POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO-SIECIOWE

Automatyczne grupowanie wyników wyszukiwania w bibliotekach cyfrowych

Adam Dudczak

Poznańskie Centrum Superkomputerowo-Sieciowe

IV Warsztaty “Biblioteki Cyfrowe”

Poznań, 2007

Plan prezentacji

- Problemy związane z wyszukiwaniem informacji
- Na czym polega grupowanie wyników ?
- Przykłady zastosowania automatycznego grupowania wyników
- Grupowanie wyników w bibliotekach cyfrowych
- Podsumowanie

Wyszukiwanie - problemy

- Skąd wiedzieć jak zapytać?
- Zapytania bardzo ogólne
  - najczęściej **1,2 wyrazy**
- **Ogólne zapytanie = dużo wyników**
- Jak sobie z tym poradzić?

Na czym polega grupowanie?

- Odnaleźć grupy dokumentów o podobnej treści
- Nadać grupom nazwy dobrze charakteryzujące jej elementy składowe
- Jeden dokument może znajdować się w kilku grupach

Na czym polega grupowanie?

- Każdy dokument to zbiór:
  - wyrazów, znaków interpunkcyjnych
- Może też zawierać treść nie tekstową:
  - zdjęcia, grafiki ...
- Musimy określić jakie cechy dokumentu są dla nas istotne
- Jak będziemy określać podobieństwo dokumentów?

Na czym polega grupowanie?

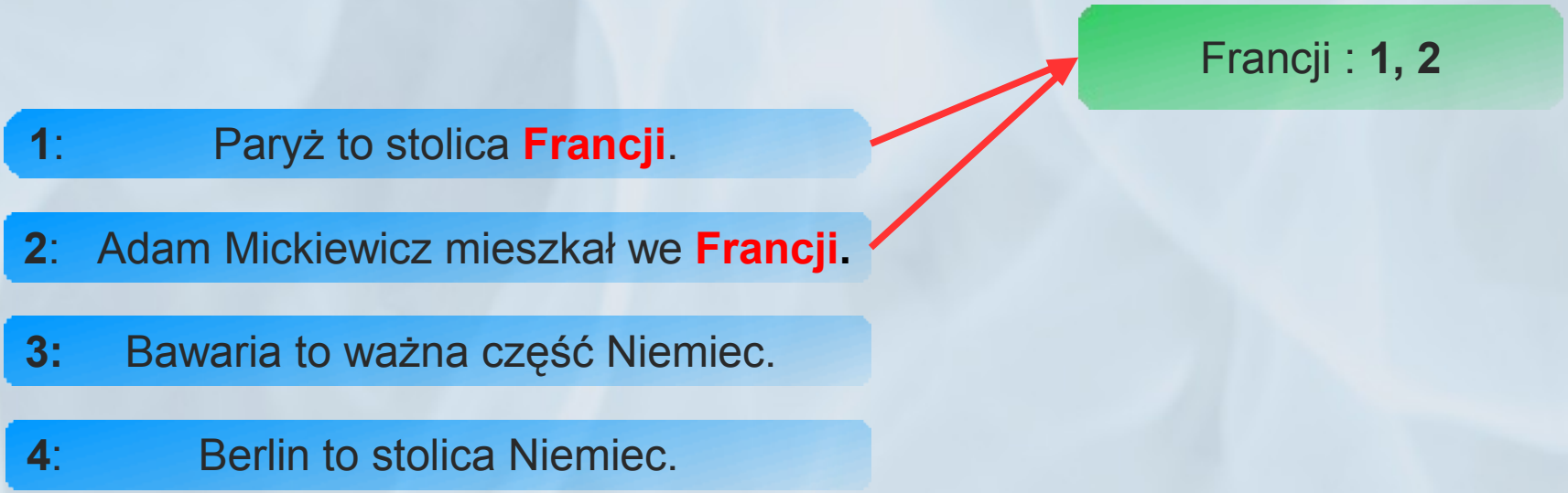
1: Paryż to stolica Francji.

2: Adam Mickiewicz mieszkał we Francji

3: Bawaria to ważna część Niemiec.

4: Berlin to stolica Niemiec.

Na czym polega grupowanie?



Na czym polega grupowanie?

1: Paryż to stolica **Francji**.

2: Adam Mickiewicz mieszkał we **Francji**.

3: Bawaria to ważna część Niemiec.

4: Berlin to stolica Niemiec.

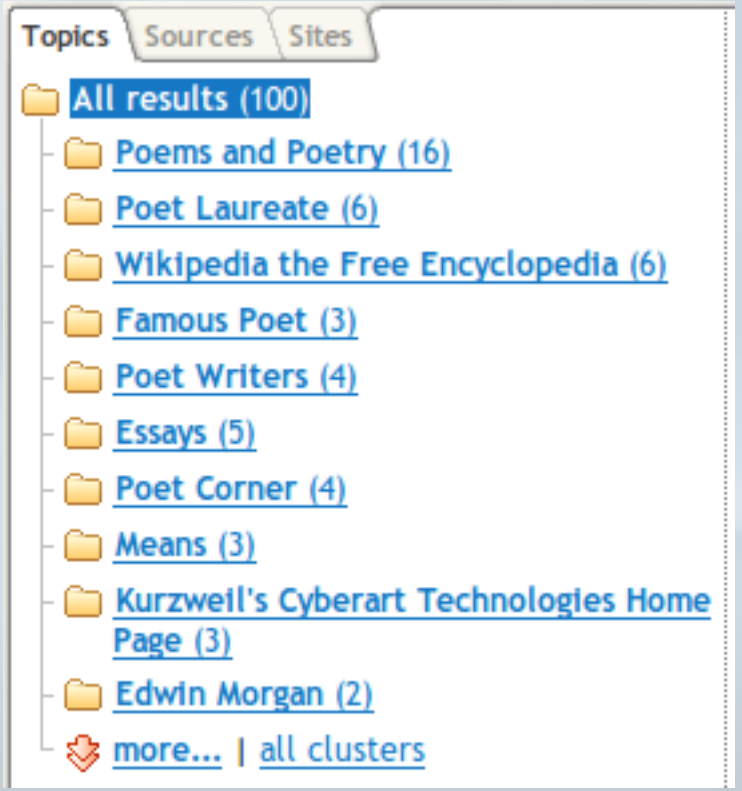
Francji : **1, 2**

Niemiec : **3, 4**

Na czym polega grupowanie?

Na czym polega grupowanie?

- Rodzaje grupowania
  - **płaskie**
  - hierarchiczne



Na czym polega grupowanie?

- Rodzaje grupowania
  - płaskie
  - **hierarchiczne**

Przykłady zastosowania

- Grupowanie wyników wyszukiwarek
- **Clusty** : http://clusty.com
- **Carrot^2** : http://carrot2.org
  - produkt polski :)
  - wyszukiwarka
  - narzędzie do grupowania wyników o otwartym kodzie źródłowym

Grupowanie wyników w bibliotekach cyfrowych

- Tworzenie grup na podstawie opisów bibliograficznych
  - tylko najważniejsze informacje
  - kontrolowane słowniki wartości

Grupowanie wyników w bibliotekach cyfrowych

- Wyniki pierwszych eksperymentów są obiecujące
  - wykorzystanie Carrot^2
  - danych z indeksów portalu FBC
- Grupowanie
  - w oparciu o jeden atrybut (przykład prosty)
  - w oparciu o wiele atrybutów
- Grupowanie z wykorzystaniem treści publikacji

Problemy, wyzwania, dalsze prace...

- Grupowanie po datach
- Ujednolicenie wartości w opisach bibliograficznych
- Dostępność warstwy tekstowej i jej jakość
- Nacisk na zagadnienia związane z przetwarzaniem języka polskiego
- Wdrożenie na FBC i w dLibrze 5.0?

Podsumowanie

- Zastosowanie automatycznego grupowania wyników może znacznie ułatwić wyszukiwanie zasobów
- Dzięki wysokiej jakości opisów bibliograficznych istnieje duża szansa, że algorytmy grupowania sprawdzą się w bibliotekach cyfrowych
- Trzeba jednak pamiętać iż metody automatyczne to „tylko” metody automatyczne

Podziękowania

Twórcom projektu Carrot^2 :

- p. Dawidowi Weissowi
- p. Stanisławowi Osińskiemu

http://carrot2.org

POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO-SIECIOWE

Dziękuję za uwagę!