A Project Report on

DO WE SPEAK THE SAME WHAT WE COMPOSE? A STATISTICAL PERSPECTIVE.

Submitted to the Department of Information Technology

For the partial fulfilment of the degree of B.E. in Information Technology

by

Rajat Kumar Agarwal (111308038)
Chandrima Bhattacharya (111308009)
Goutam Debnath (111308041)

Registration number:110813039, 110813009, 110813043 of 2013-17

B.E. 3rd Year

Under the supervision of

Dr. Malay Bhattacharyya

Department of Information Technology
INDIAN INSTITUTE OF ENGINEERING SCIENCE AND TECHNOLOGY, SHIBPUR
May, 2016

Department of Information Technology
Indian Institute of Engineering Science and Technology, Shibpur

CERTIFICATE

This is to certify that the work presented in this report entitled "**Do we speak the same as we compose? A statistical perspective.**", submitted by Rajat Kumar Agarwal, Chandrima Bhattacharya, Goutam Debnath having the examination roll number 111308038, 111308009 and 111308041 has been carried out under my supervision for the partial fulfilment of the degree of Bachelor of Technology in Information Technology during the session 2015-16 in the Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur.

11/05/2016

Dr. MALAY BHATTACHARYYA
Assistant Professor
Department of Information Technology
Indian Institute of Engineering Science
and Technology, Shibpur

Assistant Professor
Department of Information Technology
Indian Institute of Engineering
Science and Technology
Shibpur, Howrah-711103. (W.B.)

Dean (Academic)
Indian Institute of Engineering Science
and Technology, Shibpur

DR. ARINDAM BISWAS
Head of the Department
Department of Information Technology
Indian Institute of Engineering Science
and Technology, Shibpur

Dr. Arindam Biswas
Associate Professor & Head
Department of Information Technology
Indian Institute of Engineering &
and Technology, Shibpur
Howrah-711103(W.B), India

Date: 12/05/2016

Acknowledgements

I would like to acknowledge the help of the following individuals without whom this project couldn't have been completed. I would firstly like to thank Dr. Malay Bhattacharyya for giving us the project. 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.

Rajat Kumar Agarwal
Goutam Debnath
Chandrima Bhattacharya

Date: 12/04/2016

RAJAT KUMAR AGARWAL
CHANDRIMA BHATTACHARYA
GOUTAM DEBNATH
Department of Information Technology
Indian Institute of Engineering Science
and Technology, Shibpur

Abstract

In the last few centuries we have seen great scientific and technical advancements. With these advancements, we have noticed a global trend of documentation of researches as journal papers, and also protecting ideas and knowledge as patents. Conferences held worldwide and papers presented and documented are the source of understanding more about the new scientific discoveries. But now we see that when a research paper is published and what is presented in the conference on that research, the content weightage, etc. is not the same. There happen to be some topics which are more stressed upon. Here in this project, 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). We have tried to find the similarities in the content and hence made an attempt to understand the differences also.

Contents

1. INTRODUCTION

According to the Oxford Dictionary, research is "The systematic investigation into and study of materials and sources to establish facts and reach new conclusions". 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.

This given project is based 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.

2. RELATED WORKS

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 – the literature published, a conference power point presentation and what is spoken during the conference.

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 thirty total articles for comparing.

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

Serial Number	Name of presentation	Person presenting
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 1: List of datasets and authors presenting during the conference.

4. PROPOSED APPROACH

4.1 FREQUENCY ANALYSIS

We have selected TagCrowd and have created the tag crowd for each dataset to visualize the crowd and the frequency of each.

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

Table 2: TagCrowd Options used for analysis

4.2 SIMILARITY ANALYSIS

We selected ten lectures delivered by different authors on their papers in some international conference. We then collected the published documents of the respective papers and the presentations made by the authors in support of their lectures. We ignored all the charts, images, numerical expressions and equations and only collected the text(i.e., words).

We prepared our data set in the form of triplets :
i) Paper published in document form (a .txt file)
ii) Text collected from the presentation(ppt) prepared on the paper (a .txt file)
iii) Text extracted from the video lecture delivered on the paper (a .txt file)
So, we have ten such triplets for statistical analysis.

In order to estimate the similarity between different forms of the same paper, we have decided on the following strategy:
Step 1:
All the non-word characters and numbers (regular expression patterns) are removed from the document.

Step 2:
The document is splitted into tokens(words). A list of all distinct words is prepared for each of the documents.

Step 3:
We convert the document into vector and calculate each of its component values (tf-idf of tokens).
The document is transformed into a vector in Euclidean vector space. Each distinct word of the document represents a dimension in n-dimensional Euclidean vector space.

Say, d is a document with n words. It may be represented as a vector X in n dimensional vector space.

Hence, $\vec{X} = \vec{X}(x_1, x_2, \ldots x_n)$

The values of each of the components are computed using the tf-idf scheme.

Term Frequency(tf): $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.

Inverse Document Frequency(idf): Estimate the rarity of a term in the whole document collection. (If a term occurs in all the documents of the collection, its IDF is zero.)

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

with $|D|$: cardinality of D, or the total number of documents in the corpus $|\{j : t_i \in d_j\}|$: number of documents where the term t_i appears (viz. the document frequency) (that is $n_{i,j} \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to use $1 + |\{j : t_i \in d_j\}|$

Normalized tf: tf count is normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j .

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j , and the denominator is the sum of number of occurrences of all terms in document d_j , that is, the size of the document $|d_j|$.

Tf-idf: The tf-idf weight of a term is the product of its tf weight and its idf weight. So, the tf-idf weight of a term t_i in document d_j is given by

$$\text{tf-idf}_{i,j} = tf_i * idf_j$$

As I have mentioned before, each component of a document vector denotes the tf-idf value of a term.

Despite its strength, TF-IDF has its limitations. In our own experiment, TF-IDF could not equate the word 'drug' with its plural 'drugs' categorizing each instead as separate words and slightly decreasing the word's weight value. For large document collections, this could present an escalating problem.

STEP 4:

We compute cosine similarity between two different document forms of the same paper (i.e., between paper and ppt, between ppt and video and between paper and video). Thus, we get three cosine similarity values for each of the papers and thirty values for all the ten papers in total.

Cosine Similarity of $\vec{x}$ and $\vec{y}$ is defined by

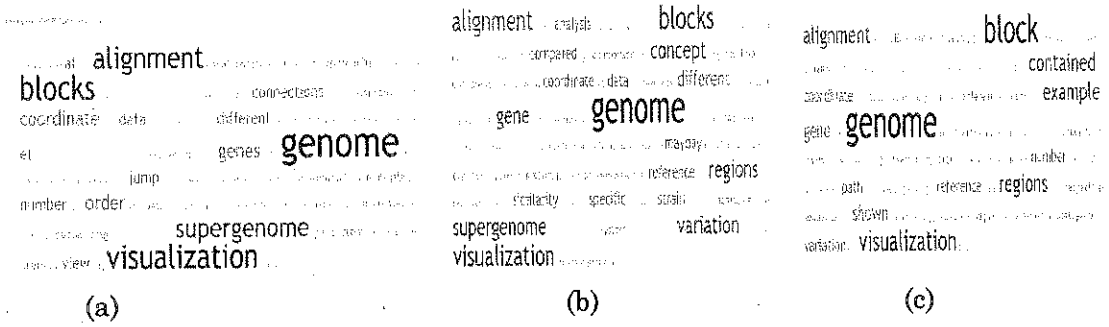
$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$
$$|\vec{y}| = \sqrt{\sum_{i=1}^{n} y_i^2}$$

5. EXPERIMENTAL RESULTS

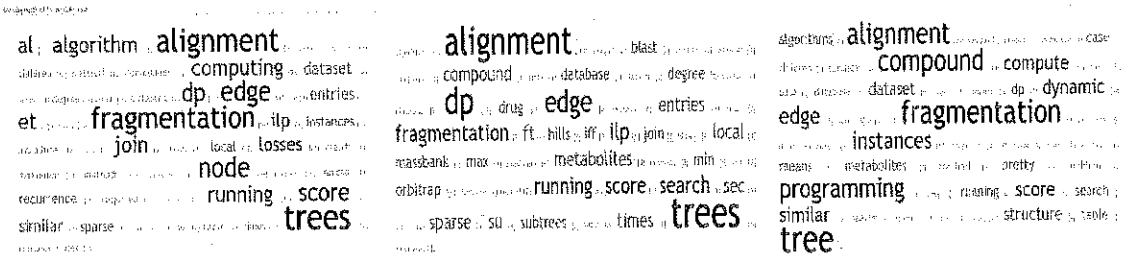
The TagCrowd data shows that the highest frequency for one set of data i.e., for the research literature, slide and presentation given. We have noticed that the highest frequency word is not the same for all three forms of data. Moreover we also notice that in videos some casual words which are spoken repeatedly by the author is shown in some cases.

Below we have attaching the tag crowd for each dataset. We have created separate tag crowds for research literature, slide and conference talk and attached as follows.



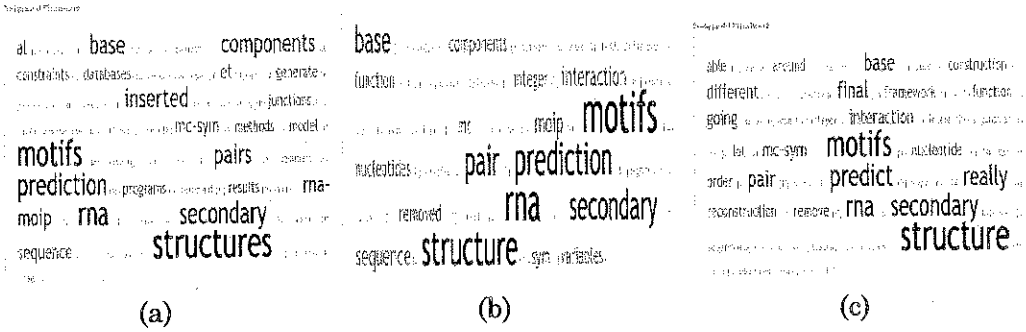
(a) (b) (c)

Figure 1: Visualization of frequent words appearing in the paper [1] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.



(a) (b) (c)

Figure 2: Visualization of frequent words appearing in the paper [2] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.



(a) (b) (c)

Figure 3: Visualization of frequent words appearing in the paper [3] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

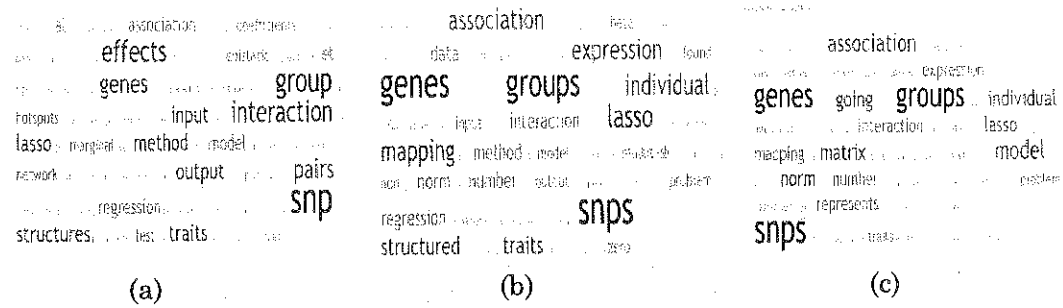


Figure 4: Visualization of frequent words appearing in the paper [4] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

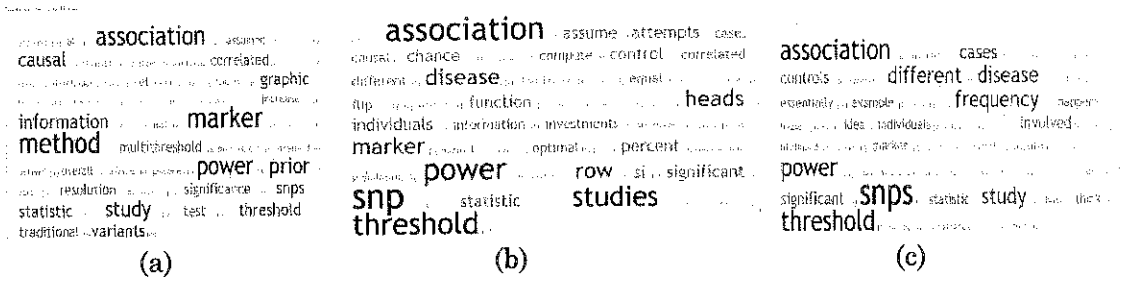


Figure 5: Visualization of frequent words appearing in the paper [5] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

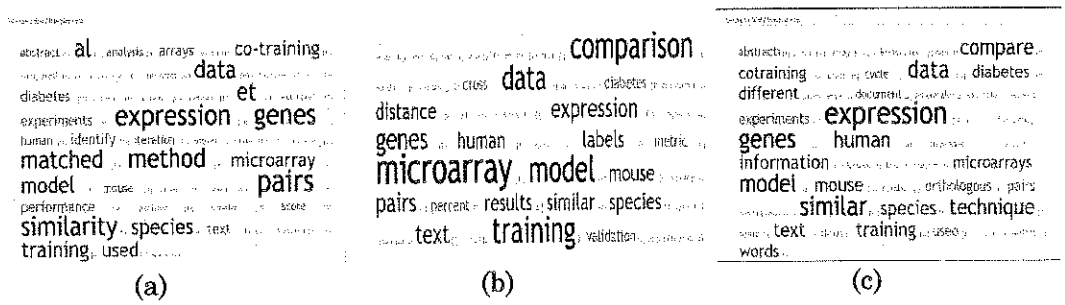


Figure 6: Visualization of frequent words appearing in the paper [6] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

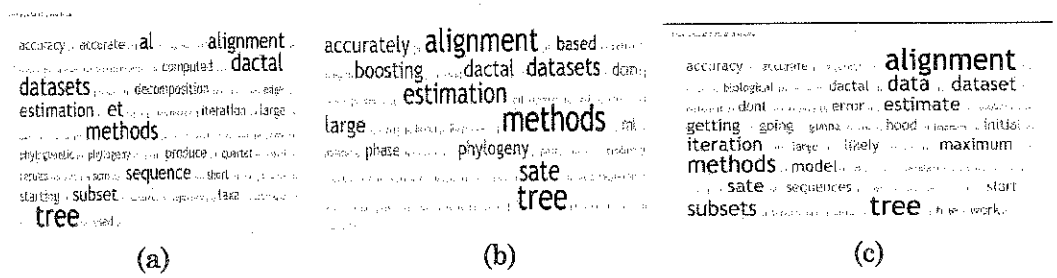
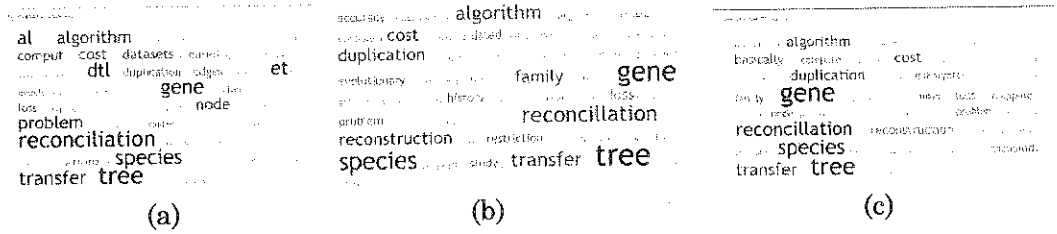


Figure 7: Visualization of frequent words appearing in the paper [7] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

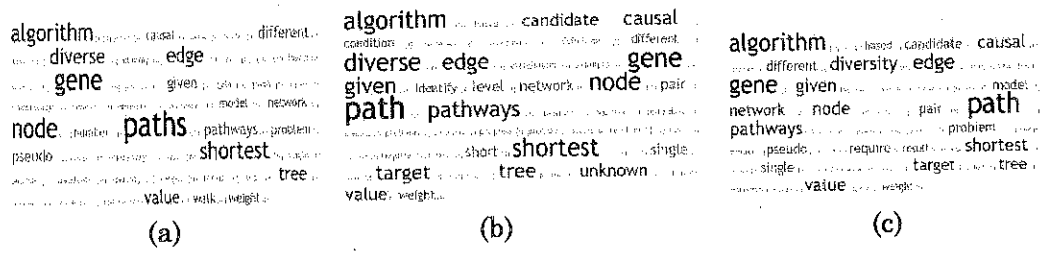


(a) (b) (c)

Figure 8: Visualization of frequent words appearing in the paper [8] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

(a) (b) (c)

Figure 9: Visualization of frequent words appearing in the paper [9] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.



(a) (b) (c)

Figure 10: Visualization of frequent words appearing in the paper [10] in the forms of (a) published paper, (b) presented slides, and (c) verbal presentation.

We have taken these thirty values and have found the cosine similarity for all the ten papers using the methods defined in Section. 4.2. 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.

Serial no.	Paper_PPT	PPT_Video	Video_Paper
1	0.718234364	0.662978071	0.869688232
2	0.704973578	0.564727738	0.880399794
3	0.650960929	0.564006724	0.875257705
4	0.687027481	0.661625297	0.87205514
5	0.679212537	0.784126295	0.737409231
6	0.615406506	0.40350298	0.822265915
7	0.055057098	0.059717816	0.6883728
8	0.579854262	0.500770032	0.701926416
9	0.367124616	0.31492547	0.737477821
10	0.891648508	0.899185584	0.83457229

Table 3: Cosine similarities values obtained by method 4.2 for paper and slide, slide and video and video and paper.

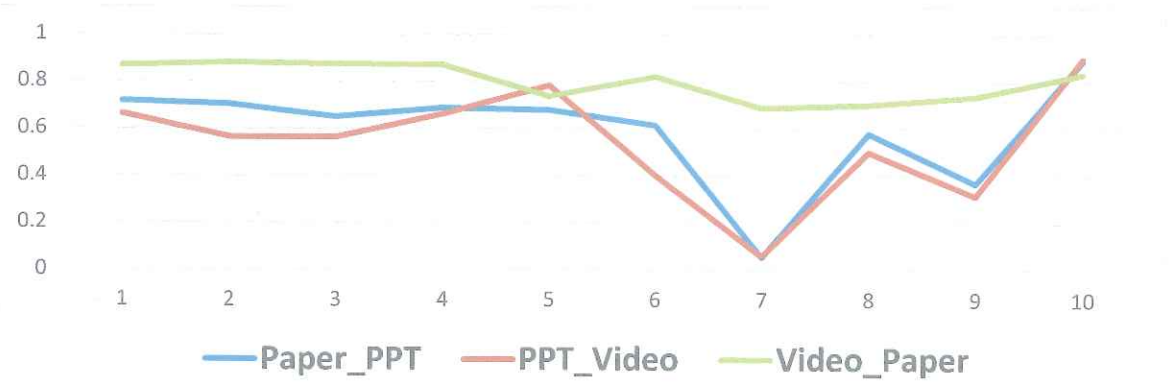


Fig 11: Graphical representation of cosine similarities between paper and slide, slide and video and video and paper

We calculate correlation between cosine similarities of different pairs of forms (that is, between Paper_PPT and PPT_Video, between PPT_Video and Video_Paper and between Paper_PPT and Video_Paper) to show whether and how strongly they are related.

Correlation coefficient $\rho_{X,Y}$ between two variables X and Y is given by

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where cov is the covariance, σ_X is the standard deviation of X.

Pairs of document forms	Correlation
Paper_PPT and PPT_Video	0.933432618
PPT_Video and Video_Paper	0.518246914
Paper_PPT and Video_Paper	0.684532761

Table 4: Correlation between the three types of documents taken in pair.

TagCrowd and the Cosine Similarities can be explained as follows:

- The difference in frequency in TagCrowd may be due to the fact that we have selected only multi authored papers. As different authors contribute to different areas, maybe the speaker spoke about the area where he/she is more comfortable about.

- The trend in cosine similarity clearly shows a strong correlation between paper and video while a loose correlation between slide and video and slide and paper. This can be explained by the fact that many times the person preparing the slide is not one of the author. Hence such a relationship.

We have also to take into consideration the following facts which we have assumed:

- 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.

- In case of multi-authored paper, the slide and the conference speech of one is technically not the same as the other author's, hence we just considered one author for analysis instead of comparing two authors.

- The parameters for TagCrowd was set taking in consideration that more than 50 as maximum value makes the analysis tough due to garbage data collected, and less than 50, does not provide enough matches for comparison. The minimum frequency was set to 2, considering that slides have very few words and more than 2 would make us miss many important words, and we are searching for frequently used words, hence setting to one wouldn't help.

6. CONCLUSIONS

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 Crowd Sourcing 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 analyse the trend in single authored paper for understanding their trend, or maybe we might analyse 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.

References

[1] *GenomeRing: Alignment visualization based on SuperGenome Coordinates* by A. Herbig, G. Jager, F. Battke, K. Nieselt. Vol. 28 ISMB 2012, pages i7–i15 doi:10.1093/bioinformatics/bts217

[2] *Fast Alignment of fragmentation tree* by F. Hufsky, K. Duhrkop, F. Rasche, M. Chimani, S. Bocker. Vol. 28 ISMB 2012, pages i265–i273 doi:10.1093/bioinformatics/bts207

[3] *Towards 3D structure prediction of large RNA molecules: An integer programming framework to insert local 3D motifs in RNA secondary structure* by Vladimir Reinharz, François Major and Jérôme Waldispühl. Vol. 28 ISMB 2012, pages i207–i214 doi:10.1093/bioinformatics/bts22

[4] *Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs* by Seunghak Lee and Eric P. Xing. Vol. 28 ISMB 2012, pages i137–i146 doi:10.1093/bioinformatics/bts22

[5] *Incorporating prior information into association studies* by Gregory Darnell, Dat Duong, Buhm Han and Eleazar Eski. Vol. 28 ISMB 2012, pages i147–i153 doi:10.1093/bioinformatics/bts235

[6] *Matching experiments across species using expression values and textual information* by Aaron Wise, Zoltán N. Oltvai and Ziv Bar-Josep. Vol. 28 ISMB 2012, pages i258–i264 doi:10.1093/bioinformatics/bts205

[7] *DACTAL: divide-and-conquer trees (almost) without* alignments by Serita Nelesen, Kevin Liu, Li-San Wang, C. Randal Linder and Tandy Warnow. Vol. 28 ISMB 2012, pages i274–i282 doi:10.1093/bioinformatics/bts218

[8] *Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss* by Mukul S. Bansal, Eric J. Alm and Manolis Kellis. Vol. 28 ISMB 2012, pages i283–i291 doi:10.1093/bioinformatics/bts225

[9] *MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins* by Fatemeh Miri Disfani, Wei-Lun Hsu, Marcin J. Mizianty, Christopher J. Oldfield, Bin Xue, A. Keith Dunker, Vladimir N. Uversky, and Lukasz Kurgan. Vol. 28 ISMB 2012, pages i75–i83 doi:10.1093/bioinformatics/bts209

[10] *A single source k-shortest paths algorithm to infer regulatory pathways in a gene network* by Yu-Keng Shih and Srinivasan Parthasarathy. Vol. 28 ISMB 2012, pages i49–i58 doi:10.1093/bioinformatics/bts212

[11] Topic title. **`http://www.iiests.ac.in`**, 2016.