USING CHAINED MACHINE LEARNING MODELS FOR SCIENTIFIC ARTICLES RECOMMENDATION

[1]RUI MAIA, [2]JOAO C. FERREIRA, [3]ANA LUCIA MARTINS

[1]Inov Inesc Inovação – Instituto de Novas Tecnologias and Instituto Superior Tecnico, Portugal
[2,3]Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
[2]Information Sciences, Technologies and Architecture Research Center (ISTAR-IUL), Portugal
[4]Business Research Centre (BRU-IUL)
E-mail: [1]rui.maia@inov.pt, [2]jcafa@iscte-iul.pt, [3]almartins@iscte-iul.pt

Abstract - Recommender systems are commonly used when it comes to online multimedia service providers or worldwide retail companies. Although, regarding educational resources, scientific papers and books, or other items with extensive textual content and description, recommendation systems are only in early development. In this paper, we propose a new approach entirely based on chained machine learning model store present and rank scientific papers. Thefirstmodel-awordembeddingsmodelsupported on a shallow neural network - is trained using a synthesized paper unit - a composition of the title, the abstract, the publishing conference or journal, and the year - that accurately captures paper's semantic information. Later we train pairwise learning to a rank model based on a support vector machine (SVM) using relevant and irrelevant papers. We show that our approach achieves state-of-art results and does not rely on any language dependent or domain knowledge. It only uses available on-line data and proves to be efficient in either user-dependent and user-independent modeling.

Index terms - Scientific papers recommendation, Machine Learning, Learning-to-Rank, Dimensionality Reduction, Technology Enhanced Learning.

I. INTRODUCTION

Many recommendation systems are used nowadays, concretely in contexts where users have numerous available choices, and it is difficult or virtually impossible to decide in an informed way about all the available options. Recommendation systems are commonly applied to multiple domains like on-line product selling or movies websites, but also on photo or social network providers. The users face a nearly unlimited set of options, which is a problem addressed by very different approaches. Although, allapproachestrytovirtuallyshrinkthesearchspacewhile maximizingtheprobability of a user choosing or accepting a relevant item or suggestion. Recommendations systems can generally be classified in one of three approaches: content-based, collaborative-based, and hybrid or heuristic approaches. Content based systemsrelyheavilyonthepropertiesofavailableitems,while collaborativeapproachestend to analyze user data and behavior in order to deliver recommendations. Hybrid and heuristic approaches cross these two fields by using items properties but also information about users and their choices. Despite the long lasting application of recommendation systems, it is assumed that therearedomainswherethesecanbefurtherdeveloped,asinrecommendationofnews, educationresources,booksorscientificarticles[1].

Theseareareasofabundanttextual and meta-data content that might be explored in order to improve recommendation's performance. Multiple papers are addressing domains where presented on Content Based Recommendation Systems (CBRecSys) 2016. Beel et al. [2] published a survey on scientific papers recommendation systems where they state that more than half - 55% - of scientific recommendation systems are content-based, while collaborative approach was used in only 18%, while graph-based recommendation approaches were used in 16% of the analyzed works. The authors reviewed more than 200 research-papers published between 1998 and 2013 and pointed twomainproblemsasforthisresearchdomain: inadequateorlackof evaluation nmethods, and lack of active (maintained) recommendation systems solutions. These problems cause a significant and negative impact on the research work [2]. They also underlined that 71% of the content-based filtering approaches did not specify the weighting scheme that was used. The application of textual representation models like the ones based on TF-IDFis not adequately documented. These models strongly depend on a language preprocessing pipeline, including stopwords removal, stemmingorlemmatization,andcommonlyarenotcorrectlyspecifiedorevenspecified at all. This stresses the fact pointed out by Beel et al. [2] that there is a recurrent and structural difficulty in reproducing and evaluating these research works. This work proposes a scientific papers recommendation system approach by leveraging chained machine learning models. The first model represents scientific papers textual and semantic content on a two-dimensional vector space using a single layer neural network. There is no need for any text preprocessing besides normalization (remove capital letters and ensure spaces between all words and punctuation marks). The information submitted to the neural network consists of the title,

abstract, conference nameandyearofeachconsideredpaper.
Thesevectorsarethenusedtotrainthesecond model, a Support Vector Machine on a pairwise rank approach. To address the previously pointed evaluation and comparison problem this work usesthedatasetoriginallybuiltbySugiyamaandKan[3]an dextendedbyAlzoghbiet al. [4]. Both authors kindly made available their datasets for this work. This paper continues by reviewing the related work. In Section 3 the problem definition and the proposed approach are presented, while Section 4 describes the experimental details, characterizes the dataset and the evaluation criteria. Section 5 presents and discusses the obtained results, and finally in Section 6, the conclusions are presented. Section 7 addresses the acknowledgments

II. RELATED CONCEPTS AND WORK

In the area of content based recommendation systems, Nart and Tasso [5] proposed an approach using graphs to represent semantic knowledge in papers. The authorsclaimed that a semantic approach might allow the creation of user models that adequately integrate users interests while maintaining high performance when compared to more traditional collaborative-based recommendation systems. Sugiyama and Kan extensively worked on scientific papers recommendation systems [6] [3]. These authors proposed a collaborative approach [7] to identify and leverage potential citation papers. The authors used language preprocessing techniques to calculate feature vectors to represent the dataset research papers and user profiles. Afterwards, by using cosine similarity to compare scientific paper and user profile feature vectors, their approach recommends those papers that are most similar to the considered user profile. In 2015, Lee et al. [8] proposed a personalized paper recommendation system that could module researchers preferences by using a collaborative-filtering. In this work, each paper is described as a bit vector that registers the presence of each domain word using a bag-of-words approach. The proposed work depends on natural language processing and domain knowledge, applying, for example, stopwords removal and stemming. Alzoghbietal. publishedmultipleworksonscholarlypapersrecommend ationsystems [9] [4] and published an updated work at Content Based Recommendation Systems Workshop (RecSys) 2016. The latest approach is based on language processing techniques and Rank Support Vector Machine model training. Alzoghbi et al. started by finding a vector representation for each dataset paper using domain knowledge and naturallanguageprocessingtechniques,inkeywordident ificationandTF-IDFprocessing.
AsforpapersranKing,theauthorsappliedaRankSVMmo delwhichpairedevery relevant and irrelevant paper for each researcher. They considered as relevant the papers that researchers had manually marked as so, and irrelevant the remaining papers of the same conferences of the relevant papers. The experiments were run over an extension of Sugiyama and Kan [3] Scholarly Dataset, where Alzoghbi et al. included publicly available metadata (for example, the abstract). Inacontent-basedorhybridrecommendationsystemapproach,theac curaterepresentationoftheavailableitems-researchpapers,inthiscontext-isacrucialfactorfor the final recommendation result. Papers are usually represented as discrete vectors as a result of processes that typically include natural language preprocessing techniques which often involves domain knowledge - and word weighting schemes (as TF-IDF). Someapproachesusecomplexlanguagepreprocessingto extractsemanticinformation from text sources. These steps require deep knowledge about the language used in theprocessedtextsources,typicallyEnglish. Grammaticalrulesanddomaindependent ontologies are also common resources in text processing and information extraction. Nonetheless, the computer science community have explored alternatives to these intensive language knowledge approaches when it comes to the continuous text representation. Mikolov et al. [10] [11] proposed an efficient approach for statistical language modeling using neural networks. The authors proposed two architectures for words representation on a continuous vector space model arguing that the similarity of word representations goes beyond the syntactic regularities as it can also describe semantic relations. Therefore, a word vector model delivers the possibility to infer knowledge using algebraic operations like the one that can be observed in the operation vector(King) + vector(Woman) – vector(Man) which would result in a vector similar to vector(Queen).

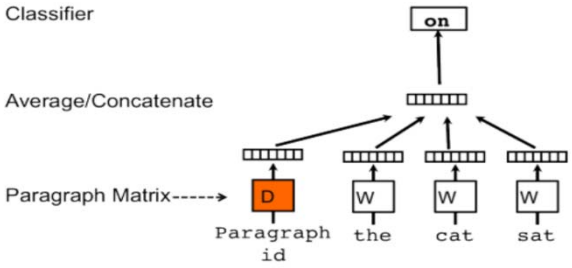

Figure 1: Using the Paragraph Vector Distributed Memory model in order to predict a fourth word given a paragraph vector and three word vectors.

In2014QuocLeandTomasMikolov[12]extendedtheWo rdVectorsworkforsentenceand paragraphcontinuously distributed vector representation. Theauthorsclaimed that their work support efficient vector modeling for pieces of text of any length as it does not rely on complex text preprocessing or parsing, neither on domain specific configuration and word weighting. Le and Mikolov proposed a Paragraph Vector Distributed Memory (PV-DM) model (see Figure 1)

as a two phase algorithm: 1) first, the calculation of word vectors and paragraph vectors. A fixed length window is sampled over each paragraph or piece of text. Each of the paragraphs and word vectors is trained using Stochastic Gradient Descend having the gradient calculated by backpropagation. 2) Second, the inference of new (unseen) paragraph vectors using the previouslygeneratedmodel. Thesuggestedapproachdoesnotrelyonanytextparsingor labelingandbenefitsfrommeaningfulsentencesandform allycorrecttexts. Wordvectors are nowadays used in multiple research areas, from language translation, to sense disambiguation and information extraction. Althoughsignificantanddiverseresearchworkshavebee nexploringrecommendationsystemsarea,fewhavebeen dedicatedtoscientificorscholarlypapersrecommendatio n. Therefore, as stated in CBRecSys 2016 [1], this area still represents an open challenge.

III. PROPOSAL

This work proposes a new personalized scientific paper recommendation system approach based on a pair of chained machine learning models. Our method spares the application of any language or domain knowledge and also excludes processes such as word removing, transformation and weighting. The proposed approach is divided into three main steps: the first to get a textual representation of each paper; a secondto get feature vectors representing each paper, and the third step to rank the list of suggestions. In order to rank suggestions using Rank Support Vector Machines, we need to consider two classes of papers, namely, the ones marked as relevant by researchers, and the ones that (regarding the same conferences) - were not considered as relevant. We follow the intuition that conferences are typically organized by themes. Therefore, we should take into account all papers from conferences or publications that had been interestingtoaresearcher(bymarkingsome papers asrelevant). In the oppositedirection, we might exclude papers from conferences were the researcher did not found any relevant item. The same assumption was taken by Alzoghbi et al. on [4]. For each researcher, 1) the first step of the method iterates over all relevant and irrelevantpapers. Ourmethodcontinuesbycreatingasynthesizedtextunitfo reachone: a textual concatenation of each paper's title and abstract, jointly with the conference (orpublication)andyear.
Attheendofthisstep,ourmethodgeneratedaconcisetextu al
representationofallthepapersusedtodescribetheresearc herpreferences. Inthescope of this work, other text units were tested as neural network input (for example, including the keywords list in the synthesized text unit). The proposed concatenation - title, abstract, conference/journal and year - showed to be the most effective and accurate text unit to represent paper's semantic information. The proposed approach follows by 2) training a shallow neural network (with only one hidden layer) to infer feature vectors that accurately describe each researcher relevantpaper,asinitiallyproposedbyLeandMikolov[12][11]. Thismethodcanarguably extract semantic information not only from independent sentences but also from complete sections of documents. Although, to the extent of our knowledge, it has not been testedandprovedpreviouslyagainstscientificcontentpub lications. Thisstepoutputis a model capable of draw feature vectors for unseen synthesized text units (or papers). The researcher's relevant and peer papers are represented in a matrix dxD, where d is the feature vector dimensions - the number of neural network outputs - and D the synthesized text units (see Figure 1), as columns. Each synthesized text unit - representing a research paper - is mapped into a unique vector with embed semantic information. The third and final step consists of getting learning to rank model. For that purposewetheapproachintegratesaSupportVectorMac hine(SVM)classifierusingPairwise Comparison [13]. In order to evaluate our approach, and to compare it with others, using the same datasetandmetrics. Forthat,wechoosetoapplyak-Foldcrossevaluation,with k = 5 and average the results for all folds, for each researcher, and finally get the average experimental result of all the researchers average.

IV. EXPERIMENT

By applying a k-Fold cross evaluation (k = 5), and averaging the results, we get the final experimental result. Concretely, for each researcher, the average result of all folds is the final researcher result, while the average of all the researchers is the final experiment result. Considering the lack of evaluation and comparison stated by Beel et al. [2], the approachwasanalyzedtwodatasetspreviouslyusedinref erenceworks: adatasetmadeavailable by Sugiyama and Kan [3] and other by Alzoghbi et al. [4]. The last consists of an extension to the first dataset by adding on-line available paper metadata. Since thisworkproposestorepresentpapersusingsemanticrich featurevectors,andthatthe synthesized text unit includes the name,and de abstract of the work, the most recent dataset [4] was the final choice for experimental evaluation.

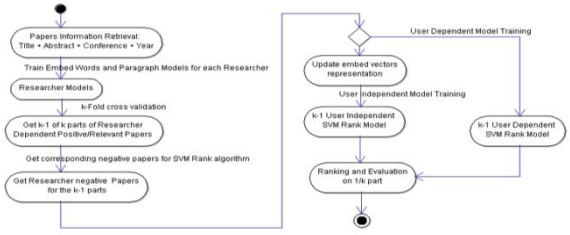

Figure 2: Approach model for user dependent and independent modeling

4.1 Dataset

In order to test the proposed approach, we analyzed two of the primary reference datasets. OnemadeavailablebySugiyamaandKan[3]consistingof candidatepapers torecommendontheperspectiveof50unidentifiedscienc eresearchers. Thecandidate papers list was built from the proceedings of the ACM Digital Library2 (ACM DL) between 2000 and 2010.. The dataset contains 100,351 papers, all written in English, which were processed according to the process described in detail in [3]. Although the datasetincludesinformationaboutcitationsandreference sbetweenpapers,wedidnot use this information in our work. The dataset gold standard identifies which papers are considered as relevant by each of the 50 researchers, who actively participated in the evaluation process. Stopwords were removed from each paper of this dataset and its feature vector calculated using TF-IDF on the stemmed paper words. Alzoghbi et al. extended [4] the Sugiyama and Kan dataset by including metadata got from candidate papers publishers websites. The extended data includes, among other properties, titles, abstracts, keywords and publishing conferences. This dataset contains 69,762 scholarly papers and keeps all the original dataset information besides the feature vectors. It includes the preferences of 48 researchers. Table 1 statistically describes this dataset. In order to get a paragraph vector representation for each paper, we needed to have a description for each one with a meaningful and objective piece of text. For that, we use the textual concatenation of title, the abstract, the publishing conference and the year. This data is taken from Alzoghbi et al. [4] dataset. For each researcher, we trained a neural network model using as input all the pointed relevant papers andall the others from the same conferences that were not considered as relevant. This follows the intuition that, not only these conferences generically have relevance for the researcher work - the domain, ontologies, methods - but also a particular distinction between relevant and irrelevant papers, for each researcher, must be taken into account the maybe light semantic differences in papers abstracts. Afterwards, for each researcher, we run a Pairwise Support Vector Machine Rank algorithm supported on a K-Fold cross-validation procedure, being K = 5. More precisely: 1) we pick the researcher relevant papers and split them into 5 folds; 2) For each fold of relevant papers, we get the other (irrelevant) papers from the same conferences; 3) Using k −1 (4) folds of relevant papers and the corresponding irrelevant papers, we train a Rank SVM model; and 4) the Rank SMV model got from the previous step will ordertheremainingpapers,concretely,the(k-1)relevantpapersandthecorresponding irrelevant papers from the same conferences.

Table 1: Scientific Papers reference Datasets

	Alzoghbi et al.	Sugiyama and Kan
Papers	69,762	100,351
Researchers	48	50
Average relevant papers per Researcher	71	
Average Relevant/Irrelent ratio per Researcher	1.2%	

4.2 Measures

For evaluation, we use the Mean Reciprocal Rank (MRR) and the Normalized DiscountedCumulativeGain(nDCG)forlistsof5and10r esults. TheMRRevaluatesthe position of the first relevant result appearing in the recommendation list. Regarding thateachresearcherhasitsownmodel,withaunique recommendation istbasedonall the papers from conferences that the researcher relates to, then, MRR is described by equation 1. |R| is the researcher count, and ranki is the position of the first relevant paper found in the researcher recommendation paper list.

$$MRR = \frac{1}{|R|}\sum_{i=1}^{|R|}\frac{1}{rank_i} \quad (1)$$

DiscountedCumulativeGainanditsnormalizedversion(nDGG)aremeasuresthat evaluate a recommendation list by analyzing the presence of relevant items in the first n listelements. Forthisexperimentalprocess,weanalyzethefirst5and10i temsofthe recommendation list. Therefore, we evaluate the top 5 elements of the recommendation list,andforthetop10list. Thismethodgradesthepresenceofrelevantitemspenalizi ng lowerlistpositionsforthese. Thenormalized(nDCG)2istheresultgotbythedivision of the DCG with the ideal DCG, which is the ideal list of results were all items would be relevant.

Table 2: IndePV approach compared with state-of-art approaches on a scholarly paper recommendation problem using the same dataset.

	IndePV	DePV	WBV	PubRec	Sugiyama and Kan
MRR	0.724	0.733	0.728	0.717	0.577
nDCG@5	0.435	0.456	0.471	0.445	0.345
nDCG@10	0.366	0.383	0.391	0.382	0.285

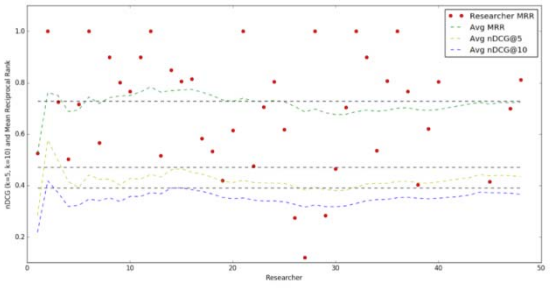

Figure 3: Mean Reciprocal Rank and Normalized Discounted Cumulative Gain experimental results. The plot shows MRR, nDCG@10 and nDGC@5 results for the 48 researchers. It also shows the global average evolution throughout the experimentation cycles. The horizontal lines represent the state-of-art results for comparison purposes.

$$nDCG = \frac{\sum_{i=1}^{|R|} \frac{rel_i}{log(i+1)}}{idealDCG} \quad (2)$$

RESULTS AND DISCUSSION

We tested our approach against Sugiyama and Kan [3] and Alzoghbi et al. [4] results, got from Alzoghbi 2016 publication. Figure 3 shows that with few iterations, our approach can be considered as valid as the other referenced results. Table 2 shows the experimental results where we can see both the user-dependent model and userindependent model. Figure 3 shows that our approach rapidly converges to the state of art results. Our method,whichextractstheembeddedsemanticinformation ofpapersandrepresentsit on paragraph vectors, is language independent since it does not rely on any language preprocessing besides punctuation normalization.

CONCLUSION

In this work, we propose a new scholarly papers recommendation approach that does not depend on language knowledge. To the best of the author knowledge, there is nopreviousworkontherepresentationofscientificpapers usingcompletetextsections for semantic meaning capture.
Thereisnoneedofdomainontologyresourcesornaturallanguageprocessingtechniques as stopwords removal, stemming or lemmatization. Our approach depends exclusively on machine learning techniques: a chained machine learning procedure. Research papers are represented in a continuous vector space, calculated using a shallow neuralnetworkfollowingtheparagraphvectorsapproach. WethenapplyaRankSVM for learning a rank model which can be either user dependent state-of-art results. We also propose a user independent,comprehensive machine learning approach, where we represent research papers as vectors got from the concatenation or average of multiple different papers representation.

Acknowledgements

The datasets analyzed and used were provided by Sugiyama and Kan, and Anas Alzoghbi et al.. We kindly thank them for making their datasets available. The possibilityofcomparisonbetweenworksdependsstrongly ontheshareofcommondatasets, which is also a share of knowledge.

REFERENCES

[1] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Lingvisticæ Investig., vol. 30, no. 1, pp. 3–26, Jan. 2007.
[2] CBRecSys, 3rd Workshop, on New, T. in, C. based Recommender, Systems, Proceedings of the, 2016.
[3] J.Beel,B.Gipp,S.Langer,C.Breitinger,Research-paperrecommendersystems: a literature survey, International Journal on Digital Libraries 17 (4) (2016) 305– 338.
[4] K. Sugiyama, M.-Y. Kan, Exploiting potential citation papers in scholarly paper recommendation, in: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL ’13, ACM, New York, NY, USA, 2013, pp. 153–162. doi:10.1145/2467696.2467701. URL http://doi.acm.org/10.1145/2467696.2467701
[5] A. Alzoghbi, V. A. A. Ayala, P. M. Fischer, G. Lausen, Learning-to-rank in researchpapercbfrecommendation: Leveragingirrelevantpapers,CBRecSys2016 (2016) 43.
[6] D. De Nart, C. Tasso, A personalized concept-driven recommender system for scientific libraries, Procedia Computer Science 38 (2014) 84–91.
[7] K. Sugiyama, M.-Y. Kan, Scholarly paper recommendation via user’s recent research interests, in: Proceedings of the 10th annual joint conference on Digital libraries, ACM, 2010, pp. 29–38.
[8] K. Sugiyama, M.-Y. Kan, A comprehensive evaluation of scholarly paper recommendation using potential citation papers, International Journal on Digital Libraries 16 (2) (2015) 91–109.
[9] J. Lee, K. Lee, J. G. Kim, S. Kim, Personalized academic paper recommendation system (2015).
[10] A. Alzoghbi, V. A. A. Ayala, P. M. Fischer, G. Lausen, Pubrec: Recommending publications based on publicly available meta-data., in: LWA, 2015, pp. 11–18.
[11] T. Mikolov, Statistical language models based on neural networks, Presentation at Google, Mountain View, 2nd April.
[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.
[13] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents., in: ICML, Vol. 14, 2014, pp. 1188–1196.
[14] R.Herbrich,T.Graepel,K.Obermayer,Largemarginrankboundariesfororordinal regression.

★★★