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

Marc Bertin\(^1\) and Iana Atanassova\(^2\)

August 11, 2017

\(^1\) CIRST - Université du Québec à Montréal (UQAM), Canada
\(^2\) CRIT - Centre Tesniere, University of Bourgogne Franche-Comte, France

2\(^{nd}\) Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) at the 40\(^{th}\) International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan

The BIRNDL proceedings are published at http://ceur-ws.org/Vol-1888/.

The video of the presentation is available at https://youtu.be/mnTmmRp1g9Y.
Research Problem

Scientific papers usually follow a specific rhetorical structure: IMRaD (Introduction, Method, Result and Discussion)

- The IMRaD structure plays an important role in determining the types of citation contexts;
- The specific domains and topics of the various journals, and also their own editorial lines, can have an effect on the direct context of citations.

Objective

Study the properties of citation contexts on a large scale to be able to create an ontology of citations that reflects the types of citations found in articles.
We propose:

- a method to **analyze citation contexts** at a large scale taking into account various criteria;
- a **multidimensional approach** to this problem which is based on clusters.

We use:

- k-means;
- hierarchical clustering.
## PLOS Dataset

<table>
<thead>
<tr>
<th>Journal</th>
<th>Articles</th>
<th>Citations</th>
<th>Citation contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLOS Biology</td>
<td>1,754</td>
<td>170,785</td>
<td>91,117</td>
</tr>
<tr>
<td>PLOS Computational Biology</td>
<td>2,560</td>
<td>243,488</td>
<td>126,870</td>
</tr>
<tr>
<td>PLOS Genetics</td>
<td>3,414</td>
<td>332,845</td>
<td>185,537</td>
</tr>
<tr>
<td>PLOS Medicine</td>
<td>926</td>
<td>72,676</td>
<td>34,819</td>
</tr>
<tr>
<td>PLOS Negl. Tropical Diseases</td>
<td>1,872</td>
<td>133,022</td>
<td>73,211</td>
</tr>
<tr>
<td>PLOS ONE</td>
<td>72,158</td>
<td>5,363,036</td>
<td>2,854,082</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>82,684</strong></td>
<td><strong>6,315,852</strong></td>
<td><strong>3,365,636</strong></td>
</tr>
</tbody>
</table>

- Published by the Public Library of Science (PLOS), in Open Access;
- XML, Journal Article Tag Suite (JATS);
- Entire corpus up to September 2013.
The Elbow Method to determine the number of clusters

Elbow with the sum of squared error:

Calinsky criterion with interval for groups between 1 and 10:
Results: K-means clustering with $k = 4$

These two components explain 65.97% of the point variability.
Results: Hierarchical Clustering

Cluster Dendrogram

dis
hclust (*, "ward.D")
Conclusion

- We observe the atypical nature of the Methods section in terms of citation contexts, and this confirms previous studies (see [6, 2, 4, 5, 1, 3]);
- One of the advantages of using the topic modeling approach is the possibility to deal with large volumes of textual data;
- 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.
Thank you!

Marc Bertin  
Assistant Professor  
ELICO  
Université Claude Bernard Lyon 1, France  
marc.bertin@protonmail.ch

Iana Atanassova  
Assistant Professor  
CRIT - Centre Tesniere,  
University of Bourgogne Franche-Comte, France  
iana.atanassova@univ-fcomte.fr
Marc Bertin and Iana Atanassova.
A study of lexical distribution in citation contexts through the IMRaD standard.

Marc Bertin and Iana Atanassova.
Multiple in-text reference phenomenon.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The distribution of references in scientific papers: an analysis of the imrad structure.

Marc Bertin, Iana Atanassova, Vincent Larivière, and Yves Gingras.
The linguistic context of citations: a cartography of the structure of scientific papers.
In AAAS Annual Meeting, San Jose, CA, February 2015. American Association for the Advancement of Science.
Marc Bertin, Iana Atanassova, Vincent Larivièreme, and Yves Gingras.

Mapping the Linguistic Context of Citations.


Marc Bertin, Iana Atanassova, Vincent Larivièreme, and Yves Gingras.

The invariant distribution of references in scientific papers.

_Journal of the Association for Information Science and Technology, 67(1):164–177, January 2016._