

---

---

---

---

---



# B1-Z1-9 Techniky pro vyhledávání textových, webových a multimediálních dokumentů: modely, algoritmy, aplikace. Optimalizace web str.

Web search engines:

- full-text indexing + link analysis
- keyword query
- full-text query

Dokument - full-text

collection - množina dokumentů

term - slovo dokumentu

Slovník - distinct slov

## Boolský model

1, preprocessing  $\left\{ \begin{array}{l} \text{stopwords} \\ \text{stemming} \end{array} \right\}$  redukuje ~80%

precision -  $P$ , že vrátený dokument je relevantní

recall -  $P$ , že relevantní dokument je vrátený

false alarm - nemal být vrátený, ale bol

false dismissal - mal být vrátený, ale nebol

Inverted index - slovo a jeho linked list dokumentů kde sa objavuje  
merge-sort style

## Rozšířený Boolský model

- váha slova  $k_x$  v dokumente  $j$  -  $w_{kj} = f_{k,j} \cdot \frac{idf_k}{\max_j idf_j}$

-  $q$  - query v DWT/KNT

$$DWT - \text{relev}(q, d, j) = 1 - \sqrt{\frac{(1-w_{1,j})^2 + (1-w_{2,j})^2 + \dots + (1-w_{i,j})^2}{t}}$$

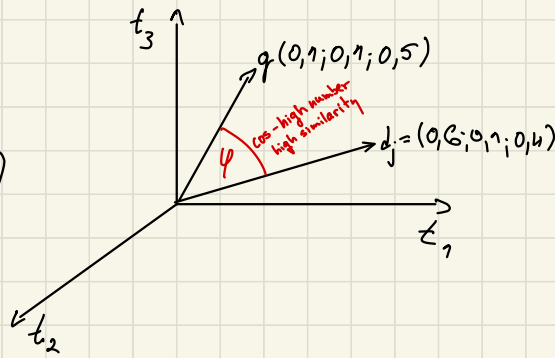
# Vektorový model

- bag of words model
- document = bag of terms (multi-set)

$f_{ij}$  - freq of term  $t_i$  in document  $d_j$   
 $tf_{ij} = f_{ij} / \max_i \{f_{ij}\}$

$df_i$  - # of documents containing term  $t_i$   
 $idf_i$  - inverse  $\rightarrow = \log_2(n/df_i)$

$$w_{ij} = tf_{ij} idf_i$$



# SEO optimalizace

- filenames - relevantné názvy, URL čo najkratšie
- <title> tag - najdôležitejšia info, max 64 znakov
- <headers> tag -
- <meta> - google už nepoužíva
- textové modifikátory - <strong>, <i>, ...
- popis obrázkov
- vnútorná štruktúra linkov
- keyword generation, selection

# BI-Z1-10 Vyhledávání v multimed. DB, podobnostní vyhledávání podle obsahu, podob. dotazování, agregační operátory, indexování metrické podobnosti, aproximativní vyhledávání

- loose structure, loose semantics

## Content based

- feature extraction - vector, set, ordered set
- similarity measure
- query - by - example

## Podobnostné funkcie

- vektorová vzdialenosť -
- adaptívna vzdialenosť - signatures, gen. sets
- sekvencia vzdialenosť - time series, strings

## Global features

- scalable color
- color structure

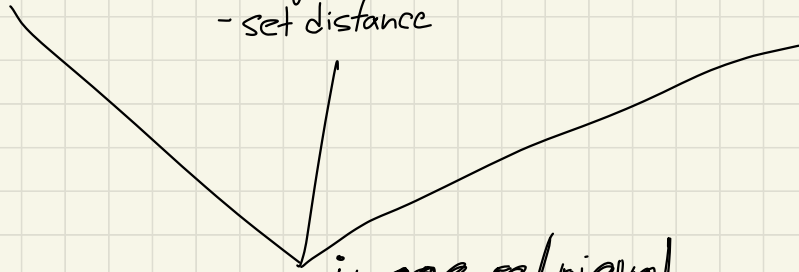
## Local features

- SIFT or SURF - interest points
- image descriptor
- set distance

## Time series

- shape retrieval
- dynamic time warping dist
- longest common subsequence

image retrieval



# Audio retrieval

- spektrálny rozklad
- melody / score - set of 2D points
  - ↳ distance - Earth mover's dist
- všeobecne náročný výpočet - lowerbound filtrácia na irelevantné prvky

## Similarity query

- similarity ordering, range
- $kNN$ ,  $kRNN$

## Operators

- similarity joins, self-joins
- $(k)$  closest pairs
- skyline operators - subset of elements that aren't dominated by other
- top-k operator - agregácia rankov do jedného
  - Fagin alg.
  - Threshold alg.

# Indexování metrické podobnosti

- pivot tables - AESA, LAESA
- hierarchical structures - GNAT, M-Tree
- hashed indexes - D-index
- hybrid structures - PM-tree, M-index

metric performance - distance computation

- I/O cost
- internal cost
- realtime cost