

Measuring interdisciplinarity of scholarly objects using an Author-Citation-Text model with a new measure

Min-Gwan Seo, Seokwoo Jung, Kyung-min Kim, and Sung-Hyon Myaeng
Korea Advanced Institute of Science and Technology (KAIST)
swbliss1201@gmail.com

Contents

1. Introduction

- ▶ Interdisciplinary research
- ▶ Previous Model and Measure
- ▶ Problems and Approach

2. Proposed Method

- ▶ Scholarly Object Model
- ▶ Interdisciplinarity Measure

3. Evaluations

- ▶ Dataset
- ▶ Model
- ▶ Measure

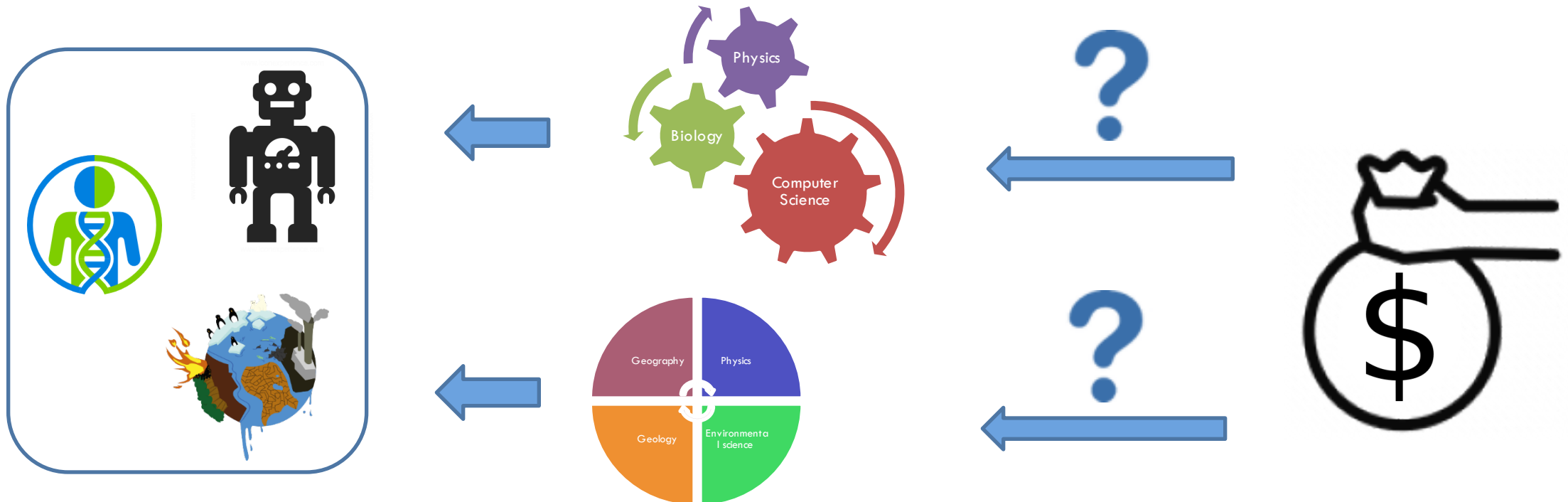
4. Conclusion

- ▶ Summary
- ▶ Contribution
- ▶ Future Work

Introduction | Interdisciplinary Research

► The necessity of interdisciplinarity measure

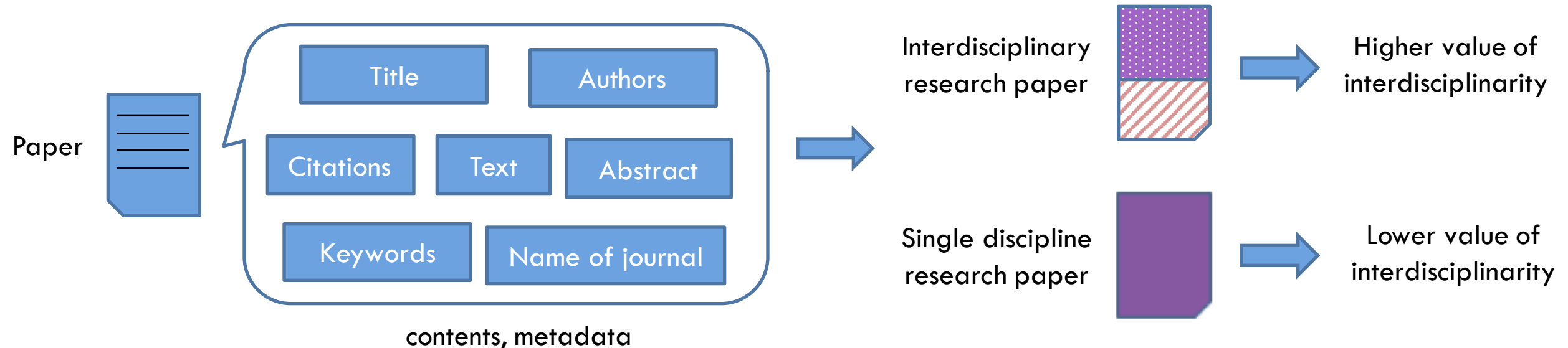
- Interdisciplinary research, team, project, policy: a key tool for the huge problem
- Limited resources: need to determine which project to support



Introduction | Interdisciplinary Research

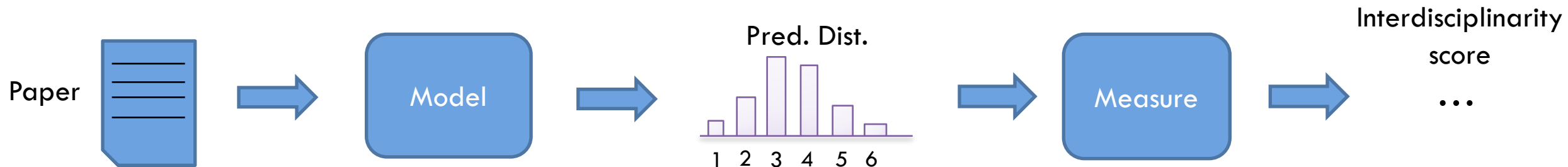
► Interdisciplinarity of papers, journals and conferences

- The problem to measure **the degree of interdisciplinarity** of a given paper, journal, conference based on its **data(contents, metadata)**



Introduction | Previous Model and Measure

- ▶ Previous works: Scholarly object model and interdisciplinarity measure
 - ▶ Model: predict the distribution of the disciplines from the given scholarly object
 - ▶ Measure: calculate the degree of interdisciplinarity from the distribution



Introduction | Previous Model and Measure

- ▶ Previous Model for the scholarly object

- ▶ Method

- ▶ Counting the disciplines of citing papers (Porter et al, 2009; Leydesdorff et al, 2009; Rafols et al, 2010)

- ▶ Counting the major disciplines of author (Aydinoglu et al, 2015)

- ▶ Other metadata (Morillo et al, 2001)

- ▶ Percentage of documents with only the main section and no secondary sections, by journal

- ▶ Average number of sections per document, by journal

- ▶ Percentage of different sections in each journal

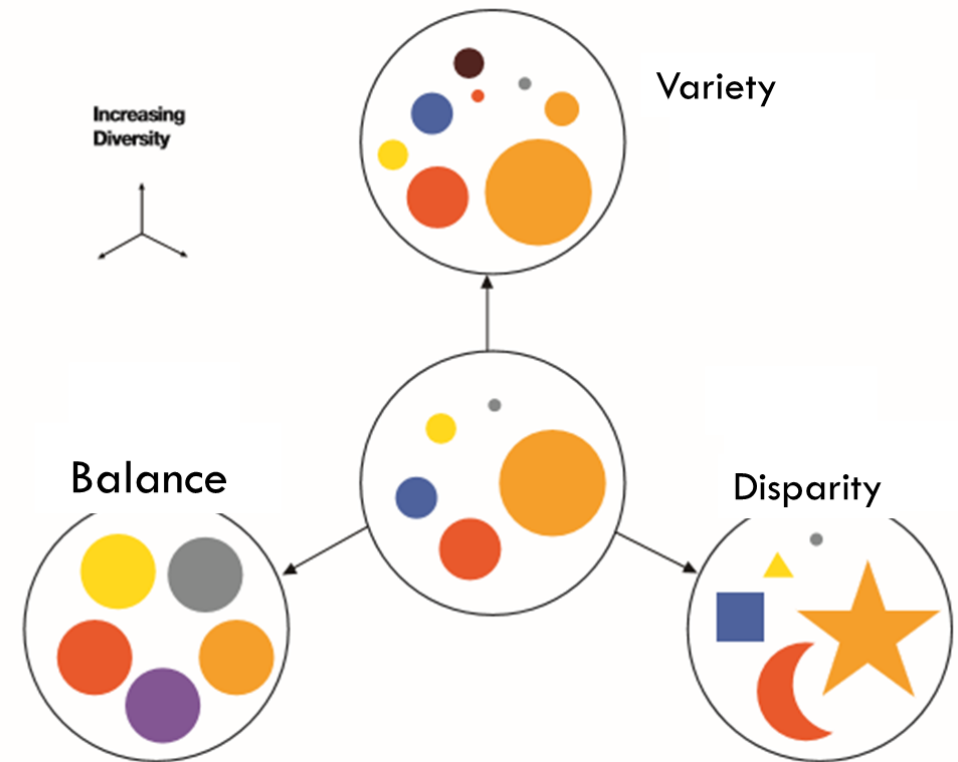
- ▶ In many previous works, they just use the citation information

Introduction | Previous Model and Measure

▶ Previous interdisciplinarity measure

▶ Diversity from Ecology and Economics

- ▶ Variety: The number of discipline
- ▶ Balance: Evenness among disciplines
- ▶ Disparity: Difference between disciplines



Introduction | Previous Model and Measure

► Previous interdisciplinarity measure

Attribute	Name of Interdisciplinarity Measure	Form
Variety	Category count	N
Evenness (Variety / Balance)	Shannon Entropy	$H = - \sum_{i=1}^n p_i \log p_i$
	Gini-Simpson Index	$S = 1 - \sum_{i=1}^n p_i^2$
Disparity	Total Dissimilarity	$A = \sum_{ij} (1 - s_{ij})$
Diversity (Disparity / Variety / Balance)	Diffusion Value	$V = 1 - \sum_{ij} (p_i p_j s_{ij})$
	Stirling's diversity	$H_{st} = \sum_{i,j} d_{ij} p_i p_j = 1 - \sum_{i,j} s_{ij} p_i p_j$

Introduction | Problems and Approaches

► Problems

► Previous model

- Used data separately, mainly used only citation information, text data is ignored
- Citation count cannot reflect the discipline information exactly

► Previous measure

- Few disciplines with high distribution
shows low score

► Measure evaluation

- Previous interdisciplinarity measures are not compared
because of the lack of baseline data



Introduction | Problems and Approaches

▶ Proposed Model

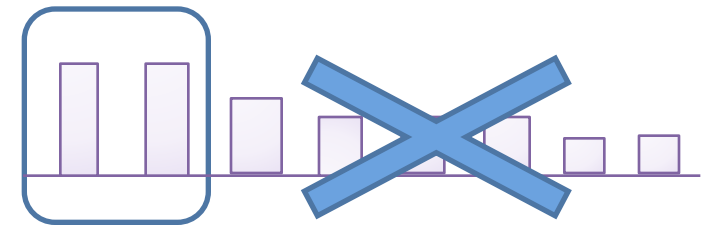
- ▶ Use author, citation, text(= abstract) information simultaneously to predict the distribution of disciplines

▶ Proposed Measure

- ▶ Focus some salient disciplines and ignore other disciplines

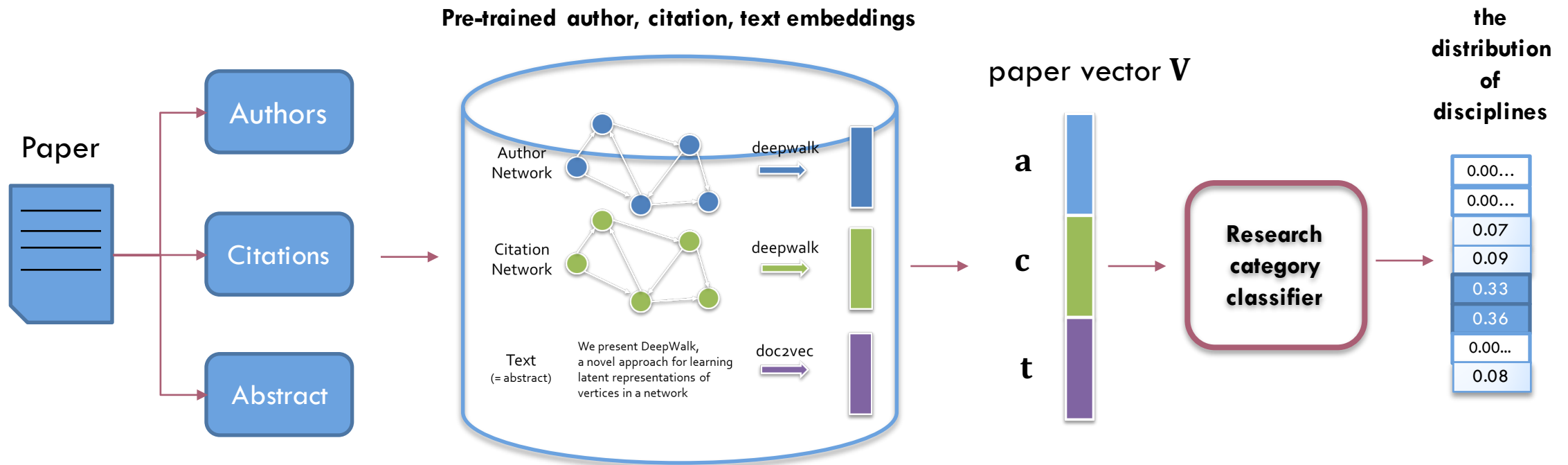
▶ Evaluation

- ▶ Analyze the effect of each features for the model
- ▶ Compare interdisciplinarity measures based on the human annotated data



Proposed Method | Scholarly Object Model

► Proposed Scholarly Object Model: Prediction Flow



a, c vectors are the mean value of all author vectors and citation vectors in the paper

Proposed Method | Scholarly Object Model

▶ Proposed Model: Scholarly Object Model

- ▶ A joint data model with author, citation, text data for predicting the distribution of disciplines
- ▶ Author
 - ▶ The major disciplines of authors, co-author relation
- ▶ Citation
 - ▶ Commonly used attributes, mainly cited papers will be different according to the discipline
- ▶ Text
 - ▶ Some well-known methodology used without citation → reflect this information from text data
- ▶ Author, Citation, Text data is organized by network/document embedding

Proposed Method | Scholarly Object Model

► Word embedding(word2vec)

- learn the latent representation(vector) of **words** reflects the meaning of the word
- Trained by predicting the output word from the input(context) words (CBOW)

predicting the output(context) words from the input word (Skip-gram)

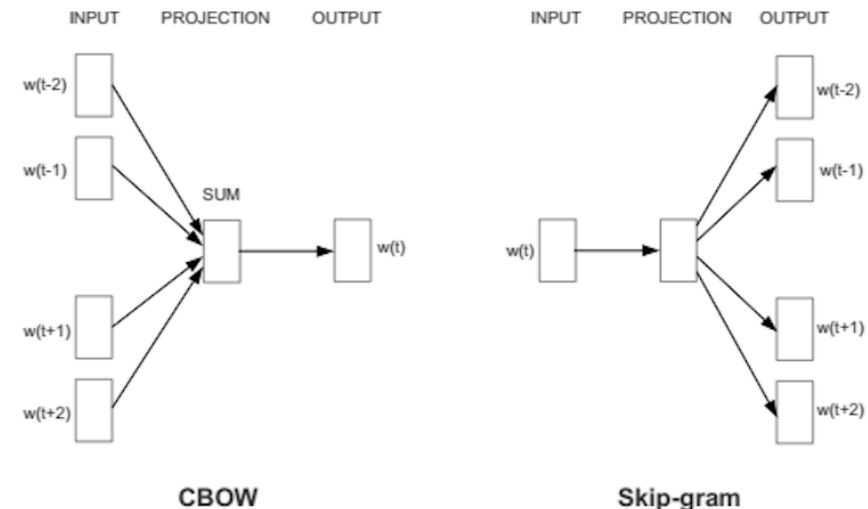
- Skip-gram

Maximize $\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$

- CBOW

Maximize $\frac{1}{T} \sum_{t=1}^T \log p(w_t | w_{t-c}, \dots, w_{t+c})$

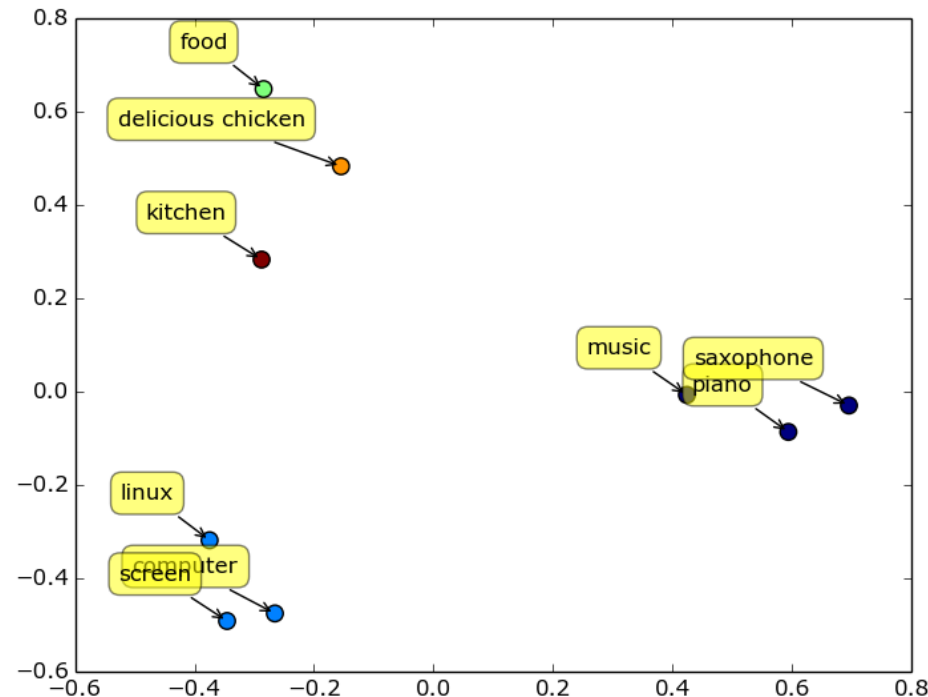
Where
$$p(w_o | w_I) = \frac{\exp(v'_{w_o} v_{w_I})}{\sum_{w=1}^W \exp(v'_w v_{w_I})}$$



Proposed Method | Scholarly Object Model

► Word embedding(word2vec)

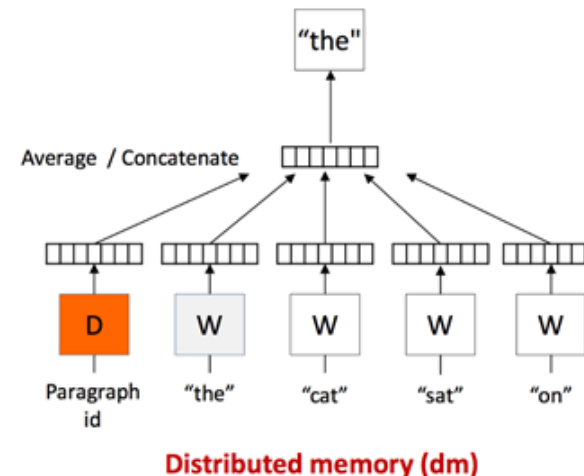
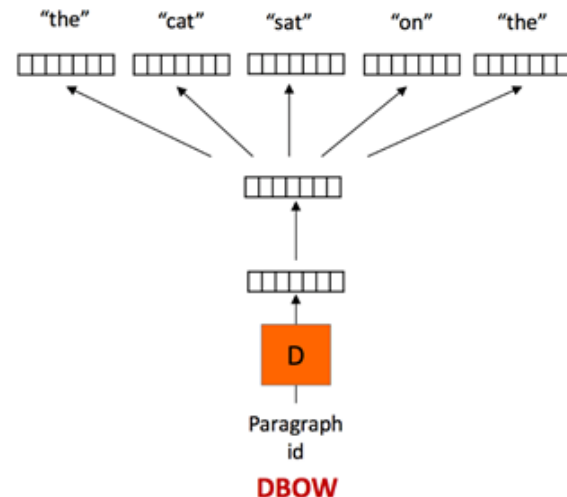
- Similar word vectors are located closely in the vector space



Proposed Method | Scholarly Object Model

► Document embedding(Doc2vec)

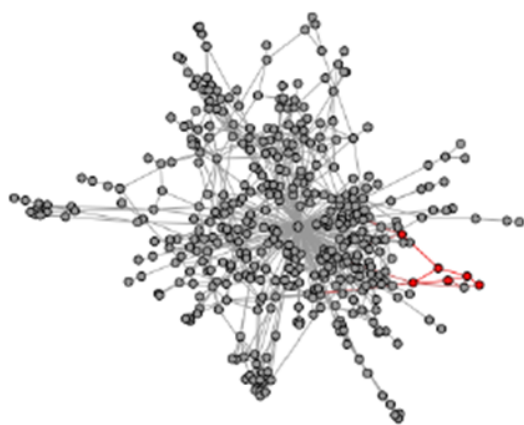
- learn the latent representation(vector) of **a document** reflects the meaning of the document
- Trained by predicting the output word from the document vector (DBOW)
 - predicting the output(context) words from the input word with the document vector (DM)
- Used DM for the better performance (Le et al. 2014)



Proposed Method | Scholarly Object Model

► Network embedding(Deepwalk)

- learn the latent representation(vector) of **a node** reflects the structure(neighborhood) of the graph
- From the graph, generate sentences(= sequences of nodes) by random walk, learn latent representations



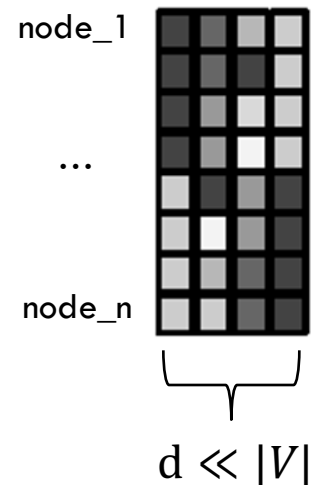
Random walk

Sequences of nodes = sentences

v_{71}	\rightarrow	v_{24}	\rightarrow	v_5	\rightarrow	v_1	\rightarrow	v_{17}	\rightarrow	v_{80}	\rightarrow
v_{92}	\rightarrow	v_2	\rightarrow	v_3	\rightarrow	v_1	\rightarrow	v_{12}	\rightarrow	v_{73}	\rightarrow
v_{37}	\rightarrow	v_{34}	\rightarrow	v_9	\rightarrow	v_1	\rightarrow	v_{10}	\rightarrow	v_{94}	\rightarrow
v_{73}	\rightarrow	v_{64}	\rightarrow	v_5	\rightarrow	v_1	\rightarrow	v_{12}	\rightarrow	v_1	\rightarrow
v_{75}	\rightarrow	v_{14}	\rightarrow	v_6	\rightarrow	v_1	\rightarrow	v_{13}	\rightarrow	v_{61}	\rightarrow

Skip-gram

Embeddings
for each node



Proposed Method | Scholarly Object Model

▶ Proposed Model: Scholarly Object Model

- ▶ Text(= abstract) data → document embedding
 - ▶ pre-trained from the paper abstract data by **doc2vec** algorithm
- ▶ Author & Citation data → network embedding
 - ▶ pre-trained from the author network, citation network by **deepwalk** algorithm
 - ▶ Author network: a node is an author and an edge is co-author relation between two authors
 - ▶ Citation network: a node is a paper and an edge is citation relation between two papers

- ▶ Construct paper vector $\mathbf{v} = \mathbf{a} \oplus \mathbf{c} \oplus \mathbf{t}$
 - ▶ \mathbf{t} is a text vector for a given paper
 - ▶ \oplus is a vector concatenation operator
 - ▶ $\mathbf{a} = \frac{\sum_i a_i}{|a|}, \mathbf{c} = \frac{\sum_i c_i}{|c|}$ where $a \in \mathbf{A}, c \in \mathbf{C}$
 \mathbf{A}, \mathbf{C} : the set of author vectors, citation vectors for a given paper
- ▶ Train multi-label(discipline) classifier based on the paper vector $\mathbf{v} = \mathbf{a} \oplus \mathbf{c} \oplus \mathbf{t}$
→ predict the distribution of the disciplines

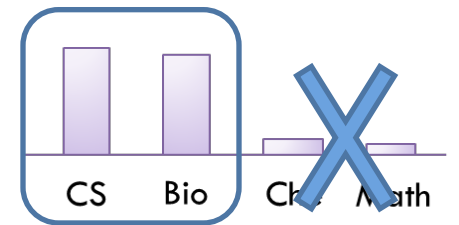
Proposed Method | Interdisciplinarity Measure

► Proposed Interdisciplinarity Measure

► Idea: focus salient disciplines of a paper for the interdisciplinarity

► Partition step: divide disciplines into salient/not salient

► largest gap, k-means(k=2) with init for partitioning



► Modified evenness of salient disciplines $\sum_i \frac{(p_i + L_1) \log(p_i + L_1)}{p_1}$ in H

► The size of salient disciplines $|H|$

► Sum of distances of salient disciplines $\sum_{i,j} d_{i,j}$

► Various distances are available: Euclidean, (1-cos)

Algorithm 1 Proposed Interdisciplinarity Measure

```
1: procedure IDSCORE( $D$ )  
2:   input: distribution of disciplines  $D$   
3:   output: interdisciplinarity score based on the salient discipline set  
4:    $H, L \leftarrow \text{partition}(D)$   
5:   return  $\frac{\sum_i (p_i + L_1) \log(p_i + L_1)}{p_1} |H| \sum_{i,j} d_{i,j} \quad \forall i,j \in H$ 
```

Proposed Method | Interdisciplinarity Measure

▶ Partitioning

- ▶ Sort the distribution of disciplines
- ▶ Automatically determine the threshold

1. Largest gap

- ▶ Divide the partitions with the largest distribution gap between two sorted disciplines

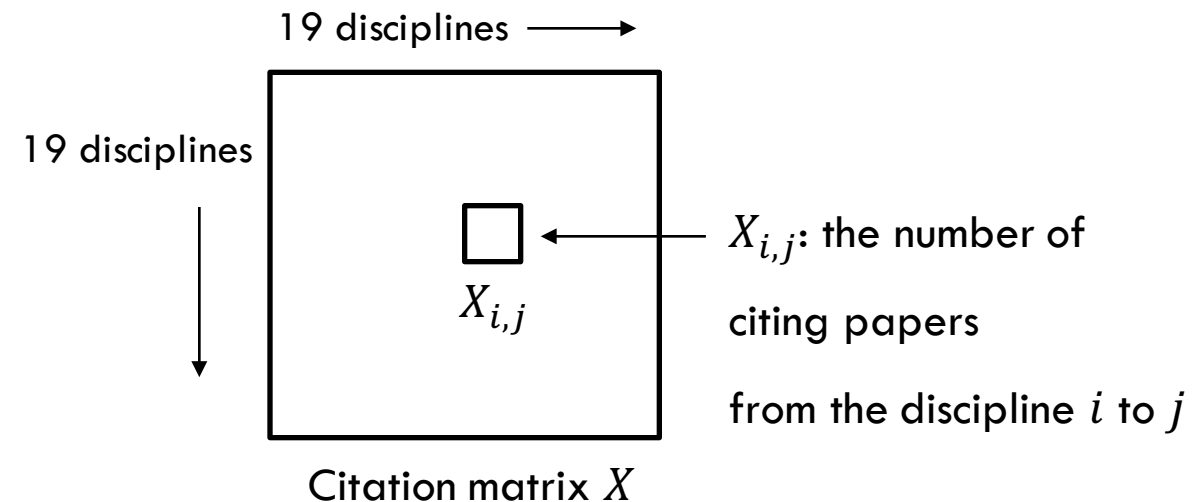
2. k-means($k=2$) with init

- ▶ 1-D k-means clustering
- ▶ Use p_{max}, p_{min} for the initial value
- ▶ Return the H, L cluster with contains p_{max}, p_{min} respectively

Proposed Method | Interdisciplinarity Measure

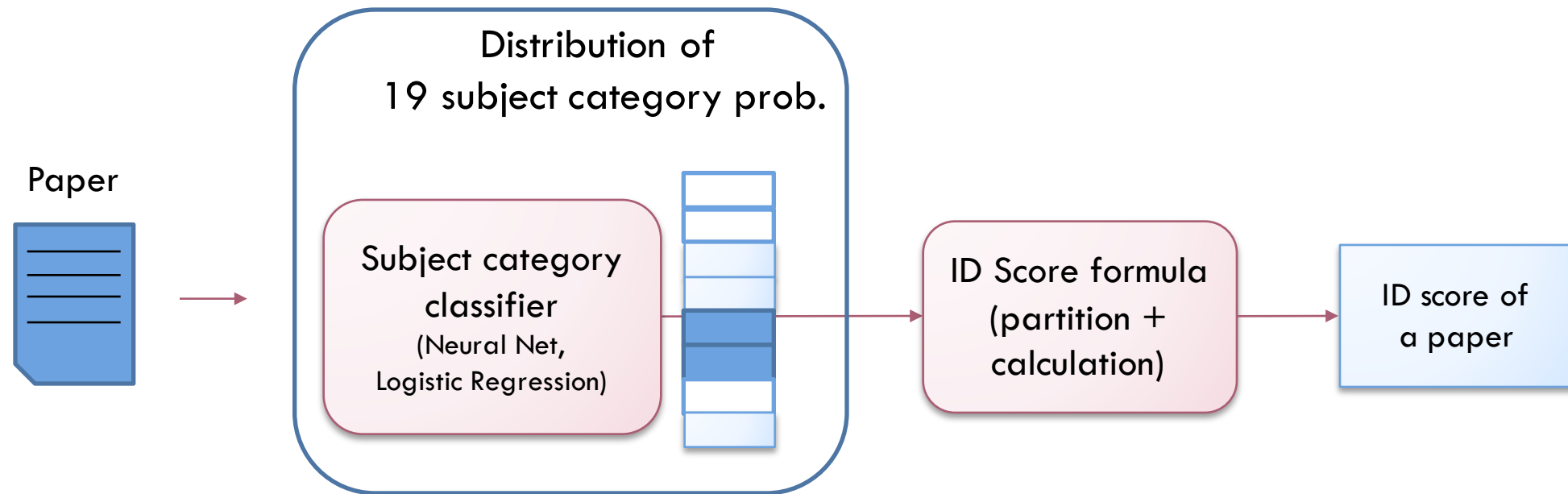
- ▶ Distance between disciplines i and j : $d_{i,j}$
 - ▶ Citation matrix X between disciplines are constructed
 - ▶ $X_{i,j}$: the number of citing papers from the discipline i to j
 - ▶ X_i : the citing vector from the discipline i to other disciplines
 - ▶ $d_{i,j}$: $1 - \cos(X_i, X_j)$

Where
$$\cos(X_i, X_j) = \frac{\sum_{k=1}^n X_{i,k} X_{j,k}}{\sqrt{\sum_{k=1}^n X_{i,k}^2} \sqrt{\sum_{k=1}^n X_{j,k}^2}}$$



Proposed Method

► Whole Interdisciplinarity Calculation Flow



Evaluations | Dataset

▶ Dataset

▶ MAS papers between 1994~2015 → Sub-collections for every 3 years

▶ The number of authors increases steadily

▶ The number of target papers decreases after 2009

Year	1994	1997	2000	2003
# of unique authors	2,345,006	2,459,321	2,561,151	2,809,515
# of whole papers	1,697,000	1,748,000	1,769,000	1,921,000
# of target papers	141,367	208,253	359,657	487,069
Year	2006	2009	2012	2015
# of unique authors	3,026,501	3,218,836	3,453,158	3,608,795
# of whole papers	1,818,000	1,698,000	1,663,000	1,648,000
# of target papers	681,916	800,585	786,753	593,515

Early papers without citation and/or abstract were eliminated

Evaluations | Dataset

► Paper attributes from MS academic search API

Id	Entity ID
Ti	Paper title
Y	Paper year
D	Paper date
CC	Citation count
AA.AuN	Author name
AA.AuId	Author ID
AA.AfN	Author affiliation name
AA.AfId	Author affiliation ID

Id	Entity ID
F.FN	Field of study name
F.FId	Field of study ID
J.JN	Journal name
J.JId	Journal ID
C.CN	Conference series name
C.CId	Conference series ID
RId	Reference ID
W	Words from paper title/abstract for full text search
E	Extended metadata (including abstract)

Evaluations | Dataset

► Discipline labeling

- The discipline of the paper is not provided
- Pre-defined 19 main disciplines and 268 sub disciplines
- Exact string matching between disciplines and paper keywords
- **Conservative labeling**
→ prediction is needed

```
{  
  "Ti": "quality of life improvement  
in patients treated with degarelix versus ...",  
  "F": [{"FId": 71924100, "FN": "medicine"}, ...],  
  "W": [..., "algorithm", "cancer", "effect", ...]  
  "E": {  
    "D": "Purpose: We used responses ...",  
  }  
  
  "categories": [0, ..., 1, 0, ...]  
}
```

matching

MS Academic Search
categories

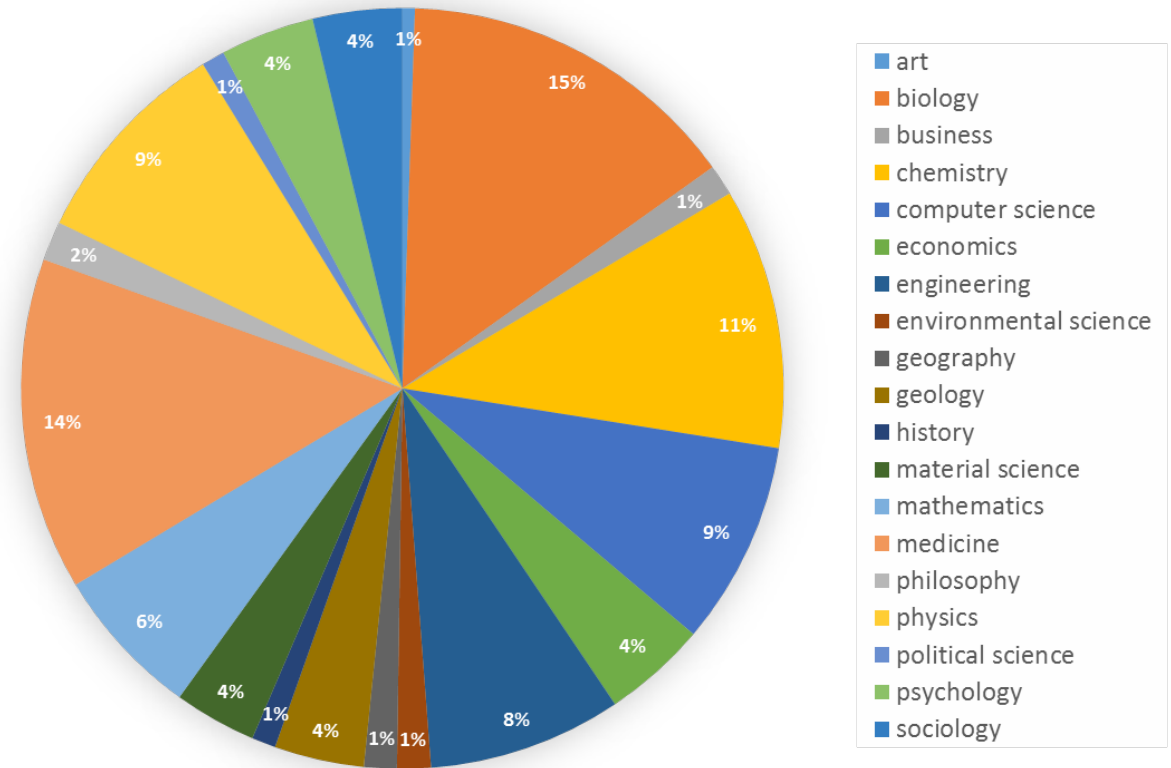
...
materials science
mathematics
medicine
...

Append category label

Evaluations | Dataset

► Dataset

- The proportion of disciplines for whole 11,344,264 papers (from 1994 to 2015 for every 3 years)
- The proportion of the highest discipline: 15%
The proportion of the lowest discipline: 1%
- Humanities (art, business, history, etc.) shows lower proportion



Evaluations | Model

▶ Proposed Model

- ▶ Goal: analyzing the performance and characteristics of the features (C, AC, ACT) and classifiers (Neural network, Logistic regression)
- ▶ 10-fold cross validation from MSA paper data per each year
- ▶ Measure: Average Label Precision(compare as label),
Average JS-divergence(compare as distribution)
- ▶ Neural network
 - ▶ 2 hidden layers, Cross-entropy, Rectified Linear Unit, AdaDelta, 128 mini-batch
- ▶ Logistic regression
 - ▶ Cross-entropy, L2 penalty, One-vs-rest scheme for the multi-label problem

Evaluations | Model

▶ JS-divergence

- ▶ Consider the label as a distribution and compare the difference between distributions (compare as distribution)
- ▶ Jensen-Shannon divergence: the similarity between two probability distributions
- ▶ $[0, 1]$ range, close to 0 if two distributions are similar

P, Q : distribution

$$M = \frac{1}{2}(P + Q)$$

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

Evaluations | Model

► JS-divergence

	Logistic regression			Neural Network		
Vector type	C	AC	ACT	C	AC	ACT
Average JS-divergence	0.2852	0.2853	0.2525	0.1532	0.1560	0.1313

► The results of two classifiers are quite different

- Author embedding degrades the prediction performance
- The orders of two predicted distribution are similar,

the shape of predicted distribution from logistic regression classifier is more even

Logistic regression



Neural network



Evaluations | Measure

▶ Setup

- ▶ Model: (C, AC, ACT) X (Logistic regression, Neural Network)
- ▶ Target interdisciplinarity measures: Entropy, Stirling's diversity, Salient(largest gap), Salient(kmeans)
- ▶ Distance: $d_{i,j} = 1 - \cos$ distance based on the citation count between the disciplines
- ▶ Construct baseline data with Human annotators
- ▶ Evaluation method
 - ▶ Spearman's rank correlation coefficient between baseline and interdisciplinarity measure

$d_i = rg(X_i) - rg(Y_i)$: difference b/w rank(X_i), rank(Y_i)
 n : the number of variables

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

Evaluations | Measure

▶ Baseline data

▶ Randomly selected 100 journals/conferences at 2015

- ▶ Hard to evaluate the interdisciplinarity of a paper based on the abstract

- ▶ Calculate the interdisciplinarity score of journal/conference by $\frac{\sum_i v_i}{|J|}$ where $v_i \in J$ v_i : a paper vector

▶ Evaluate the interdisciplinarity of the journal/conference

based on the introduction, aims and scope of the journal/conference

- ▶ 1~5 range of score

- ▶ The guideline focuses on the number of disciplines and the type of integration between disciplines


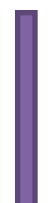
- ▶ Set the score of a journal/conference by voting among 6 human annotators

- ▶ Delete the journal/conference that cannot get three or more agreement

- ▶ Delete the journal/conference that the voting result is 3:3

- ▶ 75 journals/conferences are remained

Evaluations | Measure

 high	Classifier	Logistic regression											
	Feature	C				AC				ACT			
	Measure	ent	stir	salient(lg)	salient(km)	ent	stir	salient(lg)	salient(km)	ent	stir	salient(lg)	salient(km)
	Spearman correlation (r_s)	0.3272	0.3847	0.4101	0.4877	0.3320	0.3847	0.3511	0.4804	0.3689	0.4396	0.3686	0.5401
	P-value	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05
Difference between salient disciplines and others (from JS-div)  low	Classifier	Neural network											
	Feature	C				AC				ACT			
	Measure	ent	stir	salient(lg)	salient(km)	ent	stir	salient(lg)	salient(km)	ent	stir	salient(lg)	salient(km)
	Spearman correlation (r_s)	0.4960	0.5527	0.4422	0.4299	0.4949	0.5517	0.4625	0.4323	0.5071	0.5700	0.4826	0.5732
	P-value	<0.05	<0.05	0.0712	0.1183	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

- Proposed measure shows similar performance with the Stirling's diversity when salient disciplines have higher proportion

Accurate order & lower other disciplines (from label precision & JS-div)

Evaluations | Measure

Name of journal/conference (abbreviations)	Rank by C – Logi – Ent	Rank by ACT – NN – Salient(km)
Symposium on Discrete Algorithms (SIAM)	125	61
International Joint Conference on Artificial Intelligence (IJCAI)	679	326
International Conference on Machine Learning (ICML)	854	210
Conference on Computer Vision and Pattern Recognition (CVPR)	856	212
Conference on Human Factors in Computing Systems (CHI)	1099	979
Symposium on User Interface Software and Technology (UIST)	936	284
International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)	775	485
International Conference on Management of Data (SIGMOD)	406	124
International Conference on Data Engineering (ICDE)	488	578
International Conference on Data Mining (IDCM)	951	520
World Wide Web Conference (WWW)	981	493
Conference on Information and Knowledge Management (CIKM)	889	177
Conference on Computer and Communications Security (CCS)	165	80
International Conference on Software Engineering (ICSE)	665	547
Symposium on the Foundations of Software Engineering (FSE)	515	302
Nature	694	847
Science	588	746
Cell	483	830
Proceedings of the National Academy of Sciences of the United States of America (PNAS)	607	861

- Total 1156 journals/conferences
- High Interd
→ bigger rank
- Red: ML
- Specialized J/C shows low rank
- Rank of Nature, Science, Cell, PNAS
→ increases significantly

Conclusion

▶ Summary

▶ Model

- ▶ Using different types of data simultaneously can improve the performance of predicting the distribution of disciplines
- ▶ The author embedding based on the co-author relation only is not useful to prediction
- ▶ Reflecting the meaning of text(text embedding) can be useful for predicting the distribution

▶ Measure

- ▶ Focusing some salient disciplines can provide more similar results with the human evaluation when some salient disciplines record high proportion and/or other disciplines have even proportion

Conclusion

► Contribution

- We proposed a joint data(ACT) model to predict the distribution of disciplines
- We proposed an interdisciplinarity measure that can be used for calculating the degree of interdisciplinarity from the predicted distribution
 - Proposed measure shows more similar results with the human evaluators with more accurate predicted distribution
- We evaluated proposed approach with real-world data

THANK YOU