UNIVERSITY OF CALIFORNIA

Los Angeles

Contributions to Structured and Unstructured Data Analysis: Liquid Association Computation Acceleration and Word Similarity via Folksonomy

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

GuanI Wu

2019

ABSTRACT OF THE DISSERTATION

Contributions to Structured and Unstructured Data Analysis: Liquid Association Computation Acceleration and Word Similarity via Folksonomy

by

GuanI Wu
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2019
Professor Ker-Chau Li, Chair

In this thesis, I organize two independent projects into five chapters. The first chapter introduces Liquid Association and our proposed method to accelerate its computation. The second chapter is related to the design of the computational structure for Liquid Association website (LAP3). The third chapter is regarding the application of Liquid Association to Global Health Observatory (GHO) data. The fourth chapter describes a novel method to model the distribution of human ratings on word-similarity. The last chapter focuses on the analysis of the relationship between the knowledge-based approach and the corpus-based approach.

The dissertation of GuanI Wu is approved.

Hsian-Rong Tseng

Yingnian Wu

Qing Zhou

Ker-Chau Li, Committee Chair

University of California, Los Angeles

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Ker-Chau Li, for giving me invaluable guidance and support of my Ph.D. study and research. I would not be able to thank him enough for giving me a chance when most people would have looked away. His encouragement and teaching not only broadened my horizon but also uplifted and inspired me in many aspects of my life. I would also like to thank Robert Yuan. His friendship and guidance helped me grow into my potential. Additionally, I would like to thank the members of my committee, Qing Zhou, Yingnian Wu, and Hsian-Rong Tseng, who gave me valuable feedback and suggestions.

I would also like to thank a wonderful group of people in the Statistics Department. Nicolas Christou, who has given me endless support of study and guidance in teaching since the moment I was his TA. Akram Almohalwas, who has provided me many opportunities to work with him and learn from him. Juana Sanchez, who gave me valuable advice on teaching. I appreciate all the professors who taught the courses I took. In addition, I would like to thank Glenda Jones, who gave me limitless support since day one in my Ph.D. program. I can't thank more for what I have taken away as a student and as a TA in this amazing department from those years. Thank all of my friends and colleagues, especially, Levon, Medha, Min, and Hao, who have shared their Ph.D. journey with me.

Finally, I would like to express my immense appreciation to my family. Thank you, Julie, for letting me feel at home here. Thank you, Ruth, for being a wonderful wife all the time. Thank you, my sister, Wen-Chen, for always standing by my side and believing in me. Thank you, my mom, for raising me and supporting me all the way through a tough time. Thank you all. This one is for you.

VITA

2005	B.S. (Information Management), Kun Shan University, Tainan City, Taiwan
2007	M.S. (Information Management), Chang Gung University, Taoyuan City, Taiwan
2007-2013	Research Assistant, Institute of Statistical Science, Academia Sinica
2014-2019	Teaching Assistant, Reader and Graduate Student Researcher UCLA Department of Statistics
2014	C.Phil., Statistics, UCLA, Los Angeles, California
2017	Teaching Assistant of the Year UCLA Department of Statistics

PUBLICATIONS AND PRESENTATIONS

Hsuan-Yu Chen, Sung-Liang Yu, Bing-Ching Ho, Kang-Yi Su, Yi-Chiung Hsu, Chi-Sheng Chang, Yu-Cheng Li, Shi-Yi Yang, Pin-Yen Hsu, Hao Ho, Ya-Hsuan Chang, Chih-Yi Chen, Hwai-I Yang, Chung-Ping Hsu, Tsung-Ying Yang, Kun-Chieh Chen, Kuo-Hsuan Hsu, Jeng-Sen Tseng, Jiun-Yi Hsia, Cheng-Yen Chuang, Shinsheng Yuan, Mei-Hsuan Lee, Chia-Hsin Liu, **GuanI Wu**, Chao A. Hsiung, Yuh-Min Chen, Chih-Liang Wang, Ming-Shyan Huang, Chong-Jen Yu, Kuan-Yu Chen, Ying-Huang Tsai, Wu-Chou Su, Huei-Wen Chen, Jeremy J.W. Chen, Chien-Jen Chen, Gee-Chen Chang, Pan-Chyr Yang and Ker-Chau Li. R331W missense mutation of oncogene YAP1 is a germline risk allele for lung adenocarcinoma with medical actionability. Journal of Clinical Oncology, 33(20):2303-2310, 2015.

Pei-Ing Hwang, Huan-Bin Wu, Chin-Di Wang, Bai-Ling Lin, Cheng-Tao Chen, Shinsheng Yuan, GuanI Wu, and Ker-Chau Li. Tissue-specific gene expression templates for accurate molecular characterization of the normal physiological states of multiple human tissues with implication in development and cancer studies. BMC Genomics, 12:439–439, 2011.

Shinsheng Yuan, Sung-Liang Yu, Hsuan-Yu Chen, Yi-Chiung Hsu, Kang-Yi Su, Huei-Wen Chen, Chih-Yi Chen, Chong-Jen Yu, Jin-Yuan Shih, Yih-Leong Chang, Chiou-Ling Cheng, Chung-Ping Hsu, Jiun-Yi Hsia, Chien-Yu Lin, **GuanI Wu**, Chia-Hsin Liu, Chin-Di Wang, Kang-Chung Yang, Yi-Wei Chen, Yi-Ling Lai, Chu-Chun Hsu, Tai-Ching Lin, Tsung-Ying Yang, Kun-Cheieh Chen, Kuo- Hsuan Hsu, Jeremy J.W. Chen, Gee-Chen Chang, Ker-Chau Li, and Pan-Chyr Yang. Clustered Genomic Alterations in Chromosome 7p Dictate Outcomes and Targeted Treatment Responses of Lung Adenocarcinoma With EGFR-Activating Mutations. Journal of Clinical Oncology, 29(25):3435–3442, September 2011.

Yi-Chiung Hsu, Hsuan-Yu Chen, Shinsheng Yuan, Sung-Liang Yu, Chia-Hung Lin, **GuanI Wu**, Pan-Chyr Yang, and Ker-Chau Li. Genome-wide analysis of three-way interplay among gene expression, cancer cell invasion and anti-cancer compound sensitivity. BMC Medicine, 11(1):106, 2013.

GuanI Wu, Shinsheng Yuan, Yu-Cheng Li, Yi-Chang Lu, Ker-Chau Li. GPU Accelerated Liquid Association, Statistics and Its Interface (In revision).

GuanI Wu, Ker-Chau Li. A folksonomy-based approach for profiling human perception on word similarity (In review).

GuanI Wu, Ker-Chau Li. A Noval Method to Estimate Human Judgment on Words Similarities, Joint Statistical Meetings, poster session, 2018.

CHAPTER 1

GPU Accelerated Liquid Association

1.1 Abstract

High throughput biological assays have provided numerous data sources for studying complex interactions between multiple variables in a biological system. Many computational tools for exploring the voluminous biological data are based on pair-wise correlation between variables. Liquid Association (LA) is a novel statistical concept for inferring higher order of association between variables in a system. While LA was originally introduced to study gene-gene interaction involving three genes at a time, it can be applied for correlating biological measurements with clinical variables such as drug sensitivity profiling and patient survival time. It is computationally expensive to compute LA scores for all possible triplets in very large datasets. Here we show how to take advantage of Graphic Processing Units (GPUs) for speeding up the LA computing. Our GPU-accelerated version of LA computation (GALA) achieved nearly 200-fold improvement over the traditional CPU-alone version. A companion package in R was developed for facilitating follow-up analysis and improving user experience.

1.2 Introduction

Correlation is a simple yet powerful concept in analyzing gene expression data. Two genes with positively correlated expression profiles are likely to be functionally associated and they may participate in the same or related biological process. However, functionally associated genes may not have correlation in expression. For instance, they may not be regulated at the transcription level and they have multiple functions. Co-expressed genes may become uncorrelated or even turn into contra-expressed when the underlying cellular state changes. Liquid association (LA), as opposed to "steady" association, is designed to quantify the size and the direction of the change of correlation between two genes. LA describes the ternary relationship between variables in a system [Li02, LY04, LPY07, WSY08, SYL08, TWY10]. In gene expression study, the total computing complexity of LA is O(n^3) where n is the number of the genes. For integrated studies, it is time-consuming to compute all possible combinations from whole genome gene expression, SNP, or copy number variation data. [Li02]

To mitigate this problem, we developed a program via Compute Unified Device Architecture (CUDA) language for Graphic Processing Unit (GPU) platforms to accelerate the performance of LA score computation. A 200 times speed-up over the CPU version was obtained. A companion R package was also developed. The users can use it for visualizing the correlation changes and for conducting further analyses.

1.3 Liquid Association

In the context of gene expression, LA conceptualizes the mediation of the change in the co-expression pattern of two genes (X, Y) by a third gene Z. A positive LA score indicates that the correlation between gene X and gene Y is likely to change from being negative to positive. Conversely, a negative LA score indicates the change from positive to negative correlation. The standard procedure to obtain LA score $LA(X, Y|Z)$ requires two steps [Li02]:

1. Normal score transformation. To standardize each gene-expression profile with normal score transformation, the m values in the profile are compared with each other and their ranks $R_1, ..., R_m$ are recorded. The ranks are then used to obtain the transformed profile, $\Phi^{-1}(R_1/(m+1)), \Phi^{-1}(R_2/(m+1)), ..., \Phi^{-1}(R_m/(m+1))$, where $\Phi(.)$ is the cumulative normal distribution. Let X', Y', Z' denote the transformed profiles.

2. LA score computation. Compute the average product of the three transformed profiles,

$(X'_1Y'_1Z'_1 + ... + X'_mY'_mZ'_m)/m$. This gives the LA score $LA(X, Y|Z)$.

It is computer intensive to obtain LA scores because the number of combinations in choosing three from N genes or probes under study grows rapidly as N increases. It is typical for N to exceed 50K in commercial human gene expression chips and the number gets 10 times higher in SNP, DNA copy number, or methylation arrays. To improve user experience, we also compare the computed LA scores and save the top positive LA scores and bottom negative LA scores. This helps speed up the response time for on-line queries.

1.4 GPU Accelerated Liquid Association

GPUs were first introduced to accelerate computing speeds in computer graphics. General Purpose Computing on GPU (GPGPU) is a technique of using GPUs, which generally requires a set of stream processors and a hierarchical memory structure, to execute the computing tasks in parallel. We chose the popular CUDA language for reprograming the LA computation. The speeds of GALA running on two different GPUs will be compared to the C version running on the CPU machine in this article.

Since GPU executes in SIMT (Single Instruction Multiple Thread) mode, we must design the instruction set for each thread, the GPU kernel function, to perform LA computation for the three normal-score transformed profiles. In general, an optimized GPU kernel function consists of several steps such as utilization of shared memory for computation, effective usage of global memory bandwidth, efficient coordination of multiple threads. Our kernel function was constructed with these performance considerations.

Shared memory is the key to the reduction of global memory traffic. In order to fully utilize the shared memory, GALA partitions data into subsets so that each subset matches the size of shared memory. Coordinated by the GPU scheduler, the GPU processing elements execute a fixed number of the threads at a time and within the grouped threads, *warp*, the executed instructions must be the same at any time point. Because the size of the warp is limited, we constructed our GPU kernel function to tailor the dimensions of the matrices of the three transformed profiles declared in the shared memory. As GPU transfers data by moving one block of consecutive memory bits at a time, our input data are arranged with the memory coalescing technique to minimize the transfer counts. The GPU scheduler also determines when and which warp to be executed or placed on hold. A barrier synchronization function is employed to coordinate the parallel activities of multiple warps, thus enabling the more efficient parallel execution of threads (Algorithm 1).

Algorithm 1: The kernel function of GALA

```
kernelOfGALA $(X, Y, Z)$
    inputs : $X$ and $Y \in \Re^{k\times m}$, $Z \in \Re^{v\times m}$
    output: $LA(X, Y, Z) \in \Re^{k\times k\times v}$
    foreach $t \in v$ do
        foreach $i \in m$ by $Block\_Size$ do
            $\_\_shared\_\_\ x_i \leftarrow X[Block\_Size][Block\_Size]$;
            $\_\_shared\_\_\ y_i \leftarrow Y[Block\_Size][Block\_Size]$;
            $\_\_shared\_\_\ z_{t,i} \leftarrow Z_t[Block\_Size]$;
            $\_\_syncthreads()$;
            $LA(X, Y|Z_t) \leftarrow LA(X, Y|Z_t) + LA(x_i, y_i|z_{t,i})$;
            $\_\_syncthreads()$;
    return $LA(X, Y, Z)$;
```

Table 1.1: Parameters of the preference file, and their description

PARAMETERS	DESCRIPTION
`NUMBER_COLUMN`	Total number of columns in a dataset.
`DATATYPE_{X,Y,Z}`	0: Not normalized, 1: Normalized.
`OFFSET_COL_{X,Y,Z}`	Data starts with which column.
`GROUP_INDEX_{X,Y,Z}`	Fulfill 1 in correspondence with the number of column.
`DATAFILE_{X,Y,Z}`	A file path for {X,Y,Z}.
`MYSQL_SQL_{X,Y,Z}_0`	A SQL query for {X,Y,Z}.
`NUMBER_ROW_{X,Y,Z}`	Total number of rows in a dataset
`COMP_MODE`	0: Keep all LA triplets, 1: Remove duplicate triplets.
`NUMBER_PAIRS`	How many triplets in both top and bottom will be kept.
`OUTPUT_FILE_TAG`	A string for the file name and path of output.
`LENGTH_RESTRICTION`	An integer and indicates that at least `LENGTH_RESTRICTION` of values a row must contain for computing a LA score.

GALA allows users to prepare inputs (X, Y, Z) with either flat files or SQL commands to request data from MySQL database. Since maximum number of columns in MySQL is generally far less then number of variables in a dataset, we take each row as a variable, and require every row to have the same number of entries. Meanwhile, users are required to prepare a preference file with computational arguments such as how many of top positive/negative LA scores will be saved, locations of input and output files, etc. Table 1.1 gives a list of parameters required in the preference file. There are two options to calculate LA scores. If X, Y, Z are prepared by flat files, one can execute

```
>./gala tmp_foo_pref.txt
```

Otherwise, input data is requested from a database, one can execute

```
>./gala tmp_foo_pref.txt [IP address] [username] [password]
```

Initially, GALA dynamically declares the feasible number of threads according to the size of input. When the input is too large to be computed, GALA will split the input into smaller pieces so that each of them fits in the allowable number of threads for the kernel function. In addition, if the input size is too small, GALA will launch the kernel function with an adjusted number of threads to prevent the kernel function from running the extra threads. The output of the kernel function is an array identifier and the LA scores with allocated consecutively in the global memory. Once the kernel function was executed, GALA will perform a modified version of Quick Sort. This sorting function is used to sort the outputs from the kernel function and to filter LA scores according to the parameters of the preference file. Iterations between the kernel function and the sorting function will be continued until all LA scores are computed(Figure 1.1).

Users can install GALA on any GPU-equipped computers with `make` command. However, apart from CUDA library, it also requires users to install `libmysqlclient` before the installation as GALA allows users to retrieve data from a MySQL database. In the package, we also provide tmp_yzfiles_pref.txt and tmp_who_fml_pref.txt to demonstrate how to prepare the preference files for GALA. Finally, GALA generates two output files, `OUTPUT_FILE_TAG_`TOP.txt and `OUTPUT_FILE_TAG_`BOT.txt, of which every triplet is saved in the form of {Index of Z, Index of Y, Index of X, LA score}, and all of triplets are sorted in order.

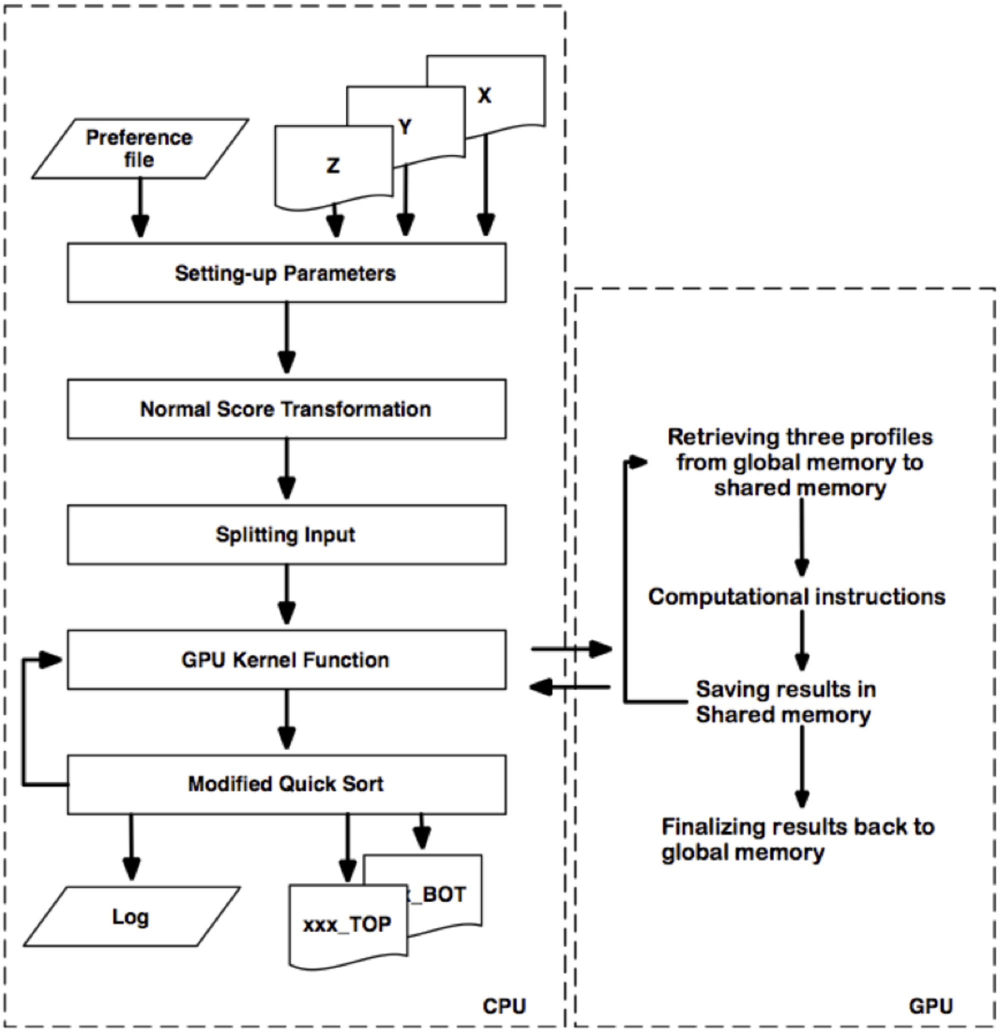

Preference file
X
Y
Z
Setting-up Parameters
Normal Score Transformation
Splitting Input
GPU Kernel Function
Modified Quick Sort
Log
xxx_TOP
_BOT
CPU
Retrieving three profiles from global memory to shared memory
Computational instructions
Saving results in Shared memory
Finalizing results back to global memory
GPU

Figure 1.1: The flowchart of GALA. The normal score transformation and sorting of computed LA scores are performed by CPU as shown on the left panel. Computation of LA scores, the most time-consuming part, is executed by GPU as shown on the right panel.

1.4.1 Performance

We demonstrated the improvement of GALA over the original LA program with eight public available gene expression datasets as Table 1.2 shows. We used two different types of GPU cards to implement GALA, Tesla M2050 which contains 448 sets of 1.3 GHz processors with 6 GB dedicated memory and Tesla M2090 which contains 512 sets of 1.3 GHz processors with 6 GB dedicated memory. On the other hand, the CPU version of LA is performed on an Intel Core i7 965 model with the clock-speed at 3.2 GHz and 6 GB main memory. Since the loading ratio between the LA-score computation and LA-score sorting was around

Table 1.2: Eight Gene Expression Datasets

ID	Sources
S1, S2	NCI-60 cancer cell line. [SVK09]
S3	Lung adenocarcinoma. [STE08]
S4	High-grade lung neuroendocrine tumors of the lung (GSE1037). [JVH04]
S5	Various human and mouse tissues (GSE1122). [SWB04]
S6	Frozen tissue of primary lung tumors (GSE3141). [BYC05]
S7	Normal human tissues from selected samples (GSE7307). [Rot07]

Table 1.3: LA Performance Comparison. The column, Complexity, is defined as the number of conditions multiplied by the square of the number of genes in log scale.

Dataset	M2090 (sec.)	M2050 (sec.)	CPU (sec.)	Complexity (log)	Subjects	Genes
S1	0.66	0.79	31	9.75	60	9,706
S2	1.24	1.42	93.01	9.97	59	12,625
S3	8.5	10.5	1049	10.95	179	22,215
S4	14.3	17.52	1774	11.17	91	40,368
S5	13.74	16.33	1566.11	11.21	143	33,689
S6	25.29	28.57	2182.37	11.51	111	54,683
S7	70.61	89.59	13695.81	12.15	473	54,675

10:1, the speed comparison for GALA would be focused on the LA-score computation only. We used the most time-demanding on-line query, i.e. finding the top LA scores of $(X, Y|Z)$ over all possible pairs of (X, Y) from an input of Z, as the submitted job and recorded the elapsed time of computing in each of the aforementioned test datasets. In addition, the elapsed time also involved the data transportation between the main memory and the global memory. In Table 1.3, the time listed under Tesla M2050 and Tesla M2090 is the elapsed time for GPU kernel function. For fair comparison, the column under CPU, only recorded the time on computing LA scores. We found that GALA outperformed CPU version and the improvement generally ranged from 40-fold to 190-fold. Moreover, the result shows that our implementation takes full advantage of GPU card upgrade. Compared to Tesla M2050, Tesla M2090 has 64 more computational cores and 17% higher memory bandwidth. Our implementation had better performance on Tesla M2090 than that on Tesla M2050 with a 17% speedup in average. In Figure 1.2, the strong linear relationship was also observed

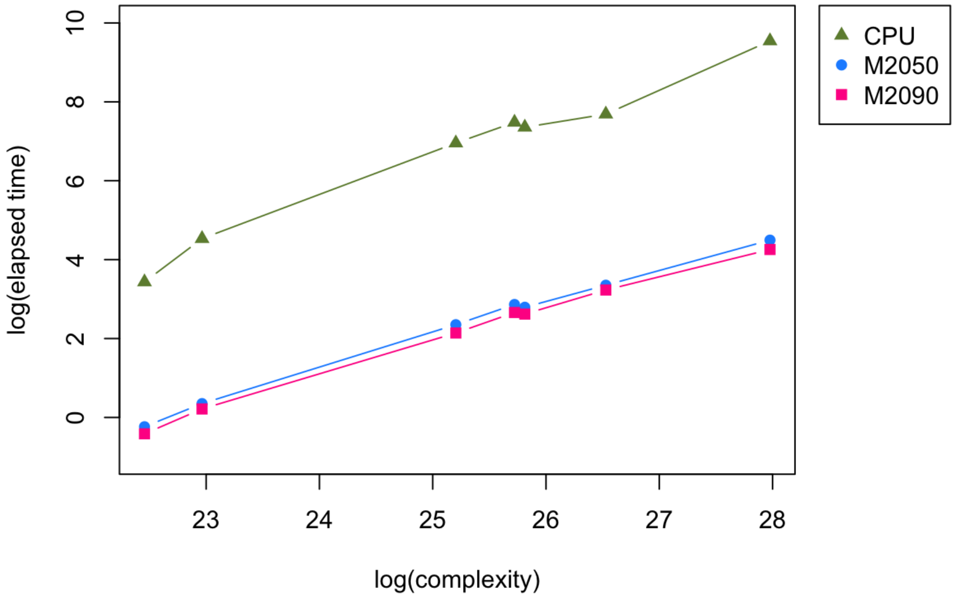


Figure 1.2: Complexity versus Elapsed Time. The x-axis is the log(Complexity) and y-axis is the log(Elapsed Time) in log scale.

between elapsed time and complexity. The relationship signals that GALA have the same performance regardless of the complexity of data.

1.4.2 LA Package in R

For encouraging the routine use of LA analysis, we also developed `la` package in R to calculate LA scores and draw LA plots for further inspection of correlation patterns. We may select one triplet from the outcomes of GALA, and employs `la` to exam the relationship among three variables. The package contains `la` function and a dataset for the demonstration of LA. `drawla` has the following arguments:

```
drawla(x, y, z, ename, xyzLabels, switch = 2, ...)
```

Three vectors `X`, `Y`, and `Z` are taken as input variables, and the order is also arranged as $LA(X, Y|Z)$. We can change the order of three vectors to observe the changes of LA plots such as $LA(Y, Z|X)$ or $LA(X, Z|Y)$. Detail description regarding arguments is listed at `http://mib.stat.sinica.edu.tw/MIB/downloads.php`. `drawla` aids the visualization of correlation between X and Y given different status of Z, where Z are split into three *status* (low, median, high). Cut points used to split Z were optimized by Algorithm 2, which maximizes log-likelihood function $l(\mu, \sigma^2; X^*, Y^*)$

$$RSS = \sum_{i=1}^{cut_1}(Y_i^* - \hat{\alpha_0} - \hat{\alpha} X_i^*)^2 + \sum_{i=cut_1+1}^{cut_2-1}(Y_i^* - \hat{\beta_0})^2 + \sum_{i=cut_2}^{n}(Y_i^* - \hat{\gamma_0} - \hat{\gamma} X_i^*)^2 \tag{1.1}$$

$$\sum_{i=1}^{cut_1}(Y_i^* - \hat{\alpha_0} - \hat{\alpha} X_i^*)^2 = Var(Y_{1:cut_1}^*)(1 - Corr(Y_{1:cut_1}^*, X_{1:cut_1}^*)) \tag{1.2}$$

$$\sum_{i=cut_1+1}^{cut_2-1}(Y_i^* - \hat{\beta_0})^2 = \sum_{i=cut_1+1}^{cut_2-1}(Y_i^* - \bar{Y}_{cut_1+1:cut_2-1})^2 \tag{1.3}$$

$$\sum_{i=cut_2}^{n}(Y_i^* - \hat{\gamma_0} - \hat{\gamma} X_i^*)^2 = Var(Y_{cut_2:n}^*)(1 - Corr(Y_{cut_2:n}^*, X_{cut_2:n}^*)) \tag{1.4}$$

,where (X^*, Y^*) denotes (X, Y) sorted by Z.

Algorithm 2: Finding cut points of LA

```
findCutsOfLA ($X, Y, Z$)
    inputs : $X, Y, Z \in \Re^{1 \times m}$
    output: $cut_1, cut_2 \in \Re^{1 \times 1}$
    Sort $\{X, Y, Z\}$ by $Z$
    foreach Try $cut_1 \in \{1, cut_2\}$ do
        $b \leftarrow cov(Y_{1:cut_1}, X_{1:cut_1})/\sigma^2_{X_{1:cut_1}}$;
        $a \leftarrow \bar{Y}_{1:cut_1} - b\bar{X}_{1:cut_1}$;
        $RSS_1 \leftarrow (Y_{1:cut_1} - a - bX_{1:cut_1})^2$;
        foreach $cut_2 \in \{cut_1 + 1, n\}$ do
            $RSS_2 \leftarrow (Y_{cut_1+1:cut_2-1} - \bar{Y}_{cut_1+1:cut_2-1})^2$;
            $c \leftarrow cov(Y_{cut_2:n}, X_{cut_2:n})/\sigma^2_{X_{cut_2:n}}$;
            $d \leftarrow \bar{Y}_{cut_2:n} - b\bar{X}_{cut_2:n}$;
            $RSS_3 \leftarrow (Y_{cut_2:n} - c - dX_{cut_2:n})^2$;
            $RSS \leftarrow RSS_1 + RSS_2 + RSS_3$;
            $l \leftarrow -\frac{m}{2}\log(2\pi RSS) + \frac{1}{2}(m-1)$;
            If $Max(l)$ return $cut_1, cut_2$;
```

1.5 Conclusion

In this chapter, we demonstrate a hybrid CPU/GPU program to obtain LA scores. The input data were arranged in a certain order for the efficient access from GPUs, and the configuration took advantage of multiple cores of GPUs to speed up the LA scores computation. We recorded the elapsed time in testing seven real datasets and compared GALA with the original LA program. GALA was much faster at execution speed regardless of the complexity of data. The use of the companion R code for visualizing the dynamic change of association between variable is illustrated. Our package can be widely applied in analyzing complex data from various scientific areas.

CHAPTER 2

Computation Structure for Liquid Association Website

2.1 Abstract

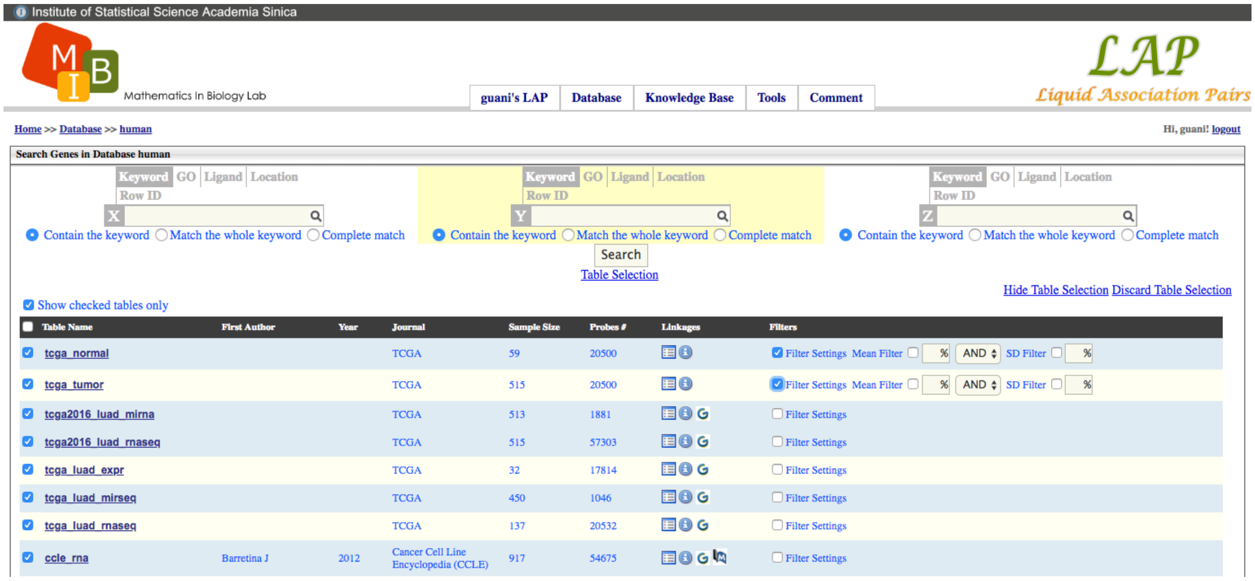
To facilitate the online analysis of gene expression data, a primitive website, LAP, was created [Yua03] long ago. As the scale of data and demands of a variety of analysis strategies are growing rapidly, newer versions of LA website with better back-end configuration are developed by the team. Our lab, Mathematics in Biology (MIB) at Institute of Statistical Science, Academia Sinica. This section introduces the essential components of the version LAP3, focusing on my contribution to back-end program and hardware configuration.

2.2 Introduction

LAP was originally developed by for facilitating the use of Liquid Association on gene expression data. However, due to the large scale of data and demands in applying various analysis strategies to LA outcomes, a new website design is required. The next generation of LAP website, LAP3, aims to provide various functions to improve the user experience. It is a collaborative project. This section introduces an overview of the system, and focusing on my contributions in the design of the core LA computation. The LAP website configuration is composed of three main components, Database, User Interface, and Computation. Each component contains multiple programming objects that provide functions and communicate across components.

Starting from the users' end, the operation flow of LAP3 follows the original design of the LAP. Users first select datasets of interest and then do the keyword search for inputs of

X, Y, and Z. (Figure 2.1). Except for X, the input boxes of Y and Z can be blank, which indicates that all of the genes from the selected databases are input genes. To find LA pairs, give scouting-genes to input box of X and leave Y and Z empty. To find LA scouting-genes, give gene pairs to input boxes of X and Y. In consist with the original paper [Li02], we will use Z to denote scouting-gene, and (X, Y) to denote LA pair for the rest of the sections. The location of the LAP is http://mib.stat.sinica.edu.tw/LAP/.



X: ● Contain the keyword ○ Match the whole keyword ○ Complete match

Y: ● Contain the keyword ○ Match the whole keyword ○ Complete match

Z: ● Contain the keyword ○ Match the whole keyword ○ Complete match

☑ Show checked tables only

☐ Table Name	First Author	Year	Journal	Sample Size	Probes #	Linkages	Filters
☑ tcga_normal			TCGA	59	20500	▤ ⓘ	☑ Filter Settings Mean Filter ☐ % AND SD Filter ☐ %
☑ tcga_tumor			TCGA	515	20500	▤ ⓘ	☑ Filter Settings Mean Filter ☐ % AND SD Filter ☐ %
☑ tcga2016_luad_mirna			TCGA	513	1881	▤ ⓘ G	☐ Filter Settings
☑ tcga2016_luad_rnaseq			TCGA	515	57303	▤ ⓘ G	☐ Filter Settings
☑ tcga_luad_expr			TCGA	32	17814	▤ ⓘ G	☐ Filter Settings
☑ tcga_luad_mirseq			TCGA	450	1046	▤ ⓘ G	☐ Filter Settings
☑ tcga_luad_rnaseq			TCGA	137	20532	▤ ⓘ G	☐ Filter Settings
☑ ccle_rna	Barretina J	2012	Cancer Cell Line Encyclopedia (CCLE)	917	54675	▤ ⓘ G	☐ Filter Settings

Figure 2.1: The portal of LAP3. Users are allowed to select multiple datasets for obtaining LA scores. We precomputed descriptive statistics so that users can select the subsets of selected datasets by applying their means or standard deviations as filters.

LAP3 has made major changes in the design of relational database schema. To store the descriptions of collected data and integrate with User Interface, we organize the information under `data_name` as five relational data tables: `data_name`_INFO, `data_name`_DESC, `data_name`_EXP, `data_name`_INDEX, and `data_name`_DATA. The purpose of `data_name`_DESC is to keep the information such as authors, data source, number of rows, etc. `data_name`_INDEX stores descriptive statistics. The rest of the data tables play the same function as the original version in [Yua03].

The computation component of LAP3 is also redesigned. Since this task was crosses the front-end design and the back-end design, I first present a sketch of the website configuration of LAP3, and then specify the communication mechanism across different component of the system. Finally, I focus on how we utilize and summarize precomputed LA results.

2.3 Configuration of LAP3

The entire system is composed of several servers as shown in Figure 2.2. We set up two web servers with one equipped with Apache `mod_proxy_balancer` to automatically assign web sessions to end users. Each web server directly connects to one standalone database for faster data access. As for computation, we set up a computer cluster and a head node, **Job Coordinator**, installed with `Oracle Grid Engine` to distribute computation tasks. That is, Job Coordinator plays a role to transform requests from web servers to executable commands for the computer cluster. Therefore, we made Job Coordinator run an internal HTTP server with developed computer programs to communicate with Web Server A and Web Server B. When Web Server A or Web Server B receives computation requests, a series of programming objects in PHP at Job Coordinator was called to process the communication across database, computer cluster, and web servers.

2.3.1 Communication Mechanism of Front-End and Back-End

We adopt Object-Oriented Programming (OOP) to develop all of the communication functions from the presentation layer to all data processing units—including database access and computation. Unlike the original version of LAP, the users are allowed to select multiple gene expression data at the same time with the same inputs of X, Y and Z to compute LA scores and Pearson correlation coefficients. To achieve this aim, we created a series of objects in PHP working with JavaScript.

One of primary objects (*package*) is created to carry a set of input instructions with the functions that translate user requests to executable commands for the computer programs at the back-end. If one submits a request with multiple gene expression data, the system will create an array of *package*, and pack them into a *carrier*, the object in charge of communication between a web server and Job Coordinator. Figure 2.3 presents the hierarchy relationship of *carrier* and *package*. Upon receiving packages from the front-end, the Job

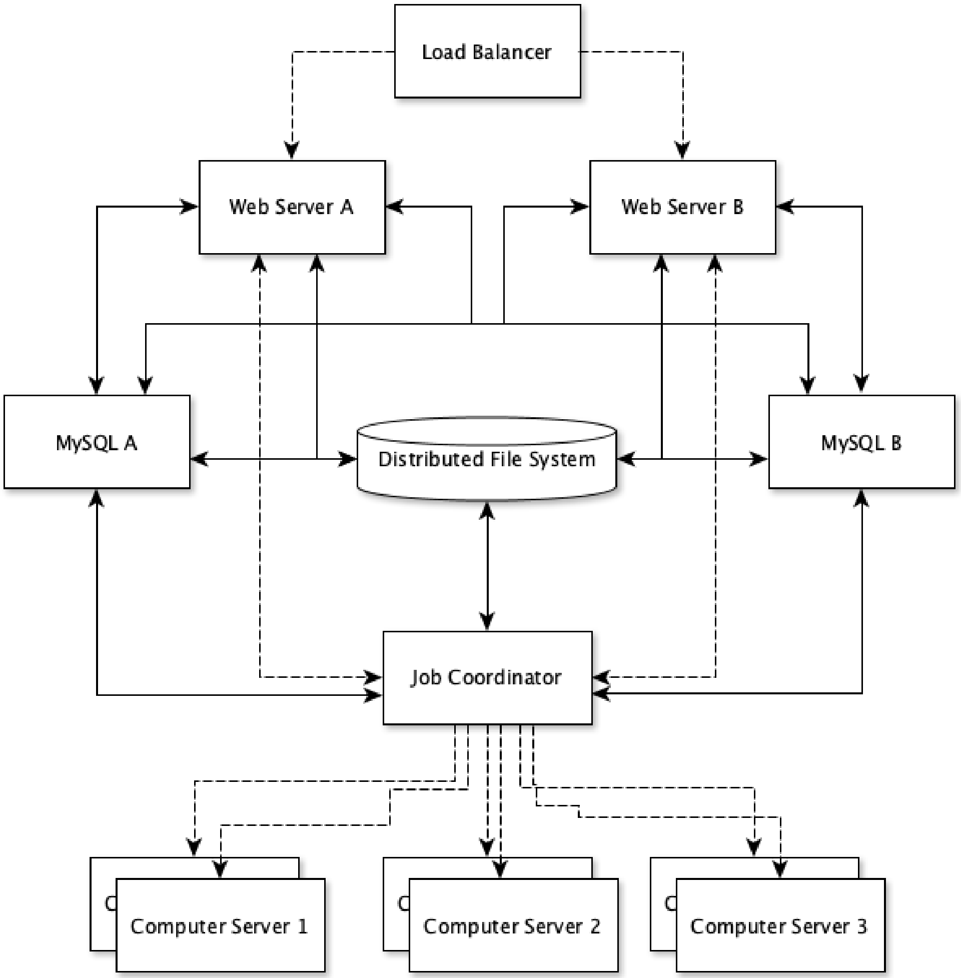


Figure 2.2: The structure of LAP3. The dashed lines are the communication signals. The solid lines indicate data transmission.

Coordinator forwarded them to the back-end servers. After computing, a corresponding objects are generated for shipping the output back to the front-end. For instance, gene symbols are now the outputs of *gene* object. LA scores and plots are generated by *LA* that inherited from a general computation object we developed. In short, the design of *package* contains computation objects and data objects, and provide the related functions.

2.3.2 Computation of LAP3

To coordinate various objects at web servers, we run an internal HTTP server at Job Coordinator with developed PHP objects and SHELL scripts. For example, for activating GALA

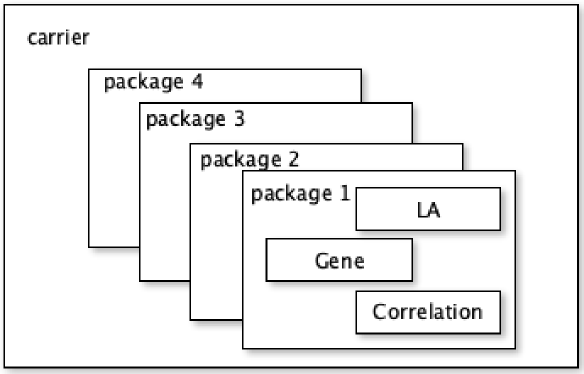


Figure 2.3: Communication Objects of LAP3.

computing, Job Coordinator can automatically generate preference files as specified in Table 1.1 and executable commands for performing the computation. As soon as the preference files are ready, Job Coordinator will submit the computation task to the computer cluster. During the computation, GALA generates a log file for tracking purpose. Once the output is ready, Job Coordinator will pass it to PHP objects at the web servers. Here each computational outcome is also organized as an object in PHP so that the UI components could easily access and display output in a systematical way.

2.4 Precomputed LA Scores

Due to GALA, we successfully reduce the elapsed time of LA computation and improve the user experience of the LAP website. Unlike the previous version, LAP3 can compute all possible combinations of gene-pairs for a dataset by giving a scouting gene. However, due to the rapidly growing data scale, to accelerate the response, we keep precomputed LA scores in a database. Moreover, to provide an overall picture of the precomputed LA scores, we propose two ways, dependent on the LA analysis strategies [Li02], to summarize precomputed LA scores.

Given a gene expression dataset, we iteratively take each gene profile as the scouting gene to compute LA scores with all possible pair combinations, and then keep 1,000 LA pairs with top/bottom LA scores in two separate files as shown in Figure 1.1 from Chapter 1

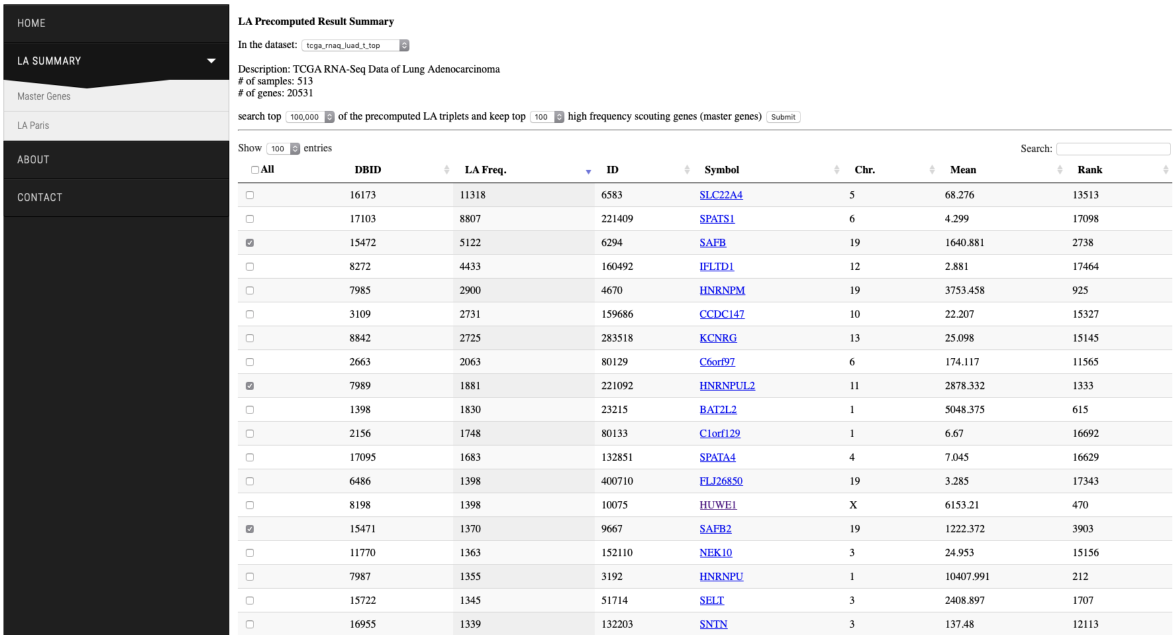
GPU Accelerated Liquid Association. We transform the two separate files into the format of MySQL, *data_name_top* and *data_name_bot*, where each data table has the number of gene profiles times 1,000 rows, and each row is composed of four columns (X, Y, Z, LA_Score). Through these precomputed results, LAP3 can accelerate the response when only a scouting gene is given.

2.4.1 Summary of Precomputed LA Triplets

We proposed **master genes** and **master gene-pairs** to summarize the precomputed LA scores. Here the master genes are the high-frequency scouting genes in the set of triplets with highest/lowest LA scores, such as 100,000 LA triplets with highest LA scores in *data_name_top*. The master gene-pairs are the high-frequency LA pairs in the set of triplets with highest/lowest LA scores. The developed website can be found at http://ws.stat.sinica.edu.tw/lax.

To demonstrate the functions, we downloaded and organized the gene expression data of lung adenocarcinoma (LUAD) with 20,531 gene profiles by 513 tumor samples, and the expression data of lung squamous cell carcinoma (LUSC) including 20,531 gene profiles by 501 tumor samples from GDC Data Portal (https://portal.gdc.cancer.gov). Through GALA, we generated two outputs, *tcga_rnaq_luad_t_top* and *tcga_luad_rnaq_t_bot* for the data of LUAD, and *tcga_rnaq_lusc_t_top* and *tcga_lusc_rnaq_t_bot* for the data of LUSC. Figure 2.4 shows the list of master genes ordered by the frequencies. One can also select the interesting genes to retrieve the corresponding LA pairs as a screenshot shown in Figure 2.5.

We collected 100 master genes from 100,000 triples with highest LA scores through *tcga_rnaq_lusc_t_top* and *tcga_rnaq_luad_t_top* respectively, and then found genes that both the lists of master genes have in common. As a result, 39 genes are found overlapping across both outcomes. Figure 2.6 and Figure 2.7 are Pearson correlation matrices of 39 genes with gene expression data of LUAD and LUSC. Interestingly, both correlation matrices show a similar pattern. Taking Figure 2.6 as an example, two groups of genes (top right and bottom left) have no linear relationship to each other. The group of genes at the top right of Figure 2.6 shows positive correlations within the group, but the group of 17 genes at the



LA Precomputed Result Summary

In the dataset: tcga_rnaq_luad_t_top

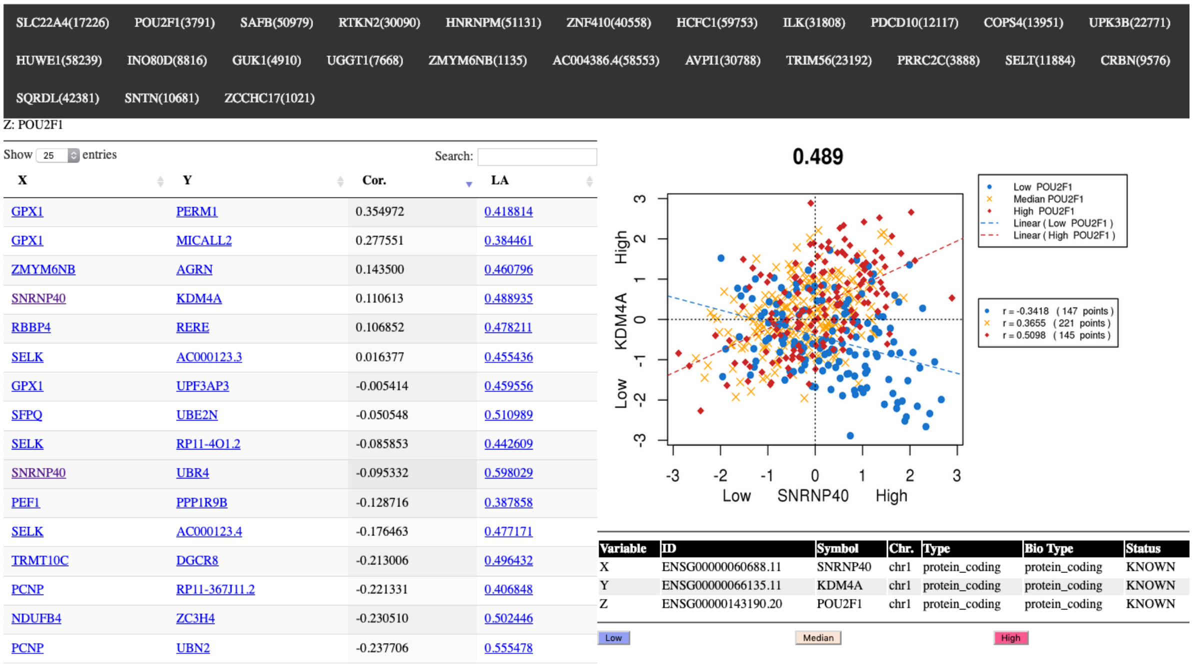
Description: TCGA RNA-Seq Data of Lung Adenocarcinoma
# of samples: 513
# of genes: 20531

search top 100,000 of the precomputed LA triplets and keep top 100 high frequency scouting genes (master genes) Submit

Show 100 entries Search:

□ All	DBID	LA Freq.	ID	Symbol	Chr.	Mean	Rank
□	16173	11318	6583	SLC22A4	5	68.276	13513
□	17103	8807	221409	SPATS1	6	4.299	17098
☑	15472	5122	6294	SAFB	19	1640.881	2738
□	8272	4433	160492	IFLTD1	12	2.881	17464
□	7985	2900	4670	HNRNPM	19	3753.458	925
□	3109	2731	159686	CCDC147	10	22.207	15327
□	8842	2725	283518	KCNRG	13	25.098	15145
□	2663	2063	80129	C6orf97	6	174.117	11565
☑	7989	1881	221092	HNRNPUL2	11	2878.332	1333
□	1398	1830	23215	BAT2L2	1	5048.375	615
□	2156	1748	80133	C1orf129	1	6.67	16692
□	17095	1683	132851	SPATA4	4	7.045	16629
□	6486	1398	400710	FLJ26850	19	3.285	17343
□	8198	1398	10075	HUWE1	X	6153.21	470
☑	15471	1370	9667	SAFB2	19	1222.372	3903
□	11770	1363	152110	NEK10	3	24.953	15156
□	7987	1355	3192	HNRNPU	1	10407.991	212
□	15722	1345	51714	SELT	3	2408.897	1707
□	16955	1339	132203	SNTN	3	137.48	12113

Figure 2.4: A screenshot of the list of master genes. This list is generated by utilizing *tcga_rnaq_luad_t_top*.



Z: POU2F1

Show 25 entries Search:

X	Y	Cor.	LA
GPX1	PERM1	0.354972	0.418814
GPX1	MICALL2	0.277551	0.384461
ZMYM6NB	AGRN	0.143500	0.460796
SNRNP40	KDM4A	0.110613	0.488935
RBBP4	RERE	0.106852	0.478211
SELK	AC000123.3	0.016377	0.455436
GPX1	UPF3AP3	-0.005414	0.459556
SFPQ	UBE2N	-0.050548	0.510989
SELK	RP11-4O1.2	-0.085853	0.442609
SNRNP40	UBR4	-0.095332	0.598029
PEF1	PPP1R9B	-0.128716	0.387858
SELK	AC000123.4	-0.176463	0.477171
TRMT10C	DGCR8	-0.213006	0.496432
PCNP	RP11-367J11.2	-0.221331	0.406848
NDUFB4	ZC3H4	-0.230510	0.502446
PCNP	UBN2	-0.237706	0.555478

Variable	ID	Symbol	Chr.	Type	Bio Type	Status
X	ENSG00000060688.11	SNRNP40	chr1	protein_coding	protein_coding	KNOWN
Y	ENSG00000066135.11	KDM4A	chr1	protein_coding	protein_coding	KNOWN
Z	ENSG00000143190.20	POU2F1	chr1	protein_coding	protein_coding	KNOWN

Figure 2.5: A screenshot of displaying LA pairs by selecting a master gene. The left column is the LA plot of $LA(SNRNP40, KDM4A|POU2F1)$.

bottom left of Figure 2.6 reveals a group of 5 genes and a group of 12 genes.

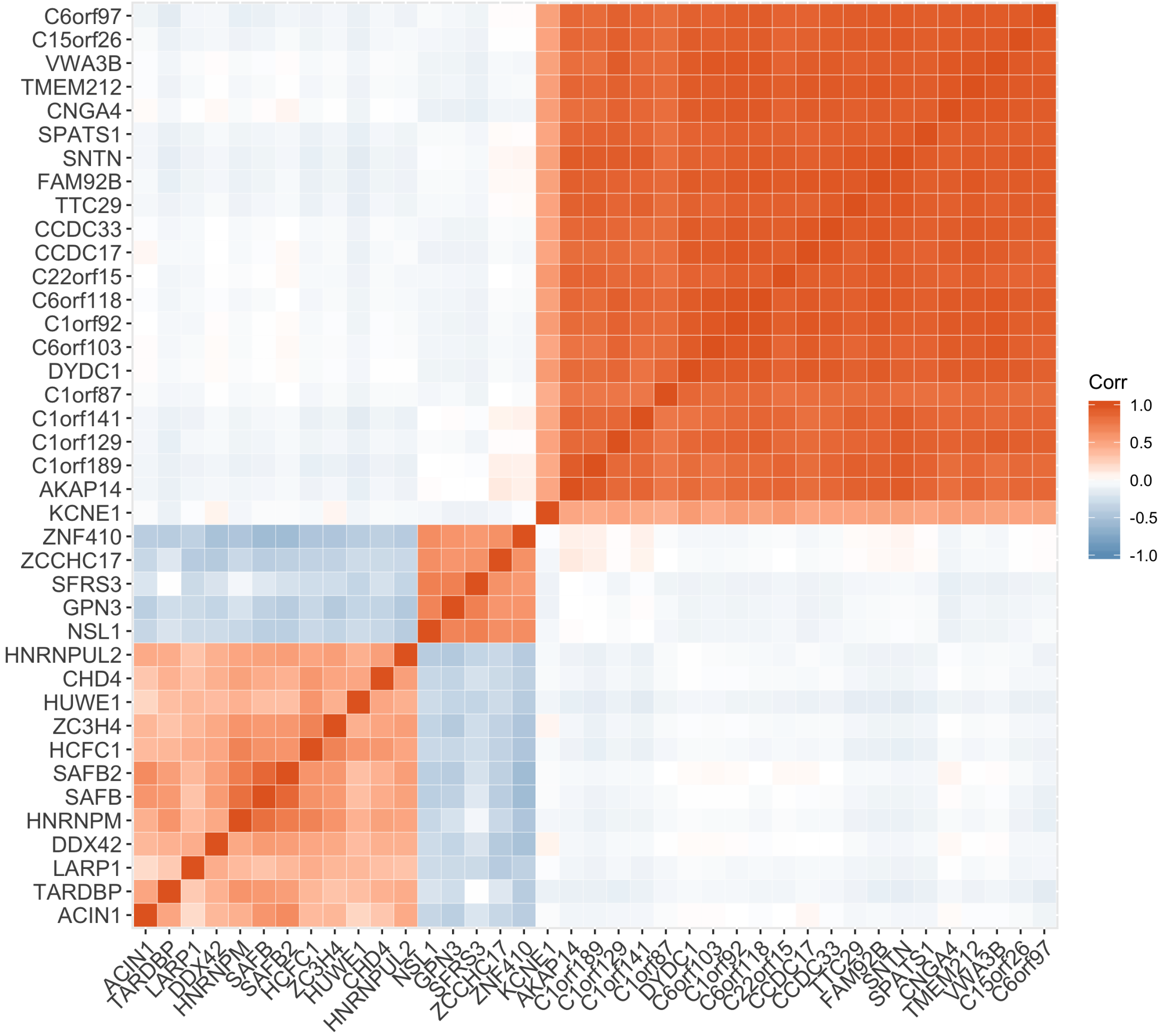


Figure 2.6: Pearson correlation matrix for the thirty-nine genes of LUSC.

Since the group of 17 genes at the bottom left of Figure 2.6 shows an interesting pattern, we further examined the pathway where these 17 genes link to via https://david.ncifcrf.gov/-summary.jsp. Table 2.1 presents 17 genes and their gene names. We found four genes linking to the spliceosome pathway (Figure 2.8), which consists of five small nuclear ribonucleoproteins and other factors to proceed with RNA splicing.

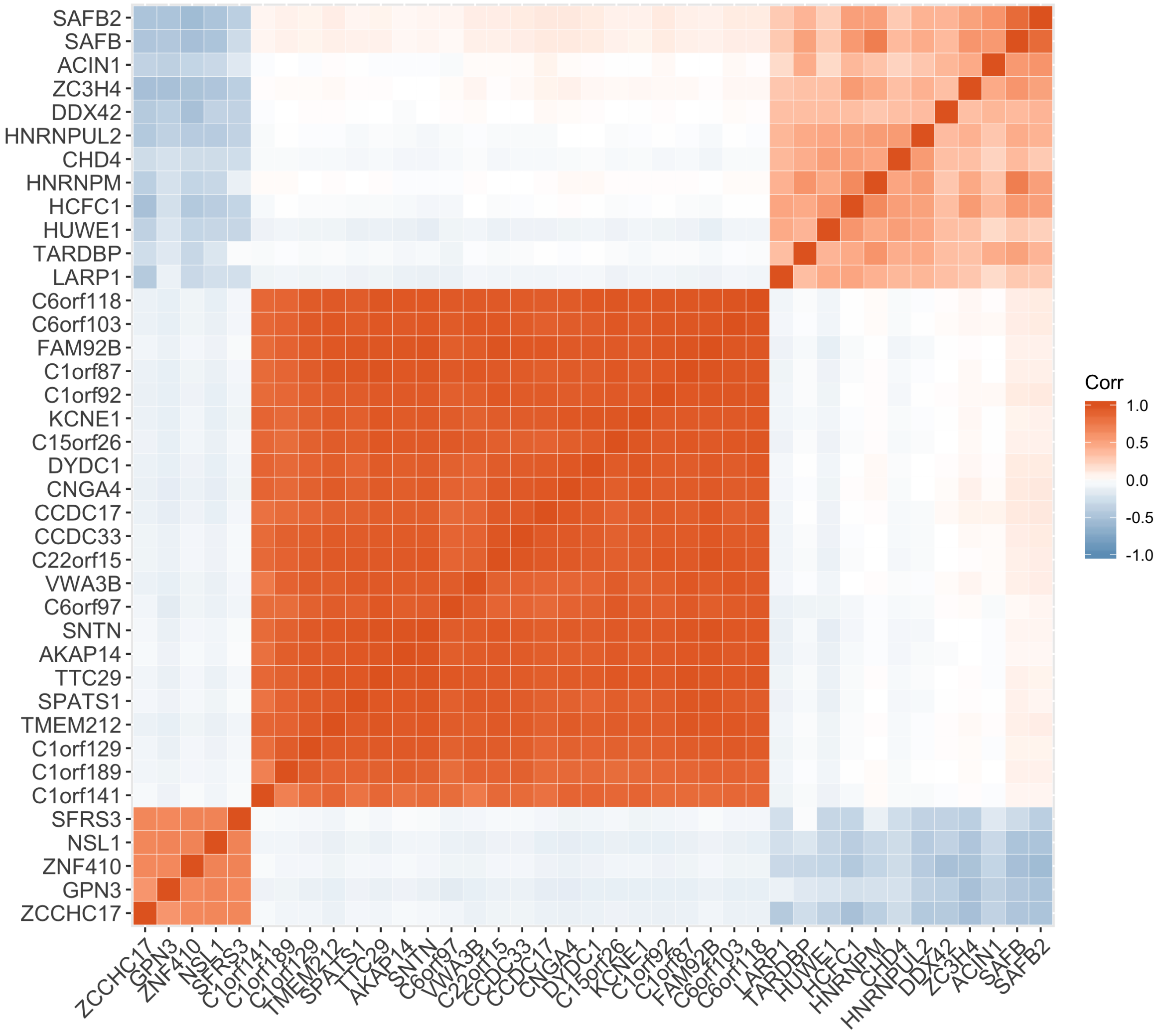


Figure 2.7: Pearson correlation matrix for the thirty-nine genes of LUAD.

Table 2.1: List of 17 Genes

Gene Symbol	Gene Name
NSL1	NSL1, MIS12 kinetochore complex component
SFRS3	serine and arginine rich splicing factor 3
GPN3	GPN-loop GTPase 3
ZCCHC17	zinc finger CCHC-type containing 17
ZNF410	zinc finger protein 410
ACIN1	apoptotic chromatin condensation inducer 1
CHD4	chromodomain helicase DNA binding protein 4
DDX42	DEAD-box helicase 42
HCFC1	host cell factor C1
HNRNPM	heterogeneous nuclear ribonucleoprotein M
HNRNPUL2	heterogeneous nuclear ribonucleoprotein U like 2
HUWE1	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase
LARP1	La ribonucleoprotein domain family member 1
SAFB2	scaffold attachment factor B2
SAFB	scaffold attachment factor B
TARDBP	TAR DNA binding protein
ZC3H4	zinc finger CCCH-type containing 4

As there increasingly attention to the spliceosome pathway in cancer research, we further examine if there is a liquid association of the four finding genes. As a result, we found an LA pattern with HNRNPM, SRSF3, and ACIN1 by using the gene expression data of LUAD and LUSC. Both LA plots (Figure 2.9 and Figure 2.10) show that HNRNPM and SRSF3 have a negative correlation for the samples with the lower expression level of ACIN1 and a positive correlation for the samples with the higher expression level of ACIN1.

In addition, we can show the LA interaction network of master genes by a graph, Figure 2.11. Each square contains the LA score for the three linked genes (circles). The size of the

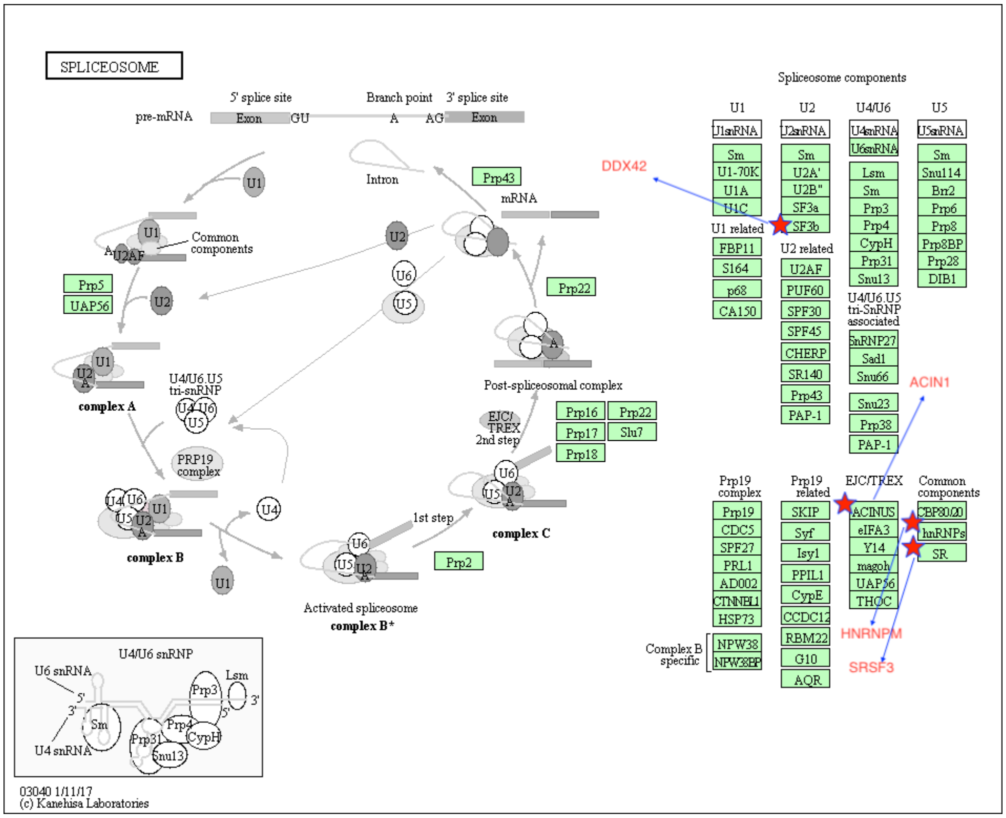


Figure 2.8: The spliceosome pathway with four genes found in the precomputed LA summary. The pathway information generated by KEGG.

circle indicates the degree of the node. Here only the genes with degrees more than one will be displayed. To conduct enrichment analysis, we provide a function for users to download a list of the selected genes, and then employ http://geneontology.org or other similar tools with the list.

Our goal of this example is to showcase that one can use the master genes to discover interesting LA pattern via our developed website without the inputs of the scouting genes or gene pairs. The further analysis and study of the biological mechanism reflecting our observation will be one of our ongoing projects. In addition, It is interesting to examine high frequency LA pairs, master pairs, in the set of triplets with highest/lowest LA scores. Figure 2.12 gives an example. One can click the numbers in **Freq.** to display the corresponding scouting genes as shown in Figure 2.13. The same functions as equipped for the list of master genes are also available here.

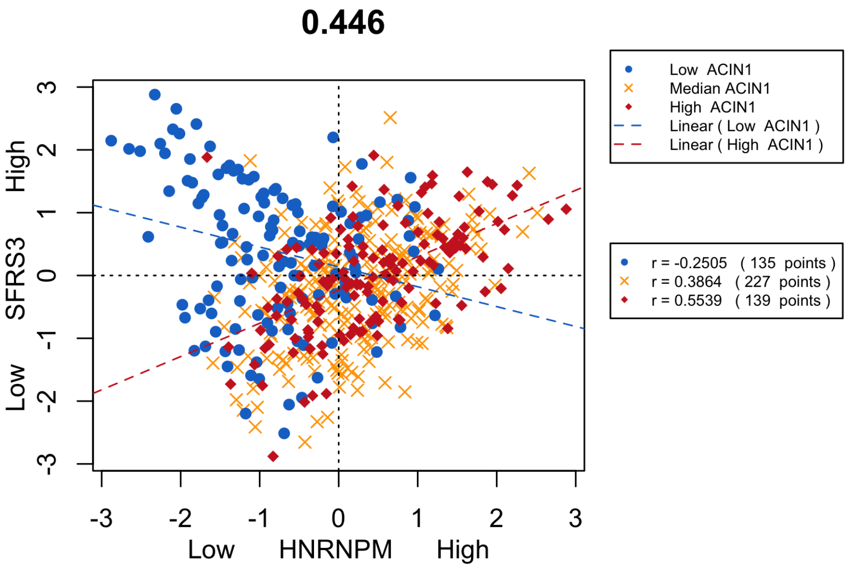


Figure 2.9: LA(HNRNPM, SRSF3|ACIN1) with LUAD.

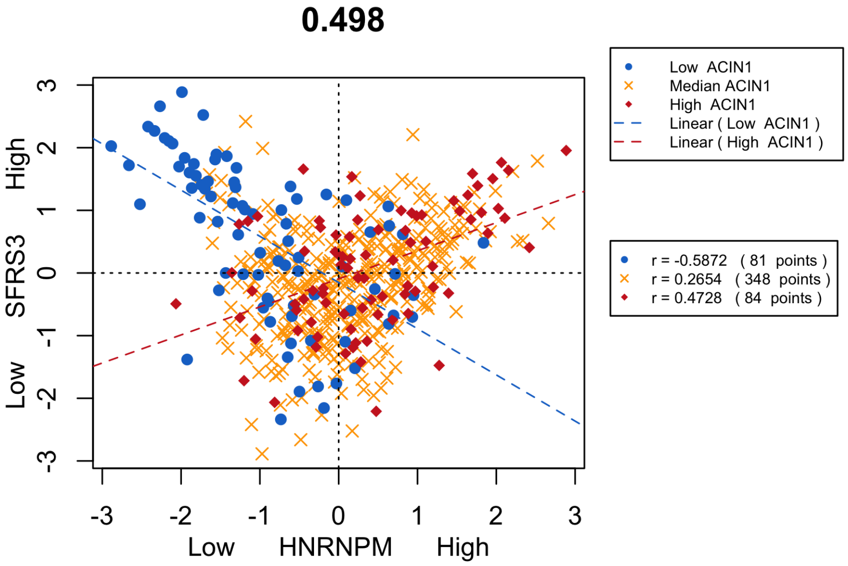


Figure 2.10: LA(HNRNPM, SRSF3|ACIN1) with LUSC.

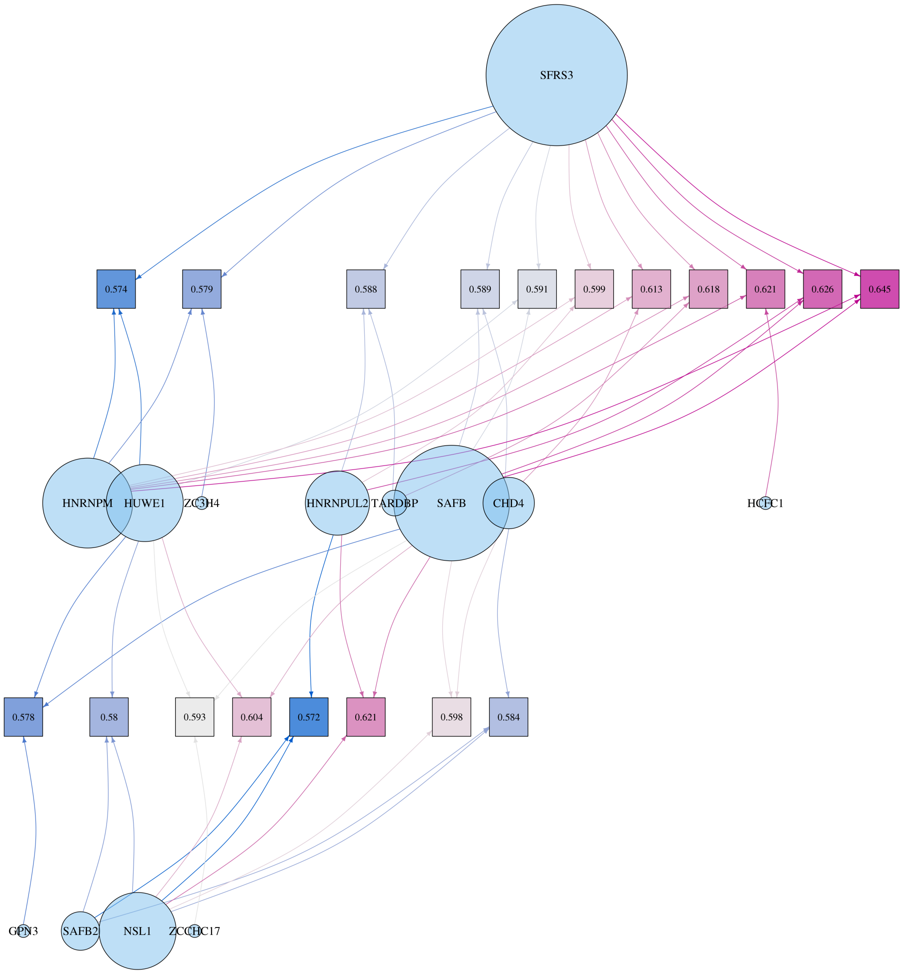
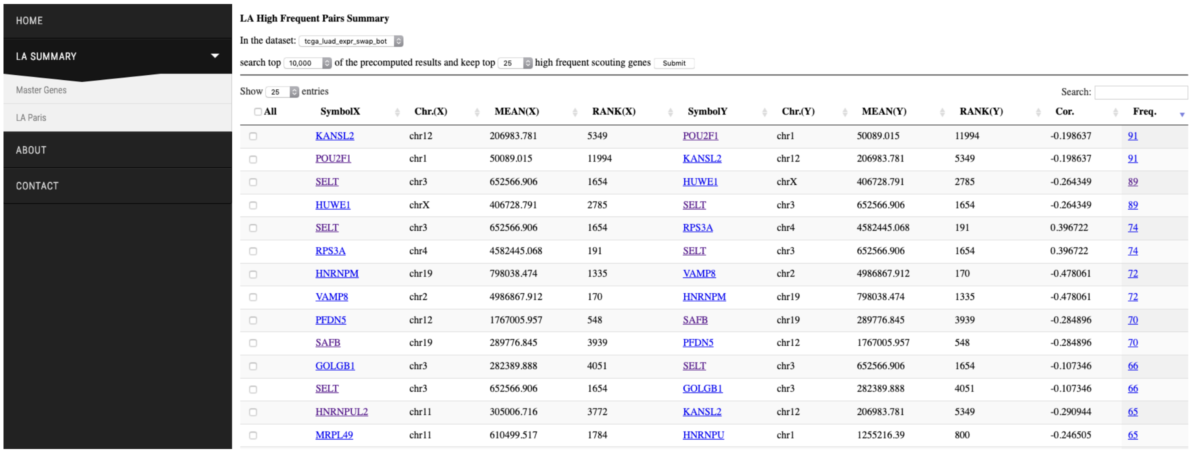


Figure 2.11: An example of LA interaction network. Each color-encoded square contains LA score for three linked genes (circles). The size of the circle indicates the degree of the node.



LA High Frequent Pairs Summary

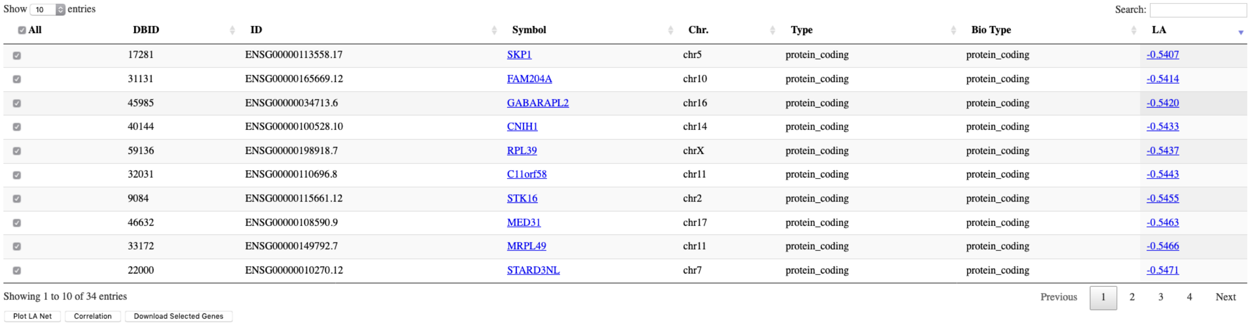
In the dataset: tcga_luad_expr_swap_bot

search top 10,000 of the precomputed results and keep top 25 high frequent scouting genes Submit

Show 25 entries Search:

□ All	SymbolX	Chr.(X)	MEAN(X)	RANK(X)	SymbolY	Chr.(Y)	MEAN(Y)	RANK(Y)	Cor.	Freq.
□	KANSL2	chr12	206983.781	5349	POU2F1	chr1	50089.015	11994	-0.198637	91
□	POU2F1	chr1	50089.015	11994	KANSL2	chr12	206983.781	5349	-0.198637	91
□	SELT	chr3	652566.906	1654	HUWE1	chrX	406728.791	2785	-0.264349	89
□	HUWE1	chrX	406728.791	2785	SELT	chr3	652566.906	1654	-0.264349	89
□	SELT	chr3	652566.906	1654	RPS3A	chr4	4582445.068	191	0.396722	74
□	RPS3A	chr4	4582445.068	191	SELT	chr3	652566.906	1654	0.396722	74
□	HNRNPM	chr19	798038.474	1335	VAMP8	chr2	4986867.912	170	-0.478061	72
□	VAMP8	chr2	4986867.912	170	HNRNPM	chr19	798038.474	1335	-0.478061	72
□	PFDN5	chr12	1767005.957	548	SAFB	chr19	289776.845	3939	-0.284896	70
□	SAFB	chr19	289776.845	3939	PFDN5	chr12	1767005.957	548	-0.284896	70
□	GOLGB1	chr3	282389.888	4051	SELT	chr3	652566.906	1654	-0.107346	66
□	SELT	chr3	652566.906	1654	GOLGB1	chr3	282389.888	4051	-0.107346	66
□	HNRNPUL2	chr11	305006.716	3772	KANSL2	chr12	206983.781	5349	-0.290944	65
□	MRPL49	chr11	610499.517	1784	HNRNPU	chr1	1255216.39	800	-0.246505	65

Figure 2.12: A screen shot of master pairs.



Show 10 entries Search:

☑ All	DBID	ID	Symbol	Chr.	Type	Bio Type	LA
☑	17281	ENSG00000113558.17	SKP1	chr5	protein_coding	protein_coding	-0.5407
☑	31131	ENSG00000165669.12	FAM204A	chr10	protein_coding	protein_coding	-0.5414
☑	45985	ENSG00000034713.6	GABARAPL2	chr16	protein_coding	protein_coding	-0.5420
☑	40144	ENSG00000100528.10	CNIH1	chr14	protein_coding	protein_coding	-0.5433
☑	59136	ENSG00000198918.7	RPL39	chrX	protein_coding	protein_coding	-0.5437
☑	32031	ENSG00000110696.8	C11orf58	chr11	protein_coding	protein_coding	-0.5443
☑	9084	ENSG00000115661.12	STK16	chr2	protein_coding	protein_coding	-0.5455
☑	46632	ENSG00000108590.9	MED31	chr17	protein_coding	protein_coding	-0.5463
☑	33172	ENSG00000149792.7	MRPL49	chr11	protein_coding	protein_coding	-0.5466
☑	22000	ENSG00000010270.12	STARD3NL	chr7	protein_coding	protein_coding	-0.5471

Showing 1 to 10 of 34 entries

Figure 2.13: A list of scouting genes. After clicking **Freq.**, this page shows a list of scouting genes.

2.5 Conclusion

LAP3 aims to facilitate the use of Liquid Association. LAP3 not only provide LA computation and visualization but also statistical toolkits for further analysis of LA scores, such as correlation analysis, hierarchical clustering, the goodness of fit test, etc. Additionally, many functions to manage users' analysis are also developed. This chapter reveals the main structure of LAP3 and describes how LAP3 works. We developed two methods to summarize the precomputed results of LA and showcased an application of utilizing LA summary to search interesting LA patterns. This project aims to help biologists exploit Liquid Association with massive gene expression data. We have finished the precomputed LA scores for couple of gene expression data and expect to complete more in the future.

Appendix A

Outputs for Gene Expression Data of LUAD and LUSC

A.1 Pearson correlation matrices for 17 genes

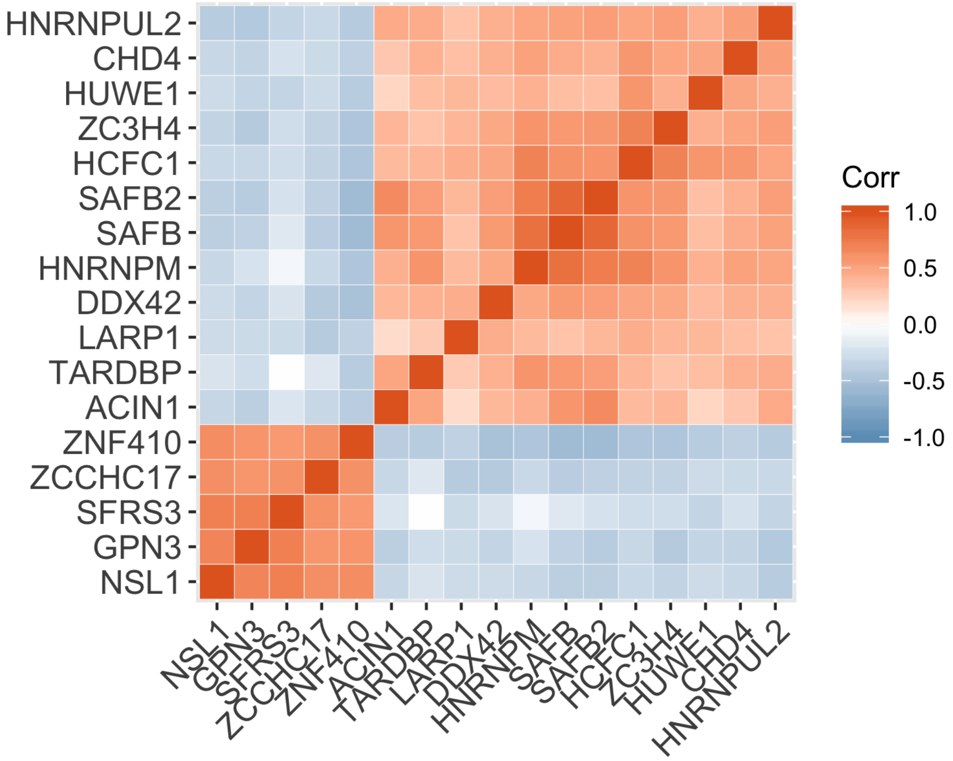


Figure A.1: Pearson correlation matrix for the seventeen genes of LUSC.

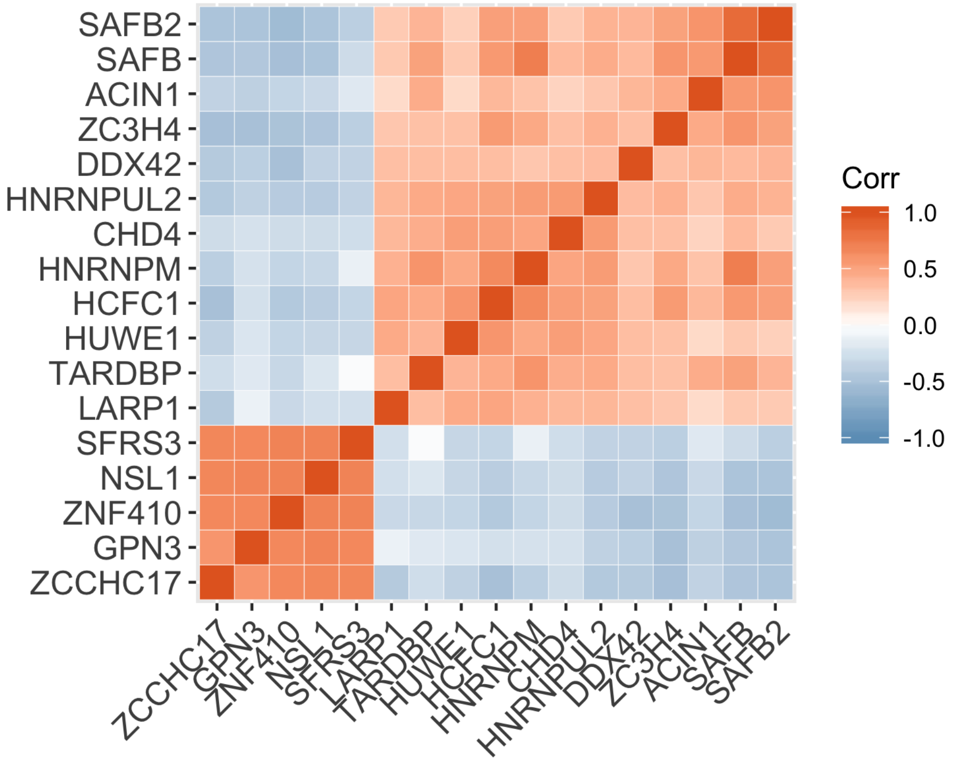


Figure A.2: Pearson correlation matrix for the seventeen genes of LUAD.

CHAPTER 3

Liquid Association on Health-Related Analysis

3.1 Abstract

Liquid Association has been successfully applied to many gene expression studies. In this chapter we extend the application of LA to data from the studying of the public health related data. Concerning the unclear relationships between general public health expenditures and its outcomes, we utilize LA to analyze the cross-nation association of public health expenditures and efficiency with the downloaded data of the Global Health Observatory (GHO). We set the public health expenditures as a given LA-scouting variable (Z) to search LA pairs (X, Y) from female-related/male-related health outcome indicators. Due to the fact that noncommunicable diseases (NCDs) account for over 60% of all deaths worldwide, we set the search domain of X to cover mortality rates by 130 types of NCDs, as well the mortality rate by combing all of NCDs. Meanwhile, the search domain of Y is all of female-related/male-related health outcome indicators. As a result, the liquid association patterns were found relating to general pubic health expenditures. The discovered associations not only agree with the former studies, but also reveal some previously unknown associations of health related indicators.

3.2 Introduction

Liquid Association was inspired by the biological process and has been successfully applied to many biological studies. Yet there is no applications of LA to other fields for LA pattern is existing across domains. In this chapter, we shift our focus from biology to other data intensive field.

Previous studies reported that the exact nature of relationships between public health expenditures and health outcomes remains unclear [CEM13, FFP15]. Public health expenditures also vary widely across different nations. Its cost/efficiency does not appear robust association with economic development [DCC17]. That is, cost/efficiency evaluation between different health care systems in different nations is a complicate issue requiring deep analysis from many perspectives and by different models [EDN95, KL13, FFP15].

In this study, we concern the cross-nation comparison of public health expenditure and efficiency using the GHO data released by [Wor17]. The GHO is the gateway of World Health Organization (WHO) providing health-related statistics in its 194 Member States. There are up to a thousand health outcome indicators including overall health status indicators, the indicators for the specific health and health-related targets of the Sustainable Development Goals.

3.3 Research Method

We downloaded the Year 2012 data from the GHO and preprocessed it in the format that can be directly used by LAP website. To explore if public health expenditure leads different impacts by genders, we organize all of health outcome indicators into three subsets: 306 female-related indicators, 306 male-related indicators and 379 gender-irrelevant indicators. Every indicator keeps the data from 194 member states. The LA-based analysis we conduct is to find LA pairs (X, Y) in the male/female-related indicators by given the LA-scouting variable GGE (Z), *General Government Expenditure on health as a percentage of total government expenditure* in the gender-irrelevant indicators. Since NCDs account for over 60% of all deaths worldwide, we restrict the search domain of X to the mortality rates of 130 types of NCDs, and the mortality rate of all of NCDs combined. Meanwhile, the search domain of Y is restricted to all of female-related/male-related indicators. Here, due to the fact that NCDs are different by genders, the search domains of X and Y are both confined to the same gender related indicators. Figure 3.1 shows the LA-based analysis we conduct.

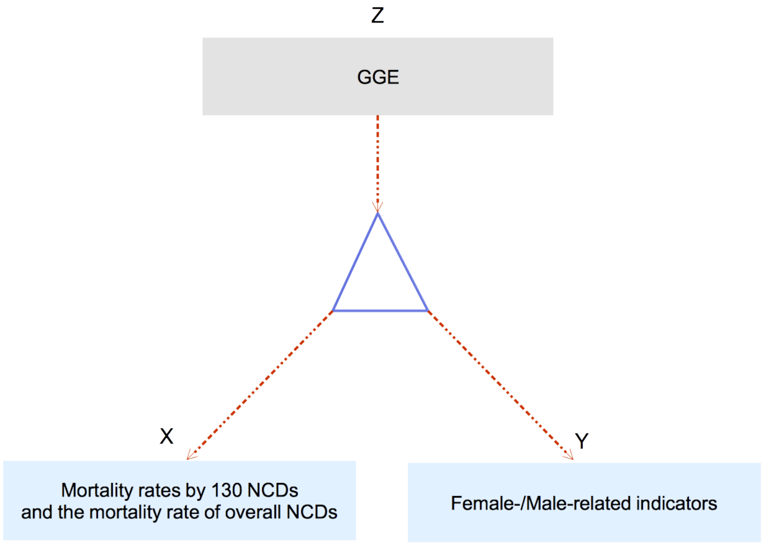

Z
GGE
X
Mortality rates by 130 NCDs
and the mortality rate of overall NCDs
Y
Female-/Male-related indicators

Figure 3.1: LA-based analysis to find LA pairs by given a LA-scouting variable (Z). The search domains of X and Y are both confined to the same gender related indicators, where X is restricted to mortality rates by 130 NCDs and the mortality rate of all of NCDs combined.

3.4 LA-based Analysis Results

We present analysis results by genders, and list the triplets with the highest LA scores in positive and negative. Not only do our findings agree with the former studies, but also reveal previously unknown associations related to GGE.

3.4.1 Finding LA Pairs (X, Y) among Female-Related Indicators

Two leading triplets with the highest positive LA score and the lowest negative LA score are listed in Table 3.1. We found that the correlation between FD75[1] and FMNCD is shown to change from negative for nations with lower GGE to positive for nations with higher GGE (Figure 3.2). Further investigation on how FMNCD correlates with female mortality rate for other age intervals, showed an interesting dynamic pattern of LA (Figure 3.3).

[1] Number of people dying between the beginning of the age group x and the beginning of the next age group $x + n$, n being the interval of the age group, given the hypothetical birth l0 = 100,000 [Wor17].

Table 3.1: LA triplets with the highest LA scores in positive and negative (Female-related indicators for X and Y)

	TOP	BOT
Z	GGE	
X	Female mortality rate by all of NCDs (FMNCD)	Mortality rate by Alzheimer's disease (FMAZ)
Y	Number of people (female) dying between ages 75 and 79 (FD75)	Mortality rate by Cervix uteri cancer (FMCC)
LA	0.336	-0.352

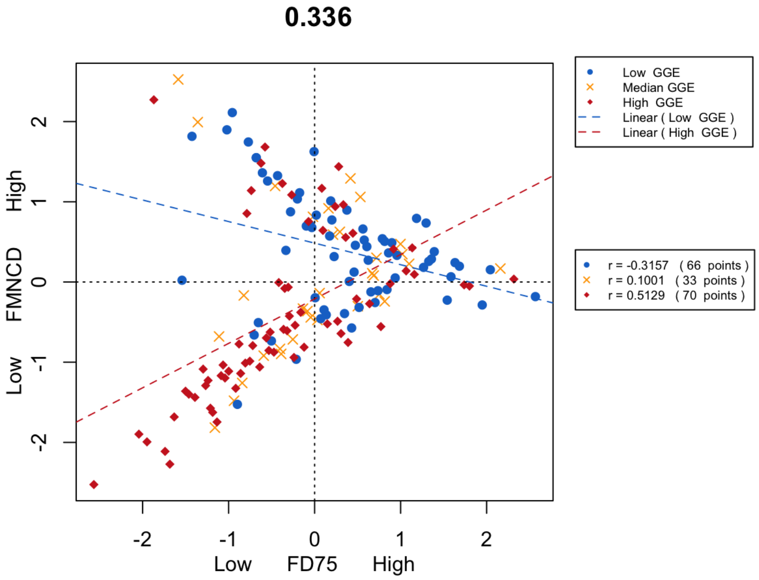


Figure 3.2: LA plot for (FD75, FMNCD, GGE). The correlation between FD75 and FMNCD is shown to change from negative to positive between low GGE nations and high GGE nations.

The dynamic pattern of LA suggests that FD75 is a health outcome indicator that reflects different efficiency between countries with higher GGE and lower GGE. Giving a deeper examination of Appendix B.1. We also found that 40% of nations with higher GGE are high-income countries and only 13% are low-income countries. On the other hand, 31% of nations with lower GGE are low-income countries and 16% are high-income countries. Meanwhile, the similar proportion regarding regions can be found between Europe and Africa in Appendix B.1. Finally, with independent tests GGE is significantly related to income-level and regions of countries.

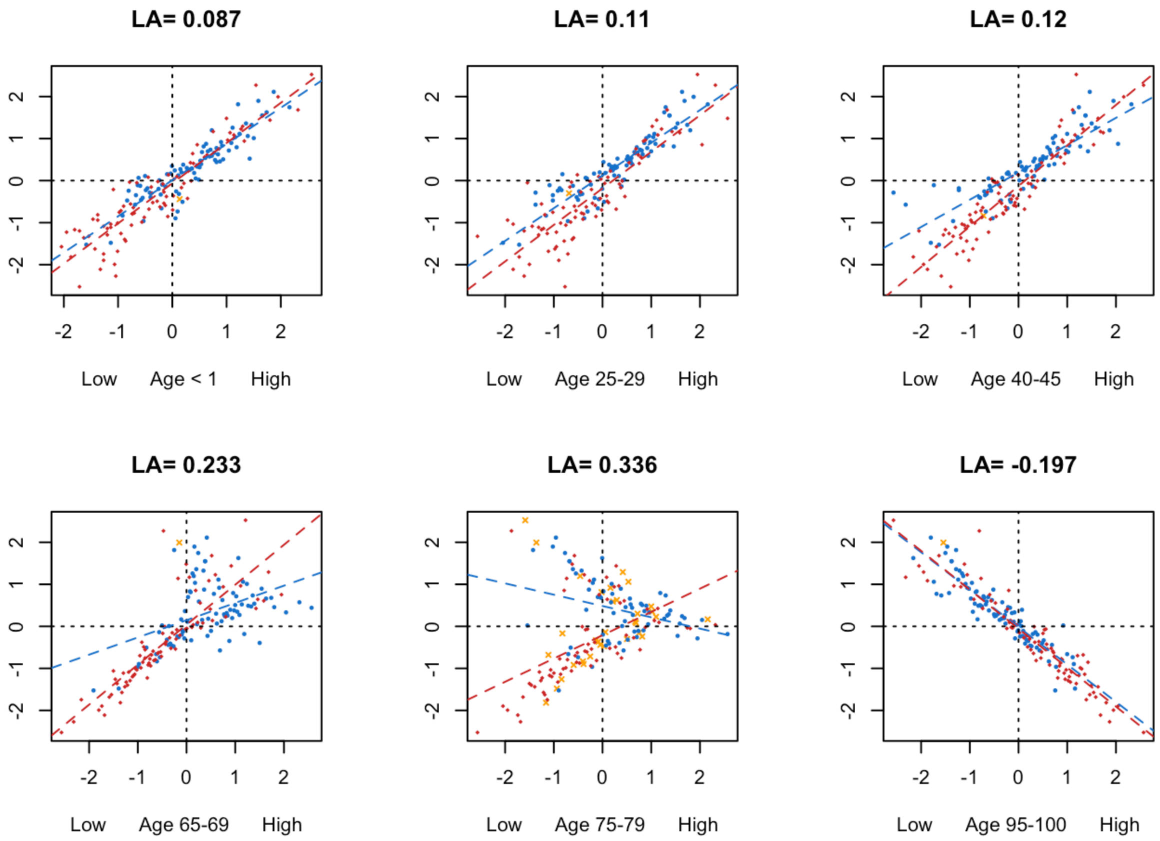


Figure 3.3: The changes of LA scores, where X axis is number of females dying between ages (x, x + 5),and Y axis is, FMNCD

The other leading LA triplet in the negative LA scores is FMAD and FMCC in Table 3.1 and its LA plot is shown in Figure 3.4. The past studies have reported that there are bidirectional inverse associations between cancer and Alzhiemer's disease [RBX05, RFX10, MAD13, OLH13, DBA12, AKU13, STL15], yet some other studies hold different opinions against it [RCA12, FWC16]. In Figure 3.4, for higher GGE, we found that the negative correlation between FMAD and FMCC that agrees with the former studies which suggested the existence of the inverse associations between Alzhiemer's disease and cancers. The similar pattern can also be found in different types of cancers, though their LA scores are slightly smaller than the leading triplet. However, the slightly higher positive correlation for nations with lower GGE in Figure 3.4—1/3 of blue dots are upper-middle- and upper-income countries—reveals an interesting association for cervix cancer is the second most common cancer in women living in less developed regions, and Alzhiemer's disease is one of leading causes of death in high-income economies [Wor17].

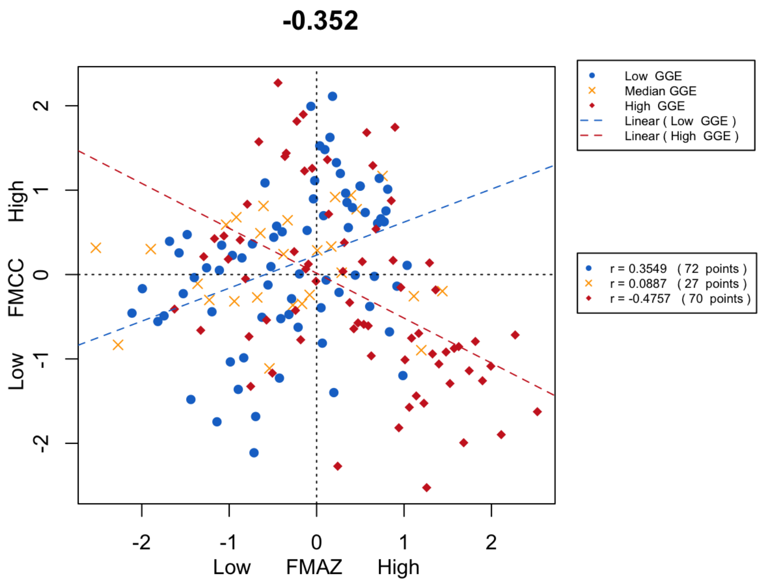


Figure 3.4: LA plot for (FMCC, FMAZ, GGE). The correlation between FMCC and FMAZ is shown to change from negative to positive between low GGE nations and high GGE nations.

3.4.2 Finding LA Pairs (X, Y) among Male-Related Indicators

The top leading triplet in Table 3.2 shows a high positive correlation (0.7021) between MRTA and MMA for the states with higher GGE, of which most are high-income countries in Europe (Figure 3.5). The study [Lut16] concludes that the high traffic accident rate deteriorated the air quality in the UK. Through this evidence of the link between MRTA and air pollution of road-traffic, our result agrees with the past studies [FMD04, MF17] that shows the high correlation between the air pollution of road-traffic and the mortality rate of Asthma.

The leading triplet of negative LA scores shows that a positive correlation between MMPD and MMD for countries with lower GGE. The past studies [FVL15, PAC16, AYR16] have reported that the gastrointestinal tract is affected in PD patients, where the digestive system is made up of the gastrointestinal tract. However, the studies were all mostly based on the data collected in the US, one of countries with high GGE, the evidence is not suffi-

Table 3.2: Triplets with the highest LA score and lowest LA score (Male-related indicators for X and Y).

	TOP	**BOT**
Z	GGE	
X	Male mortality rate by road traffic accidents (MRTA)	Male mortality rate by digestive diseases (MMD)
Y	Male mortality rate by Asthma (MMA)	Male mortality rate by Parkinson's disease (MMPD)
LA	0.298	-0.366

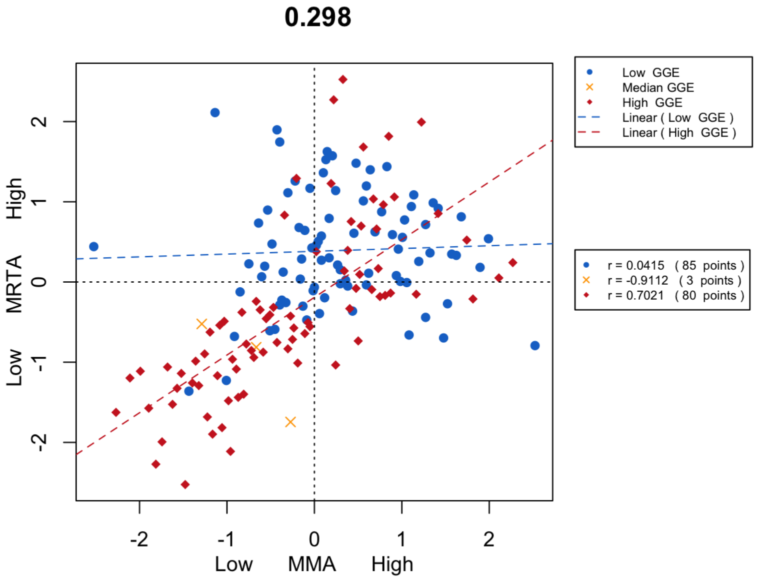


Figure 3.5: LA plot for (MMA, MRTA, GGE). The correlation between MMA and MRTA is shown to change from negative to positive between low GGE nations and high GGE nations

cient to support our finding, and the association we observed calls for further across-nation studies.

Finally, since the pattern of correlation changes between FMNCD and FD75 are observed in Figure 3.2, the similar pattern can also be found for the triplet, X = *Age-standardized (male) mortality rate by all causes* (MMNCD), Y =*Number of people (males) dying between ages 70 and 75* (MD75), and Z = GGE as shown in Figure 3.7. The similar dynamic pattern as Figure 3.3 is shown in Figure 3.8. This shift from 75-79 to 70-75 reflects the fact that the life span of male is in general shorter than female.

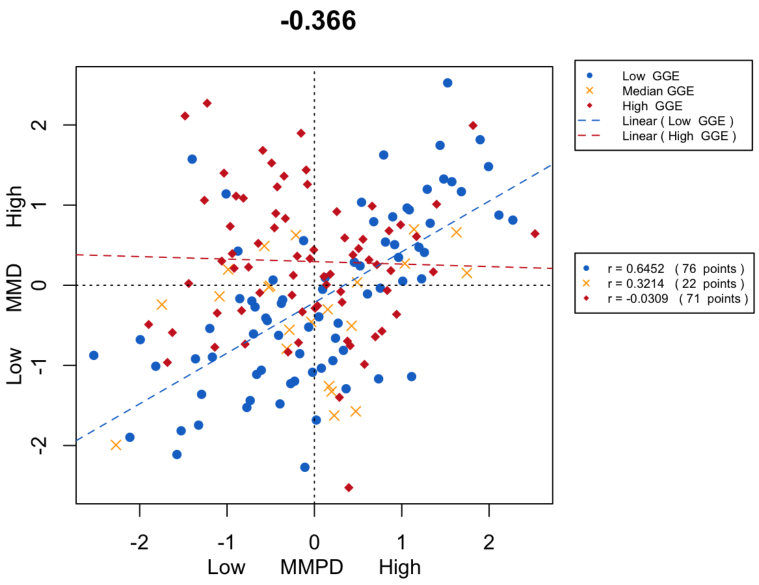


Figure 3.6: LA plot for (MMPD, MMD, GGE). The correlation between MMPD and MMD is shown to change from negative to positive between low GGE nations and high GGE nations

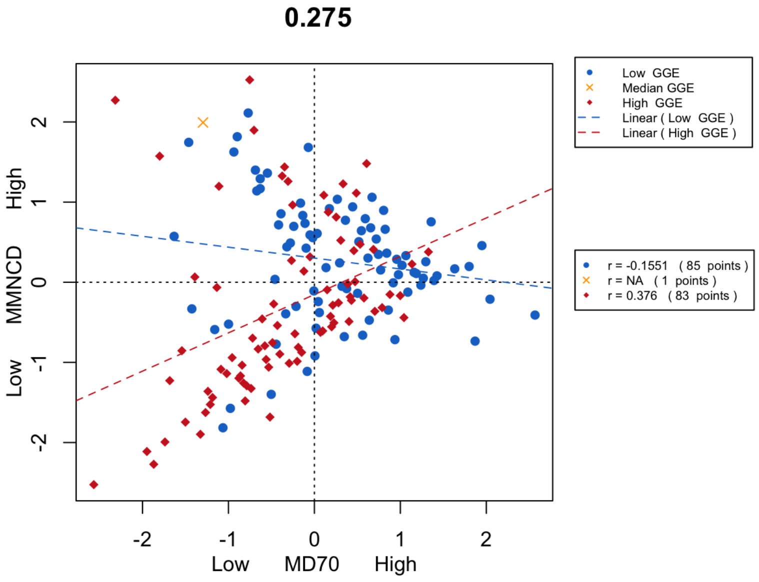


Figure 3.7: LA plot for (MD70, MMR, GGE). The correlation between MD70 and MMR is shown to change from negative to positive between low GGE nations and high GGE nations

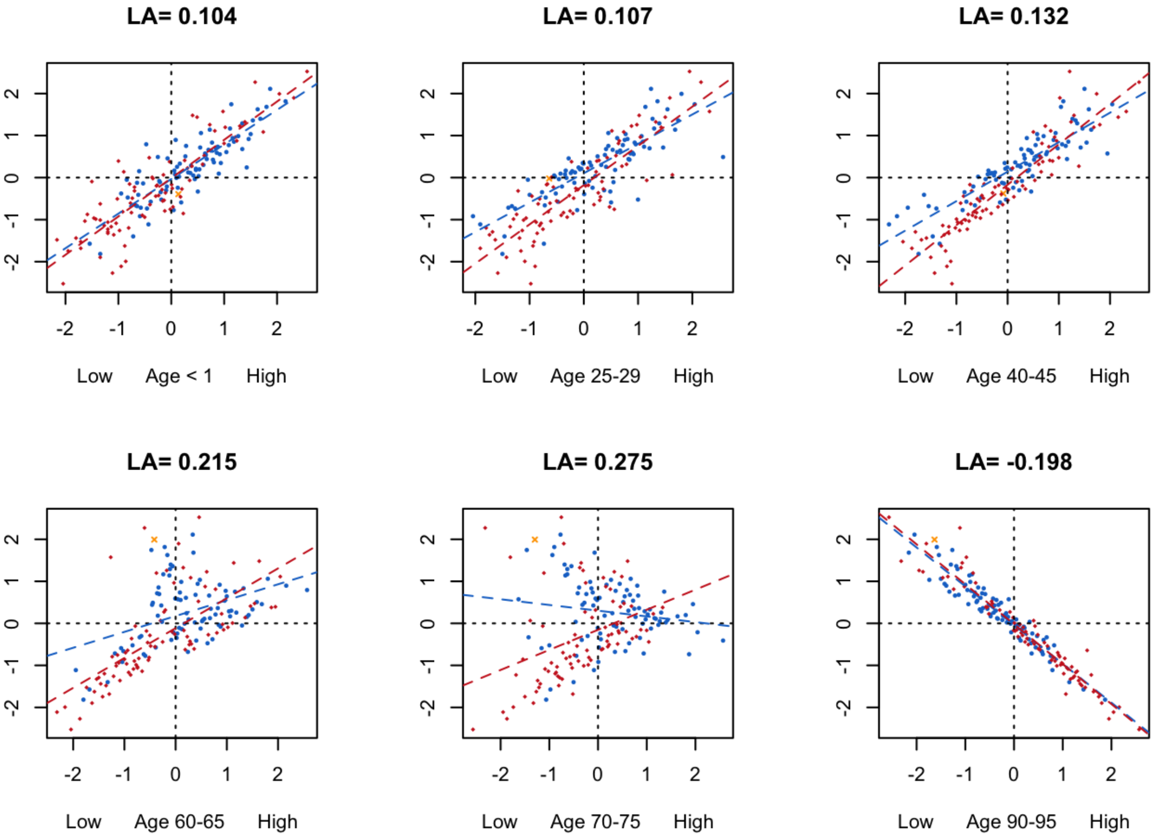


Figure 3.8: The changes of LA scores, where X axis is number of males dying between ages (x, x + 5),and Y axis is MMNCD

3.5 Conclusion

The LA-based analysis reveals that public health expenditure is related to correlation changes between health-related indicators. The correlation-change between the mortality rate by overall NCDs and female/male populations dying at distinctive age-intervals reflects that GGE might affect female and male differently. Although most our results show public health expenditures are significantly related to income levels and regions of countries, some low income countries with higher GGE are found to have the same association fashion as the developed countries.

Unlike most past studies, Liquid Association offers a higher dimension of view on the associations of variables, so that not only did our results agree with the past studies, but also discover previously unknown associations among health-related indicators. Above all, this application showcases how to apply LA-based analysis to other data intensive field.

Appendix B

Outputs for GHO example

B.1 Countries

Table B.1: Countries of LA plot for (FD75, FMNCD, GGE)

Country	INCOME	REGION
Myanmar	Low-income	South-East Asia
Timor-Leste	Lower-middle-income	South-East Asia
Chad	Low-income	Africa
Eritrea	Low-income	Africa
Azerbaijan	Lower-middle-income	Europe
Yemen	Low-income	Eastern Mediterranean
South Sudan	NA	NA
Iraq	Lower-middle-income	Eastern Mediterranean
Pakistan	Lower-middle-income	Eastern Mediterranean
Georgia	Lower-middle-income	Europe
Qatar	High-income	Eastern Mediterranean
Oman	High-income	Eastern Mediterranean
Haiti	Low-income	Americas
Venezuela (Bolivarian Republic of)	Upper-middle-income	Americas
Kuwait	High-income	Eastern Mediterranean
Angola	Lower-middle-income	Africa
Syrian Arab Republic	Lower-middle-income	Eastern Mediterranean

Table B.1 continued from previous page

Country	INCOME	REGION
Saudi Arabia	High-income	Eastern Mediterranean
Malaysia	Upper-middle-income	Western Pacific
Egypt	Lower-middle-income	Eastern Mediterranean
Kenya	Low-income	Africa
Morocco	Lower-middle-income	Eastern Mediterranean
Brunei Darussalam	High-income	Western Pacific
Lao People's Democratic Republic	Low-income	Western Pacific
Sri Lanka	Lower-middle-income	South-East Asia
Congo	Lower-middle-income	Africa
Lebanon	Upper-middle-income	Eastern Mediterranean
Nigeria	Lower-middle-income	Africa
Cambodia	Low-income	Western Pacific
Guinea	Low-income	Africa
Tajikistan	Low-income	Europe
Cyprus	High-income	Europe
Libya	Upper-middle-income	Eastern Mediterranean
Indonesia	Lower-middle-income	South-East Asia
Bhutan	Lower-middle-income	South-East Asia
Equatorial Guinea	High-income	Africa
Ecuador	Lower-middle-income	Americas
Afghanistan	Low-income	Eastern Mediterranean
Gabon	Upper-middle-income	Africa
Trinidad and Tobago	High-income	Americas
Brazil	Upper-middle-income	Americas
Bangladesh	Low-income	South-East Asia
Guinea-Bissau	Low-income	Africa

Table B.1 continued from previous page

Country	INCOME	REGION
Armenia	Lower-middle-income	Europe
Côte d'Ivoire	Lower-middle-income	Africa
Botswana	Upper-middle-income	Africa
Cameroon	Lower-middle-income	Africa
Turkmenistan	Lower-middle-income	Europe
Mozambique	Low-income	Africa
Cabo Verde	Lower-middle-income	Africa
Latvia	Upper-middle-income	Europe
Fiji	Upper-middle-income	Western Pacific
Mongolia	Lower-middle-income	Western Pacific
Maldives	Lower-middle-income	South-East Asia
United Arab Emirates	High-income	Eastern Mediterranean
India	Lower-middle-income	South-East Asia
Bolivia (Plurinational State of)	Lower-middle-income	Americas
Viet Nam	Low-income	Western Pacific
Senegal	Low-income	Africa
Bahrain	High-income	Eastern Mediterranean
Ghana	Low-income	Africa
Uzbekistan	Low-income	Europe
Algeria	Upper-middle-income	Africa
Comoros	Low-income	Africa
Mauritania	Low-income	Africa
Barbados	High-income	Americas
Albania	Lower-middle-income	Europe
Montenegro	Upper-middle-income	Europe
Mauritius	Upper-middle-income	Africa

Table B.1 continued from previous page

Country	INCOME	REGION
Uganda	Low-income	Africa
Hungary	High-income	Europe
Russian Federation	Upper-middle-income	Europe
Benin	Low-income	Africa
United Republic of Tanzania	Low-income	Africa
Philippines	Lower-middle-income	Western Pacific
Niger	Low-income	Africa
Nepal	Low-income	South-East Asia
Israel	High-income	Europe
Jamaica	Upper-middle-income	Americas
Sudan	Lower-middle-income	Eastern Mediterranean
Kazakhstan	Upper-middle-income	Europe
Poland	Upper-middle-income	Europe
Ethiopia	Low-income	Africa
Central African Republic	Low-income	Africa
Paraguay	Lower-middle-income	Americas
Gambia	Low-income	Africa
Romania	Upper-middle-income	Europe
Greece	High-income	Europe
Singapore	High-income	Western Pacific
Ukraine	Lower-middle-income	Europe
Cuba	Upper-middle-income	Americas
Estonia	High-income	Europe
Bulgaria	Upper-middle-income	Europe
Honduras	Lower-middle-income	Americas
Burkina Faso	Low-income	Africa

Table B.1 continued from previous page

Country	INCOME	REGION
Suriname	Upper-middle-income	Americas
Belize	Lower-middle-income	Americas
Kyrgyzstan	Low-income	Europe
Sierra Leone	Low-income	Africa
Finland	High-income	Europe
Ireland	High-income	Europe
Portugal	High-income	Europe
Mali	Low-income	Africa
China	Lower-middle-income	Western Pacific
Panama	Upper-middle-income	Americas
Lithuania	Upper-middle-income	Europe
Democratic Republic of the Congo	Low-income	Africa
Madagascar	Low-income	Africa
Turkey	Upper-middle-income	Europe
South Africa	Upper-middle-income	Africa
Guyana	Lower-middle-income	Americas
Slovenia	High-income	Europe
Belarus	Upper-middle-income	Europe
Tunisia	Lower-middle-income	Eastern Mediterranean
Republic of Moldova	Lower-middle-income	Europe
Malta	High-income	Europe
Serbia	Upper-middle-income	Europe
Luxembourg	High-income	Europe
The former Yugoslav republic of Macedonia	Upper-middle-income	Europe
Republic of Korea	High-income	Western Pacific

Table B.1 continued from previous page

Country	INCOME	REGION
Burundi	Low-income	Africa
Namibia	Upper-middle-income	Africa
Papua New Guinea	Lower-middle-income	Western Pacific
Djibouti	Lower-middle-income	Eastern Mediterranean
Italy	High-income	Europe
Thailand	Lower-middle-income	South-East Asia
Dominican Republic	Upper-middle-income	Americas
Lesotho	Lower-middle-income	Africa
Czech Republic	High-income	Europe
Slovakia	High-income	Europe
Belgium	High-income	Europe
Spain	High-income	Europe
Sweden	High-income	Europe
Croatia	High-income	Europe
Chile	Upper-middle-income	Americas
Togo	Low-income	Africa
Iran (Islamic Republic of)	Lower-middle-income	Eastern Mediterranean
El Salvador	Lower-middle-income	Americas
Iceland	High-income	Europe
Bahamas	High-income	Americas
Mexico	Upper-middle-income	Americas
France	High-income	Europe
Denmark	High-income	Europe
United Kingdom of Great Britain and Northern Ireland	High-income	Europe
Zambia	Low-income	Africa

Table B.1 continued from previous page

Country	INCOME	REGION
Guatemala	Lower-middle-income	Americas
Bosnia and Herzegovina	Upper-middle-income	Europe
Austria	High-income	Europe
Canada	High-income	Americas
Malawi	Low-income	Africa
Norway	High-income	Europe
Jordan	Lower-middle-income	Eastern Mediterranean
Australia	High-income	Western Pacific
Swaziland	Lower-middle-income	Africa
Peru	Upper-middle-income	Americas
Colombia	Upper-middle-income	Americas
Germany	High-income	Europe
Liberia	Low-income	Africa
Japan	High-income	Western Pacific
Nicaragua	Lower-middle-income	Americas
Netherlands	High-income	Europe
United States of America	High-income	Americas
Solomon Islands	Lower-middle-income	Western Pacific
New Zealand	High-income	Western Pacific
Switzerland	High-income	Europe
Rwanda	Low-income	Africa
Argentina	Upper-middle-income	Americas
Uruguay	Upper-middle-income	Americas
Costa Rica	Upper-middle-income	Americas

CHAPTER 4

A Folksonomy-based Approach for Profiling Human Perception on Word Similarity

4.1 Abstract

Automatic assessment of word similarity has long been considered as one important challenge in the development of Artificial Intelligence. People often have a big disagreement on how similar a pair of words is. Yet most word similarity prediction methods, taking either the knowledge-based approach or the corpus-based approach, only attempt to estimate an average score of human raters. The distribution aspect of similarity for each word-pair has been methodologically neglected, thus limiting their downstream applications in Natural Language Processing. Here, utilizing the category information of Wikipedia, we present a method to model similarity between two words as a probability distribution. Our method leverages the unique features of folksonomy. The success of our method in describing the diversity of human perception on word similarity is evaluated against the rater dataset WordSim-353. Our method can be extended to compare documents.

4.2 Introduction

Making machine understand human language is one of the ultimate goals in the development of Artificial Intelligence [Chr15]. In order to reach the goal, many different Natural Language Processing (NLP) tasks were designed. Among them, one of the fundamental upstream task is to automatically assess similarities between words. The performance of this task has direct impacts on many downstream NLP applications such as Question Answering, Information Retrieval, Topic Modeling, and Text Clustering [SG12, NAS16, WLC15], etc. The performance of computed similarity has to be evaluated against human raters, but human raters often display considerable disagreement in assigning similarity scores. As an example, see Figure 4.1 for the distribution of 16 raters' scores assigned to the pair of *life* and *lesson* from **WordSim-353** [02]. Such rating disagreements are quite common. However, most word-similarity methodologies attempt to estimate only the "average" score of human rating. The distribution aspect has been methodologically neglected, thus limiting their downstream applications in NLP.

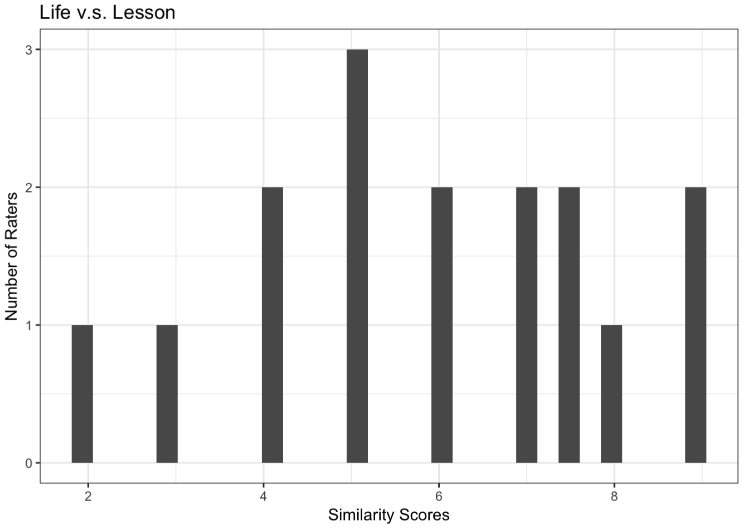
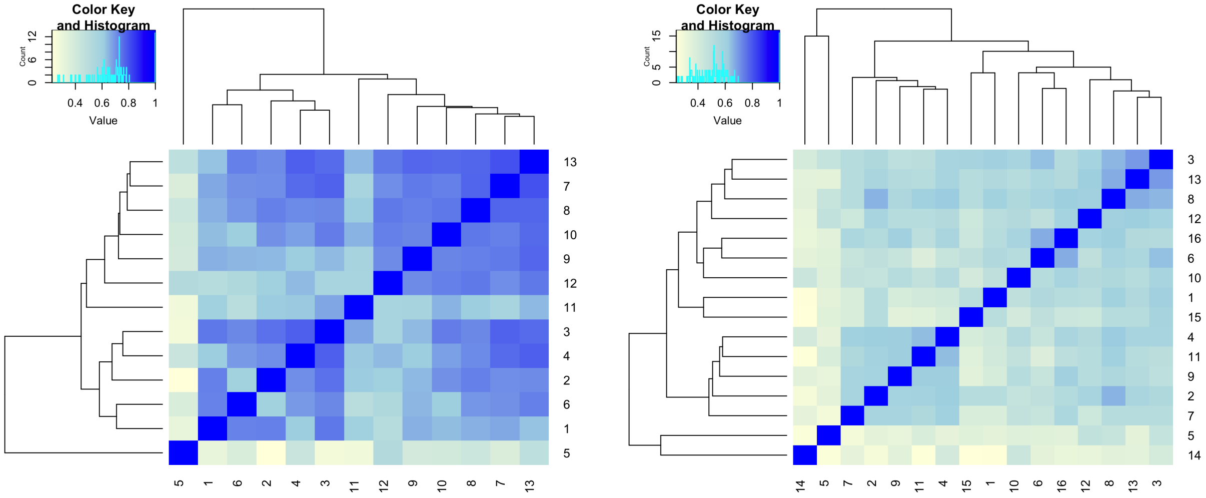


Figure 4.1: The histogram of human ratings on the comparison between *life* and *lesson*

4.3 Rating Disagreement on Word-Similarity

WordSim-353 is composed of two datasets: **WordSim-353.1**, a list of 153 word-pairs rated by 13 persons, and **WordSim-353.2**, a list of 200 word-pairs rated by 16 persons. We computed the Pearson correlation coefficient and the weighted Cohen's kappa coefficient for the similarity scores between any two raters. The results are shown in Figure 4.2 and Figure 4.3 after we ordered raters by hierarchical clustering. Rater disagreement on word-similarity is evident.

The important message we like to deliver is two-fold. First, the computer-imputed single

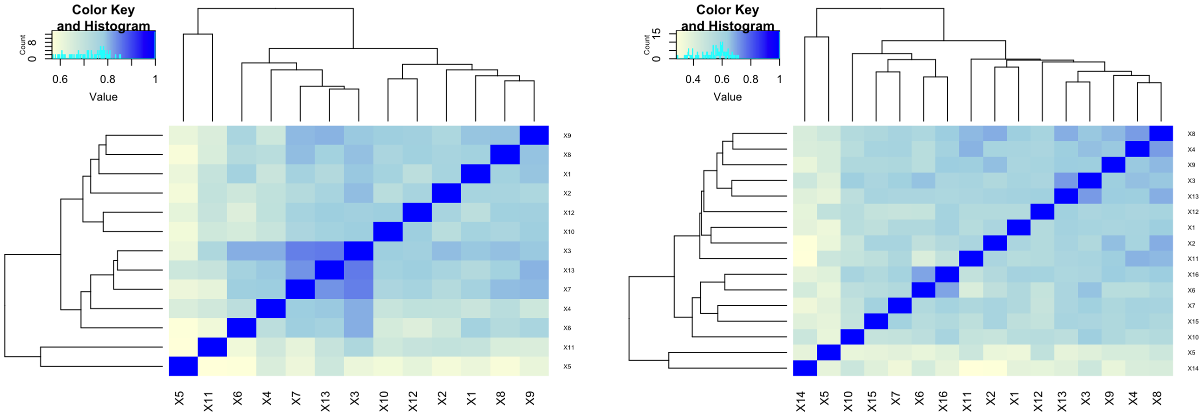


(a) WordSim-353.1

(b) WordSim-353.2

Figure 4.2: Weighted Cohen's kappa coefficient matrices for WordSim-353.1 and WordSim-353.2.

similarity score has grossly simplified the human behavior. Second, using average rater score to evaluate the performance of different word-similarity prediction algorithms is itself a problematic evaluation approach.



(a) WordSim-353.1

(b) WordSim-353.2

Figure 4.3: Pearson correlation matrices for WordSim-353.1 and WordSim-353.2.

4.4 Leveraging Folksonomy for Distribution Quantification of Word Similarity

To reflect the more realistic human behaviors, we propose that in lieu of assigning a single similarity score, a better computer task would be to assign a probability distribution to each word-pair, $(p_0, p_1, \ldots, p_d, \ldots, p_\delta)$, where p_d denotes the probability of similarity score d, and δ is the highest allowable score. To evaluate the performance of a computer algorithm, we should employ common statistical criteria that are designed for the distribution against distribution comparison.

4.4.1 Category Information of Wikipedia

Wikipedia organizes the categories of articles via folksonomy, which is a collaborative tagging system allowing users to tag articles with multiple category notions [AE09]. Links between categories do not impose any specification on relations such as *is-a*, *is-part-of*, *is-an-example-of*, etc. Figure 4.4 illustrated how Wikipedia category is organized into a Directed Acyclic Graph (DAG). It is typical to find multiple roots linking to the title of an article.

In contrast to the traditional centralized classification, folksonomy may directly reflect the diversity of article contributors in their personal styles of vocabulary management, which in turn are influenced by a variety of factors including cultural, social or personal bias. At this writing, about 70,000 editors—from expert scholars to casual readers—regularly edit Wikipedia. (March 2, 2019 https://en.wikipedia.org/wiki/Wikipedia:About)

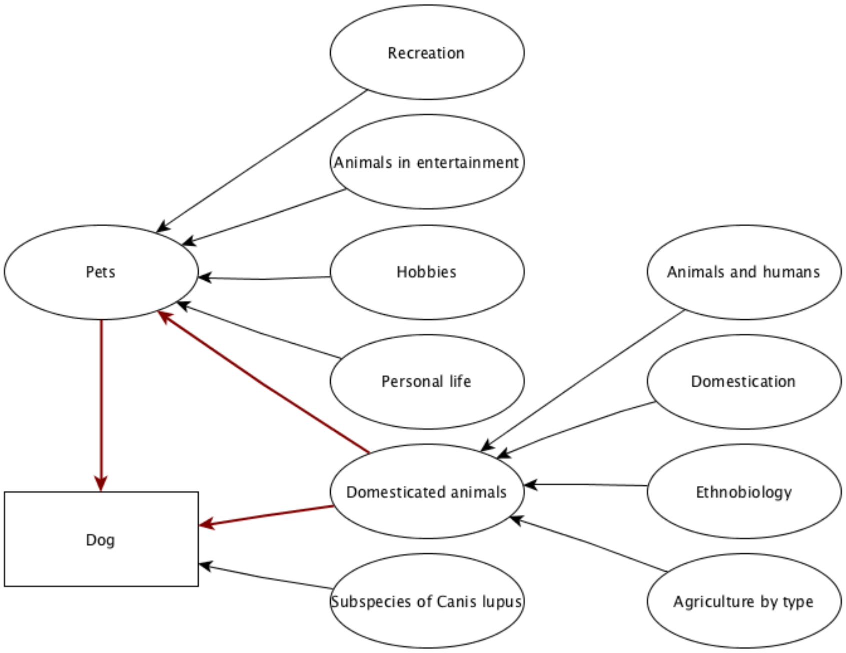


Figure 4.4: An example of Wikipedia category structure, where rectangle indicates a title of an article, and ellipses are categories. The graph is drawn based on the data downloaded from https://wiki.dbpedia.org/data-set-36.

4.4.2 Distribution Quantification of Word-Similarity

We propose a method to assign a probability distribution to a pair of words (W_1, W_2). First, we find the set of conceptual paths $X = \{X_1, \ldots, X_N\}$ linking to W_1, and also find the set of conceptual paths $Y = \{Y_1, \ldots, Y_M\}$ linking to W_2. We delete paths in X that are disconnected from any path in Y, and vice versa. We then compute a similarity score c_{ij} for each path pair (X_i, Y_j) to generate a matrix as shown in Table 4.1. The probability of similarity score d, denoted by p_d, is set to be the proportion of path pairs with $c_{ij} = d$.

We propose Equation 4.1 to calculate the similarity score for (X_i, Y_j).

$$sim(C_i, C_j) = 1 - \frac{(K_i + K_j)}{L_i + L_j} \propto L_i + L_j - K_i - K_j \tag{4.1}$$

Table 4.1: Matrix of Similarity Degrees Between Sets of Conceptual Paths

Y \ X	X_1	X_2	...	X_N
Y_1	c_{11}	c_{12}	...	c_{1N}
Y_2	c_{21}	c_{22}	...	c_{2N}
$\vdots$	...	...	$\ddots$	$\vdots$
Y_M	c_{M1}	c_{M2}	...	c_{MN}

As illustrated by Figure 4.5, L_i is the number of nodes on the path from C_i to its root node R_i, and L_j is the number of nodes on the path from C_j to its root node R_j. K_i is the number of nodes on the path from C_i to C_k, and K_j is the number of nodes on the path from C_j to C_k.

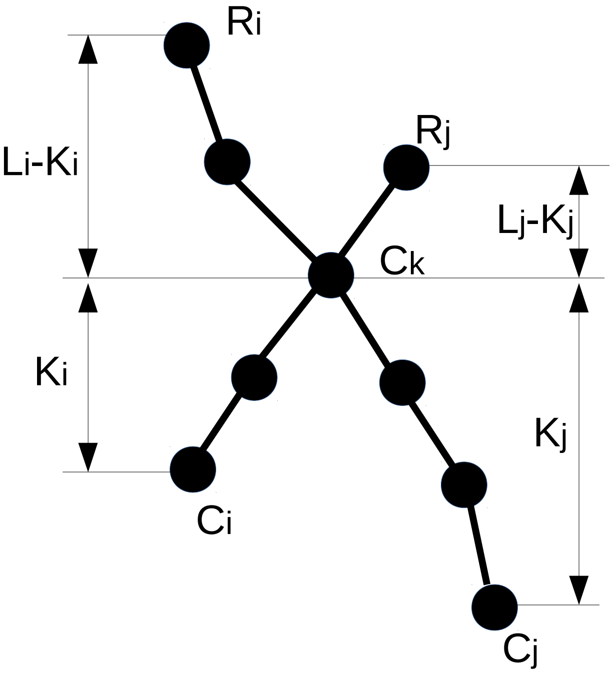


Figure 4.5: Calculating similarity between two conceptual paths via node counting.

In our implementation, we set L_i and L_j as constants and let $L_i = L_j = L$. There are two reasons. First, nodes that are too far away from C_i, C_j are often un-informative. Second, due to the large number of conceptual paths in X and Y, we must alleviate computational complexity. This leads to

$$c_{ij} = 2L - K_i - K_j \tag{4.2}$$

4.4.3 Implementation

Since there are over one million categories contained in Wikipedia, it would be a challenge to collect data directly from Wikipedia. Fortunately, DBpedia has collected and organized Wikipedia data in a way easier for us to use [ABK07]. We downloaded two datasets, *article-categories* and *skos-categories*; the former keeps the links between articles and categories, and the latter stores links between categories. Since the downloaded databases are stored in Triplestore format, *subject-predicate-object*, we set up Apache Jena Fuseki as an in-house SPARQL server for access by our main program.

Figure 4.6 illustrated how we implement our method. After inputing a pair of target words (W_1, W_2), the program will start with stemming the words, and check if they can be found in *article-categories*. If not, the program will search the disambiguation database and return a category closest to the target word. After stemming, the program sends the linked categories as the input to Search Subcategories. This phase recursively searches superior categories of given categories until the search reaches the maximum number of depth we set initially. Once the search is done, the system generates a plain file in Jason format for displaying the output as a taxonomy-like graph on the website. Through the same procedure, the program generates the other plain file in the same format for the other target word. Finally, we use the distribution quantification method described earlier to generate the probability distribution $(p_0, p_1, \ldots, p_d, \ldots, p_\delta)$ for (W_1, W_2).

We developed a website to implement our method, http://ws.stat.sinica.edu.tw/wikiCat. Given a pair of words, it provides a summary table and two taxonomy-like graphs for the input words as shown in Figure 4.7. Every node in the graph represents a category, and it can be clicked to show its superior categories hidden underneath.

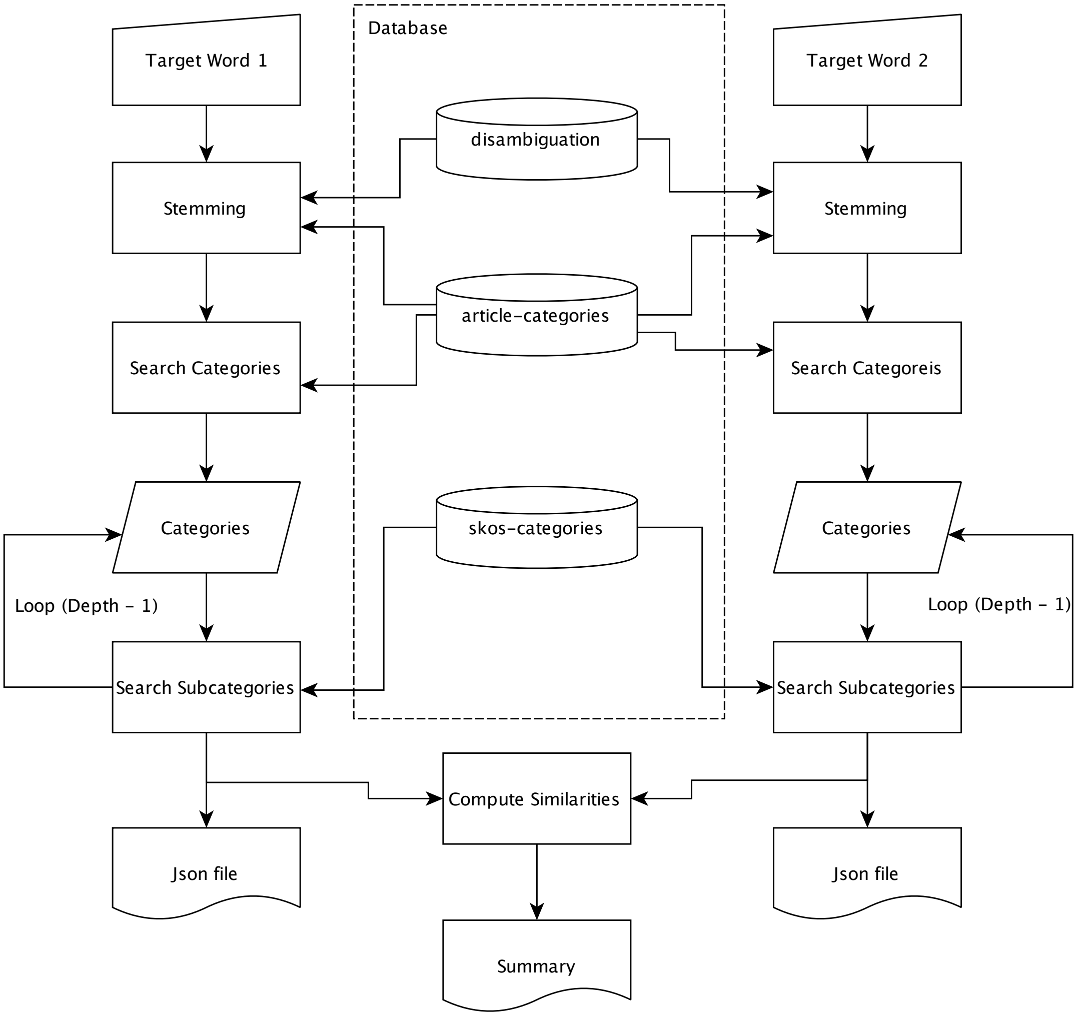
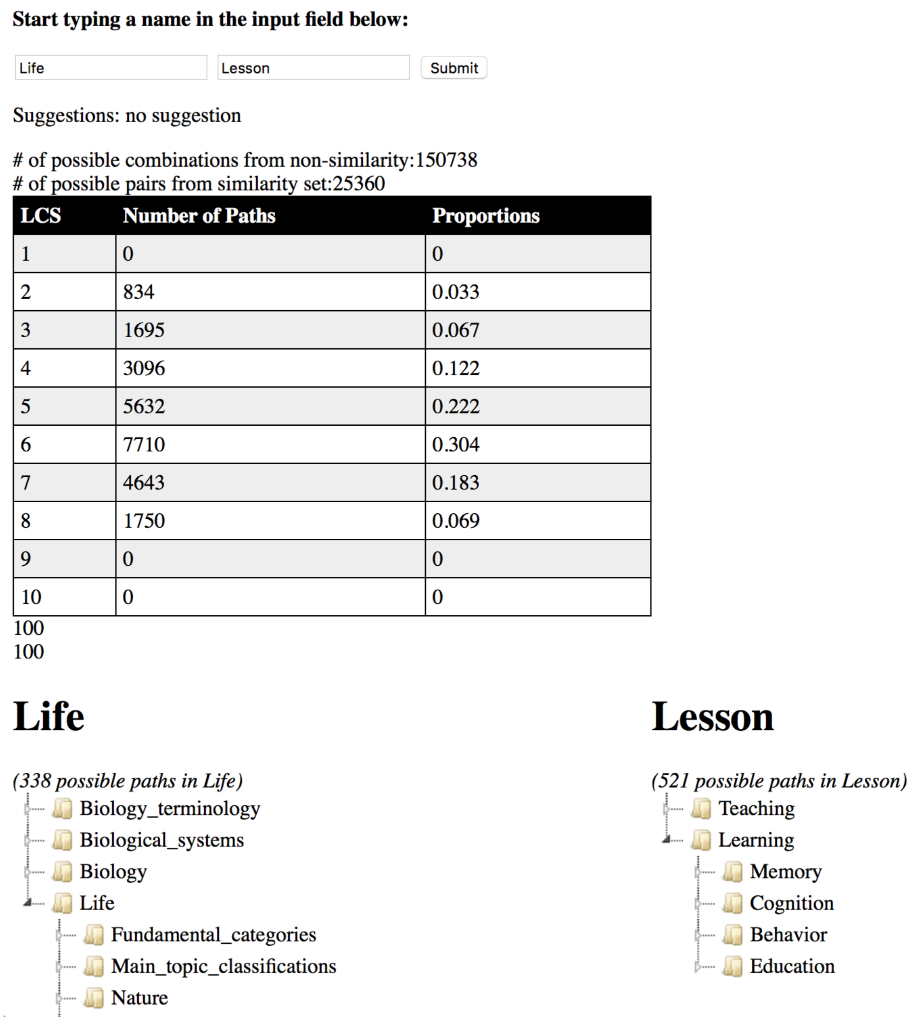


Figure 4.6: The flowchart of the developed main program



Start typing a name in the input field below:

Life

Lesson

Suggestions: no suggestion

# of possible combinations from non-similarity:150738
# of possible pairs from similarity set:25360

LCS	Number of Paths	Proportions
1	0	0
2	834	0.033
3	1695	0.067
4	3096	0.122
5	5632	0.222
6	7710	0.304
7	4643	0.183
8	1750	0.069
9	0	0
10	0	0

100
100

Life

(338 possible paths in Life)

- Biology_terminology
- Biological_systems
- Biology
- Life
  - Fundamental_categories
  - Main_topic_classifications
  - Nature

Lesson

(521 possible paths in Lesson)

- Teaching
- Learning
  - Memory
  - Cognition
  - Behavior
  - Education

Figure 4.7: A screen shot of the developed website.

4.5 Experiment

We use WordSim-353 to evaluate the performance of our method. We set $L = 5$ in order to be consistent with the scale used in WordSim-353 (from 0 to 10), so that our program will yield a probability distribution $(p_0, p_1, ..., p_{10})$ for each word-pair(W_1, W_2). To see how our probability distribution agrees with the score distribution of WordSim-353 raters, Kolmogorov-Smirnov statistic (K-S statistic) between two distributions is used. We perform the following procedure 1000 times to get a p-value. A p-value smaller than 0.05 indicates significant disagreement between the two distributions.

1. Simulating 13 (16, respectively) scores from the distribution $(p_0, p_1, ..., p_{10})$ for the word pair (W_1, W_2) from WordSim-353.1 (from WordSim-353.2, respectively).

2. Computing Kolmogorov-Smirnov distance between $(p_0, p_1, ..., p_{10})$ and the distribution of simulated scores.

After 1000 simulations, the p-value for (W_1, W_2) is given by the proportion of times that the observed K-S statistic exceeds the simulated K-S distance. As it turns, around 50% of word-pairs showed agreement between human rating and our computer rating (Figure 4.8). Given that the raters of WordSim-353 were from a generation before the inception of Wikipedia, we consider this result supports the potential of our folksonomy-based approach in reflecting human judgment diversity. Figure 4.9 showed some cases that our folksonomy-based method agreed very well with human rating.

We further split the word pairs into two groups, AG (agreement, word pairs with p-value > 0.05) and DIS (disagreement, word pairs with p-value < 0.05). We examined the variance of human rater scores for each word-pair and plot the distribution for AG group and DIS group separately for comparison (Figure 4.10). We found AG group of word pairs tend to have larger variance than the DIS group. This indicates our approach may overestimate the degree of divergence in human rating, provided that the small group of raters participating WordSim-353 did not under-represent the true diversity of human behavior.

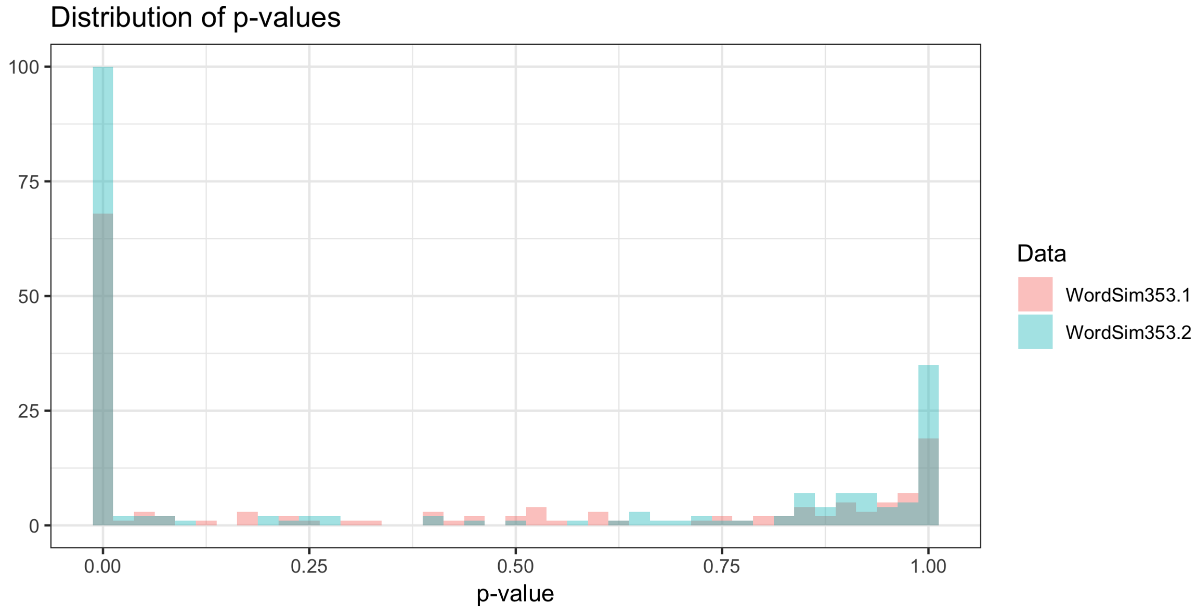


Figure 4.8: Histograms of p-values for WordSim-353.1 and WordSim-353.2. 53.59% of word-pairs have p-values greater than 0.05 in WordSim-353.1 and 48% in WordSim-353.2.

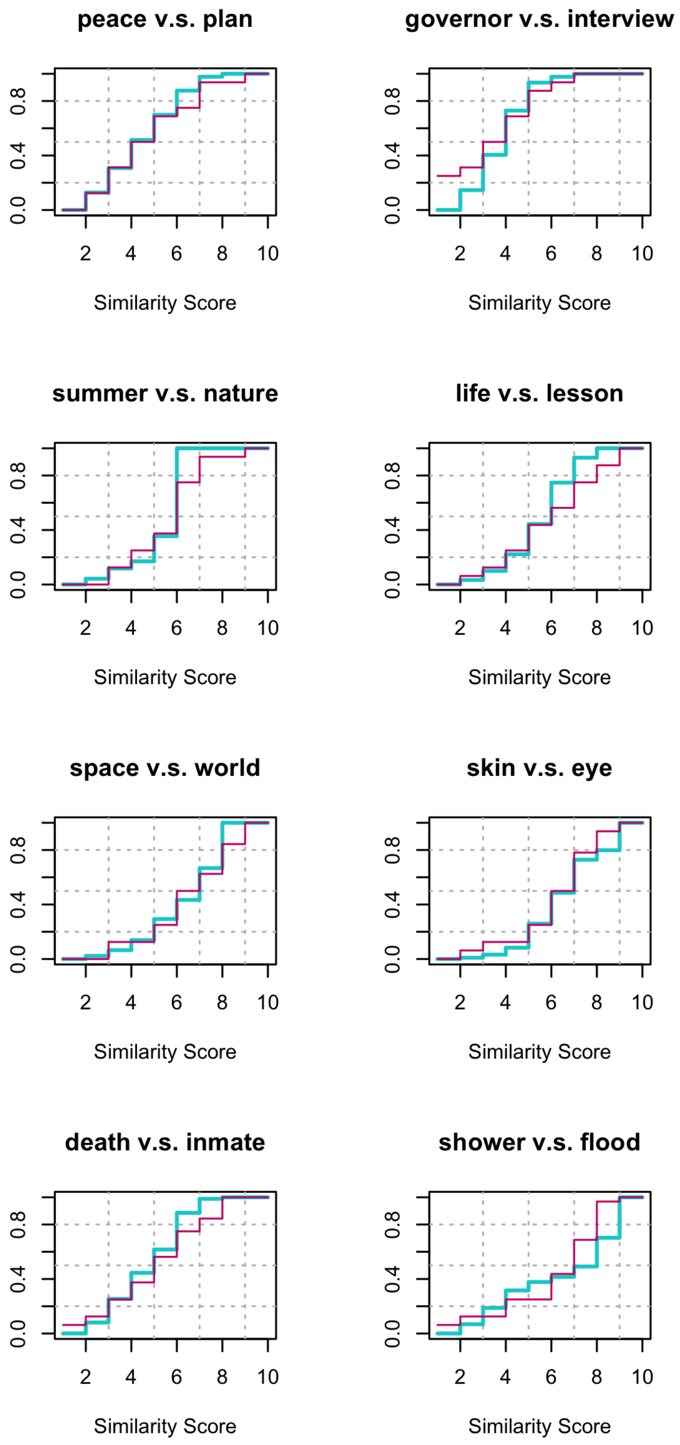


Figure 4.9: Eight cases that our method agreed well with human rating. The red lines are CDF by human rating and the blue lines are CDF by our folksonomy-based method.

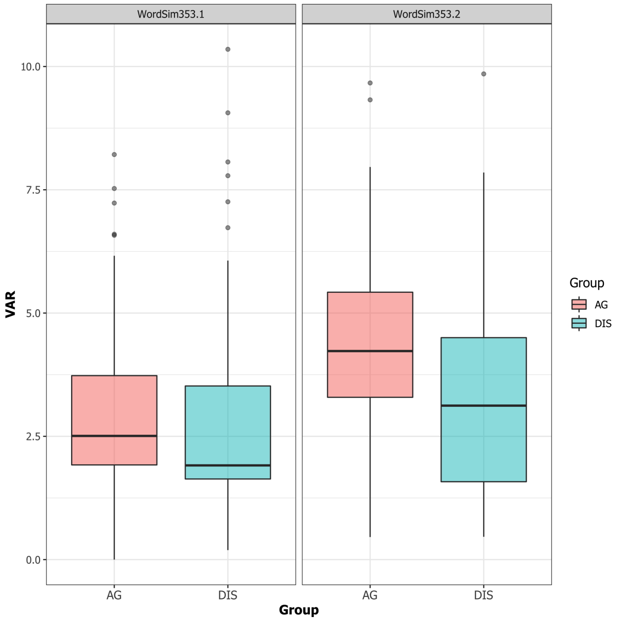


Figure 4.10: Boxplots for variances of similarity scores across 13 raters (WorSim-353.1) and 16 raters (WordSim-353.2). Word-pairs are split into two groups, AG (agreement, $p > 0.05$) and DIS (disagreement, $p < 0.05$).

Table 4.2: Lists of top 10 words with highest tf-idf scores

talk.politics	talk.politics	sci.med
178908	178860	59319
president	oath	widex
masks	garrett	resound
attorney	gain	aids
federal	ingres	programmable
gas	nixon	hearing
reno	powers	loss
yesterday	office	ear
departments	personal	ahead
janet	monetary	sloping
children	indictment	reprogramed

4.6 Application in Document Similarity Comparison

Our method can be extended for comparing documents. As a word can be mapped to multiple conceptual paths, a document will be mapped to an even bigger set of conceptual paths. As an example, we select three documents (*talk.politics.178908*, *talk.politics.178860* and *sci.med.59319*) from The 20 Newsgroups dataset [Lan95]. We further employed tf-idf (term frequency-inverse document frequency) [SM86] to extract the feature words of documents. Only top 10 words with highest tf-idf were kept (Table 4.2). We merge conceptual paths of these words to form a bigger set of representative conceptual paths for each document. Then we applied the same procedure as described in 3.2 to yield a probability distribution of similarity scores between two documents.

In this example, we set $L = 4$ to yield a probability distribution $(p_0, p_1, \ldots, p_8)$ for comparing two documents as shown in Table 4.3. Here PP is *talk.politics.178908* v.s. *talk.politics.178860*, PM1 is *talk.politics.178908* v.s. *sci.med.59319* and PM2 is *talk.politics.178860* v.s. *sci.med.59319*. Evidently, the probability distributions for (*talk.politics.178908*, *sci.med.59319*) and (*talk.politics.178860*, *sci.med.59319*) have low probabilities on high similarity scores (6, 7, 8). In contrast, we observe relatively higher probabilities being assigned to high similarity scores for (*talk.politics.178908*, *talk.politics.178860*).

Table 4.3: Probability Distributions of Document Similarity

Scores[a]	**PP**	**PM1**	**PM2**
0	0	0	0
1	0	0	0
2	0.1236742	0.2240363	0.2725498
3	0.1616162	0.3133787	0.3924248
4	0.1674242	0.245805	0.2225693
5	0.1511995	0.2126984	0.1124561
6	0.1440657	0.00408163	0
7	0.1337121	0	0
8	0.1183081	0	0

[a] Similarity scores. Here we set L = 4.

4.7 Conclusion

Human perception on word similarity can be very discordant. Against the common trend of assigning a single score of similarity by most computer algorithms, we request a new computer task of assigning a probability distribution of similarity for each word pair. Leveraging the rich information embroidered behind the principle of free expression and empowered by user diversity of folksonomy, we design an approach that exploited the category tagging system of Wikipedia articles to perform the task. The good performance of our method is illustrated against two word similarity datasets with scores assigned by human raters. For future works, we plan to modify our word similarity scoring formula by path-dependent weight adjustment for broadening the application in document comparison.

To sum up, our contributions are fivefold. First, we take a first step in redirecting the task of word similarity from single score assignment to probability distribution assignment. Second, we are the first to recognize the rich information contained in folksonomy can be exploited to describe the diversity of human perception on word similarity. Third, we developed a method to perform the new task, and created a website to implement our method and allow for on-line word comparison by the public. Finally, our word similarity method can be directly extended for document similarity comparison.

CHAPTER 5

Comparison between Knowledge-based and Corpus-based Approaches to Word Similarity Prediction

5.1 Abstract

Methods automatically judging word similarity generally fall into two categories, knowledge-based and corpus-based approaches. However, the connection between the outcomes of the two approaches remains unclear. The corpus-based approach generates word vectors by training models with a large training corpus. To obtain a similarity score between two words, the dot product of the two word vectors is computed. Instead of the dependence on which corpus to use, the knowledge-based approach requires a preexisting knowledge base. This section aims to compare their prediction performance via regression and factor analysis. We found that the outputs of the two approaches indeed reflect disjointed perceptions human raters employed in the word-similarity tasks. Therefore, we proposed a way to easily distinguish what word pairs that two approaches yield consistent/inconsistent predictions.

5.2 Introduction

Methods of judging word similarity generally fall into two categories, corpus-based and knowledge-based approaches [HRJ15]. The corpus-based approach was founded on the maxim, *You should know a word by the companies it keeps* [Fir57], which has shown remarkable performance on different word-similarity tasks. Landauer et al. proposed Latent

Semantic Analysis (LSA) that employs singular value decomposition to generate vectors as word representations [TPD98]. Since then, many methods were proposed to generate word vectors. Bengio et al. published a series of papers using neural network techniques [YRP03]. The team of Tomas Mikolov proposed the continuous bag of words (CBOW) and skip grams (also known as word2vec) [TKG13] and Jeffrey et al. proposed Global Vectors for Word Representation (GloVe) [PSM14]. These methods need to be fed with a large corpus to train models in order to generate word vectors. To obtain a similarity score between two words, the dot product of the two word vectors is computed.

Instead of the dependence on which corpus to use, the knowledge-based approach requires a pre-existing knowledge base. WordNet is the most common knowledge base employed by the majority of methods developed in this realm. WordNet collects over 150,000 English words, and organizes them into cognitive synonyms (synsets). These synsets are connected through conceptual, semantic and lexical relations such as hyponyms, hypernyms, meronyms, holonyms [Geo95]. Wu and Palmer proposed a method that exploited ontology/taxonomy to compute similarity scores based on Least Common Subsumer (LCS) [ZM94]. Many methods based on LCS, known as the edge-counting-based approach, were proposed [TBK06, YZD03, HBB14]. Another type of knowledge base approach used features of words to assess the similarities [Amo77, AM03, EGA06].

In order to evaluate the performance of word-similarity methods, many test collections have been proposed (Table 5.1). Each test collection provides a list of word-pairs rated by multiple human raters. Generally, Pearson correlation coefficients of word-similarity methods' output and the average score of human rating indicates the performance of the word-similarity methods. However, there is a lack of statistical methodologies to measure differences between word-similarity methods [FTR16]. Therefore, we conducted a series of analysis to distinguish the outcomes of knowledge-based and corpus-based approaches and show both approaches reflect distinctive perceptions that human employed in word-similarity tasks. Finally, we proposed a better way to assess the performance of different word-similarity methods with distinctive approaches.

Table 5.1: Word similarity test collections

Test Collection	Number of Word Pairs	Source
RG	65	[RG65]
MC	30	[MC91]
WS-353	353	[FGM01]
YP-130	130	[YP06]
MTurk-287	287	[RAG11]
MTurk-771	771	[HDG12]
MEN	3000	[BBB12]
RW	2034	[LSM13]
Verb	144	[BRK14]
SimLex	999	[FRA14]
SemEval-2017 Task 2	500	[CPC17]

5.3 Research Methods

Since designing word-similarity prediction methods is one of fundamental task in AI, many word-similarity performance tests have been available for the public. However, many word-pairs in the test collections are the comparison of words in different part-of-speech, such as *day* vs *sunny*, *dirt* v.s. *dirty*, *dirty* v.s. *friends*, etc. In addition, some earlier test collections did not have enough number of word pairs for reaching statistical significance. To simplify the analysis, we focus only on the similarity between nouns and use the 666 Noun-Noun pairs of SimLex-999 (SimLex-666) and WordSim-353 for the following analysis. Here, SimLex-999 is a test collection designed to measure the performance of word-similarity methods on semantic similarity. It provides average human ratings of 666 Noun-Noun pairs, 222 Verb-Verb pairs (SimLex-222), and 111 Adjective-Adjective pairs (SimLex-111)

We execute four different knowledge-based methods—Tversky's, Rodriguez's, MA Hadj Taieb's, and X-Similarity[1]—and two popular corpus-based methods—word2vec and GloVe—with SimLex-666 and WordSim-353. In addition, as our folksonomy-based word similarity method can work with a wide knowledge base, we fed it with both hypernyms of WordNet and the

[1] We initially selected MA Hadj Taieb's method and X-Similarity as representative methods for edge-counting-based and feature-based. However, X-Similarity was published in 2006, earlier than MA Hadj Tajeb's. Therefore, we add two more popular feature-based methods into the analysis to increase the prediction power of the knowledge-based approach.

category information of Wikipedia respectively, and computed the expected similarity scores for word pairs of SimLex-666 and WordSim-353. The outputs with hypernyms of WordNet is denoted by wordnet.E, and the other with Wikipedia is denoted by wiki.E. In total, eight measurements (outputs of word-similarity methods) were obtained for a word pair.

5.3.1 Regression Analysis

In this section, we compare the prediction power of two approaches and examine whether they can perform equally well on SimLex-666 and WordSim-353. Since most word-similarity prediction methods only attempt to estimate an average score of human raters, we take the average scores as the response variable Y. We selected two similarity measurements, Hadj-Taieb's (X) and GloVe (Z), to represent two approaches respectively due to best performance in their kind, and start with a linear regression model 5.1.

$$Y = \alpha X + \beta Z + \epsilon_1 \tag{5.1}$$

We wish to know what percent of variation for the response variable Y not explained by $X(Z)$ is explained by $Z(X)$ with both WordSim-353 and SimLex-666. Applying the same procedure (Equation 5.2-Equation 5.6) to both test collections and obtain partial R^2s, we obtain the outcomes shown in Figure 5.1a and Figure 5.1b.

From 5.1, we obtain R^2

$$X = \beta' Z + \epsilon_2 \tag{5.2}$$

$$Y = \gamma_z \epsilon_2 + \epsilon_3 \tag{5.3}$$

From 5.3, we obtain $R^2_{\gamma_z}$

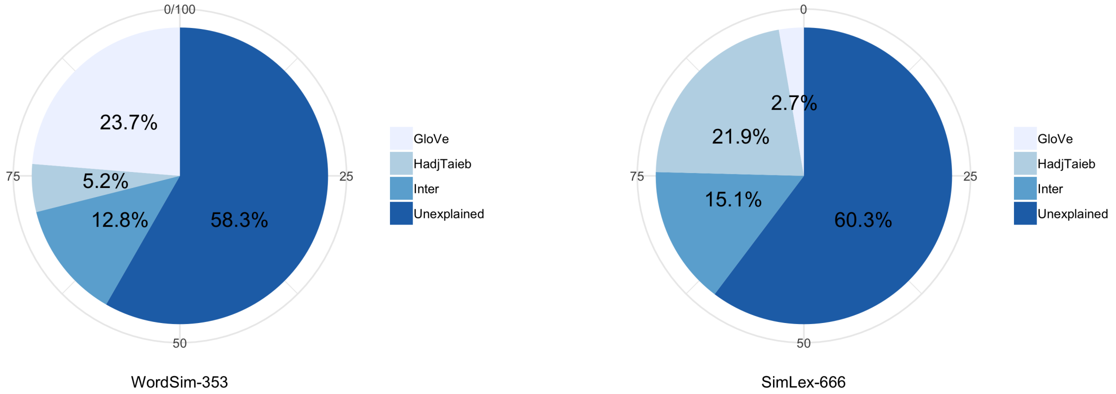
$$Z = \alpha' X + \epsilon_4 \tag{5.4}$$

$$Y = \gamma_x \epsilon_4 + \epsilon_5 \tag{5.5}$$

From 5.5, we obtain $R^2_{\gamma_x}$

$$R^2_{inter} = R^2 - R^2_{\gamma_z} - R^2_{\gamma_x} \tag{5.6}$$

Accordingly, for WordSim-353 GloVe shows a better prediction power than HadjTaieb's method, whereas HadjTaieb's method explains more variation than GloVe for SimLex-666. However, both plots show over 50% of unexplained variations. We have also conducted the same procedure for other methods. As a result, they have an even higher ratio in unexplained variation.



(a) WordSim-353

(b) SimLex-666

Figure 5.1: The percentage of R^2 contributed by GloVe and HadjTaieb's method for WordSim-353 and SimLex-666.

Two approaches show inconsistent performance across WordSim-353 and SimLex-666. Regardless of human ratings, we further examine whether or not a common factor of approaches exists with factor analysis.

5.3.2 Factor Analysis

We take the word-similarity measurements of SimLex-666 and WordSim-353 obtained by 8 methods as observable variables and generate the correlation matrices as shown in Figure 5.2 and 5.3 for conducting factor analysis. Here we employed **factanal( )** function in the *psych* R package to produce maximum likelihood factor analysis for 2 common factors (Table 5.2 and Table 5.3) at 0.05 significant level. In both tables, F is the estimated loading factors and F^* denotes the rotated (with varimax) estimated loading factors.

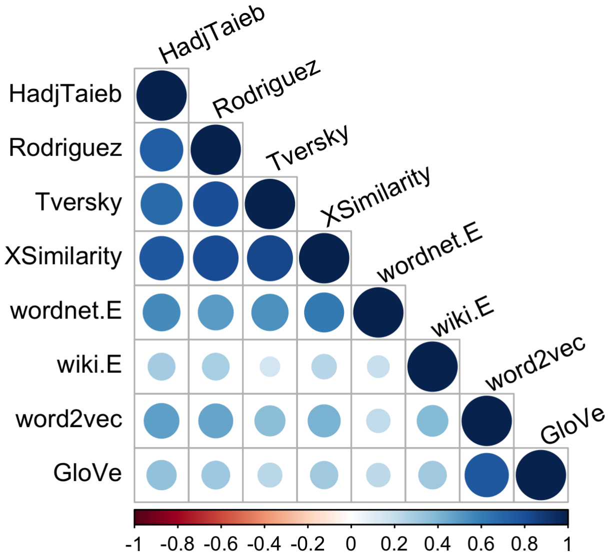


Figure 5.2: The correlation matrix of eight word-similarity measurements with WordSim-353.

In both Table 5.2 and Table 5.3, all the measurements are positive in the first loading factor F_1 that indicates the first loading factor might reflect a common factor of all variables. Additionally, the high loading of word2vec and GloVe suggest the corpus-based approach explains more variation in this factor. Therefore, we might label it a *common sense* factor. For the second loading factor F_2, the measurements of knowledge-based approach are all positive, whereas the rest are all negative. This factor reflects a clear difference between two approaches.

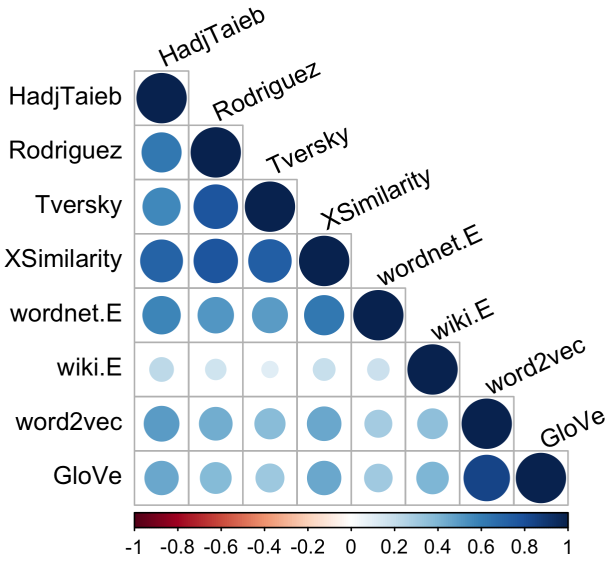


Figure 5.3: The correlation matrix of eight word-similarity measurements with SimLex-666.

Table 5.2: Factor Analysis of WordSim-353

Variable	F_1^*	F_2^*	F_1	F_2	Specific variances
HadjTaieb	0.756	0.355	0.516	0.647	0.316
Rodriguez	0.834	0.296	0.497	0.732	0.217
Tversky	0.865	0.185	0.397	0.790	0.218
XSimilarity	0.911	0.226	0.449	0.824	0.119
wnet.E	0.630	0.086	0.243	0.588	0.596
wiki.E	0.178	0.354	0.387	0.083	0.843
word2vec	0.216	0.974	0.997	-0.037	0.005
GloVe	0.13	0.759	0.767	-0.067	0.407
Cumulative Variance	0.365	0.598	0.333	0.66	p-value = 0.157

Table 5.3: Factor Analysis of SimLex-666

Variable	F_1^*	F_2^*	F_1	F_2	Specific variances
HadjTaieb	0.695	0.319	0.665	0.378	0.415
Rodriguez	0.843	0.204	0.658	0.566	0.247
Tversky	0.825	0.134	0.59	0.592	0.302
XSimilarity	0.868	0.265	0.722	0.55	0.177
wnet.E	0.628	0.169	0.504	0.411	0.577
wiki.E	0.099	0.406	0.388	-0.156	0.825
word2vec	0.297	0.829	0.847	-0.243	0.224
GloVe	0.233	0.938	0.897	-0.359	0.066
Cumulative Variance	0.397	0.646	0.459	0.646	p-value = 0.18

Depending on the feeding data our word-similarity method generate disparate outcomes. For instance, as WordNet is a collection of well definite knowledge, the measurement of our word-similarity method inclines to those methods of knowledge-based approach, whereas the category information of Wikipedia as folksonomy is more close to the information the corpus-based approach attempts to extract. This result suggests that two approaches assess word-similarity from distinctive perspectives.

5.3.3 Liquid Association Analysis

Hitherto our analysis suggests both the knowledge-based approach and the corpus-based approach indeed reflect distinctive human perceptions. To provide a comprehensive view of the comparison, this section is to explore what word-pairs two approaches yield consistent similarity scores. A simple scatter plot might be useful for finding those word pairs. However, the simple scatter plot cannot tell on what conditions the word-pairs are judged consistently or inconsistently by different approaches. Therefore, we propose to use Liquid Association in conjunction with our word-similarity method to detect those conditions.

As our word-similarity method yields a probability function for a word pair, we can obtain a cumulative distribution function, $F(i) = P(\text{similarity score} \leq i)$, where $i = \{0, 1, 2, 3, \ldots, 10\}$. A low similarity score with high probability, such as large $F(2)$, indicates an weak association between words. To apply LA, we set the measurements of GloVe as X, the measurements of HadjTaieb's method as Y, and $F(2)$ as Z to draw LA plots as shown in Figure 5.4. Here we fed our word-similarity method with the category information of Wikipedia based on the observation from Factor Analysis.

The LA plot shows that the correlation between two measurements is positive for the word-pairs with lower $F(2)$. For word-pairs with higher $F(2)$, both methods come up with inconsistent scores in WordSim-353. We compare the word-pairs with lower $F(2)$ and higher $F(2)$ based on Figure 5.4. We found that the word-pairs with higher $F(2)$ have larger variance of human rating. That is, two approaches tend to disagree with each other for word-pairs that may lead higher disagreement.

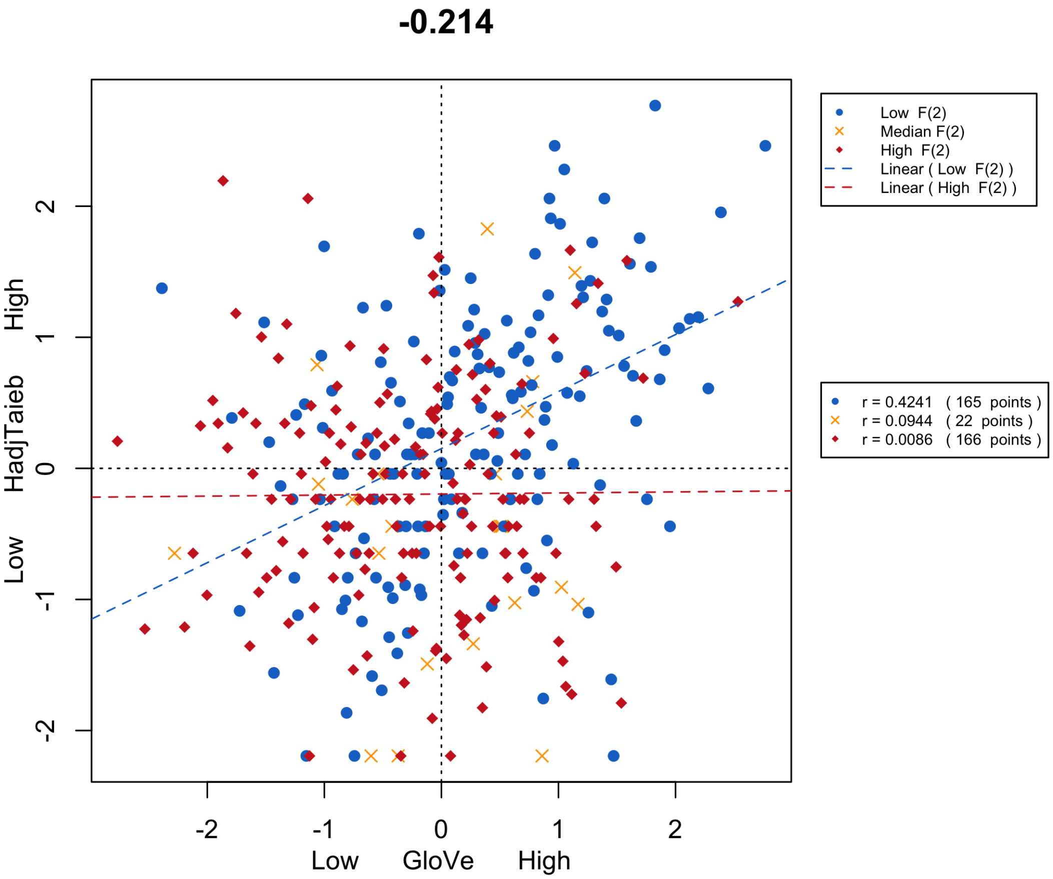


Figure 5.4: Liquid Association Plot for Three Measurements to WordSim-353.

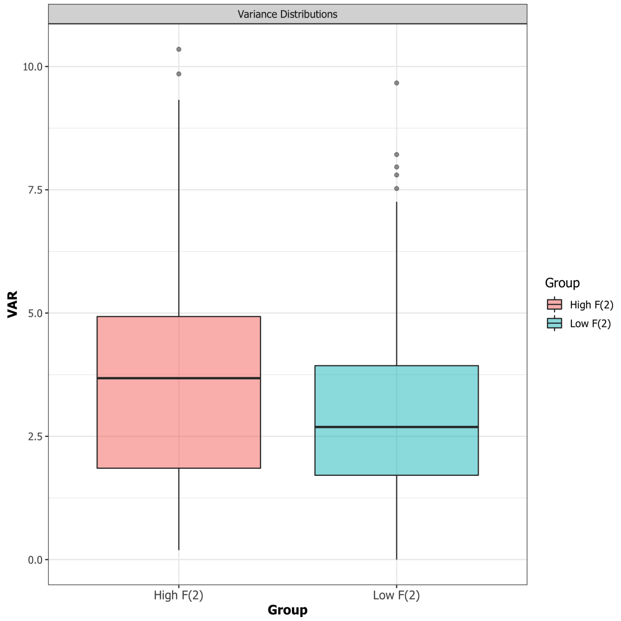


Figure 5.5: Variance Distributions of high $F(2)$ and low $F(2)$.

5.4 Discussion

In this chapter, we use partial R^2s to explore what percentage of variations two approaches can explain. Except for the portion of unexplained variation, the knowledge-based approach works better in SimLex666, the test collection for measuring the performance of word-similarity methods in semantic similarity, whereas the corpus-based approach does better in WordSim353, the test collection for measuring the performance in word relativity. Both approaches might reflect distinctive human perceptions in the judgment of word-similarity.

Through factor analysis, two factor model show two approaches share a common factor, *common sense* factor, which is dominated by the corpus-based approach. Additionally, a clear cut between two approaches are observed on the second loading factor. This result suggests that methods assess word-similarity differently by the approaches they adopt.

Since the knowledge-based approach and the corpus-based approach, in our analysis, reflect distinctive human perceptions in the judgment of word-similarity, we conclude that human raters employ three distinctive perceptions: *common sense*, *definite knowledge* and *sentiment* to judge word-similarity. As what Figure 5.6 shows, a human rater may use three perceptions to judge word-similarity, yet there might be a small portion of overlap between definite knowledge and common sense (a gray area).

Although two approaches reflect distinctive human perceptions, we use LA in conjunction with our folksonomy word-similarity method could detect what word-pairs are judged consistently or inconsistently by the knowledge-based approach and the corpus-based approach. Therefore, our proposed method can effectively measure the performance of word-similarity methods.

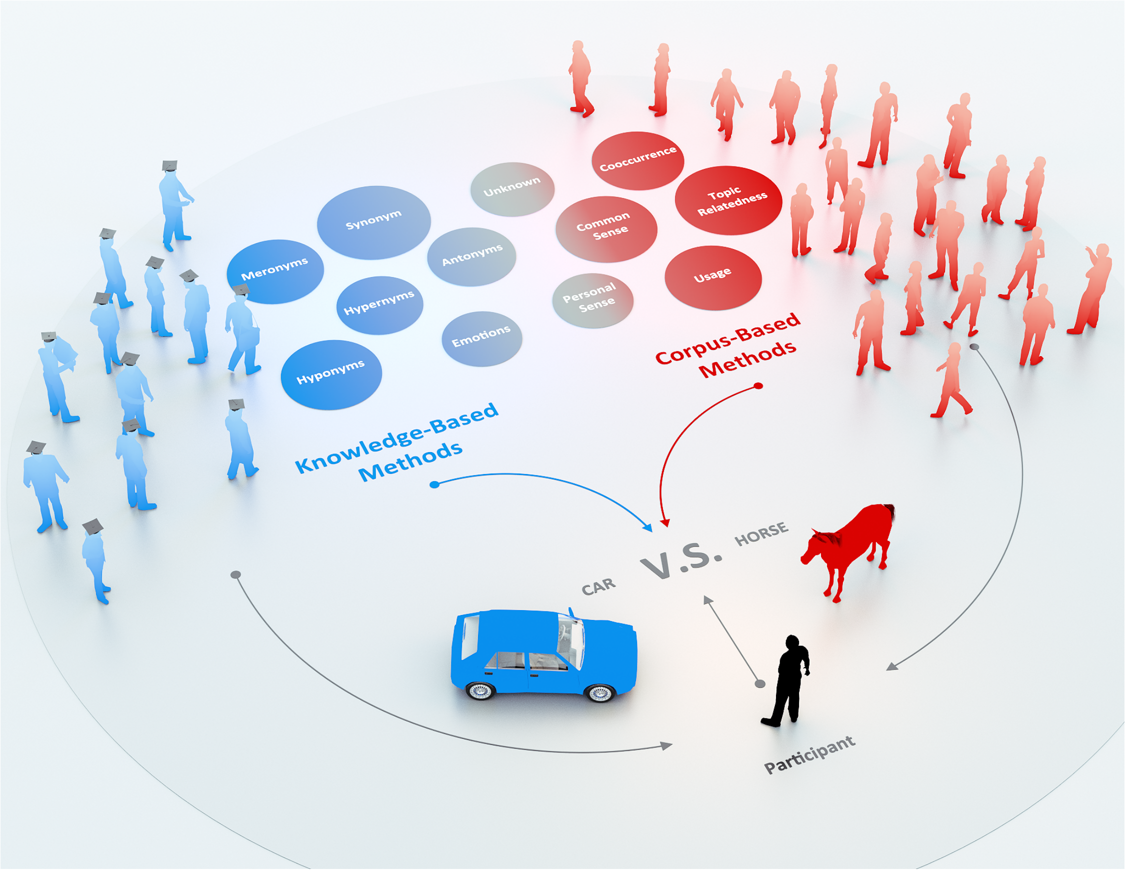


Figure 5.6: Three distinctive perceptions: definite knowledge (blue circles), common sense (red circles), and sentiment (gray circles). When a participant (human rater) assess the comparison between CAR and HORSE, one may use knowledge-based and corpus-based approaches to reflect human judgments.

REFERENCES

[02] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. "Placing Search in Context: The Concept Revisited." *ACM Trans. Inf. Syst.*, **20**(1):116–131, January 2002.

[ABK07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "DBpedia: A Nucleus for a Web of Open Data." In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pp. 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[AE09] Aniket Kittur and Ed H. Chi, Bongwon Suh. "What's in Wikipedia?: mapping topics and conflict using socially annotated category structure." In *The SIGCHI Conference on Human Factors in Computing Systems*, pp. 1509–1512, 2009.

[AKU13] Igor Akushevich, Julia Kravchenko, Svetlana Ukraintseva, Konstantin Arbeev, Alexander Kulminski, and Anatoliy I Yashin. "Morbidity risks among older adults with pre-existing age-related diseases." *Experimental gerontology*, **48**(12):1395–1401, December 2013.

[AM03] Andrea Rodriguez and Max J Egenhofer. "Determining Semantic Similarity among Entity Classes from Different Ontologies." *IEEE Transactions on Knowledge and Data Engineering*, **15(2)**:442–456, April 2003.

[Amo77] Amos Tversky. "Features of Similarity." *Psycological Review*, **84(4)**:327–352, July 1977.

[AYR16] Syed A Ali, Ning Yin, Arkam Rehman, and Verline Justilien. "Parkinson Disease-Mediated Gastrointestinal Disorders and Rational for Combinatorial Therapies." *Medical Sciences*, **4**(1):1, March 2016.

[BBB12] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. "Distributional Semantics in Technicolor." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pp. 136–145, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[BRK14] Simon Baker, Roi Reichart, and Anna Korhonen. "An Unsupervised Model for Instance Level Subcategorization Acquisition." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 278–289, Doha, Qatar, 2014. Association for Computational Linguistics.

[BYC05] Andrea H. Bild, Guang Yao, Jeffrey T. Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M. Lancaster, Andrew Berchuck, John A. Olson Jr, Jeffrey R. Marks, Holly K. Dressman, Mike West, and Joseph R. Nevins. "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." *Nature*, **439**:353, November 2005.

[CEM13] Mirella Cacace, Stefanie Ettelt, Nicholas Mays, and Ellen Nolte. “Assessing quality in cross-country comparisons of health systems and policies: Towards a set of generic quality criteria.” *Health System Performance Comparison: New Directions in Research and Policy*, **112**(1):156–162, September 2013.

[Chr15] Christopher D. Manning. “Computational Linguistics and Deep Learning.” *Computational Linguistics*, **41(4)**:701–707, December 2015.

[CPC17] Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. “SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity.” In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 15–26, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[DBA12] Jane A Driver, Alexa Beiser, Rhoda Au, Bernard E Kreger, Greta Lee Splansky, Tobias Kurth, Douglas P Kiel, Kun Ping Lu, Sudha Seshadri, and Phillip A Wolf. “Inverse association between cancer and Alzheimer’s disease: results from the Framingham Heart Study.” *The BMJ*, **344**:e1442, 2012.

[DCC17] Joseph Dieleman, Madeline Campbell, Abigail Chapin, Erika Eldrenkamp, Victoria Y Fan, Annie Haakenstad, Jennifer Kates, Yingying Liu, Taylor Matyasz, Angela Micah, Alex Reynolds, Nafis Sadat, Matthew T Schneider, Reed Sorensen, Tim Evans, David Evans, Christoph Kurowski, Ajay Tandon, Kaja M Abbas, Semaw Ferede Abera, Aliasghar Ahmad Kiadaliri, Kedir Yimam Ahmed, Muktar Beshir Ahmed, Khurshid Alam, Reza Alizadeh-Navaei, Ala’a Alkerwi, Erfan Amini, Walid Ammar, Stephen Marc Amrock, Carl Abelardo T Antonio, Tesfay Mehari Atey, Leticia Avila-Burgos, Ashish Awasthi, Aleksandra Barac, Oscar Alberto Bernal, Addisu Shunu Beyene, Tariku Jibat Beyene, Charles Birungi, Habtamu Mellie Bizuayehu, Nicholas J K Breitborde, Lucero Cahuana-Hurtado, Ruben Estanislao Castro, Ferran Catalia-Lopez, Koustuv Dalal, Lalit Dandona, Rakhi Dandona, Pieter de Jager, Samath D Dharmaratne, Manisha Dubey, Carla Sofia e Sa Farinha, Andre Faro, Andrea B Feigl, Florian Fischer, Joseph Robert Anderson Fitchett, Nataliya Foigt, Ababi Zergaw Giref, Rahul Gupta, Samer Hamidi, Hilda L Harb, Simon I Hay, Delia Hendrie, Masako Horino, Mikk Jürisson, Mihajlo B Jakovljevic, Mehdi Javanbakht, Denny John, Jost B Jonas, Seyed M. Karimi, Young-Ho Khang, Jagdish Khubchandani, Yun Jin Kim, Jonas M Kinge, Kristopher J Krohn, G Anil Kumar, Hassan Magdy Abd El Razek, Mohammed Magdy Abd El Razek, Azeem Majeed, Reza Malekzadeh, Felix Masiye, Toni Meier, Atte Meretoja, Ted R Miller, Erkin M Mirrakhimov, Shafiu Mohammed, Vinay Nangia, Stefano Olgiati, Abdalla Sidahmed Osman, Mayowa O Owolabi, Tejas Patel, Angel J Paternina Caicedo, David M Pereira, Julian Perelman, Suzanne Polinder, Anwar Rafay, Vafa Rahimi-Movaghar, Rajesh Kumar Rai, Usha Ram, Chhabi Lal Ranabhat, Hirbo Shore Roba, Joseph Salama, Miloje Savic, Sadaf G Sepanlou, Mark G Shrime, Roberto Tchio Talongwa, Braden J Te Ao, Fabrizio Tediosi, Azeb Gebresilassie Tesema, Alan J Thomson, Ruoyan Tobe-Gai, Roman Topor-Madry, Eduardo A Undurraga, Tommi

Vasankari, Francesco S Violante, Andrea Werdecker, Tissa Wijeratne, Gelin Xu, Naohiro Yonemoto, Mustafa Z Younis, Chuanhua Yu, Zoubida Zaidi, Maysaa El Sayed Zaki, and Christopher J L Murray. "Evolution and patterns of global health financing 1995–2014: development assistance for health, and government, prepaid private, and out-of-pocket health spending in 184 countries." *The Lancet*, **389**(10083):1981–2004, May 2017.

[EDN95] J Elola, A Daponte, and V Navarro. "Health indicators and the organization of health care systems in western Europe." *American Journal of Public Health*, **85**(10):1397–1401, October 1995.

[EGA06] Euripides G.M. Petrakis, Giannis Varelas, Angelos Hliaoutakis, and Paraskevi Raftopoulou. "X-Similarity: Computing Semantic Similarity between concepts from different ontologies." *Journal of Digital Information Management*, **4**(4):233–237, 2006.

[FFP15] Bianca K. Frogner, H.E. Frech, and Stephen T. Parente. "Comparing efficiency of health systems across industrialized countries: a panel analysis." *BMC Health Services Research*, **15**(1):415, September 2015.

[FGM01] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. "Placing Search in Context: The Concept Revisited." In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pp. 406–414, New York, NY, USA, 2001. ACM.

[Fir57] Firt, J. R. "A Synopsis of Linguistic Theory 1930-55." *Studies in Linguistic Analysis(special volume of the Philological Society)*, pp. 1–32, 1957.

[FMD04] Elspeth C Ferguson, Ravi Maheswaran, and Mark Daly. "Road-traffic pollution and asthma – using modelled exposure assessment for routine public health surveillance." *International Journal of Health Geographics*, **3**:24–24, 2004.

[FRA14] Felix Hill, Roi Reichart, and Anna Korhonen. "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation." Technical report, eprint arXiv:1408.3456, 2014.

[FTR16] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. "Problems With Evaluation of Word Embeddings Using Word Similarity Tasks." *ArXiv e-prints*, May 2016.

[FVL15] Alfonso Fasano, Naomi P. Visanji, Louis W C Liu, Antony E. Lang, and Ronald F. Pfeiffer. "Gastrointestinal dysfunction in Parkinson's disease." *The Lancet Neurology*, **14**(6):625–639, June 2015.

[FWC16] Daryl Michal Freedman, Jincao Wu, Honglei Chen, Ralph W Kuncl, Lindsey R Enewold, Eric A Engels, Neal D Freedman, and Ruth M Pfeiffer. "Associations between cancer and Alzheimer's disease in a U.S. Medicare population." *Cancer Medicine*, **5**(10):2965–2976, October 2016.

[Geo95] George A. Miller. “WordNet: a lexical database for English.” *Communications of the ACM*, **38**(11):39–41, November 1995.

[HBB14] Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, and Abdelmajid Ben Hamadou. “Ontology-based Approach for Measuring Semantic Similarity.” *Eng. Appl. Artif. Intell.*, **36**(C):238–261, November 2014.

[HDG12] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. “Large-scale Learning of Word Relatedness with Constraints.” In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’12, pp. 1406–1414, New York, NY, USA, 2012. ACM.

[HRJ15] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. “Semantic Similarity from Natural Language and Ontology Analysis.” *Synthesis Lectures on Human Language Technologies*, **8**(1):1–254, May 2015.

[JVH04] Michael H Jones, Carl Virtanen, Daisuke Honjoh, Tatsu Miyoshi, Yukitoshi Satoh, Sakae Okumura, Ken Nakagawa, Hitoshi Nomura, and Yuichi Ishikawa. “Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles.” *The Lancet*, **363**(9411):775–781, March 2004.

[KL13] Tae Kuen Kim and Shannon R. Lane. “Government Health Expenditure and Public Health Outcomes: A Comparative Study among 17 Countries and Implications for US Health Care Reform.” *American International Journal of Contemporary Research*, **3(9)**:8–13, 2013.

[Lan95] Ken Lang. “Newsweeder: Learning to filter netnews.” In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.

[Li02] Ker-Chau Li. “Genome-wide coexpression dynamics: Theory and application.” *Proceedings of the National Academy of Sciences*, **99**(26):16875–16880, 2002.

[LPY07] Ker-Chau Li, Aarno Palotie, Shinsheng Yuan, Denis Bronnikov, Daniel Chen, Xuelian Wei, Oi-Wa Choi, Janna Saarela, and Leena Peltonen. “Finding disease candidate genes by liquid association.” *Genome Biology*, **8**(10):R205–R205, 2007.

[LSM13] Minh-Thang Luong, Richard Socher, and Christopher D. Manning. “Better Word Representations with Recursive Neural Networks for Morphology.” In *CoNLL*, Sofia, Bulgaria, 2013.

[Lut16] Lutz Sager. “Estimating the effect of air pollution on road safety using atmospheric temperature.” Technical report, Grantham Research Institute on Climate Change and the Environment, 2016.

[LY04] Ker-Chau Li and Shinsheng Yuan. “A functional genomic study on NCI’s anticancer drug screen.” *The Pharmacogenomics Journal*, **4**:127, March 2004.

[MAD13] Massimo Musicco, Fulvio Adorni, Simona Di Santo, Federica Prinelli, Carla Pettenati, Carlo Caltagirone, Katie Palmer, and Antonio Russo. "Inverse occurrence of cancer and Alzheimer disease." *Neurology*, **81**(4):322, July 2013.

[MC91] George A. Miller and Walter G. Charles. "Contextual correlates of semantic similarity." *Language and Cognitive Processes*, **6**(1):1–28, January 1991.

[MF17] Pier Mannuccio Mannucci and Massimo Franchini. "Health Effects of Ambient Air Pollution in Developing Countries." *International Journal of Environmental Research and Public Health*, **14**(9):1048, September 2017.

[NAS16] Nilupulee Nathawitharana, Damminda Alahakoon, and Daswin De Silva. "Using semantic relatedness measures with dynamic self-organizing maps for improved text clustering." *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2662–2671, 2016.

[OLH13] S.-M. Ou, Y.-J. Lee, Y.-W. Hu, C.-J. Liu, T.-J. Chen, J.-L. Fuh, and S.-J. Wang. "Does Alzheimer's Disease Protect against Cancers? A Nationwide Population-Based Study." *Neuroepidemiology*, **40**(1):42–49, 2013.

[PAC16] Andrée-Anne Poirier, Benoit Aubé, Mélissa Côté, Nicolas Morin, Thérèse Di Paolo, and Denis Soulet. "Gastrointestinal Dysfunctions in Parkinson's Disease: Symptoms and Treatments." *Parkinson's Disease*, **2016**:6762528, 2016.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[RAG11] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. "A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis." In *Proceedings of the 20th International World Wide Web Conference*, pp. 337–346, Hyderabad, India, March 2011.

[RBX05] C. M. Roe, M. I. Behrens, C. Xiong, J. P. Miller, and J. C. Morris. "Alzheimer disease and cancer." *Neurology*, **64**(5):895, March 2005.

[RCA12] Sabrina Realmuto, Antonio Cinturino, Valentina Arnao, Maria Mazzola, Chiara Cupidi, Paolo Aridon, Paolo Ragonese, Giovanni Savettieri, and Marco D'Amelio. "Tumor Diagnosis Preceding Alzheimer's Disease Onset: Is There a Link Between Cancer and Alzheimer's Disease?" *Journal of Alzheimer's Disease*, **31**:177–182, July 2012.

[RFX10] C. M. Roe, A. L. Fitzpatrick, C. Xiong, W. Sieh, L. Kuller, J. P. Miller, M. M. Williams, R. Kopan, M. I. Behrens, and J. C. Morris. "Cancer linked to Alzheimer disease but not vascular dementia." *Neurology*, **74**(2):106, January 2010.

[RG65] Herbert Rubenstein and John B. Goodenough. "Contextual Correlates of Synonymy." *Commun. ACM*, **8**(10):627–633, October 1965.

[Rot07] Roth R. “Human body index - transcriptional profiling.”, 2007.

[SG12] Nadella Sandhya and A. Govardhan. “Analysis of Similarity Measures with WordNet Based Text Document Clustering.” In Suresh Chandra Satapathy, P. S. Avadhani, and Ajith Abraham, editors, *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012*, pp. 703–714. Springer Berlin Heidelberg, 2012.

[SM86] Gerard Salton and Michael J McGill. “Introduction to modern information retrieval.” 1986.

[STE08] Kerby Shedden, Jeremy MG Taylor, Steve A Enkemann, Ming S Tsao, Timothy J Yeatman, William L Gerald, Steve Eschrich, Igor Jurisica, Seshan E Venkatraman, Matthew Meyerson, Rork Kuick, Kevin K Dobbin, Tracy Lively, James W Jacobson, David G Beer, Thomas J Giordano, David E Misek, Andrew C Chang, Chang Qi Zhu, Dan Strumpf, Samir Hanash, Francis A Shepherd, Kuyue Ding, Lesley Seymour, Katsuhiko Naoki, Nathan Pennell, Barbara Weir, Roel Verhaak, Christine Ladd-Acosta, Todd Golub, Mike Gruidl, Janos Szoke, Maureen Zakowski, Valerie Rusch, Mark Kris, Agnes Viale, Noriko Motoi, William Travis, and Anupama Sharma. “Gene Expression-Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study: Director’s Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma.” *Nature medicine*, **14**(8):822–827, August 2008.

[STL15] Hai-bin Shi, Bo Tang, Yao-Wen Liu, Xue-Feng Wang, and Guo-Jun Chen. “Alzheimer disease and cancer risk: a meta-analysis.” *Journal of Cancer Research and Clinical Oncology*, **141**(3):485–494, March 2015.

[SVK09] Uma T Shankavaram, Sudhir Varma, David Kane, Margot Sunshine, Krishna K Chary, William C Reinhold, Yves Pommier, and John N Weinstein. “CellMiner: a relational database and query tool for the NCI-60 cancer cell lines.” *BMC Genomics*, **10**:277–277, 2009.

[SWB04] Andrew I. Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P. Cooke, John R. Walker, and John B. Hogenesch. “A gene atlas of the mouse and human protein-encoding transcriptomes.” *Proceedings of the National Academy of Sciences*, **101**(16):6062–6067, 2004.

[SYL08] Wei Sun, Shinsheng Yuan, and Ker-Chau Li. “Trait-trait dynamic interaction: 2D-trait eQTL mapping for genetic variation study.” *BMC Genomics*, **9**(1):242, May 2008.

[TBK06] T. Slimani, B. Ben Yaghlane, and K. Mellouli. “A New Similarity Measure based on Edge Counting.” In *World Academy of Science, Engineering and Technology*, volume 17, pp. 232–236, December 2006.

[TKG13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." In *Workshop at International Conference on Learning Representations*, January 2013.

[TPD98] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. "An Introduction to Latent Semantic Analysis." *Discourse Processes*, **25**:259–284, 1998.

[TWY10] Shang-Kai Tai, Guanl Wu, Shinsheng Yuan, and Ker-Chau Li. "Genome-wide expression links the electron transfer pathway of Shewanella oneidensis to chemotaxis." *BMC Genomics*, **11**(1):319, May 2010.

[WLC15] Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. "A semantic approach for text clustering using WordNet and lexical chains." *Expert Systems with Applications*, **42**(4):2264–2275, March 2015.

[Wor17] World Health Organization. "Global Health Observatory (GHO) data.", 2017.

[WSY08] Tongtong Wu, Wei Sun, Shinsheng Yuan, Chun-Houh Chen, and Ker-Chau Li. "A method for analyzing censored survival phenotype with gene expression data." *BMC Bioinformatics*, **9**(1):417, October 2008.

[YP06] Dongqiang Yang and David M. W. Powers. "Verb Similarity on the Taxonomy of Wordnet." In *In the 3rd International WordNet Conference (GWC-06), Jeju Island, Korea*, 2006.

[YRP03] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. "A neural probabilistic language model." *The Journal of Machine Learning Research*, **3**:1137–1155, 2003.

[Yua03] Shinsheng Yuan. *Some Contributions in Computational Biology: Medical Imaging and Gene Expression.* PhD thesis, University of California, Los Angeles, 2003.

[YZD03] Yuhua Li, Zuhair A. Bandar, and David McLean. "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources." *IEEE Transactions on Knowledge and Data Engineering*, **15(4)**:871–882, August 2003.

[ZM94] Zhibiao Wu and Martha Palmer. "Verbs semantics and lexical selection." In *ACL 94 Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics Stroudsburg, June 1994.