zdarza się to rzadko. Kluczowym zagadnieniem dla strategii kolejnych klas jest ustalenie, dla którego aspektu tematyki wyszukiwania sformułować klasę wprowadzaną do systemu jako pierwszą.



Tablica 2: Strategia kolejnych klas.

W literaturze przedmiotu proponuje się dwa możliwe rozwiązania, które w praktyce bardzo niewiele różnią się od siebie (zob. Harter, 1986, p. 177-180). Według pierwszego z nich najpierw tworzy się klasę

dla najbardziej specyficznego, charakterystycznego terminu (ang. *the most specific concept first)*. Często takim terminem jest nazwa własna instytucji, osoby, miejsca, technologii, itp. (np. w przypadku wyszukiwania na temat nauki języka angielskiego metodą Callana, terminem specyficznym jest „metoda Callana").

Druga wytyczna, według której ustala się początkową klasę, to wybór terminu, po wprowadzeniu którego system wygeneruje najmniejszą liczbę rekordów (ang. *the lowest postings first*). Często, lecz nie zawsze, ten termin jest jednocześnie terminem najbardziej specyficznym. Niektórzy autorzy rozróżniają dwie odrębne strategie wyszukiwania ze względu na te dwie metody wyboru pierwszej klasy (termin najbardziej specyficzny, czy o najmniejszej liczbie odwołań).

W piśmiennictwie fachowym czasami omawiany jest jeszcze trzeci rodzaj strategii kolejnych klas, który w terminologii angielskiej nosi nazwę *succesive fraction* (strategia kolejnych komponentów). Tak jak poprzednie, polega ona na wprowadzeniu pojedynczej klasy do systemu, aby następnie zawężać otrzymany wynik poprzez dodawanie kolejnych klas. Różnica polega na tym, iż w pierwszej klasie zawarte są również formalne cechy poszukiwanych dokumentów, takie jak język, data publikacji, typ dokumentu.

Niezależnie od kryteriów wyboru pierwszej wprowadzanej klasy oraz od sposobu jej sformułowania, uzyskany w pierwszej kolejności wynik będzie charakteryzował się wysoką kompletnością i niską dokładnością. Wprowadzanie następnych klas ma na celu poprawienie współczynnika dokładności. Każdorazowo uzyskany wynik podlega ocenie przez wyszukującego, który dołącza kolejne klasy do momentu, kiedy uzyskany zbiór rekordów uzna za satysfakcjonujący.

W niniejszym artykule uznano, iż nie ma potrzeby wyróżniania trzech odrębnych strategii ze względu na specyfikę pierwszej wprowadzanej klasy. W każdym z opisywanych wypadków schemat postępowania jest podobny: w pierwszej kolejności wybiera się klasę najbardziej charakterystyczną dla danego wyszukiwania, aby następnie poprzez kolejne operacje poprawiać współczynnik

dokładności. Dlatego też wytyczne, których nazwy angielskie brzmią „the most specific concept first", „the lowest postings first" oraz „succesive fractions" zostały uznane za reguły, wchodzące w obręb jednej strategii, nazywanej tu strategią kolejnych klas.

Przykład 3

Poszukiwane są informacje na temat zbiorów na tradycyjnych nośnikach papierowych znajdujących się w Bibliothèque Nationale de France. Wyszukiwanie ma dotyczyć całokształtu zagadnień związanych z wszelkimi dokumentami papierowymi (książki, czasopisma, stare druki, kartografia, i in.), takich jak udostępnianie, polityka gromadzenia, stan zachowania, itd. Poszukiwanie powinno ograniczyć się do artykułów z czasopism w języku angielskim, które ukazały się w ciągu ostatnich 10 lat (1996-2006).

Bazą, w której można spodziewać się największej liczby materiałów na temat bibliotek jest *LISA (Library and Information Science Abstracts)*. Instrukcja wprowadzana do systemu jako pierwsza dotyczy klasy „Bibliothèque Nationale de France". Przeszukiwany jest indeks deskryptorów (ang. *descriptors)*, co powinno zapewnić większą precyzję wyszukiwania. Analiza terminów w tezaurusie pozwoliła ustalić, że używane są dwa deskryptory opisujące interesującą bibliotekę: „Bibliotheque Nationale de France", oraz „Bibliotheque Nationale". Te dwa terminy oraz powszechnie używany skrót nazwy biblioteki „BNF" składają się na pierwszą instrukcję: „DE=("Bibliotheque Nationale de France" or "Bibliotheque Nationale" or "BNF")".

W wyniku wyszukiwania otrzymano 369 rekordów. To duży zbiór, dla którego można wprowadzić ograniczenia. Specyfika tematu wyszukiwania powoduje, że większość publikacji jest w języku francuskim. Jeśli zbiór byłby zbyt mały, można by zastanawiać się nad sensownością wprowadzania ograniczenia językowego – być może pomocne byłoby czerpanie informacji z dostępnych w LISA abstraktów w języku angielskim, a nie z oryginalnych tekstów artykułów. Jednak w tym przypadku nie ma takiej konieczności, tak wiec instrukcja wyszukiwania zostanie uszczegółowiona poprzez podanie języka publikacji (język angielski), daty publikacji (1996-2006) oraz określenia typu publikacji (artykuły z czasopism).

Otrzymany wynik to 31 rekordów. Wstępny przegląd otrzymanych rekordów pozwolił ustalić, że niektóre z publikacji dotyczą Bibliothèque Nationale w Quebec. System wyszukuje wyrażenie „Bibliotheque Nationale" nie tylko jako wyrażenie samoistne (według tezaurusa oznaczające Bibliotekę Narodową w Paryżu), ale również jako część innych deskryptorów, w tym wypadku „Bibliotheque Nationale du Quebec". Wobec tego zdecydowano się wykluczyć w instrukcji nazwę „Bibliotheque Nationale du Quebec" za pomocą operatora NOT. Używając tego operatora wprowadzono również termin „book review", aby wyeliminować recenzje książek (poszukiwane miały być jedynie artykuły z czasopism). Instrukcja wprowadzona do systemu ma więc postać: DE=("Bibliotheque Nationale de France" OR "Bibliotheque Nationale" OR "BNF") AND LA=("English") AND PT=(journal article) NOT DE=("Bibliotheque Nationale du Quebec" OR "Book review"). Dodatkowo oznaczona została data publikacji (1996-2006).

Uzyskany wynik to 28 rekordów. Jak do tej pory w instrukcji uwzględniono najbardziej specyficzny aspekt zapytania informacyjnego dotyczący Bibliothèque Nationale, lecz nie zawarto terminów, które dotyczyłyby zbiorów w wersji papierowej (drukowanej, ale nie tylko, np. rękopisów). Jednak wybór terminów odzwierciedlających ten aspekt zapytania informacyjnego jest problematyczny. Zbiory, których dotyczyć ma wyszukiwanie, to książki, czasopisma, kartografia, ryciny, muzykalia, rękopisy i inne. Istnieje wiele różnych form takich zbiorów, jak i wiele terminów, którymi się je określa. Ponadto istnieje duże prawdopodobieństwo, że artykuły dotyczące zbiorów w formie innej niż papierowa, będą zawierały również informacje na temat tradycyjnych zbiorów, np. artykuł o dygitalizacji zbiorów kartograficznych, może zawierać informacje o przechowywaniu czy udostępnianiu map.

Tak więc, z uwagi na to, iż wynik otrzymany na tym etapie wyszukiwania jest stosunkowo niewielki, a precyzyjne sformułowanie klasy dotyczącej zbiorów papierowych trudne, zdecydowano, że ostatnia faseta, która miała zawierać terminy identyfikujące rodzaje zbiorów na tradycyjnych nośnikach papierowych, nie zostanie wcale wprowadzona do wyszukiwania. Wszystkie uzyskane rekordy zostaną przeanalizowane w poszukiwaniu relewantnych informacji. Trudno jednak jednoznacznie ocenić ich relewancję. W zapytaniu nie został sprecyzowany żaden konkretny aspekt, w którym przedstawione miałyby być informacje dotyczące zbiorów

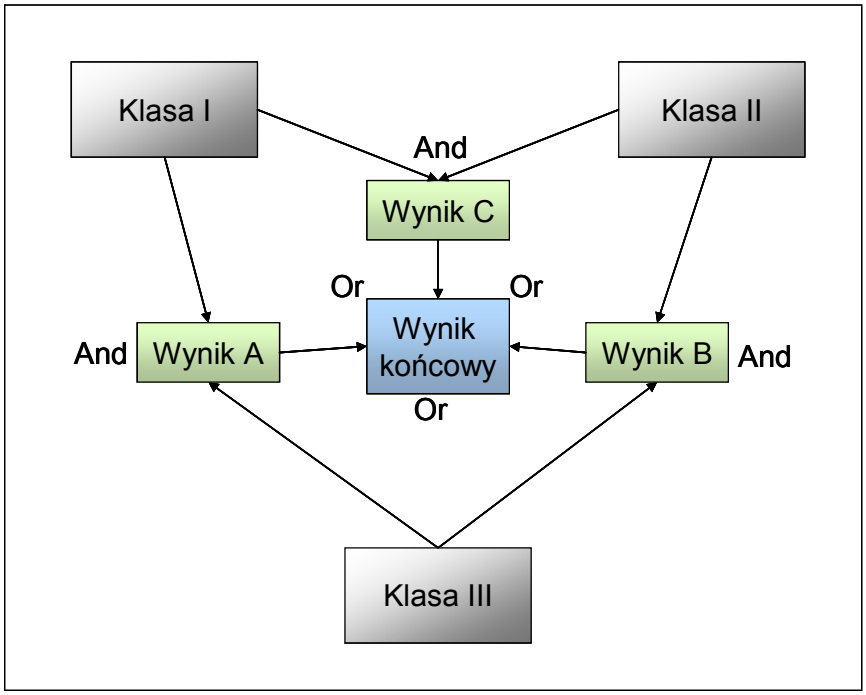
tradycyjnych, tak więc za relewantne można uznać, np. artykuły omawiające postęp we wprowadzaniu opisów bibliograficznych książek do OPAC-a, artykuły dotyczące architektury i budynków biblioteki, które mogą zawierać informacje o sposobach przechowywania i rozmieszczenia zbiorów, itp. Z dużym prawdopodobieństwem jako nierelewantne odrzucić można cztery artykuły dotyczące: współpracy przy katalogowaniu zasobów; kulturalnej działalność biblioteki, archiwizowania zawartości Internetu oraz organizację strony internetowej Biblioteki. Dokonanie oceny pozostałych dwudziestu czterech rekordów wymaga zapoznania się z pełnymi tekstami artykułów.

Strategia kolejnych klas jest użytecznym podejściem, które najczęściej stosowane jest w dwóch sytuacjach. Można posługiwać się nią wtedy, gdy temat wyszukiwania jest na tyle specyficzny, iż wprowadzenie klas odzwierciedlających wszystkie jego aspekty mogłoby skutkować zbyt małym zbiorem wyników. Umiejętne formułowanie i wybór kolejnych klas pozwoli otrzymać wynik o zadowalających współczynnikach zarówno dokładności jak i kompletności. Druga sytuacja, w której ta strategia jest niezwykle przydatna ma miejsce, gdy wyszukujący nie jest pewien jak sformułować jedną z klas wyszukiwania. Powodem tego może być nieznajomość wyrażeń JIW używanego w bazie, brak wyrażeń reprezentujących pojęcie istotne dla zapytania, bądź też specyfika danego aspektu wyszukiwania, który trudno jest przedstawić precyzyjnie. Zapoznanie się z opisami rekordów, które zostały otrzymane w wyniku wcześniej wprowadzonych instrukcji, może pomóc w odpowiednim sformułowaniu kolejnej klasy lub wykazać, że taka klasa jest zbędna. Strategia ta wymaga interakcji z systemem i poświęcenia więcej czasu niż w przypadku stosowania strategii formowania klas.

STRATEGIA PODWÓJNYCH KLAS

Strategia podwójnych klas (ang. *pairwise facet strategy*) jest kolejną strategią opierającą się na metodzie tworzenia klas dla

różnych aspektów problemów wyszukiwawczych. Jest jedną z rzadziej stosowanych strategii. Używana jest w sytuacji, kiedy poszukiwane są zagadnienia bardzo specyficzne, a wyszukujący jest przekonany, że zbiór wyników dla instrukcji odzwierciedlającej wszystkie aspekty zapytania informacyjnego będzie zbyt mały. Specyfika wyszukiwania nie pozwala jednak, aby, jak to było czynione przy użyciu strategii kolejnych klas, pominąć jeden z aspektów tematyki wyszukiwania.



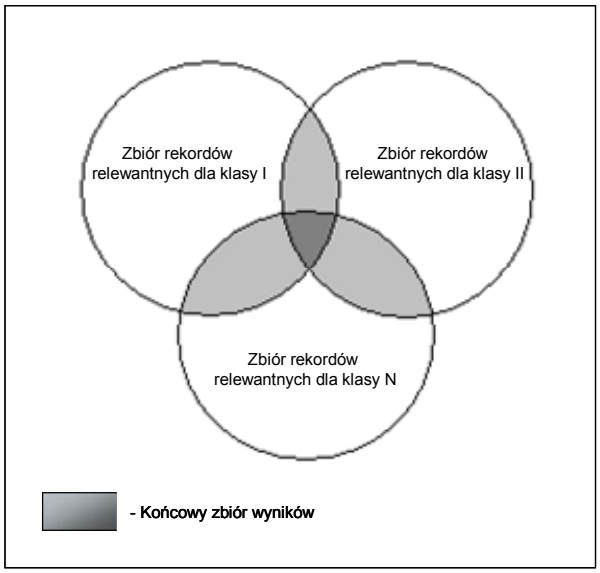
Tablica 3: Strategia podwójnych klas.

Tworzenie instrukcji dla wyszukiwania z użyciem takiej strategii odbywa się w następujący sposób: dla wyszukiwania, które można przedstawić za pomocą trzech klas, pierwsza część instrukcji zawiera klasę pierwszą i drugą połączone operatorem AND; kolejny człon instrukcji to iloczyn klasy drugiej i trzeciej, a następny trzeciej i pierwszej. Jeśli wyszukiwanie zawiera więcej klas dalsza metoda postępowania jest analogiczna, aż do momentu uzyskania instrukcji uwzględniającej połączenie iloczynu każdych dwóch klas. W ten sposób otrzymuje się trzy (lub więcej) zbiory wyników cząstkowych. Suma tych zbiorów stanowi końcowy zbiór wyników wyszukiwania.

Proces tworzenia strategii podwójnych klas ilustruje tablica 3, można też przedstawić go za pomocą formuł rachunku zbiorów:

(Klasa I ∩ Klasa II) ∪ (Klasa II ∩ Klasa III) ∪ (Klasa III ∩ Klasa I).

Wynik końcowy (zob. tablica 4 – kolor szary) zawiera również podzbiór, który stanowi iloczyn wszystkich wyników cząstkowych (na rysunku oznaczony kolorem ciemnoszarym), jednak dla sytuacji, w których uzasadnione jest używanie tej strategii może on być zbiorem pustym. Jeśli tak nie jest, systemy wyszukiwawcze stosowane współcześnie w bazach danych powinny wyeliminować powtarzające się rekordy.



Tablica 4: Zbiór wyników strategii podwójnych klas.

Wyszukiwanie przy użyciu strategii podwójnych klas można realizować na dwa sposoby. Można zawrzeć w instrukcji wszystkie klasy w sposób opisany powyżej i wprowadzić do systemu za jednym razem. Taki sposób postępowania jest przydatny, gdy użytkownik ponosi koszty związane z czasem korzystania z baz danych. W takiej sytuacji trzeba niezwykle starannie przygotować instrukcję wyszukiwawczą.

Drugi sposób polega na wprowadzaniu do systemu instrukcji cząstkowych, z których każda zawiera iloczyn dwóch klas. Pozwala to na zweryfikowanie poprawności tych klas, zastanowienie się nad

ewentualnym zmodyfikowaniem terminów składających się na poszczególne klasy, itp. Sposób ten wymaga jednak poświęcenia wyszukiwaniu większej ilości czasu.

Jak już zostało to wspomniane, strategia podwójnych klas stosowana jest w przypadku bardzo specyficznych zapytań informacyjnych. Dlatego w tej części pracy przedstawiony zostanie tylko jeden przykład wyszukiwania, wzorowany na przykładzie omówionym w książce Stephena Hartera (zob. Harter, 1986, p. 180-181).

Przykład 4

Wyszukiwanie ma pozwolić na dotarcie do materiałów na temat fizjologicznych reakcji mięśni twarzy pod wpływem bodźców emocjonalnych. Mają to być publikacje z ostatnich pięciu lat, napisane w języku angielskim. Istotne jest dotarcie do jak największej liczby publikacji dotyczących tego tematu.

Problem wyszukiwawczy można przedstawić przy pomocy trzech klas. Pierwsza z nich dotyczy mięśni twarzy, druga reakcji fizjologicznych, a trzecia emocji, czy bodźców emocjonalnych. Każda z klas jest tak samo istotna.

Wyszukiwanie przeprowadzone jest w bazie *MEDLINE*. W celu uzyskania wysokiego współczynnika kompletności przeszukiwane są wszystkie indeksowane pola tekstowe (prefiks TX). Zgodnie z założeniami wyszukiwania wprowadzono ograniczenia daty publikacji (2001-2006) oraz języka (angielski). Wykorzystano także specjalne opcje bazy MEDLINE, aby zawęzić wyszukiwanie do artykułów dotyczących organizmu człowieka (ang. *human*).

Ponadto ustalono dodatkowe ograniczenie ze względu na specyfikę strategii podwójnych klas. Zgodnie z nią, do zbioru wyników wyszukiwania celowo włączane są rekordy, których relewancja w stosunku do zapytania informacyjnego nie jest oczywista (wyniki cząstkowych wyszukiwań nie odzwierciedlają wszystkich aspektów problemu wyszukiwawczego). Dlatego też, aby ułatwić użytkownikowi podjęcie decyzji, czy dana publikacja podejmuje interesujące go zagadnienia, ograniczono wyszukiwanie do rekordów zawierających abstrakt. W bazie takiej jak *MEDLINE* nie powinno to nadmiernie zawęzić poszukiwania.

Klasy wprowadzane do instrukcji wyszukiwawczej zawierają następujące terminy. Klasę dotyczącą mięśni twarzy wyodrębniają dwa wyrażenia: „facial musculature” oraz „facial muscle*”. Wyrażenia klasy reakcji fizjologicznych to: „physiological responding”, „physiological reaction”, „physiological response*”. Natomiast klasa emocji zawiera terminy: „emotion” oraz „emotions” (w tym wypadku nie posłużono się maskowaniem „*” ponieważ włączyłoby to do wyszukiwania takie terminy, jak „emotional”, które na razie nie zostaną uwzględnione). Jeśli tak przeprowadzone wyszukiwanie okaże się niewystarczające, wtedy fasety będzie należało uzupełnić terminami takimi jak „anger”, „fear”, „facial expression”, itd.

Zdecydowano, że wyszukiwanie odbywać się będzie poprzez wprowadzanie do systemu instrukcji cząstkowych.

Instrukcja wprowadzona jako pierwsza brzmi: ("facial muscle*" OR "facial musculature") AND ("physiological responding" OR "physiological reaction" OR "physiological response*"). Zbiór wyników dla tej instrukcji to tylko 3 rekordy. Taki wynik dowodzi słuszności zastosowanej strategii, gdyż iloczyn wszystkich trzech klas wyszukiwania (w przypadku użycia strategii formowania klas) byłby równy temu wynikowi, bądź mniejszy od niego.

Następna wprowadzona instrukcja zawiera następujące terminy: ("Physiological responding" OR "Physiological reaction" OR "Physiological response" OR "Physiological responses") AND (emotion OR emotions). Skutkowała ona znacznie większym niż w poprzednim wypadku wynikiem, gdyż odszukanych zostało 38 rekordów.

Trzecia instrukcja uwzgędnia następujące połączenie terminów: ("facial muscle*" OR "facial musculature") AND (emotion OR emotions). System odnalazł 41 rekordów.

Kolejnym etapem jest połączenie wszystkich trzech instrukcji za pomocą operatora „OR”. Większość współczesnych baz danych ułatwia wykonanie tej operacji przy pomocy historii wyszukiwania, tak więc nie ma potrzeby ręcznego wpisywana instrukcji do systemu. Wynik wyszukiwania wynosi 78 rekordów. Liczba ta nie jest równa sumie liczby rekordów odnalezionych w wyszukiwaniach cząstkowych (82 rekordy), gdyż system pominął rekordy powtarzające się.

Tak jak to już zostało powiedziane, w przypadku zastosowania tego rodzaju strategii ocena relewancji rekordów nie jest kwestią

jednoznaczną. Wynik zawiera wiele rekordów, w przypadku których istnieje duże prawdopodobieństwo, że okażą się relewantne. Informacje zawarte w abstraktach pomogą określić stopień tego prawdopodobieństwa, lecz ostateczna ocena może zostać wydana dopiero po zapoznaniu się z pełną treścią wyszukanych materiałów. Jednak analizując otrzymane wyniki można stwierdzić, że wyszukiwanie zidentyfikowało rekordy, które wydają się odpowiadać zapytaniu informacyjnemu, a które nie znalazłby się w zbiorze iloczynu wszystkich klas. W większości przypadków w tych rekordach (w szczególności w ich abstraktach) użyte były terminy, które odczytane w określonym kontekście są synonimami poszukiwanych terminów, jednak w innych kontekstach mogą być inaczej rozumiane (np. wyrażenie „biological reactivity" użyte w abstrakcie jednego z rekordów jest synonimem „physiological responding").

Można zastanawiać się, czy bardziej efektywne nie byłoby jednak przeprowadzenie wyszukiwania za pomocą strategii formowania klas przy użyciu ogólniejszych terminów. W celu sprawdzenia tej hipotezy do jednej z instrukcji cząstkowych dołączono termin bardziej ogólny „facial expression". Instrukcja przybrała następującą postać: (("facial expression" OR "facial muscle*" OR "facial musculature") AND (emotion OR emotions)).

Wynik wyszukiwania to aż 623 rekordy, podczas gdy instrukcja, która nie zawiera terminu „facial expression" pozwoliła na zidentyfikowanie jedynie 41 rekordów. Tak liczny zbiór wyników mógłby stanowić argument za użyciem strategii formowania klas. Jednak pobieżne zapoznanie się z próbką wyników tego wyszukiwania pozwoliło stwierdzić, że terminem „facial expression" opisanych zostało wiele rekordów dotyczących takich zagadnień, jak: wywołane emocjami zmiany ukrwienia skóry twarzy, zmiany rozmiarów źrenicy, itd. Te zagadnienia nie wchodzą w zakres problematyki wyszukiwania. Duża różnica między wynikami omawianych wyszukiwań wydaje się wskazywać na fakt, że termin „facial expression" jest znacznie ogólniejszy niż terminy poprzednio użyte. W związku z tym nie należy

oczekiwać, że strategia formowania klas przy użyciu tego terminu zapewni wyższy współczynnik dokładności niż cechujący przeprowadzone powyżej wyszukiwanie.

Powyższy przykład ilustruje przydatność strategii podwójnych klas. Podsumowując, możemy wyróżnić następujące wyznaczniki sytuacji, kiedy jej zastosowanie może przynieść dobre efekty: 1. Wszystkie aspekty wyszukiwanego zagadnienia są tak samo istotne i nie można sobie pozwolić na pominięcie żadnego z nich; 2. Użytkownik w równym stopniu kładzie nacisk na precyzję i kompletność wyszukiwania – nie można sobie pozwolić na takie sformułowanie instrukcji, które będzie się charakteryzować dobrym poziomem jednego ze współczynników, lecz, zgodnie z zachodzącą między nimi korelacją, zbytnio obniży drugi z nich; 3. Iloczyn klas reprezentujących to zagadnienie zawiera zbyt małą liczbę rekordów; 4. Terminy ogólniejsze od użytych, stosowane są do indeksowania rekordów dotyczących wielu innych zagadnień, nie należących do zakresu wyszukiwania. Użycie ich skutkowałoby zbyt licznym zbiorem rekordów.

Użytkownik rozpoczynający wyszukiwanie na dany temat zazwyczaj nie jest świadomy, czy jego problemem wyszukiwawczy odpowiada podanej powyżej charakterystyce, ponieważ ustalenie tego wymaga interakcji z systemem. Dlatego też strategia podwójnych klas rzadko wykorzystywana jest przy pierwszym podejściu do wyszukiwania. Jej użycie jest zazwyczaj skutkiem niepowodzenia innych strategii.

STRATEGIA POMNAŻANIA ODWOŁAŃ

Strategia pomnażania odwołań (ang. *citation pearl growing strategy*) znacznie różni się od opisanych wcześniej strategii. Stosowanie tej strategii wymaga, aby użytkownik znał jeden lub kilka dokumentów relewantnych w stosunku do zapytania informacyjnego. Dalsze postępowanie polega na zidentyfikowaniu przy pomocy wyszukiwania prostego rekordów znanych użytkownikowi pozycji, a

następnie zapoznaniu się z terminami użytymi do ich indeksowania. Kolejnym krokiem jest budowa klas wyszukiwania na podstawie zidentyfikowanych terminów, tak aby wśród wyszukanych znalazły się również znane wcześniej publikacje. Po wprowadzeniu takiej instrukcji do systemu, użytkownik zapoznaje się z wyszukanymi rekordami. Jeśli kompletność wyszukiwania nie jest satysfakcjonująca, wybiera kolejne rekordy, które ocenił jako relewantne i na podstawie ich opisów uzupełnia instrukcję wyszukiwawczą. Ten proces może być powtarzany wielokrotnie, aż do momentu, kiedy kompletność wyszukiwania osiągnie wymagany poziom.

Przykład 5.1

Poszukiwane są materiały, które zawierają przewidywania na temat przyszłego rozwoju Internetu i jego wpływu na życie codzienne. Znany jest jeden artykuł, który porusza te zagadnienia:

Minsky, Marvin; Ferren, Bran; Mountford, Joy; Gemperle, et al. (2000). *The future of the Internet. Hang tough, surfers: here come adventures unimagined, dangers undreamed of, and a towering wave of chaos to test your nerve*, "Discover" vol. 21, iss. 11, p. 55-59.

Temat wyszukiwania łączy w sobie zarówno zagadnienia nauk ścisłych (informatyka), jak i problematykę socjologiczną. Z tego powodu wyszukiwanie przeprowadzone zostanie symultanicznie w kilku bazach. Wybrane zostały bazy konsorcjum *Proquest*: *ABI/INFORM Global, ProQuest Computing, ProQuest Science Journals, ProQuest Social Science Journals, ProQuest Telecommunications*. Pierwsza z tych baz nie należy do żadnej z wymienionych dziedzin, jej profil dotyczy zagadnień gospodarczych. Została jednak włączona do grupy przeszukiwanych baz, ponieważ indeksuje artykuły z prasy codziennej różnych krajów, nie tylko z zakresu gospodarki. Jest prawdopodobne, że poszukiwana tematyka była podjęta także w tego typu piśmiennictwie.

Wyszukiwanie ma być przeprowadzone w kilku bazach o diametralnie różnej specyfice. Każda z tych baz ma różne indeksy dostosowane do charakterystyki zagadnień, które uwzględnia. Wyszukiwanie w polach opisu bibliograficznego, abstraktu i pełnego tekstu skutkowałoby zbyt dużym zbiorem wyników, gdyż terminy, za pomocą których można przedstawić tematykę wyszukiwań (np. „life conditions", „Internet", itd.) są powszechnie używane w różnych kontekstach. Dlatego też zdecydowano się zastosować strategię pomnażania odwołań.

Przy użyciu wyszukiwania prostego udało się odnaleźć rekord znanego wcześniej artykułu. Stwierdzono, że został on opisany trzema terminami: „Internet", „Forecasts" oraz „Social aspects". Te trzy wyrażania połączone operatorem „AND" zawarte zostały w instrukcji wyszukiwawczej wprowadzonej do systemu (użyto prefiksu „SUB" oznaczającego *Subject term*).

Wynik wyszukiwana to 12 rekordów, wszystkie dotyczą zadanej tematyki. Jest to niezbyt liczny zbiór, dlatego też możliwe było zapoznanie się z opisami wszystkich wyszukanych materiałów. Na tej podstawie zdecydowano się w następujący sposób rozszerzyć instrukcję wyszukiwawczą: „SUB(Internet or "World Wide Web") AND SUB(Forecasts or Future or "Technological change") AND SUB("Social aspects" or Lifestyles)". Są to terminy, które uznano za najbardziej adekwatne do tematyki wyszukiwania. W zależności od rezultatów można będzie dołączyć terminy bardziej ogólne (np „information technology") lub bardziej szczegółowe (np „smart houses").

Wynik wyszukiwania to 22 rekordy. Ponieważ poszukiwane dokumenty nie dotyczą faktów, lecz przewidywań co do rozwoju wszechstronnego narzędzia, jakim jest Internet, tematyka odnalezionych rekordów jest bardzo różnorodna – od ataków terrorystycznych za pośrednictwem Internetu, przyszłości przemysłu czasopiśmienniczego, po „inteligentne" domy czy kuchnie. Wśród wyszukanych rekordów 4 pozycje to recenzje książek dotyczących zadanej tematyki. Nie ma jednak powodu, aby traktować je jako nierelewantne, gdyż choć same prawdopodobnie nie zawierają poszukiwanych informacji, to pozwalają na zidentyfikowanie obszernych publikacji z tego zakresu. Jako nierelewantne można ocenić dwa rekordy. Są to artykuły zawierające krótkoterminowe przewidywania na temat rozwoju telefonii komórkowej z wykorzystaniem Internetu, lecz opublikowane były w latach 1997, 2000 i dotyczą minionego już okresu. Można by się starać wykluczyć podobne publikacje poprzez zawężenie wyszukiwania do publikacji z ostatnich kilku lat, lecz spowodowałoby to pominięcie interesujących materiałów z prognozami długoterminowymi.

Wysoka dokładność wyszukiwania pozwala sądzić, iż w bazie pozostały jeszcze relewantne rekordy (zgodnie z korelacją między współczynnikiem kompletności a dokładności). Kontynuacja wyszukiwania według zaprezentowanego schematu zależy od stopnia satysfakcji użytkownika dotychczasowymi wynikami. Do rozbudowania instrukcji wyszukiwania można użyć zidentyfikowane, lecz pominięte wcześniej, terminy wyszukiwawcze, jak i przeanalizować wyniki uzyskane w czasie ostatniej operacji. Jeśli

pozwalają na to możliwości systemu, warto zapisywać w profilu użytkownika rekordy wyszukane w wyniku każdej operacji – umożliwi to łatwe zidentyfikowanie rekordów, których nie uwzględniały rezultaty poprzedniego wyszukiwania. Omówienie powyżej przykładowego wyszukiwania zostanie zakończone na tym etapie, ponieważ przeprowadzone już operacje wystarczająco dobrze ilustrują schemat działania według strategii pomnażania odwołań.

Przykład 5.2

Niniejsza część nie zawiera przykładu podobnego do przedstawionych wcześniej, czyli ilustrujących problemy wyszukiwawcze, dla których optymalną metodą postępowania była prezentowana strategia. Poniżej przedstawiona jest sytuacja, w której pozornie słusznym wyborem byłby wybór strategii pomnażania odwołań, lecz - jak wykaże wyszukiwanie - tym razem okazuje się ona całkowicie nieprzydatna.

Poszukiwane są publikacje zajmujące się pojęciem relewancji w informacji naukowej. Szczególnie przydatne byłyby artykuły zawierające przegląd piśmiennictwa na ten temat oraz publikacje traktujące temat bardzo szeroko, odnosząc się również do innych dziedzin wiedzy. Znany jest jeden artykuł tego rodzaju:

Saracevic, Tefko (1975): *Relevance: a Review of and a Framework for the Thinking on the Notion in Information Science*, "Journal of the American Society for Information Science", vol. 26, iss. 6, p. 321-343.

Artykuł ten dotyczy pojęcia relewancji w informacji naukowej, a wyszukiwanie powinno obejmować różne dziedziny nauk. Dlatego też wyszukiwanie będzie prowadzone w kilku bazach jednocześnie: specjalistycznej bazie informacji naukowej *Library, Information Science & Technology Abstracts* oraz kilku innych, zawierających piśmiennictwo naukowe z różnych dziedzin – *Academic Search Premier, MasterFILE Premier, SocINDEX with Full Text.* Wybór tych właśnie baz był również podyktowany faktem, że są one udostępniane przez konsorcjum *EBSCOHost,* a co za tym idzie można przeszukiwać je symultanicznie przy użyciu jednego interfejsu. Przeszukując jednocześnie kilka baz nie można oprzeć się na jednym tezaurusie, dlatego też zdecydowano się zastosować strategię pomnażania odwołań.

Wyszukiwanie proste pozwoliło odnaleźć artykuł Tefko Saracevica. Został on zaindeksowany przy pomocy następujących terminów: “classification”, “communication in science”, “documentation”, “information science”, “logic”, “knowledge management”.

Terminy użyte w opisie tego artykułu zostały połączone operatorem „AND” i wprowadzone do systemu (przeszukiwany jest indeks *Subject terms)*. Wynikiem wyszukiwania był tylko jeden rekord, zawierający opis znanego już artykułu T. Saracevica. Wyszukiwanie ma dotyczyć pojęcia relewancji w różnych naukach, dlatego z instrukcji usunięto termin „information science”. Niestety nie zmieniło to wyniku wyszukiwania. Zastosowano rozmaite połączenia zidentyfikowanych terminów, które mogłyby odzwierciedlać kontekst poszukiwanych zagadnień. Rezultat inny od uzyskanego początkowo przyniosło wyszukiwanie za pomocą instrukcji: „SU (classification AND "communication in science")” – 5 rekordów, SU ("communication in science" AND documentation) – 8 rekordów, SU (documentation AND "information science" AND logic) – 3 rekordy, SU ("communication in science" AND "knowledge management") – 5 rekordów, SU ("documentation" AND "information science") – 757 rekordów. Analiza wyników tych wyszukiwań, poza ostatnim, wykazała, że nie został odnaleziony żaden relewantny rekord poza znanym już artykułem. Wynik ostatniego wyszukiwania jest tak duży, że jego analiza wymagałaby poświęcenia nieracjonalnie dużej ilości czasu.

Baza, w której znaleziono pierwszy rekord jest bazą specjalistyczną, tak więc wyszukujący ma prawo oczekiwać, że rekordy są tu odpowiednio zaindeksowane. Jednak w tym wypadku najwyraźniej tak się nie stało. Nie oznacza to, że termin „relewancja” nie jest stosowany w tym systemie. Przeprowadzono wyszukiwanie proste wprowadzając do systemu termin „relevance” w polu *Subject terms* (nie uwzględniony w opisie artykułu Saracevica). Wynik wyszukiwania to 419 rekordów. Pobieżne zapoznanie się z odnalezionymi artykułami pozwala stwierdzić, że wiele z nich jest relewantnych. W tym konkretnym wypadku, choć przed przystąpieniem do wyszukiwania nie można było tego stwierdzić, nawet wyszukiwanie proste przynosi lepsze rezultaty niż strategia pomnażania odwołań.

Powyższe przykłady przedstawiają problemy związane ze stosowaniem strategii pomnażania odwołań. Strategia ta, w przeciwieństwie do strategii omówionych wcześniej, polega na

rozpoczęciu wyszukiwania od zidentyfikowania zbioru wyników o wysokim współczynniku dokładności, a niskiej kompletności, aby następnie rozszerzać poszukiwania do momentu, kiedy liczba odnalezionych relewantnych rekordów zadowoli użytkownika. Taka strategia jest użyteczna przede wszystkim wtedy, kiedy baza nie posiada tezaurusa dostępnego dla użytkowników, bądź kiedy przeszukiwanie prowadzone jest symultanicznie w kilku bazach. Przydatna jest również w sytuacji, kiedy wyszukujący nie jest obeznany ze słownictwem używanym w tezaurusie, a problem wyszukiwawczy można przedstawić za pomocą wielu niejednoznacznych terminów, tak jak było to w przykładzie 5.1. Należałoby bardzo dokładnie zapoznać się z tezaurusem (bądź kilkoma w przypadku wyszukiwania symultanicznego), aby zidentyfikować deskryptory, czy hasła, które są adekwatne do tematu wyszukiwania. Ponadto stosując tę metodę można poznać specjalistyczne terminy których wyszukujący wcześniej nie znał, a które pozwolą na odpowiednie rozszerzenie zbioru wyników.

Strategię pomnażania odwołań można stosować także jako swoistego rodzaju „drogę na skróty". Jeśli wyszukujący zna publikację odpowiadającą tematyce, może na podstawie jej opisu szybko odnaleźć kilka relewantnych pozycji. Jednak formułując instrukcję jedynie na podstawie opisów zawartych w odnalezionych rekordach należy mieć świadomość niebezpieczeństwa łatwego pominięcia relewantnych pozycji.

Specyfiką strategii pomnażania odwołań jest to, że jej powodzenie w większym stopniu niż to było we wcześniej opisanych strategiach zależy od precyzji i konsekwencji zaindeksowania dokumentów znanych użytkownikowi. Jednak nawet wyszukiwanie w bazie o uznanej renomie nie gwarantuje, że stopień szczegółowości każdego z opisów będzie miał poziom satysfakcjonujący dla użytkownika. Dlatego też ta strategia nie jest zalecana dla osób o małym doświadczeniu w wyszukiwaniu. Użytkownik o niewielkiej wiedzy na temat wyszukiwanego zagadnienia, języka informacyjno-

wyszukiwawczego systemu, a przede wszystkim podchodzący bezkrytycznie do zidentyfikowanych opisów, może stracić dużo czasu na przeprowadzenie wyszukiwania dającego bardzo mierne rezultaty.

STRATEGIA INDEKSÓW CYTOWAŃ

Zupełnie odmienna od wcześniej opisanych jest strategia indeksów cytowań (ang. *citation indexing strategy*). Jej specyfika nie wyraża się w sposobie konstruowania instrukcji, lecz polega na wykorzystaniu możliwości śledzenia cytowań, dostępnej w niektórych bazach. Podstawą tego sposobu wyszukiwania publikacji jest założenie, że istnieje semantyczny związek między pracą cytowaną a cytującą.

W piśmiennictwie przedmiotu można spotkać się z wyróżnianiem trzech strategii cytowań: cytowania publikacji (ang. *cited publication*), cytowania autora (ang. *cited author*) oraz współcytowania autorów (ang. *cocited authors*). Wyszukiwanie prowadzone według każdego z tych schematów charakteryzuje się własną specyfiką. Jednak różnice między nimi porównywalne są do różnic między wyszukiwaniem przy użyciu różnych indeksów w obrębie jednej strategii (np. przy użyciu słownictwa kontrolowanego lub słów kluczowych w tytule). Dlatego też w niniejszym artykule te trzy strategie zostaną uznane za odmiany jednej strategii i przedstawione w kolejnych przykładach.

Przykład 6.1

Poszukiwane są publikacje, które dotyczą metod wyszukiwania informacji – strategii, taktyk, czy inaczej zdefiniowanych przez autorów sposobów odnajdywania potrzebnych informacji. Ta tematyka została omówiona w artykule: Bates, Marcia (1979). Information Search Tactics, "Journal of the American Society for Information Science", vol. 30, iss. 4, p. 205-215. Jest to na tyle istotna i znana publikacja, że można założyć, iż w wielu pracach, w których podjęto te zagadnienia powinna być przywołana.

Do prześledzenia cytowań odnoszących się do tej publikacji wybrano *Science Citation Index Expanded,* bazę wielodziedzinową, której producentem jest *Thomson Scientific* (poprzednia nazwa: *Institute for Scientific*

Information[3]). Rekord dotyczący danego artykułu został zidentyfikowany na podstawie nazwiska autorki, daty publikacji oraz numeru stron, na których się znajduje. Kwerenda dała rezultat w postaci 39 publikacji, w których wskazany artykuł był cytowany.

Wszystkie odnalezione rekordy to publikacje z zakresu informacji naukowej, dotyczące wyszukiwania informacji. Bez zapoznania się z pełnymi tekstami artykułów nie można jednoznacznie określić, które z nich zawierają opisy efektywnych metod wyszukiwania informacji. Kilka z wyszukanych pozycji dotyczy korzystania lub projektowania narzędzi ułatwiających wyszukiwanie informacji, lecz w tego typu publikacjach również mogą pojawić się zagadnienia będące tematem wyszukiwania.

Przykład 6.2

Poszukiwane są materiały na temat badań bibliometrycznych, a w szczególności - rozproszenia artykułów specjalistycznych w czasopiśmiennictwie z różnych dziedzin. Autor, który jako pierwszy zajął się badaniami tego zjawiska to Samuel C. Bradford. Większość jego prac dotyczyła tej właśnie tematyki, dlatego wyszukiwane będą publikacje, w których zacytowano którykolwiek z jego artykułów.

Wyszukiwanie przeprowadzone zostało w bazie *Library, Information Science & Technology Abstracts,* w której jednym z dostępnych narzędzi jest indeks cytowań. Trudnością w tego rodzaju wyszukiwaniu jest to, iż wybrana baza (podobnie jak *Science Citation Index Expanded)* nie pozwala na wyszukiwanie w cytowaniach jedynie według nazwiska autora. Po wpisaniu instrukcji „WA Bradford S*" (maskowanie imienia wprowadzono w celu odnalezienia zarówno tych cytowań, w których użyto pełnego imienia Bradforda, jak i tych gdzie zapis uwzględnia jedynie inicjały), wyświetlona została lista wszystkich publikacji tego autora. System wymaga, aby użytkownik oznaczył jakie publikacje danego autora mają zostać wyszukane. W tym wyszukiwaniu wybrane zostały wszystkie wyświetlone przez system pozycje. Rezultatem tak skonstruowanego wyszukiwania jest 20 rekordów.

Większość wyszukanych artykułów dotyczy badań bibliometrycznych. Wyjątkiem jest rekord, który omawia historię organizacji *ASLIB* w latach

[3] Inne produkty tej znanej instytucji to najobszerniejsze bazy cytowań literatury naukowej: *Science Citation Index, Arts and Humanites Citation Index* oraz *Social Science Citation Index*. Niestety w trakcie powyższych badań nie uzyskano do nich dostępu.

1924-1950, a więc w czasie kiedy publikował S. C. Bradford. Trzy z odnalezionych rekordów dotyczą zagadnień badania piśmiennictwa dostępnego w Internecie (ang. *webometrics*), jest to jednak tematyka bardzo blisko związana z zagadnieniami bibliometrii.

Niestety, jak pozwoliły ustalić testy, nie wszystkie cytowania występujące w artykułach dostępnych w bazie *LISTA* zostały wprowadzone do systemu. Tak więc należy sądzić, że przy wyszukiwaniu w tej bazie za pomocą strategii cytowań wiele artykułów relewantnych mogło zostać pominiętych. Dla porównania można dodać, że analogiczne wyszukiwanie w bazie *Science Citation Index Expanded* pozwoliło zidentyfikować 73 rekordy.

Przykład 6.3

Poszukiwane są materiały na temat pszczół – zarówno materiały dotyczące ich zachowań społecznych, jak anatomii, żywienia i innych zagadnień. Znane są nazwiska dwojga naukowców, którzy specjalizują się w tym temacie. Są to Claudia Dreller, oraz Wolfgang H. Kirchner. Wyszukanie przeprowadzone będzie przy pomocy strategii cytowań, a dla zwiększenia dokładności, będą wyszukiwane publikacje, w których przytoczono prace obydwojga autorów.

Wyszukiwanie przeprowadzane jest w wielodziedzinowej bazie *Scopus*. Baza posiada wiele użytecznych narzędzi wyszukiwania, m.in. indeks cytowań. Wyszukiwanie w indeksowanych cytowaniach można prowadzić podając jedynie nazwisko autora – nie ma potrzeby uściślania jakie publikacje są poszukiwane. W instrukcji podano tylko nazwiska autorów (pominięto imiona, czy ich inicjały), licząc na to, że zestawienie tych dwóch nazwisk pojawi się jedynie w relewantnych pracach. Wynik wyszukiwania to 23 rekordy.

Wśród wyszukanych rekordów jedynie dwa są nierelewantne. Oba opisują zagadnienia związane z owadami, lecz innymi niż pszczoły – jeden z nich dotyczy świerszczy, a drugi pluskwiaków.

Analiza słów kluczowych użytych w opisie znalezionych rekordów pozwoliła przekonać się o użyteczności zastosowanej strategii. Do opisu rekordów użyto wielu różnych terminów, m.in.: „honey bee", „honeybee", „Apis mellifera", „stingless bees", „Apidae: Meliponini", „apis". Wyszukiwanie za pomocą słów kluczowych powinno uwzględnić każdy z tych terminów oraz warianty ich pisowni, a nie wszystkie terminy mogą być oczywiste dla

użytkownika w momencie rozpoczynania wyszukiwania. Ponadto niektóre rekordy dotyczące pszczół nie zostały zaindeksowane żadnym terminem opisującym ten gatunek, a co za tym idzie nie zostałyby zidentyfikowane przy wyszukiwaniu przez słowa kluczowe.

Przedstawione powyżej przykłady pokazują trzy sposoby użycia strategii indeksów cytowań. Najczęściej w tego rodzaju wyszukiwaniu śledzi się cytowania konkretnej pracy, której znaczenie jest fundamentalne dla danych zagadnień. Strategia wyszukiwania cytowań dla danej pracy zazwyczaj stosowana jest w przypadku stosunkowo wąskiej tematyki wyszukiwań (indeksy cytowań nie uwzględniają wydawnictw zwartych, a tematyka artykułów przeważnie dotyczy szczegółowych problemów z danych dyscyplin badawczych). Nie jest to oczywiście regułą, gdyż artykuł może na przykład zawierać przegląd piśmiennictwa z danej dyscypliny.

Ustalanie cytowań całości piśmiennictwa danego autora nie jest powszechnie stosowaną metodą. Wynika to z faktu, że autorzy, nawet specjalizujący się w konkretnych zagadnieniach badawczych, rzadko ograniczają swoje zainteresowania i publikacje do jednej tylko problematyki. Można liczyć, iż autor taki jak na przykład Tefko Saracevic we wszystkich swoich publikacjach zajmował się kwestiami z zakresu informacji naukowej, jednak nie zawsze będą to prace dotyczące zagadnienia relewancji, które najczęściej wiąże się z jego nazwiskiem. Dlatego też zazwyczaj wyszukiwanie według cytowań wszelkich prac autora stosuje się w wypadku, kiedy dany twórca opublikował niewiele prac, bądź też kiedy temat wyszukiwania obejmuje szeroki zakres zagadnień.

Ustalanie cytowań dla kilku autorów pozwala na odszukanie rekordów dla tematyki bardziej szczegółowo zdefiniowanej niż w przypadku śledzenia cytowań według nazwiska jednego autora, a bardziej ogólnej niż przy identyfikowaniu publikacji cytujących dany artykuł. Prawdopodobieństwo, iż autor konkretnej pracy odwoła się do publikacji kilku badaczy specjalizujących się w jednym obszarze

badawczym, a nie będzie ona dotyczyć tematu ich specjalizacji, jest zdecydowanie mniejsze.

Zaletą strategii cytowań jest to, że pozwala na odnajdywanie relewantnych rekordów bez konieczności identyfikowania leksykalnych elementów (deskryptorów, haseł przedmiotowych, słów użytych w tytule, czy innych) związanych z opisem poszukiwanych materiałów. Jeśli kwerenda przeprowadzana jest w bazie, w której śledzenie cytowań nie jest jedyną metodą wyszukiwania, wtedy ta strategia może też służyć jako punkt wyjścia do wyszukiwania przy użyciu którejś ze strategii klas.

Najbardziej znane bazy cytowań bibliograficznych, takie jak *Science Citation Index, Arts and Humanites Citation Index, Social Science Index,* a także *Scopus,* to bazy, których profil obejmuje różne dyscypliny z zakresu nauk ścisłych, humanistycznych, czy społecznych. Przeprowadzone w nich wyszukiwanie może więc łączyć w sobie zagadnienia z pogranicza kilku dziedzin. Wyszukiwanie przeprowadzone w tych bazach pozwala ustalić nie tylko w jakich pracach przywołano dany artykuł, ale też zidentyfikować, bez sięgania do pełnego tekstu danej pracy, odwołania bibliograficzne w niej zawarte. Tak więc wyszukiwanie pozwala na zapoznanie się z piśmiennictwem przedmiotu poprzedzającym jak i następującym po danej pracy.

Powszechnie przyjęte jest, iż w pracach naukowych należy uwzględnić publikacje, w których wcześniej podejmowano zagadnienia, stanowiące przedmiot danej pracy. Staranny przegląd piśmiennictwa przedmiotu uważa się nawet za jeden z wyznaczników jakości prac naukowych. Wyszukiwanie, które polega na zidentyfikowaniu publikacji odwołujących się do istniejącego już piśmiennictwa, w dużym stopniu zawęża krąg wyszukiwania do takich właśnie prac.

Warunkiem korzystania z tej strategii wyszukiwania jest znajomość prac lub autorów, których twórczość ma duże znaczenie dla

poszukiwanych zagadnień. A zatem, wymaga ona uprzedniej dobrej znajomości piśmiennictwa danej dziedziny.

ZAKOŃCZENIE

W literaturze przedmiotu można spotkać się z omówieniami różnych strategii wyszukiwania, wyróżnianych ze względu na różne etapy, lub aspekty procesu wyszukiwania. Jedno z najkompletniejszych opracowań tej tematyki zawarte jest w książce Stephana Hartera *Online Information Retrieval* (zob. Harter, 1986, p. 170 -204). Powyżej przeanalizowano sześć strategii wyszukiwania, które uznano za najbardziej użyteczne przy kwerendach przeprowadzanych we współczesnych systemach informacyjno-wyszukiwawczych dostępnych online. W przypadku każdej z omówionych strategii można wskazać sytuacje i problemy wyszukiwawcze, dla których zastosowanie danej strategii wydaje się szczególnie uzasadnione. Wybór konkretnej strategii wyszukiwania podyktowany jest nie tylko specyfiką zapytania informacyjnego, ale również możliwościami dostępu do baz danych, jakimi dysponuje wyszukujący, narzędziami wyszukiwawczymi dostępnymi w danych systemach, celem, w jakim wykorzystane mają być wyszukane rekordy oraz wieloma innymi czynnikami. Duża liczba zmiennych, które należy uwzględnić obierając daną strategię, powoduje, że nie można arbitralnie określić reguł wyboru strategii, które znajdowałyby zastosowanie w przypadku każdego wyszukiwania określonego typu. Jest to potwierdzeniem tezy, że proces wyszukiwania ma charakter heurystyczny i nie można skodyfikować jego etapów w sposób umożliwiający odwzorowanie ich za pomocą algorytmów.

BIBLIOGRAFIA

Harter, Stephen (1986). *Online Information Retrieval: concepts, principles, and techniques*, San Diego (et al.): Academic Press, 259 p.

Meadow, Charles T. (1992). *Text information retrieval systems*, San Diego [et al]: Academic Press, 302 p.

Szczepańska, Anna (2006). Strategia, heurystyka i taktyka wyszukiwania informacji. Próba uporządkowania pojęć, „Przegląd Biblioteczny", R. 74, z. 2, s. 165- 187.

Artykuł przygotowywany dla: PRZEGLĄD BIBLIOTECZNY 2007 z. 2
PL ISSN 0033-202X

ANNA SZCZEPAŃSKA
Warsaw University
Institute of Information and Book Studies
e-mail: aszczepanska@uw.edu.pl

THE BASIC INFORMATION RETRIEVAL STRATEGIES AND THEIR USE IN PRACTICE

KEYWORDS: Information retrieval strategies. Briefsearch. Search problems. Building blocks strategy. Successive facet strategy. Pairwise facets strategy. Citation pearl growing strategy. Citation indexing strategy. Search statements. Databases.

ABSTRACT: The author presents six most common information retrieval (IR) strategies used in contemporary IR systems: briefsearch, building blocks strategy, successive facet strategy, pairwise facets strategy, citation pearl growing strategy, and citation indexing strategy. She describes search principles characteristic of each strategy and provides examples of relevant queries made in databases available at Warsaw University y Library. She attempts to show which search strategies provide most appropriate solutions to specific search problems and which of their drawbacks the researchers should be aware of.