Metody indeksowania dokumentów tekstowych

Paweł Szołtysek

21 maja 2009

Agenda

Agenda
Wstęp
Indeksowanie
Metoda List Prostych
Metoda Saltona
Metoda List Inwersyjnych
Metoda Łańcuchowa
Metoda Chowa
Metoda Grafów
Podsumowanie

Czym jest wyszukiwanie informacji?

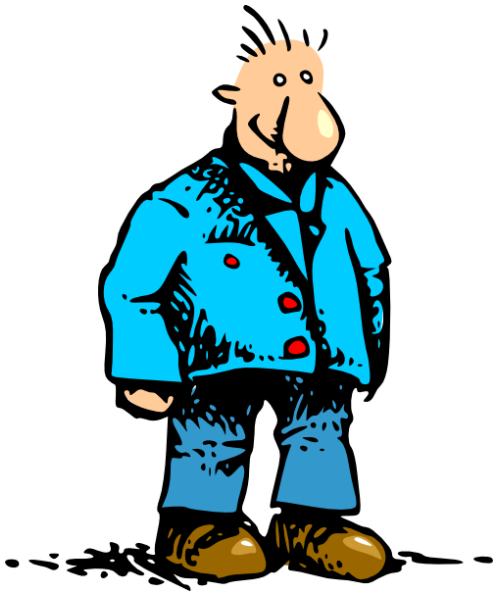
Wyszukiwanie informacji to wyszukiwanie w pewnym zbiorze (np. dokumentów tekstowych) według sformułowanej kwerendy, zawierających niezbędne dla użytkownika informacje.

Czym jest wyszukiwanie informacji?

Wyszukiwanie informacji to wyszukiwanie w pewnym zbiorze (np. dokumentów tekstowych) według sformułowanej kwerendy, zawierających niezbędne dla użytkownika informacje.

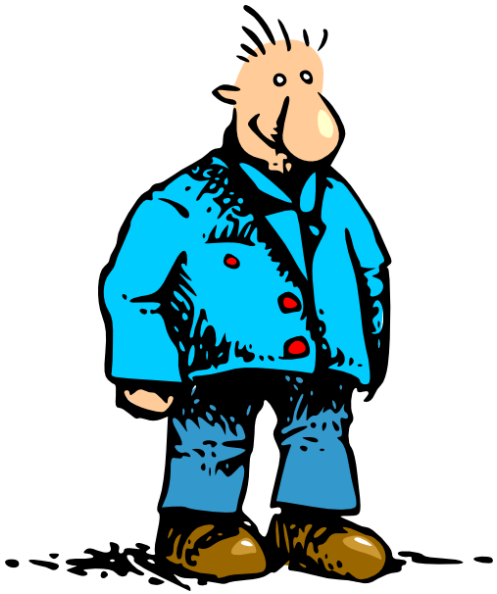
Czym jest wyszukiwanie informacji?

Wyszukiwanie informacji to wyszukiwanie w pewnym zbiorze (np. dokumentów tekstowych) według sformułowanej kwerendy, zawierających niezbędne dla użytkownika informacje.

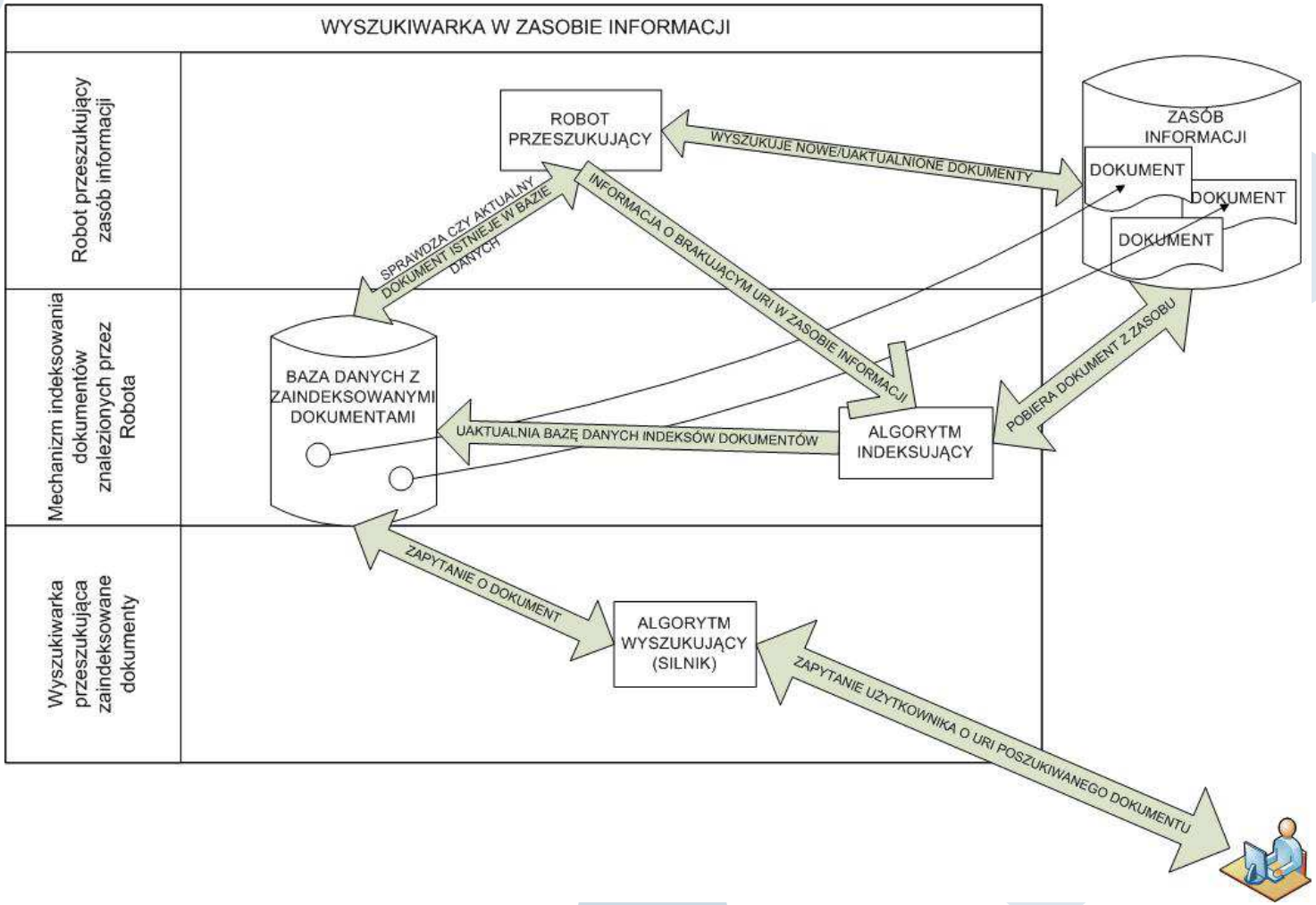


Czym jest wyszukiwanie informacji?

Wyszukiwanie informacji to wyszukiwanie w pewnym zbiorze (np. dokumentów tekstowych) według sformułowanej kwerendy, zawierających niezbędne dla użytkownika informacje.



Proces wyszukiwawczy.



Czym jest indeksowanie?

Indeksowanie jest najważniejszą operacją umożliwiającą wyszukiwanie informacji.

Czym jest indeksowanie?

Indeksowanie jest to proces tworzenia i utrzymywania indeksu umożliwiającego obniżenie czasu dostępu do danych; polega na określeniu tematu lub przedmiotu i wyrażeniu go w języku informacyjno-wyszukiwawczym.
Baza danych wyszukiwarki posiada właściwe dla siebie możliwości formułowania zapytań.

Podstawy indeksowania.

Indeksowanie musi być jednorodne.
Indeksowanie powinno dawać możliwość dowolnego konstruowania zapytań.
Indeksowanie powinno uwzględniać różne możliwości formułowania zapytań.
Indeksowanie NIE musi być odwrotne.
Indeksowanie NIE musi być dokładne.

Podstawowe problemy indeksowania

- ✔ Istnieją słowa, które, chociaż mają taką samą budowę składniową, róznią się semantycznie.
- ✔ Istnieją słowa, które, chociaż mają takie same znaczenie semantyczne, różnią się budową składniową.
- ✔ Istnieją słowa, które się bardzo często powtarzają, ale niewiele wnoszą do tekstu w sensie strukturalnym.
- ✔ Istnieją różne formy gramatyczne!
- ✔ Dokumenty są tworzone na podstawie doświadczenia i inteligencji człowieka.
- ✔ Dokumenty wcześniej zaindeksowane mogą podlegać zmianom.

Podstawowe problemy indeksowania

- ✔ Istnieją słowa, które, chociaż mają taką samą budowę składniową, różnią się semantycznie.
- ✔ Istnieją słowa, które, chociaż mają takie same znaczenie semantyczne, różnią się budową składniową.
- ✔ Istnieją słowa, które się bardzo często powtarzają, ale niewiele wnoszą do tekstu w sensie strukturalnym.
- ✔ Istnieją różne formy gramatyczne!
- ✔ Dokumenty są tworzone na podstawie doświadczenia i inteligencji człowieka.
- ✔ Dokumenty wcześniej zaindeksowane mogą podlegać zmianom.

Podstawowe problemy indeksowania

- ✔ Istnieją słowa, które, chociaż mają taką samą budowę składniową, różnią się semantycznie.
- ✔ Istnieją słowa, które, chociaż mają takie same znaczenie semantyczne, różnią się budową składniową.
- ✔ Istnieją słowa, które się bardzo często powtarzają, ale niewiele wnoszą do tekstu w sensie strukturalnym.
- ✔ Istnieją różne formy gramatyczne!
- ✔ Dokumenty są tworzone na podstawie doświadczenia i inteligencji człowieka.
- ✔ Dokumenty wcześniej zaindeksowane mogą podlegać zmianom.

Podstawowe problemy indeksowania

- ✔ Istnieją słowa, które, chociaż mają taką samą budowę składniową, różnią się semantycznie.
- ✔ Istnieją słowa, które, chociaż mają takie same znaczenie semantyczne, różnią się budową składniową.
- ✔ Istnieją słowa, które się bardzo często powtarzają, ale niewiele wnoszą do tekstu w sensie strukturalnym.
- ✔ Istnieją różne formy gramatyczne!
- ✔ Dokumenty są tworzone na podstawie doświadczenia i inteligencji człowieka.
- ✔ Dokumenty wcześniej zaindeksowane mogą podlegać zmianom.

Podstawowe problemy indeksowania

- ✔ Istnieją słowa, które, chociaż mają taką samą budowę składniową, różnią się semantycznie.
- ✔ Istnieją słowa, które, chociaż mają takie same znaczenie semantyczne, różnią się budową składniową.
- ✔ Istnieją słowa, które się bardzo często powtarzają, ale niewiele wnoszą do tekstu w sensie strukturalnym.
- ✔ Istnieją różne formy gramatyczne!
- ✔ Dokumenty są tworzone na podstawie doświadczenia i inteligencji człowieka.
- ✔ Dokumenty wcześniej zaindeksowane mogą podlegać zmianom.

Podstawowe problemy indeksowania

- ✔ Istnieją słowa, które, chociaż mają taką samą budowę składniową, różnią się semantycznie.
- ✔ Istnieją słowa, które, chociaż mają takie same znaczenie semantyczne, różnią się budową składniową.
- ✔ Istnieją słowa, które się bardzo często powtarzają, ale niewiele wnoszą do tekstu w sensie strukturalnym.
- ✔ Istnieją różne formy gramatyczne!
- ✔ Dokumenty są tworzone na podstawie doświadczenia i inteligencji człowieka.
- ✔ Dokumenty wcześniej zaindeksowane mogą podlegać zmianom.

Czynniki wpływające na indeksowanie.

Złączenie
Zapisywanie
Wielkość
Szybkość
Przechowywanie
Tolerancja
Język dokumentu.
Słowa kluczowe.

Czynniki wpływające na indeksowanie.

Złączenie
Zapisywanie
Wielkość
Szybkość
Przechowywanie
Tolerancja
Język dokumentu.
Słowa kluczowe.
Users will sacrifice accuracy for speed!

Metoda List Prostych (Forward Indexing)

Metoda List Prostych = Metoda Przeglądu Zupełnego

Indeks stanowi dokument (opis) wprost zapisany w bazie.
Pytanie jest zawarte w opisie dokumentu, jeśli wszystkie deskryptory pytania występują w opisie.
Wyszukiwanie po całym indeksie, zgodnie z regułami bazy danych w której się znajduje.
Najprostsza, najwolniejsza metoda.
Istnieją pewne sposoby przyśpieszania działania, np grupowanie dokumentów czy beczkowanie termów

Metoda List Prostych (Forward Indexing)

Metoda List Prostych = Metoda Przeglądu Zupełnego

Indeks stanowi dokument (opis) wprost zapisany w bazie.

Pytanie jest zawarte w opisie dokumentu, jeśli wszystkie deskryptory pytania występują w opisie.

Wyszukiwanie po całym indeksie, zgodnie z regułami bazy danych w której się znajduje.

Najprostsza, najwolniejsza metoda.

Istnieją pewne sposoby przyśpieszania działania, np grupowanie dokumentów czy beczkowanie termów

Metoda List Prostych (Forward Indexing)

Metoda List Prostych = Metoda Przeglądu Zupełnego

Indeks stanowi dokument (opis) wprost zapisany w bazie.

Pytanie jest zawarte w opisie dokumentu, jeśli wszystkie deskryptory pytania występują w opisie.

Wyszukiwanie po całym indeksie, zgodnie z regułami bazy danych w której się znajduje.

Najprostsza, najwolniejsza metoda.

Istnieją pewne sposoby przyśpieszania działania, np grupowanie dokumentów czy beczkowanie termów

Metoda List Prostych (Forward Indexing)

Metoda List Prostych = Metoda Przeglądu Zupełnego
Indeks stanowi dokument (opis) wprost zapisany w bazie.
Pytanie jest zawarte w opisie dokumentu, jeśli wszystkie deskryptory pytania występują w opisie.
Wyszukiwanie po całym indeksie, zgodnie z regułami bazy danych w której się znajduje.
Najprostsza, najwolniejsza metoda.
Istnieją pewne sposoby przyśpieszania działania, np grupowanie dokumentów czy beczkowanie termów

Metoda List Prostych (Forward Indexing)

Metoda List Prostych = Metoda Przeglądu Zupełnego
Indeks stanowi dokument (opis) wprost zapisany w bazie.
Pytanie jest zawarte w opisie dokumentu, jeśli wszystkie deskryptory pytania występują w opisie.
Wyszukiwanie po całym indeksie, zgodnie z regułami bazy danych w której się znajduje.

Najprostsza, najwolniejsza metoda.

Istnieją pewne sposoby przyśpieszania działania, np grupowanie dokumentów czy beczkowanie termów

Metoda List Prostych (Forward Indexing)

Metoda List Prostych = Metoda Przeglądu Zupełnego
Indeks stanowi dokument (opis) wprost zapisany w bazie.
Pytanie jest zawarte w opisie dokumentu, jeśli wszystkie deskryptory pytania występują w opisie.
Wyszukiwanie po całym indeksie, zgodnie z regułami bazy danych w której się znajduje.
Najprostsza, najwolniejsza metoda.
Istnieją pewne sposoby przyśpieszania działania, np grupowanie dokumentów czy beczkowanie termów

Metoda Saltona

Rozszerzenie metody list prostych.

Dokumenty są dzielone na grupy tematyczne (klasteryzowane).
Każda grupa jest opisana koniunkcją deskryptorów (z wagami).
Wyszukiwanie najpierw interesującą nas grupę dokumentów, a następnie jak w MLP.

Metoda Saltona

Rozszerzenie metody list prostych.

Dokumenty są dzielone na grupy tematyczne (klasteryzowane).

Każda grupa jest opisana koniunkcją deskryptorów (z wagami).

Wyszukiwanie najpierw interesującą nas grupę dokumentów, a następnie jak w MLP.

Metoda Saltona

Rozszerzenie metody list prostych.
Dokumenty są dzielone na grupy tematyczne (klasteryzowane).
Każda grupa jest opisana koniunkcją deskryptorów (z wagami).
Wyszukiwanie najpierw interesującą nas grupę dokumentów, a następnie jak w MLP.

Metoda List Inwersyjnych (Inverted Indexing)

Różnica: przypisanie dokumentów do deskryptora, a nie deskryptora do dokumentów.

Indeksowanie trwa dłużej, ale zapytanie to tylko znalezienie rekordów oraz operacja różnicy zbiorów.

Wyszukiwanie: znalezienie zbioru dokumentów dla każdego z deskryptorów, a następnie wybranie z nich części wspólnej.

Bardzo często wykorzystywana metoda indeksowania dokumentów.

Występujące modyfikacje skupiają się na dwóch priorytetach: pamięci (przedziały, negacja, redukcja) i czasie (ustawienie kolejności).

Metoda List Inwersyjnych (Inverted Indexing)

Różnica: przypisanie dokumentów do deskryptora, a nie deskryptora do dokumentów.

Indeksowanie trwa dłużej, ale zapytanie to tylko znalezienie rekordów oraz operacja różnicy zbiorów.

Wyszukiwanie: znalezienie zbioru dokumentów dla każdego z deskryptorów, a następnie wybranie z nich części wspólnej.

Bardzo często wykorzystywana metoda indeksowania dokumentów.

Występujące modyfikacje skupiają się na dwóch priorytetach: pamięci (przedziały, negacja, redukcja) i czasie (ustawienie kolejności).

Metoda List Inwersyjnych (Inverted Indexing)

Różnica: przypisanie dokumentów do deskryptora, a nie deskryptora do dokumentów.
Indeksowanie trwa dłużej, ale zapytanie to tylko znalezienie rekordów oraz operacja różnicy zbiorów.
Wyszukiwanie: znalezienie zbioru dokumentów dla każdego z deskryptorów, a następnie wybranie z nich części wspólnej.
Bardzo często wykorzystywana metoda indeksowania dokumentów.
Występujące modyfikacje skupiają się na dwóch priorytetach: pamięci (przedziały, negacja, redukcja) i czasie (ustawienie kolejności).

Metoda List Inwersyjnych (Inverted Indexing)

Różnica: przypisanie dokumentów do deskryptora, a nie deskryptora do dokumentów.
Indeksowanie trwa dłużej, ale zapytanie to tylko znalezienie rekordów oraz operacja różnicy zbiorów.
Wyszukiwanie: znalezienie zbioru dokumentów dla każdego z deskryptorów, a następnie wybranie z nich części wspólnej.
Bardzo często wykorzystywana metoda indeksowania dokumentów.
Występujące modyfikacje skupiają się na dwóch priorytetach: pamięci (przedziały, negacja, redukcja) i czasie (ustawienie kolejności).

Metoda List Inwersyjnych (Inverted Indexing)

Różnica: przypisanie dokumentów do deskryptora, a nie deskryptora do dokumentów.
Indeksowanie trwa dłużej, ale zapytanie to tylko znalezienie rekordów oraz operacja różnicy zbiorów.
Wyszukiwanie: znalezienie zbioru dokumentów dla każdego z deskryptorów, a następnie wybranie z nich części wspólnej.
Bardzo często wykorzystywana metoda indeksowania dokumentów.
Występujące modyfikacje skupiają się na dwóch priorytetach: pamięci (przedziały, negacja, redukcja) i czasie (ustawienie kolejności).

Metoda Łańcuchowa

Stanowi pewnego rodzaju połączenie metody list prostych i inwersyjnych.

Indeks jest tworzony jak w metodzie list prostych, ale przy każdym deskryptorze jest podany odnośnik do następnego jego wystąpienia.

Ponadto tworzona jest lista deskryptorów z pierwszym jego wystąpieniem oraz łączną ilością ich wystąpień w całym indeksie.

Proces indeksowania jest znacznie dłuższy niż w przypadku metody list prostych, ale wyszukiwanie jest bardzo szybkie.

Wyszukiwanie: dla jednego deskryptora znajduje się go w liście deskryptorów, oraz kolejno czyta dokumenty w których występuje.

Przy większej ilości deskryptorów, zaczyna się od termu który posiada najmniejszą ilość wystąpień w całym indeksie.

Prosta, ale bardzo efektywna metoda wyszukiwania.

Modyfikacje jakie można zastosować opierają się na podwójnym łańcuchowaniu oraz łańcuchowaniu grup obiektów.

Metoda Łańcuchowa

Stanowi pewnego rodzaju połączenie metody list prostych i inwersyjnych.

Indeks jest tworzony jak w metodzie list prostych, ale przy każdym deskryptorze jest podany odnośnik do następnego jego wystąpienia.

Ponadto tworzona jest lista deskryptorów z pierwszym jego wystąpieniem oraz łączną ilością ich wystąpień w całym indeksie.

Proces indeksowania jest znacznie dłuższy niż w przypadku metody list prostych, ale wyszukiwanie jest bardzo szybkie.

Wyszukiwanie: dla jednego deskryptora znajduje się go w liście deskryptorów, oraz kolejno czyta dokumenty w których występuje.

Przy większej ilości deskryptorów, zaczyna się od termu który posiada najmniejszą ilość wystąpień w całym indeksie.

Prosta, ale bardzo efektywna metoda wyszukiwania.

Modyfikacje jakie można zastosować opierają się na podwójnym łańcuchowaniu oraz łańcuchowaniu grup obiektów.

Metoda Łańcuchowa

Stanowi pewnego rodzaju połączenie metody list prostych i inwersyjnych.
Indeks jest tworzony jak w metodzie list prostych, ale przy każdym deskryptorze jest podany odnośnik do następnego jego wystąpienia.
Ponadto tworzona jest lista deskryptorów z pierwszym jego wystąpieniem oraz łączną ilością ich wystąpień w całym indeksie.
Proces indeksowania jest znacznie dłuższy niż w przypadku metody list prostych, ale wyszukiwanie jest bardzo szybkie.
Wyszukiwanie: dla jednego deskryptora znajduje się go w liście deskryptorów, oraz kolejno czyta dokumenty w których występuje.
Przy większej ilości deskryptorów, zaczyna się od termu który posiada najmniejszą ilość wystąpień w całym indeksie.
Prosta, ale bardzo efektywna metoda wyszukiwania.
Modyfikacje jakie można zastosować opierają się na podwójnym łańcuchowaniu oraz łańcuchowaniu grup obiektów.

Metoda Łańcuchowa

Stanowi pewnego rodzaju połączenie metody list prostych i inwersyjnych.
Indeks jest tworzony jak w metodzie list prostych, ale przy każdym deskryptorze jest podany odnośnik do następnego jego wystąpienia.
Ponadto tworzona jest lista deskryptorów z pierwszym jego wystąpieniem oraz łączną ilością ich wystąpień w całym indeksie.
Proces indeksowania jest znacznie dłuższy niż w przypadku metody list prostych, ale wyszukiwanie jest bardzo szybkie.
Wyszukiwanie: dla jednego deskryptora znajduje się go w liście deskryptorów, oraz kolejno czyta dokumenty w których występuje.
Przy większej ilości deskryptorów, zaczyna się od termu który posiada najmniejszą ilość wystąpień w całym indeksie.
Prosta, ale bardzo efektywna metoda wyszukiwania.
Modyfikacje jakie można zastosować opierają się na podwójnym łańcuchowaniu oraz łańcuchowaniu grup obiektów.

Metoda Łańcuchowa

Stanowi pewnego rodzaju połączenie metody list prostych i inwersyjnych.
Indeks jest tworzony jak w metodzie list prostych, ale przy każdym deskryptorze jest podany odnośnik do następnego jego wystąpienia.
Ponadto tworzona jest lista deskryptorów z pierwszym jego wystąpieniem oraz łączną ilością ich wystąpień w całym indeksie.
Proces indeksowania jest znacznie dłuższy niż w przypadku metody list prostych, ale wyszukiwanie jest bardzo szybkie.
Wyszukiwanie: dla jednego deskryptora znajduje się go w liście deskryptorów, oraz kolejno czyta dokumenty w których występuje.
Przy większej ilości deskryptorów, zaczyna się od termu który posiada najmniejszą ilość wystąpień w całym indeksie.
Prosta, ale bardzo efektywna metoda wyszukiwania.
Modyfikacje jakie można zastosować opierają się na podwójnym łańcuchowaniu oraz łańcuchowaniu grup obiektów.

Metoda Łańcuchowa

Stanowi pewnego rodzaju połączenie metody list prostych i inwersyjnych.
Indeks jest tworzony jak w metodzie list prostych, ale przy każdym deskryptorze jest podany odnośnik do następnego jego wystąpienia.
Ponadto tworzona jest lista deskryptorów z pierwszym jego wystąpieniem oraz łączną ilością ich wystąpień w całym indeksie.
Proces indeksowania jest znacznie dłuższy niż w przypadku metody list prostych, ale wyszukiwanie jest bardzo szybkie.
Wyszukiwanie: dla jednego deskryptora znajduje się go w liście deskryptorów, oraz kolejno czyta dokumenty w których występuje.
Przy większej ilości deskryptorów, zaczyna się od termu który posiada najmniejszą ilość wystąpień w całym indeksie.
Prosta, ale bardzo efektywna metoda wyszukiwania.
Modyfikacje jakie można zastosować opierają się na podwójnym łańcuchowaniu oraz łańcuchowaniu grup obiektów.

Metoda Chowa (1)

Opiera się na podziale wszystkich dokumentów na grupy, które są opisane przez k zadanych deskryptorów ze zbioru n dostępnych. (Grup więc będzie $\binom{n}{k}$.) Zwykle k odpowiada średniej długości najczęściej zadawanych pytań.

Zapisujemy w odpowiedniej kolejności wybrane deskryptory, bazując na nich tworzymy grupy. Ponieważ znamy ich kolejność, bezpośrednio po otrzymaniu zapytania możemy określić w jakiej grupie znajduje siędany ciąg deskryptorów - $\binom{1}{k} - \sum_{v=1}^{k-1} \frac{n-j_v-1}{k-v+1}\binom{n-j_v-1}{k-v} - (n - j_k)$.

Indeksowanie jest proste - wystarczy określić występowanie kolejnych deskryptorów w dokumentach i dodać dokument do konkretnej grupy.

Metoda Chowa (1)

Opiera się na podziale wszystkich dokumentów na grupy, które są opisane przez k zadanych deskryptorów ze zbioru n dostępnych. (Grup więc będzie $\binom{n}{k}$.) Zwykle k odpowiada średniej długości najczęściej zadawanych pytań.

Zapisujemy w odpowiedniej kolejności wybrane deskryptory, bazując na nich tworzymy grupy. Ponieważ znamy ich kolejność, bezpośrednio po otrzymaniu zapytania możemy określić w jakiej grupie znajduje siędany ciąg deskryptorów - $\binom{1}{k} - \sum_{v=1}^{k-1} \frac{n-j_v-1}{k-v+1}\binom{n-j_v-1}{k-v} - (n - j_k)$.

Indeksowanie jest proste - wystarczy określić występowanie kolejnych deskryptorów w dokumentach i dodać dokument do konkretnej grupy.

Metoda Chowa (1)

Opiera się na podziale wszystkich dokumentów na grupy, które są opisane przez k zadanych deskryptorów ze zbioru n dostępnych. (Grup więc będzie $\binom{n}{k}$.) Zwykle k odpowiada średniej długości najczęściej zadawanych pytań.
Zapisujemy w odpowiedniej kolejności wybrane deskryptory, bazując na nich tworzymy grupy. Ponieważ znamy ich kolejność, bezpośrednio po otrzymaniu zapytania możemy określić w jakiej grupie znajduje siędany ciąg deskryptorów - $\binom{1}{k} - \sum_{v=1}^{k-1} \frac{n-j_v-1}{k-v+1}\binom{n-j_v-1}{k-v} - (n - j_k)$.

Indeksowanie jest proste - wystarczy określić występowanie kolejnych deskryptorów w dokumentach i dodać dokument do konkretnej grupy.

Metoda Chowa (2)

Wyszukiwanie opiera się na znajdowaniu odpowiedniej grupy. Jeśli wprowadziliśmy więcej deskryptorów niż k, wyznaczamy iloczyn zbiorów zawartości grup. Jeśli natomiast jest ich mniej, bierzemy sumę po wszystkich interesujących nas grupach.

Ten rodzaj wyszukiwania jest najszybszy dla zapytań które mają długość deskryptorów k, a akceptowalny dla tych które mają ich więcej. Jeśli jest ich mniej, czas się znacznie wydłuża.

Metoda Chowa (2)

Wyszukiwanie opiera się na znajdowaniu odpowiedniej grupy. Jeśli wprowadziliśmy więcej deskryptorów niż k, wyznaczamy iloczyn zbiorów zawartości grup. Jeśli natomiast jest ich mniej, bierzemy sumę po wszystkich interesujących nas grupach.

Ten rodzaj wyszukiwania jest najszybszy dla zapytań które mają długość deskryptorów k, a akceptowalny dla tych które mają ich więcej. Jeśli jest ich mniej, czas się znacznie wydłuża.

Metoda Grafów

Nowa jakość w indeksowaniu dokumentów.

Podczas procesu indeksowania tworzony jest graf dokumentu.
Wykorzystywane są do tego metody wcześniej wymienione.
Przy realizowaniu zapytania, wykorzystywane jest określanie względnej pozycji dwóch termów w dokumencie.
Strukturą jest graf, mamy więc olbrzymie możliwości dotyczące rozbudowy algorytmu jego generowania, wykorzystywania oraz wyciągania różnych statystyk.

Metoda Grafów

Nowa jakość w indeksowaniu dokumentów.

Podczas procesu indeksowania tworzony jest graf dokumentu.

Wykorzystywane są do tego metody wcześniej wymienione.

Przy realizowaniu zapytania, wykorzystywane jest określanie względnej pozycji dwóch termów w dokumencie.

Strukturą jest graf, mamy więc olbrzymie możliwości dotyczące rozbudowy algorytmu jego generowania, wykorzystywania oraz wyciągania różnych statystyk.

Metoda Grafów

Nowa jakość w indeksowaniu dokumentów.
Podczas procesu indeksowania tworzony jest graf dokumentu.
Wykorzystywane są do tego metody wcześniej wymienione.
Przy realizowaniu zapytania, wykorzystywane jest określanie względnej pozycji dwóch termów w dokumencie.
Strukturą jest graf, mamy więc olbrzymie możliwości dotyczące rozbudowy algorytmu jego generowania, wykorzystywania oraz wyciągania różnych statystyk.

Metoda Grafów

Nowa jakość w indeksowaniu dokumentów.
Podczas procesu indeksowania tworzony jest graf dokumentu.
Wykorzystywane są do tego metody wcześniej wymienione.
Przy realizowaniu zapytania, wykorzystywane jest określanie względnej pozycji dwóch termów w dokumencie.

Strukturą jest graf, mamy więc olbrzymie możliwości dotyczące rozbudowy algorytmu jego generowania, wykorzystywania oraz wyciągania różnych statystyk.

Podsumowanie

Nie istnieje jeden najlepszy algorytm indeksowania.

Wybór algorytmu i jego parametrów powinien być uzależniony od tego do czego ma on służyć.

Podczas implementacji należy zwracać uwagę czy założone przez nas ograniczenia na indeks zostaną spełnione.

Podsumowanie

Nie istnieje jeden najlepszy algorytm indeksowania.

Wybór algorytmu i jego parametrów powinien być uzależniony od tego do czego ma on służyć.

Podczas implementacji należy zwracać uwagę czy założone przez nas ograniczenia na indeks zostaną spełnione.

Podsumowanie

Nie istnieje jeden najlepszy algorytm indeksowania.
Wybór algorytmu i jego parametrów powinien być uzależniony od tego do czego ma on służyć.
Podczas implementacji należy zwracać uwagę czy założone przez nas ograniczenia na indeks zostaną spełnione.

Literatura

- ✔ Mieczysław Alojzy Kłopotek: *Inteligentne wyszukiwarki internetowe*, Akademicka Oficyna Wydawnicza Exit, Warszawa 2001.
- ✔ K. Rzerzicha, W. Skórski, P. Szołtysek: *Projecting of the text document searching system: indexation using graph structures*, Information Systems Architecture and Technology. Web Information Systems: Models, Concepts & Challenges.
- ✔ www.adamiko.republika.pl
- ✔ *Systemy Wyszukiwania Informacji*, http://www.swi.wsti.pl/