Int J Digit Libr (2015) 16:25–36
DOI 10.1007/s00799-014-0132-0

Towards robust tags for scientific publications from natural language processing tools and Wikipedia

Michał Łopuszyński · Łukasz Bolikowski

Received: 10 September 2014 / Accepted: 11 October 2014 / Published online: 28 October 2014

Abstract In this work, two simple methods of tagging scientific publications with labels reflecting their content are presented and compared. As a first source of labels, Wikipedia is employed. A second label set is constructed from the noun phrases occurring in the analyzed corpus. The corpus itself consists of abstracts from 0.7 million scientific documents deposited in the ArXiv preprint collection. We present a comparison of both approaches, which shows that discussed methods are to a large extent complementary. Moreover, the results give interesting insights into the completeness of Wikipedia knowledge in various scientific domains. As a next step, we examine the statistical properties of the obtained tags. It turns out that both methods show qualitatively similar rank–frequency dependence, which is best approximated by the stretched exponential curve. The distribution of the number of distinct tags per document follows also the same distribution for both methods and is well described by the negative binomial distribution. The developed tags are meant for use as features in various text mining tasks. Therefore, as a final step we show the preliminary results on their application to topic modeling.

Keywords Tagging document collections · Natural language processing · Wikipedia · ArXiv preprint collection

M. Łopuszyński (✉)· Ł. Bolikowski
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Pawińskiego 5a, 02-106 Warsaw, Poland
e-mail: m.lopuszynski@icm.edu.pl

Ł. Bolikowski
e-mail: l.bolikowski@icm.edu.pl

1 Introduction

Text mining methods and techniques are increasingly important in the design and deployment of digital library systems. They automatically generate additional value from the stored information, which improves the way the content may be searched, presented and consumed by the end user [32]. In this work, we present a study of two methods for enriching scientific publications with compact and uniform set of tags reflecting their content. The first method is based on Wikipedia and the second approach relies on the noun phrases detected by the natural language processing (NLP) tools. Both methods are applied to the document collection consisting of abstracts from the ArXiv preprint server [2]. The motivation behind this study is threefold.

First, we would like to generate compact and meaningful features for document content representation in text mining tasks, which will go beyond the basic bag of words approach. The developed tags can serve as such features, which later on can be employed for various applications, e.g., determining document similarity, clustering and topic modeling. After the appropriate filtering and ranking, the obtained tags can also be used as keyphrases, summarizing the document. In this work, we briefly demonstrate the potential of obtained tags by using them, instead of bag of words representation, in latent Dirichlet allocation topic modeling method.

Our second goal is the comparison of the two approaches to tagging publications with labels reflecting their content. We employed two methods, abbreviated hereafter NP and WIKI. The NP approach relies on the tags' dictionary generated from noun phrases detected in the analyzed corpus using NLP tools. The WIKI method relies on the filtered set of Wikipedia multi-word entries. The tags generated by both of the methods are to a large extent independent. Therefore, a comparison of both obtained results reveals strengths and shortcomings of the NP and WIKI tags. As a side effect, on the basis of such comparison one may draw conclusions about the completeness of Wikipedia knowledge in examined scientific domains. Even though Wikipedia is often used in analysis of scientific texts, its completeness with respect to domain-specific vocabulary involved could be questioned. Many topics may be out of the scope of interest of the average Internet user, i.e., Wikipedia reader and author. When considering texts from particular field of science, the relative efficiency of the WIKI method to complementary NP approach may be used as figure of merit, crudely reflecting the knowledge coverage in Wikipedia with respect to the examined discipline.

The third goal of this work is the analysis of the statistical properties for the obtained tags. We look at the distributions of the number of different tags per document. We also examine if the Zipf's law is valid for the rank–frequency curves of labels detected by both methods. It is also interesting to check, if the aforementioned statistical properties are qualitatively similar for the NP and WIKI tags.

The paper is organized as follows. Related work is discussed in Sect. 2. In Sect. 3, the employed datasets are described. Afterward, in Sect. 4, we provide the details of both tagging procedures—the one based on Wikipedia (WIKI) and the complementary approach based on the noun phrases (NP). Comparison of both methods is the subject of Sect. 5. Statistical properties of the obtained tags are investigated in Sect. 6. The sample application of both NP and WIKI tags to the topic modeling problem is presented in Sect. 7. The results are summarized in Sect. 8.

This paper contains an extended version of the material presented on the Linking and Contextualizing Publications and Datasets Workshop, during the conference Theory and Practice of Digital Libraries 2013 [18].

2 Related work

In the first of our tagging methods, we used noun phrases as source of tags reflecting the content of a given text. This was inspired by methods employing NLP, in particular noun phrases detection, to keyword extraction problem. Barker and Cornacchia [5] filtered noun phrases according to the frequency of a head noun to obtain the best keyphrases summarizing a document. Hulth [12] employed noun phrases and part of speech patterns in algorithms for supervised keyword extraction. Chuang, Manning and Heer [9] conducted recently a large-scale research on the properties of human-assigned keyphrases. They provided solid data, confirming the intuition about the importance of noun phrases. In their study, almost 65 % of the manually assigned keyphrases were either a noun phrase or were contained in it. They also showed that a vast majority of human-assigned keyphrases consisted of multiple words (75 %, for the experiment when human experts are presented one document at a time). Therefore, in our study we also focus on multi-word tags.

Our second approach to tagging is based on Wikipedia. Wikipedia is currently very often used in studies on conceptualizing and contextualizing document collections. There is no doubt that it constitutes a very useful source of semi-structured knowledge. To name just a few recent examples of research, applications of Wikipedia knowledge in text mining tasks include: extracting keywords [13], clustering [28,29], assigning readable labels to the obtained document clusters [23,24] and facilitating classification [31]. When it comes to tagging, a method similar to ours, although much more advanced, was used by Mendes and coworkers [19]. They created DBpedia Spotlight—a system which uses DBpedia URIs to tag documents. Moreover, it allows to configure the annotations to the user needs through the DBpedia Ontology [17] and dedicated quality measures.

3 Employed datasets

The ArXiv repository [2] was started in 1991 by a physicist, Paul Ginsparg. Originally, it was intended to host documents from the domain of physics. However, later on it gained popularity in other areas. Currently, it hosts entries from physics, mathematics, computer science, quantitative biology, quantitative finance and statistics. The content is not peer reviewed; however, many documents are simply preprints, published later on in scientific journals or presented in conferences. In this work, we analyze the ArXiv publications metadata harvested via OAI/PMH protocol up to the end of March 2012. This made up to over 0.7 million of abstracts. For our study, the distribution of the manuscripts across domains is of high interest. For this purpose, we used `<setSpec>` field of the ArXiv XML format, which gives coarse-grained information about the field of document. All the ArXiv coarse-grained categories together with their full names are presented in Table 1. The percentage of documents in each category is displayed in Fig. 1. The presented values do not add up to 100 % since multiple categories per document are allowed. In this study, we have also employed Wikipedia. From its dump dated 2013.01.02, the lexicon of all page titles was created and used later on in the tagging procedure.

4 Processing methods

Our process of tagging ArXiv abstracts consisted of three phases—generating the preliminary dictionary, cleaning the dictionary and tagging itself. Only the first phase differentiated the two analyzed methods, that is, the approach employ-

Table 1 The ArXiv categories and their abbreviations

Abbreviation	Category full name
cs	Computer Science
math	Mathematics
nlin	Nonlinear Sciences
physics-astro-ph	Astrophysics
physics-cond-mat	Condensed Matter Physics
physics-gr-qc	Physics—General Relativity and Quantum Cosmology
physics-hep-ex	High Energy Physics—Experiment
physics-hep-lat	High Energy Physics—Lattice
physics-hep-ph	High Energy Physics—Phenomenology
physics-hep-th	High Energy Physics—Theory
physics-math-ph	Mathematical Physics
physics-nucl-ex	Nuclear Physics—Experiment
physics-nucl-th	Nuclear Physics—Theory
physics-quant-ph	Quantum Physics
physics-physics	Physics—Other Fields
q-bio	Quantitative Biology
q-fin	Quantitative Finance
stat	Statistics

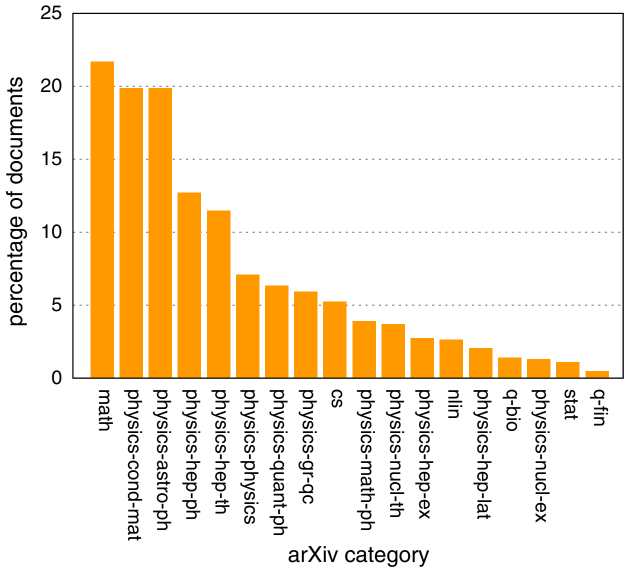


Fig. 1 The percentage of documents marked with various ArXiv categories. Note that, since multiple categories per paper are possible, the sum of the numbers above exceeds 100 %. The labels for categories are explained in Table 1

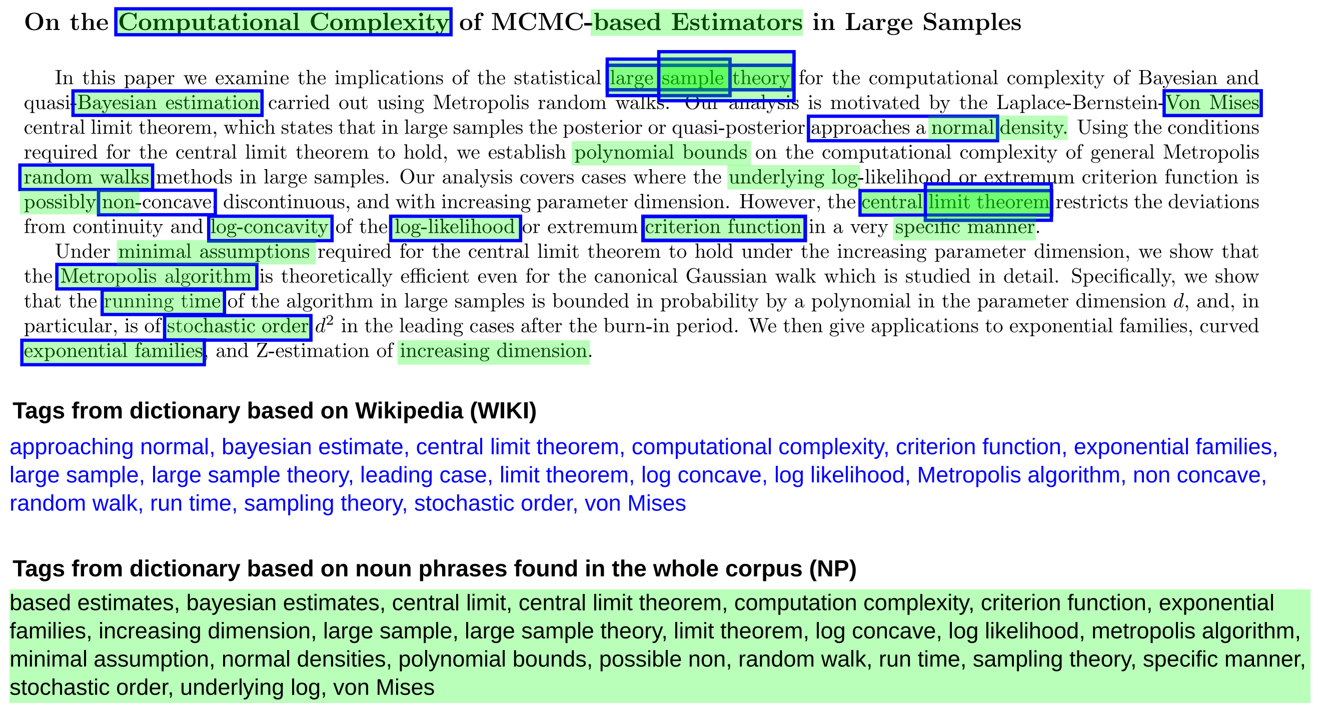
ing Wikipedia (WIKI) and the procedure making use of the noun phrases (NP).

1. *Generating the preliminary dictionary* During this stage we generated the preliminary version of the dictionary, used later on as a lexicon for tagging. For the WIKI case, simply all multi-word titles of articles form Wikipedia dump were extracted. Full texts of Wikipedia articles were not used. For the NP method all the abstract from ArXiv corpus were analyzed using general purpose natural language processing library OpenNLP [1], detecting all the noun phrases containing two or more words. Noun phrases occurring in fewer than four documents were excluded from the dictionary.
2. *Cleaning the dictionary* Clearly, on this level both dictionaries contain a lot of non-informative entries. Therefore, we apply a cleaning procedure to both preliminary tag sets. For each tag we remove the initial and final words, if they belong to the set of stopwords. The labels which contain only one word after such filtering are removed. Then we use a simple heuristic observation that good label candidates usually do not contain stopword in the middle; see the study [26] for more details. One notable exception here is the word *of*. We drop all the entries according to this heuristic rule. Naturally, many far more sophisticated algorithms can be employed here. One example might be matching a grammatical pattern devised to select true keywords, which could be employed, when the knowledge about the part-of-speech classification is available [3,14]. However, the simple stopword method worked well enough for us, especially because we mostly aimed at labels for further applications in machine learning and hence we could afford having a certain fraction of "bogus labels". The generated dictionaries after the cleaning procedure contained around 5 million entries for the WIKI method and 0.3 million for the NP case.
3. *Tagging* Finally, we tag the analyzed corpus of the ArXiv abstracts with the filtered dictionaries, obtained previously. In the process of tagging, we make use of the Porter stemming [25], to alleviate the problem of different grammatical forms. For the WIKI case, all abstracts that contain a sequence of words that stems to the same roots as the label contained in the WIKI dictionary are tagged with this label. Similarly in the NP case, however, the dictionary generated from noun phrases is used instead of the lexicon created from titles of Wikipedia articles.

To illustrate the process of tagging, we presented a sample abstract from domains `math` and `stats` tagged with both WIKI and NP methods; see Fig. 2.

5 Comparison of the WIKI and NP tags across domains

As a first step in the comparison of the WIKI and NP methods, we calculated the average number of tags per document. This quantity was examined across different disciplines, and the results are presented in Fig. 3. The disciplines in Fig. 3 are



On the Computational Complexity of MCMC-based Estimators in Large Samples

In this paper we examine the implications of the statistical large sample theory for the computational complexity of Bayesian and quasi-Bayesian estimation carried out using Metropolis random walks. Our analysis is motivated by the Laplace-Bernstein-Von Mises central limit theorem, which states that in large samples the posterior or quasi-posterior approaches a normal density. Using the conditions required for the central limit theorem to hold, we establish polynomial bounds on the computational complexity of general Metropolis random walks methods in large samples. Our analysis covers cases where the underlying log-likelihood or extremum criterion function is possibly non-concave, discontinuous, and with increasing parameter dimension. However, the central limit theorem restricts the deviations from continuity and log-concavity of the log-likelihood or extremum criterion function in a very specific manner.

Under minimal assumptions required for the central limit theorem to hold under the increasing parameter dimension, we show that the Metropolis algorithm is theoretically efficient even for the canonical Gaussian walk which is studied in detail. Specifically, we show that the running time of the algorithm in large samples is bounded in probability by a polynomial in the parameter dimension d, and, in particular, is of stochastic order d^2 in the leading cases after the burn-in period. We then give applications to exponential families, curved exponential families, and Z-estimation of increasing dimension.

Tags from dictionary based on Wikipedia (WIKI)

approaching normal, bayesian estimate, central limit theorem, computational complexity, criterion function, exponential families, large sample, large sample theory, leading case, limit theorem, log concave, log likelihood, Metropolis algorithm, non concave, random walk, run time, sampling theory, stochastic order, von Mises

Tags from dictionary based on noun phrases found in the whole corpus (NP)

based estimates, bayesian estimates, central limit, central limit theorem, computation complexity, criterion function, exponential families, increasing dimension, large sample, large sample theory, limit theorem, log concave, log likelihood, metropolis algorithm, minimal assumption, normal densities, polynomial bounds, possible non, random walk, run time, sampling theory, specific manner, stochastic order, underlying log, von Mises

Fig. 2 The results of tagging for sample ArXiv abstract from domains `math` and `stat`. WIKI labels are marked with blue rectangles and NP tags are denoted with green background

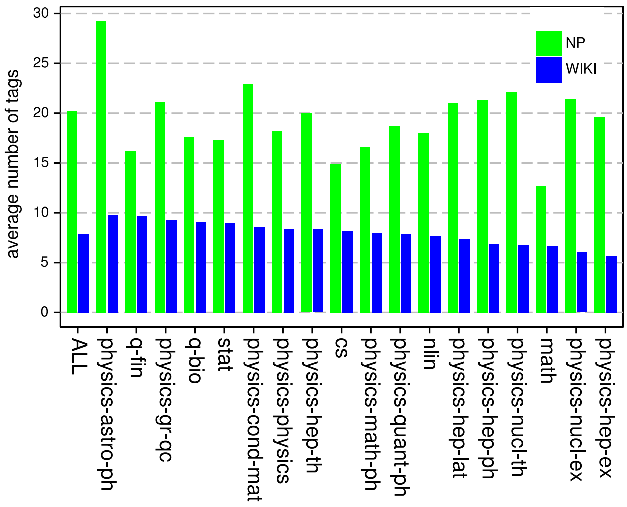


Fig. 3 Average number of tags per article for the WIKI and NP cases separated into ArXiv categories. Note that the categories are sorted according to the average number of WIKI tags in descending order

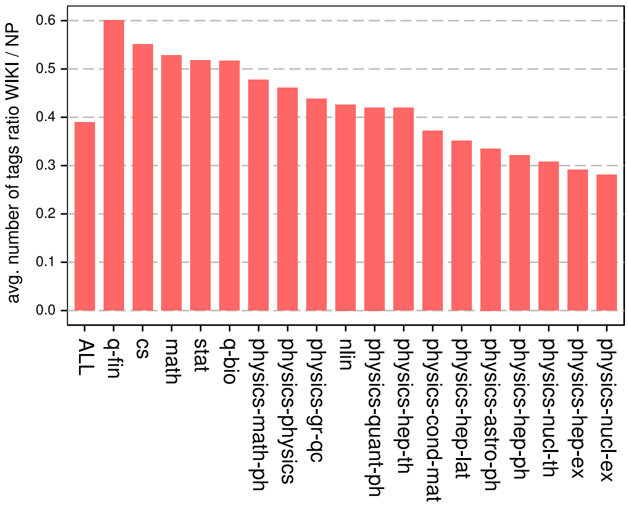


Fig. 4 The ratio of the average number of the WIKI tags to the number of the NP tags for different ArXiv categories. The categories are sorted in the order of descending ratio

sorted according to the average result for the WIKI method in ascending order. This allows us to observe that both methods are weakly correlated. In other words, if the WIKI method gives a high number of tags for a certain category, it does not imply that the NP approach yields a high average as well. This observation can be quantified by calculating the correlation coefficient between the average number of the WIKI and NP tags for each category, which indeed turns out to have a very low value of $\rho = 0.13$. Another conclusion from Fig. 3 is that clearly the NP method yields higher number of tags across all the domains. The average number of WIKI tags is roughly in the range from 0.3 to 0.6 of the NP result. The exact ratios for all the domains are visualized in Fig. 4. The bar chart is sorted according to the descending ratios. The sequence of disciplines can be, to a certain extent, intuitively understood. The leading categories, such as computer science and

Table 2 Comparison of the top ten most frequent tags in four categories

Top WIKI	Top NP	Top WIKI-only	Top NP-only
cs			
Lower bound	Lower bound	State of art	Large scale
Upper bound	Upper bound	Degrees of freedom	Interference channel
Polynomial time	Polynomial time	Point of view	Time algorithm
Et al	Et al	Object oriented	Proposed algorithm
Sensor network	Sensor network	Quality of service	Proposed method
Logic programming	Logic programming	Order of magnitude	Hoc network
Wireless network	Wireless network	Game theory	Considered problem
Real time	Real time	Reed Solomon	Wireless sensor
Network coding	Network coding	Multi agent system	Channel state
Ad hoc	Ad hoc	Multi user	Capacity region
math			
Lie algebra	Lie algebra	Calabi Yau	Give rise
Differential equation	Differential equation	Navier Stokes	Higher order
Moduli space	Moduli space	Point of view	Initial data
Lower bound	Lower bound	Non negative	Infinitely many
Field theory	Field theory	Cohen Macaulay	New proof
Finite dimensional	Finite dimensional	Algebraically closed	Over field
Sufficient condition	Sufficient condition	Degrees of freedom	Value problem
Upper bound	Upper bound	Self dual	Large class
Lie group	Lie group	Gromov Witten	Time dependence
Two dimensional	Two dimensional	Answered question	Mapping class
physics-nucl-ex			
Cross section	Cross section	Equation of state	Heavy ion
Au Au	Heavy ion	Center of mass	Au collisions
Heavy ion collision	Au Au	Order of magnitude	Ion collision
Form factor	Au collisions	Degrees of freedom	Au Au collision
Beta decay	Ion collision	Ultra relativistic	Transversal momentum
Elliptic flow	Au Au collision	Drell Yan	200 GeV
High energies	Heavy ion collision	Time of flight	Relativistic heavy
Experimental data	Transversal momentum	Presented first	Relativistic heavy ions
Charged particle	200 GeV	Long lived	Low energies
Nuclear matter	Form factor	National laboratory	Pb Pb

The first column (Top WIKI) denotes labels occurring in the WIKI method. The second column (Top NP) includes results produced by the NP method. The third column (Top WIKI-only) displays the most frequent tags generated by the WIKI method, but not by the NP. Finally, the fourth column shows the most frequent NP results not detected by the WIKI (Top NP-only)

quantitative finance, are probably more familiar to the average Internet user than experimental nuclear physics or high-energy physics. Thus, the coverage of the WIKI labels is also better in these domains. This indicates that various methods, relying on the knowledge from Wikipedia and verified on the computer science texts (such as, e.g., keyphrases in [13]) can have considerably lower performance when applied to documents from different scientific fields.

To further investigate the differences between the two methods, we displayed the most frequent tags generated by both methods in Table 2. In addition, we also included the most frequent tags generated uniquely by each method, to be able to better judge the differences. We have performed this analysis for three different ArXiv categories. We have selected `cs` and `math` as they have high ratio of the WIKI/NP average number of tags (we have neglected `q-fin` since there is very low number of documents from this field; see Fig. 1). We have also included `physics-nucl-ex`, as it is at the other end of the spectrum, having very low aforementioned ratio of the WIKI/NP average number of tags. There are a couple of interesting observations, which can be made from Table 2. Note that the Top WIKI and Top NP categories are identical for `cs` and `math` categories, whereas for `physics-nucl-ex` they are much different. In the latter case, the top four WIKI tags occur also in the NP results; however, the NP adds a lot of additional labels. They are mostly related to various kinds of nuclei collision processes, which apparently are too specific to be described in Wikipedia. Interestingly, the *Au–Au* tag from the WIKI corresponds to the article about one of the online auction portals and has nothing to do with gold nuclei. Another interesting property is that the WIKI method is much better at detecting surnames related to various theories, equations, etc. In particular, this is visible for `math` and the WIKI-only category, where four out of ten tags are related to surnames. This is a very desired feature and important strong point of the WIKI method.

Clearly, not all of the above tags are perfect. It can be observed that noun phrases detector sometimes yields the fragments of actual noun phrase, e.g., *hoc network* is a fragment of correct phrase *ad hoc network* and *time algorithm* comes from complexity statements, such as *polynomial time algorithm*. There are also a few tags which do not yield any information, e.g., *et al*, *point of view*, *give rise* and *initial data*. If there is a need, their impact can be reduced by improving the filtering procedure described in Sect. 4.

As a final stage of the analysis, we decided to address a question, to what extent the tags generated by the WIKI and NP methods are different? Table 2 suggests that in many categories top rank labels might be similar. Larger deviations may get introduced for the less frequent tags. To examine this phenomenon, we propose the following measures that describe the percentage of unique tags detected by each method up to rank r

$$\begin{aligned} C_{\mathrm{WIKI}}(r) &= \frac{\#(T_{\mathrm{WIKI}}(r) \setminus T_{\mathrm{NP}}(\infty))}{r}, \\ C_{\mathrm{NP}}(r) &= \frac{\#(T_{\mathrm{NP}}(r) \setminus T_{\mathrm{WIKI}}(\infty))}{r}, \end{aligned} \tag{1}$$

where $T_{\mathrm{WIKI}}(r)$ denotes the set of all tags up to rank r assigned by the WIKI method, and $T_{\mathrm{WIKI}}(\infty)$ refers to the set of all tags assigned by the WIKI method. The meaning of $T_{\mathrm{NP}}(r)$ and $T_{\mathrm{NP}}(\infty)$ is analogous, but refers to the NP approach. The $C_{\mathrm{WIKI}}(r)$ function describes the percentage of tags up to rank r, obtained from the WIKI method that were not detected by the NP approach (independently of rank). The $C_{\mathrm{NP}}(r)$ has analogous meaning for the NP case. The plots of the above quantities for a few sample ArXiv categories are presented in Fig. 5. We have selected the categories in a way that the edge cases of the fastest and the slowest growing dependencies are included. The figures clearly show that for the WIKI case the percentage of the unique tags is low, i.e., around 10 %, up to relatively high ranks, mostly $\sim 10^3 - 10^4$. This confirms the intuition that the relevant WIKI tags are indeed in majority noun phrases. On the other hand, the curves for the NP case show a different behavior; the percentage of the unique tags grows much faster in this case, indicating that they might yield much richer information. The 10 % level of unique tags is exceeded for the ranks lower than 10^2 for most categories. However, to give a definitive statement about the quality of the above tags, domain experts should be consulted.

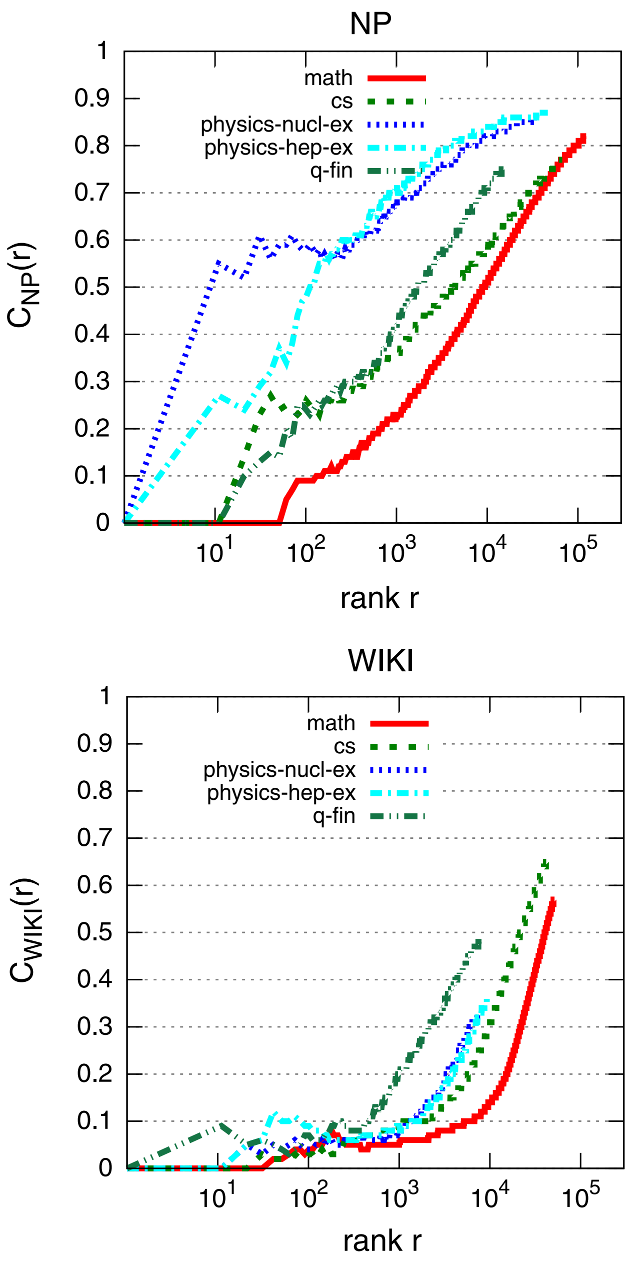


Fig. 5 The dependence of C_{NP} (*top panel*) and C_{WIKI} (*bottom panel*) on rank r, i.e., the percentage of tags up to rank r for the WIKI/NP method that were not detected by the other approach. Only a few sample categories were selected, including edge cases with the fastest and the slowest growing dependencies. See Eq. (1) and the main text for details

6 Statistical properties of the WIKI and NP tags

Tags can be expected to have similar statistical properties as ordinary words. One of the universal properties observed for words is the so-called Zipf's law, which states that the word frequency f as a function of its rank r in the frequency table should exhibit power-law behavior

$$f(r;\, A,\, N) = A\, r^{-N}, \tag{2}$$

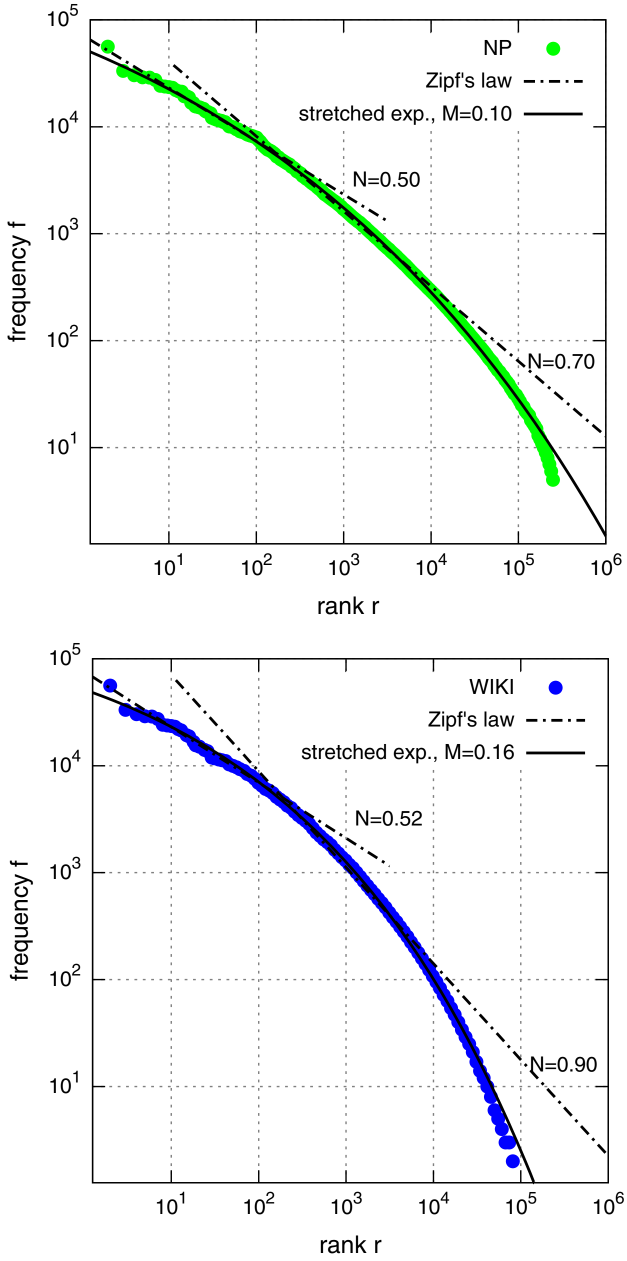


Fig. 6 Comparison of the frequency dependence on rank observed for tags obtained from both approaches—the NP (*top panel*) and the WIKI (*bottom panel*). The models fitted to the observed distributions are Zipf's law, see Eq. (2), and stretched exponential model, see Eq. (3)

where A and N are parameters. This type of simple dependency was observed not only for words, but also keyphrases, e.g., in the PNAS Journal bibliographic dataset [33]. However, the detailed investigation reveals that for large corpora, in particular when many different authors and hence different styles are involved, the simple model (2) might be insufficient to describe the frequency–rank dependence throughout the whole r variability range [20]. Sometimes, a few curves of the type (2) are necessary to accurately describe the observed distribution throughout the whole rank domain.

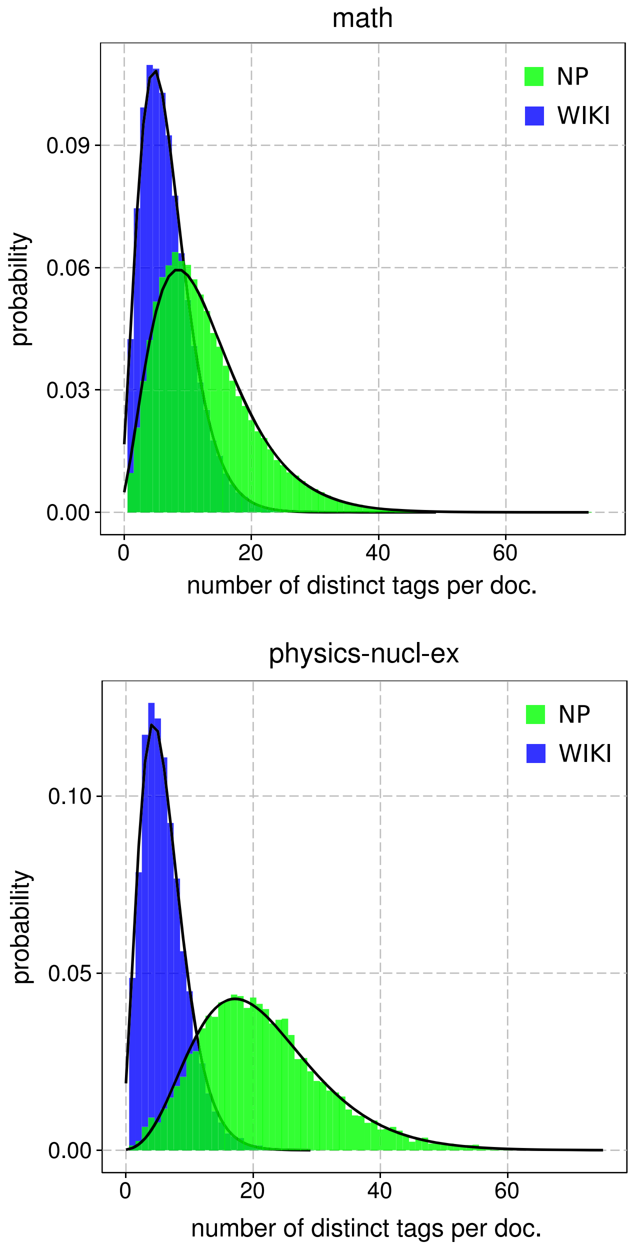


Fig. 7 Distribution for the number of tags per document within two sample ArXiv categories `math` (*top panel*) and `physics-nucl-ex` (*bottom panel*). The distribution can be well approximated by the negative binomial distribution; see Eq. (4). The *black line* represents the fits of this model to the observed data

In the case of our tags, the observed rank–frequency dependencies are presented in Fig. 6. In both cases (WIKI and NP), the crude approximation for the observed data was obtained using a combination of two Zipf type curves for different rank regimes. It turned out that up to rank 100, the values of exponent N were very similar in both cases and approximately equal to 0.5. However, for larger values the WIKI case showed more rapid decay with $N = 0.95$, as opposed to $N = 0.73$ in the NP case. Nevertheless, it is easily observed that a simple combination of the Zipf type curves does not fit the data very well. It turns out that the observed rank–frequency dependencies are much better approximated by one of the alternatives to the power law (2), namely the stretched exponential distribution. This type of distribution is used to describe a large variety of phenomena from physics to finance [15]. It was observed, e.g., for rank distributions of radio/light emission intensities from galaxies, French and US agglomeration sizes, daily Forex US-Mark price variation, etc. The stretched exponential model yields the following dependence of frequency on rank

$$f(r; C, D, M) = C \exp\left(-D\, r^{M}\right), \tag{3}$$

with C, D and M being parameters. As can be observed in Fig. 6, this model fits the data much better. Similarly to the Zipf's law, the value of the exponent for the NP case, which reads $M = 0.12$, is lower than for the WIKI, where $M = 0.19$. This indicates slower decay and "fatter tail" for the NP tags case.

Another interesting statistical property of the generated tags is the distribution of the number of distinct labels per document. It turns out that, even though the average tag counts per document are quite different for the WIKI and NP methods (see Sect. 5), the distributions in both cases come from the same family. Observed histograms can be approximated with the negative binomial distribution. According to this model, the probability of finding document with k tags reads

$$\text{Prob}(k; P, R) = \binom{k+R-1}{k} P^{R}(1-P)^{k}, \tag{4}$$

where $R > 0$ and $P \in (0, 1)$ are the parameters of the distribution. The comparison of the above model with the observed histograms can be found in Fig. 7.

7 Application of the NP and WIKI tags to topic modeling

7.1 Description of the employed methods

The tags developed and examined in the previous sections can be employed as features in various sorts of text mining algorithms. In this part, we examine their applicability to topic modeling.

Detecting topics in large collections of documents is of very high interest when building digital libraries or other document repositories. In particular, this technique can significantly improve search facilities or content discovery options [22]. The key method in the field is latent Dirichlet allocation (LDA) [6]. This is a very useful model as it does not require any *a priori* knowledge about the domain of a given document collection or manually labeled training set (unsupervised method). However, topics generated by this model are still often imperfect and not always meaningful for humans [8]. Therefore, to improve the situation, we examine running an LDA analysis not on words, but on tags, which already form comprehensible short phrases. To put it another way, we investigate the use of "set of tags" instead of "bag of words" document representation.

For the experiments in this part, the subset of 1,052 documents was selected. This set consisted of ArXiv abstracts with domain `cond-mat`, published in March 2012. The domain of condensed matter physics was selected, since it is a very strongly represented area in a given dataset (second most abundant domain, see Fig. 1). Therefore, the content is likely to be a reasonable sample of the whole body of ongoing research. Moreover, the domain expert in the field was available for this study.

To carry out the analysis, we employed the R statistical environment. It was supplemented with specialized packages for text analysis—`TM` [10] and `topicmodels` [11]. The latter contains procedures to carry out LDA analysis on a given corpus of documents.

In this part, we compare two approaches. The first method, referenced hereafter as ABS, runs the LDA procedure on abstracts of the articles, employing normal bag of words representation. The second examined technique makes use only of the list of tags extracted from the paper by the NP or WIKI methods (set of tags representation). The LDA procedure was run with topic number parameter $k = 7$. As a result, LDA yields a set of topics, which consist of words and their probabilities. The most probable words in each topic should form a reasonable description of the themes occurring in the examined document collection. The obtained topics are presented in the form of the so-called "word clouds"—Fig. 8 represents results for the ABS method, Fig. 9 shows the outcome of the approach with the NP tags and Fig. 10 provides analogous graph for the WIKI tags. The pictures display the most probable words or tags in each obtained topic, with font size being proportional to its probability.

7.2 Qualitative analysis of the obtained topics

The objective automatic analysis of the obtained topics is a difficult problem. Typically, they are ranked on the basis of probability of the held-out set of documents calculated from previously trained LDA model [6,30]. However, also other more sophisticated methods of ranking topic models were introduced [4,21]. Nevertheless, automatic methods not always correlate with human judgment [7,8]. In the work [22], the authors summarized the features describing a well-extracted topic as *sensible*, *meaningful*, *interpretable* and *coherent*. These traits are very difficult for automatic quantifi-

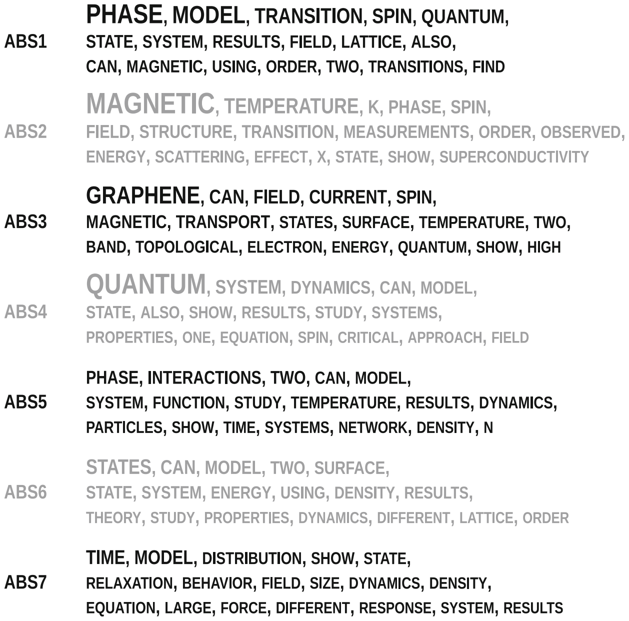

ABS1
PHASE, MODEL, TRANSITION, SPIN, QUANTUM,
STATE, SYSTEM, RESULTS, FIELD, LATTICE, ALSO,
CAN, MAGNETIC, USING, ORDER, TWO, TRANSITIONS, FIND
ABS2
MAGNETIC, TEMPERATURE, K, PHASE, SPIN,
FIELD, STRUCTURE, TRANSITION, MEASUREMENTS, ORDER, OBSERVED,
ENERGY, SCATTERING, EFFECT, X, STATE, SHOW, SUPERCONDUCTIVITY
ABS3
GRAPHENE, CAN, FIELD, CURRENT, SPIN,
MAGNETIC, TRANSPORT, STATES, SURFACE, TEMPERATURE, TWO,
BAND, TOPOLOGICAL, ELECTRON, ENERGY, QUANTUM, SHOW, HIGH
ABS4
QUANTUM, SYSTEM, DYNAMICS, CAN, MODEL,
STATE, ALSO, SHOW, RESULTS, STUDY, SYSTEMS,
PROPERTIES, ONE, EQUATION, SPIN, CRITICAL, APPROACH, FIELD
ABS5
PHASE, INTERACTIONS, TWO, CAN, MODEL,
SYSTEM, FUNCTION, STUDY, TEMPERATURE, RESULTS, DYNAMICS,
PARTICLES, SHOW, TIME, SYSTEMS, NETWORK, DENSITY, N
ABS6
STATES, CAN, MODEL, TWO, SURFACE,
STATE, SYSTEM, ENERGY, USING, DENSITY, RESULTS,
THEORY, STUDY, PROPERTIES, DYNAMICS, DIFFERENT, LATTICE, ORDER
ABS7
TIME, MODEL, DISTRIBUTION, SHOW, STATE,
RELAXATION, BEHAVIOR, FIELD, SIZE, DYNAMICS, DENSITY,
EQUATION, LARGE, FORCE, DIFFERENT, RESPONSE, SYSTEM, RESULTS

Fig. 8 "Word clouds" generated from the LDA topics obtained on the bag of words representation of abstracts (method abbreviated ABS in the main text)

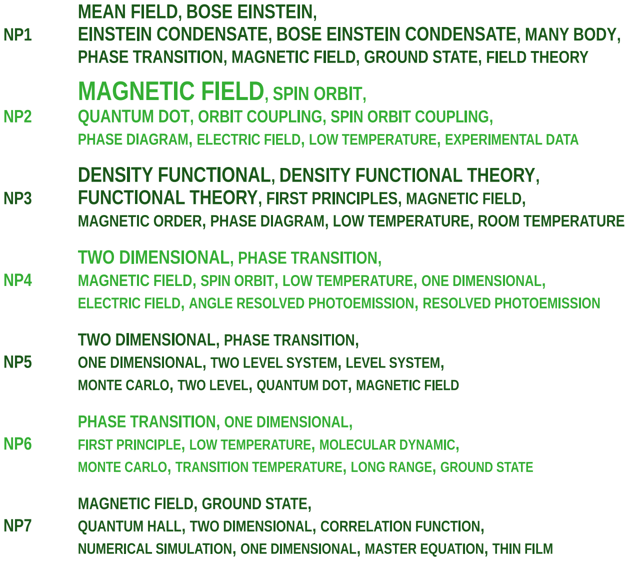

NP1
MEAN FIELD, BOSE EINSTEIN,
EINSTEIN CONDENSATE, BOSE EINSTEIN CONDENSATE, MANY BODY,
PHASE TRANSITION, MAGNETIC FIELD, GROUND STATE, FIELD THEORY
NP2
MAGNETIC FIELD, SPIN ORBIT,
QUANTUM DOT, ORBIT COUPLING, SPIN ORBIT COUPLING,
PHASE DIAGRAM, ELECTRIC FIELD, LOW TEMPERATURE, EXPERIMENTAL DATA
NP3
DENSITY FUNCTIONAL, DENSITY FUNCTIONAL THEORY,
FUNCTIONAL THEORY, FIRST PRINCIPLES, MAGNETIC FIELD,
MAGNETIC ORDER, PHASE DIAGRAM, LOW TEMPERATURE, ROOM TEMPERATURE
NP4
TWO DIMENSIONAL, PHASE TRANSITION,
MAGNETIC FIELD, SPIN ORBIT, LOW TEMPERATURE, ONE DIMENSIONAL,
ELECTRIC FIELD, ANGLE RESOLVED PHOTOEMISSION, RESOLVED PHOTOEMISSION
NP5
TWO DIMENSIONAL, PHASE TRANSITION,
ONE DIMENSIONAL, TWO LEVEL SYSTEM, LEVEL SYSTEM,
MONTE CARLO, TWO LEVEL, QUANTUM DOT, MAGNETIC FIELD
NP6
PHASE TRANSITION, ONE DIMENSIONAL,
FIRST PRINCIPLE, LOW TEMPERATURE, MOLECULAR DYNAMIC,
MONTE CARLO, TRANSITION TEMPERATURE, LONG RANGE, GROUND STATE
NP7
MAGNETIC FIELD, GROUND STATE,
QUANTUM HALL, TWO DIMENSIONAL, CORRELATION FUNCTION,
NUMERICAL SIMULATION, ONE DIMENSIONAL, MASTER EQUATION, THIN FILM

Fig. 9 "Word clouds" generated from the LDA topics obtained on documents represented as set of tags from the NP method

cation. Therefore, the only currently available method of reliable topics ranking is the manual judgement. Indeed, recent works dealing with this issue [7,8,22] employed a large number of human evaluators, gathered, e.g., using crowdsourcing services such as Amazon Mechanical Turk. Human judgement is particularly important when further applications in digital library systems are planned. Obviously, digital libraries are designed for people, and their user experience, not a mathematical criterion, has to be maximized.

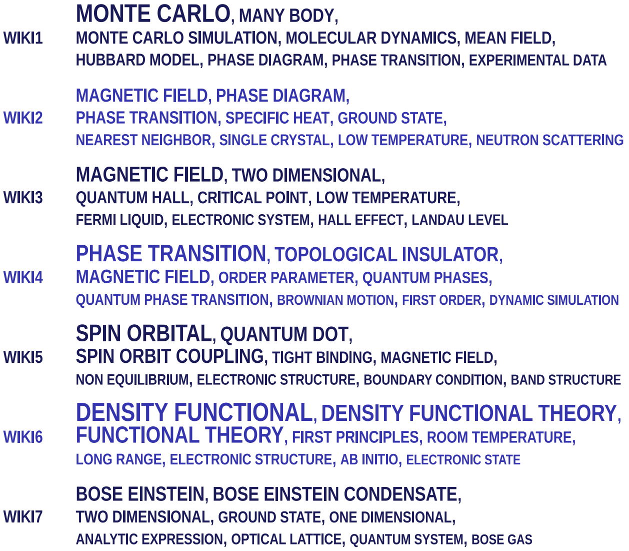

WIKI1
MONTE CARLO, MANY BODY,
MONTE CARLO SIMULATION, MOLECULAR DYNAMICS, MEAN FIELD,
HUBBARD MODEL, PHASE DIAGRAM, PHASE TRANSITION, EXPERIMENTAL DATA
WIKI2
MAGNETIC FIELD, PHASE DIAGRAM,
PHASE TRANSITION, SPECIFIC HEAT, GROUND STATE,
NEAREST NEIGHBOR, SINGLE CRYSTAL, LOW TEMPERATURE, NEUTRON SCATTERING
WIKI3
MAGNETIC FIELD, TWO DIMENSIONAL,
QUANTUM HALL, CRITICAL POINT, LOW TEMPERATURE,
FERMI LIQUID, ELECTRONIC SYSTEM, HALL EFFECT, LANDAU LEVEL
WIKI4
PHASE TRANSITION, TOPOLOGICAL INSULATOR,
MAGNETIC FIELD, ORDER PARAMETER, QUANTUM PHASES,
QUANTUM PHASE TRANSITION, BROWNIAN MOTION, FIRST ORDER, DYNAMIC SIMULATION
WIKI5
SPIN ORBITAL, QUANTUM DOT,
SPIN ORBIT COUPLING, TIGHT BINDING, MAGNETIC FIELD,
NON EQUILIBRIUM, ELECTRONIC STRUCTURE, BOUNDARY CONDITION, BAND STRUCTURE
WIKI6
DENSITY FUNCTIONAL, DENSITY FUNCTIONAL THEORY,
FUNCTIONAL THEORY, FIRST PRINCIPLES, ROOM TEMPERATURE,
LONG RANGE, ELECTRONIC STRUCTURE, AB INITIO, ELECTRONIC STATE
WIKI7
BOSE EINSTEIN, BOSE EINSTEIN CONDENSATE,
TWO DIMENSIONAL, GROUND STATE, ONE DIMENSIONAL,
ANALYTIC EXPRESSION, OPTICAL LATTICE, QUANTUM SYSTEM, BOSE GAS

Fig. 10 "Word clouds" generated from the LDA topics obtained on documents represented as set of tags from the WIKI method

Therefore, our results of the LDA analysis were inspected by an expert in the field of condensed matter. Most apparent issue, noticeable for human expert in the presented word clouds, is that the tag-based methods yield narrower and more easily interpretable topics. For example, both tag-based methods (NP and WIKI) recognized topics related to Bose–Einstein condensation—see NP1 and WIKI7 in Figs. 9 and 10, respectively. This is indeed a heavily investigated field, both experimentally and theoretically (Nobel Prize in Physics in 2001 was awarded for the work in this field). Similarly, both NP and WIKI methods recognized the Quantum Hall effect and related research (again, heavily investigated phenomenon—Nobel Prizes in Physics in 1998 and 1985 were awarded for the work related to this field); see NP7 and WIKI3. Furthermore, the NP and WIKI methods recognized the density functional theory as a separate topic (NP3, WIKI6), which is indeed the case, as it is currently the most powerful theoretical approach to predict properties of solids from first principles of quantum mechanics. The remaining topics easily identifiable for the human expert are spin–orbit effects, especially in low-dimensional structures (NP2, WIKI5) and many body simulation methods (WIKI1). Out of these topics, only one can be (with slight difficulty) identified by the expert in the abstract-based results. This is the ABS6 in Fig. 8, which is related to density functional theory, except that the ABS method yields only one more sharp topic that could be clearly resolved by human expert. It is the ABS3, which can be interpreted as graphene-related research. Admittedly, it is a hot topic in recent years—Nobel Prize 2010 in Physics was awarded for the pioneering work in this field. Other than that, the ABS topics described rather general and broad themes. The fact that so few of the results from ABS and NP/WIKI methods coincide seems to be related to the deficiencies of the LDA method. As it was pointed out, e.g., in [27], the distribution of words in documents can be much sharper than distribution in fairly general topics. This effect is captured by models, where active features get multiplied and renormalized, e.g., [16,27]. It is possible that such methods would locate sharp correlations between words *density, functional, theory* or *spin, orbit, coupling*.

The next important conclusion from the presented results is that the efficiency of topic detection in tag-based methods is heavily affected by the quality of the input tags. Both tag-based methods (NP and WIKI) missed the very important graphene topic ABS3. This was because of the tags' construction. They were by design restricted to capture multi-word features. Even though the detected tag set contained terms such as *bilayer graphene* or *graphene nanoribbon* they were too specific and rare to generate separate topic. Recent studies [9] show that human-generated keywords contain mostly multiple words; however, it turns out that important unrecoverable information may be contained in single word tags. Therefore, including unigram tags in the presented methods (without cluttering the tag dictionary with all nouns or all single word entries from Wikipedia) is a very interesting field of further research.

As far as the number of tags per document is concerned, it does not seem to influence the topics quality very strongly. The average number of WIKI tags in the field of `physics-cond-mat` was slightly lower than 40 % of the number of NP tags (compare Sect. 5, in particular Fig. 4). However, the quality of topics obtained from both NP and WIKI methods is comparable. Moreover, in many cases the correspondence between the topics produced by both approaches can be established, e.g., NP1–WIKI7, NP2–WIKI5, NP3–WIKI6 and NP7–WIKI3.

Another interesting aspect of the presented comparison is related to the topic repetition effect. This is one of the flaws encountered in the LDA results, when one theme resolved by human appears as multiple LDA topics [8]. Our experiments indicate that both tag-based approaches and plain abstract method seem equally prone to this difficulty. For example, ABS1 and ABS2 or NP4 and NP5 are related to very similar areas.

7.3 Analysis of computational cost

Clearly, the computational performance is definitely secondary to the topics quality. However, in particular for practical applications, it can also be a significant factor. Therefore, we decided to compare also the NP/WIKI methods and the ABS approach is terms of computational time. In the tag-based methods document representation is much more compact. Therefore, the analysis time reduces significantly. The time-consuming phases in the analysis are creation of the term-document matrix, which scales as $O(n\bar{m})$, and the LDA computation itself, which can be implemented as $O(kn\hat{m})$, where n is the number of documents, k is the number of topics, $\bar{m}$ is the average number of terms per document and $\hat{m}$ is the average number of distinct terms per document. The tag representation has considerably lower $\bar{m}$ and $\hat{m}$ factors, as the document has much fewer tags than plain words. Since these asymptotic considerations do not take into account prefactors, we also complement the analysis with real-life timings. They are presented in Table 3, which shows the average execution duration for the five runs of the described LDA computations on a modern laptop. The results indicate that the NP method is over two times faster than ABS, and WIKI is over four times faster than ABS on the examined dataset. In addition, the greater speed is also associated with the lower memory footprint. This issue may become important in analyzing large datasets, when analysis of the whole available text (not only tags) might easily get infeasible. Another situation when the time is crucial occurs when the LDA results are needed in real time, e.g., for rapid analysis of user search query results. In such cases, reducing a text to a small set of tags could be very beneficial.

Table 3 Execution time for the examined methods

	ABS	NP	WIKI
Timing (s)	39.8	19.4	8.6
Relative timing	1.0	0.46	0.21

The average from the five runs is given. The relative timings are with respect to the slowest method (ABS)

Overall, we find that the examined NP and WIKI tags are useful features for applications in LDA. They generate interpretable, narrow topics and reduce computing resources needed to obtain results. Their applications to other text mining problems, such as evaluating document similarity or clustering, seems promising.

8 Summary and conclusions

In this paper, we have compared two methods of tagging scientific publications. First, abbreviated WIKI was based on the multi-word entries from Wikipedia. Second, referenced as NP, relied on the noun phrases detected by the NLP tools. We have focused on the effectiveness of each method across domains and on the statistical properties of the obtained labels. Since the tags are meant for applications as features in text mining tasks, we have shown a sample application to topic modeling using the LDA approach.

When it comes to the effectiveness of the above tagging methods, it turned out that the NP approach yields higher average number of tags per document. The difference is by a factor between two and three with respect to the WIKI case. This strongly depends on the domain. The WIKI tags coverage is better in the areas more relevant to the Internet community, such as computer science or quantitative finance than in more exotic domains such as nuclear experimental physics. The ratio could be interpreted as a crude measure of Wikipedia knowledge completeness across domains. None of the methods is clearly superior than the other. Both have their strengths and weaknesses. The NP method is more prone to inaccuracies of underlying NLP tools, sometimes detecting incomplete phrases. It also produces more uninformative tags. The WIKI method yields generally fewer tags, but is much more effective in detecting tags including surnames, which are often important in names of theorems, equations or effects in science.

As far as the statistical properties are concerned, it turned out that both the WIKI and NP methods exhibit qualitatively very similar behavior. Observed dependencies of the tag frequency on the tag rank deviate from the Zipf's law. However, it can be well approximated with the so-called stretched exponential model. The investigation of the distribution of the number of distinct tags per document revealed that in both the WIKI and NP cases it follows quite closely the negative binomial model.

Sample application of the prepared tags to the topic modeling using the LDA method shows promising results. The representation of the document as a "set of tags" instead of "bag of words" yielded topics that were definite and easily interpretable by humans. It also reduced the computational time required for the analysis. Both NP and WIKI tags method performed here comparably well. Overall, in our opinion, the presented tags seem useful complement to the "bag of words" representation. We plan further refinement of both methods (in particular, extending them with unigrams) and further experiments with their applications in text mining tasks.

Acknowledgments This research was carried out with the support of the "HPC Infrastructure for Grand Challenges of Science and Engineering (POWIEW)" Project, co-financed by the European Regional Development Fund under the Innovative Economy Operational Programme.

References

1. Apache OpenNLP, http://opennlp.apache.org
2. arXiv preprint server, http://arxiv.org
3. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Data mining for improving textbooks. SIGKDD Explor. Newsl. **13**(2), 7 (2012). doi:10.1145/2207243.2207246
4. AlSumait, L., Barbar, D., Gentle, J., Domeniconi, C.: Topic significance ranking of lda generative models. In: Buntine, W., Grobelnik, M., Mladeni, D., Shawe-Taylor, J. (eds.) Machine learning and knowledge discovery in databases, lecture notes in computer science, p. 67. Springer, Berlin (2009). doi:10.1007/978-3-642-04180-8_22
5. Barker, K., Cornacchia, N.: Using noun phrase heads to extract document keyphrases. In: H. Hamilton (ed.) Advances in artificial intelligence, lecture notes in computer science, vol. 1822, p. 40. Springer, Berlin (2000). doi:10.1007/3-540-45486-1_4
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993 (2003)
7. Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Neural Information Processing Systems, vol. 22, p. 288 (2009)
8. Chuang, J., Gupta, S., Manning, C., Heer, J.: Topic model diagnostics: Assessing domain relevance via topical alignment. In: S. Dasgupta, D. Mcallester (eds.) Proceedings of the 30th International Conference on Machine Learning (ICML-13), vol. 28, p. 612. JMLR Workshop and Conference Proceedings (2013)
9. Chuang, J., Manning, C.D., Heer, J.: Without the clutter of unimportant words descriptive keyphrases for text visualization. ACM Trans. Comput. Hum. Interact. **19**(3), 19:1–19:29 (2012). doi:10.1145/2362364.2362367
10. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. J. Stat. Softw. **25**(5), 1 (2008)
11. Grün, B., Hornik, K.: Topicmodels: an R package for fitting topic models. J. Stat. Softw. **40**(13), 1 (2011)
12. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing, EMNLP '03, p. 216. Association for Computational Linguistics, Stroudsburg, PA, USA (2003). doi:10.3115/1119355.1119383
13. Joorabchi, A., Mahdi, A.E.: Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. J. Inf. Sci. **39**(3), 410 (2013). doi:10.1177/0165551512472138
14. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Nat. Lang. Eng. **1**(01), 9 (1995). doi:10.1017/S1351324900000048
15. Laherrre, J., Sornette, D.: Stretched exponential distributions in nature and economy: fat tails with characteristic scales. Eur. Phys. J. B **2**(4), 525 (1998). doi:10.1007/s100510050276
16. Larochelle, H., Lauly, S.: A neural autoregressive topic model. In: NIPS, p. 2717 (2012)
17. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia—a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web J. (2014, to appear)
18. Łopuszyński, M., Bolikowski, Ł.: Tagging scientific publications using wikipedia and natural language processing tools. In: Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi, J. Schirrwagen (eds.) Theory and Practice of Digital Libraries—TPDL 2013 Selected Workshops, Communications in Computer and Information Science, vol. 416, p. 16. Springer International Publishing (2014). doi:10.1007/978-3-319-08425-1_3.
19. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, I-Semantics '11, p. 1. ACM, New York, NY, USA (2011). doi:10.1145/2063518.2063519
20. Montemurro, M.A.: Beyond the Zipf-Mandelbrot law in quantitative linguistics. Phys. A Stat. Mech. Appl. **300**(34), 567 (2001). doi:10.1016/S0378-4371(01)00355
21. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, p. 100. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)

22. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL '10, p. 215. ACM, New York, NY, USA (2010). doi:10.1145/1816123.1816156
23. Nomoto, T.: WikiLabel: an encyclopedic approach to labeling documents en masse. In: Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, p. 2341. ACM, New York, NY, USA (2011). doi:10.1145/2063576.2063961
24. Nomoto, T., Kando, N.: Conceptualizing documents with Wikipedia. In: Proceedings of the fifth workshop on Exploiting semantic annotations in information retrieval, ESAIR '12, p. 11. ACM, New York, NY, USA (2012). doi:10.1145/2390148.2390155
25. Porter, M.: An algorithm for suffix stripping. Progr. Electron. Libr. Inf. Syst. **14**(3), 130 (1980)
26. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents, p. 1. John Wiley and Sons, Ltd (2010). doi:10.1002/9780470689646.ch1
27. Salakhutdinov, R., Hinton, G.E.: Replicated softmax: an undirected topic model. In: NIPS, vol. 22, p. 1607 (2009)
28. Spanakis, G., Siolas, G., Stafylopatis, A.: DoSO: a document self-organizer. J. Intell. Inf. Syst. **39**(3), 577 (2012)
29. Spanakis, G., Siolas, G., Stafylopatis, A.: Exploiting Wikipedia knowledge for conceptual hierarchical clustering of documents. Comput. J. **55**(3), 299 (2012). doi:10.1093/comjnl/bxr024
30. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, p. 1105. ACM, New York, NY, USA (2009). doi:10.1145/1553374.1553515
31. Wang, P., Hu, J., Zeng, H.J., Chen, Z.: Using Wikipedia knowledge to improve text classification. Knowl. Inf. Syst. **19**(3), 265 (2009). doi:10.1007/s10115-008-0152
32. Witten, I.H., Don, K.J., Dewsnip, M., Tablan, V.: Text mining in a digital library. Int. J. Digit. Libr. **4**(1), 56 (2004). doi:10.1007/s00799-003-0066
33. Zhang, Z.K., L, L., Liu, J.G., Zhou, T.: Empirical analysis on a keyword-based semantic system. Eur. Phys. J. B **66**(4), 557 (2008). doi:10.1140/epjb/e2008-00453