# Problem kontekstowego wyszukiwania informacji

Wyszukiwanie kontekstowe jest stosowane w wielu dziedzinach, jednak najczęściej spotykane jest w wyszukiwarkach internetowych, gdzie użytkownik po wpisaniu szukanej frazy otrzymuje wiele wyników w postaci stron internetowych. Wyszukiwarki analizują ogromne zbiory dokumentów, dlatego znalezienie w nich pożądanych informacji staje się niezwykle trudne. Na wyszukiwanie dokumentów ma wpływ przede wszystkim ich reprezentacja, dwuznaczność słów używanych we wpisywanych frazach oraz sposób indeksowania. Z tego względu niektóre wyszukiwarki mają zawężony zakres wyszukiwania, np. tylko do publikacji naukowych.

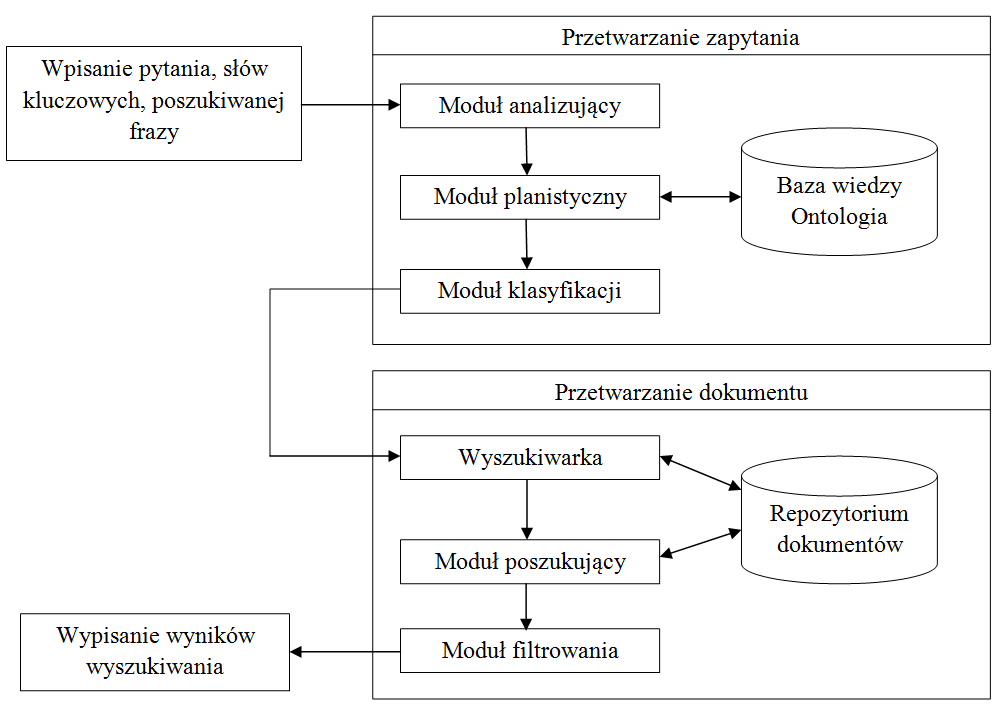
W literaturze przedmiotu można znaleźć wiele publikacji dotyczących metod wyszukiwania stosowanych w wyszukiwarkach internetowych. Jako wyniki wyszukiwania w takich wyszukiwarkach są wypisywane listy całych dokumentów. Najczęściej stosowany model wyszukiwania to tak zwany model Boolowski [24], który umożliwia wyszukiwanie po słowach kluczowych znajdujących się w dokumentach stosując operatory logiczne. Innym znanym modelem jest „Vector Space Model” [26], w którym zapytanie, jak i dokumenty są reprezentowane przez wielowymiarowy wektor. Wektor ten może być wykorzystany do obliczenia podobieństwa między dokumentem a wpisaną frazą. Znany jest także model probabilistyczny [26], którego głównym założeniem jest to, że dla danego zapytania należy nadać większą wagę dokumentom, które zawierają słowa wpisane w kontekście wyszukiwania.

## Systemy typu „pytanie – odpowiedź”

W [27] została opisana architektura systemu „pytanie-odpowiedź”, który może być uznany za typowy system umożliwiający poszukiwanie informacji w dużych zbiorach tekstu w odpowiedzi na pytania zadawane w języku naturalnym. By to umożliwić, przetwarza on zapytania zapisane w języku naturalnym do odpowiedniej postaci rozumianej przez system. Następnie na tej podstawie wyszukuje w swoim repozytorium dokumenty, które stanowią odpowiedź na zadane pytanie. Obie te fazy składają się z szeregu scharakteryzowanych poniżej operacji.

Wspomniany wyżej system rozpoczyna działanie w momencie, gdy użytkownik wpisze wyszukiwaną frazę lub pytanie. Następnie dokonywane są: klasyfikacja podanej frazy i jej przetwarzanie. Przetwarzaniem frazy zajmuje się osobny komponent - analizator, który identyfikuje najważniejsze informacje z danej frazy, np. wyróżnia czasowniki, przyimki, specyfikuje ogólny temat frazy. Po zakończeniu fazy analizy rozpoczyna działanie moduł sprawdzający typ pytania - celem jego pracy jest przedstawienie pytania w ustrukturyzowany, logiczny sposób. Kolejnym angażowanym komponentem jest baza wiedzy wykorzystywana, gdy kontekst wyszukiwania jest podobny do wcześniej wpisanych. Kolejnym ważnym komponentem systemu jest moduł klasyfikacji, określający kategorię odpowiedzi, której użytkownik szuka. Moduł ten bezpośrednio współpracuje z analizatorem, ponieważ kategoria odpowiedzi jest ustalana na podstawie wpisanego przez użytkownika wyrażenia Na tym kończy się proces klasyfikowania pytania zadanego przez użytkownika.

Po przetworzeniu i sklasyfikowaniu pytania, drugim głównym etapem pracy systemu typu „pytanie-odpowiedź” jest przeszukiwanie dokumentów. Moduł ten jest odpowiedzialny za wybranie zestawu dokumentów i wyekstrahowanie odpowiednich fragmentów tekstów w zależności od wpisanego przez użytkownika pytania. W większości systemów (np. przeglądarki internetowe) uzyskiwanymi wynikami są wskazania na konkretne dokumenty, bez identyfikacji poszczególnych zdań, czy obszarów w tych dokumentach. Jednym z modułów bardzo istotnych w tym etapie jest wyszukiwarka - jest to bowiem moduł, który bezpośrednio współpracuje z repozytorium dokumentów oraz z modułem klasyfikacji. Kolejny moduł określony w [27], za pomocą określonych zasad przekształceń tworzy nowe pytania, usuwa znaki interpunkcyjne z wpisanego wyrażenia oraz wydobywa istotne dokumenty z repozytorium tekstów. Kolejny moduł wybiera i filtruje dokumenty odpowiadające na pytania postawione przez użytkownika. Wydobywane są najlepsze odpowiedzi i na koniec wyświetlane są użytkownikowi. Rysunek 2.1. przedstawia schemat całego procesu.



Rysunek .. Architektura systemu typu „pytanie – odpowiedź”

Powyższy opis dotyczy wielu systemów, których głównym zadaniem jest wyszukiwanie informacji w określonym zbiorze danych. Uzyskiwane wyniki nie zawsze będą tym, czego użytkownik poszukuje, dlatego stosowane są różne metody optymalizacji, poprawienia jakości i skuteczności wyszukiwania, coraz częściej z użyciem podanego przez użytkownika kontekstu wyszukiwania.

## Wyszukiwarki kontekstowe

Wyszukiwarka kontekstowa (Contextual Search Engine) w większości przypadków buduje zapytania kierowane do repozytorium na podstawie wpisanej przez użytkownika frazy. Użytkownik w ten sposób tworzy kontekst wyszukiwania. Niektóre z nich dodatkowo, jako kontekst wyszukiwania, przyjmują historię odwiedzanych przez użytkownika stron, czy pobranych dokumentów z danej wyszukiwarki. Są to ukryte informacje, które wyszukiwarki wykorzystują podczas budowania zapytań do repozytoriów danych. Ma to na celu udoskonalanie zapytań tak, aby pobierały te informacje, które użytkownik faktycznie próbuje odnaleźć [28].

Pierwszą z wykonywanych czynności wyszukiwarki kontekstowej jest wyodrębnienie słów kluczowych z kontekstu wpisanego przez użytkownika. W pracy [8] przedstawiono rozwiązanie problemu wydobywania słów kluczowych z wyrażenia wpisanego przez użytkownika. Z formułowaniem zapytań wiąże się wiele problemów, między innymi niedopasowanie słownictwa. Sytuacja taka ma miejsce, gdy na dany temat można znaleźć wiele pasujących słów lub gdy jedno słowo ma wiele znaczeń. Ponadto prawdopodobieństwo tego, że dwóch wyszukujących wybierze dokładnie te same słowa kluczowe dla zadanego pojęcia wynosi od 0,1 do 0,2 [5]. Podejście do rozwiązania problemu polega na wyodrębnieniu czasowników, rzeczowników oraz wyrażeń przyimkowych do osobnych grup i na ich podstawie dokonywania wyszukiwania.

Aby budować dokładniejsze zapytania można korzystać z bazy danych słów bliskoznacznych opisanej w [1]. Baza ta została zbudowana wykorzystując algorytm SSA. Algorytm ten wykorzystuje bazę danych WorldNet. Po wpisaniu dwóch słów znajdowana jest ich bliskoznaczność w skali od 0 do 1 (wielkość bezwymiarowa). Gdy bliskoznaczność osiągnie wartość 1 oznacza to, że słowa są synonimami. Pary podobnych słów są wstawiane do omawianej bazy danych. Tabela 2.1. przedstawia wyniki algorytmu SSA dla wybranych par słów w języku angielskim.

|  |  |  |  |
| --- | --- | --- | --- |
| L.p. | Słowo 1. | Słowo 2. | Bliskoznaczność |
| 1. | abstract | outline | 1 |
| 2. | accepted | received | 1 |
| 3. | active | alive | 1 |
| 4. | add | appena | 1 |
| 5. | advance | progress | 1 |
| 6. | agency | office | 1 |
| 7. | double | raise | 0.93 |
| 8. | authentication | validation | 0.91 |
| 9. | deletion | removal | 0.89 |

Tabela .. Wyniki działania algorytmu SSA dla wybranych par słów (za [1])

Drugą czynnością wykonywaną przez większość wyszukiwarek kontekstowych jest przeszukiwanie repozytorium tekstów na podstawie otrzymanego zapytania i kontekstu. W pracy [13] omówione zostały trzy różne wyszukiwarki kontekstowe:

1. Query-Rewriting (ang. QR) – wysyła proste zapytanie do standardowego mechanizmu wyszukiwania. Standardowy mechanizm wyszukiwania ogranicza się do obsługi prostych pytań lub wyrażeń. Działa on podobnie do silników Google czy Yahoo! [13].
2. Rank-Biasing (ang. RB) – wysyła skomplikowane (złożone z różnych operatorów) zapytania do odpowiednio zmodyfikowanego pod tego typu zapytania mechanizmu wyszukiwania.
3. Iterative Filtering MetaSearch (IFM) – wysyła wiele prostych zapytań do wielu standardowych mechanizmów wyszukujących.

W wyszukiwarkach pierwszego z podanych rodzajów wykorzystywany jest model „Vector Space Model” tworzący wektor, który został opisany na początku niniejszego rozdziału. Problemem wyszukiwarek stosujących ten model jest to, że wektor ten może zawierać więcej terminów niż mechanizm wyszukiwania jest w stanie obsłużyć. Dodatkowo, im więcej terminów jest dodawanych do wektora za pomocą koniunkcji, tym bardziej skomplikowane otrzymujemy zapytanie, które może nie dawać oczekiwanych wyników.

Drugi rodzaj wyszukiwarek sortuje terminy z wektora kontekstu, a następnie wybiera terminy za pomocą funkcji rankingującej [7] i w ten sposób buduje złożone zapytania.

Trzeci rodzaj wyszukiwarek korzysta z szablonów do budowania zapytań, a także, podobnie jak poprzednie wyszukiwarki, z wektora parametrów zawierającego kontekst. Posiadając szablon zapytania oznaczonego literą q oraz wektor parametrów [a, b, d, e], można wygenerować następującą listę zapytań: qa, qb, qd, qe, qab, qbd, qde, qabd, qabe, qabde. Należy to rozumieć tak, że zapytanie q otrzymuje w parametrach termy z wektora. Nie próbuje się stosować wszystkich możliwych kombinacji zapytań z parametrami, ponieważ może to spowolnić działanie wyszukiwarki, a prawdopodobnie nie pozwoli uzyskać zadowalających wyników.

Warto jednak podkreślić fakt, że wyniki badań w pracy [13] określają metodę IFM, jako generującą wyniki mniej zadowalające użytkownika, niż otrzymane metodą RB.

## Systemy zorientowane na wyszukiwanie

W wielu pozycjach literaturowych można znaleźć opisy systemów indeksujących, czy też wyszukujących, których autorzy skuteczność swych rozwiązań opierają na budowaniu rozległych indeksów danych będących przedmiotem przeszukiwania. Systemy te można podzielić na dwie grupy:

1. specjalizowane wyszukiwarki internetowe, np. Google Scholar, CiteSeerX, Scirus – są to bezpłatne wyszukiwarki, których głównym zadaniem jest przeszukiwanie publikacji naukowych w celu znalezienia wpisanych przez użytkownika słów kluczowych. Implementują one model Boolowski. Wadą wyszukiwarek tego typu jest brak możliwości wyszukiwania w wydzielonym przez użytkownika zbiorze dokumentów oraz brak możliwości wpisywania złożonych zapytań do zbioru danych.
2. oprogramowania indeksujące i wyszukujące informacje w dokumentach na stałe wczytanych do baz systemowych, w tym:

* Mendeley Desktop – główne działanie systemu to indeksowanie i organizowanie dokumentów PDF (głównie publikacji naukowych). Jedną z rzadko spotykanych możliwości tego typu systemów jest dodawanie własnych (zgromadzonych wcześniej) dokumentów do repozytorium oraz dowolne grupowanie ich. Można także importować dokumenty z różnych stron internetowych. Dodatkowo istnieją statystyki czytelnictwa, które pomagają użytkownikom tego systemu komunikować się ze sobą nawzajem. System ten istnieje zarówno w wersji desktopowej, jak również webowej. Wyszukiwanie odbywa się po metadanych, takich jak abstrakt, tytuł, autorzy, słowa kluczowe, itp. [24].
* ScentIndex – działanie tego systemu jest oparte na tworzeniu dynamicznego indeksu wczytanych do systemu książek. ScentIndex służy przede wszystkim do ułatwienia czytania oraz szukania informacji po słowach kluczowych. Użytkownik podczas czytania książki może wpisać słowa kluczowe, które zostaną podświetlone i wyszukane w aktualnie przeglądanej książce. Książki można także przeszukiwać po alfabetycznym indeksie dodawanym do każdej z nich [4].
* Vaidurya, Medline – systemy te także dotyczą problemu grupowania, indeksowania oraz wyszukiwania w dokumentach informacji za pomocą wpisanej przez użytkownika frazy, ale są one ograniczone do tematyki medycznej. Ich repozytoria danych są hierarchizowane za pomocą słownika z hasłami medycznymi Mesh. W systemie Medline dodatkowo jest możliwość zaawansowanego szukania. Polega ona na tym, że użytkownik wpisuje słowa kluczowe, system w odpowiedzi wysyła listę wybranych dokumentów zawierających dane słowa kluczowe. Użytkownik może wybrać istotne dla siebie dokumenty, a system automatycznie zacznie przeglądać repozytorium danych w poszukiwaniu dokumentów podobnych, do tych wskazanych przez użytkownika. Obydwa systemy korzystają z modelu „Vector Space Model”. [14, 15, 31]

Wszystkie opisane powyżej systemy są rozbudowane i posiadają dużo funkcjonalności, jednakże w żadnym z nich nie zdefiniowano języka zapytań, którego można by było użyć do kontekstowego szukania informacji. Systemy te implementują zwykle wyszukiwanie za pomocą modelu Boolowskiego, którego działanie polega na wpisaniu słów kluczowych i łączeniu ich za pomocą operatorów AND i OR. Wynikami opisanych powyżej systemów są zazwyczaj dokumenty albo wskazanie miejsca (poprzez podświetlenie) występowania danego słowa (ScentIndex). W każdym z nich pomijany jest kontekst zdań zawarty w dokumentach. Nie ma także możliwości wyszukiwania całych zdań w odniesieniu do kontekstu wpisanego przez wyszukującego.

## Skuteczność i efektywność wyszukiwania dla metod heurystycznych

Według [17] jednymi z najlepszych metod wyszukiwania są metody heurystyczne wykorzystujące algorytmy genetyczne. Algorytm genetyczny można bowiem wykorzystać na wiele sposobów w celu poprawienia skuteczności wyszukiwania. Podawanymi przykładami tego typu rozwiązań jest modyfikacja reprezentacji dokumentów, czy też przekształcanie wyrażenia będącego kontekstem wyszukiwania. Z kolei w [7] opisano nową metodę odkrywania funkcji rankingowej za pomocą programowania genetycznego. Funkcja rankingowa jest funkcją stosowaną w wyszukiwarkach internetowych do przydzielenia rangi danemu dokumentowi znalezionemu podczas wyszukiwania. Funkcje rankingowe stosuje się głównie przy wyszukiwaniu informacji spośród obszernych zbiorów dokumentów.

W pracy [17] opisano także wykorzystanie programowania genetycznego do znajdowania najlepszego dopasowania dokumentów do wpisanego przez użytkownika zapytania. Rozwiązanie to bazuje na dwóch głównych kryteriach skuteczności wyszukiwania:

* stosunku liczby istotnych (właściwych) wybranych dokumentów do całkowitej liczby dokumentów w całym repozytorium (ang. recall).
* stosunku liczby istotnych (właściwych) wybranych dokumentów do całkowitej liczby wybranych dokumentów (ang. precision).

W efekcie głównym elementem zastosowanej tu funkcji oceniającej jest określanie poprawności najlepszego dopasowania, poprzez wyliczanie dwóch powyższych wielkości dla wyszukanych dokumentów. Do wykonania tych obliczeń określana jest liczba dokumentów, które użytkownik chce zobaczyć. Autorzy korzystają z modelu „Vector Space Model”, ponieważ dokumenty i zapytania są reprezentowane w nim poprzez wektory wielowymiarowe. Używając korelacji cosinusowej można używać tego modelu do wykrywania podobieństwa (dopasowania). O samym porównywaniu dokumentów oraz wykorzystaniu tego modelu można dokładniej przeczytać w pracy [19], w której przedstawiono proponowane rozwiązanie na zbiorach danych Wikipedii.

Z kolei w pracy [3] uwagę poświęcono głównie rozwiązaniu dwóch problemów dotyczących wyszukiwania stron internetowych. Pierwszym z nich jest podobieństwo stron WWW znajdowanych na pierwszych pozycjach rezultatów wyszukiwania, natomiast drugi problem dotyczy stron, które znajdują się na bardzo niskich pozycjach i prawdopodobieństwo przeczytania ich przez użytkownika jest bliskie zeru. Główną ideą w omawianej pracy jest grupowanie stron o podobnej tematyce, a następnie uruchomianie algorytmu genetycznego. Algorytm genetyczny zaczyna działanie, w momencie, kiedy w każdej grupie stron zostaną zidentyfikowane te o najwyższej randze względem wpisanego przez użytkownika zapytania. Autorzy pracy przedstawili sposób postępowania przy dokonywaniu oceny uzyskanych zbiorów stron przez algorytm genetyczny. Ocenę ową oparli na porównaniu zbioru stron o najlepszej punktacji (według funkcji rankingowej) ze zbiorem stron wyprodukowanym przez algorytm genetyczny. Porównanie dotyczyło:

1. sumy punktów ze stron w każdej grupie,
2. liczby grup reprezentowanych w zbiorze stron,
3. średniej odległości Euklidesowej pomiędzy wszystkimi parami stron w zbiorze (dokładniej między parami wektorów reprezentujących pojedyncze strony),
4. średniej odległości Hamminga pomiędzy wszystkimi parami stron w zbiorze (odległość ta jest liczona, jako suma słów kluczowych, dla których wystąpienia są obecne na jednej stronie, a nie występują na innych).

Metoda opisana w punkcie a) pozwala na mierzenie jakości zbioru według punktacji stron, w punkcie b) brany jest pod uwagę zestaw przedstawicieli poszczególnych grup. Metody określone w punktach c) i d) pozwalają na mierzenie niepodobieństwa stron ze zbioru wynikowego. Stosowanie metod grupowania oraz algorytmu genetycznego pozwoliło na uzyskanie rozwiązań dobrej jakości w krótkim czasie (kilka sekund) [17]. Wszystkie przedstawione powyżej problemy i metody ich rozwiązywania bazują na algorytmach genetycznych w celu poprawienia jakości i skuteczności wyszukiwanych informacji w odpowiedzi na wpisane przez użytkownika wyrażenie.