Podobnostní vyhledávání: současný stav a výzvy do budoucna

(výroční zpráva SRG za rok 2010)

SIRET research group (SRG), KSI MFF UK

http://siret.ms.mff.cuni.cz

David Hoksza, 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**, **RNA**, chemické struktury
  - **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)

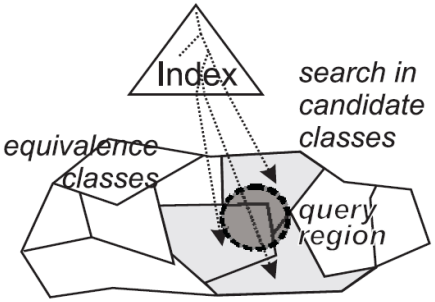
modelování

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

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

výzva: segmentace objektu, vyšší sémantika, lokální podobnost

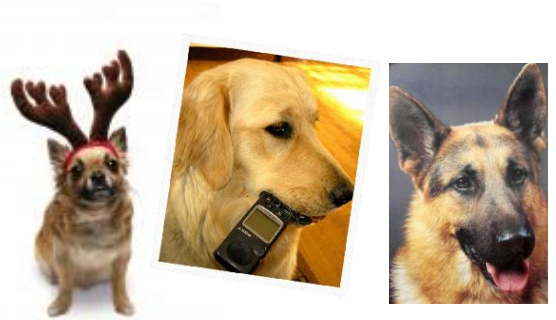
indexování



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: Hoksza, Lokoč, Pokorný, Skopal
  - doktorandi: Bartoš, Galgonek, Novák
- podobnostní vyhledávání v komplexních databázích
  - obecné metody rychlého vyhledávání
    - metrické metody — Lokoč, Novák, Skopal
    - zobecněné (nemetrické) metody — Lokoč, Skopal
  - nové typy podobnostních dotazů — Lokoč, Skopal
  - doménově specifické oblasti
    - multimédia (vyhledávání obrázků) — Lokoč, Skopal
    - proteiny (klasifikace, predikce) — Galgonek, Hoksza
    - RNA (podobnost, klasifikace) — Hoksza
    - Hmotnostní spektrometrie — Novák
    - podobnost XML , webových stránek — Bartoš, Pokorný

Vybrané výsledky SRG, 2010

	modelování	indexování	dotazování
obecné		T. Skopal. Where are you heading, metric access methods? A provocative survey, SISAP 2010, Istanbul, Turkey, **ACM DL** J. Lokoč, T. Skopal. On Applications of Parameterized Hyperplane Partitioning, SISAP 2010, Istanbul, Turkey, **ACM DL**	J. Kasarda, M. Nečaský, T. Bartoš. Generating XForms from an XML Schema, NDT 2010, Prague, Czech Republic, **Springer** J. Kasarda, T. Bartoš. Semi-Automatic Transformation of an XML Schema to Xforms, ISD 2010, Prague, Czech Republic, **Springer**
doménově specifické	D. Hoksza, J. Galgonek. Alignment-Based Extension to DDPIn Feature Extraction, IJCB 1(1), 2010, **ACTA Press** J. Galgonek, D. Hoksza. SProt - From Local to Global Protein Structure Similarity, BIBMW 2010, Hong Kong, China, **IEEE**	J. Novák, T. Skopal, D. Hoksza, J. Lokoč. Improving the Similarity Search of Tandem Mass Spectra Using Metric Access Methods, SISAP 2010, Istanbul, Turkey, **ACM DL**	
doktorské práce	D. Hoksza. Similarity Search in Protein Databases	J. Lokoč. Tree-based Indexing Methods for Similarity Search in Metric and Nonmetric Spaces	

B. Bustost, T. Skopal. Beyond the Metric Space Model, ACM SIGSPATIAL Special, 2(2):20-23, 2010, **ACM**

J. Novák, D. Hoksza. Similarity Search and Posttranslational Modifications in Tandem Mass Spectra, BIBMW 2010, Hong Kong, China, **IEEE**

Běžící granty

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

- GAČR P202/11/0968, 2011-2014
  Large-scale Nonmetric Similarity Search in Complex Domains,
  - řešitel Skopal

Spřátelené týmy

- Masarykova univerzita v Brně, ČR
  - prof. Zezula
- Vysoká škola chemicko-technologická v Praze
  - dr. Svozil
- University of Chile, Santiago, Chile
  - prof. Navarro, dr. Bustos
- RWTH Aachen university
  - dr. Beecks
- Universidad Michoacana, Mexiko
  - prof. Chávez
- University of Bologna, Itálie
  - prof. Ciaccia, prof. Patella
- University of California, Riverside, USA
  - prof. Keogh