The Word-Category RIG Matrix

April 2020 by Neslihan Suzen, PhD student at the University of Leicester (ns433@leicester.ac.uk / suzenneslihan@hotmail.com)

Supervised by Prof Alexander Gorban and Dr Evgeny Mirkes

Getting Started

This file describes the **Word-Category RIG Matrix** for the Leicester Scientific Corpus (LSC) [1], the procedure to build the matrix and introduces the Leicester Scientific Thesaurus (LScT) with the construction process. The Word-Category RIG Matrix is a 103,998 by 252 matrix, where rows correspond to words of Leicester Scientific Dictionary-Core (LScDC) [2] and columns correspond to 252 Web of Science (WoS) categories [3, 4, 5]. Each entry in the matrix corresponds to a pair *(category,word).* Its value for the pair (c_k, w_j) shows the Relative Information Gain (RIG) on the belonging of a text from LSC to the category c_k from observing the word w_j in this text. The CSV file of Word-Category RIG Matrix in the published archive is presented with two additional columns of the sum of RIGs in categories and the maximum of RIGs over categories (last two columns of the matrix). So, the file 'Word-Category RIG Matrix.csv' contains a total of 254 columns.

This matrix is created to be used in future research on quantifying of meaning in scientific texts under the assumption that words have scientifically specific meanings in subject categories and the meaning can be estimated by information gains from word to categories.

LScT (Leicester Scientific Thesaurus) is a scientific thesaurus of English. The thesaurus includes a list of 5,000 words from LScDC. We consider ordering the words of LScDC by the sum of their RIGs in categories. That is, words are arranged in their informativeness in the scientific corpus LSC. Therefore, meaningfulness of words is evaluated by words' average informativeness in the categories. We have decided to include the most informative 5,000 words in the scientific thesaurus.

Words as a Vector of Frequencies in WoS Categories

Each word of LScDC is represented as a vector of frequencies in WoS categories. Given the collection of LSC texts, each entry of the vector consists of the number of texts containing the word in the corresponding category.

It is noteworthy that texts in a corpus do not necessarily belong to a single category, as they are likely to correspond to multidisciplinary studies, specifically in a corpus of scientific texts. In other words, categories may not be exclusive. There are 252 WoS categories and a text can be assigned to at least 1 and at most 6 categories in the LSC.

Using the binary calculation of frequencies, we introduce the presence of w_j in texts of the category c_k. Let D_k be a set of texts in the category c_k. We create a vector of frequencies for each word, namely $\overrightarrow{w_j}$ where dimensions are categories in the corpus. We denote the binary frequency by b_{ij}. If the word w_j occurs in the text once or more time, $b_{ij} = 1$. Otherwise, $b_{ij} = 0$. Each component of the vector is defined by

$$w_{jk} = \sum_{d_i \in D_k} b_{ij}$$

where d_i is a text in the corpus and i is index for texts in corpus. In other words, w_{jk} is the number of texts containing the word w_j in the category c_k. Therefore, each word in the corpus is represented as a vector of binary frequencies denoted by

$$\overrightarrow{w_j} = (\mathrm{w}_{j1}, \mathrm{w}_{j2}, \ldots, \mathrm{w}_{jK})$$

where K is the number of categories in the corpus. The collection of vectors, with all words and categories in the entire corpus, can be shown in a table, where each entry corresponds to a word and a category (see Table 1).

Table 1 The structure of dictionary representation by frequencies w_{jk}

Category / Word	c_1	c_2	…	c_K
w_1	w_{11}	w_{12}	…	w_{1K}
w_2	w_{21}	w_{22}	…	w_{2K}
⋮	⋮	⋮		⋮
w_{N}	w_{N1}	w_{N2}	…	w_{NK}

This table is build for the LScDC with 252 WoS categories and presented in published archive with this file. The value of each entry in the table shows how many times a word of the LScDC appears in a WoS category. The occurrence of a word in a category is determined by counting the number of the LSC texts containing the word in a category.

Words as a Vector of Relative Information Gains Extracted for Categories

In this section, we introduce our approach to representation of a word as a vector of relative information gains for categories under the assumption that meaning of a word can be quantified by their information gained for categories.

For each category, c_k, a function is defined on texts that takes the value 1, if the text belongs to the category c_k, and 0 otherwise. For each word, w_j, a function is defined on texts that takes the value 1 if the word w_j belongs to the text, and 0 otherwise. We use for these functions the same notations c_k and w_j. Consider the LSC as a probabilistic sample space (the space of equally probable elementary outcomes). For the Boolean random variables, c_k and w_j, the joint probability distribution, the entropy and information gains can be defined as follows.

The information gain about the category c_k from the word w_j, $IG(c_k, w_j)$, is the amount of information on the belonging of a text from the LSC to the category c_k from observing the word w_j in the text. It can be calculated as [6]:

$$IG(c_k, w_j) = H(c_k) - H(c_k|w_j)$$

where $H(c_k)$ is the Shannon entropy of c_k and $H(c_k|w_j)$ is the conditional entropy of c_k given the observing of the word w_j. Entropies $H(c_k)$ and $H(c_k|w_j)$ are

$$H(c_k) = -P(c_k)\log_2 P(c_k) - P(\overline{c_k})\log_2 P(\overline{c_k})$$

where $P(c_k)$ is the probability that the text belongs to the category c_k, $P(\overline{c_k})$ is the probability that text does not belong to the category c_k and

$$\begin{aligned} H(c_k|w_j) = P(w_j)\left[-P(c_k|w_j)\log_2 P(c_k|w_j) - P(\overline{c_k}|w_j)\log_2 P(\overline{c_k}|w_j)\right] \\ + P(\overline{w_j})\left[-P(c_k|\overline{w_j})\log_2 P(c_k|\overline{w_j}) - P(\overline{c_k}|\overline{w_j})\log_2 P(\overline{c_k}|\overline{w_j})\right] \end{aligned}$$

where $P(w_j)$ is the probability that the word w_j appears in a text from the corpus, $P(\overline{w_j})$ is the probability that the word w_j does not appear in a text from the corpus; $P(c_k|w_j)$ is the probability that a text belongs to the category c_k under the condition that it contains the word w_j, $P(\overline{c_k}|w_j)$ is the probability that a text does not belong to the category c_k under the condition that it contains the word w_j, $P(c_k|\overline{w_j})$ is the probability that a text belongs to the category c_k under the condition that it does not contain the word w_j, and $P(\overline{c_k}|\overline{w_j})$ is the probability that a text does not belong to the category c_k under the condition that it does not contain the word w_j.

We used the Relative Information Gain (RIG) providing a normalised measure of the Information Gain with regard to the entropy of c_k. This provides the ability of comparing information gains for different categories. RIG is defined as

$$RIG(c_k, w_j) = \frac{IG(c_k, w_j)}{H(c_k)}.$$

For simplicity, we denote $RIG(c_k, w_j)$ by RIG_{jk}. Given the word w_j, RIG_{jk} is used to form vector $\overrightarrow{RIG_j}$, where each component of the vector corresponds to a category. Therefore, each word is represented as a vector of relative information gains. It is obvious that the dimension of vector for each word is the number of categories (K). For the word w_j, this vector is $\overrightarrow{RIG_j} = (\mathrm{RIG}_{j1}, \mathrm{RIG}_{j2}, \ldots, \mathrm{RIG}_{jK})$.

The set of $\overrightarrow{RIG_j}$ vectors is used to form the **Word-Category RIG Matrix**, in which each column corresponds to a category c_k and each row corresponds to a word w_j. Each component RIG_{jk} corresponds to a pair (c_k, w_j) and its value is the RIG from the word w_j to the category c_k. The structure of the Word-Category RIG Matrix is demonstrated in Table 2.

Table 2 The structure of the Word-Category RIG Matrix

Word \ Category	c_1	c_2	…	c_K
w_1	RIG_{11}	RIG_{12}	…	RIG_{1K}
w_2	RIG_{21}	RIG_{22}	…	RIG_{2K}
⋮	⋮	⋮		⋮
w_N	RIG_{N1}	RIG_{N2}	…	RIG_{NK}

In Word-Category RIG Matrix, a row vector represents the corresponding word as a vector of RIGs in categories. We note that in the matrix, a column vector represents RIGs of all words in an individual category. If we choose an arbitrary category, words can be ordered by their RIGs from the most informative to the least informative for the category. As well as ordering words in each category, words can be ordered by two criteria: sum and maximum of RIGs in categories. The top n words in this list can be considered as the most informative words in the scientific texts. For a given word w_j, the sum S_j and maximum M_j of RIGs are calculated from the Word-Category RIG Matrix as:

$$S_j = \sum_{k=1}^{K} RIG_{jk}$$

and

$$M_j = \max_{k=1,\dots,K}(RIG_{jk}).$$

RIGs for each word of LScDC in 252 categories are calculated and vectors of words are formed. We then form the Word-Category RIG Matrix for the LSC. The sum S_j and maximum M_j of RIGs in categories are calculated and added at the end of the matrix (last two columns of the matrix). The Word-Category RIG Matrix for LScDC with 252 categories, the sum of RIGs in categories and the maximum of RIGs over categories can be found in the database.

Leicester Scientific Thesaurus (LScT)

Leicester Scientific Thesaurus (LScT) is a list of 5,000 words from the LScDC [2]. Words of LScDC are sorted in descending order by their S_j, and the top 5,000 words are selected to be included in the LScT. We consider these 5,000 words as the most meaningful words in the scientific corpus. In other words, meaningfulness of words evaluated by words' average informativeness in the categories and the list of these words are considered as a 'thesaurus' for science. The LScT with S_j can be found as CSV file with the published archive.

Published archive contains following files:

1. **Word_Category_RIG_Matrix.csv:** A 103,998 by 254 matrix where columns are 252 WoS categories, S_j and M_j (last two columns of the matrix), and rows are words of the LScDC. Each entry in the first 252 columns is RIG from the word to the category. Words are ordered as in the LScDC.
2. **Word_Category_Frequency_Matrix.csv**: A 103,998 by 252 matrix where columns are 252 WoS categories and rows are words of the LScDC. Each entry of the matrix is the number of texts containing the word in the corresponding category. Words are ordered as in the LScDC.
3. **LScT.csv:** List of words of LScT with their S_j values.
4. **Text_No_in_Cat.csv:** The number of texts in categories**.**
5. **Categories_in_Documents.csv:** List of WoS categories for each document of the LSC.
6. **README.txt:** Description of Word-Category RIG Matrix, Word-Category Frequency Matrix and the LScT and forming procedures.
7. **README.pdf** (same as 6 in PDF format)

References

[1] Suzen, Neslihan (2019): LSC (Leicester Scientific Corpus). figshare. Dataset. https://doi.org/10.25392/leicester.data.9449639.v2

[2] Suzen, Neslihan (2019): LScDC (Leicester Scientific Dictionary-Core). figshare. Dataset. https://doi.org/10.25392/leicester.data.9896579.v3

[3] Web of Science. (15 July). Available: https://apps.webofknowledge.com/

[4] WoS Subject Categories. Available: https://images.webofknowledge.com/WOKRS56B5/help/WOS/hp_subject_category_terms_tasca.html

[5] Suzen, N., Mirkes, E. M., & Gorban, A. N. (2019). LScDC-new large scientific dictionary. *arXiv preprint arXiv:1912.06858*.

[6] Shannon, C. E. (1948). A mathematical theory of communication. Bell system technical journal, 27(3), 379-423.