# Clustering of scientific citations in Wikipedia

Finn Årup Nielsen

The Lundbeck Foundation Center for Integrated Molecular Brain Imaging
at DTU Informatics and Neurobiology Research Unit, Copenhagen University
Hospital Rigshospitalet
Lyngby and Copenhagen, Denmark
fn@imm.dtu.dk

**Abstract.** The instances of templates in Wikipedia form an interesting data set of structured information. Here I focus on the `cite journal` template that is primarily used for citation to articles in scientific journals. These citations can be extracted and analyzed: Non-negative matrix factorization is performed on a (article × journal) matrix resulting in a soft clustering of Wikipedia articles and scientific journals, each cluster more or less representing a scientific topic.

## 1 Introduction

The category system and the use of templates in Wikipedia provide interesting data sets of structured information. A number of reports have come out that use the category graph in automatic text processing, e.g., [1–3]. DBpedia databases Wikipedia template information and associated Internet services enable database-like queries [4]. I have previously reported results of relatively simple statistical analysis about a single Wikipedia template — the `cite journal` template — counting the number of overall outbound scientific citations and comparing it to the citation statistics *Journal Citation Reports* from the company *Thomson Scientific* [5]. Other researcher have considered more advanced statistical models in the form of multivariate analysis [6, 7]. They build a matrix from intrawiki links and submit it to numerical algorithms. Here I will take a similar approach but construct the matrix from data associated with the scientific citation template rather than wikilinks. The present work will show an example on how to make multivariate statistical analysis on the structured data in a Wiki, and in this particular case provide an overview of how science is represented in Wikipedia.

## 2 Method: From XML via matrices to topic visualization

A Perl script extracted the instances of the `cite journal` templates from bzipped XML files of the English Wikipedia downloaded from the Internet server

`download.wikipedia.org`. Another Perl script extracted the name of the journal from the `journal` field in the template, and at the same time tried to match the name to a 'canonical' journal name. For the matching a small XML file — originally built for the neuroinformatics Brede Database [8] — listed the canonical name and variations in the names for so far 255 different journals, e.g., the entry for the journal with the canonical name *Proceedings of the Royal Society of London, Series B, Biological Sciences* listed 12 other variations for the name including the PubMed abbreviation *Proc R Soc Lond B Biol Sci*. These 255 journals comprised a large part of the top cited journals from Wikipedia, and thus the script normalizes very many citations to a canonical name, but indeed far from all variations to lesser cited journals are resolved. There are a number of other issues that prevents the databasing of the citations to be particular exact: Special cases of journal naming make it hard to match all journal names with a canonical journal name, e.g., *Mutation Research* are actually three (or four) different journals, wrt. to ISSN. Cited 'journals' may not be scientific journals, but, e.g., newspapers. Citations that occur multiple times in the same Wikipedia article to the same item (by the `<ref name="anchor"/>` construct) were only counted once.

A (article × journal) data matrix is built up where each column corresponds either to a canonical journal name or the journal name as written in the citation of the Wikipedia articles. Each row corresponds to a Wikipedia article. The $(i, j)$ elements in the matrix is set to the number of times the $i$th article cites the $j$th journal. Most of the elements in the matrix are zero.

Clustering of the constructed matrix is performed with the multiplicative update rules of the non-negative matrix factorization (NMF) as put forward by Lee and Seung [9]. The algorithm for the 'Euclidean distance' runs with 50,000 iterations. This particular multivariate analysis resembles several other methods such as the one used by Buntine in his Wikipedia analysis [6] as well as Bellomi and coworkers' analysis [7]. NMF splits the data matrix $\mathbf{X}$(article × journal) into three other matrices $\mathbf{X} = \mathbf{WH} + \mathbf{U}$. Whereas $\mathbf{U}$ is just the residual matrix, the factorized matrices $\mathbf{W}$ and $\mathbf{H}$ form the interesting matrices that may be expected to represent specific scientific topics characterized by their citation patterns: A specific column in $\mathbf{W}$ can be interpreted to contain the loadings of articles on a specific 'topic' that the cluster represents, and a specific row in $\mathbf{H}$ contains loadings for journals on that topic. The NMF results not in a hard clustering where the items are assigned exclusively to one cluster, — rather in a soft two-way clustering. Using the Kleinberg terminology [10], the $\mathbf{W}$ matrix contains loadings for Wikipedia 'hub' articles, whereas $\mathbf{H}$ contains 'authoritative' journal articles. One advantage of the Lee and Seung's 'Euclidean distance' version of the NMF algorithm is that no multiplications take place with the full reconstructed data matrix, i.e., the product matrix $\mathbf{WH}$. This is in contrast to 'divergence' version, that in my implementation is much slower and use more memory for these kinds of data sets.

The initialization of the NMF algorithm requires the specification of the number of clusters, i.e., the number of columns in the $\mathbf{W}$ matrix and the number

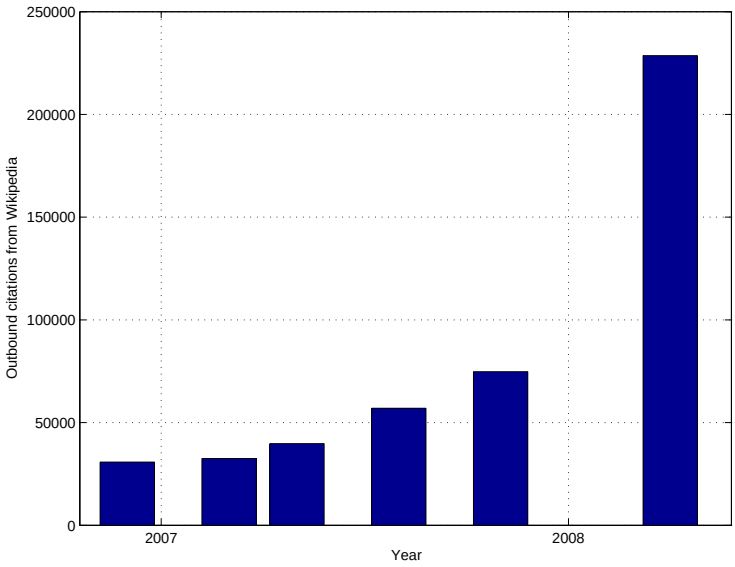


**Fig. 1.** Number of outbound scientific citations as counted from the use of the `cite journal` template for different dumps of the English Wikipedia. A sharp rise is seen from the 2007 dumps to the 2008 dump due to citations added by a bot.

of rows in the **H** matrix. I make the NMF algorithm run with different number of clusters: From one to twenty. Each run will be independent of the other and they can be run in parallel on a computer cluster. Many results appear when running the NMF with different number of clusters, and a so-called 'cluster bush' visualization can be used to get a overview of the relationship between the different clusterings [11]. In this kind of plot each cluster is rendered as a circle and the amount of overlap between two clusters is indicated with the thickness of a line.

The NMF algorithm is run and the cluster bush visualization is made in Matlab with functions from the Brede Toolbox [12].

## 3 Results

Examing the full count of scientific citations from Wikipedia a marked increase becomes apparent with a rise in the number of citations from 2007 to the examined dump of March 2008, see Figure 1: From 74,776 citations in the October 2007 dump to 228,593 in the March 2008 dump.

Whereas astronomy journals received comparably many citations from Wikipedia in the 2007 dumps, and journals such as *The Journal of Biological Chemistry* had relatively few citations when compared to the *Journal Citation Reports*, this citation pattern is now very much changed: Wikipedians have con-

| Citations | Journal name |
|---|---|
| 16739 | The Journal of Biological Chemistry |
| 12779 | PNAS |
| 8772 | Genome Research |
| 7561 | Nature |
| 4007 | Nature Genetics |
| 3928 | Genomics |
| 3689 | Science |
| 3511 | Gene |
| 3380 | Biochemical and Biophysical Research Communications |
| 3043 | Molecular and Cellular Biology |
| 2975 | Cell |
| 2261 | The EMBO Journal |

**Table 1.** Most cited journals from Wikipedia in the 12th March 2008 dump.

structed the bot *ProteinBoxBot* that automatically builds infoboxes and citations in Wikipedia articles. Thus a very large number of citations to protein/gene work has been added, and with the March 2008 dump *The Journal of Biological Chemistry* can be found as the most cited journal, see Table 1. Scientific articles cite also this journal the most, according to *Journal Citation Reports*.

The conversion of the information in the templates to a matrix representation results in matrices size $(23595 \times 18194)$ and $(43073 \times 23096)$ for the October 2007 and March 2008 dump, respectively. The densities of the constructed matrices are 0.01%–0.02% depending on the dump version of Wikipedia. The number of columns in the matrices would have been smaller and the density higher if the matching of journal names was more complete. The number of articles using the `cite journal` template has almost doubled in less than half a year between the two dumps. This increase is likely due to the large number of articles added for proteins/genes by *ProteinBoxBot*. Many of these articles have no other text besides the text added by the bot and the citations are not in-text citations.

Figure 2 displays a cluster bush visualization of the NMF results for the October 2007 dump for NMF, and for clarity only the NMF runs with one to seven clusters are shown: The bottom row displays the run with just one cluster, where the Wikipedia articles *List of molecules in interstellar space* and *Extinction (astronomy)* are the largest hubs. *The Astrophysical Journal* and *Astronomy & Astrophysics* are the largest authoritative journals for this astrophysical cluster. As the NMF model size increases, i.e., more clusters get added, this topic continues to be a cluster of its own. The new clusters that arise are related to medical sciences, intelligence, human leukocyte antigen and bacteria. For these runs of NMF the columns corresponding to the cross-disciplinary journals *Nature*, *Science* and *PNAS* were excluded.

With the present algorithm a few of the clusters represent very restricted topics, e.g., in one case the article *Henry George Fourcade* and the journal *The Photogrammetric Record* constituted a single cluster. Another cluster that is also

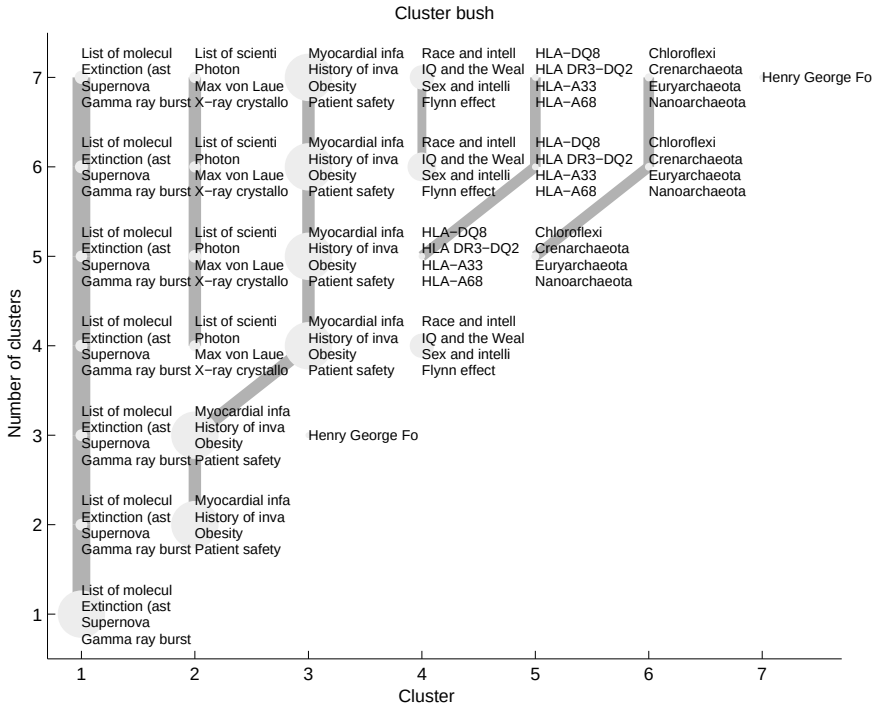


**Fig. 2.** 'Cluster bush' visualization of results for non-negative matrix factorization (NMF) of the scientific citations in the 18th October 2007 dump of the English Wikipedia. Each circle denotes a cluster. The lowest row displays the results of an NMF run with one cluster, second lowest row the results for NMF with two clusters, etc. The text on the nodes are the Wikipedia articles that are associated with high loadings in the factorized matrices of the NMF. The lines between the nodes indicate how much the clusters overlap.

dominated by single items has the article about the group of genes *Solute carrier family* and the journal *Pflügers Archiv European Journal of Physiology*.

Applying NMF on the March 2008 dump results in components that are overwhelmingly affected by the large number of citations in the protein/gene articles. A run of NMF with twenty clusters resulted in only three clusters that did not exhibit an association with genes: One cluster centered around solar system astronomy with the journal *Icarus* as the primary authoritative journal and *Uranus* as the top hub Wikipedia article, another cluster centered around *The Astrophysical Journal*, and the third as a medical clusters with *New England Journal of Medicine* and *The Lancet* as top authorities and *Myocardial infarction* as the Wikipedia hub article. The rest of the seventeen clusters were all related to proteins and genes or other closely related topics within biology and biochem-

| Cluster | Wikipedia hub articles | Authoritative journals |
|---|---|---|
| 'Cancer' | RBL2 | Oncogene |
| | MYB | Cancer Research |
| | ERG (gene) | Int. J. Cancer |
| | EPS8 | Gene & Development |
| 'Immunology' | DNA vaccination | The Journal of Immunology |
| | CCL21 | The Journal of Experimental Medicine |
| | HLA-DQ8 | Tissue Antigens |
| | HLA-DQA1 | Eur. J. Immunol. |
| 'Blood' | Acute myeloid leukemia | Blood |
| | Serpin | British Journal of Haematology |
| | CEBPE | The Journal of Clinical Investigation |
| | CD34 | The Journal of Experimental Medicine |
| 'Virology' | Papillomavirus | The Journal of Virology |
| | HHV Infected Cell ... | Virology |
| | Poliovirus | Journal of Molecular Biology |
| | RELB | AIDS Res. Hum. Retroviruses |

**Table 2.** The top Wikipedia hubs articles and authoritative journals with respect to clusters from a non-negative matrix factorization with twenty clusters.

istry. Many of these clusters are mostly driven be a single journal, i.e., a single element in each row of the **H** matrix are much larger than the rest of the elements, whereas the **W** matrix shows a much more equal loading over Wikipedia articles within each cluster, e.g., one cluster interpretable as a 'virology' cluster would have *The Journal of Virology* as the dominating authoritative journal.

A few examples of items in a sample of clusters from an NMF run with twenty clusters are shown in Table 2. These kinds of results may be written to an HTML page and put on the web to serve as an online overview of how science is cited from Wikipedia.

## References


1. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In: Advances in Web Intelligence. Volume 3528 of Lecture Notes in Computer Science., Berlin/Heidelberg, Springer (2005) 380–386
2. Strube, M., Ponzetto, S.P.: WikiRelate! computing semantic relatedness using Wikipedia. In: Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence, Menlo Park, Califonia, AAAI Press (2006) 1419–1424
3. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, Menlo Park, California, AAAI Press (2006) 1301–1306

4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: The Semantic Web. Volume 4825 of Lecture Notes in Computer Science., Springer (2008) 722–735
5. Nielsen, F.Å.: Scientific citations in *Wikipedia*. First Monday **12**(8) (August 2007)
6. Buntine, W.: Static ranking of web pages, and related ideas. In Beigbeder, M., Yee, W.G., eds.: Open Source Web Information Retrieval. (October 2005)
7. Bellomi, F., Bonato, R.: Network analysis of Wikipedia. In: Proceedings of Wikimania 2005 — The First International Wikimedia Conference. (2005)
8. Nielsen, F.Å.: The Brede database: a small database for functional neuroimaging. NeuroImage **19**(2) (June 2003) Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY.
9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In Leen, T.K., Dietterich, T.G., Tresp, V., eds.: Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, Cambridge, Massachusetts, MIT Press (2001) 556–562
10. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Technical Report 10076(91892), IBM (May 1997)
11. Nielsen, F.Å., Balslev, D., Hansen, L.K.: Mining the posterior cingulate: Segregation between memory and pain component. NeuroImage **27**(3) (2005) 520–532
12. Nielsen, F.Å., Hansen, L.K.: Experiences with Matlab and VRML in functional neuroimaging visualizations. In Klasky, S., Thorpe, S., eds.: VDE2000 - Visualization Development Environments, Workshop Proceedings, Princeton, New Jersey, USA, April 27–28, 2000, Princeton, New Jersey, Princeton Plasma Physics Laboratory (April 2000) 76–81