From cortisone to graphene: 60 years of breakthroughs in PubMed publications

André Panisson, Marco Quaggiotto

Data Science Laboratory, ISI Foundation, Torino, Italy

This report refers to the methodology used to create the visualization entitled "From cortisone to graphene: 60 years of breakthroughs in PubMed publications", submitted to the Data Visualization Challenge of the 2014 ACM Web Science Conference.

Biomedical science advances can often be described by brief periods of discovery followed by decades of study and research. This analysis aims to show an overview of the breakthroughs inferred from the titles of 21.5 millions of PubMed publications. For this visualization, we built a timeline of terms that have experienced a sudden increase in popularity in PubMed publication titles. The position of the terms is related to the years in witch the breakthrough has occurred, the size and color of the word are indicative of the popularity of the term. Authors that contributed significantly to the breakthroughs have been linked to their areas of research.

In order to create the visualization we:

1. Extracted words and bigrams from publication titles
2. Analyzed the occurrence of such terms in the years between 1950 and 2010
3. Detected peaks of activity for each term, identifying the years with high acceleration in the term usage
4. Filtered out terms with low entropy as they are too generic
5. Identified related authors (with at least 100 publications, of which at least 50 contain the given term in the title)
6. Created a weighted graph with years, terms and authors
7. Used a force-directed layout to position the terms next to their years and authors next to their related terms

In the next sections, we discuss this methodology in detail.

1 Dataset

PubMed comprises more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books. In this work, we used metadata for the complete set of all PubMed records, including title, authors, and year of publication. All data provided originates from NLMs PubMed database (as downloaded April 24, 2013 from the NLM FTP site) and was retrieved via the Scholarly Database[1;2].

The MEDLINE®/PubMed® dataset contains articles starting before 1900, however the number of publications is very small for each year in the 19th century (less than 1000 per year). After 1900, the number of publications per year grows above 1000, but a big increase starts in the 1940s, when the number of publications jumps from 3357 in 1944 to 17148 in 1945. Figure 1 shows a timeline with the number of publications in the dataset for each year, starting from 1900.

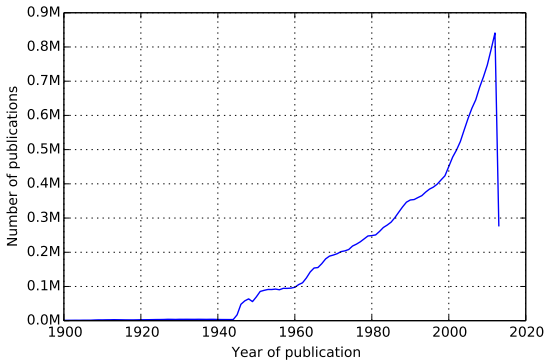

Figure 1: Number of publications in the MEDLINE®/PubMed® dataset starting from 1900

In order to limit the analysis to years with high number of publications, we decided to show in the visualization only data from 1950 to 2010.

2 Text analysis

Our method starts with text analysis of the article titles. In order to analyze the article titles, we use the "bag of words" approach, and we create a matrix $A \in \mathbb{R}^{N \times M}$, where N is the number of articles and M is the number of terms extracted from the dataset. Each position $a_{i,j}$ represents the number of times the term j appears in the title of article i.

In order to extract the terms, we use the steps that we show in the following. We illustrate each step with the example of the title `'Drugs and the high cost of health care.'`:
(1) choose a list of stop words (common words in English and in article titles): `[is, and, the, of, ..., letter, article, conference, ...]`
(2) extract a list of words from the article title: `[drugs, and, the, high, cost, of, health, care]`

(3) lemmatize the words of the list (e.g., `drugs` becomes `drug`): `[drug, and, the, high, cost, of, health, care]`
(4) remove the stop words and create n-grams with maximum size of 2 by concatenating words only if they are not separated by a stop word: `[drug, high, cost, health, care, high cost, health care]`

For each article title, we apply the above process and add the resulting terms to a counter. The result is a dictionary that associates each term with the number of times the term appears in the dataset. We select only the 100 thousand most common terms and proceed to the construction of matrix A, where the number of terms M is 100 thousand.

3 Time series analysis

Since we want to select terms that could represent discoveries and breakthroughs, we analyze the time-series of the occurrences of each term over the years.

For each term, we start by creating a time series with the number of term occurrences in each year. For example, the term **hiv** exhibits a time series of occurrences as shown in Figure 2.

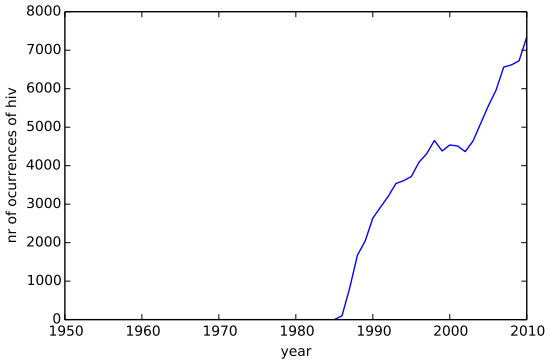

Figure 2: Time series with number of occurrences of the term **hiv** from 1950 to 2010.

In this figure, we see the increasing occurrence of the term **hiv**, but this time series is affected by the increasing number of articles in the dataset. In order to remove the effect of increasing number of articles, we normalize the frequencies by dividing the term occurrence by the sum of all term occurrences in each year. This gives us the time series with the relative frequency of the terms, in the following referred as $f(t)$. For the term **hiv**, we show its $f(t)$ in Figure 3.

We derive the following time series from $f(t)$:

$$\delta(f(t)) = f(t) - f(t-1)$$
$$\delta^2(f(t)) = \delta(f(t)) - \delta(f(t-1))$$

In this example, $\delta^2(f(t))$ corresponds to the second discrete derivative of $f(t)$, i.e., the years where the frequency acceleration is at its maximum.

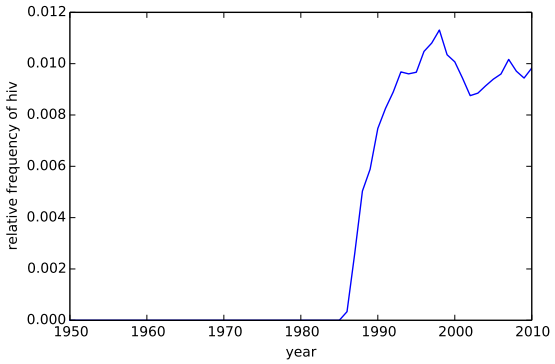

Figure 3: Time series with relative frequency ($f(t)$) of the term **hiv** from 1950 to 2010.

$$\varphi(f(t)) = \frac{f(t)}{f(t-1)+\varepsilon}$$

$$\omega(f(t)) = 1 - e^{\frac{1-\varphi(f(t))}{C}}$$

$$g(t) = \delta^2(f(t))\delta(f(t))\omega(f(t))$$

ε is a small number used to avoid zero division. The parameter C is used to give less weight to frequencies that jump from high values to higher values, and more weight to frequencies that jump from very low values to high values. In our experiments, we used $C = 10$. The resulting time series $g(t)$ is given by the product of three time series: $\delta^2(f(t))\delta(f(t))\omega(f(t))$. We multiply $\delta^2(f(t))$ - the second discrete derivative of $f(t)$ - by $\delta(f(t))$ - the first discrete derivative of $f(t)$, since we want the years where there is a high acceleration in the term usage and also the years where the increase in term usage is positive. We multiply by $\omega(f(t))$ in order to give less weight to the increases that have a relative small change from $t-1$ to t. Finally, we set all negative values to 0 in order to keep only positive peaks.

$g(t)$ has peaks in the years where the frequency goes from near zero to a high value. In order to classify the terms by their frequency jumps, we use a measure of entropy. In Figure 4, for some terms, we show in blue the initial time series with the relative frequency -$f(t)$, in red the resulting $g(t)$ for the term, and we show next to the term name the resulting entropy calculated over $g(t)$.

Using this methodology to analyze the time series of each term, we are able to extract two types of information:

1. Most of the terms with high entropy (next to 1) are terms that are used in all years, or that have no visible jump in the frequency. By selecting only terms with low entropy and high accumulated frequency, we have a good chance to extract terms that represent discoveries and breakthroughs from 1950 to 2010.

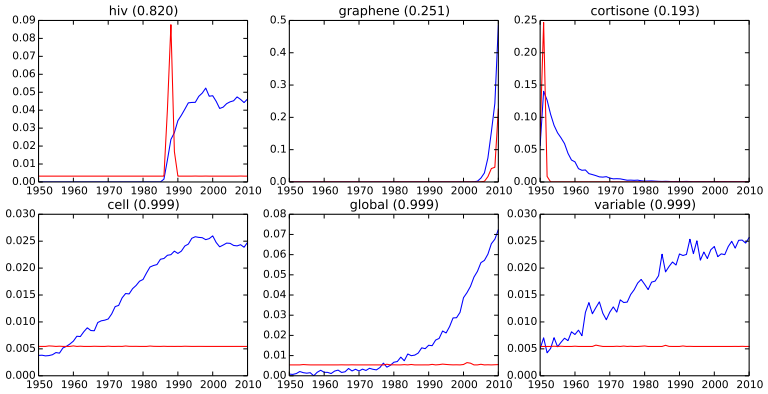

Figure 4: For some terms, we show (i) in blue the time series with the relative frequency ($f(t)$), (ii) in red the processed $g(t)$ for the term, and (iii) next to the term name, the resulting entropy calculated over $g(t)$.

2. Terms with low entropy have high peaks that can be associated to the year of their frequency increase. The values in $g(t)$ associated to each year can be used as a weight to associate a given term to the year of its discovery.

3.1 A small validation

One simple way to validate our method for finding the discoveries associated to each year is to check if terms associated to Nobel Prizes are present in the extracted terms. The 1950 Nobel Prize in Medicine was given to Edward Calvin Kendall, Tadeus Reichstein and Philip Showalter Hench "for their discoveries relating to the hormones of the adrenal cortex, their structure and biological effects". The term associated to this discovery is "cortisone". The 2010 Nobel Prize in Physics was given to Andre Geim and Konstantin Novoselov "for groundbreaking experiments regarding the two-dimensional material graphene", and the associated term is "graphene". These two terms motivated the title of our work: "From cortisone to graphene". We were able to verify that both terms are present in the visualization. Other terms like "cell", "global" and "variable" have high frequency in the dataset, but since they present low entropy for their corresponding $g(t)$, they were excluded from the visualization.

4 Author selection

The MEDLINE®/PubMed® dataset contains around 10.7 million author names associated to the articles comprised between 1950 and 2010. Each author name is expressed as 'Surname, Name' (e.g., 'Smith, John') or as 'Surname, intitals' (e.g., 'Smith, J'). From the articles and author names, we can build a matrix $B \in \mathbb{R}^{N \times O}$, where N is the number of articles and O is the number of author names. Each position $b_{i,j}$ is set to 1 if the author name j is associated to article i, and is set to 0 otherwise.

The author name associated to the highest number of articles is 'Suzuki, T'. Clearly, this author name represents more than one real author. In fact, the page of PubMed Advanced Search shows that there are tens of authors with surname 'Suzuki' and initial 'T'. In order to create a list of authors associated to each term, we wanted to avoid using author names that do not refer to a single author. For this, we also used a measure of entropy for each author. The idea behind it is that author names that are associated to a high number of terms are most probably different authors, since authors are committed to few topics. The entropy of each author is calculated using the following equation:

$$H_j = -\sum_{i=1}^{M} p(i|j) \log(p(i|j))$$

where $p(i|j)$ is the probability of having a term i given an author j. From this, we calculate a scaled entropy:

$$E_j = 1 - \frac{H_j}{H_{max}}$$

Figure 5 shows an histogram with the distribution of number of authors for each bin of E_j values and the number of removed author names when selecting only author names with E_j above a threshold t ($E_j > t$). We verified that setting $E_j > 0.25$, we still keep more than 99.7% of the author names and, at the same time, we remove all top author names that refer to more than one real author.

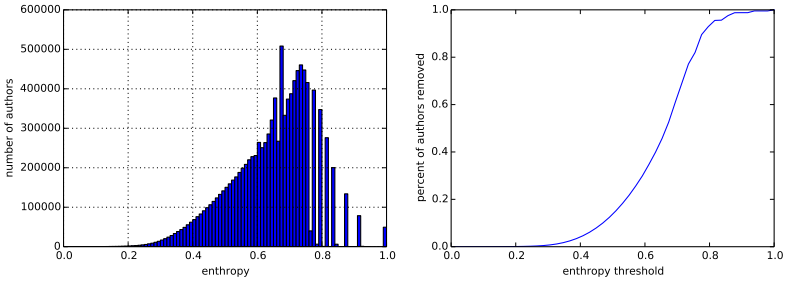

Figure 5: In the left, an histogram showing the distribution of number of author names for each bin of E_j values (we used 100 bins). In the right, the number of removed author names when selecting only author names with E_j above an entropy threshold t ($E_j > t$).

From the list of author names with $E_j > 0.25$, we identified only the authors with least 100 publications. Next, use the document-term matrix A to calculate the co-occurrence of authors and terms: $C = A^T B$, with $C \in \mathbb{R}^{M \times O}$. From the matrix C we select only the rows that refer to the terms chosen for the visualization. Finally, for each row, we select only the top 2 authors which used at least 50 times the given term in their articles.

5 Graph creation and layout

We used the years, terms and authors to create a weighted graph as follows:

- V_1 is the set of nodes representing the years, i.e., from 1950 to 2010 (61 nodes).
- V_2 is the set of nodes representing the selected terms.
- V_3 is the set of nodes representing the selected authors.
- E_{12} is a set of weighted edges connecting years with terms. The weights are proportional to the values of the $g(t)$ corresponding to each term.
- E_{22} is a set of weighted edges connecting terms with terms. The weights are proportional to the co-occurrence of the terms in article titles.
- E_{23} is a set of weighted edges connecting terms with authors. The weights are proportional to the number of times an author used the term in article titles.
- V is the union of V_1, V_2 and V_3.
- E is the union of E_{12}, E_{22} and E_{23}.

G is a graph with nodes V and edges E. This graph was exported to a file and imported to a graph exploration tool (in this case, Gephi[3]). Finally, we set the year nodes to fixed positions, and used two different layout algorithms to distribute spatially the term and author nodes: Force Atlas and Label Adjust. The size and color of term nodes were set proportionally to the term frequency, as shown in Figure 6.

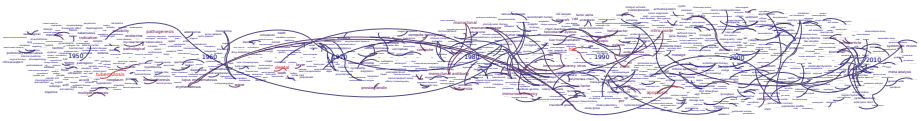

Figure 6: The graph G with year nodes, term nodes, author nodes and their connections. We don't show year-term and term-term connections, since there are too many connections between these nodes.

6 Future improvements

Many techniques for automatic term recognition (ATR) are available in the literature[4]. The approach for term extraction used in this study is very simple, and a better approach could improve the results of this work.

For example, the terms `human`, `immunodeficiency` and `virus` are part of a single multi-word term: `'human immunodeficiency virus'`. However, the term `immunodeficiency` is extracted as a relevant term. `'human'` and `'virus'` are not recognized as relevant, since they are too broad. However, both `'human immunodeficiency'` and `'immunodeficiency virus'` are extracted as relevant terms. The full multi-word term is not extracted, since we limited the n-grams

to a maximum of 2 words. We can find other examples of terms that are affected by this problem, and it could be solved by using c-value for term extraction [5], a method that combines linguistic and statistical information to extract multi-word terms.

Another problem results from the choice of using lemmatization in the term extraction phase. For example, the term `'ionize radiation'` refers to *ionizing radiation*, but the word `'ionizing'` was lemmatized in the process, and was maintained in this form for all the process to the visualization. This problem could also be avoided by using better term extraction approaches.

Since this is an unsupervised method to extract terms and construct a visualization, the method is susceptible to errors and noise in the data. For example, the terms `'july'` and `'possibility'` were considered as relevant terms, even if they can't be associated to any discovery or breakthrough. A semi-supervised methodology, where the help of a domain expert could be used to clean up the data, would be preferred in this case.

Another way to validate the terms extracted by our approach is using the UMLS vocabulary, which is also extracted from the MEDLINE®/PubMed® dataset [6]. This vocabulary integrates over 730,000 biomedical concepts, and some visual approaches were proposed for exploring semantic groups in this data [7]. Such vocabulary can be used to verify the validity and relevance of terms.

References

[1] R. P. Light, D. E. Polley, and K. Borner, "Open data and open code for big science of science studies," in Proceedings of International Society of Scientometrics and Informetrics Conference, pp. 1342–1356, 2013.

[2] G. L. Rowe, S. A. Ambre, J. W. Burgoon, W. Ke, and K. Borner, "The scholarly database and its utility for scientometrics research," Scientometrics, vol. 79, May 2009.

[3] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks.," in ICWSM, The AAAI Press, 2009.

[4] K. Kageura and B. Umino, "Methods of automatic term recognition: A review," Terminology, vol. 3, no. 2, pp. 259–289, 1996.

[5] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the C-value/NC-value method," International Journal on Digital Libraries, vol. 3, no. 2, pp. 115–130, 2000.

[6] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," Nucleic Acids Research, vol. 32, no. Database-Issue, pp. 267–270, 2004.

[7] O. Bodenreider and A. T. McCray, "Exploring semantic groups through visual approaches," Journal of Biomedical Informatics, vol. 36, no. 6, pp. 414 – 432, 2003. Unified Medical Language System.