Semantic Search-by-Examples for Scientific Topic Corpus Expansion in Digital Libraries

SERecSys Workshop at ICDM 2017, New Orleans, USA

Hussein T. Al-Natsheh[1,2] Lucie Martinet[1,3] Fabrice Muhlenbach[4]
Fabien Rico[5] Djamel A. Zighed[1]

[1] Université de Lyon, Lyon 2, ERIC EA 3083, 5 Avenue Pierre Mendès France – F69676 Bron Cedex – France
[2] CNRS, Institut des Sciences de l'Homme FRE 3768, 14 avenue Berthelot – F69363 Lyon Cedex 07 – France
[3] CESI EXIA/LINEACT, 19 Avenue Guy de Collongue – F69130 Écully – France
[4] Université de Lyon, UJM-Saint-Étienne, CNRS, Laboratoire Hubert Curien UMR 5516 – F42023 Saint Étienne – France
[5] Université de Lyon, Lyon 1, ERIC EA 3083, 5 Avenue Pierre Mendès France – F69676 Bron Cedex – France

18 November 2017

Example : Multivariate Analysis or Machine Leaning ?

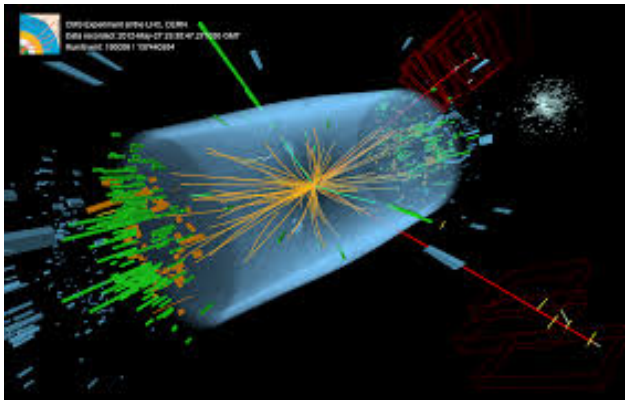


FIGURE – High Energy Physics Research Discipline

During my internship at CERN :

- What I know as **machine learning** in computer science and data mining
- Was referred to **multivariate analysis** in high energy physics

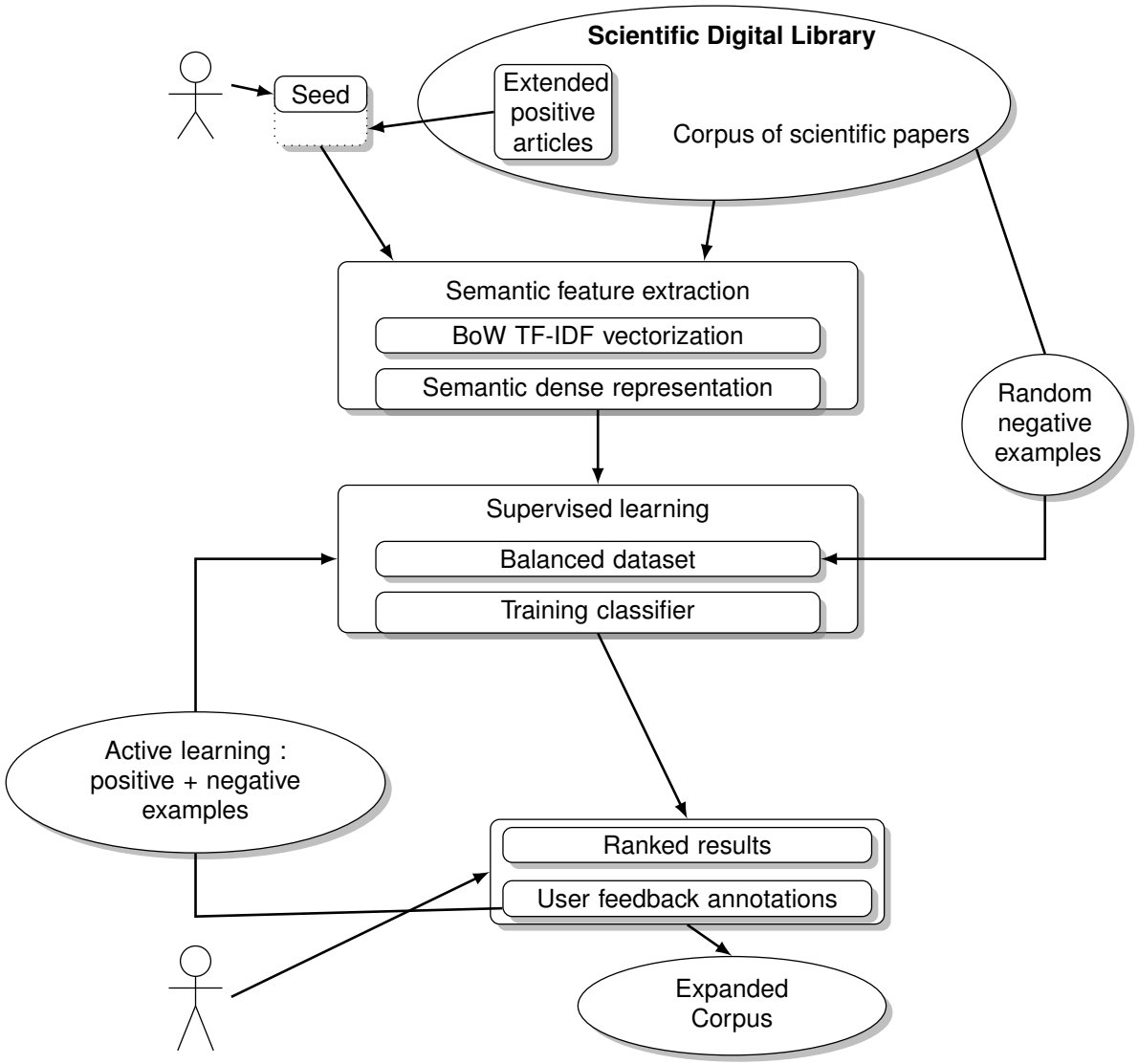
Problem Statement

Recommending semantically relevant documents from different disciplines

- **Different terminologies** used over different related disciplines
- Term usage **evolves** over time
- Total **number of articles** of a scientific topic
- No standard topic **categorization over publishers**
- **Limited** number of **keywords** per article
- Multi-disciplines expert **manual labeling** solution is too expensive

Outline

Model Overview



Vectorization

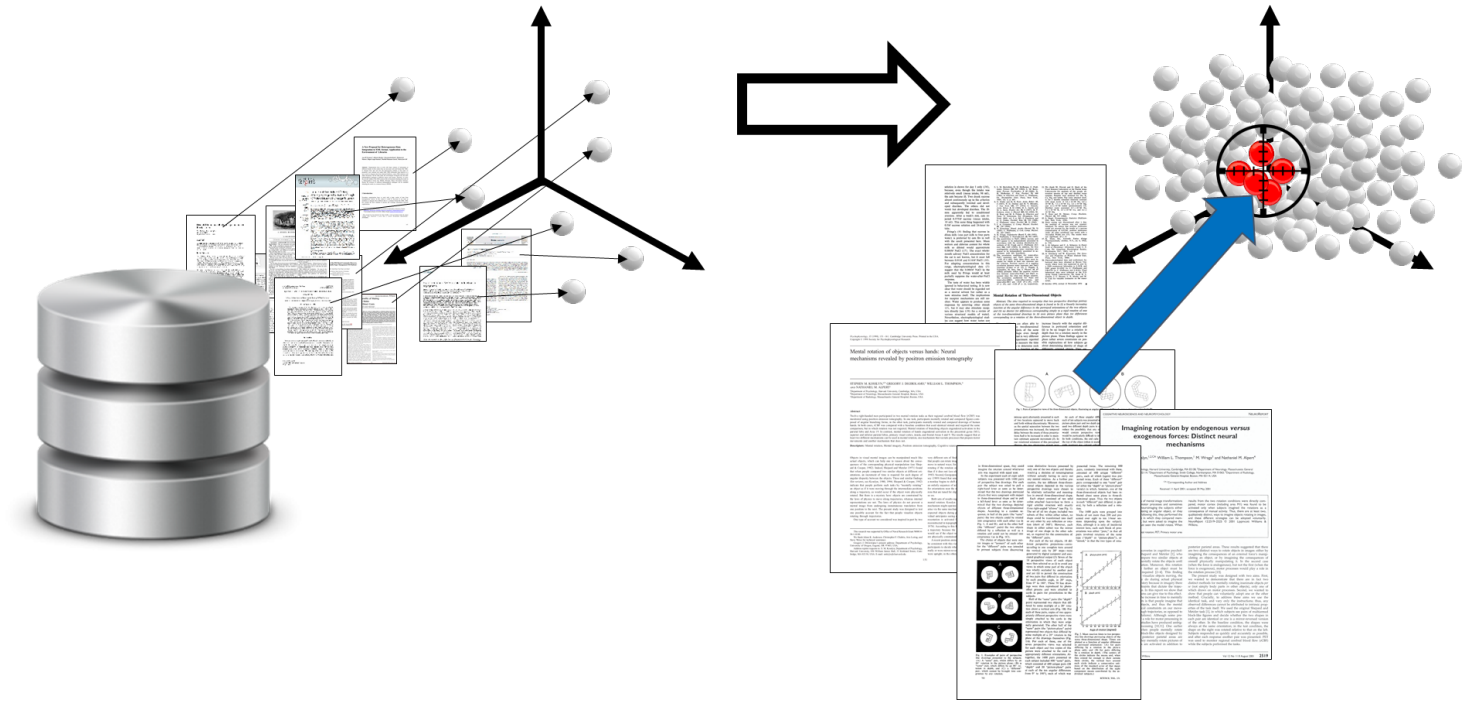


FIGURE – Document vectorization of the corpus and the seed articles (semantics feature extraction)

Eperementally Finding the Best Vector Transformer

Tuning vectorization parameters

$$AICS = \frac{\sum_{i=1}^{n-1} cosine\_similarity(list[i], list[i+1:n])}{number\_of\_comparisons} \quad (1)$$

$$argmax_{transformer}(AICS_{positive\_list} - AICS_{random\_list}) \quad (2)$$

Learning Process

Balanced Dataset Generation

- Positive examples as the seed articles and topic name keyword matching
- Negative examples as a random sampling from the corpus other than positives

Supervised Classifier

- Logistic regression or ensemble learning classifier
- Experimentally Designed (cross validation on the dataset)
- Rank the corpus by the regression model (or class belonging prob) prediction value

Corpus Semantic Expansion

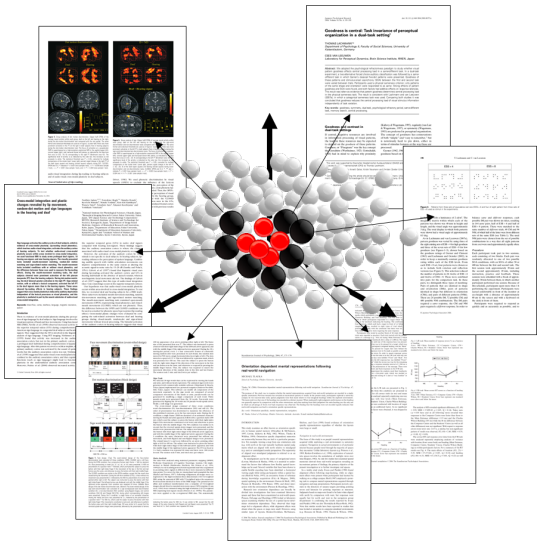


FIGURE – Find the semantics-based expanded corpus

Issues

- Where to cut (thresholding / set the boundaries)
- Validation and user feedback
- Utilizing user feedback for an active learning ?

Open Meta-data Digital Library with +16 Million Publication

FIGURE – Excellence Initiative of Scientific and Technical Information

Out of many document types, e.g., slides, posters and conference articles, we only considered **English research papers** that were published **after 1990** with sufficient abstract size (35 to 500 words). The extracted meta-data dataset contains more than **+4 Million** articles.

Sport Sciences Topic : Mental Rotation

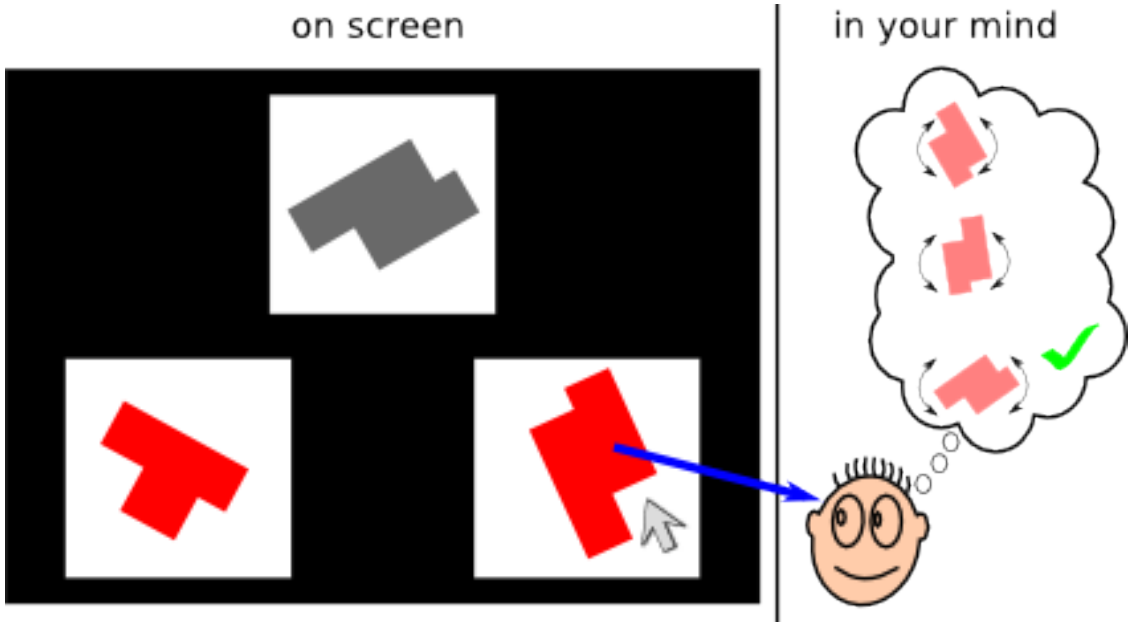


FIGURE – Mental Rotation. (Source : psytoolkit.org)

Multi-disciplinary topic : health sciences, psychology, cognitive sciences...

SSbE for Mental Rotation Use Case

Seed Articles ***182 Articles***

- **Title**
- **Abstract**
- Year of publication
- DOI
- Keywords
- Authors

Expanded *Positive* Articles

Using "mental rotation" as search keywords in ISTEX search engine to find more positive examples with matches in :

- **Title**
- **Abstract**
- Keywords

199 articles were found

Experimentally Decided Design Parameters

Bag of Word

word ngram range* : (1, 2)

Min term frequency : 20

Max term frequency percentage : 95%

Filtering out stop-words or not : yes

SVD Decomposition

Using TF–IDF transformer or not : yes

Dense vector size* : 150

Regression

- Random Forest
- 500 estimators

* : constrained by the memory size

Three Models Comparative Evaluation

Evaluation of the top ranked recommendations for each of the three systems :

- More-Like-This method : *MLT*
- Partial SSbE model (without active learning) : $SSbE_p$
- SSbE model with active learning : *SSbE*

Sample of Interesting Recommended Articles

Some interesting multi-disciplinary research topics found in some recommended relevant articles by the system. The articles do not contain the terminology "mental rotation" :

- The studies on abilities to read a **map in different orientations** or **Movement Trajectory**
- **Sign language** and lip-reading used by deaf signers are actions that require some mental rotation abilities for reading the manual communication
- **Gaze aversion** benefiting cognitive performance, e.g., visualspatial imagination, unlike face-to-face communication
- **Specific reading disorder** due to a specific difficulty in orienting and focusing.

Other terminologies : Spatial imagery, visualspatial imagination, spatial abilities, mental imagery ...

Expert Annotation Results

TABLE – Confusion matrix of the two domain expert judgment of both of $SSbE_p$ (SSbE without active learning) and MLT method on 100 results randomly picked from the top 200. *S* corresponds to $SSbE_p$ and *M* corresponds to MLT. CND indicates that the expert Can Not Decide

	relevant		CND		irrelevant		**Total**	
Method	S	M	S	M	S	M	S	M
relevant	**8**	2	3	3	0	0	**11**	5
cannot decide	10	1	10	4	17	4	37	9
irrelevant	2	0	13	5	37	**81**	52	**86**
Total	**20**	3	26	12	54	**85**	100	100

TABLE – Cohen's kappa scores for annotation of the two domain experts. The table shows results for different combination of annotation labels. The scores are rounded to 4 decimals

Labels	**Cohen's kappa score**
[relevant, irrelevant]	0.90
[relevant, cannot decide]	0.18
[cannot decide, irrelevant]	0.28
[relevant, irrelevant, cannot decide]	0.38

MLT versus SSbE

TABLE – Frequencies of the evaluation scores values for both the $SSbE_p$ method and the MLT method. The blue score labels are good while the red score labels are bad

Score	1	0.75	0.5	0.25	0
$SSbE_p$	**8**	13	12	30	37
MLT	2	4	4	9	**81**

Sentence Semantic Similarity

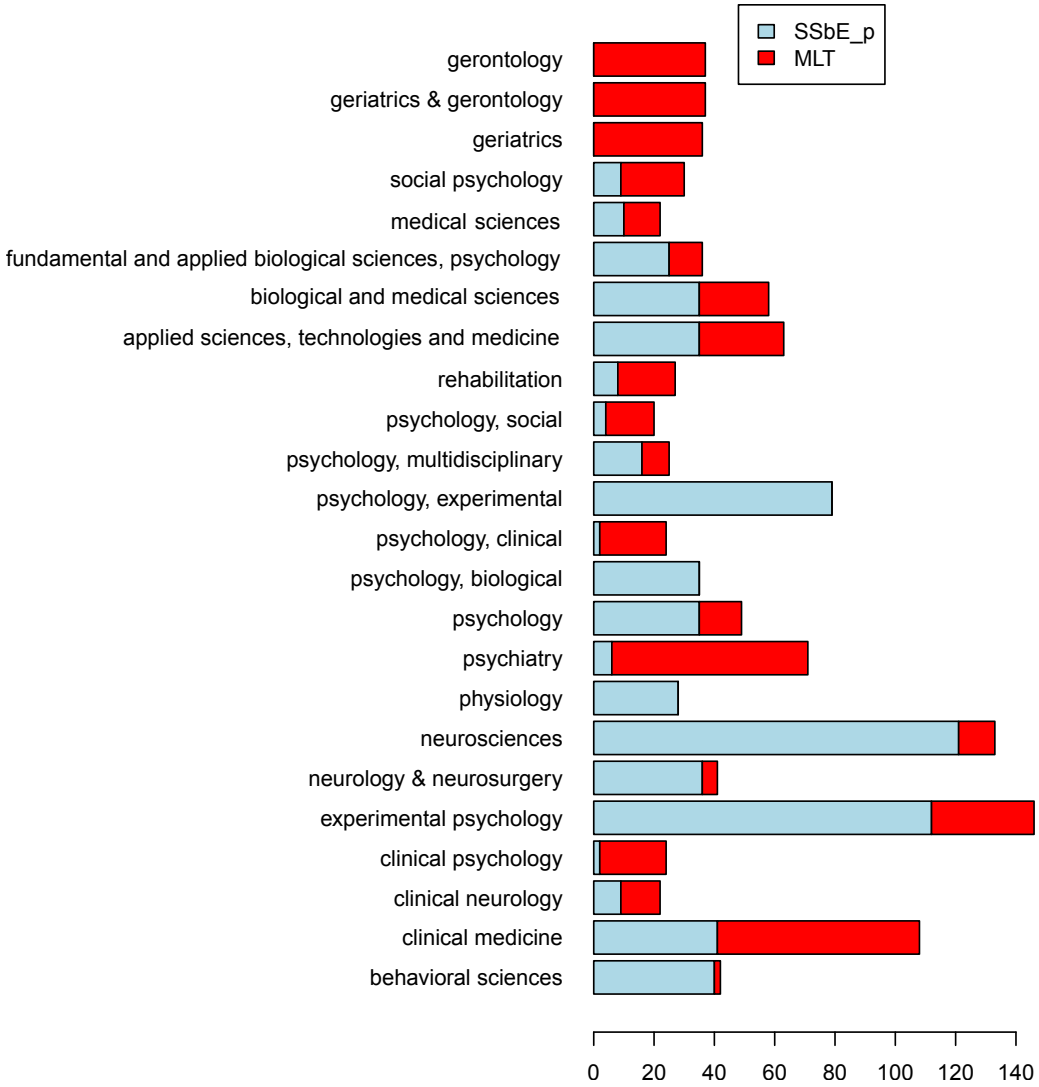
TABLE – Comparative results of the 3 methods using sentence semantic relatedness measure based on count of pairs with score higher than 3.0 out of 5.0

Method	*MLT*	$SSbE_p$	*SSbE*
Count of pairs	124	217	382

TABLE – Comparative results of the 3 methods on the top 959 results of each method using a test set extracted from the digital library meta-data that was hidden from our experiment. The number of 959 results were selected as a result of excluding extended positive articles, which have been used in the training phase, from the top 1000 results of $SSbE_p$. * :The total number of the MLT results is 391 articles

Method	*MLT**	$SSbE_p$	*SSbE*
matches count	1	1	6
rank of them	1	851	24, 82, 227, 567, 699, 929

Diversity Analysis



Open-Access Data and Code

Available on GitHub

SSbE Model and Experiments : https ://github.com/ERICUdL/ISTEX_MentalRotation

Sentence Similarity Estimator : https ://github.com/natsheh/sensim

Conclusion and Future Work

Conclusion

- SSbE was able to recommend relevant articles from different disciplines
- Active learning enhances the recommendation
- Sentence semantic similarity was a good evaluation measure for the different method

Future Work

- Study the use of semantic networks for expanding the positive examples
- Semantic topic auto tagging based on standard topic category
- Compare the topic tagging with topic modeling

Acknowledgment

- This work was kindly funded by **ISTEX project**.
- The use-case study was funded by the **Centre de Recherche et d'Innovation sur le Sport** who also technically contributed though the expert annotation process by :
  - Dr. Patrick Fargier
  - Prof. Raphaël Massarelli.
- We would like also to thank **ARC6 of the French Region Auvergne-Rhône-Alpes** that funds the current PhD studies of the first author