which should be cited to refer to this work.

Empirical analysis on a keyword-based semantic system

Zi-Ke Zhang[1], Linyuan Lü[1], Jian-Guo Liu[1,2], and Tao Zhou[1,2,a]

[1] Department of Physics, University of Fribourg, Chemin du Musée 3, 1700 Fribourg, Switzerland
[2] Department of Modern Physics, University of Science and Technology of China, 230026 Hefei Anhui, P.R. China

Abstract. Keywords in scientific articles have found their significance in information filtering and classification. In this article, we empirically investigated statistical characteristics and evolutionary properties of keywords in a very famous journal, namely *Proceedings of the National Academy of Science of the United States of America* (PNAS), including frequency distribution, temporal scaling behavior, and decay factor. The empirical results indicate that the keyword frequency in PNAS approximately follows a Zipf's law with exponent 0.86. In addition, there is a power-low correlation between the cumulative number of distinct keywords and the cumulative number of keyword occurrences. Extensive empirical analysis on some other journals' data is also presented, with decaying trends of most popular keywords being monitored. Interestingly, top journals from various subjects share very similar decaying tendency, while the journals of low impact factors exhibit completely different behavior. Those empirical characters may shed some light on the in-depth understanding of semantic evolutionary behaviors. In addition, the analysis of keyword-based system is helpful for the design of corresponding recommender systems.

PACS. 89.75.-k Complex systems – 05.65.+b Self-organized systems – 05.10.-a Computational methods in statistical physics and nonlinear dynamics

1 Introduction

The study on semantics has a long history from its birth by Breal in 1893. It has been acquainted as a branch of glossology. The modern semantic theory begins with the book, *Course in General Linguistics*, authorized by Saussure [1]. As pointed out by Graemes [2], semantics does not aim at making description of every word in the natural language, but establishing the fundamental of descriptive meta-language, according to which we can record and unify the procedure of content description.

The traditional semasiology analyzes the evolutionary properties of the acceptation mainly from the historical viewpoint, whereas the modern theory extends the horizon to the selection of new expressions, the existing and vanishing of phrases, systematicness of acceptation and the meaning of sentences. Recently, as a new interdisciplinary issue, semiotic dynamics has attracted more and more attention from different scientific communities. Compared with the traditional glossology and semasiology, the semiotic dynamics treats word and morpheme as the basic unit of content, and focuses on the understanding of how our communication pattern affects the human semantic system, as well as the underlying mechanism of evolution, emergence, self-organization and self-adaptation of the semantic system [3–5]. Therefore, semiotic dynamics not only extends the research scope of traditional semasiology,

but also contributes to the understanding of the characteristics of human language system, including the evolving properties, competition between different terms, the birth and fashion of new words, and so on [6,7].

The first step of the study on semiotic dynamics is to extract the representative morphemes, such as tags and keywords of text, and find out their relations. The mainstream methods include the Vector Space Model (VSM) [8–11] and the Ontology-Based Model (OBM) [12,13]. VSM is an algebraic model, which describes text documents as vectors of identifiers. In VSM, a document is represented as a vector, and each dimension corresponds to a separate term. Several methods have been developed to calculate the values, and one of the well-known ways is the *TI-IDF* weighting [14]. The weight vector for document d can be defined as:

$$V_d = [W_{1,d}, W_{2,d}, ..., W_{N,d}]^T, \tag{1}$$

where

$$W_{t,d} = tf_t \log \frac{|D|}{|t \in d|} \tag{2}$$

tf_t is the frequency of term t in document d. $|D|$ is the total number of documents, and $|t \in d|$ is the number of documents containing the term t. The online recommender system *Fab* [15] is a typical application of VSM. However, VSM neglects the semantic content and thus its accuracy is sensitive to the word-cutting algorithm. Comparatively,

---

[a] e-mail: `zhutou@ustc.edu`

OBM uses ontologies to describe the relationship between terms. An ontology is a set consisted of abstracts, concepts and relations by which we wish to conceptualize for the target world. The most typical kind of ontology in the web has a taxonomy and an interface rules set. The taxonomy defines the classes of terms and relations among them, while interface rules make the terms more useful and meaningful to users [17]. An ontology-based lexical database, namely *WordNet* [18], is generic ontology and free for research purposes. There are also many limitations in the OBM, for the relations between morphemes cannot be changed after the definition of a domain ontology [12,13]. In addition, the keywords in the text with special functions are usually confined within a previously defined set of words, which update generally slower than the frontier of the corresponding subjects. For instance, the articles in Physical Reviews (A–E, L) are labeled by PACS Numbers, which can only be selected from a standard set. The analysis on these kinds of semantic systems can partly exhibit the correlations and statistical evolutionary properties of keywords [19,20], however, the establishment of this set of words involves strong external disturbances, which hinder the understanding of essential evolving properties driven by the semantic system itself.

There are many ways to identify the semantic characteristics of academic articles which contains plentiful language signs. Thereinto the keywords, being seriously selected by authors and/or editors, could properly represent the main content of the corresponding article. Hence the semantic analysis of keywords can not only avoid the above limitations in VSM and OBM, but also shed some light on the in-depth understanding of the macroscopic evolutionary properties of scientific activities. In this paper, based on the data of a very famous scientific journal, we investigated the frequency distribution, the temporal scaling behavior, and the decay factor of keywords.

The rest of this paper is organized as follows. In Section 2, we introduce the empirical data. Section 3 shows the empirical results, including the Zipf's plot of keyword frequency, the temporal scaling behavior, and the decay factor. Finally, we summarize our findings and outline some open problems of related topics in Section 4.

2 Data

In order to ensure the authority and representation of our empirical analysis, we choose a journal, namely *Proceedings of the National Academy of Science of the United States of America* (PNAS), which is a very famous scientific journal among the world. PNAS was found in 1915 with one volume per year. It publishes the original research articles and reports the important academic activities. We have applied a Java script program to automatically download the keywords of each article in PNAS from the *Web of Science*. Since the articles published from 1915 to 1990 do not have keywords, our analysis is limited in the collected data from 1991 to 2006 (the documents without keywords, such as *Correction* and *Addition*, are not considered in our analysis), which is consisted of 46 149 articles
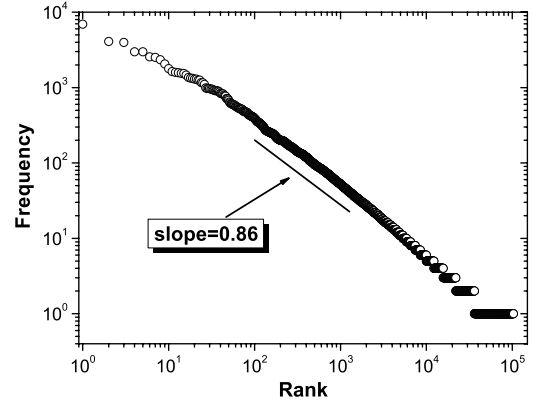
Fig. 1. The frequency of keywords in PNAS follows a Zipf's law with exponent 0.86 ± 0.01.

and 466 470 keywords. Those keywords are the combination of two parts: the ones added by authors, and the ones proposed by editors (namely Keywords Plus). Note that, some keywords are very popular and have been used in many articles, thus the number of distinct keywords, 102 992, is much smaller than the number of keyword occurrences (i.e., 466 470). Hereinafter, when referring to the number of keywords, we mean the total number of keyword occurrences. For example, if there are two articles, one has keywords A, B and C, while the other has keywords C and D. Then, we say there are 5 keywords, and 4 distinct keywords. Data of some other journals are also analyzed for comparison (see below). To be comparable, we also extract the data from 1991 to 2006.

3 Statistical analysis

3.1 Zipf's law of keywords' occurrences

In 1930s, Zipf put forth a famous law of frequency distribution of phrases, namely *Zipf's Law* [21], which has been widely used to characterize the distributions of firm size [22,23], city scale [24], wealth [25,26], earthquake strength [27], and so on. Rank the phrases in a descending order according to their occurring frequency in a text, Zipf found a power-law relation between the rank, n, and its corresponding frequency, P_n, as:

$$P_n = n^{-\alpha}. \tag{3}$$

As shown in Figure 1, the frequency distribution of keywords in PNAS approximately follows a Zipf's law with exponent 0.86 crossing 4 magnitudes. Most keywords are of low frequencies, while a few popular keywords appear very frequently. Up to 2006, the most popular keyword, *Expression*, has been used for 6927 times. Meanwhile, there are 66 782 (64.84%) distinct keywords used only once. As shown in Table 1, this Zipf's law is universally existed for various scientific journals in different subjects.

3.2 Scaling between the number of distinct keywords and the total number of keywords

A keyword in a new publication is either a new one or has appeared in a prior article. Denote τ the cumulative

Table 1. Statistics of several journals from different subjects, including Appl. Phys. Lett. (APL), British J. Pharmacology (BJP), EMBO J. (EMBOJ), Annals of Neurology (AN), SIAM J. Appl. Math. (SIAM), Chin. Sci. Bull (CSB), Czech. J. Phys. (CJP), J. Chem. Soc. Pakistan (JCSP). α is the exponent in the Zipf's plot and λ is the scaling exponent defined in equation (4). IF stands for the impact factor of the journal in 2007.

Journal title	IF	α	λ
APL	3.596	1.01 ± 0.01	0.683 ± 0.008
BJP	3.767	0.92 ± 0.01	0.753 ± 0.006
PNAS	9.598	0.86 ± 0.01	0.750 ± 0.007
EMBOJ	8.662	0.86 ± 0.01	0.753 ± 0.003
AN	8.813	0.83 ± 0.02	0.716 ± 0.005
SIAM	1.026	0.58 ± 0.02	0.825 ± 0.005
CSB	0.77	0.51 ± 0.01	0.857 ± 0.013
CJP	0.423	0.48 ± 0.01	0.912 ± 0.002
JCSP	0.095	0.39 ± 0.01	0.916 ± 0.004

Fig. 2. The scaling relation between τ and $N(\tau)$. The 16 points, from left to right, represent the cumulative data. That is to say, the leftmost point corresponds to the cumulative value up to the year 1991, the second point from left denotes the cumulative value up to the year 1992, etc.

number of keywords, and $N(\tau)$ the corresponding cumulative number of distinct keywords. Figure 2 presents a power-law relation between τ and $N(\tau)$ during the evolving process from the year 1991 to 2006. The dash line, with slope 0.750 ± 0.007, is the linear fitting of the data, that is to say,

$$N(\tau) = c\tau^{\lambda}, \qquad (4)$$

where $\lambda \approx 0.75$, and c is a constant. From equation (4), one can derive that the growing rate of distinct keywords is $c\lambda\tau^{\lambda-1}$, where τ is the number of keywords. When $\lambda = 1$, there exists a linear relation between the number of newly added distinct keywords and that of the newly added keywords, and the growing rate is a constant c. When $\lambda < 1$, the growing rate of distinct keywords will decrease with the increasing of the total number of keywords. Actually, if the number of distinct keywords is N, the probability that the next keyword has not been used before (i.e., distinct) is equal to $c^{2\lambda-1}\lambda N^{1-\frac{1}{\lambda}}$, which will decrease with the increasing of N when $\lambda < 1$. The data from some other scientific journals indicate the universality of this scaling law (see Tab. 1).

Surprisingly, some recently empirical studies demonstrate the extensive existence of this kind of scaling law, with the same form as equation (4), in the web tag systems [28–30]. Note that, the collaborative tagging system is an open and optional system where each user can optionally modify the tags in the system. In contrast, the keywords in articles are considered seriously by the authors and editors, so the keyword-based systems are more canonical and serious. However, both tags and keywords follow the same scaling law. This result indicates a possibly universal law for the generic semantic systems.

3.3 Decaying behavior of the most popular keywords

The decay factor r_t of a keyword describes the collective decay of attention, which can be defined as [31]:

$$r_t = \frac{\log N_t - \log N_{t-1}}{\log N_1 - \log N_0}, \qquad (5)$$
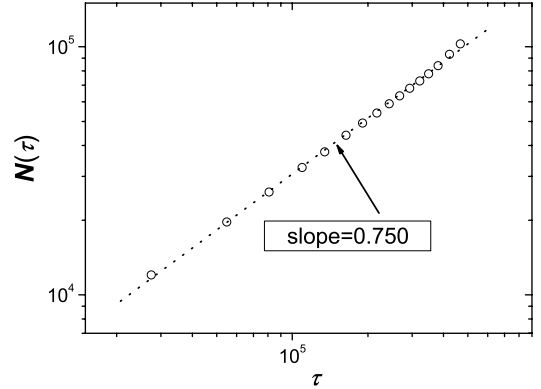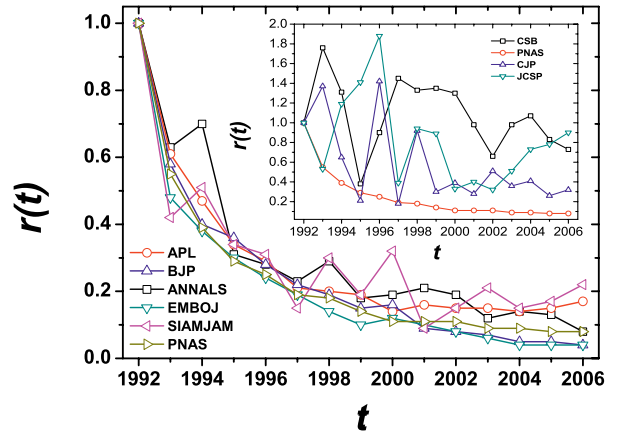

Fig. 3. (Color online) The decay factor r_t as a function of time (year resolution) for different journals. The inset compares PNAS and several local journals with much lower impact factors. The full titles of the journals can be found in Table 1.

where N_t denotes the cumulative occurring frequency of the monitored keyword at time t with year resolution. N_0 is the occurring frequency in the first year (i.e., the year 1991). In order to reduce the fluctuation, when analyzing the decay factor, we use the aggregated data of several keywords, thus the equation (5) should be rewritten as:

$$r_t = \frac{E(\log N_t) - E(\log N_{t-1})}{E(\log N_1) - E(\log N_0)}, \qquad (6)$$

where $E(\cdot)$ denotes the average over the monitored set of keywords. We analyze the decay factor of ten most popular keywords (top-10 keywords for short) in the year 1991. As shown in Figure 3, r_t of PNAS (red circles) decays very fast in the first three years, and then slows down. The decay factor almost decreases to a half in 1993. Actually, its decaying trend can be well fitted by an exponential function as
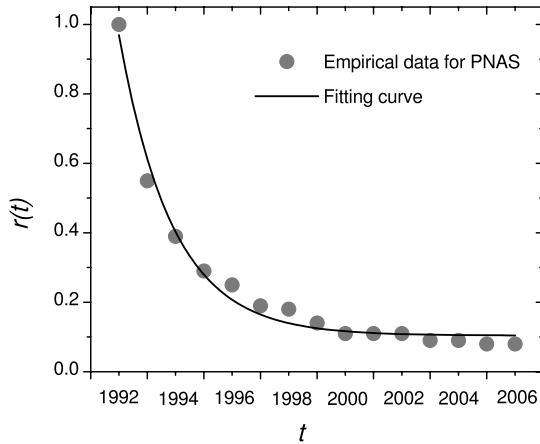
$$y = A_1 e^{-x/t_1} + y_0, \qquad (7)$$

Fig. 4. The decaying trend of top-10 keywords in the year 1991 for PNAS. The circles represent empirical result obtained by equation (6), while the solid curve corresponds to the fitting function shown as equation (7).

where $y_0 = 0.10 \pm 0.01$, $A_1 = 1.47 \pm 0.09$, $t_1 = 1.88 \pm 0.13$, and the time x varies from 1 (the year 1992) to 15 (the year 2006). The fitting curve versus empirical result is shown in Figure 4. This decaying trend can be used to quantify the broadness of interests of a journal. For a journal with high impact factor, it is possible and reasonable that r_t decays very fast in the early stage since it mainly publishes the newest progress in natural science with some new concepts.

We also empirically study the decaying behavior of top-10 keywords for several top journals in different subjects, from biology to mathematics. As shown in Figure 3, all those decaying curves display similar tendency. In contrast, as shown in the inset of Figure 3, r_t of three local journals with relatively lower scientific impacts have far different shapes compared with those top journals. Actually, the decay factor r_t exhibits large fluctuation and no obvious decaying tendency can be observed even in a long period of time (15 years). A possible reason is that those journals with low impacts do not publish as many newest progresses as top journals.

4 Conclusion and discussion

In this paper, we empirically investigated the statistical characteristics and the evolutionary properties of keywords in a very famous journal, namely PNAS, including the frequency distribution, the temporal scaling behavior, and the decay factor. Firstly, the empirical results indicate that the frequency distribution of keywords in PNAS approximately follows a Zipf's law with exponent 0.86, which means only a few keywords are used frequently in PNAS, whereas most of keywords are used unusually. Secondly, there is a power-low correlation between the number of distinct keywords and the total number of keyword occurrences. We have also investigated the data of some other journals in different subjects, which strongly indicate the universality of those two statistical properties. In addition,

we studied the decay factor of the most popular keywords. Interestingly, the top journals, though from far different subjects, exhibit very similar decaying behavior that can be approximately fitted by an exponential function. While the journals with lower impact factors exhibit very different behaviors, actually, no obvious decaying tendency is observed.

The studies of systems with collaborative keywords are also relevant to the recent progress on the design of recommender systems. Actually, with the advance of *Web2.0* technique, a great number of recommendation algorithms were applied to some on-line resource-sharing systems [33], which can recommend music, films, books and news to users according to their historical activities. Up to now, the most accurate algorithm is content-based [34]. However, those content-based methods are practical only if the items have well-defined attributes, which can be extracted automatically. The traditional content analyzing approach, based on cutting the content word by word, is often impractical since its computational complexity is too high for the huge-size database. In contrast, the structure-based algorithm has less complexity but also lower accuracy [35–37]. Because the keywords of an article can express, to some extent, the main content of this article, an algorithm with low complexity and high accuracy is expected by properly integrating the recommendations drawn from the keyword-article bipartite graph and the author-article bipartite graph (see Ref. [35] how to get recommendations from a bipartite graph).

In addition, a Keyword-Based Collaboration Network (KBCN) can be constructed based on the definition that two keywords are connected if they appear together in at least one article. More characters about the structural organization of a keyword-based semantic system can be analyzed with the help of KBCN (see Refs. [38–40] how to construct and analyze collaboration networks). Especially, the in-depth understanding of the hierarchical organization [41], the community structure [42] and the motif density [43] are crucial for the classification of research areas and the evaluation on the strength of interdisciplinary studies.

References

1. F. De Saussure, *Course in General Linguistics* (McGraw-Hill, New York, 1966)
2. A.J. Greimas, *Semantique Structurale* (Larousse, Paris, 1966)
3. S. Golder, B.A. Huberman, J. Inform. Sci. **32**, 198 (2006)
4. L. Steels, F. Kaplan, Lect. Notes on Artificial Intelligence **1674**, 679 (1999)
5. L. Steels, IEEE Intell. Syst. **21**, 32 (2006)

6. M.A. Nowak, N.L. Komarova, P. Niyogy, Nature **417**, 611 (2002)
7. E. Lieberman, J.B. Michel, J. Jackon, T. Tang, M.A. Nowak, Nature **449**, 713 (2007)
8. N. Belkin, W.B. Croft, Commun. ACM **35**, 29 (1992)
9. J. Mostafa, S. Mukhopadhyay, M. Palakal, W. Lam, ACM Trans. Inform. Syst. **15**, 368 (1997)
10. M. Kobayashi, K. Takeda, ACM Computing Surveys **2**, 144 (2000)
11. S. Lin, ACM SIGIR Conference **8**, 241 (1998)
12. R. Neches, R.E. Fiches, T. Finin, T. Guber, R. Patil, T. Senator, W.R. Swartout, AI Magazine **12**, 36 (1991)
13. S. Staab, R. Studer, *Handbook on Ontologies* (SpringerVerlag, 2004)
14. G. Salton, *Automatic Text Processing* (Addison-Wesley, 1989)
15. M. Balabanovic, Y. Shoham, Commun. ACM **40**, 66 (1997)
16. S. Roser, B. Bauer, *Lecture Notes in Computer Science* **3844**, 355 (2006)
17. T. Berners-Lee, J. Hendler, O. Lassila, Sci. Ame. **284**, 34 (2001)
18. G. Miller. Commun. ACM **38**, 39 (1995)
19. J.G. Liu, Y.Z. Dang, Z.T. Wang, Physica A **366**, 578 (2006)
20. J.G. Liu, Z.G. Xuan, Y.Z. Dang, Q. Guo, Z.T. Wang, Physica A **377**, 302 (2007)
21. G.K. Zipf, *Selective Studies and the Principle of Relative Frequency in Language* (MIT Press, Cambridge, 1932)
22. M.H.R. Stanley, S.V. Buldyrev, S. Havlin, R.N. Mantegna, M.A. Salinger, H.E. Stanley, Economics Lett. **49**, 453 (1995)
23. R.L. Axtell, Science **293**, 1818 (2001)
24. K.T. Rosen, M. Resnick, J. Urban Economics **8**, 165 (1980)
25. M. Levy, S. Solomon, Physica A **242**, 90 (1997)
26. Y.B. Xie, B.H. Wang, B. Hu, T. Zhou, Phys. Rev. E **71**, 046135 (2005)
27. P. Bak, C. Tang, J. Geophys. Res. **94**, 15635(1989)
28. C. Cattuto, A. Baldassarri, V.D.P. Servedio, V. Loreto, e-print `arXiv: 0704.3316`
29. C. Cattuto, V. Loreto, L. Pietroner, Proc. Natl. Acad. Sci. USA **104**, 1461 (2007)
30. H. Halpin, V. Robu, H. Shepherd, *Proc. 16th Int. Conf. WWW*, 2007, p. 211
31. F. Wu, B.A. Huberman, Proc. Natl. Acad. Sci. USA **104**, 17599 (2007)
32. R. Albert, A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
33. G. Adomavicius, A. Tuzhilin, IEEE Trans. Know. & Data Eng. **17**, 734 (2005)
34. M.J. Pazzani, D. Billsus, Lect. Notes Comput. Sci. **4321**, 325 (2007)
35. T. Zhou, J. Ren, M. Medo, Y.C. Zhang, Phys. Rev. E **76**, 046115 (2007)
36. Y.C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, F. Yang, Europhys. Lett. **80**, 68003 (2007)
37. T. Zhou, L.L. Jiang, R.Q. Su, Y.C. Zhang, Europhys. Lett. **81**, 58004 (2008)
38. M.E.J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001)
39. P.P. Zhang, K. Chen, Y. He, T. Zhou, B.B. Su, Y.D. Jin, H. Chang, Y.P. Zhou, L.C. Sun, B.H. Wang, D.R. He, Physica A **360**, 599 (2006)
40. T. Zhou, B.H. Wang, Y.D. Jin, D.R. He, P.P. Zhang, Y. He, B.B. Su, K. Chen, Z.Z. Zhang, J.G. Liu, Int. J. Mod. Phys. C **18**, 297 (2007)
41. E. Ravasz, A.L. Barabási, Phys. Rev. E **67**, 026112 (2003)
42. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002)
43. W.K. Xiao, J. Ren, F. Qi, Z.W. Song, M.X. Zhu, H.F. Yang, H.Y. Jin, B.H. Wang, T. Zhou, Phys. Rev. E **76**, 037102 (2007)