Text mining for notability computation

Gil Francopoulo [1], Joseph Mariani [2], Patrick Paroubek [2]
1 LIMSI, CNRS, Université Paris-Saclay + Tagmatica (France)
2 LIMSI, CNRS, Université Paris-Saclay (France)

gil.francopoulo@wanadoo.fr, joseph.mariani@limsi.fr, pap@limsi.fr

Abstract

In this article, we propose an automatic computation for the notability of an author based on four criteria which are: production, citation, collaboration and innovation. The algorithms and formulas are formally presented, and then applied to a given scientific community: the Natural Language Processing (NLP) group of scientific authors gathering 48,894 people. For this purpose, a large corpus of NLP articles produced from 1965 up to 2015 has been collected and labeled as NLP4NLP with 65,003 documents. This represents a large part of the existing published articles in the NLP field over the last 50 years. The two main points of the approach are first that the computation combines pure graph algorithms and NLP systems. The second point deals with the interoperability aspects both for the corpus and the tools.

Keywords: Natural Language Processing, Bibliometrics, Scientometrics, Citation analysis, Content analysis, Informetrics

1. Introduction

The *notability* of an author is a rather fuzzy notion, and trying to compute such a notion seems a non-sense. However, we will try to demonstrate that a computational approximation is feasible. Notability is defined in Wikipedia as “the property of being worthy of notice, having fame, or being considered to be of a high degree of interest, significance, or distinction[1]”. We are not going to compute a ranking as a hit parade of the “best” authors, but our intent is to provide a picture of the Natural Language Processing (NLP) ecosystem and acknowledge the contributions of the members of this community[2], while stressing that those contributions may have various aspects. The approach is to apply NLP tools on scientific texts related to NLP itself, taking advantage of the fact that we are well informed about the domain ourselves, a very useful skill for appreciating the pertinence of the results returned by automatic tools when dealing with author names and domain terminology.

2. Corpus

Our research began by gathering a large corpus of NLP scientific articles covering documents produced from 1965 to 2015. This corpus gathers a large content of our own research field, i.e. NLP, covering both written and spoken sub-domains and extended to a limited number of corpora, for which Information Retrieval and NLP activities intersect. This corpus was collected at LIMSI-CNRS (France) and is named NLP4NLP (Francopoulo et al, 2015). It contains currently 65,003 documents coming from various conferences and journals with either public or restricted access. This represents a large part of the existing published articles in our field, aside from the workshop proceedings and the published books. The number of sub-corpora is 34 (e.g. LREC). These corpora are made of 558 conference venues[3] (e.g. LREC 2014) and journal issues (e.g. LRE 2013). The number of different authors is 48,894 and the number of author-article combinations is 183,348. More details may be found on line in D-Lib magazine[4] and on our web site[5].

3. Interoperability

The interoperability is achieved at three levels: corpus format, tool managed formats and tool implementation.

3.1 Corpus format

The format for the corpus is the one which is implemented by the ACL Anthology[6] with the meta-data structured as a BibTex and the content as a PDF file. This decomposition in two parts is widely used within our community.

3.2 Tool managed formats

The tools are based on international standards. Internally, the NLP parser uses an ISO-LMF dictionary (Francopoulo et al, 2006). The output conforms to the international standards which are ISO-MAF (aka ISO 24611) and ISO-SynAF (aka ISO 24615).

3.3 Tool implementation

Concerning the tools, all the programs are 100% Java codes (conforming both version 7 or 8). There is nothing non-portable like shell script or C-Language portions. The code does not rely on any external library, thus the application is considered as “freestanding”. The only requirement to run is, of course, the availability of a Java Runtime Machine. The application runs on Windows and Linux, and because of the property of “freestanding”, the code may be packaged[7] into a single archive and pushed to a cloud, in other terms, the code is “cloud ready”. The code makes an heavy use of the multi-threading in Java, and thus benefits from the multi-core architecture of the modern computers. The code is open source.

[1] https://en.wikipedia.org/wiki/Notability

[2] We consider here NLP as including both written and spoken language processing.

[3] The count may be slightly different depending on the way joint conferences are considered. The number of venues is 577 when joint conferences are counted for two.

[4] www.dlib.org/dlib/november15/francopoulo/11francopoulo.html

[5] www.nlp4nlp.org

[6] http://aclweb.org/anthology

[7] This operation has been done occasionally.

4. Outlines

The notion of notability is not strictly associated to the number of papers published by an author. Some authors publish a lot but are not much cited in regard to their production. Conversely some authors did not publish a lot but are profusely cited. In our domain, the most famous example is Kishore A Papineni who published only 16 papers according to our corpus, invented the BLEU score for machine translation evaluation (Papineni et al, 2002) and whose article is the most cited over the whole history of the NLP archives with more than 1500 citations, either with a positive, neutral or negative polarity. Another feature is the collaboration aspect, especially with regards to the whole career of a researcher: does the author work within an active network of colleagues over time, or does he work with a small group of people, such as his/her students? Another point concerns the ability to create some new concepts, algorithms or data which have a great influence afterwards within the NLP field. Of course, this last point is difficult to measure and we will make the hypothesis that an approximation is the ability to introduce for the first time a term which becomes popular afterwards.

5. Known limitations

Our study, and more precisely our computation, is based on a large and fully populated corpus but it is a demarcated domain, namely NLP and our computations stick to this data. The benefit of such an approach is that the computed results are homogeneous, and thus provide a good picture of the NLP ecosystem. The disadvantage of such an approach is that we do not take into account external references in NLP articles to other communities like psychology or mathematics. Conversely, we do not study the reverse references and impact of NLP upon other domains like business oriented publications when referring to NLP applications or opportunities, for instance. Another limitation concerns the type of material that we count. We base our computations on published scientific articles in conferences and journals with peer review. We do not have access to thesis and books, so we cannot count them. We do not consider workshops as they may differ in the way the reviewing is conducted. We also do not take into account demo presentations, round table abstracts and prefaces as the abstract and reference sections are generally missing, a peculiarity which may also introduce a statistical bias. But more importantly, and especially in the private economic sector, a big amount of energy in our domain is devoted to program development and linguistic description, and if these authors do not publish[8], we cannot consider their work.

6. Related works

There are numerous works in the literature on scientific corpora. Important early landmarks include works by (De Solla Price, 1965), (Xhignesse et al, 1967) and (Pinski et al, 1976). See also (Banchs, 2012) (Radev et al, 2013) and (Mariani et al, 2015) for modern bibliographic references.

Concerning notability, a first and direct approach is to consider somebody as notable when this is an entry in Wikipedia. However, this position does not resolve the problem but just jumps to another question which is how to determine what should be an entry within Wikipedia. In fact, the rules are rather complex and based on a compromise between two positions: the 'inclusionism' and the 'deletionism'[9], the only point of agreement being that the entry should have reliable sources. The other serious problem is that our authors are, for most of them, not entries within Wikipedia. Another strategy is to parse citations and to compute an H-Index (or Hirsch number) which attempts to measure the productivity and citation (Hirsch, 2005). The definition is that an author with an index *h* has published *h* papers each of which has been cited in other papers at least *h* times, but this index does not take into account the collaborative and innovative aspects. There is also the i10-Index introduced in Google Scholar[10] defined as the number of publications which have at least 10 citations from other authors, but this index has the same limitations as the H-Index.

7. Main properties

The main factors we take into account are:

- **Production,** defined as the number of articles published by the author.
- **Citation,** defined as the number of citations of the papers published by the author within the domain of study.
- **Collaboration,** as how central is the author within the collaboration network.
- **Innovation,** as the impact of the terms that the author introduced in the research domain.

8. Production

We rank the authors with respect to the number of articles they publish within the NLP4NLP corpus. The number of articles is important. Of course there are notable exceptions like Kishore A Papineni, as mentioned above, but in general, for the top ten, the more an author publishes, the more he is cited. When dealing with the most prolific authors of our domain like Shrikanth S Narayanan (338 articles) or Hermann Ney (322 articles), it is worth noting that their publication rate is impressive (resp. 15.4 and 10.4 articles per year) as well as the length of their period of publication (resp. 22 and 31 years).

9. Citation

Citation is another indicator to assess the level of quality and influence of people and documents (Borgman et al, 2002)(Moed, 2005). From the reference section of each document, the 314,071 citations has been automatically extracted by means of a « robust key » in order to deal with the typographical variations that inevitably appear, see (Mariani et al, 2014) for details. It should be noted that we only count internal references from an NLP4NLP article to an NLP4NLP article, the variations in form of the reference section prohibiting any other reliable counting. The 10 most cited documents are as follows:

[8] In private companies, the employees are often not allowed to publish. They can file a patent and possibly contribute to changes in our every life through final products, but we cannot count these contributions.

[9] https://en.wikipedia.org/wiki/Notability_in_the_English_Wikipedia

[10] https://en.wikipedia.org/wiki/Google_Scholar

Title	Corpus	Year	Authors	#References	Rank
Bleu: a Method for Automatic Evaluation of Machine Translation	acl	2002	Kishore A Papineni, Salim Roukos, Todd R Ward, Wei-Jing Zhu	1516	1
Building a Large Annotated Corpus of English: The Penn Treebank	cl	1993	Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz	1145	2
Moses: Open Source Toolkit for Statistical Machine Translation	acl	2007	Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst	856	3
A Systematic Comparison of Various Statistical Alignment Models	cl	2003	Franz Josef Och, Hermann Ney	853	4
SRILM - an extensible language modeling toolkit	isca	2002	Andreas Stolcke	833	5
Statistical Phrase-Based Translation	hlt, naacl	2003	Philipp Koehn, Franz Josef Och, Daniel Marcu	830	6
The Mathematics of Statistical Machine Translation: Parameter Estimation	cl	1993	Peter E Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer	815	7
Minimum Error Rate Training in Statistical Machine Translation	acl	2003	Franz Josef Och	722	8
Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models	csal	1995	Chris Leggetter, Philip Charles Woodland	565	9
Suppression of acoustic noise in speech using spectral subtraction	taslp	1979	Steven F Boll	561	10

Table 1: 10 most cited documents

The ten most cited authors are as follows:

Name	Rank	#References	Nb of papers written by this author	Ratio #references / nb of papers written by this author	Percentage of self-citations
Hermann Ney	1	5201	343	15.163	17.554
Franz Josef Och	2	4099	42	97.595	2.220
Christopher D Manning	3	3946	116	34.017	5.094
Philipp Koehn	4	3115	41	75.976	2.536
Andreas Stolcke	5	3086	130	23.738	7.388
Dan Klein	6	3077	99	31.081	7.540
Michael John Collins	7	3063	53	57.792	3.657
Mark J F Gales	8	2549	195	13.072	19.145
Salim Roukos	9	2504	67	37.373	2.196
Chin-Hui P Lee	10	2334	215	10.856	18.509

Table 2: 10 most cited authors

10. Collaboration

The collaboration computations of today are based on works conducted in the 50s on the analysis of large organization networks. The aim was to choose the best structure so that the information flow could be fluent enough, taking into account various properties like robustness, for instance preventing two sub-networks to be isolated when one employee becomes sick. Here research analysis is used for Science indicators. In graph theory, there exist several types of centrality measures (Freeman, 1978)(Milojevic, 2014) classified into three main categories: *closeness, degree* and *betweenness centralities,* with some variants. The *Closeness distance* has been introduced in Human Sciences to measure the efficiency of a Communication Network (Bavelas, 1948 and Bavelas, 1950). It is based on the shortest geodesic distance between two authors regardless of the number of collaborations between the two authors. The *Closeness centrality* is computed as the average closeness distance of an author with all other authors belonging to the same connected component. More precisely, we use the *harmonic centrality* which is a refinement introduced recently by (Rochat, 2009) of the original formula to take into account the whole graph in one step instead of each connected component separately. The *degree centrality* is simply the number of different co-authors of each author, i.e. the number of edges attached to the corresponding node. The *betweenness centrality* is based on the number of paths crossing a node and reflects the importance of an author as a bridge across different sets of authors (or sub-communities). To these three main categories, a more modern family could be considered: PageRank with PageRank-related methods like Eigenfactor (Brin et al, 1998)(Waltman et al, 2014) but these algorithms are too complex to implement. It should be added that all these measures have first been developed for unweighted networks while weighted ones have been studied but their interpretation is difficult and we will not explore this direction.

The *degree centrality* is dedicated solely to measure the local collaboration of a given author, neglecting the fact that this author collaborate (or not) with authors who themselves collaborate a lot. In other words, this centrality does not inform us on the involvement of an author within a community.

The *betweenness centrality* is a measure of the robustness of a network. The score measures the control of a given node over the whole network, and so measures the power of "gatekeepers", but due to the fact that we do not take into consideration the question: what would have happened if an author had not written the article, this centrality is not well suited for our objective.

The *harmonic centrality* is the most interesting because it takes into account the relative distance (in number of edges in the graph) of an author with all the other authors: the more central he is, the higher score he gets. This computation does not presuppose a network with a single and strong center: there could be various local centers. The score just reflects the distance of an author with the center of a « cloud » of well-connected collaborators.

With the convention that d(X,Y) is the geodesic (i.e. shortest) distance from an author X to an author Y, the exact formula is as follows:

$$\textbf{harmonic centrality of X} = \sum_{d(X,Y)<\infty, X\neq Y} 1/d(X,Y)$$

11. Innovation

As said earlier, we make the hypothesis that an approximation of an author's innovation is the ability to introduce for the first time a term which becomes popular afterwards. The body of the articles has been processed by an NLP parser (TagParser, (Francopoulo, 2007)) and the technical terms were extracted following a "contrastive approach" (Drouin, 2004)(Mariani et al, 2014), excluding city names, laboratory names and author's names, unless they correspond to a specific algorithm or method. A rapid linguistic study has been conducted to regroup the most frequent terms like "HMM" vs "Hidden Markov Model", thus these strings are considered as synonyms. We then computed when and who introduced new terms, as a mark of the innovative ability of the authors, which provide an estimate of their contribution to the advances of the scientific domain. We make the hypothesis that an innovation is induced by the introduction of a term which was previously unused in the community and then became popular. The score depends on the number of uses over time. Among the 48,894 authors, a small minority of them (7,982) do not use any technical term. Thus, we consider the 40,912 authors (48,894-7,982) who used the 3M different terms contained in those documents and appearing as 23M occurrences. Among these 3M terms, 2,703 are present in the first proceedings (1965), which we consider as part of the initial background and as the starting point for the introduction of new terms, and 282,860 occur in the 2015 corpora. We then take into account the terms which are present in 2015 but not in 1965. For each of these terms, starting from the second year (1966), we determine the author(s) who introduced the term, referred to as the "inventor(s)" of the term. This may yield several author's names, as the papers could be co-authored or the term could be mentioned in more than one paper on the given year.

As a convention in the following algorithm presentation, an external usage of a term is the usage of this term by other people than its "inventor". This is important because we want to exclude names of systems or data which are specific to a specific team without any spreading within the community. Following this convention, an external document is a document whose authors are different from the inventor of the term. The exact algorithm to compute an innovation score for an author is as follows:

```
Preamble:
Let T, the set of terms and let A, the set of authors:
Every author a (from A) invented a certain number of terms (from T)
which form the set Na (possibly empty) of terms.

Algorithm:
Step#1: whose aim is to compute termScore(t), which is the score of
term t, as follows:
For all terms, t in T:
  termScore(t)= 0
  For all the years:
        If this year is the first year
        Then
        termScore(t)+=nbOfDocsOfTheTerm/nbOfDocsOfTheYear
        Else
        termScore(t)+=nbOfExternDocsOfTheTerm/nbDocsOfTheYear
Step#2: whose aim is to compute the author score.
For all authors, a in A:
      authorScore(a)= 0
      For all the terms t of the set Na
           authorScore(a) += termScore(t)
```

12. Measure of notability

A rank is computed for each author for all the four properties mentioned above. A normed index is then computed as:

|normed index| = value (rank) / value (first rank)

Finally, our measure of notability is a composite hybrid measure defined as an arithmetic mean between the four normed ranks:

notability = (∑ (|collaboration rank| + |production rank| + |citation rank| + |innovation rank|)) / 4.

It should be noted that more complex rankings and means are technically possible but we do not see the rationale for such precisions. For instance, a percentile ranking could be computed in order to prune extreme values, but there is no rationale to justly prune these scores. In the same vein, there is no rationale to assign a different weight for each of our four properties when computing the composite hybrid measure, thus we consider them of equal importance. Finally, given the approximation attached to each of the measures, we globalized the final ranking by only considering the first decimal.

13. Final Results

The final table shows, on the left side, the four ranking and the right side gives the notability computed as a composite hybrid measure as defined in the last paragraph, with the convention that the names are presented according to the notability ranking:

Author name	Production		Citation		Collaboration		Innovation			Notability	
	normed index	rank	normed index	rank	normed index	rank	normed index	rank		globalized normed index	rank
Hermann Ney	0.958	2	1.000	1	0.989	5	0.300	21		1.0	1
Lawrence R Rabiner	0.226	110	0.448	20	0.879	204	1.000	1		0.8	2
Shrikanth S Narayanan	1.000	1	0.484	15	0.990	3	0.059	472		0.8	2
Chin-Hui P Lee	0.601	5	0.620	5	0.992	2	0.237	38		0.8	2
Mari Ostendorf	0.489	13	0.391	34	1.000	1	0.415	5		0.7	5
Li Deng	0.536	9	0.592	9	0.956	12	0.165	93		0.7	5
John H L Hansen	0.832	3	0.350	43	0.906	89	0.140	128		0.7	5
Andreas Stolcke	0.363	30	0.740	4	0.949	18	0.138	131		0.7	5
Mark J F Gales	0.545	8	0.607	8	0.921	50	0.088	280		0.7	5
Alex Waibel	0.578	6	0.404	30	0.973	9	0.192	65		0.7	5

Table 3: Final results: 10 top authors according to the notability measure

14. Discussion

Another direction of study is to start from this notability results and to compute the relations between these most notable authors and try to answer to questions like: do they cite each other, or do they belong to separate communities? Another track is to study the relation between these notable authors and the topics and sub-domains of the NLP community. For somebody who knows our domain, an immediate comment may be expressed: all these authors mainly publish in the sub-domain of speech rather than on texts. This point seems to correlate with the level of production associated with each of the two sub-domains.

15. Conclusion

In this analysis exercise, we demonstrated the possibility to compute a measure of notability based on production, citation, collaboration and innovation. This experiment can therefore be applied easily to any other scientific and technical domain. However, we are aware that our computations do not address the notability outside a given domain. This is out of reach: such a work would require a volume and diversity comparable to the one of Google Scholar, which is not our current situation.

16. Bibliographical References

Banchs, R. (ed.) 2012 *Proceedings of the ACL 2012 Special Workshop on Rediscovering 50 Years of Discoveries, Jeju, Korea*.

Bavelas, A. (1948) “A mathematical model for small group structures.” *Human Organization* 7: 16-30.

Bavelas, A. (1950) “Communication patterns in task oriented groups.” *Journal of the Acoustical Society of America* 22: 271-282.

Brin S., Page L. (1998), The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems,* 30(1-7), 107-117.

Borgman, C.L., Furner J. (2002), Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology,* 36, 3-72.

De Solla Price, D.J. (1965) Networks of scientific papers. *Science* 149(3683), 510-515.

Ding, Y., Rousseau, R., Wolfram, D. (2014), Measuring Scholarly Impact: methods and practice (ed), Springer.

Drouin, P. 2004. Detection of Domain Specific Terminology Using Corpora Comparison, in *Proceedings of LREC 2004,* 26-28 May 2004, Lisbon, Portugal.

Freeman, L.C. (1977). A set of measures based on betweenness. *Sociometry* 40: 35-41.

Freeman, L.C. (1978) Centrality in Social Networks, Conceptual Clarifications. *Social Networks.* 1 (1978/79) 215-239.

Francopoulo, G., George, M., Calzolari N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006), Lexical Markup Framework (LMF), *Proceedings of LREC 2006,* Genoa, Italy.

Francopoulo, G. (2007), TagParser: well on the way to ISO-TC37 conformance. *ICGL (International Conference on Global Interoperability for Language Resources),* Hong Kong, PRC.

Francopoulo, G., Mariani, J., Paroubek, P. (2015) NLP4NLP: The Cobbler’s Children Won’t Go Unshod, *4th International Workshop on Mining Scientific Publications (WOSP2015), Joint Conference on Digital Libraries* 2015, June 24, 2015, Knoxville, USA.

Hirsch, J.E (2005) An index to quantify an individual’s scientific research output. *Proceedings of the national Academy of Sciences of the United States of America,* 15 Nov 2005.

Mariani, J., Paroubek, P., Francopoulo, G., Hamon, O. (2014), Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis, *Proceedings of LREC* 2014, 26-31 May 2014, Reykjavik, Iceland.

Mariani, J., Paroubek, P., Francopoulo, G., Hamon, O. (2015), Rediscovering 15+2 Years of Discoveries in Language Resources and Evaluation. *Language Resources and Evaluation,* Springer (to appear).

Moed, H.F (2005), Citation Analysis in research evaluation, ISBN: 978-1-4020-3713-9, Springer.

Milojevic, S. (2014), Network Analysis and Indicators, in (Ding et al, 2014).

Papineni, K.A, Roukos, S., Ward, T.R., Zhu, W-J. (2002), BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA.

Pinski, G., Narin F. (1976), Citation influence for journal aggregates of scientific publications: Theory with application to the literature of physics. Information Processing & Management, 12(5).

Radev D.R, Muthukrishnan Pradeep, Qazvinian Vahed, Abu-Jbara, Amjad (2013). The ACL Anthology Network Corpus, *Language Resources and Evaluation* 47: 919–944, Springer.

Rochat, Y. (2009), Closeness centrality extended to unconnected graphs: The harmonic centrality index. *Applications of Social Network Analysis (ASNA),* 2009, Zurich, Switzerland.

Waltman, L., Yan, E. (2014), PageRank-Related Methods for Analyzing Citation Networks, in (Ding et al, 2014).

Xhignesse, L.V., Osgood, C.E. (1967), Bibliographical citation characteristics of the psychological journal network in 1950 and in 1960. *American Psychologist*, 22(9), 778-791.