Tensorowe przetwarzanie zbiorów danych o dużych rozmiarach[1]

Andrzej Szwabe, Paweł Misiorek, Michał Ciesielczyk

Instytut Automatyki i Inżynierii Informatycznej
Politechnika Poznańska (PP)
Piotrowo 3a, 60-965 Poznań
Email: {imie.nazwisko}@put.poznan.pl

22 kwietnia 2016

[1] Wyniki prac prowadzonych w ramach projektu sfinansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/01/D/ST6/06788

Plan Prezentacji

- Motywacja do badań nad metodami tensorowymi w kontekście Big Data
- Istniejące metody tensorowego przetwarzania danych a specyfika Big Data
- Proponowane podejście a specyfika Big Data
- Badania nad hierarchicznym przetwarzaniem multitensorowym a 'źródłowe' obszary badawcze
- Badany scenariusz aplikacyjny
- Wybór uzyskanych wyników
- Mikrousługi jako przyjęty paradygmat architekturalny
- Proponowane podejście a tzw. '4V Big Data'
- Podsumowanie
- Wybór literatury

Reprezentacja zbioru krotek w przestrzeni tensorowej

Fundamentem każdej tensorowej metody reprezentacji danych jest założenie, że dane wejściowe to zbiór n-atrybutowych krotek, przy czym każdej krotce przypisana jest wartość (interpretowalna jako prawdopodobieństwo, wartość logiczna, itp.).

Sposób reprezentacji zbioru krotek w przestrzeni tensorowej można zdefiniować jako:

$$\Gamma = (n, \mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(n)}, \Lambda, \psi),$$

gdzie $\mathcal{V}^{(i)}$, $(i = 1, \ldots, n)$ to zbiór możliwych wartości i-tego atrybutu n-atrybutowej krotki,
Λ to zbiór krotek postaci $(v^{(1)}, \ldots, v^{(n)})$, takich, że $v^{(i)} \in \mathcal{V}^{(i)}$
a $\psi : \mathcal{V}^{(1)} \times \cdots \times \mathcal{V}^{(n)} \to \mathbb{R}$ to funkcja przypisująca wartości poszczególnym krotkom ze zbioru Λ.
W celu wyrażenia dowolnego zbioru n-atrybutowych krotek w postaci tensora, określa się przestrzeń tensorową $\mathcal{T} = \mathcal{I}^{(1)} \otimes \cdots \otimes \mathcal{I}^{(n)}$, przy czym $\mathcal{I}^{(i)}$ to baza przestrzeni rzędu $|\mathcal{V}^{(i)}| = n_i$ służąca indeksowaniu elementów ze zbioru $\mathcal{V}^{(i)}$.
Dowolny zbiór n-atrybutowych krotek (w tym pojedyncza krotka), może być wyrażony jako punkt w przestrzeni tensorowej $\mathcal{T}$.

Motywacja do badań nad metodami tensorowymi w kontekście Big Data

- Zastosowania ML w Big Data wymagają uniwersalizacji reprezentacji danch wejściowych, przy zachowaniu:
  - ścisłej interpretowalności modelu danych[2]:
  - możliwości weryfikacji zestawów cech opisujących zjawisko w celu znalezienia cech 'kooperujących' z perspektywy wyniku klasyfikacji[3]
- Każdy model oparty na iloczynie tensorowym daje naturalną możliwość reprezentacji każdego wyrażalnego w systemie zdarzenia (próbki/krotki) jako iloczynu logicznego→tensorowego) współwystępowania wartości atrybutów określających to zdarzenie.
- Proponowany model - jako oparty na hierarchicznej sieci tensorów ('połączonych' relacjami kontrakcji wymiarów indeksowania) - daje możliwość:
  - weryfikacji wszystkich sposobów uogólnienia ('uproszczenia') modelu - w alternatywny sposób determinujących wyrażalność poszczególnych koniunkcji atrybutów,
  - integracji wyników przetwarzania uzyskanych dla wszystkich alternatywnych sposobów uogólnienia modelu.

[2] Matwin S.: *Uczenie się maszynowe: cztery lekcje - i co dalej?*, Biuletyn PSSI 2/2013

[3] Draminski M., Dabrowski M., Diamanti K., Koronacki J., Komorowski J.: *Discovering Networks of Interdependent Features in High-Dimensional Problems*. In: Japkowicz N., Stefanowski J. (eds) Big Data Analysis: New Algorithms for a New Society (2016)

Istniejące metody tensorowe a specyfika Big Data

- Metody tensorowe od kilku lat uznawane są za skuteczne narzędzie analizy danych wielorelacyjnych; do najpopularniejszych należą metody oparte na redukcji wymiarowości z użyciem metod dekompozycji[4] oraz iteracyjnej ortogonalizacji (algorytm *Higher-Order Orthogonal Iteration*)[5].
- Dotychczasowe badania nad zastosowaniem tensorowego modelu danych w obszarze Big Data wykazały dwie jego podstawowe wady:
  - występowanie zjawiska tzw. eksplozji kombinatorycznej[6] – przetwarzanie danych dużych rozmiarów wymaga modelowania wielu koniunkcji wartości atrybutów,
  - konieczność określenia a-priori (nielicznego) zbioru wymiarów indeksowania tensora[7].
- W wyniku tych ograniczeń większość do tej pory przedstawionych zastosowań metod tensorowych ogranicza się do modelowania arbitralnie zdefiniowanych przypadków tensorów rzędu 3 (np. w celu modelowania danych RDF).

---

[4] Kolda, T. G., Bader, B. W.: Tensor decompositions and applications, SIAM Review 51(3): 455–500 (2009)

[5] de Lathauwer L., De Moor B., Vandewalle J.: *On the Best Rank-1 and Rank-(R_1, R_2, . . . , R_N) Approximation of Higher-Order Tensors*. SIAM J. Matrix Anal. Appl. 21(4), 1324-1342 (2000)

[6] Vervliet, N., et al.: *Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in Big Data analysis*, IEEE Signal Processing Magazine 31(5): 71–79 (2014)

[7] Nickel M., Tresp V.: *An Analysis of Tensor Models for Learning on Structured Data*, in H. e. Blockeel (Ed.), Machine Learning and Knowledge Discovery in Databases, Vol. 8189 of LNCS, Springer, pp. 272–287 (2013)

Proponowane podejście a specyfika Big Data

Metoda **Multi-Order Principal Component Analysis (MOPCA)** opiera się na proponowanym multitensorowym modelu reprezentacji danych:

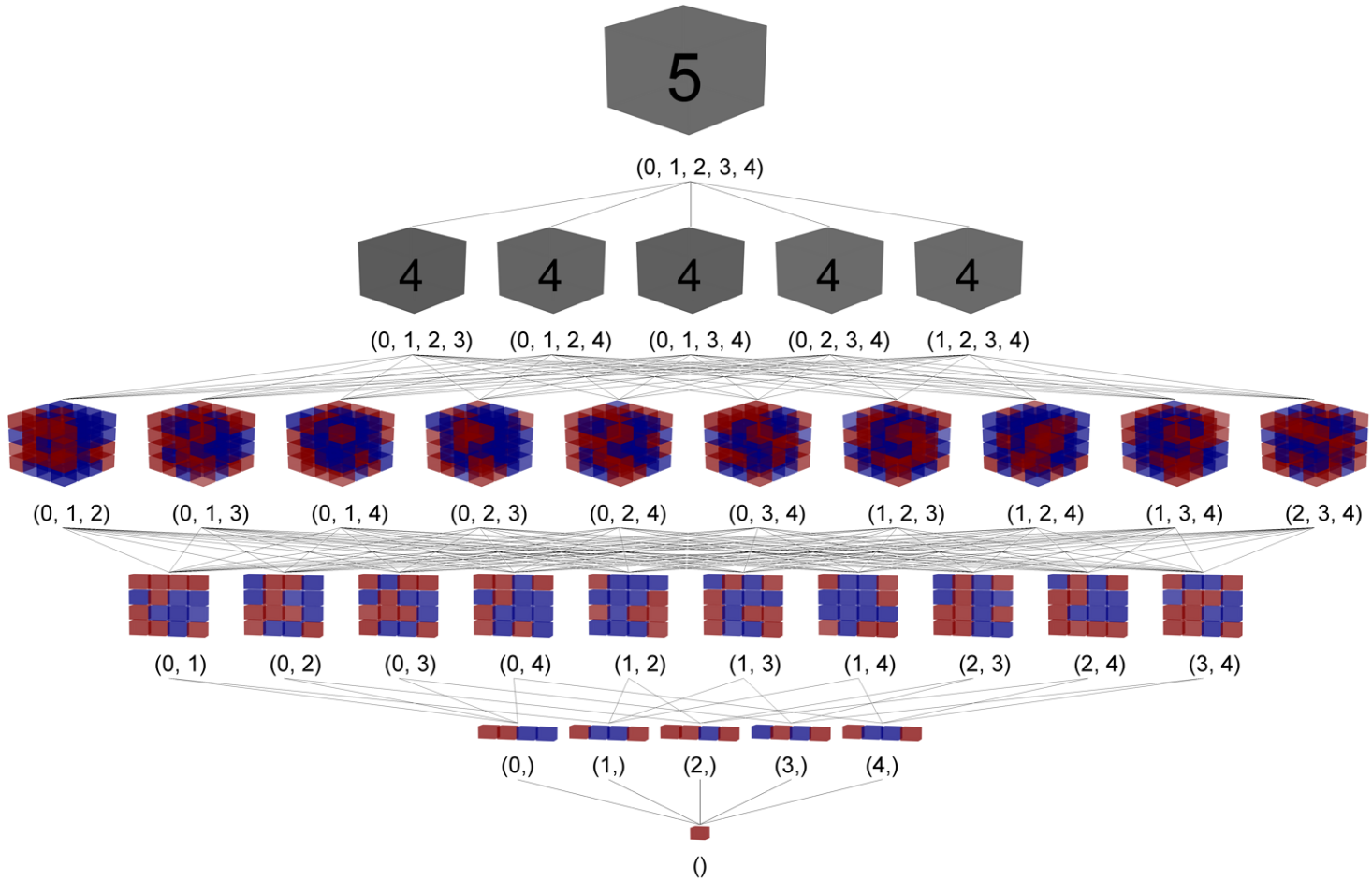
- zorientowanym na predykcję wartości logicznych przypisanych zdarzeniom,
- opartym na krotkowym formacie ujednoliconej reprezentacji heterogenicznych danych wejściowych (równoważnym tablicy ‘przykłady na atrybuty’),
- obejmującym reprezentację wartości nieznanych i niepewnych,
- opartym na **hierarchicznej strukturze tensorowej** umożliwiającej:
  - dekompozycję odpowiedzi systemu wg podzbiorów atrybutów wspólnie ‘determinujących’ wynik klasyfikacji zdarzeń,
  - ‘hierarchiczne’ przetwarzanie danych zgodne z zasada brzytwy Ockhama[8],
  - integrację modelu reprezentacji danych z modelem ich generalizacji – dzięki interpretowalności w ’języku logiki’,
  - modelowanie średnich zarówno w skali ’makro’ jak i skali ’mikro’[9],
  - modelowanie zależności zarówno na poziomie przypadków indywidualnych jak i różnych poziomów uogólnień[10],
  - ‘pogodzenie statystyki z eksplanacyjnością’ – dzięki hierarchicznej reprezentacji danych na różnych poziomach ich uogólnienia.

[8] Thagard, P.: *Coherence, Truth, and the Development of Scientific Knowledge*. Philosophy of Science 74 (1):28-47 (2007)

[9] *Reinventing Society in the Wake of Big Data* - Edge’s interview with Alex “Sandy” Pentland (Posted August 30, 2012)

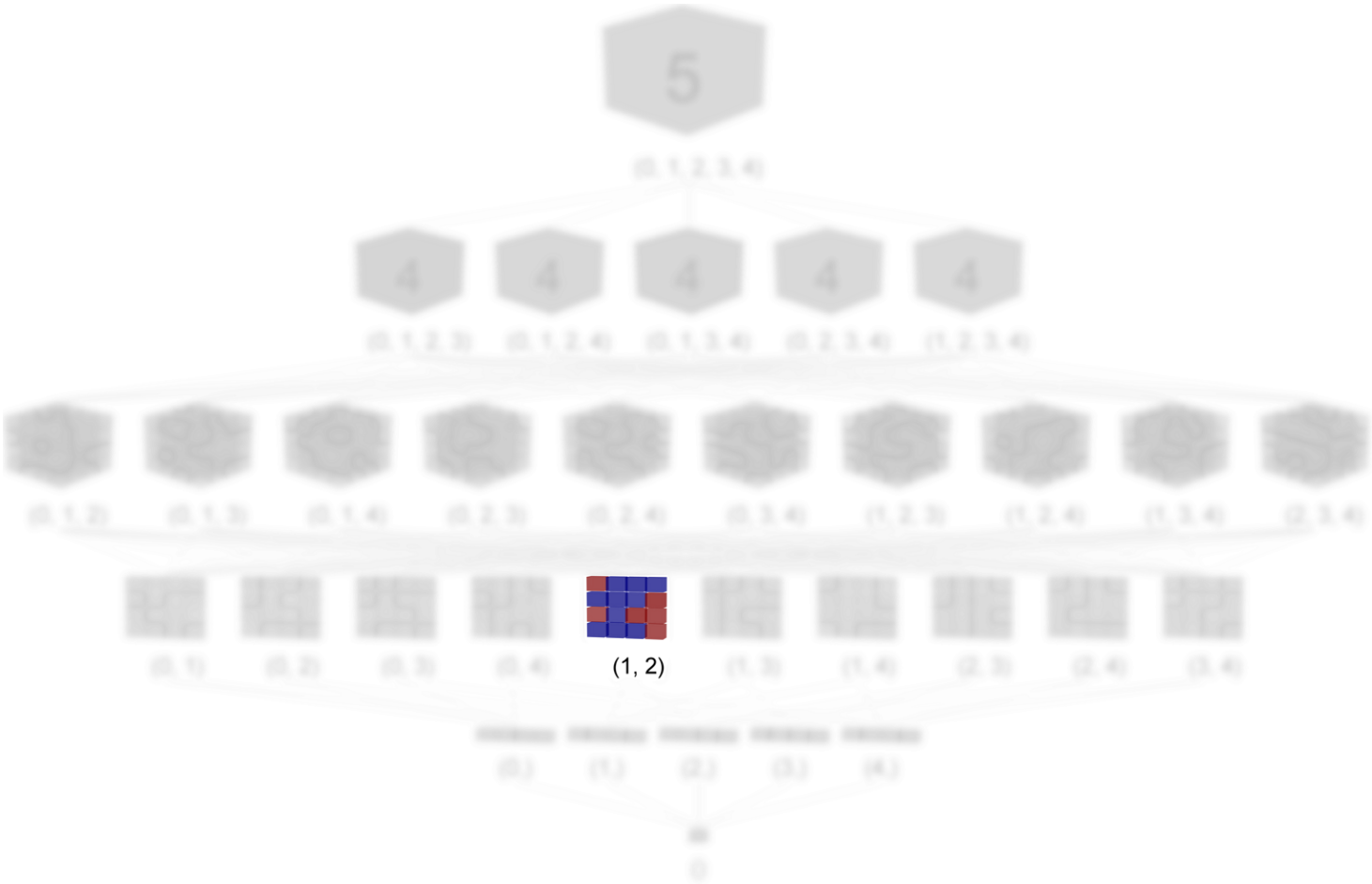
[10] Pietsch,W.: *Big Data? The New Science of Complexity*. In: 6th Munich-Sydney-Tilburg Conference on Models and Decisions (Munich; 10–12 April 2013)

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (1/5)



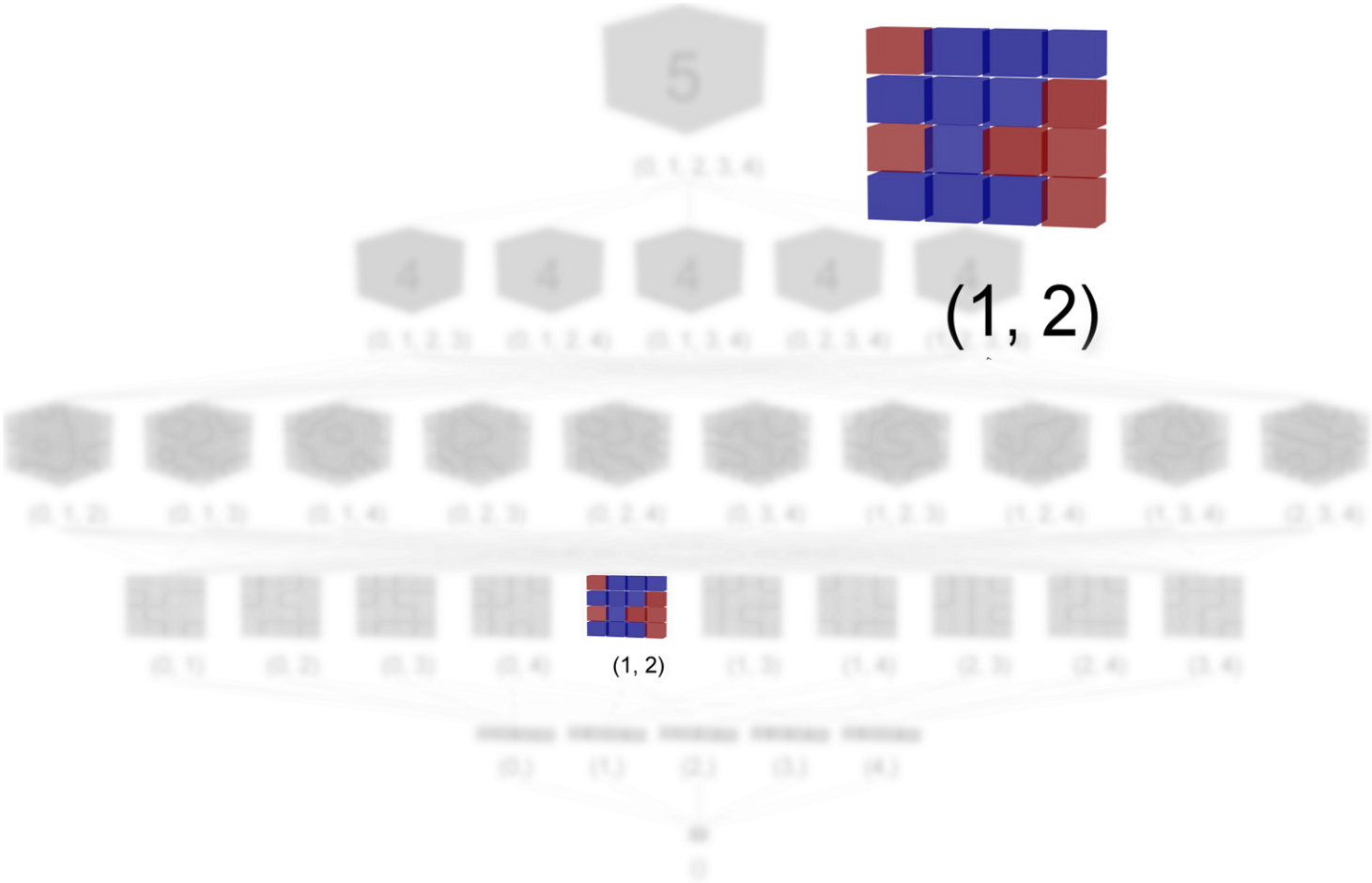
Rysunek: Wizualizacja sieci multitensorowej MOPCA dla przypadku reprezentacji 5 cech.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (2/5)



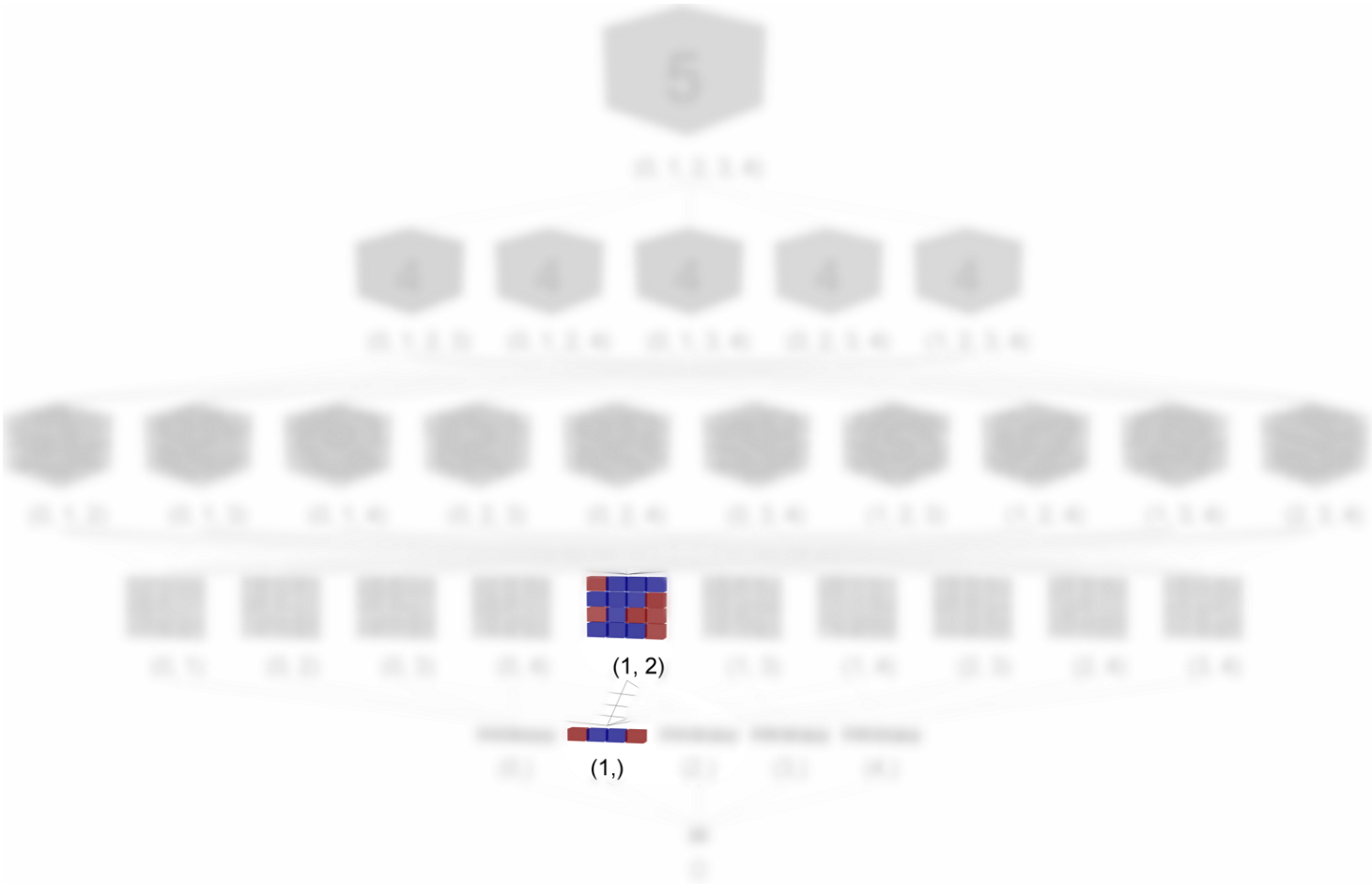
Rysunek: MOPCA a obszar badań nad metodami ’macierzowymi’ (np. na bazie SVD): składnik modelu MOPCA - monotensor rzędu 2.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (2/5)



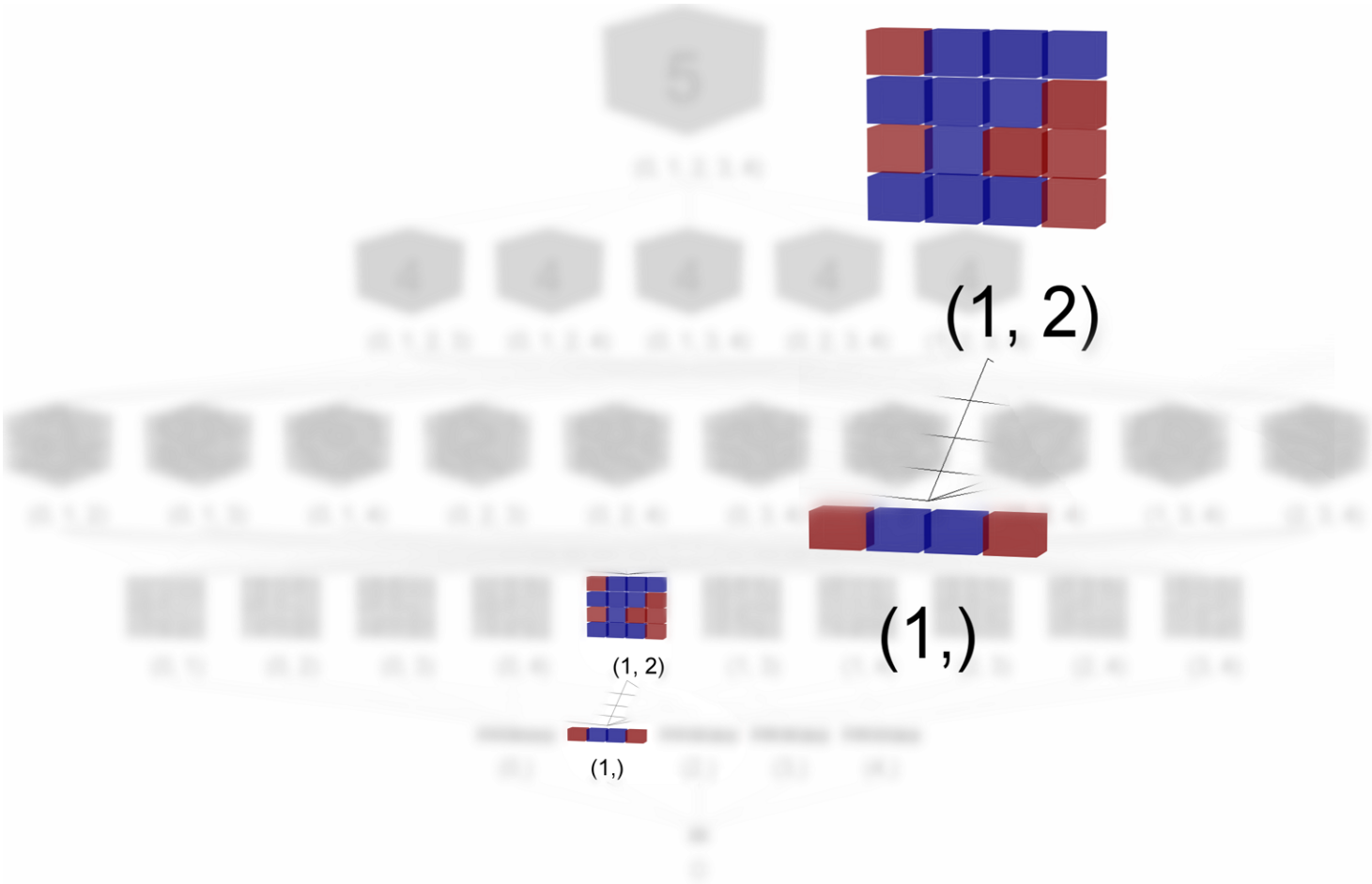
Rysunek: MOPCA a obszar badań nad metodami ’macierzowymi’ (np. na bazie SVD): składnik modelu MOPCA - monotensor rzędu 2.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (3/5)



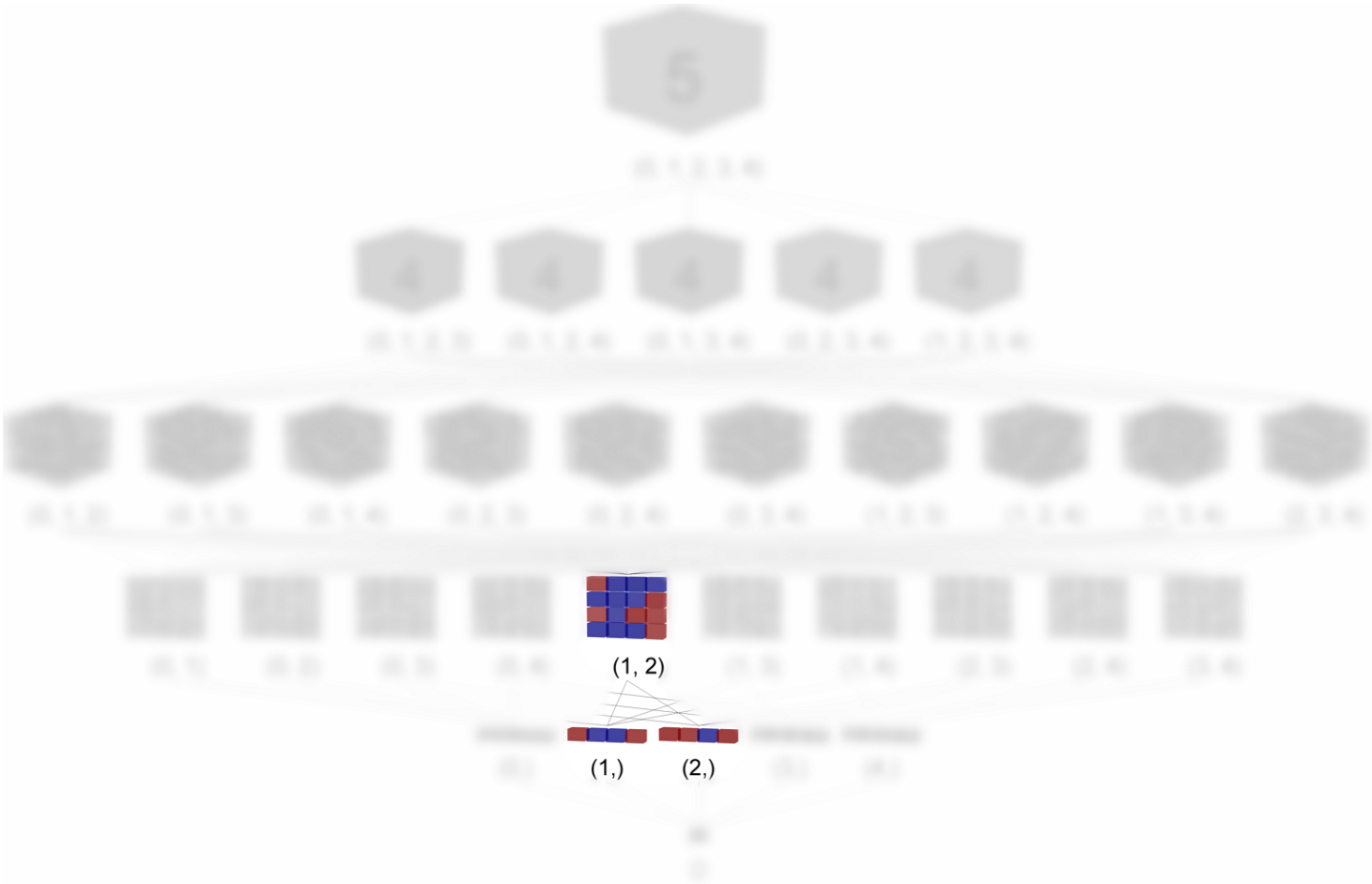
Rysunek: MOPCA a PCA: dwa składniki modelu MOPCA: monotensor rzędu 2 i jeden z wektorów wartości oczekiwanych.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (3/5)



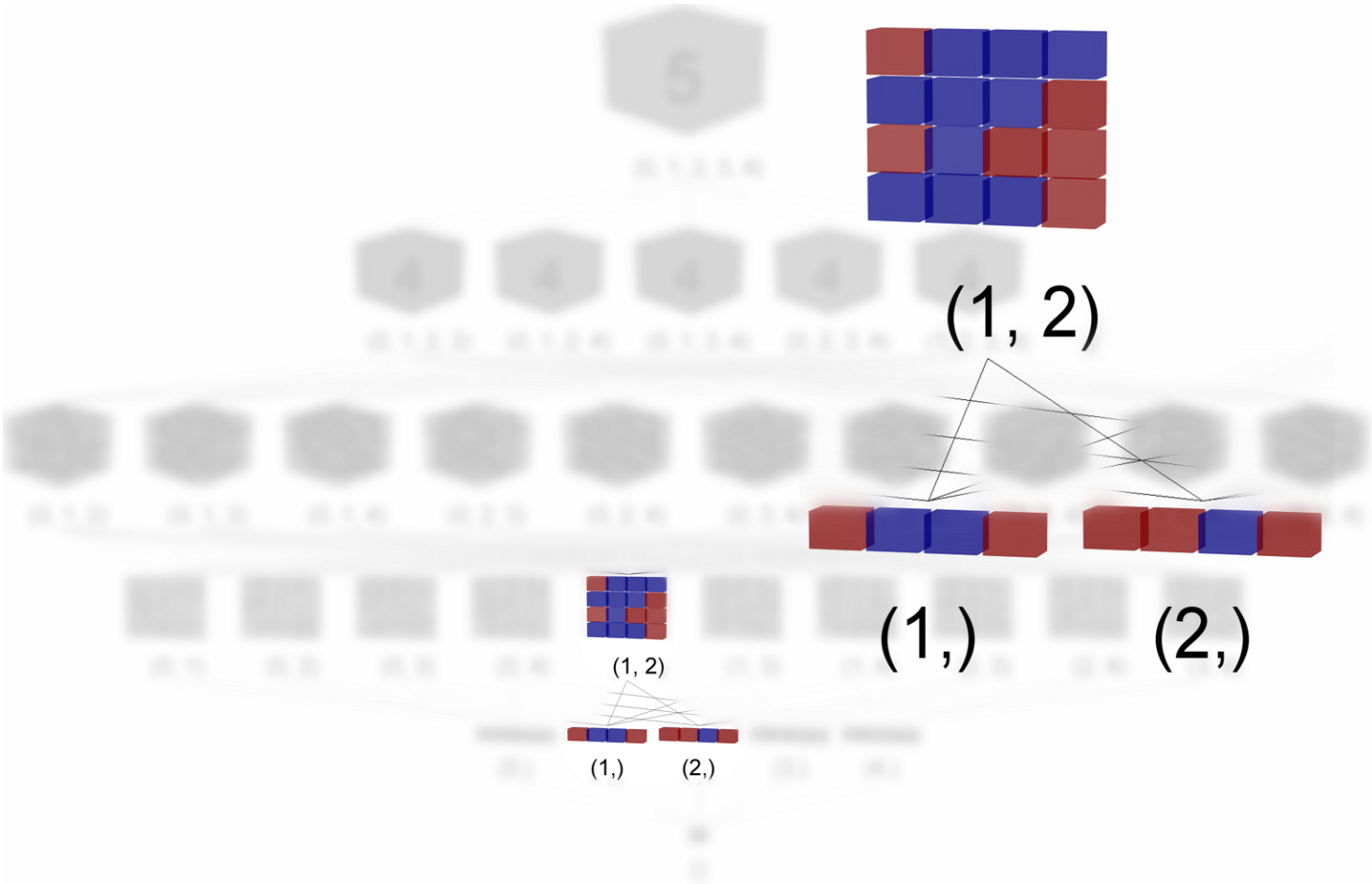
Rysunek: MOPCA a PCA: dwa składniki modelu MOPCA: monotensor rzędu 2 i jeden z wektorów wartości oczekiwanych.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (4/5)



Rysunek: MOPCA a 2-way PCA: trzy składniki modelu MOPCA: monotensor rzędu 2 i obydwa wektory wartości oczekiwanych.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (4/5)

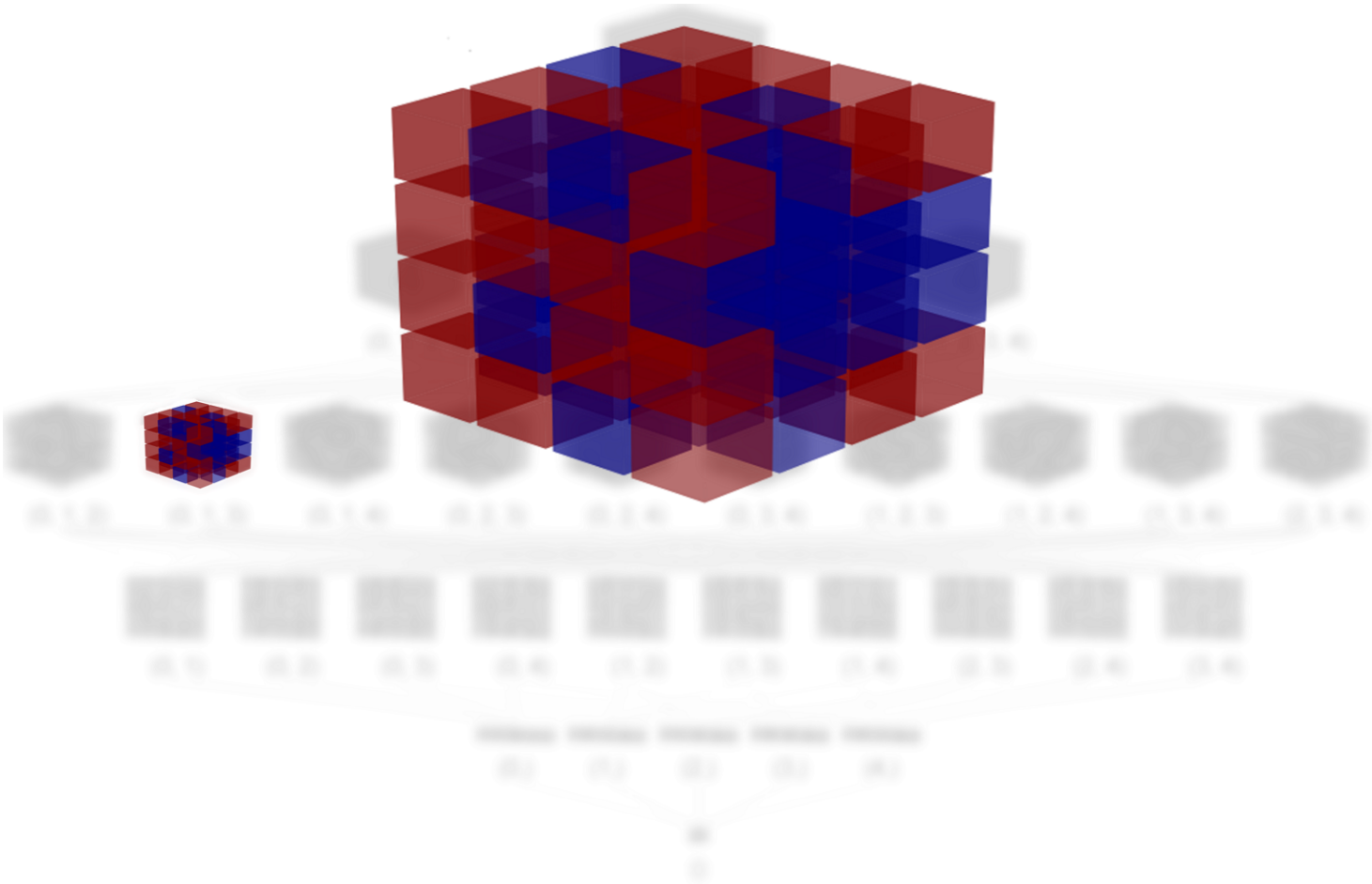


Rysunek: MOPCA a 2-way PCA: trzy składniki modelu MOPCA: monotensor rzędu 2 i obydwa wektory wartości oczekiwanych.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (5/5)

Rysunek: MOPCA a obszar badań nad metodami ’mono-tensorowymi’: HOSVD, HOOI, multi-way PCA, etc.

Badania nad hierarchicznym przetwarzaniem multitensorowym a ’źródłowe’ obszary badawcze (5/5)



Rysunek: MOPCA a obszar badań nad metodami ’mono-tensorowymi’: HOSVD, HOOI, multi-way PCA, etc.

Wybrany scenariusz aplikacyjny

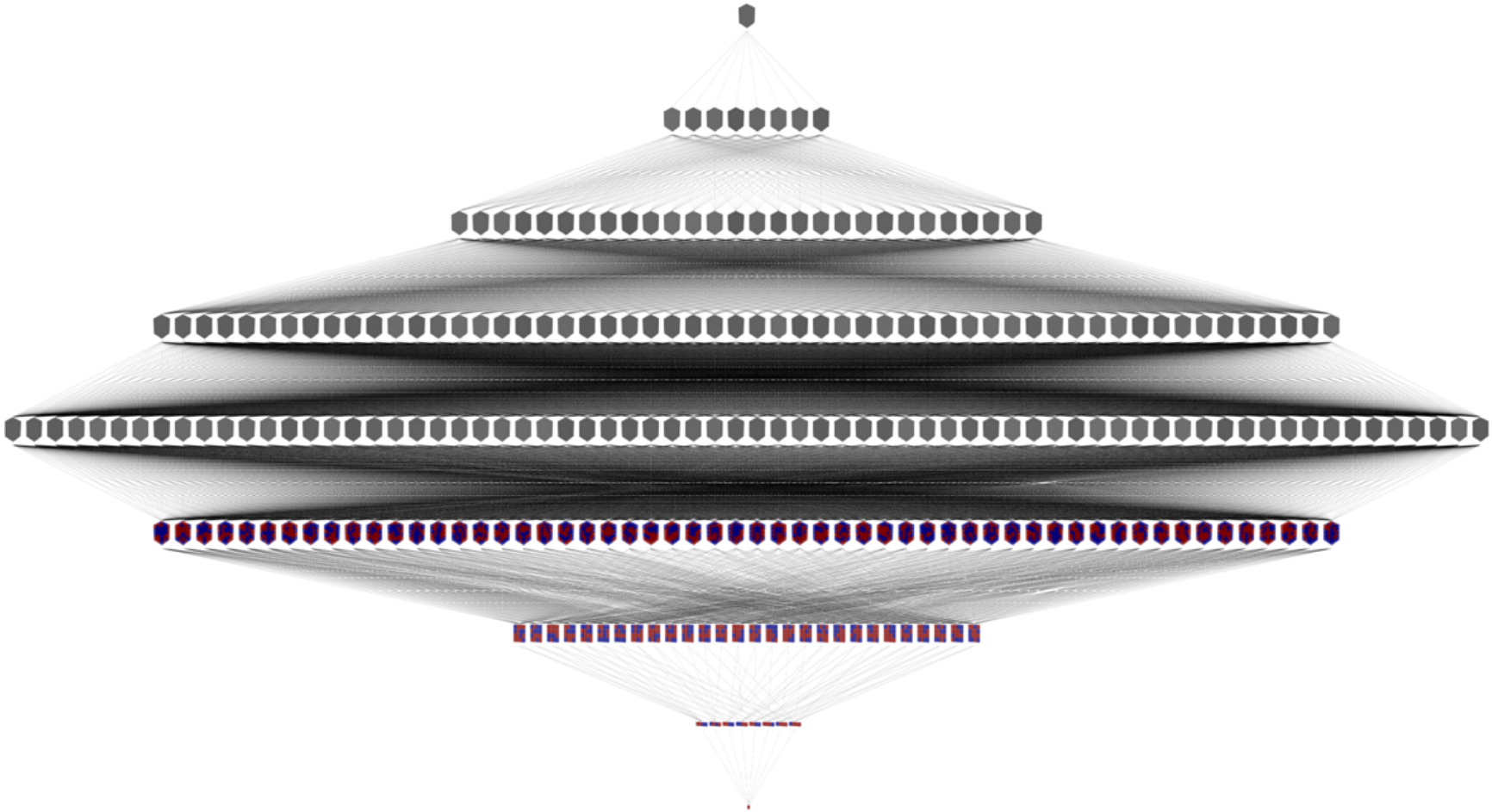
Scenariusz działania systemu licytacyjnego uczestniczącego w aukcjach RTB (ang. Real-Time Bidding), których przedmiotem jest powierzchnia reklamowa w Internecie[11] uważany jest za przykład praktycznego zastosowania technologii Big Data[12].

- Zbiór danych konkursów iPinYou Contest "Where Computational Advertising Meets Big Data"[13] zawiera ponad 100 milionów wieloatrybutowych krotek (zgromadzonych podczas kilku tygodni działania największego chińskiego systemu licytacyjnego klasy RTB-DSP).
- Klasyfikacja binarna zdarzeń wymaga trybu strumieniowego - czas odpowiedź systemu na zapytanie nie może przekroczyć 20 ms.
- Uzyskana we prezentowanych eksperymentach wartość 'wspolczynnika kompresji' reprezentacji danych (ok. 25 tys.) jest jednym z przejawów skalowalności proponowanego podejścia:
  - Teoretyczne wymagania pamięciowe wyrażone liczbą unikatowych zapytań/komórek: 9,8 mld
  - Zredukowane wymagania po (zoptymalizowanej) redukcji wymiarowości: 391056 zapytań/komórek

[11] Zhang W., Yuan S., Wang J., Shen X.: *Real-Time Bidding Benchenmarking with iPinYou Dataset*. Technical report. University College London (2014)

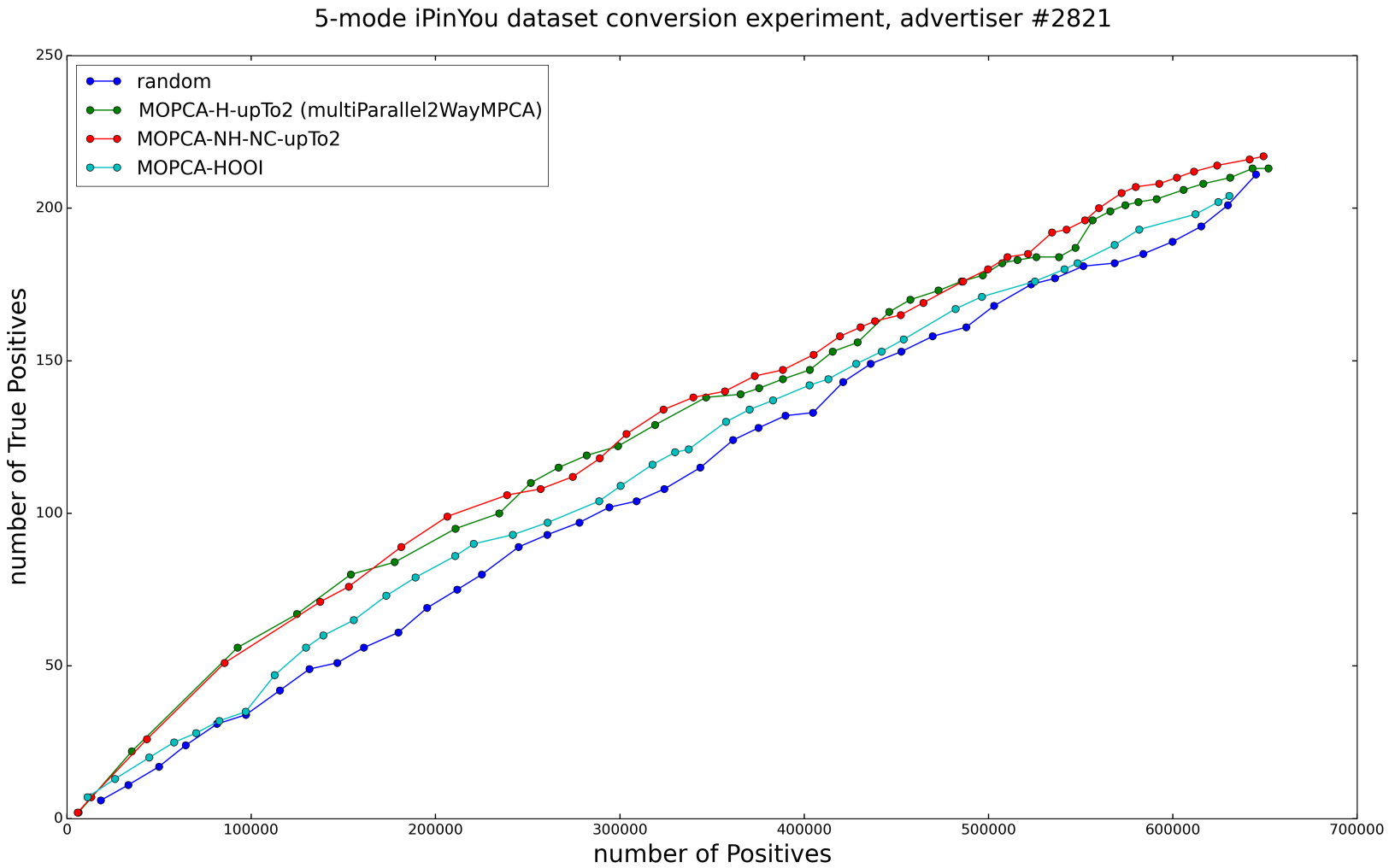
[12] Bulger M., Taylor G., Schroeder R.: *Data-Driven Business Models: Challenges and Opportunities of Big Data*, Oxford Internet

Hierarchiczny model multitensorowy MOPCA w przypadku reprezentacji 8 cech

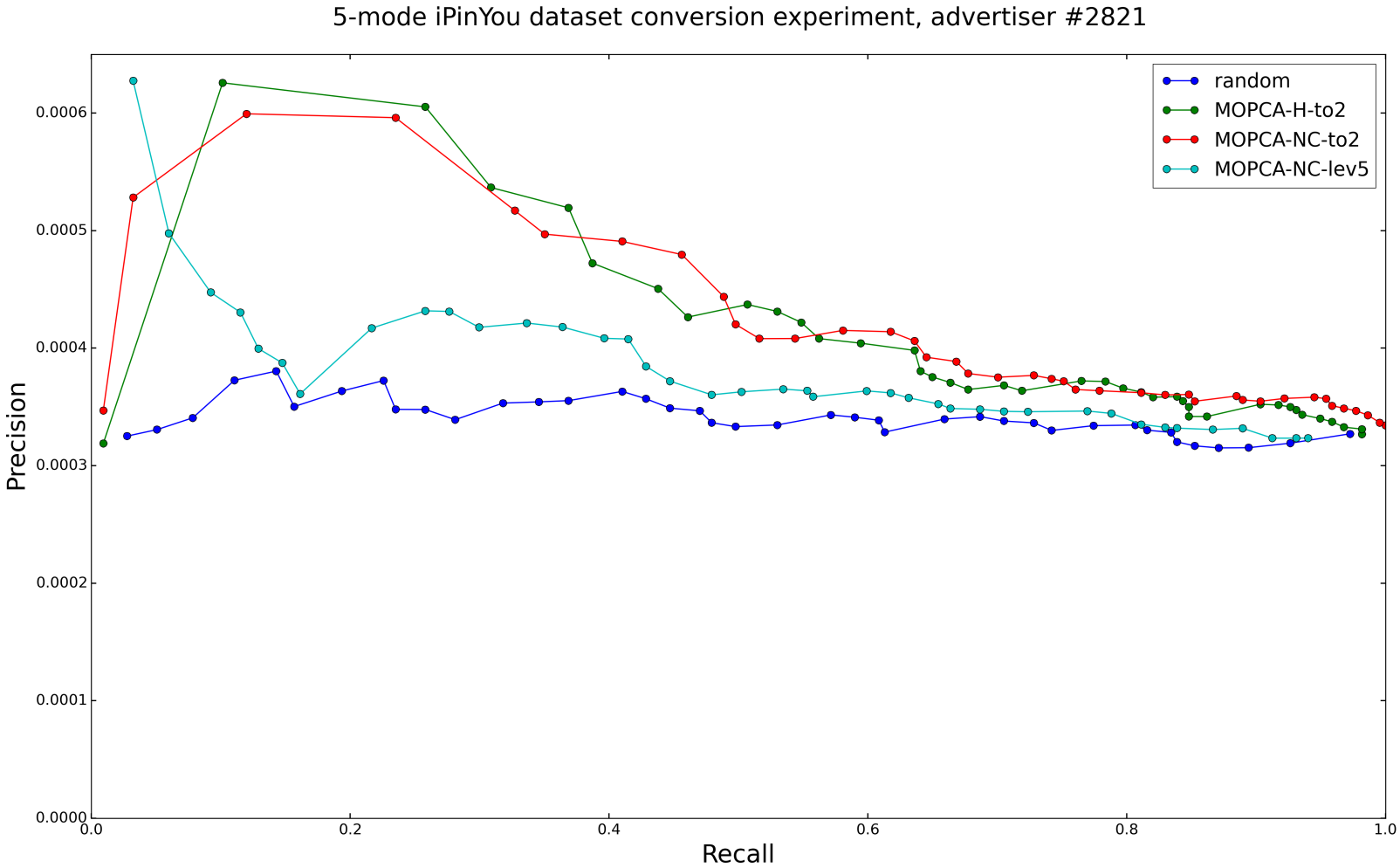


Rysunek: Wizualizacja sieci multitensorowej MOPCA dla badanego eksperymentalnie przypadku.

Wybór uzyskanych wyników (1/2)



Wybór uzyskanych wyników (2/2)



Mikrousługi jako przyjęty paradygmat architekturalny

- Mikrousługi (ang. microservices) dominują w komercyjnych systemach "usługowych" klasy Big Data (Google Dataflow, Netflix, NASA[14])
  - Technologia MapReduce – wydajne narzędzie do gromadzenia danych (szczególnie tekstowych) – nie odpowiada wymaganiom wieloetapowego przetwarzania danych wielowymiarowych[15].
  - Architektury mikrousługowe umożliwiają dostęp do funkcji bazowych oraz swobodne łączenie ('reduce') wyników w odpowiedzi na złożone zapytania.
- W kontekście "Big Data Analytics Beyond Hadoop"[16] istotna jest realizowalność sekwencyjnych etapów przetwarzania danych i realizacji zapytań.
- Właściwy MOPCA hierarchiczny model reprezentacji danych skutkuje złożonością topologii komponentów przetwarzających dane $\rightarrow$ celowością implementacji zgodnie z paradygmatem mikrousługowym.
- Wstępna implementacja MOPCA (ok. 5 tys. linii w języku Python) bazuje m.in. na bibliotekach Numpy, Multiprocessing, Tornado i jest inspirowana m.in. biblioteką TensorFlow przygotowaną przez Google Brain Team.

[14] Schnase J., et al., *MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service*, Computers, Environment and Urban Systems, (January 2014)

[15] Buck, J. B., et al.*SciHadoop: Array-based query processing in Hadoop.* In Proc. of 2011 international conference for high performance computing, networking, storage and analysis (SC '11), pp. 1–11 (2011)

[16] Agneeswaran, V.S.: *Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives*, FT Press (2014)

Proponowane podejście a tzw. cztery V Big Data (1/2)

- Volume
  - skala scenariusza aplikacyjnego RTB (Real-Time Bidding): realizacja procesu decyzyjnego w ramach mikroaukcji, których przedmiotem są odsłony reklamowe w Internecie: od kilkuset tysięcy do kilku milionów ofert spersonalizowanych odsłon na sekundę
  - model reprezentacji danych i algorytm realizacji zapytań umożliwiający realizację odpowiedzi dla bilionów unikatowych definicji zdarzeń
  - skalowalność implementacji dzięki rozpraszalnej architekturze mikrousługowej
  - skalowalność wymagań pamięciowych reprezentacji danych dzięki hierarchicznej ('wielopoziomowej') redukcji wymiarowości
  - skalowalna jakość realizacji zapytań (uwzględniająca reżim czasowy) dzięki addytywności wyników cząstkowych pochodzących z monotensorów
- Velocity
  - obsługa zmiennych strumieni zdarzeń, mniej niż 20 ms na odpowiedź
  - uwzględnienie znaczników czasowych w algorytmie MOPCA

[16] Japkowicz N., Stefanowski J.: *A Machine Learning Perspective on Big Data Analysis*. In: Japkowicz N., Stefanowski J. (eds) Big Data Analysis: New Algorithms for a New Society, pp. 1-31, Springer (2016)

Proponowane podejście a tzw. cztery V Big Data (2/2)

- Variety
  - oferty RTB opisane heterogenicznymi danymi - reprezentującymi czas, położenie internauty, urządzenie, adres odwiedzanej strony internetowej - wzbogaconymi zewnętrznymi danymi o użytkowniku (danymi demograficznymi, danymi o zainteresowaniach internauty, historii odwiedzin stron) oraz wynikami zastosowania popularnych metod NLP w celu budowy reprezentacji zawartości tekstowej odwiedzanych stron WWW)
  - wyrażalny w "języku logiki" krotkowy model reprezentacji danych
- Veracity
  - reprezentacja/przetwarzanie danych niepełnych, zanonimizowanych oraz o niezweryfikowanej jakości
  - reprezentacja nieznanych bądź alternatywnie intrepretowalnych wartości atrybutów

[16] Japkowicz N., Stefanowski J.: *A Machine Learning Perspective on Big Data Analysis*. In: Japkowicz N., Stefanowski J. (eds) Big Data Analysis: New Algorithms for a New Society, pp. 1-31, Springer (2016)

Podsumowanie

- Zastosowanie nowatorskiego multitensorowego modelu reprezentacji i przetwarzania danych umożliwia:
  - rozproszenie reprezentacji danych interpretowalnej zarówno w ’języku logiki’, jak i w ’języku statystyki’,
  - niearbitralność eksploracji wszystkich sposobów generalizacji danych,
  - integrację analizy korelacji z analizą wartości oczekiwanych,
  - zachowanie zgodności z zasadą brzytwy Ockhama przez ‘przyznanie pierwszeństwa’ (w hierarchicznym modelu przetwarzania) tym sposobom wyrażenia danych, które wymagają mniejszej liczby atrybutów.
- Zastosowanie metody redukcji wymiarowości wektorów[17] w celu redukcji wymagań pamięciowych stanowiącej połączenie HOOI i indeksowania pseudolosowego umożliwia użyteczną dekompozycję tensorów, które zajmowałyby[18] dziesiątki/setki TB[19].
- Hierarchiczność przetwarzania metodą MOPCA wskazuje na celowość implementacji w formie rozproszonego systemu mikrousługowego.
- Uzyskane dotychczas wyniki eksperymentów przemawiaja za tym, że proponowany model reprezentacji/przetwarzania danych i jego zastosowanie w metodzie MOPCA cechuje potencjalnie istotna wartość aplikacyjna.

[17] będących iloczynami Kroneckera wektorów odpowiadających reprezentacjom w poszczególnych przestrzeniach indeksowania komórek tensora

[18] w postaci ’nieskompresowanej’

Wybór literatury

- Japkowicz N., Stefanowski J.: *Big Data Analysis: New Algorithms for a New Society*, Studies in Big Data, vol. 16, Springer International Publishing (2016)
- Matwin S.: *Uczenie się maszynowe: cztery lekcje - i co dalej?*, Biuletyn Polskiego Stowarzyszenia Sztucznej Inteligencji 2/2013.
- Japkowicz N., Stefanowski J.: *A Machine Learning Perspective on Big Data Analysis*. In: Japkowicz N., Stefanowski J. (eds) Big Data Analysis: New Algorithms for a New Society, pp. 1-31, Springer International Publishing (2016)
- Draminski M., Dabrowski M., Diamanti K., Koronacki J., Komorowski J.: *Discovering Networks of Interdependent Features in High-Dimensional Problems*. In: Japkowicz N., Stefanowski J. (eds) Big Data Analysis: New Algorithms for a New Society, pp. 285-304, Springer International Publishing (2016)
- *Reinventing Society in the Wake of Big Data* - Edge's interview with Alex "Sandy" Pentland (Posted August 30, 2012)
- Ritter, D.: *When to act on a correlation and when no to*. Harward Business Review, March 19 (2014)
- Nickel M., Tresp V.: *An Analysis of Tensor Models for Learning on Structured Data*, in H. e. Blockeel (Ed.), Machine Learning and Knowledge Discovery in Databases, Vol. 8189 of Lecture Notes in Computer Science, Springer, pp. 272–287 (2013)
- Thagard, P.: *Coherence, Truth, and the Development of Scientific Knowledge*. Philosophy of Science 74 (1):28-47 (2007)
- Bulger M., Taylor G., Schroeder R.: *Data-Driven Business Models: Challenges and Opportunities of Big Data*, Oxford Internet Institute (September 2014)

Wybór literatury

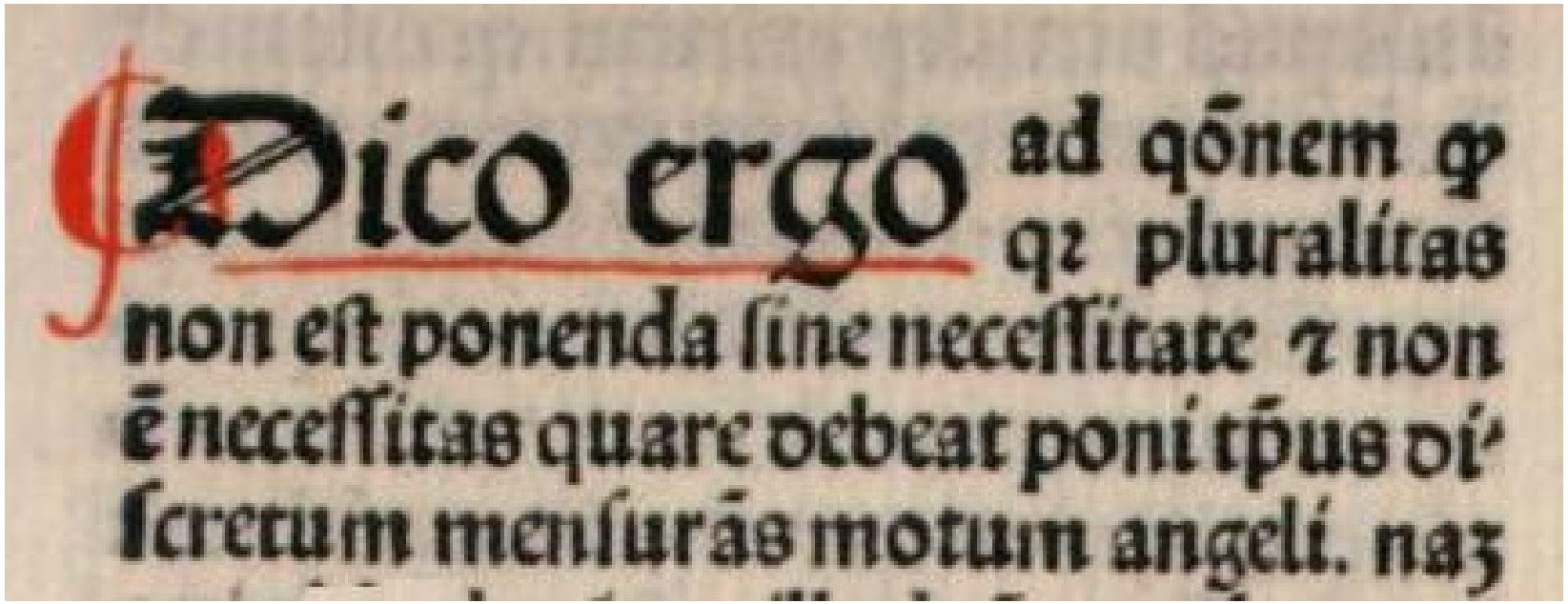
- Vervliet, N., Debals, O., Sorber, L., De Lathauwer, L.: *Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in Big Data analysis*, IEEE Signal Processing Magazine 31(5): 71–79 (2014)
- Cichocki, A.: *Era of Big Data processing: A new approach via tensor networks and tensor decompositions*, CoRR: abs/1403.2048 (2014)
- Zhang W., Yuan S., Wang J.: *Optimal real-time bidding for display advertising*. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data mining (KDD '14). ACM, New York, USA, pp. 1077-1086 (2014)
- Zhang W., Yuan S., Wang J., Shen X.: *Real-Time Bidding Benchenmarking with iPinYou Dataset*. Technical report. University College London (2014)
- iPinYou Global RTB Bidding Algorithm Competition Dataset, *http://contest.ipinyou.com/data-release.html* (dostęp: 03.2016)
- Bulger M., Taylor G., Schroeder R.: *Data-Driven Business Models: Challenges and Opportunities of Big Data*, Oxford Internet Institute (September 2014)
- Kroonenberg P. M.: *Three-mode principal component analysis: Theory and applications*, Vol. 2, DSWO press (1983)
- Tucker, L. R.: *Some mathematical notes on three-mode factor analysis*, Psychometrika 31(3): 279–311 (1966)
- De Lathauwer L., De Moor B., Vandewalle J.: *On the Best Rank-1 and Rank-($R_1, R_2, \ldots, R_N$) Approximation of Higher-Order Tensors*. SIAM J. Matrix Anal. Appl. 21(4), 1324-1342 (2000)

Wybór literatury

- De Lathauwer L., De Moor B., Vandewalle, J.: *A multilinear singular value decomposition*, SIAM Journal on Matrix Analysis and Applications 21: 1253–1278 (2000)
- Bro, R., Smilde, A. K.: *Centering and scaling in component analysis*, Journal of Chemometrics 17(1): 16–33, (2003)
- Lu, H., Plataniotis, K. N., Venetsanopoulos, A. N.: *A Survey of Multilinear Subspace Learning for Tensor Data*, Pattern Recognition 44(7): 1540–1551 (2011)
- Lu, H., Plataniotis, K. N., Venetsanopoulos, A. N.: *MPCA: Multilinear principal component analysis of tensor objects*, IEEE Trans. Neural Netw., 19 (1), 18–39 (2008)
- Kolda, T. G., Bader, B. W.: *Tensor decompositions and applications*, SIAM Review 51(3): 455–500 (2009)
- Pietsch,W.: *Big Data? The New Science of Complexity*. In: 6th Munich-Sydney-Tilburg Conference on Models and Decisions (Munich; 10–12 April 2013)
- Schnase J., et al., *MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service*, Computers, Environment and Urban Systems, (January 2014)
- Agneeswaran, V.S.: *Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives*, FT Press (2014)
- Szwabe, A., Misiorek, P., Ciesielczyk, M.: *Multilinear Filtering Based on a Hierarchical Structure of Covariance Matrices*, Schedae Informaticae 24: 98–107 (2015)

Duns Scotus: *Pluralitas non est ponenda sine necessitate*[20]

Dziękujemy za uwagę!



[20] Duns Scotus, 'Ordinatio', commons.wikimedia.org/wiki/File:Pluralitas.jpg

Slajdy dodatkowe

MOPCA a istniejące podejścia do tensorowej reprezentacji i przetwarzania danych

- Wszystkie dotychczasowe metody tensorowe dotyka problem “strukturalności” immanentnie arbitralnej wieloliniowości reprezentacji danych:
  - Wybór cech zdarzeń jako wymiarów w wieloliniowym (tensorowym) modelu ich reprezentacji i przetwarzania jest kluczowy z perspektywy jakości predykcji.
  - W tym kontekście mówi się o tzw. reprezentacji danych prze- i niedostrukturyzowanej (ang. overstructured vs understructured data representation)[21].
  - Hierarchiczna dekompozycja reprezentacji w połączeniu z modelem estymacji wartości oczekiwanych jest skutecznym sposobem na rozwiązanie tego problemu.
- Dotychczasowe metody tensorowego modelowania i przetwarzania danych “izolują” problem reprezentacji danych na bazie dekompozycji od problemu tzw. centrowania danych - pomimo, że powszechnie wiadomo, że niewłaściwe jest zaniedbywanie wpływu wartości oczekiwanych w systemie bazującym na reprezentacji kowariancji/korelacji.
- MOPCA to pierwsze podejście, w przypadku którego rola centrowania danych nie jest zmarginalizowana do "technicznej", a jest centralnym elementem modelu.

[21] Nickel M., Tresp V.: *An Analysis of Tensor Models for Learning on Structured Data*, in H. e. Blockeel (Ed.), Machine Learning and Knowledge Discovery in Databases, Vol. 8189 of LNCS, Springer, pp. 272–287 (2013)

MOPCA - główne elementy koncepcji

Metoda Multi-Order Principal Component Analysis - MOPCA - realizująca analizę głównych składowych dla hierarchii struktur tensorowych:

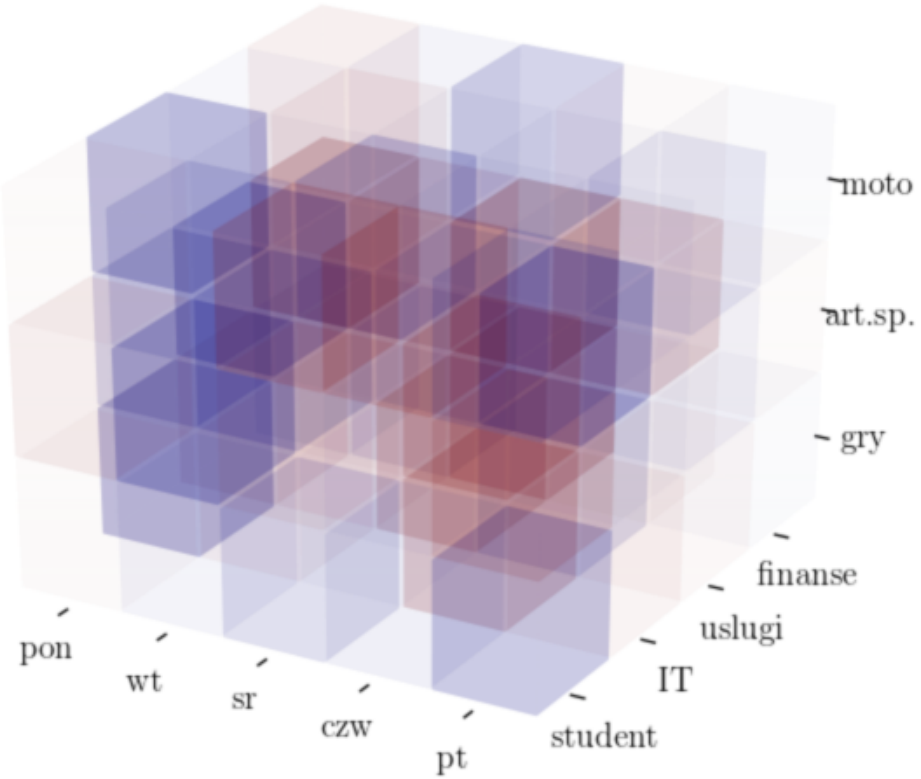
- hierarchiczna reprezentacja i przetwarzanie danych w sieci multi-tensorowej
- modelowanie i estymacja wartości oczekiwanych (ang. *centring*)
- hierarchiczne podejście do wieloliniowej regularyzacji spektralnej
- aplikowalność techniki próbkowania losowego i redukcji wymiarowości podprzestrzeni próbkowania
- algorytm realizacji zapytań do systemu implikowany modelem reprezentacji danych

MOPCA - dekompozycja danych w strukturze multitensorowej

- Dwa poziomy dekompozycji danych:
  - dekompozycja na hierarchię tensorów (tzw. ‘monotensorów’):
    - ★ umożliwiająca modelowanie i weryfikację różnych sposobów uogólnień,
    - ★ posiadająca interpretację językiem logiki,
    - ★ umożliwiająca zrównoleglenie obliczeń;
  - dekompozycja danych w ramach poszczególnych ‘monotensorów’:
    - ★ realizująca predykcję wartości ‘średnich’ dla danego sposobu uogólnienia zapytania (współwystępowania wartości w ramach określonego zbioru atrybutów),
    - ★ realizawana w kolejności wynikającej z hierarchii multitensorowej,
    - ★ realizowana razem z redukcją wymiarowości.

’Klasyczny’ model tensorowy

’Klasyczne’ podejście do modelowania tensorowego: użycie pojedynczego tensora jako jedynego środka reprezentacji danych.



Rysunek: Przykład (b. niewielkiego) tensora o trzech kierunkch indeksowania komórek, tj. służacego reprezentacji trójek

MOPCA - hierarchiczny model multitensorowy

- Proponowane podejście zakłada hierarchiczne modelowanie wartości oczekiwanych (średnich) i odchyleń od średnich (wariancji/kowariancji) z użyciem sieci multitensorowej:
  - Kolejność przetwarzania danych w strukturze multitensora zgodna jest z zasadą używania 'makrośrednich' do wyznaczenia zależnych od nich 'mikrośrednich' (odchyleń od średnich wyrażalnych dzięki wprowadzeniu nowego atrybutu) i determinuje topologię sieci multitensorowej (implikując m.in. nieaplikowalność architektury MapReduce).
  - Wyniki predykcji uzyskane dla danego tensora używane są jako estymaty wartości oczekiwanych podczas przetwarzania monotensorów na wyższych poziomach hierarchii multi-tensorowej (służących wyrażeniu odchyleń od średnich wyrażalnych w wyniku dodania nowego atrybutu, tj. w wyniku inkrementacji rzędu monotensora).
- Realizacja zapytania określonego w dowolnym podzbiorze iloczynów kartezjańskich atrybutów polega na równoległym (tj. architekturalnie rozpraszalnym) 'odpytaniu' (z użyciem tensorowego iloczynu wewnętrznego) właściwych elementów hierarchii multi-tensorowej (monotensorów) oraz zsumowaniu otrzymanych wyników.

MOPCA - nowe podejście do modelowania wartości oczekiwanych

Modelowanie wartości oczekiwanych:

- reprezentowanie średnich (wartości oczekiwanych) jako podstawa analizy głównych składowych (PCA) - relacja między PCA a SVD,
- uogólnienie PCA do przypadku wieloliniowego - MPCA,
- różne podejścia do tzw. centrowania danych:
  - 'klasyczne' PCA: dekompozycja po odjęciu wartości średnich poszczególnych cech,
  - reprezentowanie średnich w przypadku wielowymiarowym - różnica między tzw. overall centring[22] (odejmowaniem średnich we wszystkich wymiarach) a centrowaniem danych tylko w aktualnie przetwarzanym wymiarze (MPCA)[23]
- nowe podejście bazujące na hierarchicznym modelu przetwarzania - **hierarchical prediction-based centring** - w szczególności na:
  - używaniu - jako estymat wartości oczekiwanych - wyników przetwarzania "monotensorówńiższych rzędów (w miejsce wartości wejściowych),
  - wyznaczaniu wartości średnich na podstawie 'rzadkiej' (ang. sparse) (w miejsce 'gęstej', tj. tablicowej) reprezentacji danych: na podstawie tylko tych krotek wejściowych, dla których dany atrybut ma znaną wartość.

[22] Kroonenberg P. M.: *Three-mode principal component analysis: Theory and applications*, Vol. 2, DSWO press (1983)

[23] Lu, H., Plataniotis, K. N., Venetsanopoulos, A. N.: *MPCA: Multilinear principal component analysis of tensor objects*, IEEE Trans. Neural Netw., 19 (1), 18–39 (2008)

MOPCA - zastosowanie regularyzacji spektralnej

Metoda oparta na naprzemiennej filtracji spektralnej (Alternating Spectral Filtering):

- mająca cechy wspólne z metodami HOOI[24] i Multilinear PCA[25], ale:
  - realizowana dla każdego z elementów hierarchii tensorów z osobna,
  - realizująca ‘centrowanie danych’ z użyciem wyników przetworzenia tensorów ‘poprzedzających’ dany tensor w hierarchii,
  - umożliwiająca modelowanie ’wiarygodności’ ortogonalnych kierunków użytych do wynikowej reprezentacji zależności,
- realizowana równolegle z redukcją wymiarowości dla poszczególnych kierunków indeksowania:
  - wykorzystująca elementy metody Randomized SVD opartej na indeksowaniu losowym,
  - uwzględniająca wpływ losowości na zbieżność algorytmu.

[24] de Lathauwer L., De Moor B., Vandewalle J.: *On the Best Rank-1 and Rank-(R_1, R_2, . . . , R_N) Approximation of Higher-Order Tensors*. SIAM J. Matrix Anal. Appl. 21(4), 1324-1342 (2000)

[25] Lu, H., Plataniotis, K. N., Venetsanopoulos, A. N.: *MPCA: Multilinear principal component analysis of tensor objects*, IEEE Trans. Neural Netw., 19 (1), 18–39 (2008)

MOPCA - zastosowanie regularyzacji spektralnej

- Realizacja analizy głównych składowych w poszczególnych 'monotensorach':
  - przetworzenie danej struktury tensorowej - z użyciem autorskiej regularyzacji widmowej inspirowanej algorytmem HOOI (Higher Order Orthogonal Iteration)[26] użytej w celu predykcji wartości średnich/oczekiwanych dla zdarzeń opisanych zbiorem atrybutów odpowiadających wymiarom danego tensora
  - globalne 'centrowanie' danych (rozszerzenie tzw. 'overall centring')[27],
  - 'centrowanie' determinowane przez hierarchiczny algorytm przetwarzania
  - redukcja wymiarowości oparta na losowym indeksowaniu.

[26] de Lathauwer L., De Moor B., Vandewalle J.: *On the Best Rank-1 and Rank-(R_1, R_2, . . . , R_N) Approximation of Higher-Order Tensors*. SIAM J. Matrix Anal. Appl. 21(4), 1324-1342 (2000)

[27] Kroonenberg P. M.: *Three-mode principal component analysis: Theory and applications*, Vol. 2, DSWO press (1983)

Zastosowanie metod przetwarzania języka naturalnego

- Ekstrakcja terminów z tagów:
  - poprawki ‘literówek’ oraz rozwinięcie skrótów
  - filtrowanie stopwords
  - podział tagów na wyrażenia
- Identyfikacja zasobów z DBpedii:
  - Named Entity Recognition (DBpedia Spotlight)
  - zapytania SPARQL
- Eksploracja powiązań pomiędzy zasobami:
  - analiza syntaktyczna opisów z DBPedii (NLTK)
  - wyszukiwanie synonimów (WordNet)
  - analiza kategorii
- Wzbogacenie krotek online

Schemat semantycznych danych wzbogacających

```
(userTagPrefix, userTagWords, dbpediaURI, dbpediaCategories,
    dbpediaDescription, dbpediaDescriptionWithSynonyms)
```

- gdzie:
  - *userTagPrefix* – kategoria ze zbioru {“*Long-term interest*”, “*In-market*”, “*Demographic/gender*”},
  - *userTagWords* – frazy pochodzące z nazw tagów (np. “*real estate*”),
  - *dbpediaURI* – URI (lub lista URI) zasobu z DBPedii najbardziej pasującego do *userTagWords*,
  - *dbpediaCategories* – kategorie odpowiadające *dbpediaURI*,
  - *dbpediaDescription* – rzeczowniki pochodzące z komentarza (`rdfs:comment`) oraz streszczenia (`dbpedia/ontology:abstract`) dla *dbpediaURI*.
  - *dbpediaDescriptionWithSynonyms* – rzeczowniki z pola *dbpediaDescription* rozszerzone o synonimy z WordNet.

Przykładowe dane wzbogacające krotki

```
(
  'Long-term interest',
  'IT',
  'http://dbpedia.org/resource/Information_technology',
  ('Media technology', 'Information technology'),
  ('information', 'technology', 'application', [...]),
  (
    ('information', 'info', 'information', 'information', 'data',
      'information', 'information', 'selective information',
      'entropy'),
    ('technology', 'engineering', 'engineering', 'engineering
      science', 'applied science', 'technology'),
    ('application', 'practical application', 'application',
      'application', 'coating', 'covering', 'application',
      'application program', 'applications programme', 'lotion',
      'application', 'application', 'diligence', 'application'),
    [...]
  )
)
```

Listing 1: Fragment krotki z danymi dla tagu 'Long-term interest/IT'.

Analiza syntaktyczna

- dane tekstowe dla zasobów z DBpedii:
  - streszczenie (*dbpedia.org/ontology/abstract*)
  - komentarz (*rdfs:comment*)
- tokenizacja (NLTK)
- ekstrakcja nazw własnych - NER (DBpedia Spotlight)
- lematyzacja - WordNetLemmatizer (NLTK)
- tagowanie - Standard treebank POS tagger (NLTK)
- wyszukiwanie synsetów (WordNet)