Volume 65, Issue 5, 2021

Journal of Scientific Research

of

The Banaras Hindu University

Finding Research Trends of Countries and Research Collaboration Opportunities in Computer Science Using Text Mining and Clustering Analysis

Taniya Seal*[1]

*[1]Department of Computer Science and Engineering, University of Calcutta, Email-taniyaseal1992@gmail.com

Abstract : **The occurrence of academic journals and conferences are getting intensively larger to publish the up-to-date research results on each particular research topics. The number of journals and conferences tends to be increased year by year. Moreover a journal or conference information is mostly accessible on the Internet and it contains not only topics but also geographical areas with which research topics are associated. This information is considered as a valuable source to understand the research trends of countries. In this paper, the state of detailed computer science research in different research fields is presented. The main aim is to develop methods for not only research topics but also the research collaboration associated with those topics related with each document and the relationships among the countries by using text mining and clustering analysis. This trend analysis for a particular research field plays vital role to those newcomers who are seeking for future research directions and possible collaborative research opportunities.**

Index Terms: **Country Clustering, Document similarity network, Modularity analysis, Research Opportunities, Topic extraction, Trend Analysis.**

I. Introduction

Advancement of a research field can be accomplished by scientific research. The analysis of research using feasible method such as trend analysis help us to understand the current state of research and in which direction it may be going. For a particular research field, this method describes its history, present status and predict future directions using statistical tools applied on bulk of papers published in peer-reviewed journals. This trend analysis for a particular research field plays vital role to those newcomers who are seeking for future research directions and collaborative research opportunities.

The availability of journals and conferences publications on the up-to-date research results on each particular research topics increased year by year and these are mostly accessible on the Internet. These contain not only topics but also the geographical areas with which research topics are associated. This information is considered as a valuable source to understand the research trends of countries that can be extracted from each document. In this paper, the state of detailed computer science research in different research field is presented. The objective of this paper is to figure out the research trend by extracting research field information that can clearly identify not only the research topic but also the research approach associated with those topics from a large amount of academic information. The proposed method attempts to identify topical concentration levels on research articles and approach concentration level for each research topic. For this purpose, the journal or conference information is represented by topic vectors for research topics through Latent Dirichlet Allocation (LDA) topic modeling. The expertise and diversity can be identified by the standard deviation of each document's topic vector. Therefore, if the standard deviation is low, it is determined to be highly concentrated because it is biased toward a specific topic. If it is high, it shows that diversity is high. The clustering of country based on research approach for a particular research topic find similar research approach of all countries in that cluster. This enables us to determine the relevance of research as well as research approaches between countries. It provides useful results on common interests and research collaboration opportunities.

The network is constructed through modularity analysis proposed by Blondel et al. (2008) and the main topics of each cluster are identified. Similarly main research approaches are identified from each sub cluster of each topic cluster. Leading country is recorded for each sub cluster where research approach is common to the countries present in the sub cluster. The

[*] Corresponding Author

DOI: 10.37398/JSR.2021.650539

changes in international major researches with their approach over time are measured and the countries that lead each research approach are identified. The method also finds out how each theme with their approach changes over time, so the trends of interest in a topic can be seen. So mainly the aim is to find the relationships among the countries by using text mining and clustering analysis.

The paper is structured as follows. In the section II, the reviewed of previous relevant contributions regarding topic modeling and document clustering are given. Then I have presented the main theoretical background of topic extraction model, graph construction and modularity analysis in section III to V respectively. Afterward, the proposed method in section VI. Next results are discussed in section VII. Finally, conclusions and future works are described in section VIII.

II. Related Work

In this section the related works on document representation and topic extraction, document similarity measurement and trend analysis based on topic model are presented.

A. *Document Representation and Topic Extraction*

Documents hold important statistical relationships useful for many applications such as classification, finding document similarity, relevance judgments, etc. on large collections of data. Salton and Buckley (1988) proposed one of the most popular corpus representation schemes TF-IDF. Here each document is represented by a vector of real numbers representing ratios for counts of words in the document. In probabilistic topic modeling, starting from SVD (Singular Value Decomposition) inspired LSI (Latent Semantic Indexing) to generative mixture models generate documents in terms of word distributions as mixture components using the hidden structure of data. It is followed by topic modeling algorithm LDA, which is more flexible in constructing documents by sampling topic for each word and using Dirichlet priors for random variables. Deerwester et al. (1990) introduced LSI. The drawbacks of TF-IDF are detected in LSI by performing a SVD of the terms by documents matrix of TF-IDF scheme. A generative mixture models with latent variables was proposed by Hoffman (1999), known as PLSI. PLSI improved LSI as it samples each word in a document from multinomial distributions over words which serve as mixing components. Figure 1 represents Graphical model of PLSI with the help of plate notation. The shaded circles in this notation denote observed variable each and the non-shaded denote hidden variable.

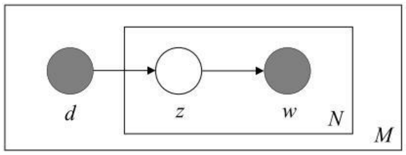


Figure 1: Graphical model of PLSI represented by plate notation.

The main problem of this model is that it learns the topic distributions from seen documents only and it becomes difficult to use it for unseen documents. Another problem to PLSI model is that the size of the model is linearly dependent on the size of the corpus and therefore, it helps in over-fitting. Blei et al. (2003) introduced this topic models. It uses Dirichlet prior on topic distribution instead of multinomial probability vectors for each document. It solves the problem of PLSI using of k-parameterized hidden random variables as described later. Figure 2 represents Graphical model of LDA with the help of plate notation.

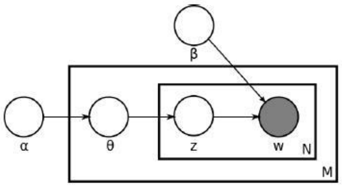


Figure 2: Graphical model of LDA in the plate notation.

As an application of LDA Wang et al. (2018) used LDA topic model to find out the topics in the aggregated tweets for each user of a community.

B. *Measuring document similarity*

Amado et al.(2018) have shown K-means and its many alternative forms like bisecting K-means and spherical K-means have been used widely on document datasets . For measuring the distance of two probability distributions, Gretarsson et al. (2012) and Wei et al. (2010) used Kullback-Leibler (KL) divergence for making such comparisons. Finding similarity of terms based on coupled correlation analysis is proposed by Kuhn et al.(2010). To compute similarity between two documents in the vector space the cosine similarity is also used in many cases. If two vectors d_1 and d_2 are given, cosine similarity is defined as:

$$\cos(\mathrm{d}_1, \mathrm{d}_2) = \frac{\mathrm{d}_1 . \mathrm{d}_2}{\|\mathrm{d}_1\| \times \|\mathrm{d}_2\|}$$

where "." represents the dot product operation between two vectors,"×" represents the cross product operation between two vectors and $\|\mathrm{d}_1\|$ and $\|\mathrm{d}_2\|$ are lengths of the two vectors d_1 and d_2 respectively. Similarity based algorithms are used for finding document similarity and assign documents to each cluster. The documents in same cluster are similar to each other and dissimilar to documents in other clusters. Based on the underlying methodology of the algorithm, final solution's structure, clustering algorithms are classified into different categories- Agglomerative algorithm and Partitional algorithm. According to Agglomerative algorithm each object is assigned to each cluster and merged pairs of clusters repeatedly until a certain stopping criterion is reached. Some methods have been introduced for finding the next pair of clusters that have to be merged, such as group average (UPGMA) proposed by Jain and Dubes (1988), singlelink proposed by Sneath and Sokal(1973), complete link proposed by King (1967), CURE proposed by Guha et al. (1998), ROCK proposed by Guha et al.(1999) and

CHAMELEON proposed by Jolliffe (2002). In Hierarchical algorithms, at the top of the clustering a single all-inclusive cluster and at the leaves single point clusters are shown. Some examples of partitional algorithms are k-Means proposed by MacQueen (1967), k-Medoids proposed by Kaufman and Rousseeuw(2005), graph partitioning based algorithm proposed by Zahn(1971) and Strehl and Ghosh (2000) , spectral partitioning based algorithm proposed by Boley (1998) and Ding et al.(2001). In partitional algorithms clusters are produced by dividing the entire dataset into either a predetermined or an automatically derived number of clusters.

C. *Trend analysis based on topic model*

Sohrabi et al.(2019) have shown evaluation of Research Trends in Knowledge Management. It analyzed the content of validated journal articles related to Knowledge Management (KM) in more than 18,000 papers of the Web of Science (WoS) database and then it provides the most recent specific trends in KM field using text mining and burst detection. Recently Liu et al.(2020) found out hot research topics and scientific trends in clinical psychology based on topic models.

III. TOPIC EXTRACTION & REPRESENTATION MODEL

Latent Dirichlet Allocation is a generative statistical model which represents documents as random mixtures of hidden topics, where each topic is a Dirichlet distribution of words in the vocabulary of the corpus. Documents in a corpus are comprised with collections of words that have been sampled from a Dirichlet distribution of topics specific to that document. The variables used in this paper and their descriptions are described in Table I.

Table I. Variables used in this paper and their description

Variable	Description
K	It represents number of topics that I want to extract from the documents.
V	It represents number of words in the vocabulary.
M	It represents number of documents used in corpus.
$N_{d=1......M}$	It represents number of words in each document d = 1,….,M.
N	It represents total number of words over all documents i.e. N= $\sum_{d=1}^{M} N_d$.
$\alpha_{k=1.....K}$	It represents prior weight of topic k within a document. Generally, keep the value less than 1 same for all topics.
α	It is K-dimensional vector taking all α_k values and make it a single vector.
$\beta_{w=1......V}$	It represents prior weight of word w within a topic. Generally, keep the value less than 1 same for all words.
β	It is V dimensional vector taking all β_w values, make it a single vector.
$\theta_{d=1...M,k=1....K}$	It represents probability of topic k present in document d.
$\theta_{d=1.....M}$	It is K-dimensional vector of probabilities, sum of which must be 1, i.e. topics distribution in document d.
$z_{d=1...M,w=1....N_d}$	It identifies the topic of word w = 1….N_d in the document d, the value between 1 and K.
z	It is N dimensional vector of integers between 1 and K, identifies topic of all words in all documents.
$w_{d=1...M,k=1....K}$	It identifies a word w in document d, the value between 1 and V .
w	It is N dimensional vector of integers between 1 and V , identifies all words in all documents.
D	It is M dimensional vector of w, identifies a corpus that is a collection of M documents.

A. *Dirichlet Distribution*

The probability density function in respect of Lebesgue measure over the Euclidean space R^{k-1} with Dirichlet distribution of order $k \geq 2$with parameters $\alpha_1, \dots\dots.., \alpha_k > 0$ is defined by equation 1:

$$f(\theta_1, \dots\dots, \theta_k;\ \alpha_1, \dots., \alpha_k = \frac{1}{B(\alpha)} \prod_{i=1}^{k} \theta^{\alpha_i - 1} \qquad (1)$$

where $\sum_{i=1}^{k} \theta_i = 1$ and $\theta_i \geq 0$ for all $i \in [1, k]$.

The multivariate beta function using the gamma function is shown in equation 2:

$$B(\alpha) = \frac{\sum_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma \sum_{i=1}^{k} \alpha_i} \qquad (2)$$

$\Gamma(x)$ represents the Gamma function. This Dirichlet is a probability distribution over distributions, very convenient on the simplex. It has statistical support for finite dimension and is conjugate to the multinomial distribution. For this reason, Dirichlet is chosen as the distribution. It also ensures that with the increasing size of the corpus, the model size does not grow linearly.

B. Multinomial Distribution

Suppose for n independent trials k possible mutually exclusive outcomes with their probabilities $p_1, \ldots\ldots, p_k$ are calculated. Let the random variables are the occurrences of outcome number i in the n trials, then the vector $X = (X_1, \ldots\ldots\ldots, X_k)$ follows a multinomial distribution. The probability mass function of this multinomial distribution using the gamma function is given below in equation 3:

$$f(X_1, \ldots\ldots\ldots, X_k; p_1, \ldots\ldots\ldots, p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i x_i + 1} \prod_i^k {p_i}^{x_i} \qquad (3)$$

C. Formal description of LDA

The following generative process associated with corpus D consists with M documents where the each document consists with N words, is described below.

For each document w in a corpus D:

1. Choose $\theta \sim \mathrm{Dir}(\alpha)$ i.e. randomly choose a distribution over topics. From this it can be understood that each document can represent multiple topics in different proportions.
2. For each word out of N:
    a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$. It means that a topic is selected randomly from the distribution over topics.
    b) Choose a word w_n from $p(w_n | z_n, \beta)$ that is a multinomial probability conditioned on the topic z_n.

As discussed, one of the main objectives of topic modeling is to find the hidden structure. In LDA, topic distribution in every document and topic assignment of each word to a topic in a document, are the hidden structures which needs to be estimated. Therefore, topic modeling in LDA can be interpreted as the reverse process of the generative process defined above and the problem goes down to inferring the parameters describing the hidden structure. This process can be defined by a joint probability distribution including the hidden variables. This is used to calculate the conditional probability of the latent variables where given the observed variables. This is also known as analysis of the posterior distribution. The joint distribution of topic mixture θ, a set of N topics **z**, a set of N words **w** given α and β is expressed by the equation 4.

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta) \qquad (4)$$

Here $p(\theta|\alpha)$ is simply the equation 1, $p(z_n|\theta)$ is θ_i.

The marginal distribution of a document after integrating over θ and summing over z are expressed by the equation 5.

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)\, p(w_n|z_n, \beta) \right) d\theta \quad (5)$$

Finally, the probability of a corpus is calculated by taking the product of the marginal probabilities of single documents which is expressed by the equation 6:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_n} p(z_n|\theta_d)\, p(w_{dn}|z_{dn}, \beta) \right) d\theta \qquad (6)$$

D. Graph Construction

After extracting topic from each document, a graph is constructed. The nodes of graph represent documents and edges represent the similarity between documents. Document similarity is defined here using the distance between documents' topic distributions. Edge lengths are presented using multidimensional scaling. Clusters are formed by the documents who have closer distance to each other.

1) Document similarity measurement

To measure the similarity between two document topic distributions, Hellinger distance is used. The Hellinger distance H for two discrete distributions $P = p_1, \ldots\ldots\ldots, p_k$ and $Q = q_1, \ldots\ldots\ldots, q_k$ is defined in equation 7 :

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2} \qquad (7)$$

2) Multidimensional scaling

Multidimensional scaling (MDS) proposed by Wasserman and Faust (1994) is a technique of non-linear dimensionality reduction that assigns each node in two-dimensional space such that more similar nodes are placed closer together. The implementation of MDS used here for minimizing the loss function of "stress", a residual sum of squares using the Euclidean distance between points.

E. Modularity Analysis

Modularity is used to calculate the strength of division of a network into groups or clusters. Here modularity analysis is used to identify the research topics that are common to the countries in each cluster and identify the leading countries for those topics. Lambiotte et al. (2009) have shown the range of the modularity is $[-1,1]$. In case of weighted graph, modularity is defined in equation 8:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i}{2m} k_j \right] \delta(c_i, c_j) \qquad (8)$$

where

A_{ij} denotes the edge weight between nodes i and j,

The sum of the weights of the edges associated to nodes i and j is k_i and k_j respectively,

The sum of all of the edge weights in the graph is 2m,

c_i and c_j are the communities of the nodes, δ is delta function defined as equation 9.

$$\delta = \begin{cases} 0 \ if \ i \neq j \\ 1 \ if \ i = j \end{cases} \tag{9}$$

Here I use the Louvain method to maximize this Q value. This method extracts communities from large networks. Two phases of this method are repeated iteratively. Firstly each node of the network is given to its own community. The change in modularity is calculated for each node i by removing node i from its own community and shifting it to the community of each neighbor j of i. Two steps makes the calculation for finding the value very easy, these are:

Step 1: Removing node i from its original community

Step 2: Inserting i to the community j.

The equation for Step 2 is given by Li and Schuurmans(2011) shown in equation 10 :

$$Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{10}$$

Here $\sum_{\mathrm{in}}$ is sum of all the weights of the links which are inside the community C. $\sum_{\mathrm{tot}}$ is sum of all the weights of the links which are connected with the nodes of community i. The sum of weights of the links connected to node i is $\mathrm{k_i}$. The sum of the weights of the links between i and other nodes of the community where i can move into is $\mathrm{k_{i,in}}$ and sum of the weights of all links in the network is m. After calculating the Q value for all communities where it is connected to, C is moved into the community that gives the greatest modularity increase. In case of no increase in modularity, C will be kept in its original community. All nodes go through this process repeatedly and sequentially until no increase of modularity takes place. After reaching to this local maximum of modularity, the first phase comes to an end.

In the second phase of the algorithm, all nodes in the same community group together and build up a new network whose nodes are the communities from the previous phase. Self-loops of the new community node represents links between nodes of the same community and weighted edges between communities represent the links from multiple nodes in the same community to a node in a other community. After the new network is generated, the second phase comes to an end and the first phase can be further applied to this new network.

The resolution limit of modularity proposed by Fortunato and Barthelemy(2007) defines a limit on the size of the smallest community that one can obtain by modularity optimization. Lower the resolution limit I get more communities and higher to get less communities.

IV. Methodology

In this section description of dataset, data pre-processing, network generation, country extraction, research topic and research approach selection from the dataset and workflow of proposed work are presented.

A. Datasets

I have used a dataset consisting of 875 published Journal and Conference papers during five years (2014-2018), downloaded from IEEE Xplore and ScienceDirect digital library (Retrieved from https://ieeexplore.ieee.org/Xplore/home.jsp and https://www.sciencedirect.com) . E-resources of IEEE Xplore and ScienceDirect are subscribed under the Department of Computer Science and Engineering, University of Calcutta from where the data has been taken. These papers are collected using keywords Artificial Intelligence (AI), Information System (IS) and Very Large-Scale Integration (VLSI), three commonly research areas of Computer Science.

B. Data pre-processing, network generation, country extraction, research topic and research approach selection and modularity analysis

All text data are pre-processed before applying LDA i.e punctuations are removed, words are converted to lowercase and stop words are removed. All documents across the three topics are processed together. Here all documents are collected from renowned websites using keyword that always does not mean that the main topic of the document is that keyword. Sometimes the document describes another topic with the help of said keywords. For this reason topic modelling is used to find out the proportion of the topics in the document. For each corpus, a vocabulary is constructed by mapping each unique word across the whole text to an index. Each document is presented as a list of tuples listing the index of each word used in the document and the frequency of word occurrence. An LDA model takes this vocabulary, the collection of documents presented as lists of tuples, and fixed number of topics i.e three as input.

Next Corpus vs. Topic probabilities vector that finds research topic probabilities over whole dataset, Document vs. Topic probabilities vector that finds topic probabilities over each document, Word vs. Topic probabilities vector that finds topic probabilities over each word, are generated. The similarity of every pair of documents is calculated using the Hellinger distance between topic probabilities vectors of their corresponding documents. These topic probabilities vectors of each document of size 3 are obtained from Document vs. Topic probabilities vector having 3 columns for 3 research topics and 875 rows for 875 documents. After applying a threshold value of 0.2 on Hellinger distance, distances are stored in a square matrix named "sim". That means the Hellinger distance will be recorded if the calculated distance is below 0.2. Reason for choosing lower threshold is lower distance increases the similarity between two documents in document similarity network.

One of the main objectives of the work is country clustering based on research themes as well as research approach depending on research similarity networks of all countries, which provides useful results on common interests and research collaboration opportunities. So country extraction from documents is an important one. Country associated with every document is collected from the affiliation of the authors as in most of the affiliation country is mentioned where author developed his research concept. To increase accuracy for finding those country a larger training dataset is needed in name entity recognition(NER) system. But my proposed method for finding country need no such training dataset. Rather NER finds all the country present in our document and to find particular country present in affiliation of authors I have to further apply regular expression on the document. Besides in this method I directly applied regular expressions because affiliations are mentioned in most of the journal and conference papers with a specified format. Previously I got document cluster from document similarity network and with every document, country is associated. So I can easily cluster the countries using the country name instead of document name in the similarity network where each cluster represents a research topic. Above mentioned process is repeated for further clustering the countries based on research approach for each topic. Countries are extracted from those documents and clustered according to their common aspect i.e research approach.

In LDA, each document of a corpus is modelled as a finite mixture over an underlying set of topics. To find research topic probabilities from each document, all words are extracted from the documents. Similarly words are extracted in between the section 'Introduction' and the section 'Conclusion' for finding the research approach used in the document. I use Morrison and George (1995) categorization of research approaches. The categories of four main research approaches are descriptive(de), developmental(dl), formulative(fe), and evaluative(ev). For this purpose portable document format (PDF)documents retrieved from each topic cluster are converted to word files for getting the font and size of section 'Introduction' and section 'Conclusion'. The two words 'Introduction' and 'Conclusion' may present in the body of section, or subsection or any other else of the document which are not in same font and size with the section. So to remove the confusion of selecting the start and end positions for extracting the words I have to do this.

Lastly, a weighted graph is constructed using MDS. Modularity analysis is used to find each cluster from country clustering network and also identify leading countries and main research approach from each cluster for every year. A graph is drawn to show how each theme changes over time, so the trends of interest in a research topic and research approach associated with those topic can be seen. It will help researchers to choose a research topic which is extensively researched over last few years and whose future scope is also good. Possible collaborative research opportunities can be found out from country clustering network.

C. *Work flow of proposed work*

In this subsection, work flow of our proposed work is discussed, that includes some major steps required for better understanding of the sequence of this work.

Input : Published Journal and Conference papers from year 2014 to 2018 in PDF format.

Step 1 : Repeat steps 2 to 14 for year 2014 to 2018.

Step 2 : Convert PDF documents to text files and some pre-processing (eg. Stop words and punctuations are removed, words are converted to lower case) are done.

Step 3 : Extract first page from all PDF documents.

Step 4 : Extract country name from author's affiliation using regular expression on different affiliation format used in different research papers and generate country id by concatenating country name and integer value in increasing order(e.g. Chaina1,India2) to make it unique.

Step 5 : Corpus vs. Topic, document vs. topic and word vs. topic probabilities vectors are generated using LDA.

Step 6 : Hellinger distance between two topic probabilities vectors is measured for a pair of documents and stored it in a square matrix named "sim".

Step 7 : Document similarity network (DSN) is constructed using adjacency matrix "sim" whose rows and columns name are file names. Country clustering network is constructed using an adjacency matrix whose row and column name is country id generated in step 4.

Step 8 : DSN and country clustering network are visualized using Gephi software.

Step 9 : Three clusters (for 3 research topics) are formed using modularity analysis. Node and edge tables are generated for each cluster.

Step 10 : The documents that are related with each node of each country cluster are retrieved and extract topic name(eg. Ai) from document name(eg.Ai1) and add it in column named category 2 of the node table. This helps to represent the topic and country of each node in graph.

Step 11 : Main topic and leading country are found out by taking the highest percentage of topic and country among all respectively.

Step 12 : Save the documents collected from each cluster.

Step 13 : Convert the PDF documents into word files and extract the words between the section 'Introduction' and section 'Conclusion' to identify the research approaches.

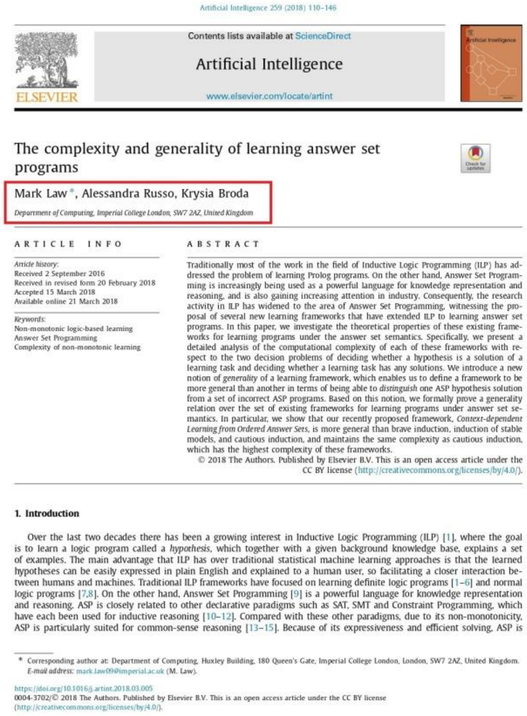
Step 14 : Repeat Step 2 to 11 for making four sub clusters(each sub-clusters represents each research approach) for each cluster. In this case, 3 clusters mentioned in Step 9 are replaced with 4 sub-clusters.

V. Result and Discussion

In this section accuracy for country extraction, year-wise main topics, research approaches and leading countries extraction from each cluster and sub-cluster using Hellinger Distance are presented pictorially.

A. Accuracy for Country Extraction

First page of a PDF is selected for country extraction from authors' affiliations which is located top left or top middle or bottom left part in 1st page of the PDF which are shown in Figure 3a, Figure 3b, Figure 3c respectively.



Artificial Intelligence 259 (2018) 110–146

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

ELSEVIER

The complexity and generality of learning answer set programs

Mark Law *, Alessandra Russo, Krysia Broda

Department of Computing, Imperial College London, SW7 2AZ, United Kingdom

ARTICLE INFO

Article history:
Received 2 September 2016
Received in revised form 20 February 2018
Accepted 15 March 2018
Available online 21 March 2018

Keywords:
Non-monotonic logic-based learning
Answer Set Programming
Complexity of non-monotonic learning

ABSTRACT

Traditionally most of the work in the field of Inductive Logic Programming (ILP) has addressed the problem of learning Prolog programs. On the other hand, Answer Set Programming is increasingly being used as a powerful language for knowledge representation and reasoning, and is also gaining increasing attention in industry. Consequently, the research activity in ILP has widened to the area of Answer Set Programming, witnessing the proposal of several new learning frameworks that have extended ILP to learning answer set programs. In this paper, we investigate the theoretical properties of these existing frameworks for learning programs under the answer set semantics. Specifically, we present a detailed analysis of the computational complexity of each of these frameworks with respect to the two decision problems of deciding whether a hypothesis is a solution of a learning task and deciding whether a learning task has any solutions. We introduce a new notion of *generality* of a learning framework, which enables us to define a framework to be more general than another in terms of being able to *distinguish* one ASP hypothesis solution from a set of incorrect ASP programs. Based on this notion, we formally prove a generality relation over the set of existing frameworks for learning programs under answer set semantics. In particular, we show that our recently proposed framework, *Context-dependent Learning from Ordered Answer Sets*, is more general than brave induction, induction of stable models, and cautious induction, and maintains the same complexity as cautious induction, which has the highest complexity of these frameworks.

© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

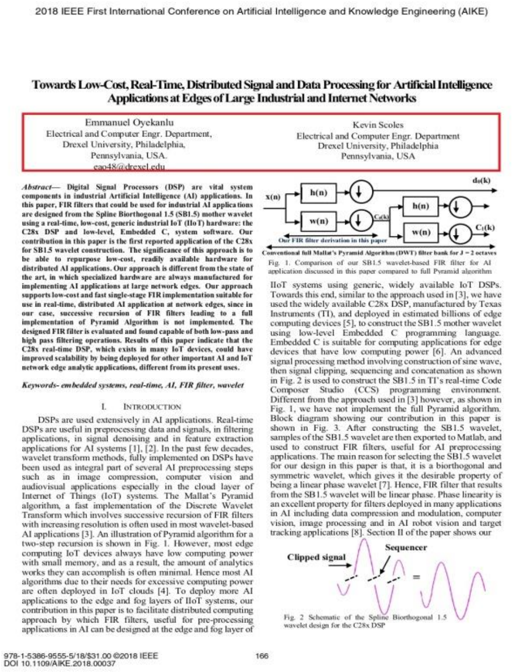
1. Introduction

Over the last two decades there has been a growing interest in Inductive Logic Programming (ILP) [1], where the goal is to learn a logic program called a *hypothesis*, which together with a given background knowledge base, explains a set of examples. The main advantage that ILP has over traditional statistical machine learning approaches is that the learned hypotheses can be easily expressed in plain English and explained to a human user, so facilitating a closer interaction between humans and machines. Traditional ILP frameworks have focused on learning definite logic programs [1–6] and normal logic programs [7,8]. On the other hand, Answer Set Programming [9] is a powerful language for knowledge representation and reasoning. ASP is closely related to other declarative paradigms such as SAT, SMT and Constraint Programming, which have each been used for inductive reasoning [10–12]. Compared with these other paradigms, due to its non-monotonicity, ASP is particularly suited for common-sense reasoning [13–15]. Because of its expressiveness and efficient solving, ASP is

* Corresponding author at: Department of Computing, Huxley Building, 180 Queen's Gate, Imperial College London, London, SW7 2AZ, United Kingdom.
E-mail address: mark.law09@imperial.ac.uk (M. Law).

https://doi.org/10.1016/j.artint.2018.03.005
0004-3702/© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(a)



2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)

Towards Low-Cost, Real-Time, Distributed Signal and Data Processing for Artificial Intelligence Applications at Edges of Large Industrial and Internet Networks

Emmanuel Oyekanlu
Electrical and Computer Engr. Department,
Drexel University, Philadelphia,
Pennsylvania, USA.
eao48@drexel.edu

Kevin Scoles
Electrical and Computer Engr. Department
Drexel University, Philadelphia
Pennsylvania, USA

***Abstract*— Digital Signal Processors (DSP) are vital system components in industrial Artificial Intelligence (AI) applications. In this paper, FIR filters that could be used for industrial AI applications are designed from the Spline Biorthogonal 1.5 (SB1.5) mother wavelet using a real-time, low-cost, generic industrial IoT (IIoT) hardware: the C28x DSP and low-level, Embedded C, system software. Our contribution in this paper is the first reported application of the C28x for SB1.5 wavelet construction. The significance of this approach is to be able to repurpose low-cost, readily available hardware for distributed AI applications. Our approach is different from the state of the art, in which specialized hardware are always manufactured for implementing AI applications at large network edges. Our approach supports low-cost and fast single-stage FIR implementation suitable for use in real-time, distributed AI application at network edges, since in our case, successive recursion of FIR filters leading to a full implementation of Pyramid Algorithm is not implemented. The designed FIR filter is evaluated and found capable of both low-pass and high pass filtering operations. Results of this paper indicate that the C28x real-time DSP, which exists in many IoT devices, could have improved scalability by being deployed for other important AI and IoT network edge analytic applications, different from its present uses.**

Keywords- embedded systems, real-time, AI, FIR filter, wavelet

I. INTRODUCTION

DSPs are used extensively in AI applications. Real-time DSPs are useful in preprocessing data and signals, in filtering applications, in signal denoising and in feature extraction applications for AI systems [1], [2]. In the past few decades, wavelet transform methods, fully implemented on DSPs have been used as integral part of several AI preprocessing steps such as in image compression, computer vision and audiovisual applications especially in the cloud layer of Internet of Things (IoT) systems. The Mallat's Pyramid algorithm, a fast implementation of the Discrete Wavelet Transform which involves successive recursion of FIR filters with increasing resolution is often used in most wavelet-based AI applications [3]. An illustration of Pyramid algorithm for a two-step recursion is shown in Fig. 1. However, most edge computing IoT devices always have low computing power with small memory, and as a result, the amount of analytics works they can accomplish is often minimal. Hence most AI algorithms due to their needs for excessive computing power are often deployed in IoT clouds [4]. To deploy more AI applications to the edge and fog layers of IIoT systems, our contribution in this paper is to facilitate distributed computing approach by which FIR filters, useful for pre-processing applications in AI can be designed at the edge and fog layer of

x(n)
h(n)
w(n)
d₀(k)
h(n)
w(n)
C₀(k)
C₁(k)
Our FIR filter derivation in this paper
Conventional full Mallat's Pyramid Algorithm (DWT) filter bank for J = 2 octaves

Fig. 1. Comparison of our SB1.5 wavelet-based FIR filter for AI application discussed in this paper compared to full Pyramid algorithm

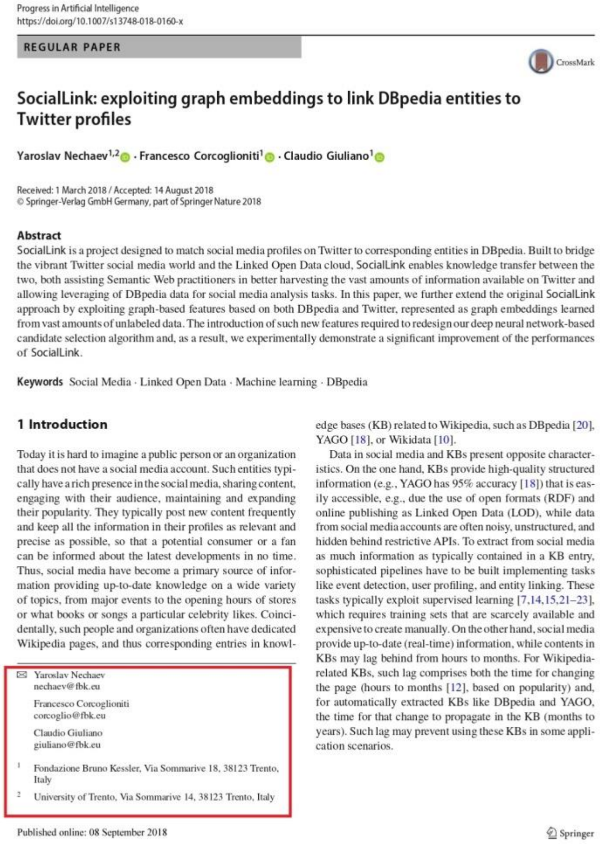
IIoT systems using generic, widely available IoT DSPs. Towards this end, similar to the approach used in [3], we have used the widely available C28x DSP, manufactured by Texas Instruments (TI), and deployed in estimated billions of edge computing devices [5], to construct the SB1.5 mother wavelet using low-level Embedded C programming language. Embedded C is suitable for computing applications for edge devices that have low computing power [6]. An advanced signal processing method involving construction of sine wave, then signal clipping, sequencing and concatenation as shown in Fig. 2 is used to construct the SB1.5 in TI's real-time Code Composer Studio (CCS) programming environment. Different from the approach used in [3] however, as shown in Fig. 1, we have not implement the full Pyramid algorithm. Block diagram showing our contribution in this paper is shown in Fig. 3. After constructing the SB1.5 wavelet, samples of the SB1.5 wavelet are then exported to Matlab, and used to construct FIR filters, useful for AI preprocessing applications. The main reason for selecting the SB1.5 wavelet for our design in this paper is that, it is a biorthogonal and symmetric wavelet, which gives it the desirable property of being a linear phase wavelet [7]. Hence, FIR filter that results from the SB1.5 wavelet will be linear phase. Phase linearity is an excellent property for filters deployed in many applications in AI including data compression and modulation, computer vision, image processing and in AI robot vision and target tracking applications [8]. Section II of the paper shows our

Fig. 2 Schematic of the Spline Biorthogonal 1.5 wavelet design for the C28x DSP

978-1-5386-9555-5/18/$31.00 ©2018 IEEE
DOI 10.1109/AIKE.2018.00037

166

(b)



Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-018-0160-x

REGULAR PAPER

CrossMark

SocialLink: exploiting graph embeddings to link DBpedia entities to Twitter profiles

Yaroslav Nechaev[1,2] · Francesco Corcoglioniti[1] · Claudio Giuliano[1]

Received: 1 March 2018 / Accepted: 14 August 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

SocialLink is a project designed to match social media profiles on Twitter to corresponding entities in DBpedia. Built to bridge the vibrant Twitter social media world and the Linked Open Data cloud, SocialLink enables knowledge transfer between the two, both assisting Semantic Web practitioners in better harvesting the vast amounts of information available on Twitter and allowing leveraging of DBpedia data for social media analysis tasks. In this paper, we further extend the original SocialLink approach by exploiting graph-based features based on both DBpedia and Twitter, represented as graph embeddings learned from vast amounts of unlabeled data. The introduction of such new features required to redesign our deep neural network-based candidate selection algorithm and, as a result, we experimentally demonstrate a significant improvement of the performances of SocialLink.

Keywords Social Media · Linked Open Data · Machine learning · DBpedia

1 Introduction

Today it is hard to imagine a public person or an organization that does not have a social media account. Such entities typically have a rich presence in the social media, sharing content, engaging with their audience, maintaining and expanding their popularity. They typically post new content frequently and keep all the information in their profiles as relevant and precise as possible, so that a potential consumer or a fan can be informed about the latest developments in no time. Thus, social media have become a primary source of information providing up-to-date knowledge on a wide variety of topics, from major events to the opening hours of stores or what books or songs a particular celebrity likes. Coincidentally, such people and organizations often have dedicated Wikipedia pages, and thus corresponding entries in knowledge bases (KB) related to Wikipedia, such as DBpedia [20], YAGO [18], or Wikidata [10].

Data in social media and KBs present opposite characteristics. On the one hand, KBs provide high-quality structured information (e.g., YAGO has 95% accuracy [18]) that is easily accessible, e.g., due the use of open formats (RDF) and online publishing as Linked Open Data (LOD), while data from social media accounts are often noisy, unstructured, and hidden behind restrictive APIs. To extract from social media as much information as typically contained in a KB entry, sophisticated pipelines have to be built implementing tasks like event detection, user profiling, and entity linking. These tasks typically exploit supervised learning [7,14,15,21–23], which requires training sets that are scarcely available and expensive to create manually. On the other hand, social media provide up-to-date (real-time) information, while contents in KBs may lag behind from hours to months. For Wikipedia-related KBs, such lag comprises both the time for changing the page (hours to months [12], based on popularity) and, for automatically extracted KBs like DBpedia and YAGO, the time for that change to propagate in the KB (months to years). Such lag may prevent using these KBs in some application scenarios.

✉ Yaroslav Nechaev
nechaev@fbk.eu

Francesco Corcoglioniti
corcoglio@fbk.eu

Claudio Giuliano
giuliano@fbk.eu

[1] Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy

[2] University of Trento, Via Sommarive 14, 38123 Trento, Italy

Published online: 08 September 2018

Springer

(c)

Figure 3: Authors' affiliation in (a) top left, (b) top middle and (c) bottom left part of 1st page of PDF.

Measured accuracy of country extraction from authors' affiliation using our proposed method on whole dataset is 0.9443. It has been seen that affiliations of authors' for 826 papers out of 875 papers are correctly identified by our proposed method.

B. Year-wise Main Topics, Research Approaches and Leading Countries Extraction from each Cluster and Sub-cluster using Hellinger Distance

Here all the above are described for year 2014. At first country clustering network with 3 clusters which represent three research topics (AI, IS, VLSI) are formed using modularity analysis taking all the documents published in year 2014. All nodes of each cluster are categorized with their corresponding research topic name. Main topic is identified by calculating highest participation among all topics.

Next all nodes of each cluster are categorized with their corresponding country name. Leading country is identified by calculating highest participation among all countries. Further a country clustering network with 4 sub-clusters which represent 4 research approaches (de, dl, fe, ev) is formed using modularity analysis taking all the documents found in each of 3 topic clusters. Main topics and leading countries are identified for each sub cluster.

1) Country Clustering on all Papers Published in Year 2014

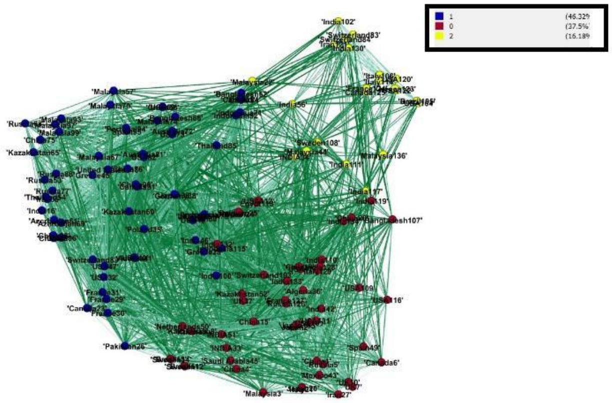


Figure 4: Country clustering on all papers published in the year 2014.

Here 3 clusters are shown with their topic proportions. They are cluster 0, cluster 1 and cluster 2 having 37.5, 46.32 and 16.18 percentage of total participation respectively. The nodes from cluster 0,cluster 1 and cluster 2 are represented with colour red ,blue and yellow respectively. Thick line between nodes represents the edge weight is small that means two documents are more similar. Here I used resolution limit 1 in modularity analysis.

2) Main Research Topic and leading country in Cluster 0

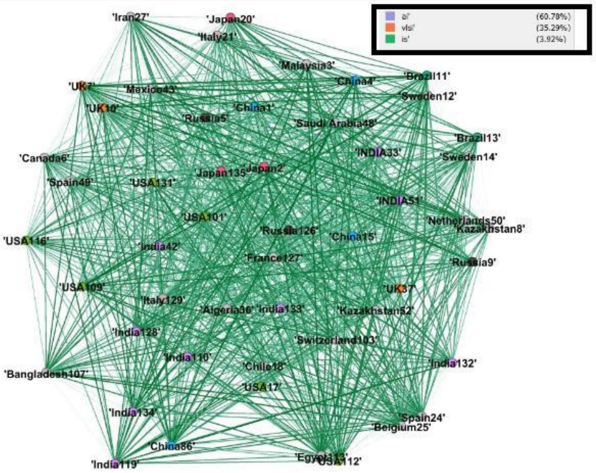


Figure 5: Main topic in cluster 0.

Now taking the documents of cluster 0, I have found out main topic and leading country in cluster 0. As we can see from the above figure that AI has highest percentage 60.78 in this cluster 0 in respect of research topic. So we can conclude that AI is the main topic for this cluster.

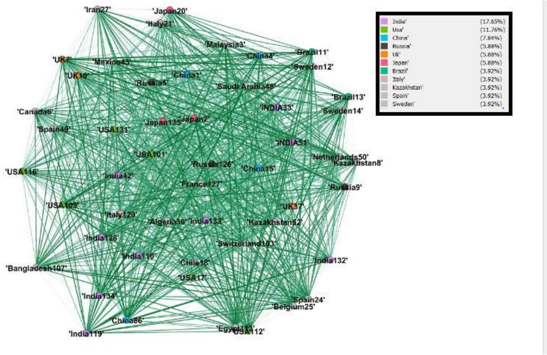


Figure 6: Leading country in cluster 0.

As we can see from the above figure that India has highest percentage in this cluster 0 in respect of country. So we can conclude that India is the leading country for this cluster.

In this way proposed method finds out IS and VLSI as main topic and USA and India as leading countries in cluster 1 and 2.

3) Main approaches and leading Countries in sub-Cluster

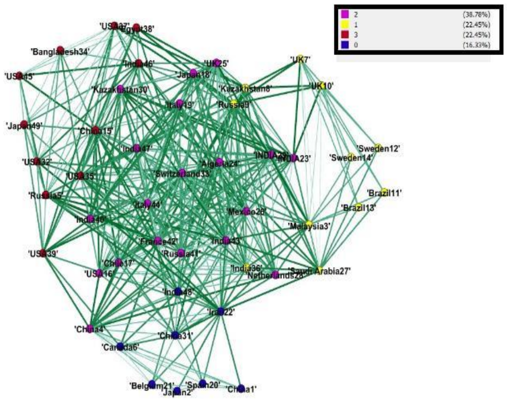


Figure 7: Country clustering on all papers present in cluster 0.

The above figure represents country clustering on all papers in cluster 0 .Here 4 sub-clusters are shown with their approach proportions. They are sub-cluster 0, sub-cluster 1,sub-cluster 2 and sub-cluster 3 having 38.78, 22.45, 22.45 and 16.33 percentage of total participation respectively.

I choose sub-cluster 2 to show the main research approach and leading country of it.

4) Main Research approach and Leading Country in Sub-Cluster2

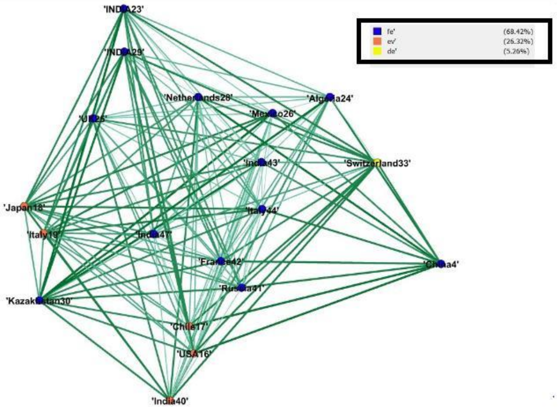


Figure 8: Main research approach in cluster AI.

As we can see from the above figure that fe(Formulative) has highest percentage 68.42 in this sub-cluster 2 in respect of research approach. So we can conclude that formulative is the main approach for this cluster.

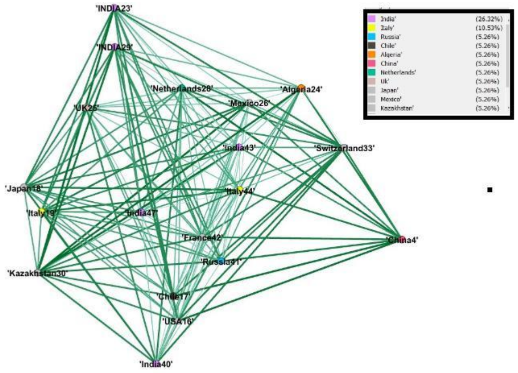


Figure 9: Leading country in cluster AI.

As we can see from the above figure that India has highest percentage in this sub-cluster 2 in respect of country. So we can conclude that India is the leading country for this cluster. Similarly I found out the best research approach and leading countries for sub-cluster 0,sub-cluster 1 and sub-cluster 3.

Whole process of 2014 is repeated on the documents of other years from 2015 to 2018.

5) Research trend analysis

In the last subsection, we have seen main research topic and leading country for cluster 0, cluster 1 and cluster 2 and these clusters are further analyzed to find the best research approach used for particular research topic. Whole process of 2014 is repeated on the documents for other years from 2015 to 2018.

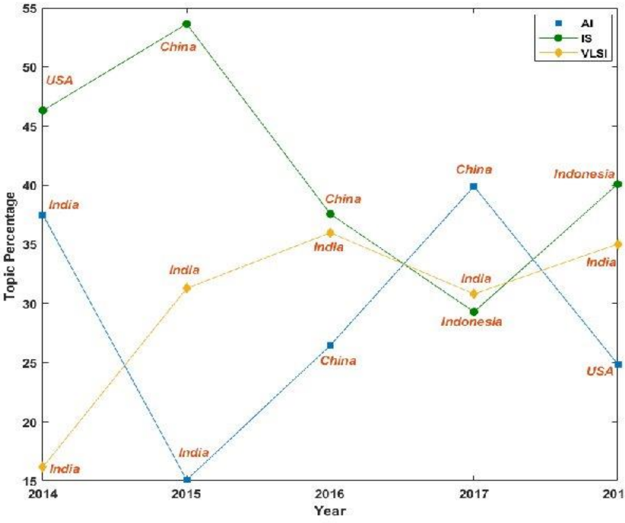


Figure 10: Time Series Graph for each topic with year 2014 to 2018.

Here it is tried to make understand how the research in three areas change with time (2014 to 2018) and highlight the most participating country.

Finally, trend of research approaches used in AI, IS and VLSI research areas from year 2014 to 2018 with the highest participating country are shown below using three graphs respectively.

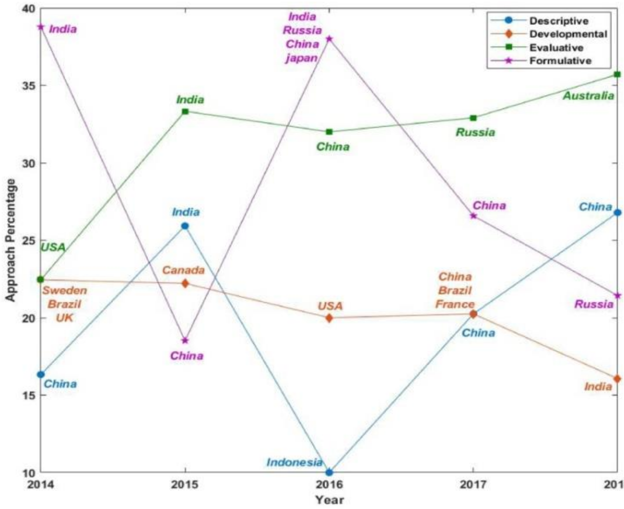


Figure 11: Time Series Graph for each research approach with year 2014 to 2018 in research area AI.

Here we can see that year wise research approaches vary with different countries for research topic AI. In overall evaluative approach is preferable as this approach is detected as a main approach for 3 years(2015,2017 and 2018) out of total 5 years. It gives researchers an idea of evaluative research approach that are mostly used in the AI research field for those 5 years. Research collaboration can be established with different countries shown in the graph.

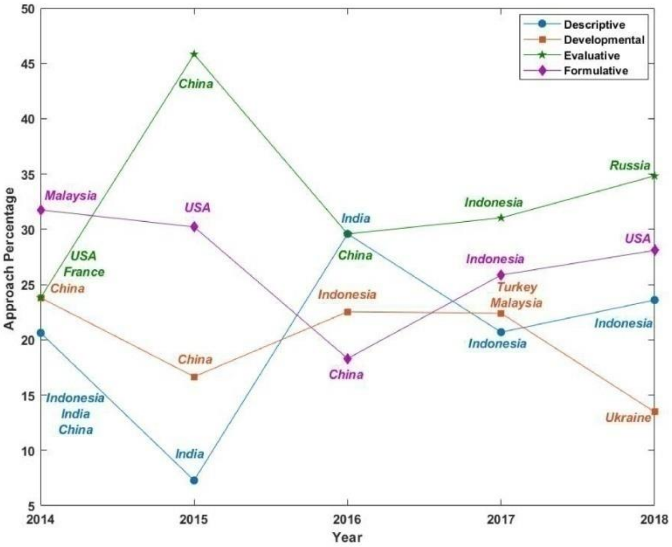


Figure 12: Time Series Graph for each research approach with year 2014 to 2018 in research area IS.

In overall evaluative approach is preferable for IS research field as this approach is detected as a main approach for 4 years(2015,2016,2017 and 2018) out of total 5 years. It gives researchers an idea of evaluative research approach that is mostly used in the IS research field for those 5 years .Research collaboration can be established with different countries shown in this figure.

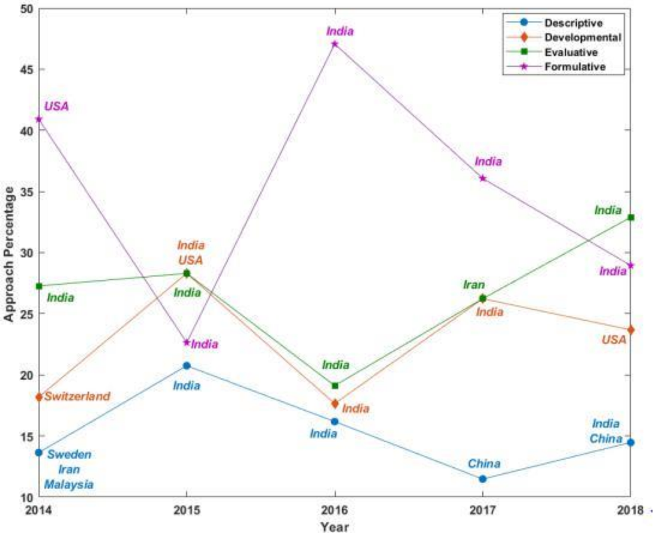


Figure 13: Time Series Graph for each research approach with year 2014 to 2018 in research area VLSI.

In overall formulative approach is preferable in VLSI research field as this approach is detected as a main approach for 3 years (2014,2016,2017) out of total 5 years . It gives researchers an idea of formulative research approach that are mostly used in the VLSI research field for those 5 years. Research collaboration can be established with different countries shown in the graph.

CONCLUSIONS AND FUTURE WORK

In this paper, Latent Dirichlet Allocation model is used for extracting information from documents which are mixture of research topics published from year 2014 to 2018.In this research, published Journal and Conference papers are collected using keywords Artificial Intelligence (AI), Information System (IS) and Very Large-Scale Integration (VLSI) from renowned digital libraries. A particular document where a above mentioned keyword is used does not mean that the document is totally based on this keyword or topic. This document may contain many other topics. So I have used LDA here for finding AI, IS and VLSI topic proportions from documents . A tool for visualizing and analyzing text corpora is utilized. It helped to understand the formation of clustering. Then a method is built up to find the geographical areas i.e. country where the research concepts are developed. Next a graph is constructed with nodes representing each country and edges with lengths drawn proportionally to the Hellinger distance between topic probability distribution of pair of documents. Using multidimensional scaling the edges are adjusted with edge weights i.e. down-weight large distances and up-weight small distances. Ultimately, it is shown that Country Clustering is most useful to determine the relevance of research between countries on their common interests and research collaboration opportunities. From figure 6,we can see that India, USA, China, Russia, UK, Japan and some more countries had interest on AI research in 2014 and their interest level can be calculated by percentage of their participation. Finally, the proposed method also finds out how each theme changes over time, so the trends of interest in a topic can be seen. From figure 10,we can see that the research done on AI was most in 2017 and for other years research done on IS was most. Similarly research approaches are analyzed on a topic for detail analysis. So in this paper it is tried to show how topic modelling and country clustering can be applied in computer science research field. Specially newcomers who are seeking for future research on AI, IS and VLSI, get an idea about research trend, highest participation of country and possible collaborative research.

In future, the following problems and ideas can help to explore the method for further research in this area.

1. Some words are not properly extracted from PDF documents. Proper extraction of words will give more accurate result as words are basic unit of topic modelling.
2. Analysis of reference discipline that gives the theoretical foundation of a research will be considered for betterment of this research.
3. Evaluation of Topic Models: Evaluation of topic models has been a challenging problem for a long time now. Introduction of coherence score has been useful, but for alternate topic modelling techniques such as sentence2cluster where semantics of a topic are different from conventional LDA based topics. No objective evaluation criteria exists to measure utility of topics. More efforts should be given into these objective function which can measure the necessity of diverse topic modelling techniques.

REFERENCES

Amado, A., Cortez, P., Rita, P. and Moro, S.(2018).Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Bussiness Economics,1(24), 1-7.

Blei, D., Ng, A. and Jordan, M.(2003).Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blondel, V., Guillaume, J.L., Lambiotte, R. and Lefebvre, E.(2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10.

Boley, D. (1998).Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4),325-344.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R.(1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Ding, C., He, X., Zha, H., Gu, M.and Simon, H.(2001).Spectral min-max cut for graph partitioning and data clustering.*2001*

1st IEEE International Conference on Data Mining (pp.107-114).

Fortunato, S. and Barthelemy, M.(2007).Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America, ser. 1, 104, 36-41.

Gretarsson, B., Odonovan, J., Bostandjiev, S., Llerer, T., Asuncion, A., Newman, D. and Smyth, D.(2012).Topic-nets: Visual analysis of large text corpora with topic modeling. *Journal ACM Transactions on Intelligent Systems and Technology*, 3(2).

Guha, S., Rastogi, R. and Shim, K.(1998).Cure: an efficient clustering algorithm for large databases.*1998 ACM SIGMOD International Conference on Management of Data*, 73-84.

Guha, S., Rastogi, R. and Shim, K.(1999). Rock: A robust clustering algorithm for categorical attributes. *1999 15th International Conference on Data Engineering. IEEE Computer Society*, 512-521.

Hofmann, T.(1999). Probabilistic latent semantic indexing.*1999 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57.

https://ieeexplore.ieee.org/Xplore/home.jsp (Retrieved on February 2019)

https://www.sciencedirect.com (Retrieved on February 2019)

Jain, A. and Dubes, R.(1988).Algorithms for Clustering Data. Prentice-Hall.

Jolliffe, I.(2002). Principal Component Analysis. Springer.

Kaufman L.and Rousseeuw, P.(2005).Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics.

King, B.(1967).Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317),86-101.

Kuhn, A.,Ducasse, S. and G´ırba, T.(2007).Semantic clustering: Identifying topics in source code. *Information and Software Technology*, 49(3)3, 230-243.

Lambiotte, R. Delvenne, J.C. and Barahona, M.(2009). Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770.

Li, W. and Schuurmans, D.(2011).Modular community detection in networks. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, ser. 1, 22, 1366-1371.

Liu, S., Zhang, R. and Kishimoto, T.(2020).Analysis and prospect of clinical psychology based on topic models: hot research topics and scientific trends in the latest decades. Psychology ,Health and Medicine, 1-13.

MacQueen, J.(1967).Some methods for classification and analysis of multivariate observations. *1967 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

Morrison, J. and George, J.F.(1995).Exploring the Software Engineering component in MIS research, *Communications of the ACM*, 7(38), 80-91.

Salton, G. and Buckley, C.(1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5),513-523.

Sneath, P.and Sokal, R.(1973).Numerical Taxonomy: The Principles and Practice of Numerical Classification. San Francisco: W. H. Freeman.

Sohrabi, B., Vanani, I.R., Jalali S.M.J.and Abedin , E.(2019).Evaluation of Research Trends in Knowledge Management : A Hybrid Analysis Through Burst Detection and Text Clustering. *Jourrnal of Information & Knowledge Management*, 18(4), 1950043-1 - 1950043-27.

Strehl, A. and Ghosh, J.(2000).A scalable approach to balanced, high- dimensional clustering of market-baskets. *2000 7th International Conference on High Performance Computing. Springer- Verlag*, 2000, 525-536.

Wang, F., Orton, K., Wagenseller, P. and Xu, K.(2018).Towards Understanding Community Interests with Topic Modelling. *IEEE Access*, 6, 24660-24667.

Wasserman, S. and Faust, K.(1994). Social Network Analysis Methods and Applications. Cambridge University Press.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M., Qian, W., Shi, L., Tan, L. and Zhang, Q.(2010).Tiara: a visual exploratory text analytic system. *2010 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 153-162.

Zahn, C.(1971).Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 20(1), 68-86.

***