Advanced Data-driven Techniques for Mining Expertise

Milena Angelova, Veselka Boeva[1] and Elena Tsiporkova[2]

1 Blekinge Institute of Technology, Karlskrona, Sweden

2 Sirris, The Collective Center for the Belgian technological industry, Brussels, Belgium

Milena Angelova

Ph.D. student at Technical University of Sofia, Plovdiv branch

Contacts:

Email: mangelova@tu-plovdiv.bg

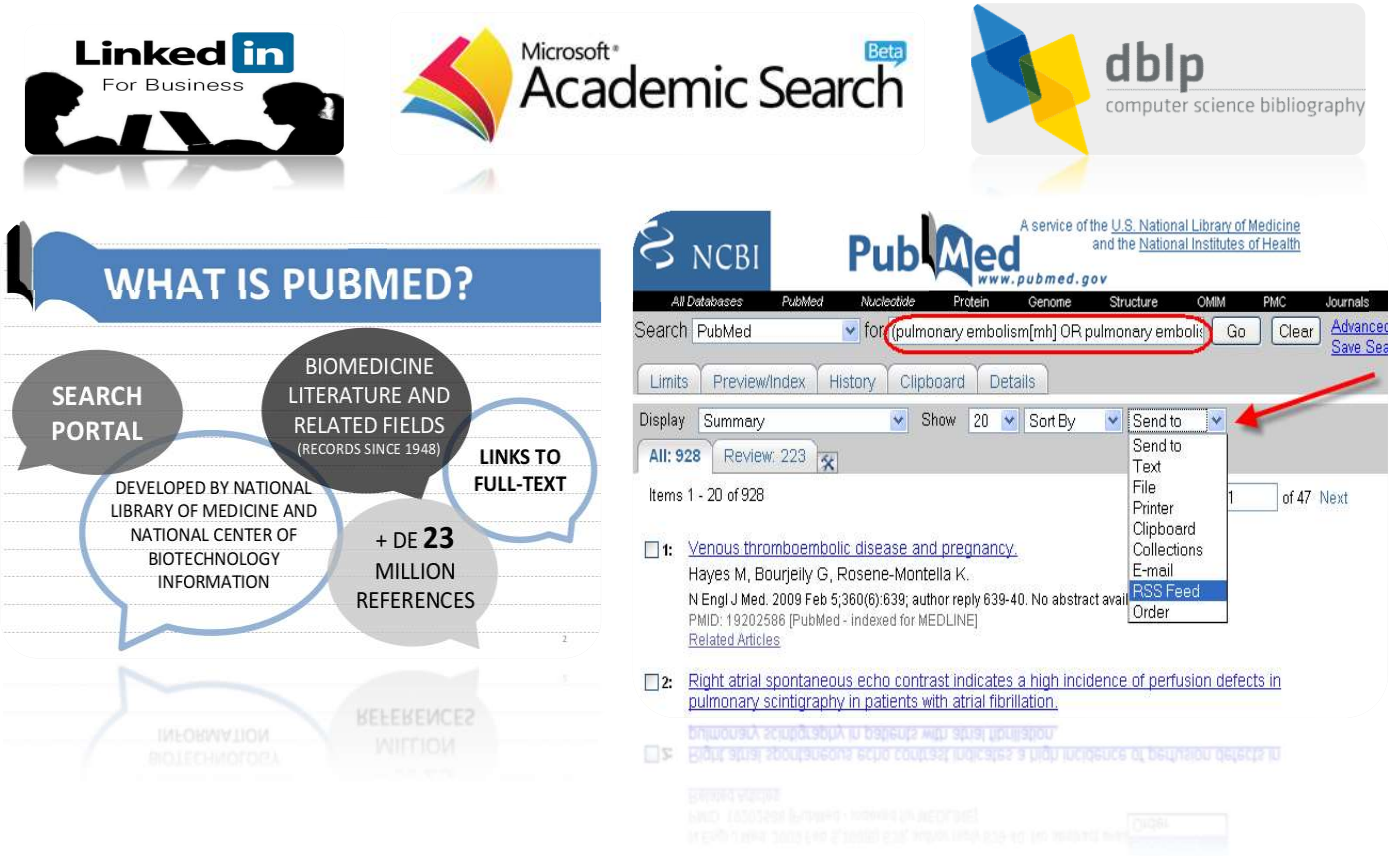
LinkedIn: https://www.linkedin.com/in/milena-angelova-7254966a

Introduction

- PubMed data
- Expert representation
  - Expertise Retrieval
    - Expert finding
    - Expert profiling
- Identification of subject experts
  - Resolve the problem with ambiguity
- Weighting method for assessing of expertise
- Estimate the expertise similarity between experts
- Formal Concept Analysis
- Evaluation metrics
- Results
  - Weighting of expertise
  - FCA based expert clustering

PubMed data

The data needed for constructing the expert profiles could be extracted from various Web sources.



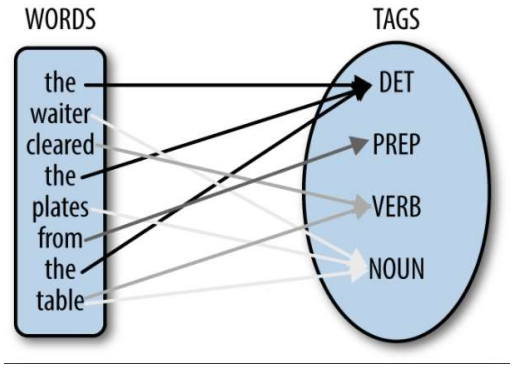
Expertise Retrieval

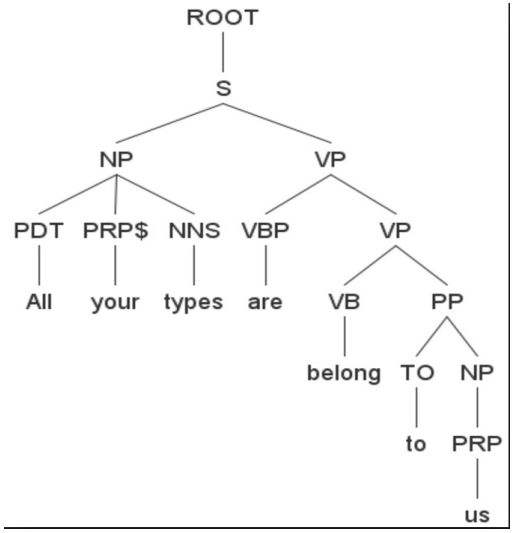
- The task of finding the right person with the appropriate skills and knowledge with respect to a topic.
- For example, given a textual topic (e.g., "Expertise Retrieval"), rank experts in descending order of expertise.

Expert profiling and

Expert finding

- Expert finding: Expert finding experts are the task of finding the right person with the appropriate skills and knowledge: "Who are the experts on topic X?"
- Expert profiling: While the task of expert finding is concerned with finding experts given a particular topic, the task of experts profiling turns around and asks "What topics do a person know about?"
- For instance, Stanford part-of-speech tagger can be used to annotate the different words.





Identification of subject experts

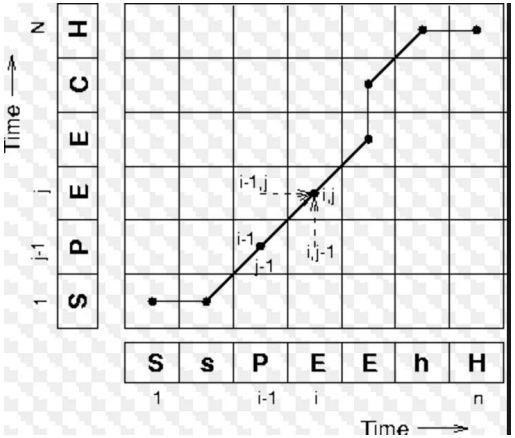
- An expert profile can be associated with information that includes: name, e-mails, address, affiliation, a list of publications, co-authors and etc.

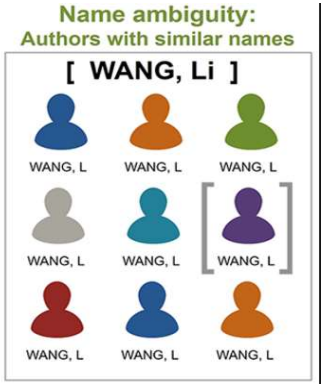
Personal data = { e-mails, name, affiliation, co-authors … }

List of keywords = { k_{i1}, k_{i2}, … k_{in} }, where i is i = 1, 2, 3, … n

The problem with ambiguity

- The expert's personal data can be used to resolve the problem with ambiguity. This problem refers to the fact that multiple profiles may represent one and the same person.
- Dynamic Time Warping based approach to deal with the ambiguity issue.





Weighting method

- There is no standard or absolute definition for accessing expertise.
- Weighting method – assess the levels of expertise of an expert to the domain-specific topics
- Each keyword k_{ij} in expert i is associated with a weight w_{ij}
- Each expert can be presented by two components:
  - a list of describing keywords
  - a vector of weights

Expertise similarity

- It is a complicated task for calculation of expertise similarity.
- Similarity between two expertise profiles as a strength of the relations between the semantic concept associated with the keywords of two compared profiles.
- Take the similarities between any pair of keywords
- Definition of semantic similarity between corresponding keywords:

$$S_{ij} = \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} W_{lm} \cdot s(k_{il}, k_{jm}),$$

Finding Similar Experts

- A small number of examples who have been used to work on similar problems
- For example , find the experts with the required expertise by entering the name of example expert and system will return a list of similar experts with close expertise
- Domain of interest can be presented by several preliminary specified subject categories and then the experts can be grouped with respect to this categories

Formal Concept Analysis

- FCA is a mathematical classification technique
  - Helps discover meaningful data in binary relations
  - Can be visualized with Concept Lattices
  - Input: A context <O, A, R>
    - O is a set of objects
    - A is a set of attributes
    - R is a binary relation between O and A
  - Mapping:
    - Common attributes of a set of objects:

$$CA(O' \subseteq O) = \{a \in A \mid \forall o \in O' : (o, a) \in R\}$$

    - Common objects of a set of attributes:

$$CO(A' \subseteq A) = \{o \in O \mid \forall a \in A' : (o, a) \in R\}$$

    - Output: Concepts <O', A'> s.t.

$$\begin{cases} CA(O') = A' \\ CO(A') = O' \end{cases}$$

Formal Concept Analysis

- A formal context consists of the set of the n experts, the set of main categories { C_1, C_2, …, C_k}
- The formal concept is defined as a pair (X, Y) , where:
  - X ⊆ $of\ experts\ and\ Y \subseteq of\ categories$, ∀ expert ∈ ∀ area in Y
  - ∀ expert ∉ X, there is a subject area in Y that does not contatin that expert
  - ∀ subject area that is not in Y, there is an expert in X who is not associated with the area
- The concept lattice represents a subset of experts belonging to a number of subject areas
- The set of all concepts partitions the experts into a set of disjoint expert areas

Evaluation metrics

- The similarity between two different expertise retrieval results, it can be assessed by:
  - Resemblance r

$$r(S_i', S_i) = |S_i' \cap S_i| / |S_i' \cup S_i|,$$

  - Containment *c*

$$c(S_i') = |S_i' \cap S_i| / |S_i'|$$

  - Silhouette Index is defined as:

$$s(C) = 1 \Big/ m \sum_{i=1}^{m} (b_i - a_i) \Big/ \max\{a_i, b_i\},$$

Data extraction and preprocessing

- Extracted a set of 4343 Bulgarian authors from PubMed
- After resolving the problem with ambiguity they are reduced to 3753 different researches.
- Each author is represented by a list of MeSH headings and a vector of weights

Results – Weighting expertise

Experts	MeSH headings
1	Kidney Transplantation; Liver Transplantation
2	Health Behavior
3	Drinking; Health Behavior; Health Knowledge, Attitudes, Practice; Program Evaluation
4	Models, Biological; Temperature; Models, Neurological; Water
5	Computer Simulation; Models, Molecular; Protons; Thermodynamics; Molecular Conformation
6	Vibration; Models, Molecular; Infrared Rays; Hydrogen Bonding
7	Monte Carlo Method; Models, Theoretical; Phase Transition; Thermodynamics
8	Photosynthesis; Quantum Theory
9	Health Behavior; Decision Support Techniques; . . .(more than 20 MeSH terms)
10	Polymorphism, Genetic

Table 1 Expert MeSH heading profiles.

Experts	MeSH heading weights
1	0.5; 0.5
2	1
3	0.25; 0.25; 0.25; 0.25
4	0.166; 0.333; 0.166; 0.333
5	0.285; 0.285; 0.142; 0.142; 0.142
6	0.5; 0.166; 0.166; 0.166
7	0.428; 0.285; 0.142; 0.142
8	0.75; 0.25
9	0.022;...; 0.045; . . . ; 0.068; . . . ; 0.25
10	1

Table 2 Expert MeSH heading weights.

Results – FCA based expert clustering

Category label	Category name	Number of authors
A	Anatomy	45
B	Organisms	101
C	Diseases	68
D	Chemicals and Drugs	158
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment	663
F	Psychiatry and Psychology	97
G	Phenomena and Processes	797
H	Disciplines and Occupations	38
I	Antropology, Education, Socialogy and Social Phenomena	14
J	Tehnology, Industry, Arguculture	20
K	Humanities	2
L	Information Science	37
M	Named Groups	1
N	Health Care	125

Table 3 Number of authors partitioned into the main MeSH categories (singleton concepts).

United categories	Number of authors
{G, N}	106
{E, N}	55
{C, G}	59
{E, L}	36
{F, N}	23
{F, I}	12
{E, G, N}	56
{E, H, J, L}	8
{G, H, L, N}	6
{E, G, I, L, N}	11
{F, G, H, I, N}	7

Table 4 Number of authors partitioned into united MeSH categories (non-singleton concepts)

Thank you for your attention!