XCleaner: Grupowanie dokumentów XML według struktury

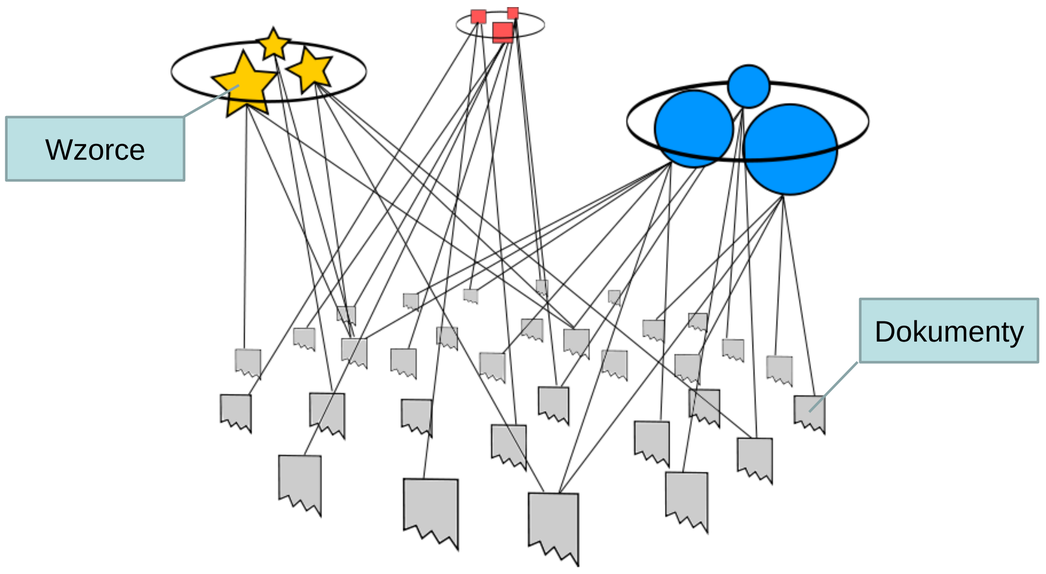
D. Brzeziński A. Leśniewska T. Morzy M. Piernik
Instytut Informatyki, Politechnika Poznańska

Motywacja

- Grupowanie pozwala między innymi:
  - Odkrywać cechy charakterystyczne zbiorów obiektów
  - Zredukować dużą liczbę danych do kilku kategorii
  - Porównywać obiekty wielokryterialnie
- Wszechobecność formatu XML
- Grupowanie XML to:
  - Grupowanie adnotacji do zdjęć
  - Grupowanie usług typu webservice
  - Tematyczne wyszukiwanie dokumentów
  - Grupowanie podobnych struktur w bioinformatyce
- **Grupowanie XML różni się od grupowania dokumentów tekstowych**
- **Wykorzystanie struktury dokumentu XML poprawia trafność grupowania w stosunku do korzystania z samej tylko treści dokumentu**

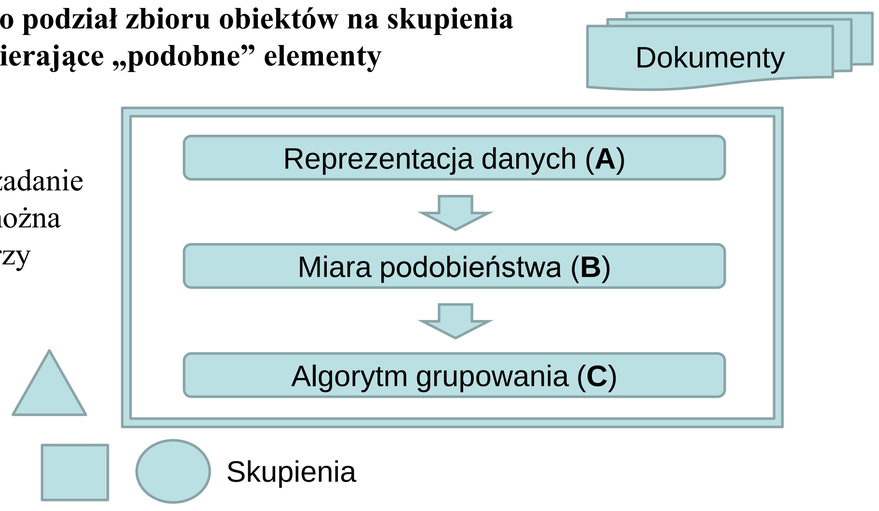
Pomysł

Grupować dokumenty według wzorców (struktur) częstych zawartych w dokumentach



Rozwiązanie

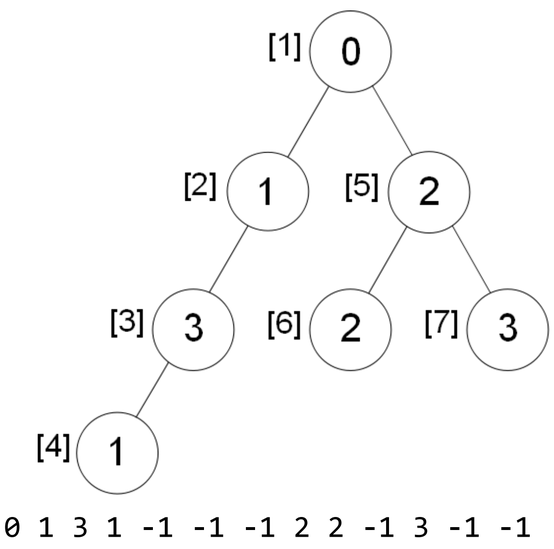
Grupowanie to podział zbioru obiektów na skupienia zawierające „podobne” elementy



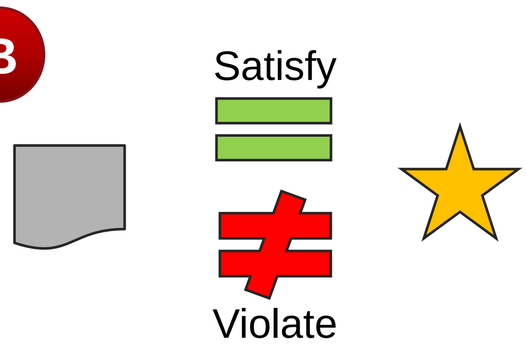
W ogólności zadanie grupowania można podzielić na trzy etapy

A

- Reprezentacja dokumentów XML jako drzewa
- Maksymalne poddrzewa częste dokumentów jako wzorce
- Algorytm CMTreeMiner do wyszukiwania poddrzew zaproponowany przez Yun Chi
- Problem NP-zupełny:
  - problemy z uzyskiwaniem poddrzew częstych dla „szerokich” dokumentów
  - główny problem to pamięć



B

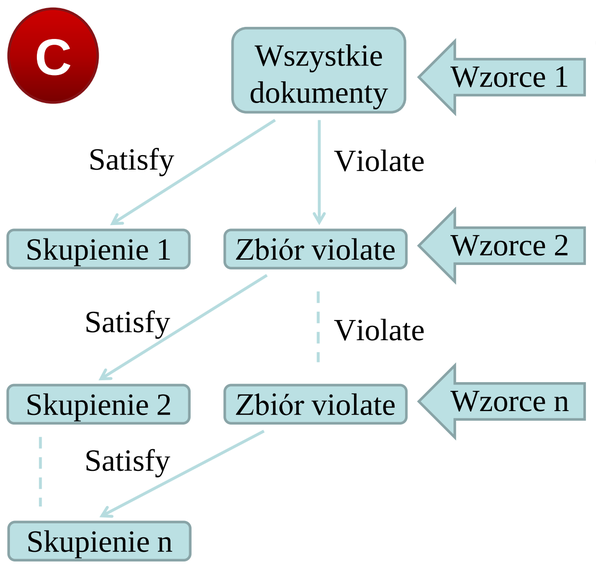


Satisfy/Violate - idea

Operator wybiera ze zbioru obiektów te, które spełniają (Satisfy) lub nie spełniają (Violate) podanego warunku

- Zamiast typowej miary podobieństwa zaproponowano operator *Satisfy/Violate*
- Operator wybiera ze zbioru dokumentów te, które zawierają podany wzorzec (Satisfy) lub go nie zawierają (Violate)
- Wzorce grupowane są w skupienia przy pomocy algorytmu AHC
- Miara podobieństwa to liczba dokumentów, w których wzorce współwystępują

C



- Pojedyncze skupienie dokumentów reprezentowane przez skupienie wzorców
- Dokumenty grupowane przez wywoływanie operatora Satisfy na kolejnych skupieniach wzorców dla dokumentów należących do zbioru Violate z poprzedniego kroku (w pierwszym kroku na wszystkich dokumentach)

Wyniki

Eksperymenty

- *Precision* i *recall* jako miary oceny jakości uzyskanych skupień
- 1 rzeczywisty i 3 sztuczne zbiory danych
- Porównanie z 2 innymi algorytmami
- Problem doboru minsup i maxsup
  - dla danych heterogenicznych: minsup = 1/k, maxsup = 1
  - dla danych homogenicznych: minsup = 1/k, maxsup ≈ minsup

Zbiór danych	Precision			Recall		
	TagOnly	Xproj	Xcleaner	TagOnly	Xproj	Xcleaner
SIGMOD	1,00	1,00	**1,00**	1,00	-	**1,00**
Heterogeneous_3	0,56	1,00	**1,00**	0,56	1,00	**1,00**
Heterogeneous_6	0,52	1,00	**1,00**	0,52	1,00	**1,00**
Homogeneous_OT	0,51	1,00	**1,00**	0,51	1,00	**0,90**
Homogeneous	0,51	1,00	**1,00**	0,51	1,00	**1,00**

Dalsze badania

- Poszukiwanie nowych typów wzorców
  - Uzyskanie podobnej jakości grupowania dla struktur lżejszych niż maksymalne poddrzewa częste
- Nowe metody pozyskiwania wzorców
- Testowanie rozwiązań w ramach konkursu INEX

Podsumowanie

- Wykorzystanie maksymalnych poddrzew częstych jako wzorców do grupowania dokumentów XML
- Operator Satisfy-Violate do grupowania wzorców
- Przypisywanie dokumentów do wzorców na prostej zasadzie występowania wzorców
- Trudności z pozyskiwaniem wzorców dla dużych zbiorów danych

Bibliografia

1) T. Dalamagas, T. Cheng, K.-J. Winkel, and T. K. Sellis, „Clustering XML documents by structure," in SETN (G. A. Vouros and T. Panayiotopoulos, eds.), vol. 3025 of Lecture Notes in Computer Science, pp. 112-121, Springer, 2004.

2) C. C. Aggarwal, N. Ta, J. Wang, J. Feng, and M. J. Zaki, „Xproj: a framework for projected structural clustering of XML documents," in KDD (P. Berkhin, R. Caruana, and X. Wu, eds.), pp. 46-55, ACM, 2007.

3) M. J. Zaki, „Eciently mining frequent trees in a forest," in KDD, pp. 7180, ACM, 2002.

4) Y. Chi, Y. Xia, Y. Yang, and R. R. Muntz, „Mining closed and maximal frequent subtrees from databases of labeled rooted trees," IEEE Trans. Knowl. Data Eng., vol. 17, no. 2, pp. 190-202, 2005.

5) A. Doucet and H. Ahonen-Myka, „Naive clustering of a large XML document collection" in INEX Workshop (N. Fuhr, N. Giovert, G. Kazai, and M. Lalmas, eds.), pp. 81-87, 2002.