Approaches to assessing the semantic similarity and future citation of publications by identifying informative terms with predictive properties

A.Kh. Khakimova[1], M.M. Charnine[2]
aida_khatif@mail.ru | mc@keywen.com
[1]ANO «Scientific and Research Center for Information in Physics and Technique», Nizhny Novgorod, Russia;
[2]FRC CSC of the Russian Academy of Sciences, Moscow, Russia

The article discusses new approaches to assessing the semantic similarity of documents in a vector space, taking into account statistically significant and informative terms. Informative terms reflect the current state of research in a certain field of research. To select informative terms, an algorithm for calculating the impact factor of the term is proposed. It is shown that informative terms allow both to evaluate the semantic similarity of texts and to predict future citations. The developed methods for assessing the semantic similarity and future impact of scientific publications can be used in the framework of "Predictive optimization", a modern technology that allows us to make decisions based on forecasts. In evaluating the activities of research and individual scientists, bibliometric indicators often play an important role. However, the use of citation-based indicators is problematic in determining the impact of recent publications. Usually, two years after the publication of most articles, they receive only a few links. The probability of future citation can be predicted using the proposed indicator - IFT.

***Keywords**: semantic similarity, informative terms, impact factor of the term, citations, statistical analysis, citation prediction.*

1. Introduction

Measuring the similarity between documents is an important component in various tasks such as document clustering, topic detection, topic tracking, question answering, information retrieval and text summarization.

For scientific articles, there are two main types of similarity measures: citation-based similarity [1] and semantic textual similarity [2]. These two types of similarity measures should correlate and maximizing this correlation is a convenient way to adjust the coefficients and parameters on which these measures depend.

Citation-based similarity measures such as bibliographic coupling (if two documents share a reference in their bibliography) and co-citation (if two documents are cited by a third document) are an integral component of many information retrieval systems. Semantic textual similarity measures analyze situations where two documents share certain words (co-word linkages [3]), phrases or ideas [4].

Latent Semantic Analysis (LSA) [5] and Generalized Latent Semantic Analysis (GLSA) [6] are the most popular techniques of Corpus-Based semantic textual similarity [2]. GLSA extends the LSA approach by focusing on term vectors instead of the dual document-term representation.

There is a problem of efficient filtering of non-informative words. LSA and GLSA suffer from noise introduced by typos and infrequent and non-informative words [6]. To solve this problem, we present a new citation-based method for efficient filtering of the core vocabulary and keeping only content bearing words. This new citation-based method is called the Impact Factor of Terms (IFT). It is described in Section 2. IFT assesses the significance and informational content of terms in scientific articles based on citation analysis of the articles with these terms. Also, IFT is useful for prediction future citations and promising topics in different subject areas such as smart energy systems.

Maximizing correlation between citation-based similarity and IFT-based semantic textual similarity is a convenient way to adjust the coefficients and parameters of the IFT method.

IFT is similar to journal impact factor (JIF) which has been used for many years and has proven effective. JIF is a scientometric index that reflects the yearly average number of citations that articles published in the last two years in a given journal received. If all articles of a journal are highly cited, then this journal has a high JIF value and is considered significant and authoritative. Similarly, if all articles with some general term are highly cited, then this term has a high IFT value and is considered significant and informative. The IFT helps to identify informative terms that indicate significant fundamental ideas. Words and terms with a constantly high IFT (for example, neural networks) denote significant ideas, interest in which is stable for many years. For such informative words, the IFT values are stably high. Also, such words have a high correlation between IFT values of the current and next year. This correlation as well as the conditions for the stability and predictability of the IFT are discussed in Section 4. Section 3 describes a collection of articles used for experiments to study the empirical properties of IFT, including its correlations. The next section gives a formal description of the IFT.

2. Impact Factor of Terms (IFT)

There are currently several journal ranking systems, but the oldest and most influential system is a journal impact factor (JIF). JIF is used as an indicator of the importance of a journal for its field.

A journal's impact factor is based on how often articles published in that journal during the previous two years (e.g. 2017 and 2018) were cited by articles published in a particular year (e.g. 2019).

The higher the JIF, the more often articles in that journal are cited by other articles. Thus, the influence factor can give an approximate idea of how prestigious the magazine is in its field of science.

The journal with the highest IF value is the one that publishes the most frequently cited articles over a two-year period. One easy way to increase JIF is to publish more review articles, which are usually cited more often than research reports [7].

Author Impact Factor (AIF) is an extension of the impact factor for authors. The AIF of an author A in year t is the average number of citations given by papers published in year t to papers published by A in a period of Δt years before year t. AIF is able to capture trends and variations in the influence of scientists over time, in contrast to the h-index, which is a measure that takes into account the entire career path [8].

We offer an extension of the impact factor idea for terms. We offer a new numerical indicator of the authority of words and terms, called the impact factor of the term (IFT).

IFT (formula 1) can be used to effectively filter the dictionary, excluding uninformative words and terms. With the help of IFT, we can identify promising topics and ideas, find implicit links between articles and texts, and discover ideologically influential sites.

$$IFT = \frac{A_t}{N_t}, \quad (1)$$

where A_t is the number of citations in articles with the term A published in year t to articles with the term A in the period Δt years to year t; Nt - total number of articles with term A for the time period Δt + 1.

Therefore, the IFT of term A in year t is the average number of references cited in articles with term A published in year t to articles with term A in the period Δt years to year t.

It follows from the IFT formula (1) that the method will certainly increase the correlation of the similarity measure of texts with their bibliographic relationship, since the IFT linearly depends on the number of bibliographic references over the past two years (or over a period of Δt years).

Various approaches to the calculation of IFT were investigated.

The modified impact factor of the term (IFTm) is the ratio of citations of articles with term A to the total number of articles with this term over 3 years.

$$IFT_m = \frac{A_{t-2} + A_{t-1} + A_t}{N}, \quad (2)$$

where A_{t-2} - the number of links to the article with the term A two years ago in same year; A_{t-1} - the number of links to the article with term A last year for the same and previous years; A_t - the number of links to the article with term A over a three-year period, including the current year; N - total number of articles with term A for three years.

Both the IFT and IFTm are considered only for articles in which the given term is in the title. Only citations from articles containing the specified term in the title are taken into account.

3. AI collection (Data Set)

In our experiments, we analyze DBLP citation network, which is a collection of articles on Artificial Intelligence from 1936 to 2017, compiled by aminer.org and referred to here as AI collection.

The citation data is extracted from DBLP (Digital Bibliography & Library Project dblp.org), ACM (Association for Computing Machinery acm.org), MAG (Microsoft Academic Graph), and other sources.

We used the V10 version released in October 2017. This data set consists of 3,079,007 articles and 25,166,994 citation relationships. For each article there is a title, authors, year of publication and links. We have processed all titles and citation relationships.

In this paper, the AI collection was analyzed in different directions described in the next Section.

4. Results of a statistical analysis of term trends

The main goal of the statistical analysis of the AI collection is to study the empirical properties of Impact Factor of Terms (IFT), including the correlation of its current and future values to assess its stability and forecast future citations.

Statistical analysis of the collection was carried out using the Trend+ author program, which built a frequency dictionary of all words and terms in the collection. Also, for each term with a frequency of more than 5, Trend+ calculated its trend indicators (trending situations), including the number of articles with this term for the year, the number of citations from other articles with this term, the IFT and IFTm indicators for the current and next year.

To calculate the correlation, situations/points were selected for different words in different years, when the values of IFT and IFTm of the current year were more than zero. There could be several such situations for one word in different years. The selected situations were divided into groups differing in the number of articles with a word over the past 3 years. According to the number of situations, the IFTm groups turned out to be larger than the IFT groups, because IFTm takes into account more citations. Fig. 1 shows graphs of the number of situations/points in these groups for calculating correlations.

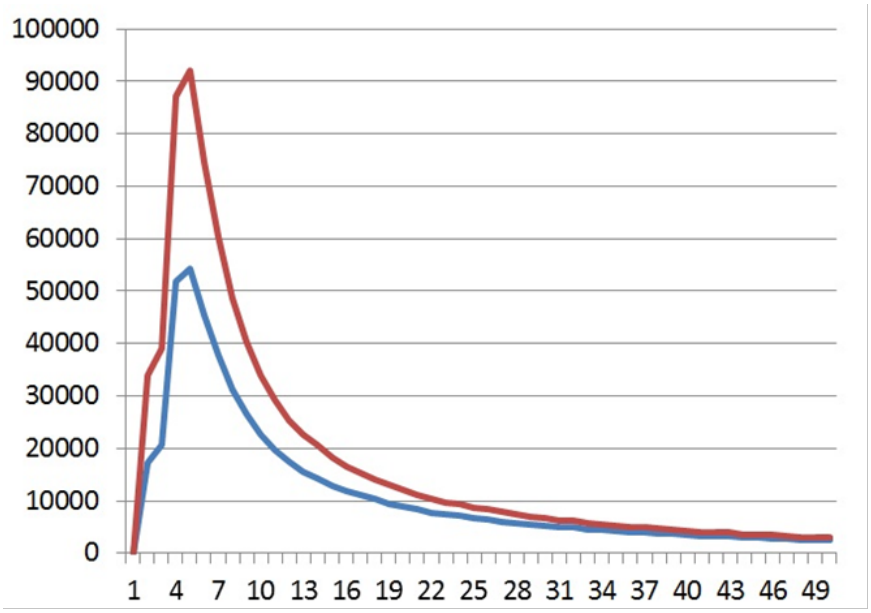

Fig. 1. Graphs of the number of points for calculating the correlations of the current and future years for the indicators IFTm (upper) and IFT, depending on the number of articles with the word in the last 3 years

In Fig. 1, the upper graph corresponds to the IFTm, and the lower IFT. The y-axis represents the number of points for calculating the correlations of the current and future years. The x-axis represents the frequency of terms, i.e. the number of articles with the term over the last 3 years. The maximum points on both graphs are achieved when the number of articles is 5, because the experiment did not analyze terms that occurred less than 5 times in the collection for all time.

On the IFT graph, the maximum number of points 54326 is reached at X = 5, and the minimum 2423 at X = 50. On the IFTm graph, the maximum number of points 91997 is reached at X = 5, and the minimum 2913 at X = 50.

For each group of trending situations/points (i.e., for each X) individually, a correlation was calculated between the current and future values of IFT and IFTm. The results of calculating the correlations are shown in Fig. 2.

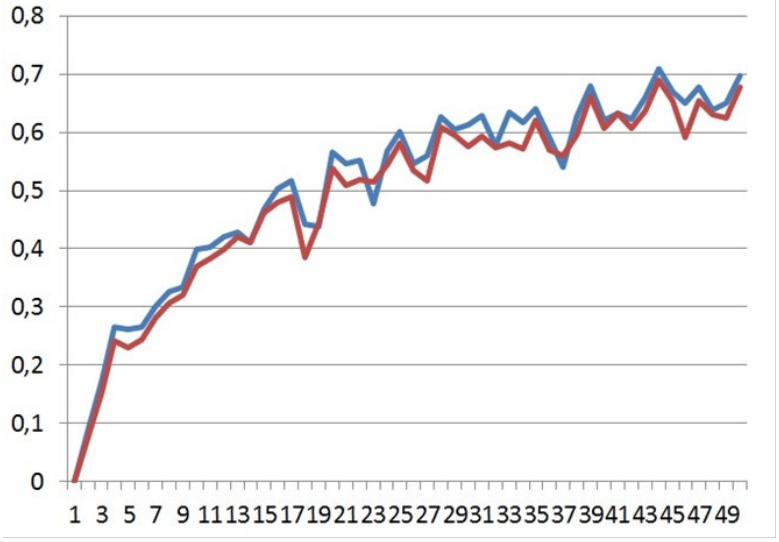

Fig. 2. Graph of IFT correlations (upper) and IFTm correlations of the current and future years depending on the number of articles with the word in the last 3 years

The upper graph is the IFT correlations, and the lower graph is the IFTm correlations.

Both graphs behave very similarly, but the correlations of the IFT (upper graph) are almost always greater than the correlations of the IFTm. The correlation on the graphs reaches 0.5 at a frequency of 17 articles over the past three years, 0.6 at 26 articles, and 0.7 at 45 articles. Thus, IFT behaves more stably and predictably than IFTm, but IFTm covers more different situations and words/terms.

The graphs show that the higher the current frequency of the term (the number of articles with the term), the higher the correlation, and therefore, the more stable the IFT behaves in time. Stable IFT allows you to accurately predict the average number of future citations, since the IFT is exactly equal to the average number of citations of articles with the specified word/term. Thus, the words/terms with a high frequencies and high IFT values

define promising topics in different subject areas such as artificial intelligence or smart energy systems.

The most stable and predictable words/terms with high IFT values are called informative terms. Informative words/terms have high frequencies and IFT meanings above a certain threshold. The type of function for filtering of non-informative words which grows with increasing IFT and frequency can be selected by maximizing the correlation between citation-based similarity and IFT-based semantic textual similarity. As a first approximation, this filtering function can be taken as the product of IFT and frequency with a certain minimum threshold for IFT.

Here are examples of the most informative words/terms in the collection of AI articles that have the largest total values of IFT multiplied by the current frequency: web (year 1982), fuzzy (1969), sensor networks (1992), neural (1962), video (1976) , social (1971), cognitive (1973), semantic (1967), clustering (1970), neural networks (1986).

These examples point to the most actively and stably developing areas of AI, and also confirm the usefulness of the proposed filtering function and its ability to evaluate the significance and information content of words/terms.

5. Predicting the citations with IFT

Prediction of citation of scientific works was studied by many researchers. The described approaches are mainly based on the analysis of a number of features, including information about the authors (number of authors, country, authors rating, etc.), features of the journal (total number of links to the journal, impact factor of the journal), article parameters (topic, volume, number of references etc.), type of research (for example, original research compared to a literature review), as well as other characteristics (reputation of institutions etc.). In addition, altmetrics are also used to predict the citation of a scientific paper.

Citation prediction methods have been proposed, for example, by Walters (2006) [9], Haslam et al. (2008) [10], Fu and Aliferis (2010) [11], Wang, Yu and Yu (2011) [12], Wang et al. (2012) [13], Didegah and Thelwall (2013) [14], Yu, Yu, Li and Wang (2014) [15], Onodera and Yoshikane (2015) [16], Cao et al. (2016) [17], Golosovsky and Solomon (2017) [18], Fiala and Tutoky (2018) [19], Bai et al. (2019) [20]. For example, Wang et al. (2013) [21] propose mathematical models that describe how publications accumulate citations over time. Using these models, the authors predict the effect of publication citation on a longer term based on a short-term publication citation history. Bornmann et al. (2013) [22] present an empirical analysis of the correlation between short-term and long-term citation indicators.

IFT evaluates the significance and informativeness of terms in scientific articles based on an analysis of the citation of articles with these terms. IFT can also be used to predict future citations of new articles.

Given the practical importance of incorporating the latest publications in evaluations of scientific performance, one of the goals of our study is to develop a model to predict the impact that recent publications will have in the long run.

Our model assumes a publication citation prediction based on the following predictors: the impact factor of significant terms (for example, authors' keywords) and the time of appearance of subsequent articles associated with implicit links to the original article.

The two predictors used are readily available, and unlike most prediction approaches, they allow you to make predictions pretty soon after the publication.

Citation forecasts have a high degree of uncertainty. Therefore, we believe that it is more important to know the likelihood that the publication will receive a certain number of links in the future. Therefore, we do not predict the average number of links that the publication should attract in the future, but we predict the probability distribution for the future number of links based on the developed mathematical probabilistic model of the dependence of the number of direct citations on terms with high IFT.

It is important to emphasize that the purpose of our work is different from the studies mentioned above. As in the above studies, we are interested in predicting the future citation. However, many indicators that have been found to correlate with the influence of citation are easy to manipulate.

For example, suppose researchers know that future citations of a publication will be predicted based, for example, on the number of pages or the number of links. In this case, authors can artificially increase the number of pages or increase the number of bibliographic references. Therefore, we consider variables that cannot be changed by the authors of the publication.

Based on IFT values, we can choose informative terms that indicate important fundamental ideas. Words and terms with a consistently high IFT indicate important ideas that have been stable for many years.

In our experiments, we analyze the DBLP citation network, which is a collection of articles on artificial intelligence from 1936 to 2017, including 3,079,007 articles and 25,166,994 links. Statistical analysis of the collection was carried out using the Trend + program, which built a frequency dictionary and trend indicators, including the number of articles with this term per year, the number of links to other articles with this term, IFT and IFTm indicators for the current and next year.

The term “Trend of the initial frequency” (TIF) is proposed - this is the number of years from the first article with a certain term to the nth article with this term. A relationship was found between TIF, IFT, and citation trends. It is shown that the higher the trends of the initial frequency, the higher the trends of fresh citation links, that is, the higher the likelihood of quick appearance of links to the article.

Of particular interest are trend terms with a large number of new articles (more than 10 articles in the previous 2 years). For trend terms, the correlation of current and future IFTm is more than 60%, which allows us to make a fairly confident forecast of IFTm (i.e. citation forecast) for the next year.

We summarize how our study differs from existing works:

- we are interested in predicting the long-term impact of citation, based solely on the impact factors of

significant terms (as mentioned above, we do not want to use variables that can be easily manipulated);
- we are interested in predicting the long-term impact of citation within one or two years after the publication;
- unlike most earlier papers, our interest is in predicting the probability distribution for the future number of links to a publication. We do not aim to give an accurate estimate of the future number of links to the publication.

Acknowledgment

The reported study was funded by RFBR according to the research projects № 18-07-00909, 19-07-00857 and 20-04-60185.

References

[1] Gipp, B. (2014). Citation-based Document Similarity. Citation-based Plagiarism Detection. Springer Fachmedien Wiesbaden, pp. 43-55.
[2] Gomaa, W.H.and Fahmy, A.A. (2013). A survey of text similarity approaches, Int. J. Comput. Appl., vol. 68, no. 13, doi: https://doi.org/10.5120/11638-7118.
[3] Leydesdor, L. (1989). Words and co-words as indicators of intellectual organization. Research Policy 18(4), pp. 209-223. DOI http://dx.doi.org/10.1016/0048-7333(89)90016-4. URL http://www.sciencedirect.com/science/article/pii/0048733389900164
[4] Charnine, M., Klimenko, S. (2015). Measuring of "Idea-based" Influence of Scientific Papers // Proceedings of the 2015 International Conference on Information Science and Security (ICISS 2015), December 14-16, Seoul, South Korea, pp.160-164.
[5] Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 104.
[6] Matveeva, I., Levow, G., Farahat, A. & Royer, C. (2005). Generalized latent semantic analysis for term representation. In Proc. of RANLP.
[7] Zaidi I, Singh S, Sinha A, Dwivedi R. (2015). Current views and implications of journal impact factor: A key note. Indian J Dent. 6(2):113-114. doi:10.4103/0975-962X.154375
[8] Pan, R., Fortunato, S. (2015). Author Impact Factor: tracking the dynamics of individual scientific impact. Sci Rep 4, 4880. https://doi.org/10.1038/srep04880.
[9] Walters, G. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact.Scientometrics, 69(3), pp. 499-510.
[10] Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., et al. (2008). What makes an article influential? Predicting impact in social andpersonality psychology. Scientometrics, 76(1), pp.169-185.
[11] Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature.Scientometrics, 85(1), pp. 257-270.
[12] Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. Scientometrics, 87(3), pp. 695-706.
[13] Wang, M., Yu, G., Xu, J., He, H., Yu, D., & An, S. (2012). Development a case-based classifier for predicting highly cited papers. Journal of Informetrics, 6(4), pp.586-599.
[14] Didegah, F., & Thelwall, M. (2013a]). Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society forInformation Science and Technology, 64(5), pp.1055-1064.
[15] Yu, T., Yu, G., Li, P.-Y. & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. Scientometrics, 101(2), pp.1233-1252.
[16] Onodera, N. & Yoshikane, F. (2015). Factors affecting citation rates of research articles. Journal of the Association for Information Science and Technology,66(4), 739-764.
[17] Cao, X., Chen, Y., Liu K.J.R. (2016). A data analytic approach to quantifying scientific impact. Journal of Informetrics, 10 (2), pp. 471-484.
[18] Golosovsky, M., Solomon S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. Physical Review E, 95 (1), p. 012324.
[19] Fiala, D., Tutoky G. (2018). PageRank-based prediction of award-winning researchers and the impact of citations. Journal of Informetrics, 11 (4), pp. 1044-1068.
[20] Wang, D., Song, C., Barabási, A.-L. (2013). Quantifying long-term scientific impact. Science, 342 (6154) , pp. 127-132.
[21] Bornmann, L., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values?an empirical comparison of five approaches including a newly developed citation-rank approach (p100). Journal of Informetrics, 7(4), pp.933-944.
[22] Bai, X., Zhang, F., Lee, I. (2019). Predicting the citations of scholarly paper. Journal of Informetrics, Volume 13, Issue 1, pp. 407-418.

About the autors

Khakimova Aida Kh., PhD, docent, Kama Institute (Naberezhnye Chelny, Russia), ANO «Scientific and Research Center for Information in Physics and Technique» (Nizhny Novgorod, Russia), E-mail: aida_khatif@mail.ru

Charnine Mikhail M., PhD, Senior Researcher, FRC CSC of the Russian Academy of Sciences, Moscow, Russia, E-mail: mc@keywen.com