Calculating distance measure for MRDM clustering

Olegas Niaksu

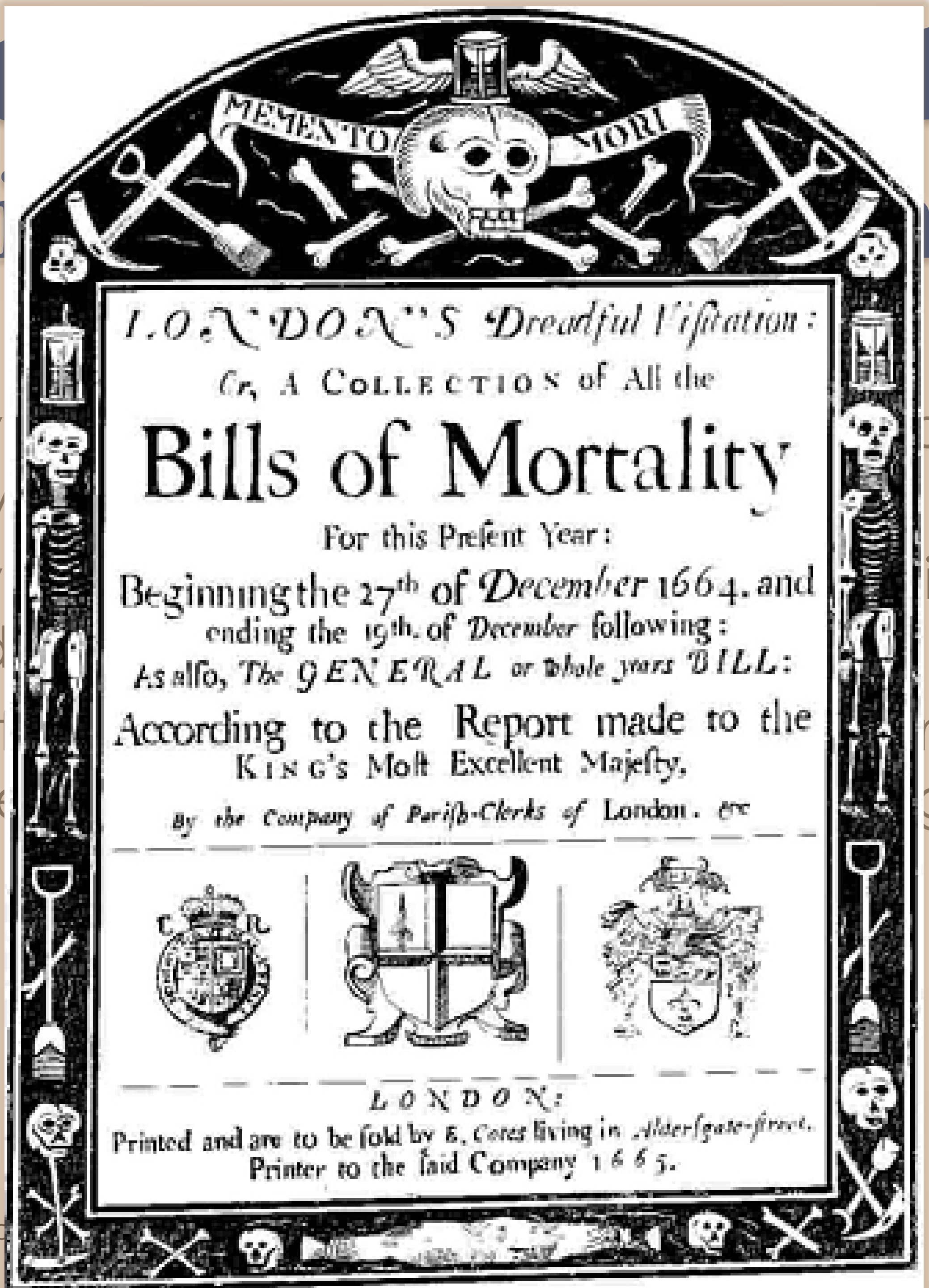
Institute of Mathematics and Informatics,
Vilnius University, Lithuania

Setting the scene - data mining in healthcare

- The very first medical statistics published: bills of Mortality (dated 1665)
- The very first data mining related publication in PubMed database is dated 1984
- Starting from the 21st century many countries have chosen e-Health as a priority national program

Sett lata
mi are

- The very bills of Mortality
- The very ion in PubMed
- Starting ries have chosen gram


MEMENTO MORI
C R

LONDON'S Dreadful Viſitation:

Or, A COLLECTION of All the

Bills of Mortality

For this Preſent Year:

Beginning the 27th of December 1664. and ending the 19th. of December following:

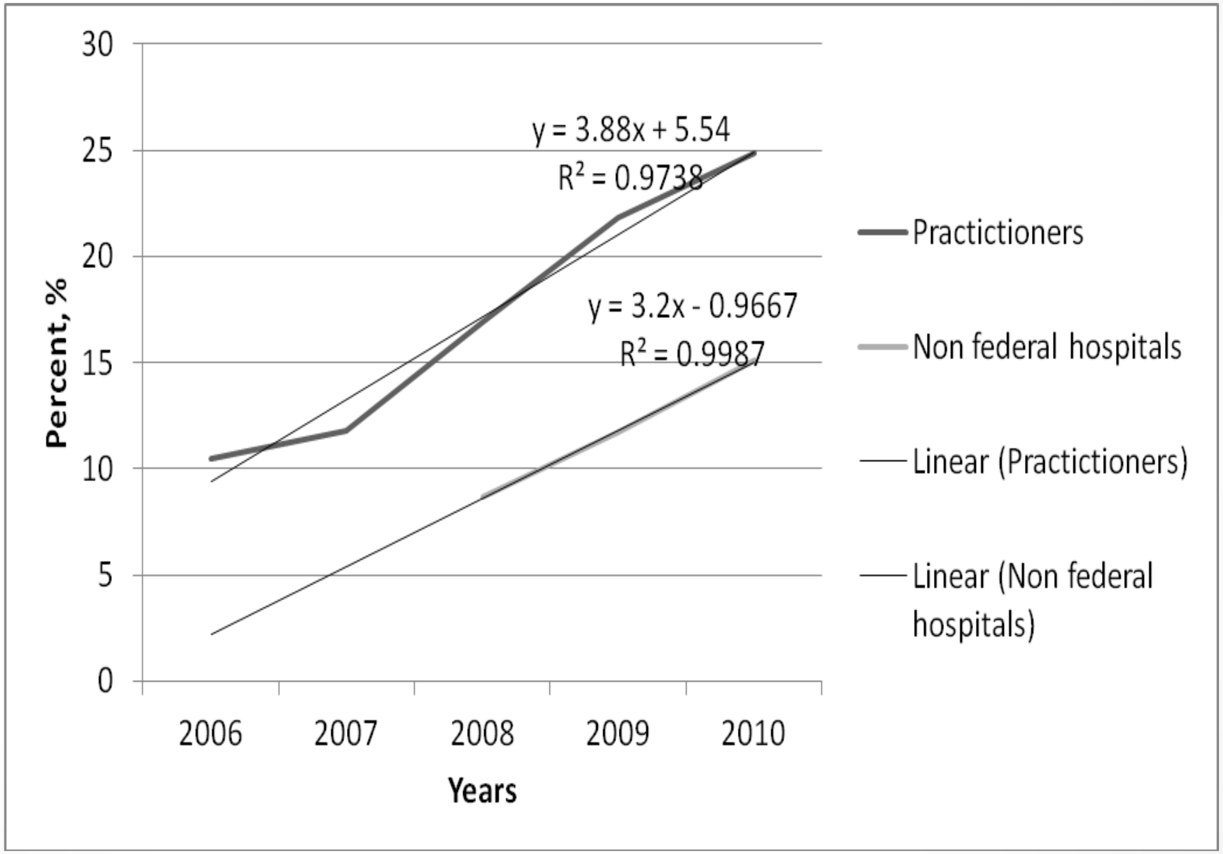
As alſo, The GENERAL or whole years BILL:

According to the Report made to the KING's Moſt Excellent Majeſty.

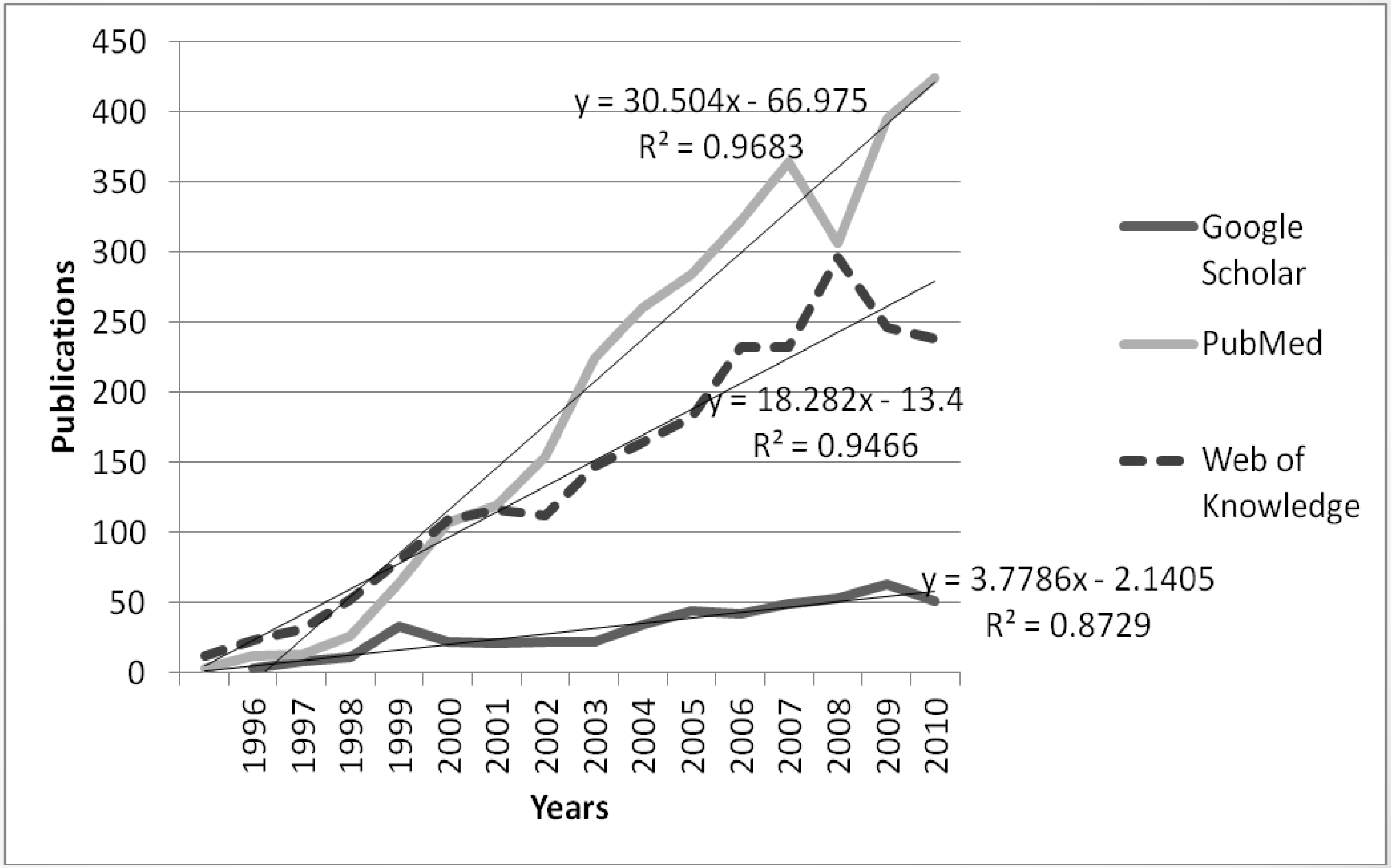
By the Company of Pariſh-Clerks of London. &c

LONDON:
Printed and are to be ſold by E. Cotes living in Alderſgate-ſtreet.
Printer to the ſaid Company 1665.

EHR adoption in the USA



Trend lines of DM applications in medicine related publications



Survey on data mining application

- In our survey we revealed, that the greatest part of medical community of tertiary hospitals have either minimal or zero awareness of the DM practical usage and its potential possibilities.
- Only 29% of healthcare representatives were able to provide any example of practical DM usage
- 86% of respondents expressed their interest in DM and even more would like to participate in international DM research projects.

Motivation

- It is of great interest to understand, what topics present in medical data mining research today, and how they do evolve?
- How to practically apply DM in multi-relational settings?

PubMed database and MESH

- PubMed database is comprised of more than 21 million citations for biomedical literature from MEDLINE, life science journals, and online books
- The Medical Subject Headings (MeSH) is a **controlled vocabulary** produced by National Library of Medicine and is used for **indexing**, **cataloging**, **and searching** for biomedical and health-related information and documents.

Intro to MESH concepts

- *Descriptor* is used to index citations in MEDLINE database, and for cataloging of publications. Most *Descriptors* indicate the subject of an indexed item, such as a journal article.
- A *Descriptor* is broader than a *Concept* and consists of a class of concepts.
- *Concepts*, in turn, correspond to a class of *Terms* which are synonymous with each other.
- Terms are attached to a broadest category of Semantic Type.

Multi-relational data mining

- Multi-relational presentation is natural
- Medical data is naturally and highly *relational. Holistic vs medical approach.*
- *Overall strategies to deal with MRDM:*
  - *Scaling down to single-table*
  - *Upgrading traditional DM algorithms*
  - *Utilizing first-order logic to induce rules (ILP)*

Clustering in multi-relational settings

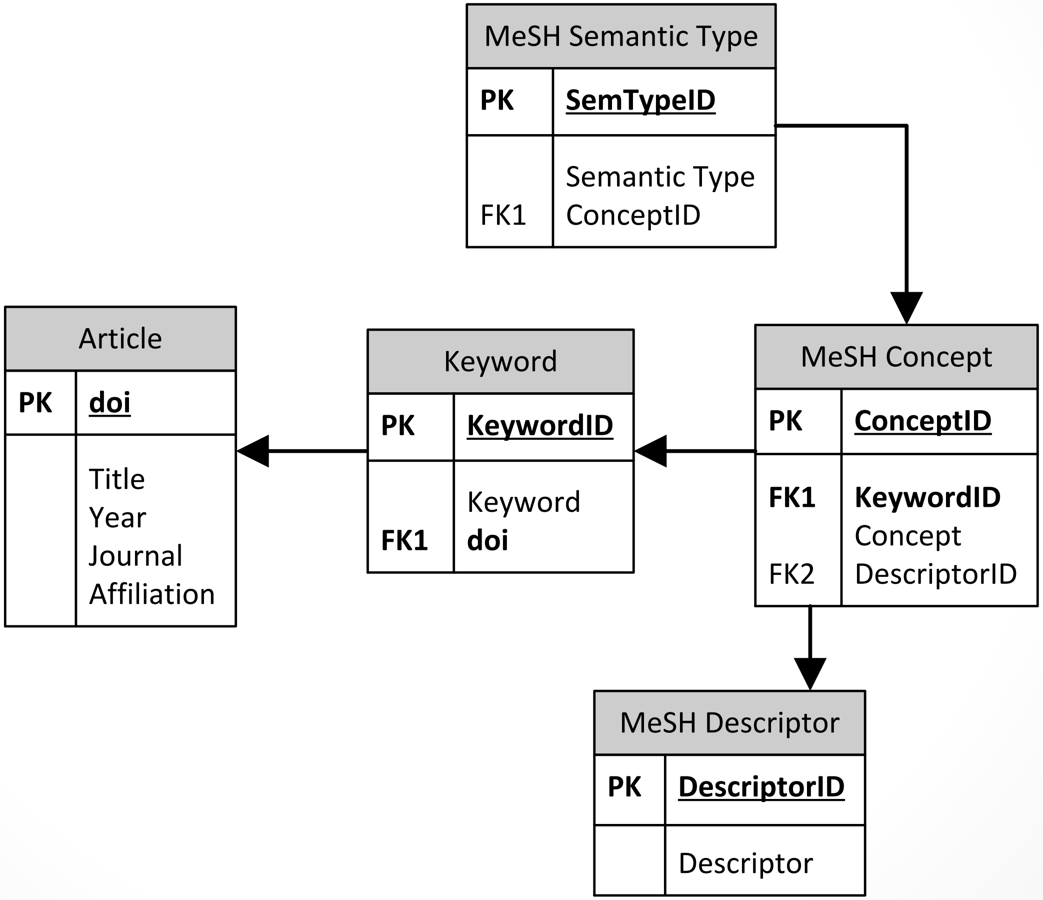
- There are many ways…
- Distance based clustering
- According to Van Laer and De Raedt, when upgrading propositional algorithm to the first-order learners type, it is important to retain as much of the original algorithm as possible, and only the key notion should be updated.
- Distance measure (dissimilarity measure) – key notion for distance based clustering approaches.

MRDM clustering approach - informally

- In our study, first-order instances of Articles A are represented by the predicate *Article* A, and the following ground atoms: Concept - C, Descriptor - D, and Semantic type - S.
- Example:
  - I = A (art1),

  with defined background knowledge BK:
  - C(art1, “Benpen”),
    D(art1, “Penicillin G”),
    S(“Benpen”, “Antibiotic”).

E-R diagram



MeSH Semantic Type

PK	**SemTypeID**
 FK1	Semantic Type ConceptID

Article

PK	**doi**
	Title Year Journal Affiliation

Keyword

PK	**KeywordID**
 FK1	Keyword **doi**

MeSH Concept

PK	**ConceptID**
FK1 FK2	**KeywordID** Concept DescriptorID

MeSH Descriptor

PK	**DescriptorID**
	Descriptor

Distance measure for MRDM clustering

- In complex data structures, there can be no objectively “best” distance or similarity measure, or at least formal proof would be too expensive.
- We propose to combine Gower and Ochiai-Barkman coefficient for a dissimilarity measure calculation

Gower general coefficient of similarity

- Gower's coefficient of similarity si is defined as follows:

$$s_{i,j} = \frac{\sum_k w_k s_{ijk}}{\sum_k w_k}$$

- For nominal variables:

s_{ijk} = 1, iff x_{ik} = x_{jk} , and s_{ijk} = 0, when $x_{ik} \neq x_{jk}$

- For numeric variables:

$$s_{ijk} = 1 - |x_{ik} - x_{jk}|/r_k$$

- Binary data type in Gower metric can be treated as a nominal data type, where, s_{ijk} = 1, iff the compared values equals to 1.

Ochiai-Barkman coefficient

To compare 2 nominal value lists in the case of comparing objects with one-to-many relations, we propose to use Ochiai-Barkman coefficient:

$$s_{l1,l2} = \frac{n(l_1 \cap l_2)}{\sqrt{n(l_1) \times n(l_2)}}$$

where l_1, l_2 – nominal value lists, n(l) – the number of elements in l.

Case study

We have constructed the following compound similarity measures to compare two instances of article A:

$$sim_{A1,A2} = \frac{w_c simC + w_d simD + w_s simS}{\mathrm{w_c} + \mathrm{w_d} + \mathrm{w_s}},$$

where

$$\mathrm{simC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij}(concept_i(A_1), concept_j(A_2))}{\sqrt{m \times n}},$$

$$\mathrm{simD} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij}(descriptor_i(A_1), descriptor_j(A_2))}{\sqrt{m \times n}},$$

$$\mathrm{distS} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij}(semantictype_i(A_1), semantictype_j(A_2))}{\sqrt{m \times n}}$$

Weights calculation

Possible options:

1. Not to use weights
2. Weights proportional to the tuples in the related entities
3. Expert based weighting
4. Other…

$$w_c = \frac{n_c}{n_c+n_d+n_s},\ w_d = \frac{n_d}{n_c+n_d+n_s},\ w_s = \frac{n_s}{n_c+n_d+n_s}$$

Case study realization

- The algorithm, calculating full dissimilarity matrix for the set of articles, has been implemented in R. Totally 2.284.453 similarity values have been calculated.
- R libraries "cluster" and "fpc" were used, for the different partitioning around medoids (PAM) implementations.
- Due to a large search space, extended by joined relations, the algorithm requires a vast computational power.
- The algorithm is well scalable, and our further step will be parallelization of this algorithm.

Case study outcomes

- For the evaluation of the overall clustering quality, cluster's *silhouette* value has been used. The *silhouette* value depicts the quality of each object's cluster.

$$silh_i = \frac{d\_min(i) - a(i)}{\max\{a(i), d\_\mathrm{min}(i)\}}$$

- In our case, the maximum achieved *silhouette* values were in the range: 0.20 - 0.30.
- Objectively, that means **the overall clustering result is unsatisfactory**, and shows that the found clusters are poorly describing the data set.

Case study outcomes

- Regretfully, there is no point of reference or golden standard to compare our results with. Therefore, comparison to other possible clustering methods is planned for further research step.
- The presented approach **has not been formally tested yet** and requires **further experiments** and formal evaluation. **Initial comparison tests** have been made by using the same use case data converted to a propositional form and applying k-means, PAM, and CLARA clustering algorithms. Still it has resulted in another set of low quality clusters, with less interesting practical information.

Discussion

- The next planned research activity will include approbation with classified multi-relational data sets and comparison to another clustering methods.
- The main known shortcoming of the implemented algorithm is its overall performance, due to the applied greedy approach.
- In other cases large data clustering algorithms CLARA or CLARANS might be used instead of PAM.

Thank you for your attention!

Any questions…?