

UNIVERSITETI I EVROPËS JUGLINDORE УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА SOUTH EAST EUROPEAN UNIVERSITY

A Model for Recommending Research Articles: A Case Study in Computer Science, Neuroscience and Biology

Nuhi BESIMI, Betim ÇIÇO, Adrian BESIMI

Outline

- Introduction
- Research Problem
- The Proposed Model
- A Case Study in Computer Science, Neuroscience and Biology
- Future Work
- Conclusion

Introduction

- I. Why big data matters?
- 2. How to extract meaningful information from unstructured data, textual data?
- 3. What machine learning algorithms are used to analyze textual documents (textual data)?
- 4. Traditional vs. Parallel/Distributed machine learning algorithms for text analysis.
- 5. What kind of solutions have been proposed for recommending research articles to researchers.

Why big data matters?

- Data on the Web is increasing rapidly.
- Big Data
 - Store huge amount of data (big data)
 - Process large data sets (big data)



Research Problem

IMPROVING THE PROCESS OF RECOMMENDING RESEARCH ARTICLES TO RESEARCHES. EASE THE PROCESS OF LITERATURE REVIEW?



Related Work

Mendeley and CiteULike [1] [2]

- Reference Management
- Collaborative Filtering (User Filtering)
- Altimetric-Driven approach [3]
 - Enhance performance for research paper recommender systems.
- Topic-Modeling approach [4]
 - Exclude the keyword and focus on the topic

Our Study

The aim of our study is to collect/retrieve and analyze research/scientific articles by applying machine learning techniques in order to recommend research articles or/and research gaps to researchers based on their research fields.

Scientific Article

- Title
- Author/s
- Year
- Abstract
- Keywords
- Content

- Contribution
- Results
- Future Work
- Importance
- Related articles

Proposed Solution



Research Questions - Model

- What is the best document representation in text mining?
- Which are the most efficient clustering algorithms used recently?
- What classification techniques are used to build the most accurate models in text mining?
- What is the difference between Neural Networks and traditional classification techniques?
- Which is the best hierarchical clustering technique for textual documents?

Solution

Why Hybrid solution?

The reason why we consider this model as hybrid solution is because it is built on top of combination of supervised and unsupervised learning algorithms.

Model – Phase 1



Model – Phase 1.1



Bag of Words vs Word2Vec Document Representation

- Terms
- Bag of Words
- Term Frequency
- Form Frequency Inverse Document Frequency

Enhanced TF-IDF models

- Word Sequences
- Graph Structure

Word2Vec

- Word Embeddings
- ► NLP
- Extract Linguistic Context of Words
- Latent Semantic Analysis (LSA)

Model – Phase 1 Results

Phase I:

- Validate the Input Data Set
- Distance between Clusters
- Total number of Clusters
- Outliers
- Cluster Labels
- Cluster List of Labels (Keywords)
- The quality of the generated training dataset will be dependent on three key factors:
 - The input dataset
 - The text representation model
 - The applied clustering algorithm

Model – Phase 1 (Important)

I. How we are going to define the number of clusters?

The Silhouette Coefficient

2. How we will measure the accuracy of the clustering algorithm?

Since it was difficult for us to have concrete measurements for the first phase we had to use an already labeled textual dataset with sports news articles. [5]

[5] Nuhi Besimi, Betim Çiço, A Model for Recommending Research Articles,
7th Information & Communication Technologies at Doctoral Student Conference
2018 (DSC), Thesaloniki, Greece.

Model – Phase 2



Generated Clusters

Model – Phase 2 Results

Phase 2:

- Model (Decision Tree, Probabilistic Model, Centroids, Neural Network)
- Our aim is to select the most efficient model based on the literature review and the experiments. This model will help us to solve tasks like:
 - Classify new research articles based on their content.
 - Recommend research articles based on search criteria.
 - Query the input dataset for potential research gaps and trend research fields recently.

Model – Phase 2 (Important)

How we are going to measure the efficiency of the model?

Model – Phase 3

Finally, we can apply hierarchical clustering on the generated clusters to extract different levels of details for specific research fields. This step will be used to extract the trend topics and to increase the accuracy of the recommendation system.



Case Study Computer Science, Neuroscience, and Biomedical.

Open Research Corpus

- Over 39 million published research papers in Computer Science, Neuroscience and Biomedical.
- http://labs.semanticscholar.org/corpus/

 Waleed Ammar et al. 2018. Construction of the Literature Graph in Semantic Scholar.
NAACL. <u>https://www.semanticscholar.org/paper/09e3cf57</u> 04bcb16e6657f6ceed70e93373a54618

Dataset

- 36GB in JSON Format
- Computer Science, Neuroscience, Biomedical
- Attributes
 - Id, title, paperAbstract, entities, s2Url, s2PdfUrl, pdfUrls, authors, inCitations, outCitations, year, venue, journalName, journalVolume, journalPages, sources.



Processing

GCE

- Machine type: custom (2 vCPUs, 16 GB memory)
- Storage: I00GB (SSD persistent disk)
- OS: Ubuntu 16.04
- MongoDB v4.0
- Scikit-learn
 - Machine Learning in Python. Simple and efficient tools for data mining and data analysis.

NLTK

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.



mongoDB



- Total number of papers: 10 000
- Total number of clusters: 37
- Number of valid clusters: 26
- Outliers: II clusters



Cluster I

I257 papers

Top keywords:

- der
- health
- disease
- medical
- evaluation

clinical oder Medical analysis between beim Richtung transport practice research th management nicht detection human des following some eine protein properties training sich acid auf therapy zur Application DNA serum das gene nursing process blood one und nach International two flow other II change approach nurses über pulmonary Cancer support after approach more hospital Cancer support after approach early zu syndrom positive cycles reactions durch bei through syr Diseases rat strains mit als evaluation drug syndrome care drug right first measure children synthesis health 1e New measurements acute services dem de impact

problems

Cluster 2

> 759 papers

Top keywords:

- treatment
- brain
- therapy
- blood
- disease

•

development action high chronic Clinical Medical aspects high chronic infection Technology acid Changes first studies Institute binding NEW compounds work GTP diseases tumors study pulmonary other strains properties means fractures tubulin complex therapy author's effect time both NA disorders transl time both NA disorders distribution **U**^{ul} results atternt LI'ed nesis comparison problems diagnosis acute between analysis test protein nasal Current method recent analysis test protein nasal Current method **Synthesis** medicine nfants patients Surgical Biomaterials children tissue syndrome case use problem surgery during ieria lung fuzzy Experimental brain disease structure infants University Engineering liver cancer C]dopamine human Inżynieria ^{lung} ków some brain Biomateriałów

- Cluster 3
 - ► 364 papers
- Top keywords:
 - patients
 - health
 - risk
 - cancer
 - compared

associated care cases new clinical system presence review total normal used control analysis both less treatment diagnosis CT Health compared before between use three found blood medical chronic life present months higher studies infection without acute hea SCLUSION factors • surgery significantly healthy CONCLUSION factors cardiac liver **patients** two data rate survival after only number her five first during to first heart age heart groups other five first during high significant hospital groups type response four stage group primary performed treated included vs test one well more cancer all study levels follow up follow-up disease early serum

increased antibodies

26

Cluster 4

> 350 papers

- Top keyword:
 - cell
 - human
 - cancer
 - tumor
 - dna

• • • •

protein tumor showed blood molecular properties receptors between growth response different both mice factor induced observed significantly production number level presence mouse tissue all inhibition surface during effect rat after studi show levels responses addition studies proliferation high role carcinoma receptor lines signaling model type mechanism **11** two ¹ normal antigens culture only increase line analysis ^{using} other time present differentiation cellular within used function death increased found suggest II virus IL factors proteins DNA apoptosis cancer control more formation effects vitro stem pathway lung expressed study genes activity expressed expression CD membrane activation results nuclear gene vivo human treatment patients

Cluster 5

- ► 312 papers
- Top keywords:
 - system
 - information
 - query
 - strategy
 - user

• • • •

esults area many modi indexing problem design developed results project each etrieval paper appropriate speci SYStems important judgments data two designed goals et types all study knowledge Section be strategiesSearch well user's users health Section qualitative ^{type} either need techniques other particular needs more relevance education document used case al provide disease both States about propos different USE develop terms While bec ed mechanisms cation because one time queru new using some analysis another USEr research network queries based di order relevant number process management space mechanism environment abstract feedback model only service example current approach informat interaction between

Cluster 6

- I7I papers
- Top keywords:
 - algorithm
 - paper
 - image
 - based
 - proposed
 - detection

learning distribution set show proposed data sensor system propo approach frequency framework estimation technique features region some application presented small all more power research species performance model detection scheme order both spatial complex novel algorithms use image result first accuracy over simulation state methods newtime algorithm test network study within case propose each paper due techniques content analysis well contrast used base mobile high le one problem any existing Neural recognition face multiple localization information sampling measurement signal rate amino structure applications compared process while using method results through

between two

Future Work

- Experiment with different textual documents representation and evaluate the models which produce high performance in accuracy.
- Generate models based on various classification techniques to identify the most proper techniques for this task.
- Make our solution Open Source

COMMENTS & RECOMMENDATIONS

??!

References

- I. T. Bogers and A. van den Bosch, "Recommending scientific articles using citeulike," Proc. 2008 ACM Conf. Recomm. Syst. RecSys '08, no. January 2008, p. 287, 2008.
- 2. A. K. M. S.T.T. M. KAMAL NIGAM, "Text Classification from Labeled and Unlabeled Documents using EM".s
- 3. Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for textdocuments classification. Journal of advances in information technology, 1(1), 4-20.
- 4. H. J. Kim, J. Kim, and J. Kim, "Semantic text classification with tensor space model-based naive Bayes," 2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc., pp. 4206–4210, 2017.
- 5. Z. Wang, L. Ma, and Y. Zhang, "A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec," 2016 IEEE First Int. Conf. Data Sci. Cybersp., pp. 98–103, 2016.
- 6. M.Al-Amin, M.S. Islam, and S. Das Uzzal, "Sentiment Analysis of Bengali Comments With Word2Vec and Sentiment Information of Words," pp. 186–190, 2017.
- 7. N. Besimi, B. Cico, and A. Besimi, "Overview of data mining classification techniques: Traditional vs. parallel/distributed programming models," 2017 6th Mediterr. Conf. Embed. Comput., pp. 1–4, 2017.
- 8. X. Liu, X. Yan, Z. Yu, G. Qin, and Y. Mo, "Keyword extraction for web news documents based on LM-BP neural network," Proc. 2015 27th Chinese Control Decis. Conf. CCDC 2015, pp. 2525–2531, 2015

References

- 9. V. S. Reddy, P. Kinnicutt, and R. Lee, "Text Document Clustering: The Application of Cluster Analysis to Textual Document," 2016.
- I0. M. Habibi and A. Popescu-Belis, "Keyword Extraction and Clustering for Document Recommendation in Conversations," IEEE Trans. Audio, Speech Lang. Process., vol. 23, no. 4, pp. 746–759, 2015.
- II. Q. Bai and C. Jin, "Text Clustering Algorithm Based on Semantic Graph Structure," pp. 312–316, 2016.
- I2 M. B. Magara, S. Ojo, T. Zuva "Toward Altmetric-Driven Research-Paper Recommender System Framework", Signal-Image Technology & Internet-Based Systems (SITIS), 2017 13th International Conference on 4-7 Dec 2017, IEEE.
- I 3 V. Chaitanya, P. K. Singh "Research articles suggestion using topic modelling", Soft Computing & Machine Intelligence (ISCMI), 2017 IEEE 4th International Conference on 23-24 NOV 2017, IEEE.