Aspekty ekstrakcji autorów na podstawie metadanych dokumentów

Łukasz Bolikowski
Piotr Jan Dendek

Interdyscyplinarne Centrum
Modelowania Matematycznego i Komputerowego,
Uniwersytet Warszawski

Spotkanie
6 lipca 2011

Agenda

- Utwór, czyli co?
  - utwór muzyczny
  - film
  - książka
  - artykuł naukowy
  - magazyn

- Forma opisu
  - katalog papierowy
  - XML
  - BibTex
- Jak powstaje opis?
  - ręcznie
    - wprowadzany przez autorów
    - weryfikowany przez pracowników bibliotecznych
  - automatycznie (pozyskiwany z dokumentu)

- Cele tworzenia opisów
  - fizyczne przechowywanie danych w sposób uporządkowany (potencjalnie użytkownik sięga po dane związane ze sobą)
  - ułatwienie dostępu użytkownikom (stworzenie jak największej ilości wysoce przydatnych kryteriów wyszukiwania)
  - agregowanie danych
    - odzyskiwanie (przybliżone) brakujących informacji z metadanych
    - tworzenie rankingu publikacji
    - tworzenie rankingu autorów
    - etc.

- Jak wygląda nagłówek artykułu naukowego?

Two Supervised Learning Approaches for Name Disambiguation in Author Citations

Hui Han
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA, 16802
hhan@cse.psu.edu

Lee Giles
School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16802
giles@ist.psu.edu

Hongyuan Zha
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA, 16802
zha@cse.psu.edu

Cheng Li
Department of Biostatistics
Harvard School of Public Health
Boston, MA, 02115
cli@hsph.harvard.edu

Kostas Tsioutsiouliklis
NEC Laboratories America, Inc.
4 Independence Way,
Princeton, NJ 08540
kt@nec-labs.com

ABSTRACT

Due to name abbreviations, identical names, name misspellings, and pseudonyms in publications or bibliographies (citations), an author may have multiple names and multiple authors may share the same name. Such name ambiguity affects the performance of document retrieval, web search, database integration, and may cause improper attribution to authors. This paper investigates two supervised learning approaches to disambiguate authors in the ci-

Keywords

Naive Bayes, Name Disambiguation, Support Vector Machine

1. INTRODUCTION

Due to name variation, identical names, name misspellings, and pseudonyms, we observe two types of name ambiguities in research papers or bibliographies (citations). The first type is that an author has multiple name labels. For example, the author "David S. John-

- Jak wygląda zakończenie artykułu naukowego?

6. ACKNOWLEDGMENTS

We would like to acknowledge partial support from NSF Grant 0121679, CCF-0305879, and helpful comments from reviewers.

7. REFERENCES

[1] Getty's ULAN (Union List of Artist's Names). http://www.getty.edu/research/conducting_research/vocabularies/ulan/.

[2] The library of congress name authority file. http://www.loc.gov/marc/authority/index.html.

[3] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, 1998.

[4] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Generative model-based clustering of directional data. In *Proceedings of The 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD)*, pages 19–28, 2003.

[5] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*.

Intelligence (ECAI), 2000.

[17] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[18] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, pages 89–98, 1998.

[19] P. Gillman. National name authority file: Report to the national council on archives. Technical Report British Library Research and Innovation Report 91, The British Library Board, 1998.

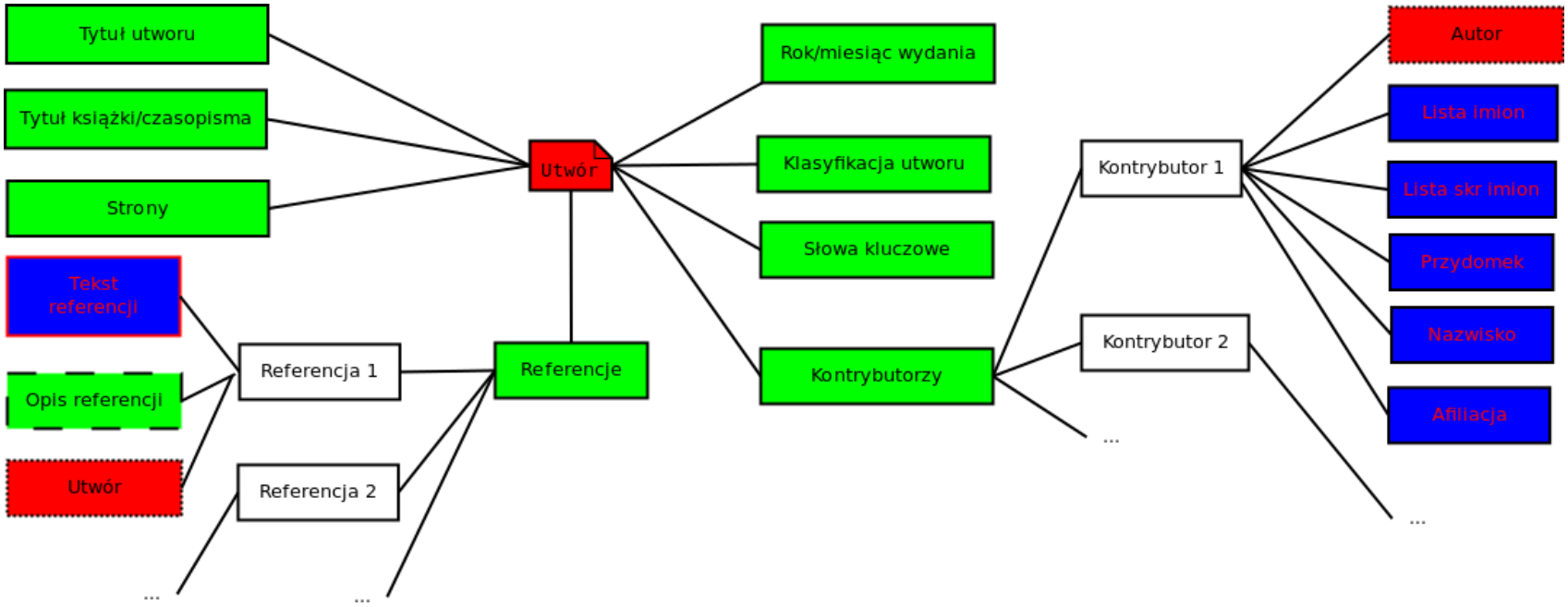
[20] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries*, pages 37–48, 2003.

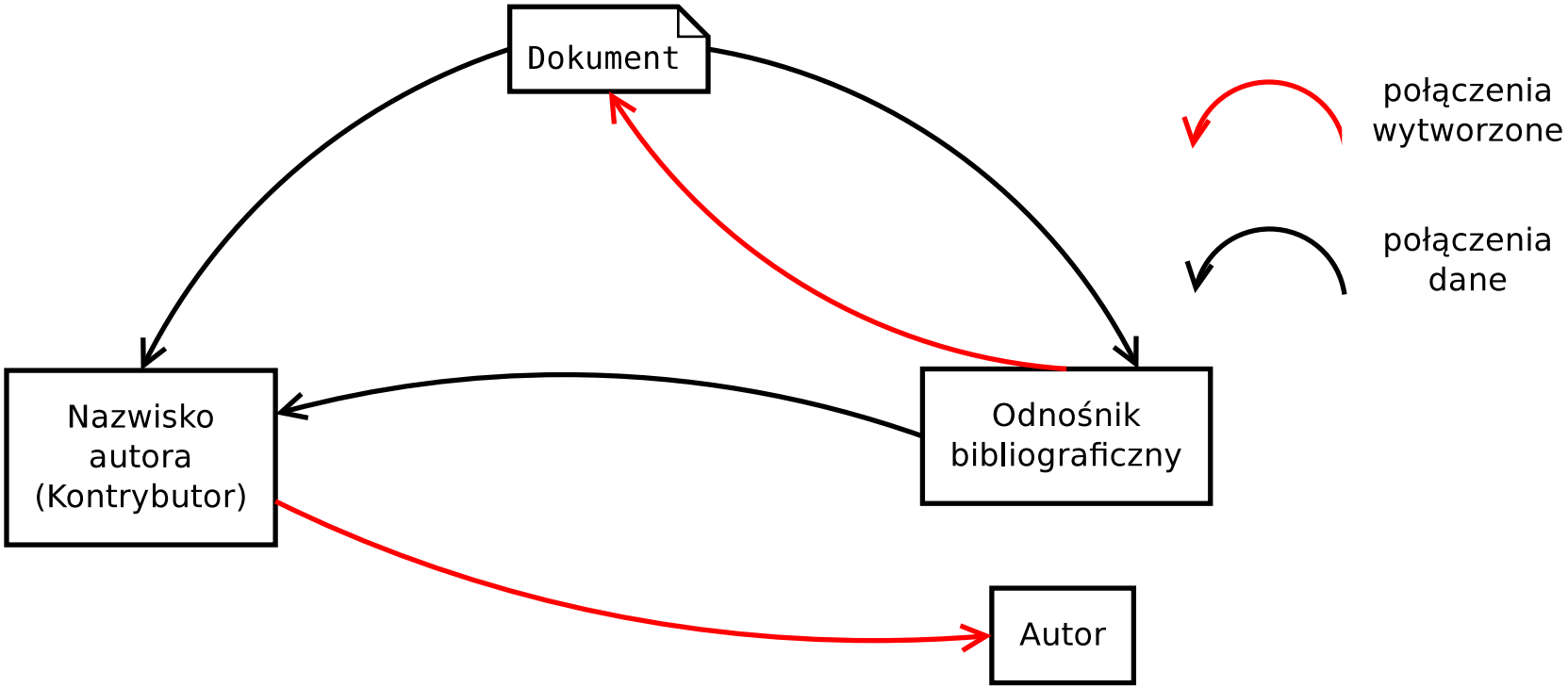
[21] H. Han, H. Zha, and C. L. Giles. A model-based k-means algorithm for name disambiguation. In *Proceedings of the 2nd International Semantic Web Conference (ISWC-03) Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, 2003.

[22] M. A. Hernandez and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9–37, 1998.

[23] T. Hofmann. Probabilistic latent semantic analysis. In

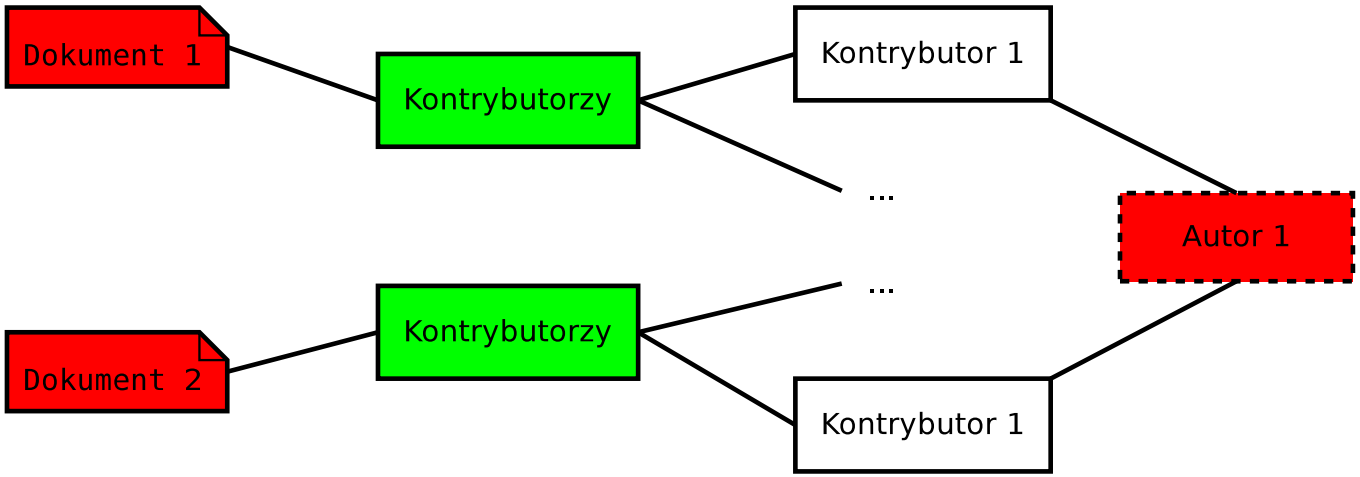
- Zawartość opisu
  - twórca/twórcy
  - wydawnictwo
  - miesiąc/rok publikacji
  - strony na których widnieje w publikacji
  - klasyfikacja dokumentu (PACS, CLC, CEJSH)
  - słowa kluczowe
  - referencje
- Jest kilka równorzędnych notacji na zapisanie w/w
- W opisie występują błędy


Tytuł utworu
Tytuł książki/czasopisma
Strony
Tekst referencji
Opis referencji
Utwór
Referencja 1
Referencja 2
...
...
Utwór
Referencje
Rok/miesiąc wydania
Klasyfikacja utworu
Słowa kluczowe
Kontrybutorzy
Kontrybutor 1
Kontrybutor 2
...
Autor
Lista imion
Lista skr imion
Przydomek
Nazwisko
Afiliacja
...



- Powiązania między dokumentami/autorami
  - jawne (np. połączenie grafów 2 dokumentów)
  - niejawne (przez nieprzetworzoną referencję)

- Cel - Odzyskanie informacji o powiązaniu autor-dzieło dzięki grafom dokumentów



- "dzięki grafom dokumentów" - tzn.?
  - wykorzystując podobieństwa w grafach poszczególnych dokumentów stworzyć grupy kontrybutorów tworzących podobnie
  - podobieństwo w grafach - wskazówka

- Przykłady wskazówek
  - Te same nazwiska autorów dokumentów
  - Ta sama dziedzina problemu
  - Zbliżony rok powstania pracy

$$\textbf{CzasPowstania}(c_1, c_2) \text{ (ciągły)} = \begin{cases} 0 & rok(c_1) = \bot \\ & \vee\ rok(c_2) = \bot \\ -1 & |rok(c_1) - rok(c_2)| > 70 \\ \left(\frac{rok(c_1) - rok(c_2)}{70}\right)^2 & w\ p.p. \end{cases}$$

$$\textbf{CzasPowstania}(c_1, c_2) \text{ (dyskretny)} = \begin{cases} 0 & rok(c_1) = \bot \vee rok(c_2) = \bot \\ -1 & |rok(c_1) - rok(c_2)| > 70 \\ 1 & w\ p.p. \end{cases}$$

$$\textbf{Czasopismo}(c_1, c_2) = \begin{cases} 0 & czasomismo(c_1) = \bot \vee czasomismo(c_2) = \bot \\ 1 & czasomismo(c_1) = czasomismo(c_2) \\ -0.1 & w\ p.p. \end{cases}$$

$$\textbf{Email}(c_1, c_2) = \begin{cases} 0 & email(c_1) = \bot \vee email(c_2) = \bot \\ 1 & email(c_1) = email(c_2) \\ -0.1 & w\ p.p. \end{cases}$$

$$\textbf{Język}(c_1, c_2) = \begin{cases} 0 & język(c_1) = \perp \vee język(c_2) = \perp \\ 0.05 & język(c_1) = eng \vee język(c_2) = eng \\ 0.1 & język(c_1) = język(c_2) \\ -1 & w\ p.p. \end{cases}$$

$$\textbf{SłowaKluczowe}(c_1, c_2) \text{ (dyskretny)} = \begin{cases} 0 & słowakluczowe(c_1) = \emptyset \vee słowakluczowe(c_2) = \emptyset \\ -1 & \frac{|słowakluczowe(c_1) \cap słowakluczowe(c_2)|}{|słowakluczowe(c_1) \cup słowakluczowe(c_2)|} < 0.25 \\ 1 & w\ p.p. \end{cases}$$

$$\textbf{SłowaKluczowe}(c_1, c_2) \text{ (ciągły)} = \begin{cases} 0 & słowakluczowe(c_1) = \emptyset \\ & \vee słowakluczowe(c_2) = \emptyset \\ \frac{|słowakluczowe(c_1) \cap słowakluczowe(c_2)|}{|słowakluczowe(c_1) \cup słowakluczowe(c_2)|} * 2 - 1 & w\ p.p. \end{cases}$$

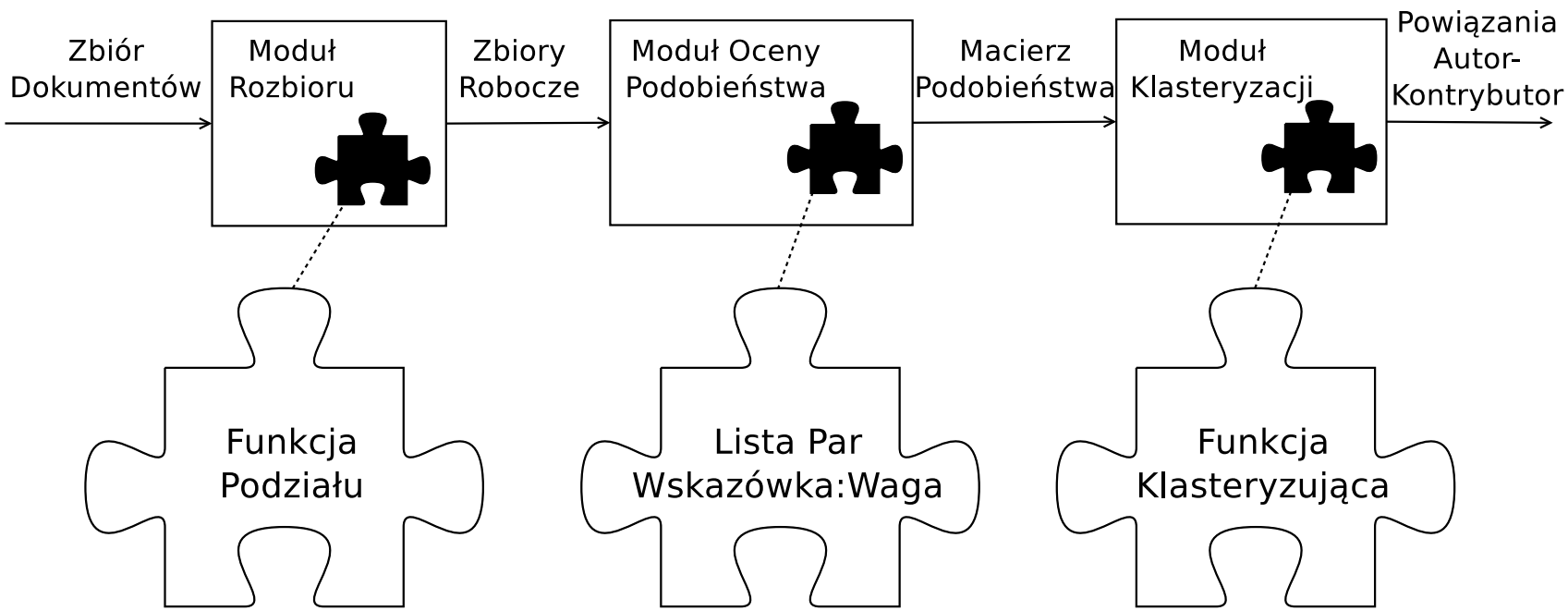
$$\textbf{Samocytowanie}(c_1, c_2) = \begin{cases} 1 & nazwisko - kanoniczne(c_1) = nazwisko - kanoniczne(reference(c_1)) \\ 0 & w\ p.p. \end{cases}$$

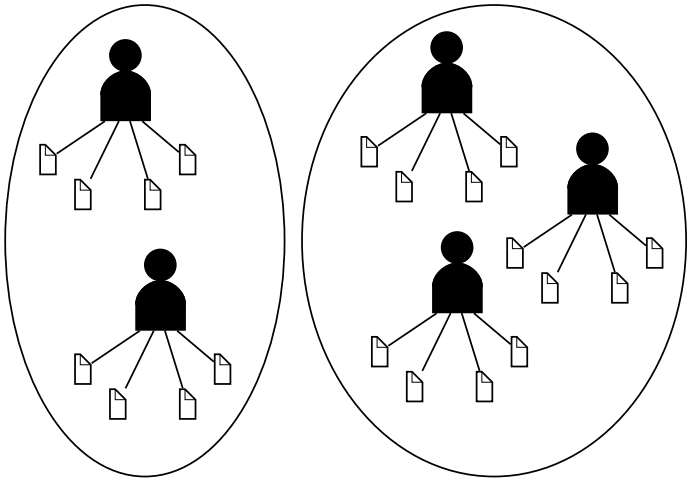
$$\textbf{Współpraca}(c_1, c_2) = \begin{cases} 0.7 & |współautor(c_1) \cap współautor(c_2)| = 1 \\ 1 & |współautor(c_1) \cap współautor(c_2)| > 1 \\ 0 & w\ p.p. \end{cases}$$

- Cechy wskazówki
  - Ciągła/Dyskretna
  - Wyważona/Spolaryzowana
  - Płaska/Głęboka

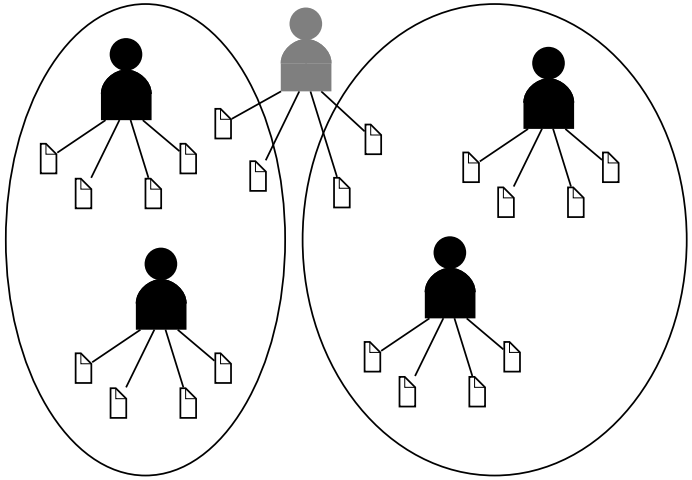
- Warunki wstępne ekstrakcji danych o autorstwie:
  1. posiadanie zbadanych, ważnych wskazówek
  2. przyporządkowanie wskazówkom wag

- Etapy ekstrakcji danych:
  1. Import danych
  2. Podział zbioru grafów na podzbiory
  3. Wyznaczenie podobieństw autorów dokumentów
  4. Klastrowanie
  5. Utrwalenie uzyskanych informacji


Zbiór Dokumentów
Moduł Rozbioru
Zbiory Robocze
Moduł Oceny Podobieństwa
Macierz Podobieństwa
Moduł Klasteryzacji
Powiązania Autor-Kontrybutor
Funkcja Podziału
Lista Par Wskazówka:Waga
Funkcja Klasteryzująca



Kontrybucje (ikony dokumentów) tego samego autora (ikona człowieka) powinny znaleźć się w tym samym zbiorze roboczym (elipsa).

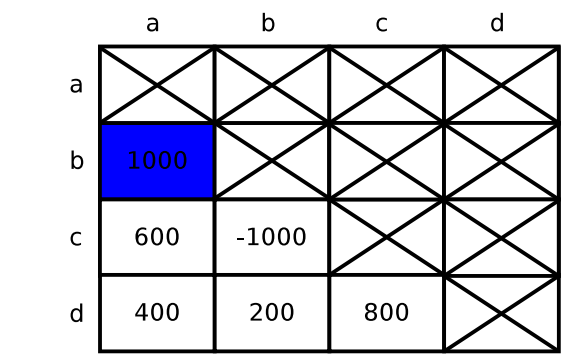


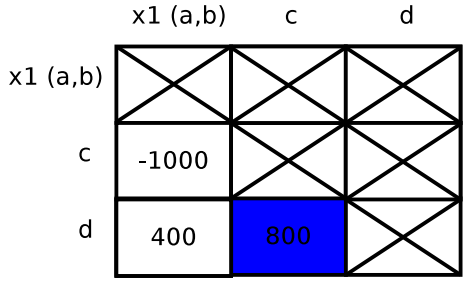
Źle skonstruowana funkcja podziału umieszcza kontrybucje tego samego autora w różnych zbiorach roboczych (wyszarzona ikona człowieka).

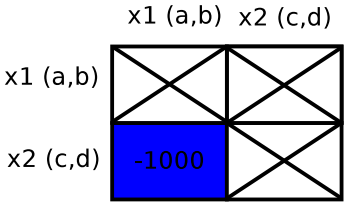
1. Weź 2 najpodobniejsze, aktywne obiekty
2. Jeśli ich podobieństwo jest niższe od zadanego progu przerwij procedurę i zwróć wynik
3. Oznacz jeden z obiektów jako nieaktywny
4. Wybrane wcześniej obiekty połącz w klaster i przelicz ich podobieństwa do innych obiektów wg. poniższego wzoru:

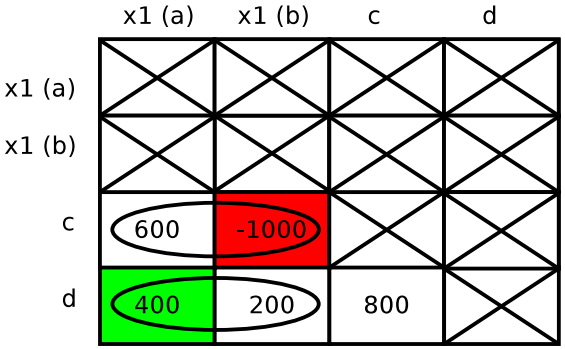
$$\underset{\substack{1<i<N,\\ i\neq a,\\ i\neq b}}{\forall} podob(c_a, c_i) = podob(c_b, c_i) = \begin{cases} -\infty & podob(c_a, c_i) < próg \\ & \vee\ podob(c_b, c_i) < próg \\ podob(c_a, c_i) & podob(c_a, c_i) > podob(c_b, c_i) \\ podob(c_b, c_i) & podob(c_a, c_i) \leqslant podob(c_b, c_i) \end{cases}$$

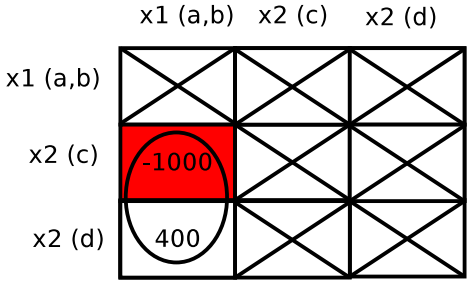
5. Przejdź do kroku pierwszego

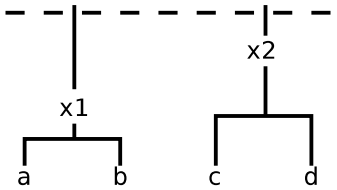

a
b
c
d
1000
600
-1000
400
200
800

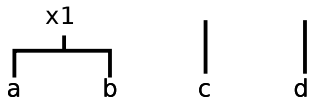

x1 (a,b)
c
d
x1 (a,b)
c
d
-1000
400
800

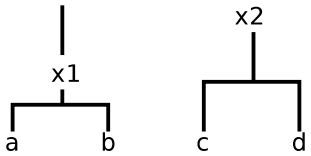

x1 (a,b)
x2 (c,d)
x1 (a,b)
x2 (c,d)
-1000


x1 (a)
x1 (b)
c
d
x1 (a)
x1 (b)
c
d
600
-1000
400
200
800


x1 (a,b)
x2 (c)
x2 (d)
x1 (a,b)
x2 (c)
x2 (d)
-1000
400


x2
x1
a
b
c
d


x1
a
b
c
d

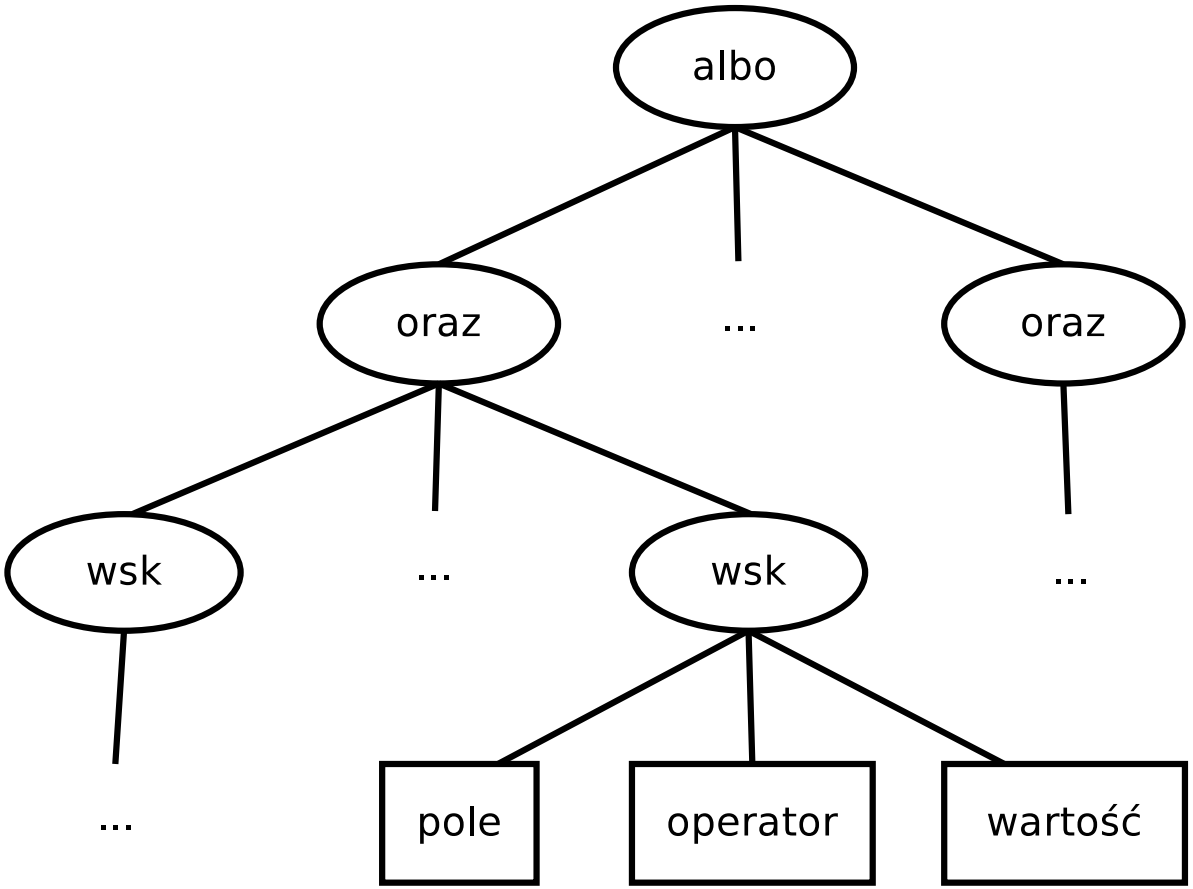

x1
x2
a
b
c
d

- W opisanym procesie pojawiają się problemy
  - wyboru istotnych wskazówek
    - przez "sprytną obserwację"
    - przez automatyczne wytworzenie
  - doboru wag dla wskazówek

- ...które trzeba rozwiązać
  - baza danych posiadająca powiązania autor-dzieła (np. Zentralblatt)

- Generacja wskazówek
  - baza zawierająca połączenia pewne
  - algorytm wytwarzania
    - Apriori
    - Programowanie genetyczne
    - Sprytna obserwacja

- sposób wytwarzania wskazówki
  - słowa kluczowe
    - zawiera *t*
    - zawiera to samo słowo kluczowe
  - rok
    - powstało w tym samym roku
    - różnica w dacie publikacji mniejsza niż 3 lata
  - *pole*
    - *dopuszczalne akcje* (wraz z zakresem, np. mniej niż [0,200] lat)


albo
oraz
...
oraz
wsk
...
wsk
...
...
pole
operator
wartość

- Efekty przy zastosowaniu programowania genetycznego
  - "Using Genetic Programming to Evaluate the Impact of Social Network Analysis in Author Name Disambiguation"
  - Levin, F.H., Heuser, C.A.

$$
\begin{aligned}
\textbf{Levin-Heuser}(c_1, c_2) \quad = \quad & (IsAbbrev = 1 \wedge NameSim \geqslant 0.94 \wedge MinRQ \leqslant 1) \\
& \vee (TitleSim \geqslant 0.39 \wedge MaxRQ \leqslant 5 \wedge NameSim \geqslant 0.87 \\
& \quad \wedge NumTitleWords \geqslant 2 \wedge IniLastName = 1) \\
& \vee (TitleWordSim \geqslant 0.23 \vee NameSim \geqslant 0.87 \wedge IsAbbrev = 1 \\
& \quad \wedge VenueSim \geqslant 0.35 \wedge MaxRQ \leqslant 3) \\
& \vee (VenueSim \geqslant 0.67 \wedge MaxRQ \leqslant 3 \wedge NameSim \geqslant 0.98) \\
& \vee (IsAbbrev = 1 \wedge NameSim \geqslant 0.45 \wedge NumTitleWords \geqslant 4 \\
& \quad \wedge MinRQ \leqslant 2 \wedge IniLastName = 1 \wedge RE = 0) \\
& \vee (IniLastName = 1 \wedge RE = 1 \wedge MaxRQ \leqslant 9) \\
& \vee (RE = 1 \wedge NameSim \geqslant 0.72) \\
& \vee (RS \geqslant 2 \wedge IniLastName = 1)
\end{aligned}
$$

- 78,5 %

- Różnicowanie wpływu wskazówki na ostateczną całościową ocenę podobieństwa
- Wskazówka idealna
  - Wskazująca TYLKO kontrybutorów będących tą samą osobą (email?)
  - Nie wskazująca na kontrybutorów będących różnymi osobami (niestety nie...)

- Metody doboru wag
  - przy generacji, metodą Apriori - poziom ufności
  - metodą AdaBoost, względem TP,TN,FP,FN,F1

- Występuje problem braku danych w metadanych dokumentów (m.in. o autorstwie)
- Do rozwiązania problemu wykorzystać można podobieństwo dokumentów (wskazówki)
- Wskazówki trzeba wytworzyć i zbadać

- Inkrementacyjne rozszerzanie danych o autorstwie
- Wykorzystanie wiedzy eksperckiej ($1,2\%$ kontrybutorów o różnym nazwisku jest tym samym autorem)
- Nakładanie się zbiorów roboczych (podzbiorów kontrybutorów)
- Transliteracje

Dziękuję

Dziękuję! — Pytania?