K-means and Hierarchical Clustering Method to Improve our Understanding of Citation Contexts

Marc Bertin[1] and Iana Atanassova[2]

[1] Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal (UQAM), Canada,
bertin.marc@gmail.com
[2] CRIT-Centre Tesnière, University of Bourgogne Franche-Comté, France,
iana.atanassova@univ-fcomte.fr

Abstract. In this paper we focus of the clustering of citation contexts in scientific papers. We use two methods, k-means and hierarchical clustering to better understand the phenomenon and types of citations and to explore the multidimensional nature of the elements composing the contexts of citations in different sections of the papers. We have analyzed a data set of seven peer-reviewed academic journals published by PLOS. The obtained clusters show that the Methods section is specific in nature, regardless of the journal. A proximity between some of the journals can be observed.

Keywords: In-text References, Bibliometrics, Citation Analysis, IMRaD Structure, Text Mining, K-means, hierarchical clustering

1 Introduction

A lot of research currently takes place around citation contexts in scientific papers. Although these themes are not recent, there is a renewed interest in this field with the implementation of different technics that come from text-mining. The main challenge of these studies is to propose a method to analyze citation contexts at a large scale taking into account various criteria.

We propose a multidimensional approach to this problem which is based on clusters. Clustering algorithms allow us to select similar contexts, that should be considered members of a cluster. This study provides new results around citation contexts and is related to two previous studies on similar problems [3, 2]. This type of approach and techniques have direct applications for the processing of texts from the social sciences (see [8]).

2 Method

We know from previous studies that the rhetorical structure of papers must be taken into account as it plays an important role in determining the types of citation contexts (see e.g. [1]). Furthermore, the specific domains and topics of the various journals, which also have their own editorial lines, can lead to variations and have an effect on the direct context of citations. For this reason, we try to obtain, using a text mining approach, the sets and subsets for determining the existence of different classes of contexts to produce a typology and better understand this issue.

2.1 Dataset

To perform this study we have analyzed a data set of seven peer-reviewed academic journals published in Open Access by the Public Library of Science (PLOS). Six of the journals are domain-specific (*PLOS Biology, PLOS Computational Biology, PLOS Genetics, PLOS Medicine, PLOS Neglected Tropical Diseases and PLOS Pathogens*) and the 7th is *PLOS ONE*, which is a general journal that covers all fields of science and social sciences. We have used for our experiments the entire data set of about 80,000 research articles in full text published up to September 2013.

The data set is in the XML JATS format, where the sections and paragraphs that are identified as distinct XML elements, as well as the in-text references that linked to the corresponding elements in the bibliography of the article. Various aspects of the processing of this corpus and the distributions of in-text references and thier contexts with respect to the IMRaD structure have been the object of previous studies [1, 3].

2.2 Protocol

We have considered the articles in the corpus that are organised following the IMRaD structure (Introduction, Methods, Results and Discussion). As this is part of the editorial requirements of the journals, the vast majority of the papers share this structure. For each journal and for each of the four section types, we have extracted a random sample of 1000 sentences that contain in-text citations. These sentences will be considered as citation contexts in our experiment.

The pre-processing of the corpus consists in removing all punctuation marks and numerical values so as not to introduce bias as to the overrepresentation of the bibliographic references. We also used a stopword list and the terms obtained were stemmatized.

2.3 Hierarchical Clustering and K-means

We analyse citation contexts using a multidimesional approach. We used two complementary approaches, hierarchical clustering and K-means, that allow us to better understand the phenomenon and types of citations and to explore the multidimensional nature of the elements composing the contexts of citations. To obtain the clusters we use the method of K-means (see [5, 6]). From an application point of view, this work is based on the use of an R library [9].

From a methodological point of view, we have used popular partitioning method: K-means clustering. However, this approach requires to know the exact

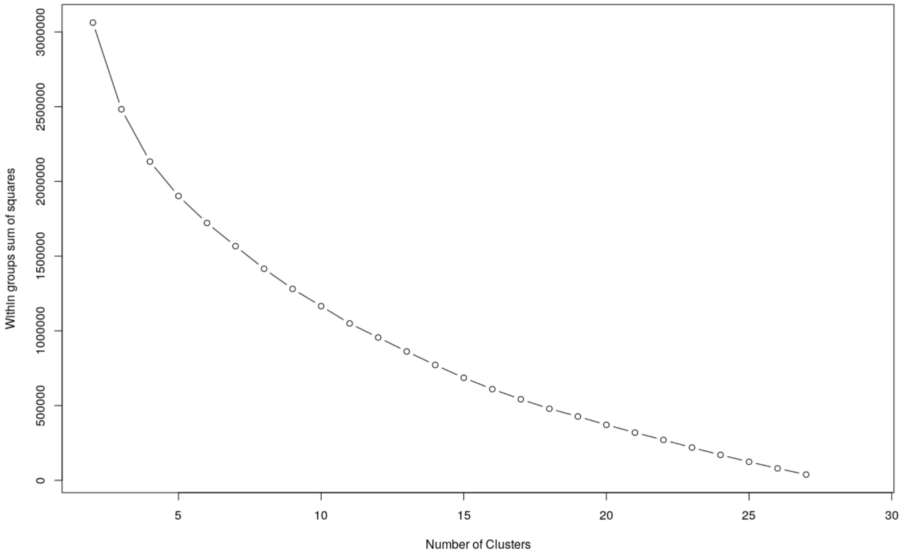

Fig. 1: Elbow in the SSE

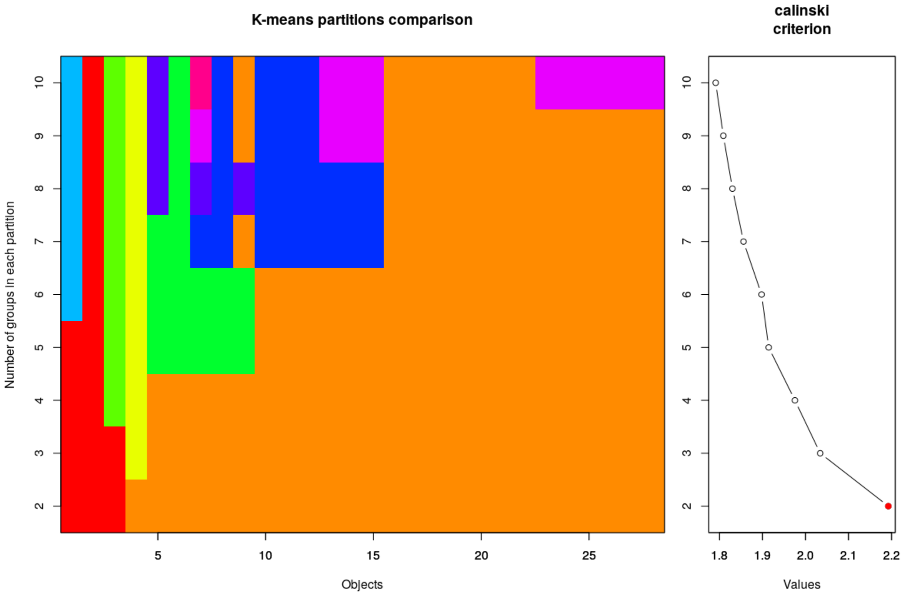

Fig. 2: Calinsky criterion

number of clusters. To determine this number, we have used a graph that shows the relation between the number of extracted clusters and their distance. This method allows us to obtain the value and to generate the clusters using the K-means method. The correct choice of k can be ambiguous. The difficulty arises from the fact that we have to find a balance between the shape and scale of the data set and the number of clusters that the user wants to obtain. We propose in this paper two approaches for diagnosing the number of clusters suitable for the data (see figure 1 and 2). The result on figure 1 uses elbow with the sum of squared error and figure 2 uses Calinsky criterion with an interval for groups between one and ten. The Elbow method is a method for interpreting and validating coherence in clusters to find the appropriate number of clusters in a data set.

3 Results

The results obtained from the K-means clustering and Hierarchical Clustering are respectively presented in the figures 3 and 4. Figure 3 shows an analysis of the data set with an arbitrary choice of the number of clusters $k = 4$. The names of the journals are coded using 4 characters, followed by one character (i, m, r or d) that corresponds to the section type. This figure shows the specific character of the Methods section (at the left), confirming earlier works [3] that underline the atypical nature of this section in terms of citation contexts. Indeed, this cluster shows that the Methods section is located in a single cluster and this regardless of the journal. The journal *ppat* (PLOS Pathogens) is also in one unique cluster that contain the different sections of the rhetorical structure of the component. In addition, *pntd* (PLOS Neglected Tropical Diseases) and *pmed* (PLOS Medicine) are both located in one and the same cluster.

The hierarchical clustering allows us to illustrate the hierarchical organisation of groups as shown on the figure 4. This visualization confirms the previous result, but offers also a hierarchical view of the clusters. The hierarchical analysis highlights the specific character of the Results sections in *pmed* and *pntd* that are in the same group as the Introduction and Discussion sections in these journals. Again, *ppat* appears in a separate cluster. The journal *pcbi* (PLOS Computational Biology) is also in a separate sub-cluster that suggests that the citation contexts in this journal are quite different from those in the other journals.

4 Discussion and Conclusion

This work emphasizes the need for a tool to identify and analyze the contexts of citations, while being aware of the multidimensional nature of these phenomena. Indeed, we have not yet addressed the problem of the categorization of quotations. Our approach aims to determine the clusters of citation contexts, so that at a different stage the topics will have to be identified and processed for a finer analysis. In order to do this, we can consider for example the syntax of citation contexts and the topics, as already proposed in [4]. One of the advantages of

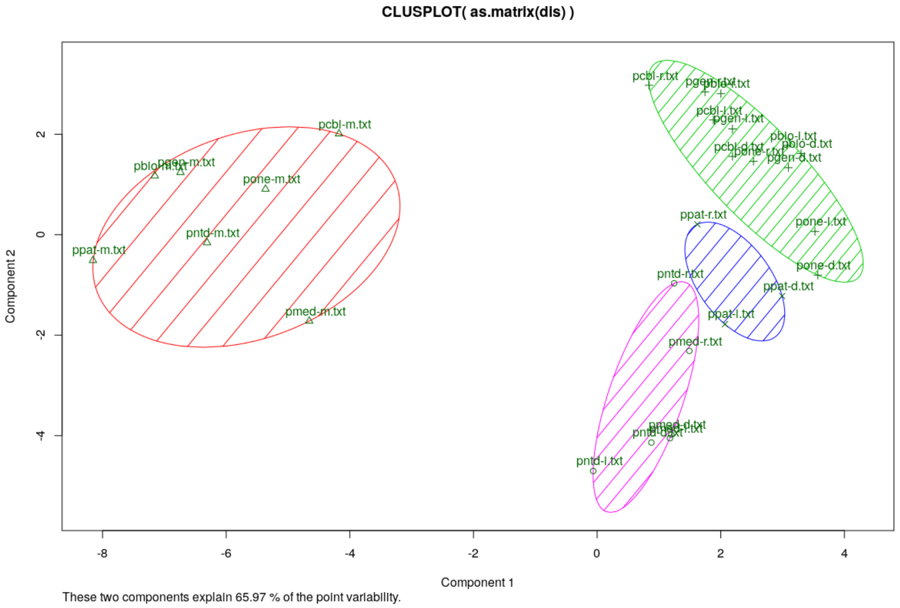

Fig. 3: Clusters: K-means clustering with $k = 4$

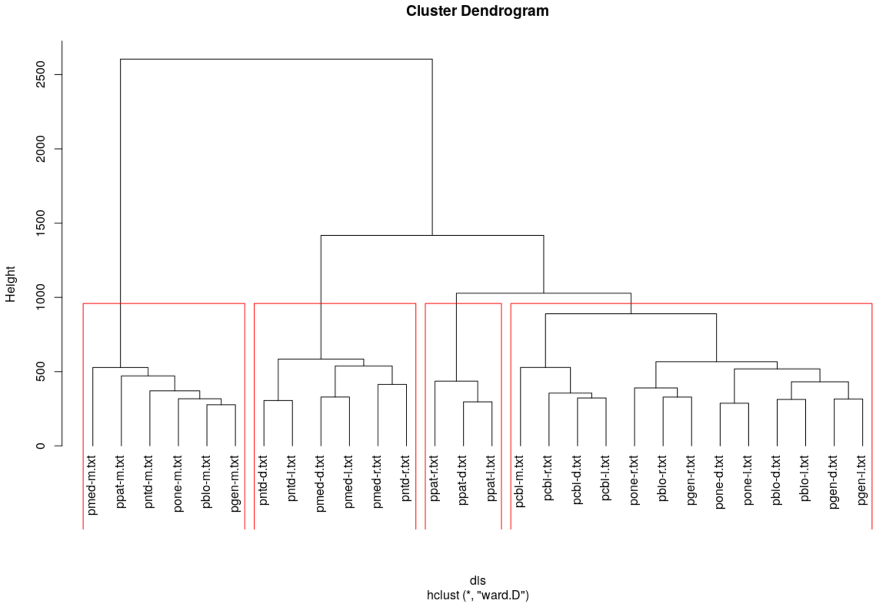

Fig. 4: Clusters: hierarchical clustering

using the topic modeling approach is the possibility to deal with large volumes of textual data.For example, this type of approach has already been used in the political text study [7].

Studying the structure of scientific papers and observing the regularities in the contexts of in-text citations is an important step towards understanding the phenomenon of citation which is central in the process of building scientific knowledge. Different types of citations exist based on the motivation to cite and the relation between the citing authors and the cited work. To be able to create an ontology of citations that reflects the types of citations found in articles it is necessary to process existing corpora and study the properties of citation contexts on a large scale.

Acknowledgments

We thank Benoit Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS data set.

References

1. Bertin, M., Atanassova, I.: A study of lexical distribution in citation contexts through the IMRaD standard. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014). pp. 5–12. Amsterdam, The Netherlands (April 13 2014)
2. Bertin, M., Atanassova, I.: Factorial correspondence analysis applied to citation contexts. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 37th European Conference on Information Retrieval (ECIR 2015). Vienna, Austria (March 29 2015)
3. Bertin, M., Atanassova, I., Larivire, V., Gingras, Y.: The invariant distribution of references in scientific papers. Journal of the Association for Information Science and Technology 67(1), 164177 (January 2016)
4. Griffiths, T.L., Steyvers, M., Blei, D.M., Tenenbaum, J.B.: Integrating topics and syntax. In: NIPS. vol. 4, pp. 537–544 (2004)
5. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1), 100–108 (1979)
6. Kintigh, K.W., Ammerman, A.J.: Heuristic approaches to spatial analysis in archaeology. American Antiquity pp. 31–63 (1982)
7. Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A., Tingley, D., Sinclair, B., Blattman, C., Corstange, D., Humphreys, M., Jamal, A., King, G., Milner, H., Mitts, T., O 'connor, B., Spirling, A.: Computer-Assisted Text Analysis for Comparative Politics. Advance Access publication February Political Analysis 4(23), 254–277 (2015)
8. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. Journal of the American Statistical Association 111(515), 988–1003 (2016)
9. Roberts, M.E., Stewart, B.M., Tingley, D.: stm: R package for structural topic models. R package version 0.6 1 (2014)