Classifying Documents Using Keyness Values of Words

Maki Miyake

Graduate School of Language and Culture, University of Osaka

Abstract

This paper aims to investigate similarities among documents or words within texts that contain similar themes. A Victorian periodical for women, called the English Woman's Journal is used for the text analysis, and, more specifically, we focus on the Passing Events articles that are included in nearly every issue of the journal. We calculate keyness values of words in order to classify the documents. The ultimate goal of the study is to extract the characteristics of the English Woman's Journal, by combining different useful methods, such as multidimensional analysis and network analysis.

Keywords: Text Mining, Keyness, Cluster Analysis

1. Introduction

Applications of multivariate statistical analyses have been one of the most effective techniques for detecting and capturing the structured patterns of intrinsic contexts within large-scale corpora such as collections of newspaper articles.

In this paper, we investigate the similarities and differences among documents or words within texts that contain similar themes. For the text analysis, we select a large-scale corpus called the Nineteenth-Century Serials Edition (NCSE)[1], which is composed of six nineteenth-century periodicals and newspapers, and is available online for free. More specifically, the Victorian periodical for women called the English Woman's Journal is used for the study. For the particular purpose of extracting the characteristics within similar themes, we focus on the "Passing Events" articles that are included in nearly every issue of the journal.

In addition to the word frequency data for each article, we simultaneously prepare several kinds of datasets calculated by "keyness values" such as Chi-square measure and Term Frequency Inverse Document Frequency. Finally, we apply hierarchical cluster analysis in order to classify the articles. The classical method attempts to form similar clusters based on a matrix of distances among items or variables. The results of the cluster analysis are hierarchically shown as a form of dendrogram.

The statistical results clearly represent similarities and differences among the articles, and the results also highlight the important words as keywords for each article.

1 http://www.ncse.ac.uk/index.html

2. The English Woman's Journal

The English Woman's Journal (EWJ) was published monthly in London for about six and a half years, from March 1858 until August 1864. Its female founders, Barbara Leigh Smith (Bodichon) and Bessie Rayner Parkes (Belloc), issued the miscellaneous women's journal which was written by women. The total number of issues is 78, and each monthly issue of the EWJ consisted of 72 pages that contained about eight articles.

"Passing Events" focused on public affairs. This section was the last to appear in each issue, and was only a few pages long. Its brevity meant that it largely summarized events in note form, and did not provide readers with much analysis. However, it does indicate the importance that the journal placed on what might otherwise appear to be masculine affairs and provided the means for its readers to find out more about them.

Provided as a repository of full-page facsimiles, the electronic edition consists of fully searchable texts generated through Optical Character Recognition (OCR). However, the OCR accuracy is very low due to the poor quality of the original documents. Christodoulakis & Brey (2008) reported the problems of accuracy in textural materials, and attempted to make use of their recently devised algorithm to edit the distance between combinations and splits, to perform approximate pattern matching for OCR texts. Although their algorithm was proved to be very effective method for applying a smaller part of the NCSE, wasn't successfully applied to the entire document which consists of about 100,000 pages.

3. Overview of the Article "Passing Events"

Before discussing datasets for a statistical method, we briefly observe the contents of the English Woman's Journal (EWJ). Every issue consists of 72 pages and contains about 10 articles. For example, "issue No. 43", which was composed from 12 articles, was published on 1st September1861. Table 1 shows the list of the title and the author if it was mentioned for each article. Some articles indicated the author's names in the last line of the article, and other articles, especially in the early issues, didn't. The first article was considered to deal with contemporary political or social issues and was usually written by Parkes, who was one of the founders. The last three articles such as "Notices of Books", "Open Council", and "Passing Events" can be found in nearly every issue of the EWJ. The content of "Passing Events" mainly focuses on public affairs and makes it possible to take a glance at events that English women of those days were interested in.

Id	Title	Author
A01	The Conditions of Working Women in England and France	Bessie R. Parkes
A02	Margaret Beaufort	-
A03	The Institutions of Hofwyl	-
A04	The Victor	E.B.P
A05	Les Feuilles de Saule (a French poem)	Amable Tastu
A06	Algerine Notes	B., M.D.
A07	Women Compositors	Emily Faithfull
A08	Fruits in Their Season	-

A09	National Association for the Promotion of Social Science	-
A10	Notices of Books	-
A11	Open Council	-
A12	Passing Events	-

Table 1 : contents of issue No. 43

As mentioned in the previous section, the electronic texts generated by OCR were very poor quality. Focusing only on the parts of "Passing Events" which are to apply cluster analysis, we manually corrected the OCR's recognition errors comparing to the facsimile formats.

Table 2 shows some basic statistics such as word tokens and types of all issues and an average and standard deviation per issue. The amount of word tokens represents the volume of the texts, while the amount of word types may correspond with the variety of vocabulary. The number of tokens is 110,731 in all issues, and the number of types is 11,173. The greater value of the token's standard deviation indicates the wide range of length within all issues.

	All issues	Average within an issue	Standard Deviation within an issue
Word Tokens	110,731	1,629.2	814.1
Word Types	11,173	702.4	262.9

Table 2 : basic statistics of "Passing Events"

Figure 2 presents the transition in the number of word tokens and types in chronological order. The longest article is issue 59 (published in January 1863) that has 4713 tokens and 325 types, while the shortest article is issue 57 (published in November 1862) that consists of 325 tokens.

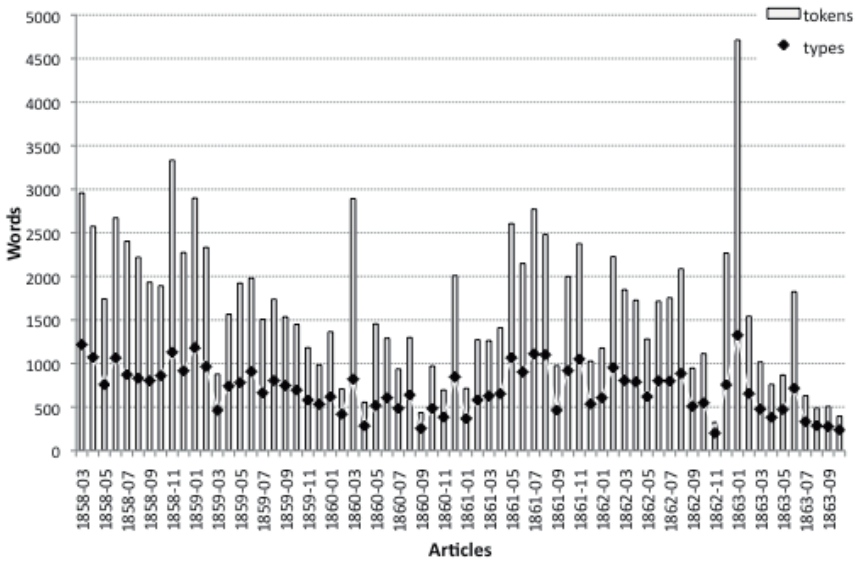

Figure 1 : word tokens and types for each issue

4. Datasets based on Keyness

Instead of using word frequencies for further cluster analysis, we prepared two sorts of datasets that are based on the weighted value of a word's so-called keyness. Keyness is a statistical index used to evaluate how significant a word is to a document. In order to calculate keyness values, the study employed two kinds of statistical measures, chi-square measure and term frequency-inverse document frequency (tf-idf). Both of these measures are widely applied in information retrieval and text mining in order to detect distinctive keywords from a set of documents.

4.1. Chi-square Measure

To detect conspicuous words which appeared in an issue contrasting with other issues, we compare an issue to the rest of the issues. Put another way, we set an issue as a target corpus, while the rest of the issues were set as a reference corpus. For example, computing the chi-square measure for the term "t" in a given issue "i" is needed to construct a contingency table as follows.

	an issue "i"	the rest of the issues	Total
word "t"	a	b	a+b
¬ (word "t")	c	d	c+d
Total	a+c	b+d	N

Table 3 : contingency table

As for the issue "i", the number of occurrences of the word "t" refers to category "a" and the category "c" is the frequency of other word types. In the same way, category "b" means the number of occurrences of the word "t" in all the issues except the issue "i", and category "d" is the frequency of other word types in the rest of the issues. The total number of word types within all the issues is represented as "N" which is equivalent to "a+b+c+d".

And the chi-square measure is calculated by the formulae:

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

We computed the chi-square measure for all word types in every issue. Besides the last four issues, every issue includes a "Passing Events" article. The total number of issues is 68 and the number of words is 11,713 in all. The result was obtained in the form of a word-document frequency matrix of which dimensions are 11,713 words×68 issues.

Since our main focus is the similarity among the issues, the data was limited to common words that occur more than five times in an issue and appears in more than five issues. Finally, cluster analysis was applied to the dataset with the size of 108 words×68 issues.

4.2. Term Frequency-Inverse Document Frequency

We define the tf-idf measure for a given word "t" within the particular issue "i" as follows:

$$tf.idf(t,i) = tf \times \log\frac{N}{df}$$

where, *tf* stands for word frequency in a document, *N* stands for the total number of documents in a given corpus, and finally *df* stands for number of documents where the word t appears.

One of major characteristics of the tf-idf weight is that the more widely and repeatedly a term occurs across documents, the less weighting is given to the term. Thus, it is possible to reduce the impact of high frequency function words such as "a", "the", and "of", and to maximize the discriminative power of low frequency keywords.

As well as the preprocessing of the chi-squared data, we chose the words in the conditions that these tf-idf value have more than 2.6 in each issue and that appear common to more than six issues, as the dataset has approximately the same size of 113 ×68 as the chi-squared dataset.

5. Hierarchical Cluster Analysis

Using the two weighted datasets based on chi-square measure and tf-idf, we applied hierarchical cluster analysis to detect relatively homogeneous clusters of issues. The method starts with each case as a separate cluster and then merges small clusters into larger ones in an agglomerative way.

In the study, we employed the Canberra distance and Ward's method to form the clusters. The Canberra distance (Lance and Williams, 1967) is calculated by the formulae:

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|}$$

As for the results of the dataset based on chi-square measure, the dendrogram for grouping the issues is shown as Figure 2. According to the dendrogram, the four larger clusters out of 68 issues can be formed at a height of around 150.

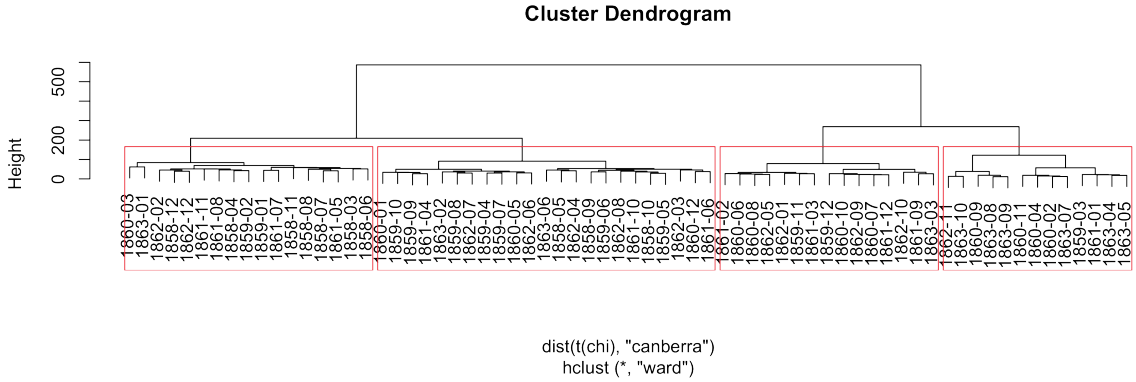

Figure 2 : issue-cluster dendrogram (chi-squered measure)

Figure 3 shows the dendrogram for the words. The branches are cut off at a height of around 150 to form four clusters consisting with the result of the articles.

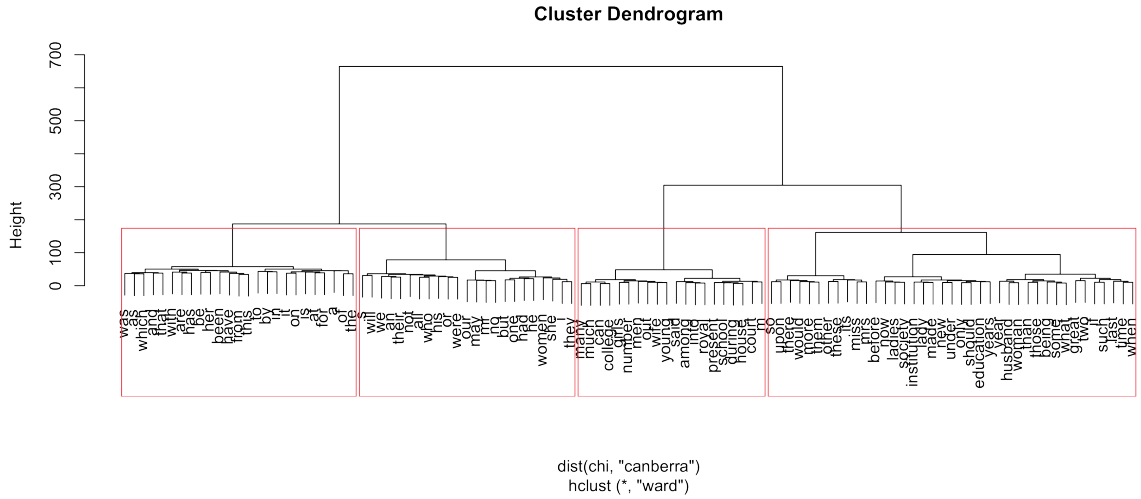

Figure 3 : word-cluster dendrogram (chi-squered measure)

Forcing on the right two clusters as shown in Figure 4, the clusters consists mainly of so-called function words such as “the”, “of”, and “an”. Some of these function words may reflect stylistic similarity between the articles. More particularly, the cluster in right side is composed of many personal pronouns such as “I”, “we”, and “she”. An interesting point here is that the word “women” is included in the personal pronoun’s cluster.

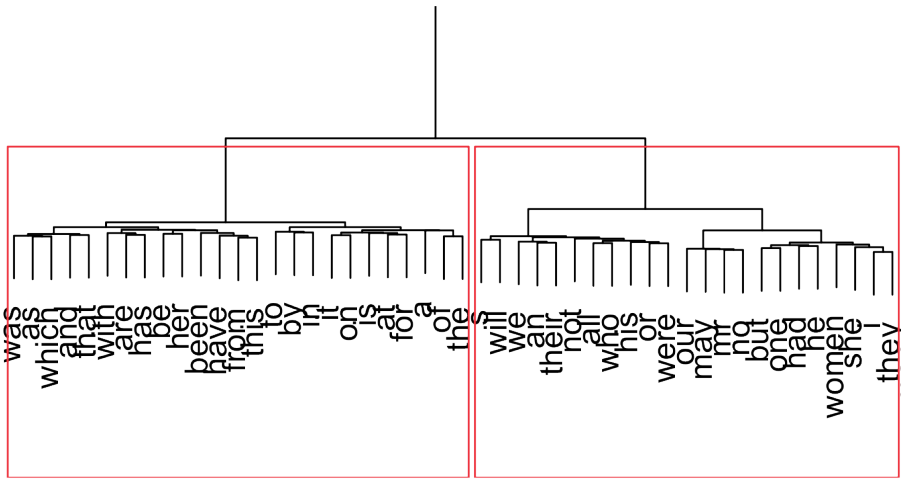

Figure 4 : clusters of function words

The results of tf-idf clustering, in turn, are different from that of chi-squared value. Figure 5 shows the dendrogram for grouping the issues, while the word clustering is shown in Figure 6. As for the issues, the seven smaller clusters can be formed at a height of 140 and the distances among issues can be seen more distinctively.

As the tf-idf tends to estimate a high frequency and common word as less important, the dataset is not included function words appearing in all the issues. The word-clusters, therefore, are classified by the words that could be possible to infer the context.

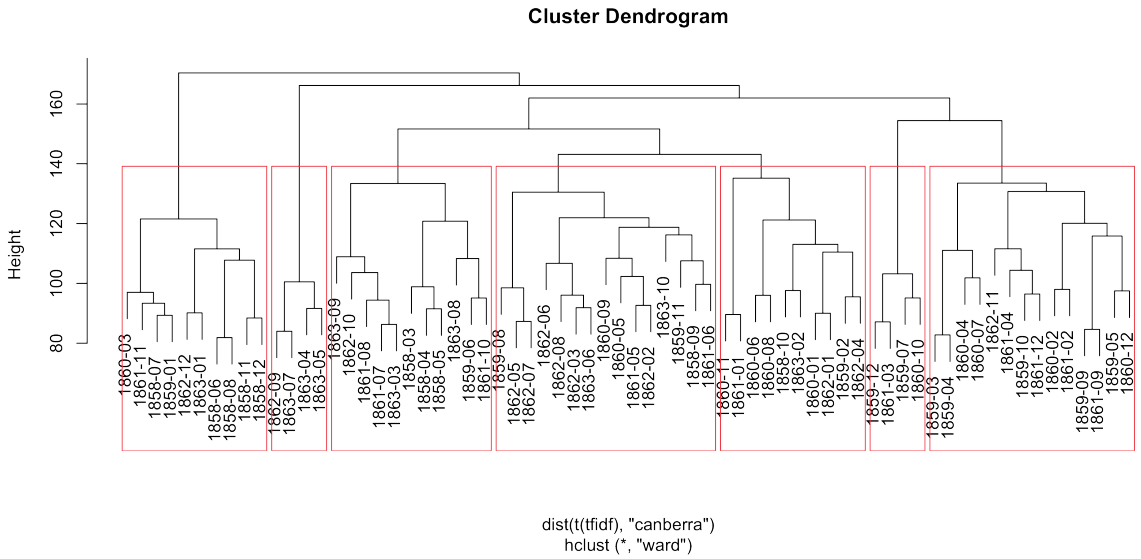

Figure 5 : issue-cluster dendrogram (tf-idf)

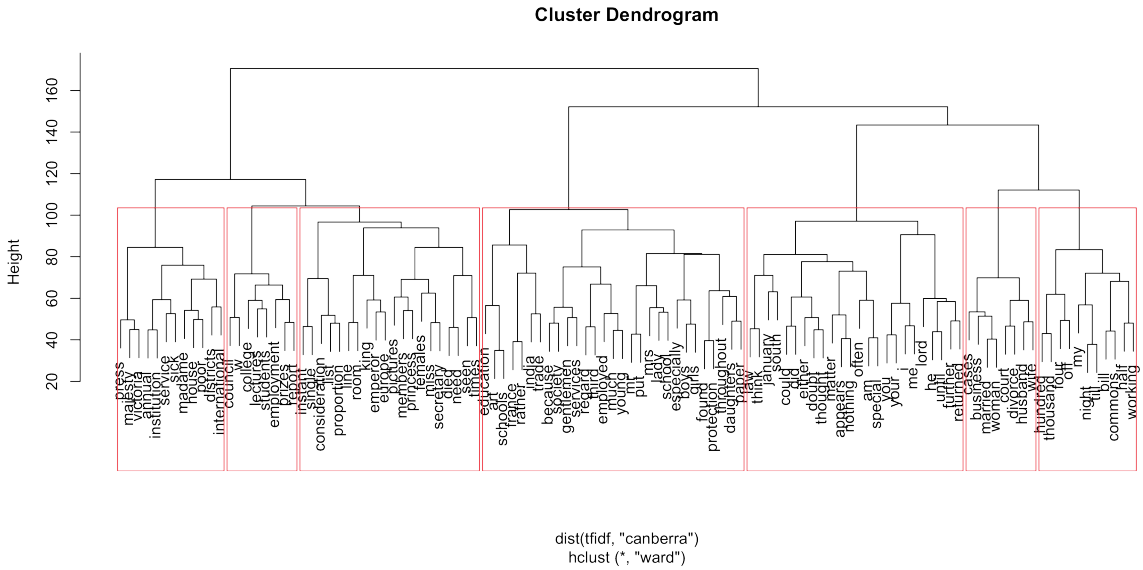

Figure 6: word-cluster dendrogram (tf-idf)

For example, a small word-cluster consists of eights words, i.e. {“cases”, “business”, “married”, “woman”, “court”, “divorce”, “husband”, “wife”} as shown in Figure 7, and the word-cluster corresponds to an issue-cluster that consists of 10 issues shown in Figure 8. In fact, some topics of these articles are concerning the new divorce act and the court.

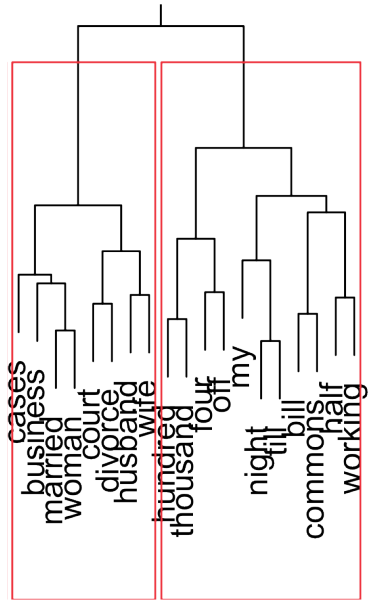

Figure 7 : elements of word-clusters

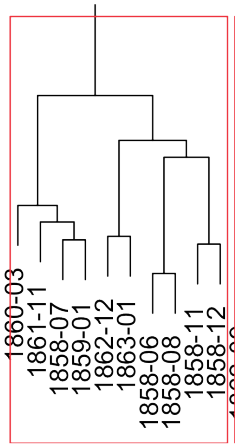

Figure 8 : elements of issue-clusters

6. Conclusions

This paper has reported on the application of hierarchical cluster analysis to investigate the nature of the English Woman's Journal based on similarities among issues and words. Employing two kinds of datasets calculated by "keyness values" of Chi-square measure and Term Frequency Inverse Document Frequency, the statistical results clearly represented similarities and differences among "Passing Events" articles, and the results also indicated the important words as keywords for each article.

Although, the clustering results will be carefully compared to the context of articles, we will continue to combine different useful methods multidimensional analysis and network analysis to elucidate the characteristics of the English Woman's Journal. And the result will be provided as a network representation that is particularly useful in visualizing large-scale linguistic knowledge resources.

References

Alexis Antonia and EllenJordan (2008). Who Wrote the Women's Movement Articles in The Saturday Review?, *Nineteenth-Century Gender Studies*, 4.3.

Christodoulakis, M. and Brey, G. (2008). Edit Distance with Single-Symbol Combinations and Splits, *Proceedings of the Prague Stringology Conference*, pp.208-217.

Lance G.N. and Williams W.T. (1967). Mixed-Data Classificatory Programs I - Agglomerative Systems, *Australian Computer Journal*, 1(1): 15–20.