Zakres tematyczny na zaliczenie EDWI

Spis treści

1. Czym są systemy IR i historia ich powstania

- W latach 50-tych badacze tacy jak Hans Peter Luhn, Eugene Garfield,Philip Bagley i Calvin Moores wypracowali pojęcie wyszukiwania informacji - *Information Retrieval* (IR)

- W 1963 roku Joseph Becker i Robert Hayes opublikowali pierwszą książkę na temat IR

- W 1983 r. Salton i McGill opublikowali książkę *Introduction to Modern Information Retrieval* omawiającą model wektorowy pozyskiwania danych

- Biblioteki były pierwszymi instytucjami wykorzystującymi systemy IR dla pozyskiwania informacji w formie przeszukiwania kart katalogowych

- Pełny opis wymaganej informacji podany przez użytkownika nie zawsze jest trafnym zapytaniem do systemu IR

- Użytkownik sieci może także sformułować swoje wymagania w formie zapytania

- Najistotniejszy dla wyszukiwania jest zbiór słów kluczowych (*keywords*) lub terminów indeksujących (*index terms*).

- Celem systemów IR jest jak najtrafniejsze wydobycie informacji istotnej dla użytkownika na podstawie podanego zapytania

---

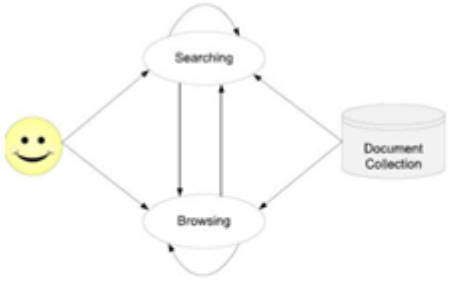
Problem wyszukiwania informacji

- System IR powinien uszeregować elementy informacji według ich istotności w zapytaniu użytkownika

- Celem systemu IR jest wydobycie wszystkich elementów istotnych dla zapytania użytkownika i jak najmniejszej ilości elementów nieistotnych

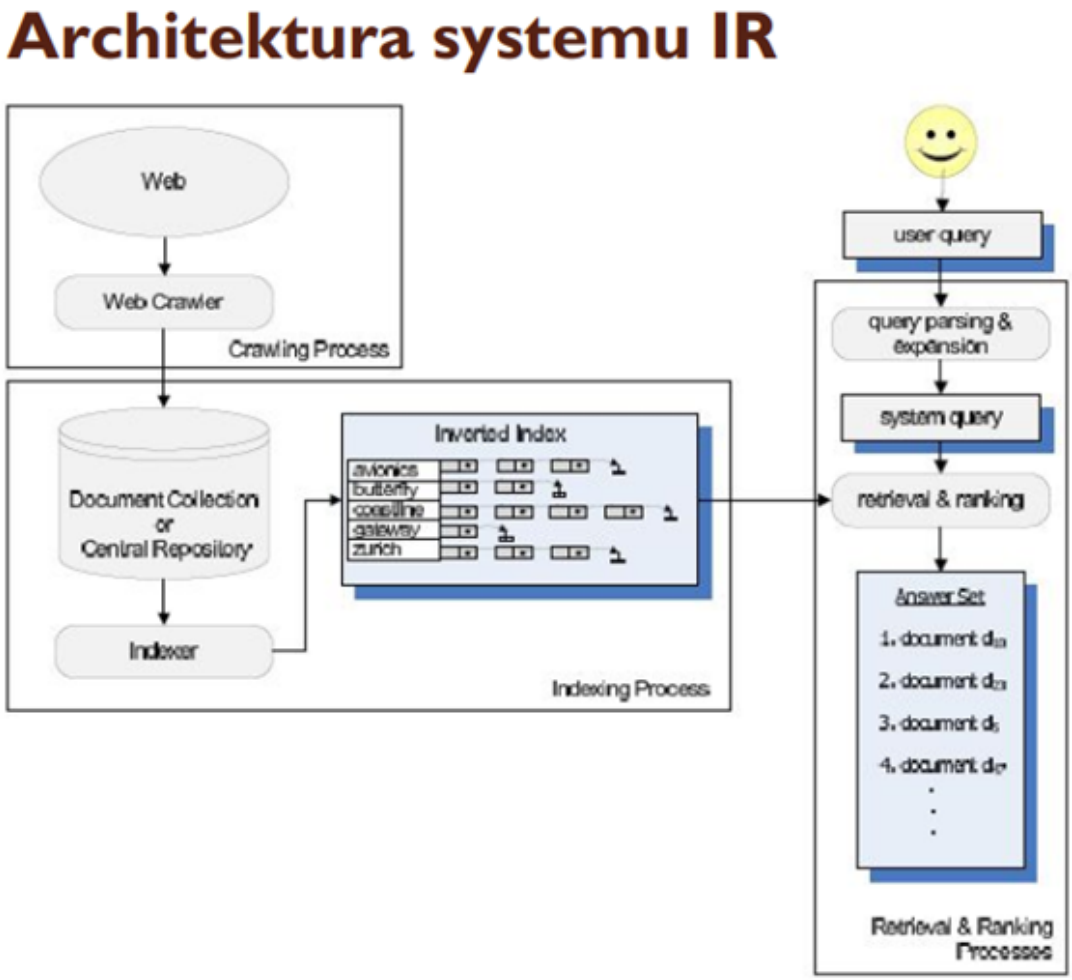
- Pojęcie istotności informacji w systemach IR jest kluczowe

2. Czym są i jaka jest różnica pomiędzy searching a query?



- Jeśli użytkownik precyzuje konkretny temat, często w formie zapytania, to mówi się że poszukuje, wyławia (*sarching*) lub pyta o informację (*querying*)

3. Szczegółowy opis architektury IR (slajd 14; wykład 1)



4. Czym jest pełzanie po stronach: crawling?

Sieć narzuca specyficzną charakterystykę wyszukiwania zbioru dokumentów – strony rozproszone w milionach witryn połączonych hiperlinkami; rozproszone dokumenty o pożądanych cechach są wydobywane i kopiowane w jedno miejsce przed ich indeksowaniem. Taki sposób wyszukiwania stron w procesie IR nazywa się „pełzaniem (po stronach)" (*crawling*).

5. Co powinien umożliwić interfejs IR?

Drugi wpływ sieci na wyszukiwanie to krytyczne znaczenie jakości i skalowalności procesu IR

Wobec przeszukiwania dużych zbiorów w sieci przewidywanie istotności danych staje się bardzo istotne

Sieć stanowi także medium do prowadzenia biznesu; strony zawierają linki do ładowania programów, adresy, numery telefonów instytucji itp.

Przy wyszukiwaniu w sieci należy eliminować spamy

Zapewnienie bezpieczeństwa, prywatności, praw autorskich i patentowych

Skanowanie i rozpoznawanie pisma przy wyszukiwaniu w różnych językach

6. Od czego zależy interakcja użytkownika z interfejsem?

- Rolą interfejsu użytkownika jest pomoc w sformułowaniu zapytania i odebraniu wymaganych informacji

- Interfejs powinien także umożliwić:
  - wybór źródła informacji,
  - zrozumienie wyników wyszukiwania,
  - śledzenie postępu w wyszukiwaniu

- Interakcja użytkownika z interfejsem wyszukującym zależy od:
  - typu postawionego zadania,
  - dotychczasowej wiedzy o użytkowniku,
  - ilości czasu i wysiłku wkładanego w proces wyszukiwania.

7. Jaka jest różnica pomiędzy information lookup i exploratory search?

- Wyszukiwanie informacji (*information lookup*) :
  - jest podobne do wyszukiwania faktów i odpowiedzi na pytania,
  - może być wypełnione przez dyskretne informacje jak liczby, daty, nazwy lub strony sieciowe,
  - pracuje poprawnie w trybie standardowych interakcji z siecią.

- Wyszukiwanie rozpoznawcze (*exploratory search*) można podzielić na zadania śledzenia i uczenia się

- Wyszukiwanie uczące się (*learning search*) wymaga:
  - więcej niż jednej akcji zapytanie-odpowiedź,
  - czasu od wyszukiwacza,
  - odczytywania wielu porcji informacji,
  - złożenia różnych treści do sformowania odpowiedzi.

8. Czym jest sensemaking i kto wymyślił takie pojęcie?

Sensemaking (Karl Weick 1969) -organizowanie wiedzy: proces iteracyjny tworzenia konceptualnej reprezentacji dużych zbiorów odpowiedzi.

9. Czym jest berry picking?

Najnowsze modele uwzględniają dynamiczne aspekty wyszukiwania:

- użytkownicy uczą się w trakcie wyszukiwania,
- potrzeby informacyjne użytkowników zmieniają się podczas przeglądania wyników wyszukiwania.

Model dynamicznego wyszukiwania nazywa się także „zbieraniem jagód" *(berry picking*) lub wyszukiwaniem zorientowanym (*orienteering*)

10. Opisz technikę “query history and revisitation”.

- Rozliczne studia prowadzone nad procesami wyszukiwania doprowadziły do wniosku, że użytkownicy często
  - reformułują zapytania z niewielkimi modyfikacjami,
  - ponownie szukają informacji znalezionych poprzednio.

- Badacze opracowali wsparcie interfejsów wyszukiwania z odwołaniem do historii zapytań i ponawianiem zapytań (*query history and revisitation*)

- Badania pokazują, że wyszukujący zwracają uwagę na kilka początkowych pozycji w rankingu wyszukanych odpowiedzi.

- Użytkownicy są zazwyczaj przekonani, że pierwsze wyniki w zbiorze odpowiedzi są lepsze niż pozostałe.

11. Czym są I kto wprowadził zapytania koniunktywne(wykład 2; s11.)

- Ok. 1997 roku Google wprowadził wyłącznie zapytania koniunktywne, które wkrótce stały się normą.

- Google dodał pojęcie informacji przybliżonej i wprowadził ranking stron (PageRank).

- W miarę rozwoju sieci pojawiły się poprawne odpowiedzi na dłuższe pytania stawiane w formie fraz.

12. Jakim hasłem określa się technikę rankingu stron stosowaną w Google?

- Ok. 1997 roku Google wprowadził wyłącznie zapytania koniunktywne, które wkrótce stały się normą.

- Google dodał pojęcie informacji przybliżonej i wprowadził ranking stron (PageRank).

- W miarę rozwoju sieci pojawiły się poprawne odpowiedzi na dłuższe pytania stawiane w formie fraz.

13. Jak nazywa się pole edycyjne zapytania wyszukującego w google?

- Standardowym interfejsem wejściowym zapytania jest pole edycyjne (*search box entry*).

14. Jak nazywa się technika rozszerzenia zapytania przy wykorzystaniu podpowiedzi?

- Nazywa się to autouzupełnianiem, autosugerowaniem, lub dynamiczną sugestią zapytań (*auto-complete, auto-suggest, or dynamic query suggestions*)

- Sugestie dotyczą często uzupełnienia wprowadzonych znaków traktowanych jako prefiksy, rzadziej podają uzupełnienia środkowych liter

15. Jak nazywa się metoda, która zakłada, że użytkownik wskaże, które dokumenty wybrano trafnie lub które termy wybrane z dokumentów są istotne?

- Metoda „**relevance feedback**" zakłada, że użytkownik wskaże, które dokumenty wybrano trafnie lub które termy wybrane z dokumentów są istotne

16. Czym jest reprezentacja fasetowa?

- **Reprezentacja fasetowa** metadanych pozwala na przypisanie wielu kategorii do pojedynczej pozycji wyników

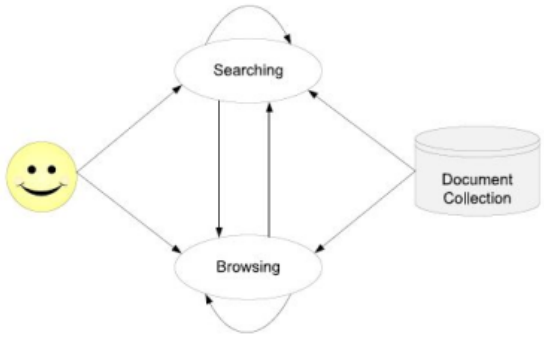
17. Jaka jest różnica pomiędzy klasyfikacją a klasteryzacją?

- Klasyfikacja (kategoryzacja) tekstów:
  Proces porządkowania informacji poprzez kojarzenie dokumentów tekstowych z klasami (kategoriami)

- **Klastering** to grupowanie pozycji wyników według pewnej miary podobieństwa
  - Grupuje razem dokumenty podobne do siebie i różne od pozostałych – np. dokumenty w języku japońskim w zbiorze publikacji głównie angielskich
- Klastering jest w pełni automatyczny, ale może dać wyniki grupowania niezgodne z intuicją użytkownika

18. Wymień 4 techniki wyników wyszukiwania danych.

Problem wyszukiwania informacji



- Jeśli użytkownik precyzuje konkretny temat, często w formie zapytania, to mówi się że poszukuje, wyławia (*sarching*) lub pyta o informację (*querying*)

- Jeżeli użytkownik formułuje wymagania szeroko lub nieprecyzyjnie to mówi się o żeglowaniu (*navigating*) lub przeglądaniu (*browsing*) dokumentów w Internecie

Sieć narzuca specyficzną charakterystykę wyszukiwania zbioru dokumentów – strony rozproszone w milionach witryn połączonych hiperlinkami; rozproszone dokumenty o pożądanych cechach są wydobywane i kopiowane w jedno miejsce przed ich indeksowaniem. Taki sposób wyszukiwania stron w procesie IR nazywa się „pełzaniem (po stronach)" (*crawling*).

19. Czym jest Term indeksujący?

- Term indeksujący:
  - W sensie ścisłym – jest to słowo kluczowe o pewnym znaczeniu, zazwyczaj rzeczownik
  - W sensie ogólnym – dowolne słowo pojawiające się w dokumencie
- Wyszukiwanie danych opiera się na termach indeksujących

20. Czym jest słownik termów?

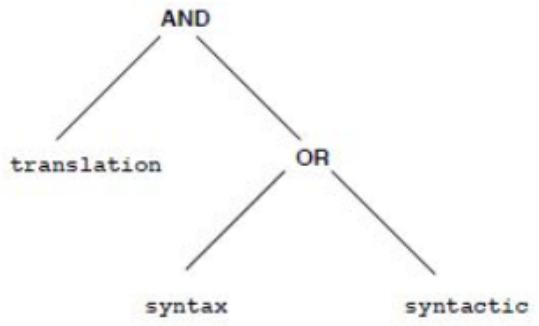
- Słownik $V = \{k_1, \ldots, k_t\}$ - zbiór różnych termów indeksujących
  - t - liczba termów w kolekcji dokumentów
  - k_i – elementarny term indeksujący

21. W jaki sposób można opisać czym jest boolowski model wyszukiwania?

Zapytania są specyfikowane jako wyrażenia boolowskie.

Indeksowanie i wyszukiwanie

- Zapytania boolowskie: - składnia w formie drzewa



- w pierwszej fazie określa się które dokumenty są do porównania,
- w drugiej fazie ocenia się istotność dokumentów,
- w trzeciej fazie wyznacza się dokładne pozycje dopasowania

22. Należy dokładnie znać model Wag TF-IDF. Wzór wraz ze zrozumieniem w jaki sposób obliczyć ww. wskaźnik

Wagi TF-IDF

- Do ustalenia wag termów w dokumentach stosuje się formułę TF-IDF (term frequency – inverse document frequency)
- **Odwrotna częstotliwość dokumentu dla** termu k_i

$$idf_i = \log \frac{N}{n_i},$$

gdzie n_i – ilość dokumentów z termem k_i, N – *wielkość kolekcji dokumentów*

- **Częstotliwość termu** k_i dla całej kolekcji dokumentów

$$tf_i = 1 + \log \sum_{j=1}^{N} f_{i,j},$$

log –logarytm o podstawie 2

23. Jakie są trzy metody normalizacji dokumentów? (wykład 3 slajd 20)

Normalizacja długości dokumentu

- Dokumenty posiadają różne długości; dłuższe z nich, a niekoniecznie bardziej istotne, mają większą szansę być wyszukane w danym zapytaniu
- Aby usunąć ten niepożądany efekt dzieli się rangę każdego dokumentu przez jego długość; ten proces nazywamy **normalizacją długości dokumentu**
- Metody normalizacji:
  - **Rozmiar w bajtach:** każdy dokument traktuje się jako strumień bajtów,
  - **Liczba słów:** dokument jest traktowany jako pojedynczy łańcuch słów do zliczenia
  - **Normy wektorowe:** dokumenty są reprezentowane jako wektory termów ważonych

24. Zakładając, że dwa dokumenty są przedstawione jako wektory, jak nazywa się miara obliczenia odległości pomiędzy nimi?

Dystans kosinusowy?

25. Jaki wzór jest wzorem podstawowym przy wykorzystaniu modelu probabilistycznego?

Ranking probabilistyczny

- Model probabilistyczny:
  - Estymuje prawdopodobieństwo, że dokument jest istotny dla zapytania użytkownika,
  - Zakłada, że to prawdopodobieństwo zależy jedynie od zapytania i reprezentacji dokumentów,
  - Idealny zbiór odpowiedzi *R*, maksymalizuje prawdopodobieństwo istotności dokumentu
- Podobieństwo dokumentu do zapytania:

$$\mathrm{sim}(d_j, q) = \frac{P(R \mid \mathbf{d}_j, q)}{P(\overline{R} \mid \mathbf{d}_j, q)},$$

gdzie R – zbiór dokumentów istotnych dla zapytania q,

$\overline{R}$ – zbiór dokumentów nieistotnych

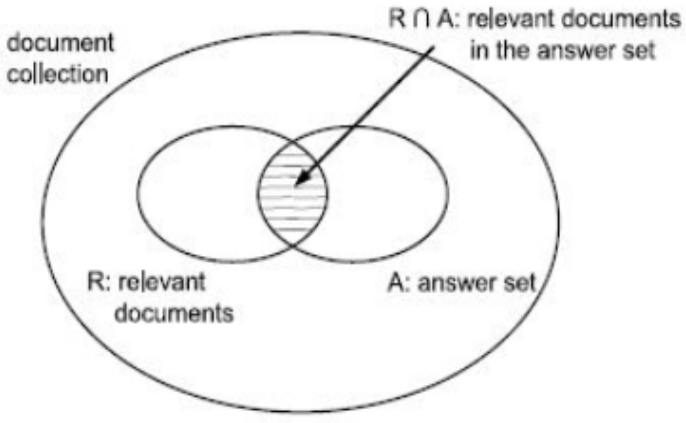
26. Jak nazywa się model, w którym dopasowanie dokumentu do termów zapytania ma charakter przybliżony?

27. Czy istnieją modele IR neuronowe?

28. Jakie parametry są kluczowe podczas oceny wyszukiwania?

29. Czym jest i jak się oblicza wartość tzw. recall przy ocenie wyszukiwania?

- Dane są:
  - *I*: żądanie informacji,
  - *R*: zbiór dokumentów istotnych dla *I*,
  - *A*: zbiór odpowiedzi dla *I*, wygenerowanych przez system wyszukiwania IR,
  - $R \cap A$: iloczyn zbiorów *R* i *A*.



- Kompletność jest częścią istotnych dokumentów (zbiór *R*), które zostały wyszukane:

$$Recall = \frac{|R \cap A|}{|R|}$$

Dokładność jest częścią wyszukanych dokumentów (zbiór *A*), które są istotne:

$$Precision = \frac{|R \cap A|}{|A|}$$

30. Czym jest miara P@5 i czym różni się od miary P@10?

Miary P@5 i P@10

- Użytkownicy wyszukiwarek cenią przede wszystkim dużo istotnych dokumentów na początku rankingu
- Miary *P@5* i *P@10* wyznaczają dokładność odpowiednio dla 5 i 10 pierwszych dokumentów; pozwalają ocenić czy użytkownicy uzyskują istotne dokumenty na początku rankingu
- Rozpatrujemy listę dokumentów dla przykładowego pytania q_1:

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

31. Czym różnią się metody klasyfikacji dokumentów z nadzorem od metod bez nadzoru?

- Algorytmy nadzorowane bazują na zbiorze testowym (uczącym)
  - Zbiór klas z przykładami dokumentów w każdej z nich
  - Przynależność do klas określają specjaliści
  - Zbiór testowy służy do uczenia klasyfikatora

- Duża liczność zbioru uczącego lepiej dostraja klasyfikator i zapobiega przeuczeniu (*overfitting*)

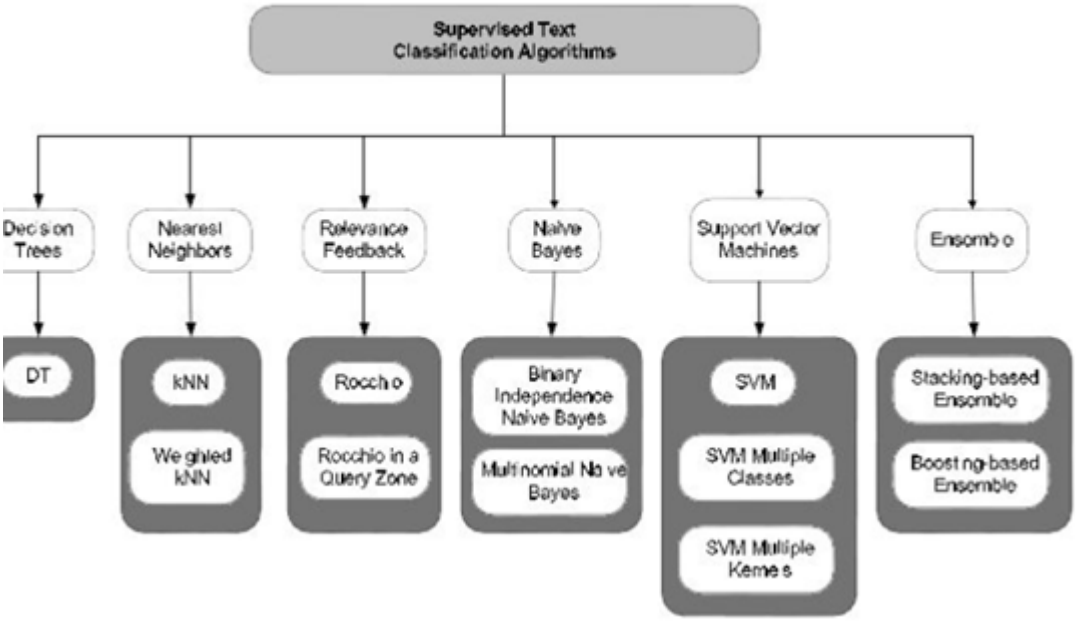
- Klasyfikator podlega ocenie jakości (walidacja krzyżowa)

W metodach nienadzorowanych etykiety klas są generowane automatycznie poprzez ustalenie środków skupień danych w odpowiednio dobranej przestrzeni atrybutów (termów) opisujących dokumenty

- Dokładniejsze są algorytmy nadzorowane:
  - jednoetykietowe – pojedyncza klasa przypisana do dokumentu,
  - wieloetykietowe – jedna lub więcej klas przypisanych do dokumentu

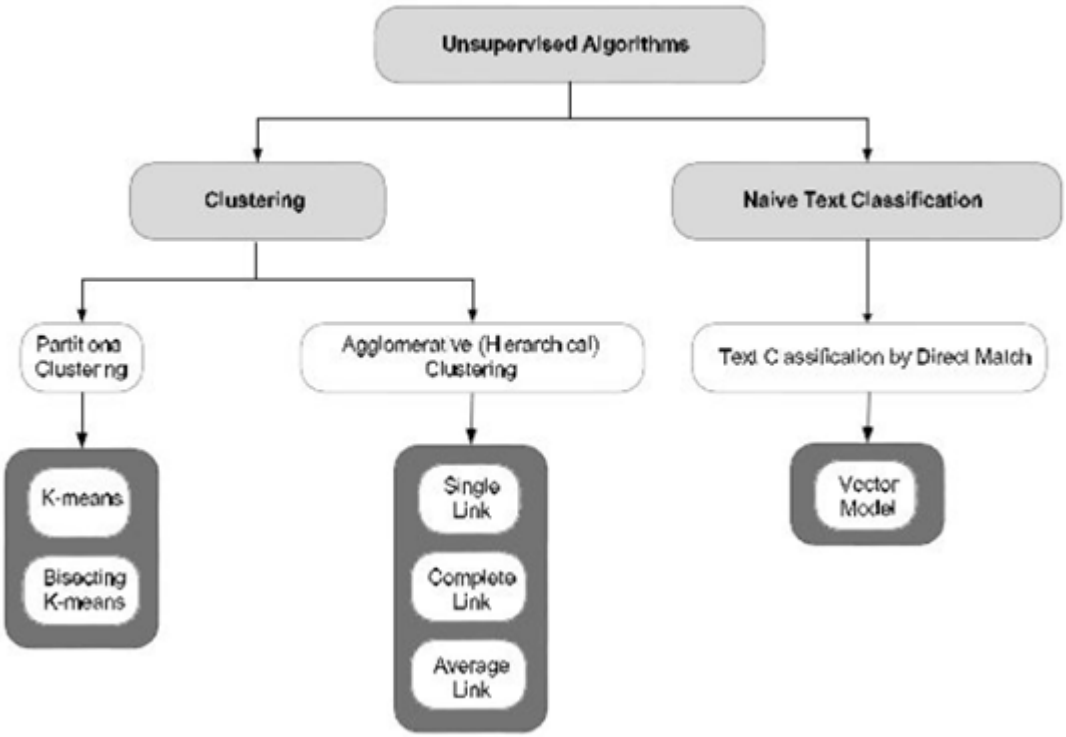
32. Wymień 3 metody klasyfikacji z nadzorem.

- Nadzorowane algorytmy klasyfikacji



33. Wymień 3 metody klasyfikacji bez nadzoru.

- Algorytmy nienadzorowane



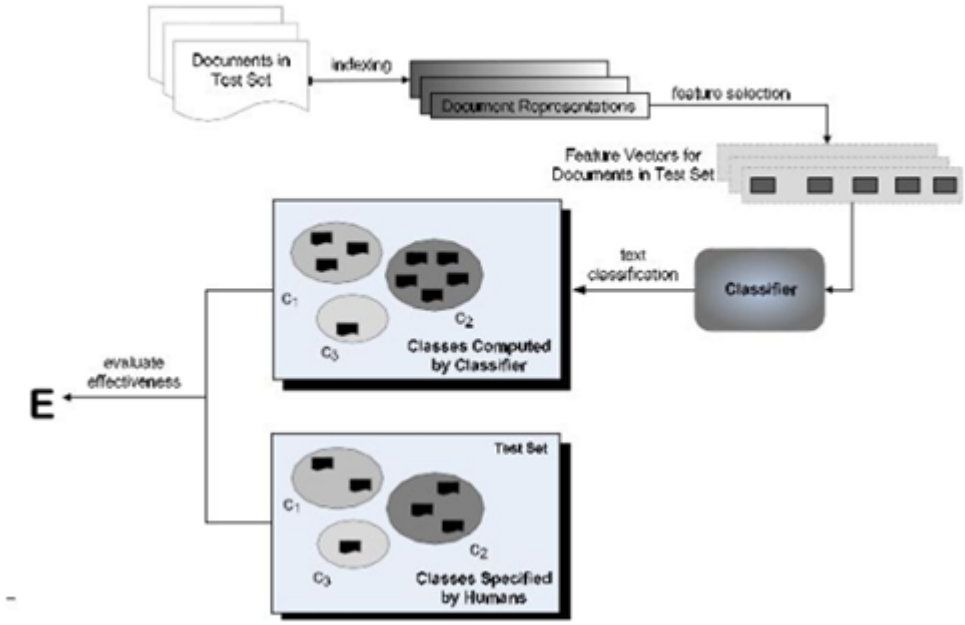
34. Jak nazywa się algorytm, dla którego prawdziwe jest: W metodach nienadzorowanych etykiety klas są generowane automatycznie poprzez ustalenie środków skupień danych w odpowiednio dobranej przestrzeni atrybutów (termów) opisujących dokumenty. Wejście algorytmu– zadana liczba K klastrów.

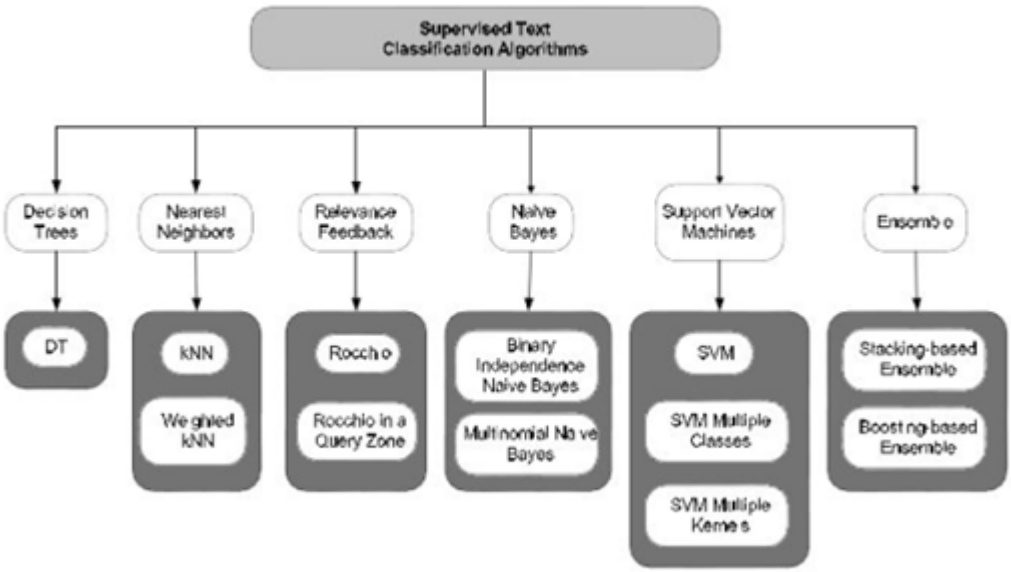
Klasyfikacja K-średnich

- W metodach nienadzorowanych etykiety klas są generowane automatycznie poprzez ustalenie środków skupień danych w odpowiednio dobranej przestrzeni atrybutów (termów) opisujących dokumenty
- Wyniki mogą być czasami niezadowalające lub różne od oczekiwanych przez użytkownika
- Metoda K-średnich:
  - Wejście – zadana liczba K klastrów,
  - Każdy klaster jest reprezentowany przez centroidę dokumentów,
  - Algorytm:
    - Przypisanie dokumentu do najbliższej centroidy,
    - Przeliczenie centroid,
    - Powtórzenie poprzednich kroków aż do stabliizacji położenia centroid

35. Jak weryfikowana jest efektywność w algorytmach nadzorowanych? (wykład 5; slajd23)

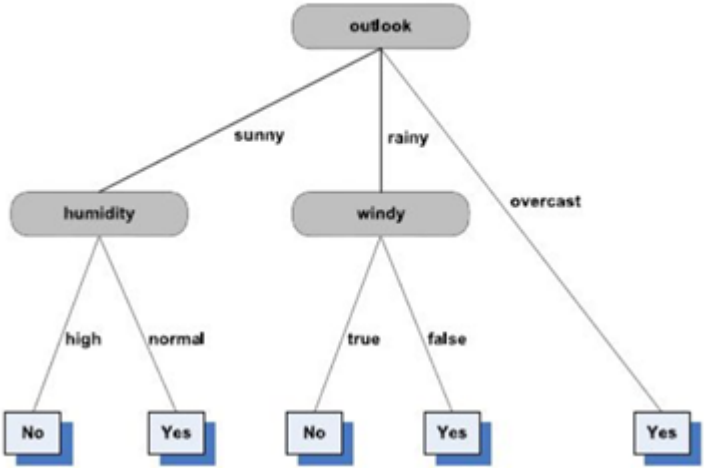
- Klasyfikacja i ocena





36. Jak działa drzewo decyzyjne (wykład 5; slajd 26)

- Przewidywanie atrybutu Play



- Węzły wewnętrzne → nazwy atrybutów
- Krawędzie → wartości atrybutów
- Trawers drzewa decyzyjnego → wartość atrybutu "Play".
- (Outlook = sunny) ∧ (Humidity = high) → (Play = no)

	Id	Play	Outlook	Temperature	Humidity	Windy
Test Instance	11	?	sunny	cool	high	false

- Predykcje bazują na zadanych przykładach
- Nowe przykłady naruszające wzorzec prowadzą do błędnych przewidywań
- Przykładowa baza danych stanowi zbiór uczący, determinujący drzewo decyzyjne DT (Decision Tree)

37. Jaka jest różnica pomiędzy algorytmem k-means, a k-NN?

Keywords:

Information retrieval IR, searching, sensemaking, crawling , query, berry picking, Term expansion, Reprezentacja fasetowa, Term indeksujący, słownik termów, wektor dokumentu, TF-IDF, racall, precision, P@5, k-means, a k-NN