Do We Present the Same What We Publish? Answering with a Statistical Contrast Mining Approach

Chandrima Bhattacharya
Department of IT
Indian Institute of Engineering Science and Technology, Shibpur
Howrah – 711103, India

Rajat Kumar Agarwal
Department of IT
Indian Institute of Engineering Science and Technology, Shibpur
Howrah – 711103, India

Goutam Debnath
Department of IT
Indian Institute of Engineering Science and Technology, Shibpur
Howrah – 711103, India

Malay Bhattacharyya*
Department of IT
Indian Institute of Engineering Science and Technology, Shibpur
Howrah – 711103, India

ABSTRACT

There is a general tendency of publishing research as a paper and then presenting it in some conferences. The paper published, the slide carried to be presented and what the speaker actually speaks in the conference if noticed carefully sometimes tends to differ in weightage. Now there happen to be some topics which are more stressed upon. Here in this paper, we try to find the relationship between the paper published, and a Conference presentation on that paper (both what is prepared as a slide as well as what is spoken) to understand the psychological behaviour of a person who is presenting a paper.

CCS Concepts

•**Computer systems organization** → **Embedded systems;** *Redundancy;* Robotics; •**Networks** → Network reliability;

Keywords

Contrast mining; word cloud; research publication

1. INTRODUCTION

According to the Oxford Dictionary, research is ŞThe systematic investigation into and study of materials and sources to establish facts and reach new conclusionsŤ. Hence we can conclude that any systematic investigation is a type of research. Now with such a broad definition, we have enormous amount of new researches on different fields going on. With the prodigious amount of research going on, we find there is an inclination towards publishing the researches.

Research is published as literature, or presented in Conferences. The idea behind the research can also be protected as a patent. Now there is a global fashion of publishing researches in journals, or trying to present an idea in a Conference. The society considers paper publication as a merit which is leading the whole concept of research turning to be a craze amongst young researchers. The whole craze of research publication has led to unethical norms used to get papers published. We find that in the new age people send the same materials to be published in multiple journals simultaneously, new journals opening which instead charge the person who is trying to publish the materials.

After research is done, and after getting published, it can also are presented in conferences. Now a very interesting question is that is the research that is published as a literature is same as that which is presented in a conference? What are the psychological thoughts behind delivering the same research content in different medium? Are all the topics from the paper taken up while giving a speech on it? Does the slide for the presented at the Conference contain a brief overview of all the topic? With that motivation, we start our project to find the similarities between the research that is published, the presentation made for the conferences, and what is spoken during the conference.

2. MOTIVATION

With changing time, we see people more interested into data analysis. Contrast mining is a new field of learning similarities and differences between related groups with the help of reverse engineering techniques. We have used the concepts of contrast mining and tried to find out the relationship between the three kinds of data Ű the literature published, a conference power point presentation and what is spoken during the conference. This given work is based

*The correspondence should be made to malaybhattacharyya@it.iiests.ac.in.

ACM KDD ODD 4.0 San Francisco, CA, USA

DOI: ...

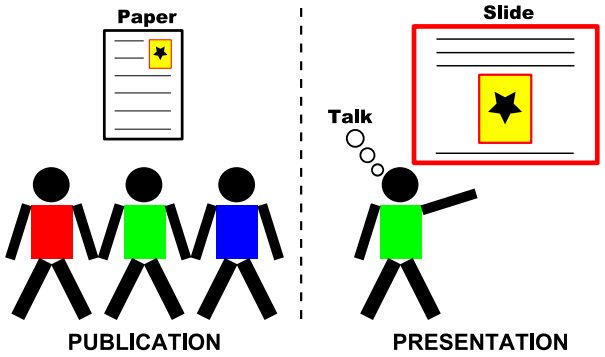

Figure 1: The three form of datasets available for contrast mining.

on Contrast mining. The three sets taken for the project, namely the paper published, the slide prepared and the way it is explained, forms a contrast set. A contrast sets can be defined as a conjunction of attributes or values that meaningfully in its distribution across different groups. Now given the data sets, we can detect the differences between contrasting groups. This approach is referred as contrast mining. This is a new field in data mining where we try to find the differences of a contrast set. This ... is explained in Fig. 1.

3. DATASET COLLECTION

The International Conference on Intelligent System for Molecular Biology (ISMB) maintains a database of the slides and presentation videos for the corresponding published papers. We took ten papers to make our dataset. Hence our dataset consists of three materials: the paper, i.e., the published literature, the conference talk and the slide presented during the conference. We have a total 30 articles for the comparative analysis, consisting of 3 sets of data for each set. The details about the dataset are provided in Table 1.

We have excluded diagrams, graphs, chart, bar graphs, tables, etc. while dataset collection, but have kept the description and captions along with it.

Crowdsourcing techniques have been used to convert the datasets. Parts of the video has been shared to be converted to data vector.

4. METHODS

We have taken the data collected and removed the stop words. The data without the stop words comprises of our data vectors for analysis. We have selected TagCrowd and have created the tag crowd for each dataset to visualize the crowd and the frequency of each. We have kept the language specification as English because all our data are in English. Moreover by grouping similar words, converting to lower case and showing frequency, we have a better representation of all the datasets present. The visual perception increases thereby. The minimum frequency is set to two. We did not consider a frequency of one because we are trying to group most frequently used words, and it is not set to five or some other number as we were missing on some important words in the slide by doing so. The frequency of words in slide is less than that of paper or videos. The maximum frequency have been set to 50. This was done after observing the pattern for 25, where we were missing many important words, and also by setting it at 75, where we were getting many unnecessary words. The parameter values considered are listed in Table 5.

Table 2: The parameters chosen for analysis with TagCrowd.

Options	Settings
Language of text	English
Maximum number of words to show	50
Minimum frequency	2
Show frequency	Yes
Group similar words	Yes
Convert to lowercase	Yes

The term frequencyŰinverse document frequency (TF-IDF) statistic is used to represent how important a word is to a document or corpus. It is defined as follows.

$$tfidf(t,d,D) = tf(t,d) * idf(t,D),$$

where

$$tf(t,d) = 0.5 * (1 + \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}),$$

and

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}.$$

Here, tf(t,d) refers to the number of times the term t occurs in document d, N refers to the number of documents in the corpus and N=|D|. As when the term is not present in a particular document will lead to division by 0, we take $1 + d \in D : t \in d\}|$. tf-idf is calculated as the product of tf(t,d) and idf(t,D). As the ratio of the idf's log function is always equal to or greater than one, we value of idf is greater than 0. As a term appears in more and more documents, the ratio inside the idf's logarithm function approaches 1, bringing the idf and tf-idf closer to 0.

5. EMPIRICAL ANALYSIS

5.1 Frequency Analysis

The TagCrowd data shows that the highest frequency for one set of data i.e., for the research literature, slide and presentation given. We have noticed that the highest frequency word is not the same for all three forms of data. Moreover we also notice that in videos some casual words which are spoken repeatedly by the author is shown in some cases. If we look at the TagCrowds created by the paper we see the terms 'et' and 'al' occurring frequently. The software cannot comprehend that it is actually 'et al' meaning 'and others'. The occupance of it is unique and found only in the papers. We find that it is not present in the other two. If we notice carefully through the three sets of TagCrowd data, we find the weightage of all the frequently used words are not the same. In set 1, we find that that SuperGenome does not have the same weightage in slide and video as it is in the paper. Likewise, in set 2, fragmentation is used more frequently in paper and while presenting, than in slide. The top 5 words used in published paper, presentd slides and verbal presentation are shown in ... Table 3.

Table 1: Dataset details. Here we have mentioned the speaker for the conference on the given topic.

Serial No.	Paper Title	Presenter
1	GenomeRing: Alignment visualization based on SuperGenome Coordinates	Alexander Herbig
2	Fast Alignment of fragmentation tree	Franziska Hufsky
3	Towards 3D structure prediction of large RNA molecules: An integer programming framework to insert local 3D motifs in RNA secondary structure	Vladimir Reinharz
4	Leveraging Input and Output structures for joint mapping of Epistatic and Marginal eQTLs	Seunghak Lee
5	Incorporating prior Information into Association Studies	Gregory Darnell
6	Matching experiments across species using expression values and textual information	Aaron Wise
7	Dactal- divide and conquer trees (almost) without alignments	Serita Nelesen
8	Efficient Algorithm for recollection problem with gene duplication, horizontal transfer and loss	Mukul S. Bansal
9	MoRFpred, a computational tool for sequence based prediction and characterization of disorder-to-order transitioning binding sites in protein	Fatemeh Miri Disfani
10	A single-source k shortest paths algorithm to infer regulatory pathway in a gene network	Yu-Keng Shih

Table 3: The 5 most frequently used words.

Serial No.	...	...	...
1			
2			
3			
4			
5			

The TagCrowd dataset is not normalized. The length of all the three sets are different. Hence the weightage of the word is not weighted. Moreover, we have selected multi-authored papers. Hence assuming that different author made different contribution. It can be seen from Fig. 2, as only one author is speaking, the frequency of the most frequent word tells us about the domain of the speaker.

5.2 Contrast Mining

We have taken these thirty values and have found the cosine similarity for all the ten papers using the methods defined in above. The cosine similarity has been found within paper and slide, slide and video and paper and video. The thirty values thus obtained are represented in a tabular format in the adjacent page. We observe that the cosine similarity between paper and video is much higher than that obtained from slide and video as well as paper and slide. A close look at the values obtained as well as the graphical representation shows us that. The cosine similarity for slide and video as well as slide and paper tends to be lower. One apparent reason maybe, we have not included images. The explanation for most diagrams and figures were both presented verbally and written in the paper. Hence the anomaly. We also notice that for set 7, the cosine similarity drops and become nearly 0 for both slide and paper as well as slide and video. This can be explained by the fact, maybe the speaker for some reason did not make the slide, hence the speech as well as the video was not much similar to the slide. This is shown in Table 4.

We calculate correlation between cosine similarities of different pairs of forms (that is, between Paper vs PPT and PPT vs Video, between PPT vs Video and Video vs Paper and between Paper vs PPT and Video vs Paper) to show whether and how strongly they are related. The Pearson correlation coefficient between two variables X and Y is given by

$$Cor(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)},$$

where $Cov()$ and $Std()$ denote the covariance and standard deviation, respectively. The values of correlation lie between -1 to 1. The trend as shown by the plot is shown in Fig. 3 in correlation shows a strong similarity between video and slide as well as paper and slide. It shows a value tending to 1, i.e. nearly correlated. This is because the slide and video as well as slide and paper shows a cosine similarity which is very near. We have also to take into consideration the following fact. The dataset was made by listening to video and converting it into text manually, as well as seeing the slide and manual conversion and document editing as per the requirement of the dataset. Hence, even though the chances of mistakes are less in case of slide and paper, there remains a high error probability during manual conversion of video. This error has been neglected during analysis of error.

Table 4: The cosine similarity values computed for pairwise documents (i.e., among the paper and presentation, video and presentation and paper and video).

Serial No.	Paper-PPT	Video-PPT	Paper-Video
1	0.718	0.663	0.870
2	0.705	0.565	0.880
3	0.565	0.564	0.875
4	0.687	0.662	0.872
5	0.679	0.784	0.737
6	0.615	0.404	0.822
7	0.055	0.060	0.688
8	0.580	0.501	0.702
9	0.367	0.315	0.737
10	0.892	0.899	0.835

Table 5: Correlation between the three types of documents taken in pair, via Peterson's Correlation.

Pairs of document	Correlation
Paper vs PPT and PPT vs Video	0.933432618
Video vs PPT and Paper vs Video	0.518246914
Paper vs PPT and Paper vs Video	0.684532761

Figure 3: The above plot is a isualization of frequent words appearing in the in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

6. CHALLENGES

The biggest challenge was getting the datasets. For published literature, finding the conferences it was spoken to, as well as finding the presentation which was given in that conference, had been one of the biggest challenge. ISMB maintaining such a database helped us solve the problem. The next biggest challenge is converting the video, i.e. the presentation by the author, to a document vector for analysis. The difference in accent, speed of presentation have created some issues and hence the document vectors are not cent percent error free. These errors have been ignored while calculation. The errors are due to manual conversion of video to text. We have considered only multi-authored paper for dataset collection. Care has been taken such that no single authored paper is considered for analysis as it might lead to variation in result.

7. CONCLUSION

Here in the project we have manually made the dataset. Even though collecting dataset from paper might seem easy, manually listening and writing from videos can be quiet cumbersome. Moreover due to difference in accent, the error probability increases hence we had to hear each and every line a number of times before finalizing. This project is just one type of contrast mining. Now with this concept, we can proceed to many different analyses. We can compare the slide and videos for all the authors in a multi-authored paper and find their differences. Moreover, we can also find the trend for single authored papers. Moreover, we can use concepts of Crowdsourcing to collect a larger dataset for a more detailed analysis of trend. The result obtained for 10 sets leads us to conclude on the fact how the psychological thought process behind speaking on a paper published work. We see a tight correlation here suggesting that mostly authors of paper tends to speak and cover most of the parts covered in paper during a seminar presentation. The close correlation between the slides prepared and words spoken, or the paper written also gives us many reasons to ponder on. Hence, continuing this project can lead us to try much new type of data similarities and help find trend amongst data for further analysis. We can analyze the trend in single authored paper for understanding their trend, or maybe we might analyze each of the author separately for multi authored paper. If we continue in this direction, we have lots of new discoveries to make in the trend of the psychological thoughts behind presenting and writing a paper.

8. ACKNOWLEDGMENTS

I would like to acknowledge the help of the following individuals without whom this work couldn't have been completed. I would firstly like to thank DR. ARINDAM BISWAS for giving us this work. It helped us gain knowledge and understanding on statistical contrast mining method. A debt of gratitude to my alma mater, Indian Institute of Engineering, Science and Technology, Shibpur and the library for its help.

9. REFERENCES


alignment blocks coordinate genome supergenome visualization
alignment analysis blocks concept different genome reference regions
alignment block contained genome example visualization
(a) (b) (c)
alignment algorithm computing edge fragmentation join node running score trees
alignment dp edge fragmentation ilp running score search trees
alignments alignment compound compute fragmentation instances programming tree
(a) (b) (c)
base components inserted motifs prediction rna secondary structures
base motifs pair prediction rna secondary structure
base final interaction motifs predict really rna secondary structure
(a) (b) (c)
association effects genes group interaction lasso output pairs snp
association expression genes groups individual lasso snps
association expression genes groups individual lasso matrix model snps
(a) (b) (c)
association causal information marker method power prior study threshold
association disease heads marker model power snp studies threshold
association different disease frequency power snps study threshold
(a) (b) (c)
al data expression genes matched method microarray model pairs similarity species training
comparison data expression microarray model pairs training
compare data expression genes human model similar species technique
(a) (b) (c)
alignment dactal datasets estimation methods sequence tree
accurately alignment boosting dactal datasets estimation methods sate tree
alignment dactal data dataset estimate methods subsets tree
(a) (b) (c)
al algorithm dtl gene node problem reconciliation species transfer tree
algorithm duplication gene reconciliation species transfer tree
algorithm gene duplication reconciliation species transfer tree
(a) (b) (c)
dataset morf morfpred prediction residues
disorder morf prediction protien
disorder method morf prediction regions validated
(a) (b) (c)
algorithm diverse gene node paths shortest
algorithm diverse edge gene path shortest
algorithm diversity gene path shortest
(a) (b) (c)

Figure 2: ...