Podobnostní vyhledávání: současný stav a výzvy do budoucna

SIRET research group (SRG), KSI MFF UK
http://siret.ms.mff.cuni.cz

Tomáš Skopal

Potřeby praxe, (potenciální) poptávka

- chytřejší vyhledávání v databázích složitých typů dat
  - ne relační databáze
    - vznikají „uměle“, vyrábí je člověk (rozumí vnitřní sémantice)
  - „signálová“ data
    - multimédia – obraz, zvuk, video
    - biologická data – proteiny, obecně molekuly
    - medicínská data – různá vyšetření, např. EEG
  - **podobnostní vyhledávání**
    - doménově závislé modelování
      - extrakce vlastností
      - podobnostní funkce jako základ vyhledávání, resp. klasifikace
    - rychlost dotazování
      - potřeba indexace (databázový problém)

Problémy v oboru (multimedia similarity search)

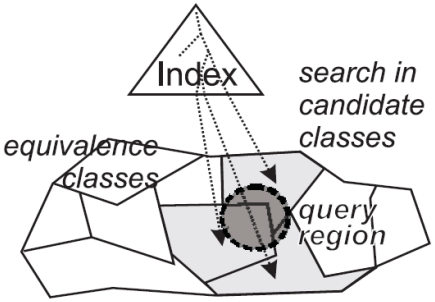
modelování

$x = [0,.3,.6,0,...,6.1]$, $\delta = L_2$

stav: globální low-level deskriptor, metrická podobnost
(**nestačí:** málo sémantiky, nerobustní)

výzva: segmentace objektu, vyšší sémantika, lokální podobnost

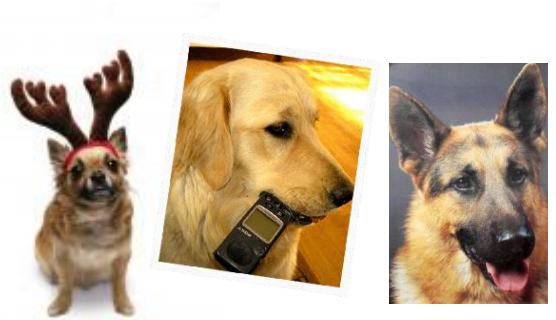
indexování



stav: metrické metody
(**nestačí:** omezují, lokální podobnost je nemetrická)

výzva 1: multireprezentace do metrického prostoru + agregace při dotazech
výzva 2: nemetrické indexy

dotazování



stav: rozsahové a kNN dotazy
(**nestačí:** jsou příliš jednoduché)

výzva: filtrace, re-ranking, atd.

SRG v kostce

- **Siret Research Group (SRG)** při KSI MFF UK
  - staff: Skopal, Pokorný
  - doktorandi: Hoksza, Lokoč, Vajbar, Galgonek
- podobnostní vyhledávání v komplexních databázích
  - obecné metody rychlého vyhledávání
    - metrické metody — Skopal, Lokoč, Hoksza
    - zobecněné (nemetrické) metody — Skopal, Lokoč
  - nové typy podobnostních dotazů — Skopal
  - doménově specifické oblasti
    - multimédia (vyhledávání obrázků) — Skopal, Lokoč
    - proteiny (klasifikace, predikce) — Hoksza, Galgonek
    - EEG sekvence (vyhledávání, klasifikace) — Vajbar
    - podobnost XML , webových stránek — Pokorný

Vybrané výsledky SRG, 2009

	modelování	indexování	dotazování
obecné	A. Eckhardt, T. Skopal, P. Vojtas. On fuzzy vs. metric similarity search in complex databases, FQAS 2009, Roskilde, Denmark, **Springer**	T. Skopal, J. Lokoč, New Dynamic Construction Techniques for M-tree. Journal of Discrete Algorithms, 7(1):62-77, 2009, **Elsevier** T. Skopal, B. Bustos, On Index-free Similarity Search in Metric Spaces, DEXA 2009, Linz, Austria, **Springer** J. Lokoč. Parallel dynamic batch loading in the M-tree, SISAP 2009, Prague, **IEEE**	T. Skopal, V. Dohnal, M. Batko, P. Zezula. Distinct Nearest Neighbors Queries for Similarity Search in Very Large Multimedia Databases, **ACM** WIDM 2009, Hong Kong, China
doménově specifické	J. Galgonek, D. Hoksza. On the Effectiveness of Distances Measuring Protein Structure Similarity, SISAP 2009, Prague, **IEEE**	D. Hoksza. DDPIn - Distance and Density Based Protein Indexing, CIBCB 2009, ISBN 978-1-4244-2783-3, Nashville, USA, **IEEE**	
	D. Hoksza, J. Galgonek. Density-Based Classification of Protein Structures Using Iterative TM-score, CSBW 2009, Washington, USA, **IEEE**		
	M.Kudělka, Y. Takama, V. Snášel, K. Klos and J. Pokorný: Visual Similarity of Web Pages. AWIC 2009, Praha, **Springer**	J. Novák, D. Hoksza. An Application of the Metric Access Methods to the Mass Spectrometry Data, CIBCB 2009, Nashville, USA, **IEEE**	

Běžící granty

- GAČR 201/09/0683, 2009 – 2011
  Similarity Searching in Very Large Multimedia Databases,
  - spoluřešitel Skopal (řešitel prof. Zezula, MU Brno)

- GAUK 18208, 2008 – 2009
  Distributed and parallel metric indexing in multimedia databases,
  - řešitel Lokoč

Konference SISAP v Praze, 29-30.8.2009

- 2. ročník mezinárodní konference (Similarity Search and Applications),
  - co-chairs: Skopal, Zezula
- 40 registrovaných z celého světa
- 4 formy příspěvků
  - 2 zvané přednášky
  - 16 dlouhých příspěvků
  - 2 postery
  - 7 demonstrací

- IEEE sborník, ACM in-cooperation status (obsah sborníku navíc v ACM DL)
- 4 nejlepší příspěvky pozvány do speciálního čísla Information Systems (Elsevier), eds: Skopal, Zezula, Dohnal
- 1. ročník (2008) – Cancun, Mexiko
  3. ročník (2010) – Istanbul, Turecko

www.sisap.org

Spřátelené týmy

- Masarykova univerzita v Brně, ČR
  - prof. Zezula
- University of Chile, Santiago, Chile
  - prof. Navarro, dr. Bustos
- Universidad Michoacana, Mexiko
  - prof. Chávez
- University of Bologna, Itálie
  - prof. Ciaccia, prof.Patella
- University of California, Riverside, USA
  - prof. Keogh