AI Methods for Building University Research Knowledge Base

Scoring scientist using knowledge databases

Henryk Rybiński, Jakub Koperwas, Łukasz Skonieczny, Marek Kozłowski, Piotr Andruszkiewicz, Wacław Struk

Institute of Computer Science, Warsaw University Of Technology

September 9, 2019

Agenda

- From Institutional Repository…
- …to University Knowledge Base.
- AI tools
  - semantic processing
    - information extraction
    - publication classifier
    - keywords extraction
    - sense indexing
  - Profiling researchers
    - person name disambiguation
    - finding and ranking an expert

Institutional Repository (2010)

- Storing scientific information (meta data with full documents) – papers, theses, reports, patents, projects
- Exposing and sharing (with clean GUI, REST API and OAI)
- Advanced querying

- Key technology: Apache Jackrabbit

(free implementation of **Content Repository API for Java - JCR)**

Knowledge Base in 2013

Ω-Ψ^R = CRIS+IR +RPS

IR

- Run by libraries
- Long-term preservation
- **Providing access t**o full text
- Dublin Core
- **Limited to OA**
- **Self-archiving**

CRIS

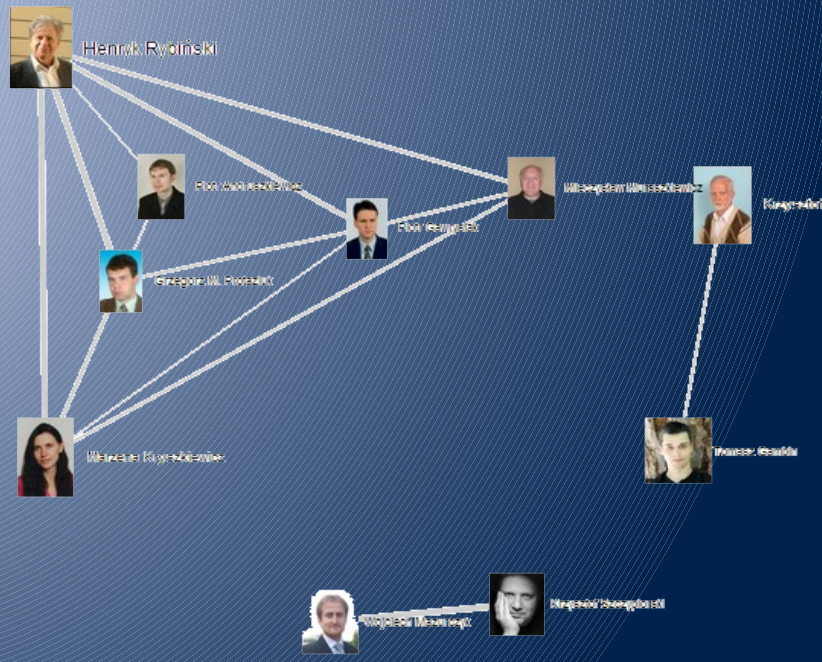
- Run by adminstration
- Covers all the research activities
- Metadata (CERIF)
- **Rather bibliogr. description**
- **Reporting**
- **Research assessment**
- **(e.g. IF vs costs)**

RPS

- Run by scientists
- Covers all the research activities
- Building profiles
- **Vizualizes the activities and cooperation**
- Social networks functions
- **No evaluation**

University Knowledge Base

- Relations and connections between objects: authors, papers, books, projects, journals, conferences
- Auto-generated information pages for researchers and units (journals and conferences in the near future)
  - with statistics regarding research activity
  - with research area word-cloud
  - with cooperation graph
- Social network elements
  - keep your profile up-to-date
  - share your scientific achievements
  - find and contact researchers



Henryk Rybiński, PhD, DSc, Professor

Professor

The Institute of Computer Science

Faculty of Electronics and Information Technology

Email: H.Rybinski@ii.pw.edu.pl

Phone: +48 22 234-7432, fax +48 22 234 6091

Room no: 204

Consultations: Monday 14.00-16.00

Researcher Report

Publications	92
PhD theses	13
Participation in projects	30
Supervised BSc and MSc theses	2
Professional activity	9

h-index*:12

MSc (1970), PhD (1974), DSc (1988), Tenured Professor (2001); Specialization: information systems, knowledge representation, data and text mining, databases, Professor, Director of the Institute (2008-), Head of Division of Information Systems (1994-2008), Co-ordinator of the Curriculum on Software Engineering and Information Systems (1994-2008), Co-ordinator of the Subject Class "Databases and Information Systems" (1995-2001), voting member of ACM and SIGMOD (1989-), Affiliate Member of IEEE (1990-1996); Member of several programme committees of international conferences and workshops, among others: IIS, ISMIS, IIPWM, AM, SWC, RSFDGRC, RSKT, TKE, PKDD, PAKDD, MCD; member of CREST Working Group; expert and consultant of many UN agencies and European Commission; Member of Informatics Committee at Polish Academy of Science (2011-); Editorial Board Member of the Journal of Intelligent Information Systems (2012-); Editorial Board Member of the International Journal of Social Network Mining (2012-); Chair of the Rector's Committee for the strategy of developing ICT infrastructure for WUT in 2013-2020.

** presented value of the Hirsch index is approximative calculation obtained in the Repository based on the scientist's publications (including autocitations) in the Repository and Internet information analysis. The value is close to the value obtained with the Publish or Perish system. In general it is higher than the value given by the Scopus or Web of Science sites. In the case of undervalued number, first of all take care of completeness of the Repository.*

Back

Henryk Rybiński, PhD, DSc, Professor

Professor

The Institute of Computer Science

Faculty of Electronics and Information Technology

Email: H.Rybinski@ii.pw.edu.pl

Phone: +48 22 234-7432, fax +48 22 234 6091

Room no: 204

Consultations: Monday 14.00-16.00

Publications

Number of records: 92.

Order by: type/author ▾ Export 0 as: Ankieta 2013 ▾

Authored books

Rybiński Henryk: Modele baz danych, vol. 63, 1987, Centrum Informacji Naukowej, Technicznej i Ekonomicznej, 156 p.
Google

Rybiński Henryk: Problem optymalizacji reorganizowania zbioru informacyjnego w systemie wyszukiwania informacji, 1976, IINTE, 92 p.
Google

Edited books

Bembenik Robert, Skonieczny Łukasz, Rybiński Henryk, Niezgódka Marek (*eds.*): Intelligent Tools for Building a Scientific Information Platform, Studies in Computational Intelligence, vol. 390, 2012, Springer, ISBN 978-3-642-24808-5, 277 p., DOI:10.1007/978-3-642-24809-2
Google

Bembenik Robert, Skonieczny Łukasz, Rybiński Henryk, Kryszkiewicz Marzena, Niezgódka Marek (*eds.*): Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions, Studies in Computational Intelligence, vol. 467, 2013, ISBN 978-3-642-35646-9, 548 p., DOI:10.1007/978-3-642-35647-6
Google

Bembenik Robert, Skonieczny Łukasz, Rybiński Henryk, Kryszkiewicz Marzena, Niezgódka Marek (*eds.*): Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation, Studies in Computational Intelligence, vol. 541, 2014, Springer International Publishing, ISBN 978-3-319-04713-3, [978-3-319-04714-0], 290 p., DOI:10.1007/978-3-319-04714-0
Google

Kryszkiewicz Marzena, Rybiński Henryk, Skowron Andrzej, Raś Zbigniew W. (*eds.*): Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence, vol. 6804, 2011, Springer, ISBN 978-3-642-21915-3, 746 p., DOI:10.1007/978-3-642-21916-0
Google

Kryszkiewicz Marzena, Peters James F, Rybiński Henryk, Skowron Andrzej (*eds.*): Rough Sets and Intelligent Systems Paradigms, Lecture Notes in Artificial

Henryk Rybiński, PhD, DSc, Professor

Professor

The Institute of Computer Science

Faculty of Electronics and Information Technology

Email: H.Rybinski@ii.pw.edu.pl

Phone: +48 22 234-7432, fax +48 22 234 6091

Room no: 204

Consultations: Monday 14.00-16.00

hindex = 12, cited by total = 379

ontologies (artificial intelligence)

pub title	cited	
Finding Reducts in Composed Information Systems	70	19/04/2014
Reducing information systems with uncertain attributes	31	19/04/2014
Computation of Reducts of Composed Information Systems	27	19/04/2014
On first order logic databases	27	03/02/2014
Dataless transitions between concise representations of frequent patterns	24	19/04/2014
Data Mining in Incomplete Information Systems from Rough Set Perspective	21	19/04/2014
Discovering Synonyms Based on Frequent Termsets	17	19/04/2014
Text onto miner - a semi automated ontology building system	14	19/04/2014
Intelligent Tools for Building a Scientific Information Platform	14	22/04/2014
Discovering Compound and Proper Nouns	12	19/04/2014
Mining spatial association rules with no distance parameter	12	19/04/2014
Knowledge sharing in default reasoning based multi-agent systems	12	19/04/2014
Towards a unifying logic formalism for semantic data models	9	19/04/2014
Discovering Word Meanings Based on Frequent Termsets	9	19/04/2014
Distributed default logic for multi-agent system	9	19/04/2014
Automatic Index Selection in RDBMS by Exploring Query Execution Plan Space	7	24/04/2014
Data Mining for Technical Operation of Telecommunications Companies: a Case Study	7	12/05/2014
Incomplete database issues for representative association rules	6	19/04/2014
Multilevel information system- Towards more flexible information retrieval systems	5	17/02/2014
Word sense discovery for web information retrieval	5	23/03/2014
Learning Mechanism for distributed default logic based MAS - implementation considerations	4	17/02/2014
Extending open source software solutions for CRM text mining	3	11/05/2014
A new approach to computing weighted attributes values in incomplete information systems	3	17/02/2014
A Distributed Decision Rules Calculation Using Apriori Algorithm	3	18/02/2014
Legitimate Approach to Association Rules under Incompleteness	3	03/02/2014
Methods and Tools for Ontology Building, Learning and Integration – Application in the SYNAT Project	3	17/02/2014
Regression - yet another clustering method	3	09/05/2014

Henryk Rybiński, PhD, DSc, Professor

Professor

The Institute of Computer Science

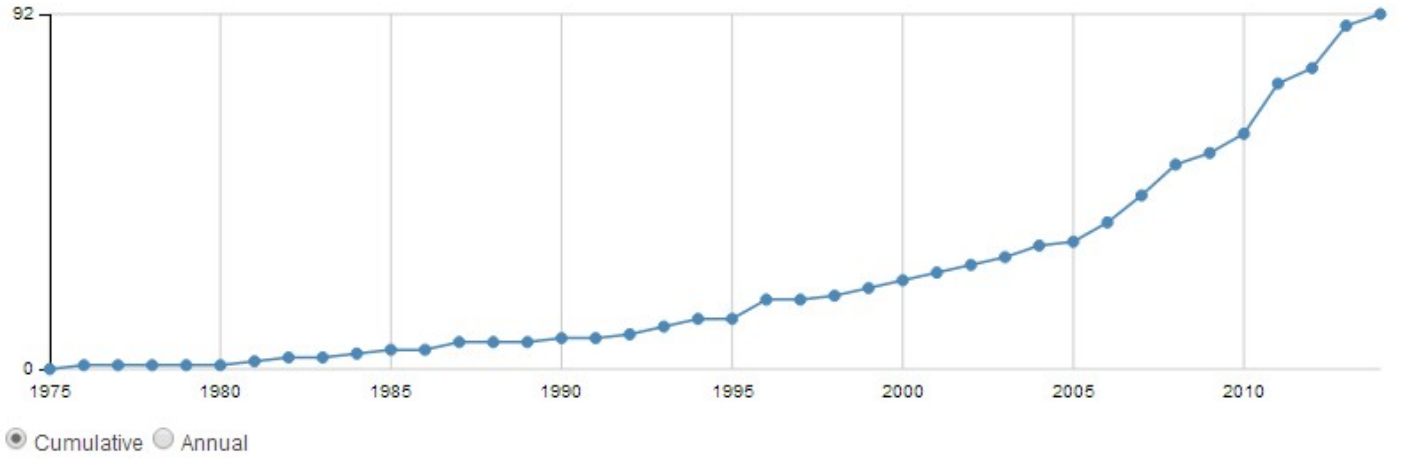
Faculty of Electronics and Information Technology

Email: H.Rybinski@ii.pw.edu.pl

Phone: +48 22 234-7432, fax +48 22 234 6091

Room no: 204

Consultations: Monday 14.00-16.00



** presented value of the Hirsch index is approximative calculation obtained in the Repository based on the scientist's publications (including autocitations) in the Repository and Internet information analysis. The value is close to the value obtained with the Publish or Perish system. In general it is higher than the value given by the Scopus or Web of Science sites. In the case of undervalued number, first of all take care of completeness of the Repository.*

Back

Problem

Manual maintenance of such a semantic network is not easy, if possible at all.

So AI methods are of crucial importance

Semantic processing

Goals:

1. enrich data by adding semantically meaningful descriptors

2. With enriched texts discover research areas (for authors, teams or institutions) and build vector representation of research

3. Given the vector representation visualize the research

AI Tools

Semantic processing

Methods

1. Tagging publications with OSJ categories
2. Keywords extraction
3. Sense indexing

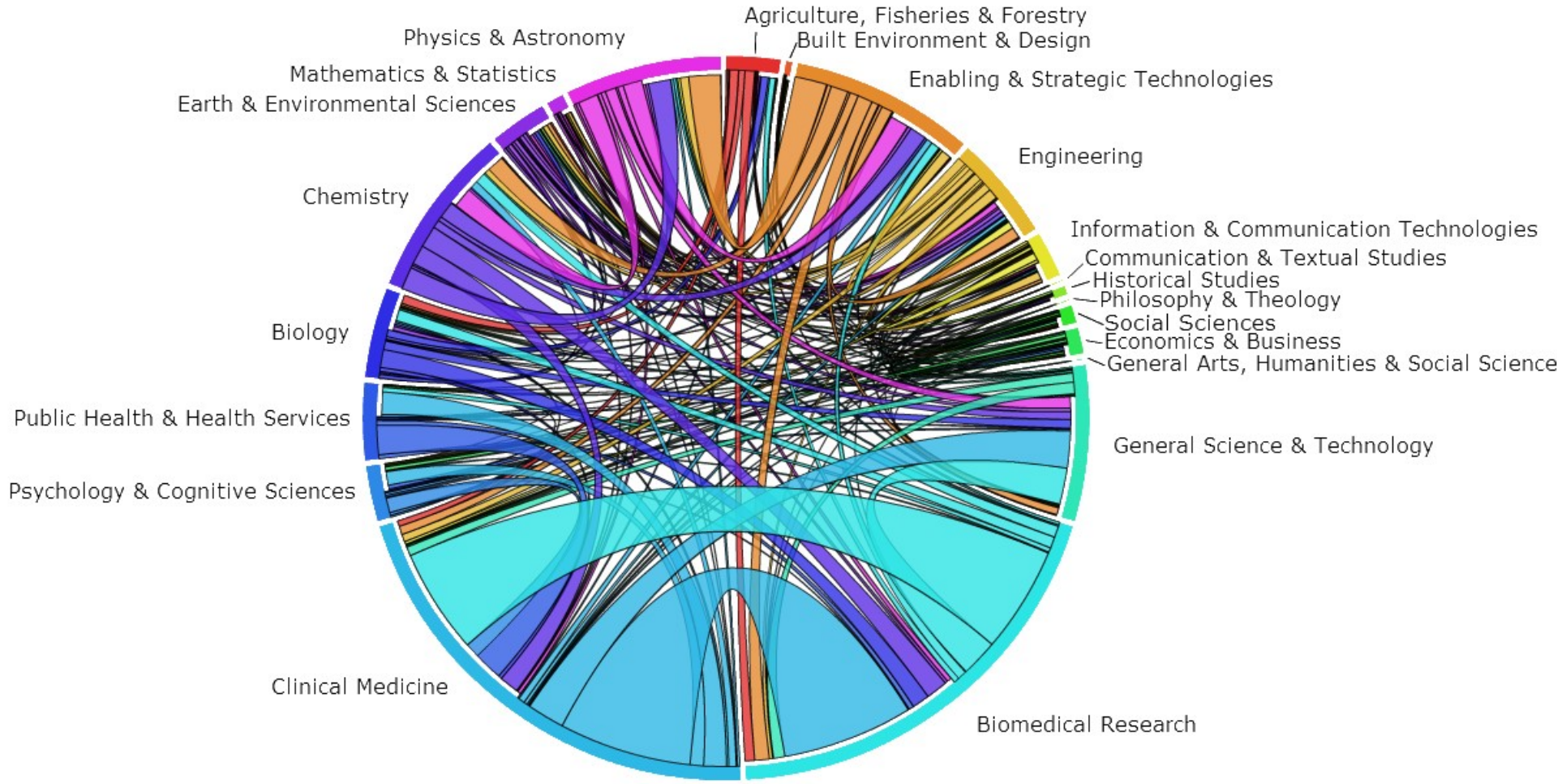
OSJ tagging

- OSJ – ontology of scientific journals
- Created by Science-Metrix, Canada, 2011
- 15000 journals, hierarchical (3 levels) classification
- Ready to use for known journals

  but…
- a classifier is needed for other journals, conference papers, and book chapters

Our solution: tree of bayesian classifiers, 85% accuracy

OSJ – level 2


Agriculture, Fisheries & Forestry
Built Environment & Design
Physics & Astronomy
Mathematics & Statistics
Earth & Environmental Sciences
Enabling & Strategic Technologies
Engineering
Chemistry
Information & Communication Technologies
Communication & Textual Studies
Historical Studies
Philosophy & Theology
Social Sciences
Economics & Business
General Arts, Humanities & Social Science
Biology
Public Health & Health Services
General Science & Technology
Psychology & Cognitive Sciences
Clinical Medicine
Biomedical Research

Keyword Extraction from Document

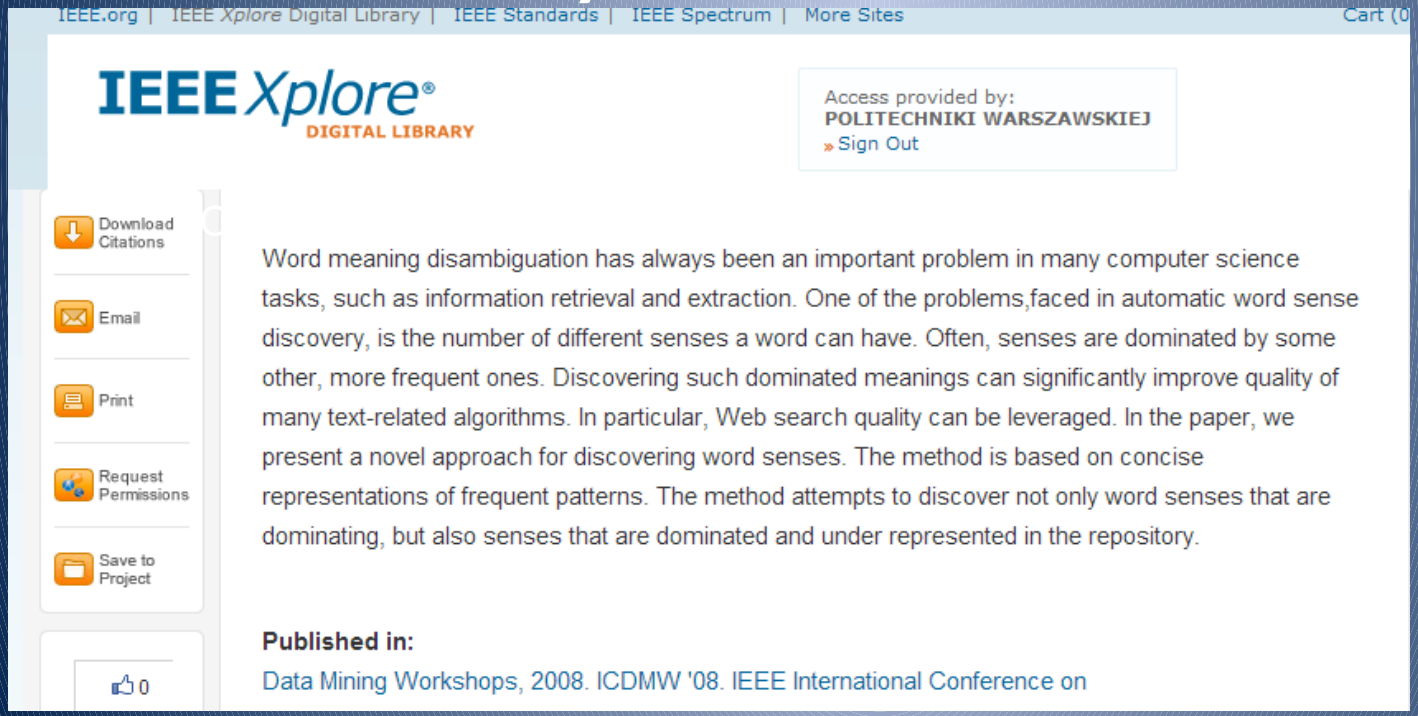
1. TKE method was proposed based on RAKE (Rose et al. 2010) and KEA (Witten et al. 1999);
2. Additionally Wikipedia is used for adding more general or synonymic keywords (senses, translations, summaries)
3. Extracted terms are enriched with context that eliminates disambiguity (SnS)

Keywords extraction

The algorithm:

1. identifies wikipedia entry corresponding to given candidate term
2. processes the entry in order to extract context: labels, senses, translation, summaries
3. evaluates source document with respect to the context found in previous step
4. taging keywords with meaning dicovered by SenseSearcher algorithm

Keywords extraction



Extracted terms:
information, information retrieval, **web search engine,**
computer science, word sense, World Wide Web

Profiling of researcher

$$keywords(d) = OSJ(d) \cup EXTRACT(d)$$

$$score(k\quad) = \sum_{d \in D(p)} rel(k,d) \times (sif(d) + 1)$$

- $rel(k,d)$ measures the relevance of keyword k with respect to document d; Usually it is the value of $tf\text{-}idf$; however, in the case of publications and technical reports the values of the OSJ keywords and the keywords provided by the authors are boosted;
- $sif(d)$ is a scientific impact factor of the document d. For the journal papers this is a linear combination of the impact factor of the journal and the citations of the document. Arbitrary values are given to other objects, like conference papers, patents, supervised theses.

Profiling of an institution

Given vectors of researchers affiliated at an institution we build a vector for the institution

Clearly, the same algorithm can be used for building an expertise vector for any subset of documents. So, given a query q we can obtain the set of documents $D(q)$ and calculate the score vector the same way as $D(p)$ for researcher p. This means that we can easily obtain an aggregated cloud of research interest for a faculty, department, as well as for a whole university. The profiles for a whole university and two different faculties can be seen in Fig. 4.

Visualization of vectors

Word clouds visualizing Unversity and 2 faculties: Chemistry, Physics

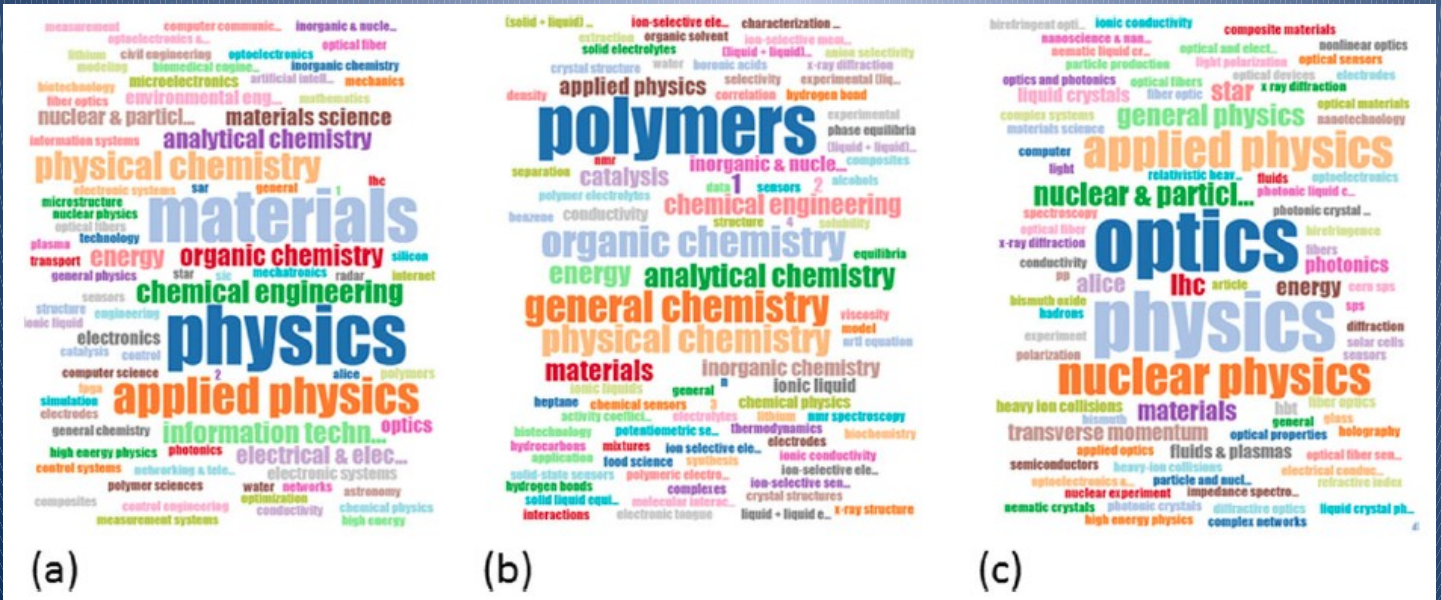


Fig. 4 Aggregated expertise clouds for a university (**a**), and two faculties (**b**) and (**c**)

Search for best experts in a domain

Step 1

– all the knowledge base resources are searched with a specified search phrase q (formulated the same way as for searching publications, theses, etc.). The result of the query q is denoted by $D(q)$;

Step 2

– a set of all persons $P(q)$ related to items from $D(q)$ is calculated as follows:

$$P(q) = \bigcup_{d \in D(q)} \{p : role(p, d) > 0\} \quad (3)$$

where the function $role(p, d)$ provides a measure of relevance of role of person p in elaborating the document d; by $role(p, d) > 0$ we mean that p has some role in d, that is, p is an author of publication d, or is a supervisor of thesis d, or is a leader of project d, etc.; in particular, the function takes into account the roles which can be *article author*, *book author*, *book editor*, *phd author*, *phd supervisor*, *master thesis author*, *master thesis supervisor*, *project member*, *project leader*; in this way we can, e.g., value the role *book author* more than *book editor*;

Search for best experts in a domain

Step 3

– for each person $p \in P(q)$ the person score measure, denoted by $Pscore(p, q)$, is calculated:

$$Pscore(p, q) = \sum_{d \in D(q): role(p,d) > 0} score(p, d, q) \qquad (4)$$

where $score(p, d, q)$ is a function expressing the importance of d with respect to query q, and in relation to p; the function is calculated according to a selected ranking algorithm;
the set of persons $P(q)$ is sorted in descending order by $Pscore(p, q)$, and a list of top n persons is presented to the user.

Search for best experts in a domain

$$score(p, d, q) = rel(d, q) \times (sif(d) + 1) \times role(p, d) \quad (5)$$

where:

- $rel(d, q)$ is a measure of relevance of d with respect to query q; here we rely on the Lucene relevance score, which uses the cosine measure, with boosted values for the fields resulting from the semantic enrichment procedures (Section 5);
- $sif(d)$ is a scientific impact factor of d; it is a linear combination of the impact factor of the journal (for the journal papers) and its citations.

Literature

KOPERWAS, J.J., SKONIECZNY, Ł., KOZŁOWSKI, M., ANDRUSZKIEWICZ, P., RYBIŃSKI, H., AND STRUK, W. 2014. AI PLATFORM FOR BUILDING UNIVERSITY RESEARCH KNOWLEDGE BASE. LNAI, SPRINGER, 405–414.

KOPERWAS, J.J., SKONIECZNY, Ł., KOZŁOWSKI, M., ANDRUSZKIEWICZ, P., RYBIŃSKI, H., AND STRUK, W. 2017. Intelligent information processing for building university knowledge base.
Journal of Intelligent Information Systems 48, 1, 141–163.

KOZŁOWSKI, M. AND RYBIŃSKI, H. 2017. Word Sense Induction with Closed Frequent Termsets. *Computational Intelligence 33*, 335–367.

RYBIŃSKI, H., SKONIECZNY, Ł., KOPERWAS, J.J., STRUK, W., STĘPNIAK, J., AND KUBRAK, W. 2017. Integrating IR with CRIS – a novel researcher-centric approach. *Program-Electronic Library and Information Systems 51*, 3, 298–321.

KOZŁOWSKI, M. AND RYBIŃSKI, H. 2018. Clustering of semantically enriched short texts. *Journal of Intelligent Information Systems*, 1–24.

Thank you!

http://omega-psir.ii.pw.edu.pl

http://repo.pw.edu.pl

http://wizzar.ii.pw.edu.pl/RepoPW/