A framework for keyphrase extraction from scientific journals

Vidas Daudaravicius

VTeX, Vilnius, Lithuania,
vidas.daudaravicius@vtex.lt

Abstract. We present a framework for keyphrase extraction from scientific journals in diverse research fields. While journal articles are often provided with manually assigned keywords, it is not clear how to automatically extract keywords and measure their significance for a set of journal articles. We compare extracted keyphrases from journals in the fields of astrophysics, mathematics, physics, and computer science. We show that the presented statistics-based framework is able to demonstrate differences among journals, and that the extracted keyphrases can be used to represent journal or conference research topics, dynamics, and specificity.

Keywords: keyphrase extraction, journal, collocation, TF-IDF

1 Introduction

Keyphrase extraction from single documents has been extensively examined for many years. Automatic keyphrase extraction concerns "the automatic selection of important and topical phrases from the body of a document" [20]. *Keyphrase extraction* is the selection of a set of phrases that are related to the main topics discussed in a given document. Document keyphrases have shown their potential for improving many natural language processing and information retrieval tasks. [9] presents a thorough survey of the state of the art in automatic keyphrase extraction, examining the major sources of errors made by existing systems and discussing the challenges ahead.

Recently, several shared tasks have been organized to evaluate the performance of various keyphrase extraction tools, including the ACL 2015 Workshop on "Novel Computational Approaches to Keyphrase Extraction" [8] and SemEval-2010 Task 5: "Automatic Keyphrase Extraction from Scientific Articles" [11]. The small number of large, publicly available data-sets of scientific texts with annotated keyphrases is a major difficulty in the keyphrase extraction research domain. Most of the data-sets for the keyphrase extraction task are not large enough [9]. For instance, [13] shows that adding a little additional training data improved the final results of the task [11], i.e., +7.4% for the F-score, raising it from 25.6 to 27.5. For the scientific domain, the data-sets amount to only a few hundred documents. This number is not sufficient to apply the widely used TF-IDF measure, and it is difficult to avoid training overfitting. The arXiv.org

Table 1. Journals of the data-set.

Journal	Abbreviation	Documents	Tokens
Journal of Functional Analysis	JFAN	2490	26M
Journal of Algebra	JABR	4713	48M
Advances in Mathematics	AIMA	2041	28M
Solar Physics	SOLA	1683	11M
Journal of Optimization Theory and Applications	JOTA	1169	7M
Astrophysics and Space Science	ASTR	2880	12M
Annals of Operations Research	ANOR	1256	10M
Acta Applicandae Mathematicae	ACAP	799	5M
Total		**17031**	**148M**

database of scientific article preprints could be a very useful source of scientific articles to compile a data-set for the keyphrase extraction task. However, not many of the articles in the database have assigned keyphrases.

Table 2. Data distribution by year.

Year	Documents	Tokens
2005	330	3M
2006	701	6M
2007	1603	14M
2008	1593	13M
2009	1608	14M
2010	1747	15M
2011	1950	17M
2012	1913	16M
2013	1564	12M
2014	2335	19M
2015	1687	16M
Total	**17031**	**148M**

Keyphrase extractions makes widespread use of multiword extraction techniques. [10] presents recent advances in multiword extraction. There are two main approaches for multiword extraction: syntax-based [15, 10], and statistics-based [6, 10]. [6] uses collocation segmentation to extract keyphrases from the Association for Computational Linguistics Anthology Reference Corpus (ACL ARC) [2] to study ACL history and research dynamics over the past 50 years. The distribution of keyphrases in the ACL ARC can be used to understand the main breakpoints of research across many years.

Little attention has been given to the extraction of keyphrases from larger sets of journals or conference papers with the aim to study research dynamics. Major conference organizers will often publish manually selected keyphrases from all accepted papers in proceedings prefaces, as a way to show trends in research. The goal of our study is to present a framework for keyphrase extraction from scientific journals in diverse research domains. While journal articles are often provided with manually assigned keywords, it is not clear how to extract statistically or syntactically significant keywords and measure their importance to the entire journal. In our study, we show that our statistics-based framework is able to demonstrate differences among journals, and can be used to represent journal or conference research, topics, dynamics, and specificity.

2 The data-set

The data-set used in our study has access to proprietary data from the VTeX production archive which is not publicly available. VTeX provides pre-publishing (copy-editing and typesetting) services to major science publishers for many years. All papers are LaTeX coded, even if some of them were originally sub-

mitted to a journal with other coding (e.g., MS Word). There were two initial requirements for the selection of journals: at least nine years of continuous typesetting at VTeX, and domain variety. We selected eight journals (see Table 1). Papers published between 2005 and 2015 in the fields of astrophysics, mathematics, physics, and computer science. Table 2 shows the yearly data distribution, which is evenly distributed except for the first two years. The journals publish different numbers of papers each year. The largest one is JABR, and the smallest one is ACAP. The total number of tokens in the corpus is 148 million. Although this data-set is far from the amount of data in reality, the size of the corpus is suficient to apply statistics and show results.

3 Framework pipeline

The pipeline of the proposed framework comprises four main steps:
1) text extraction (Section 3.1) and language detection (Section 3.2),
2) candidate keyphrase list processing (Section 3.3),
3) single article keyphrase weighting (Section 3.4), and, finally,
4) smoothing of keyphrase weights for the sets of articles (Sections 3.5 and 3.6).

3.1 LaTeX-to-text conversion

We use the open-source tool `tex2txt`[1] for the conversion from LaTeX to text, because our source files are LaTeX-based. The tool is stand-alone and does not require any other LaTeX processing tools or packages. The primary goal of the tool is to extract the correct textual information from LaTeX files.

We need to point out that this tool makes some important changes to the original text: formal notation (i.e., mathematical expressions and other formulae) is substituted with the general category tag _MATH_. This substitution reduces the amount of interruptions of irrelevant language[2], and keeps the language of the article more coherent.

3.2 Language detection

Some papers in the selected journals were written in French with an abstract in English. We used a simple word-matching technique to detect articles written in English. We use two short word lists:

FR: de, et, le, une, sur, la, les, dans, est, pour;
EN: and, the, or, is.

An article is considered to be written in English if a text contains all of the words from the EN list, and none of words from the FR list. The technique is not universal, and useful only if both English and French languages are used. There are other statistical approaches to text language identification (see [1]).

[1] See demo on-line: http://textmining.lt:8080/tex2txt.htm
[2] In our case, this is mathematical language. Other cases may include a mix of English and French paragraphs in the same article.

3.3 Collocation chains

Collocation segmentation is introduced in [7]. Collocation segmentation is a type of segmentation whose goal is to detect *collocated word sequences* and to segment a text into word sequences that we call *collocation chains*. Collocation chains can have any non-predefined length (even a single word). This definition differs from other collocation definitions that commonly use n-gram list-based approaches [19, 3, 16]. Collocation segmentation is related to collocation extraction using syntactic rules [12]. Syntax-based approaches allow us to extract collocations that are easier to describe, and the process of collocation extraction is well controlled. In our work, we use language-independent collocation segmentation for the data-set preprocessing, and the keyphrase candidate list is generated in a similar way as in [6].

Word associativity. We use a Dice score to measure the associativity between two consecutive text tokens. Dice is defined as follows:

$$\mathrm{Dice}(x_{i-1}; x_i) = \frac{2 \cdot \mathrm{TF}(x_{i-1}; x_i)}{\mathrm{TF}(x_{i-1}) + \mathrm{TF}(x_i)},$$

where $\mathrm{TF}(x_{i-1}; x_i)$ is the number of co-occurrences of x_{i-1} and x_i, and $\mathrm{TF}(x_{i-1})$ and $\mathrm{TF}(x_i)$ are the numbers of occurrences of x_{i-1} and x_i in the training corpus. If x_{i-1} and x_i tend to occur in conjunction, their Dice score will be high. The Dice score is sensitive to low-frequency word pairs (see the comparison of various associativity measures in [4]). If two consecutive words are used only once and appear together, there is a high chance that these two words are closely related and form some new concept, e.g., a proper name or a semantically closed term. A sequence of tokens is turned into a curve of Dice values between two adjacent tokens. This curve of associativity values is used to detect the boundaries of collocation chains.

Second-order derivative for collocation chain boundaries. [5] introduces the average minimum law (AML) for setting collocation chain boundaries. Actually, AML is a second- order derivative, which is applied to three adjacent associativity values, and it is defined as follows:

$$boundary(x_{i-1}, x_i) = \begin{cases} \textit{True} \mid \mathrm{Dice}(x_{i-2}; x_{i-1}) + \mathrm{Dice}(x_i; x_{i+1}) - \\ \qquad\qquad -2 \cdot \mathrm{Dice}(x_{i-1}; x_i) > 0 \\ \textit{False} \mid \text{Otherwise.} \end{cases}$$

If the second-order derivative value is positive, then two consecutive tokens are not joined into a collocation chain, and while two consecutive tokens are concatenated if the derivative value is negative.

Preprocessed collocation chains. The data-set contains 129,257 unique unigrams and 3,770,944 unique bigrams. We processed the data-set with collocation

Table 3. The list of collocation chains that end with *energy*.

accumulated ...	essentially different ...	laser ...	radiation ...
acoustic ...	excess ...	local ...	radiative ...
activation ...	excitation ...	low ...	relativistic ...
additional ...	exotic ...	lower ...	relativistic fermi ...
adiabatic ...	explosion ...	lowest ...	released ...
adm ...	extracted ...	magnetic ...	repulsive ...
alpha particle ...	false vacuum ...	mass and ...	rest ...
average ...	fermi ...	mass ...	rotational ...
beam ...	field ...	mass or ...	screening ...
binding ...	final ...	matter and ...	second ...
break ...	flare ...	matter ...	shear ...
burst ...	flowing ...	matter or ...	shock ...
calculation of ...	fluid ...	maximum ...	significant ...
cm ...	fraction of ...	mean ...	soliton ...
comparable ...	free ...	mechanical ...	solution and ...
compton ...	gas ...	minimum ...	source of ...
constant ...	graph of ...	missing ...	specific ...
correlation ...	gravitational binding ...	moller ...	spectral ...
cosmological nuclear ...	gravitational correlation ...	negative ...	standard ...
coulomb ...	gravitational ...	newtonian potential ...	state ...
cutoff ...	gravitational wave ...	nuclear ...	stored ...
cyclotron ...	gravitomagnetic ...	null ...	stress ...
dark ...	helmholtz ...	orbital ...	strong ...
decaying vacuum ...	high ...	outburst ...	sufficiently high ...
dimensionless ...	highest ...	particle ...	symmetry ...
dissipated ...	holographic dark ...	peak ...	tev ...
dominant ...	hydrodynamic explosion ...	phantom ...	thermal ...
dust matter ...	increase of ...	phantom field ...	trace of ...
edge of ...	increasing ...	photon ...	turbulence ...
effective ...	infinite ...	plasma ...	turbulent ...
effective plasma ...	initial ...	plasmon ...	unit of ...
effective potential ...	instantaneous orbital ...	position and ...	universe ...
elastic ...	interaction ...	positive ...	vacuum ...
electric ...	interaction potential ...	positron fermi ...	very high ...
electric potential ...	internal ...	possible relativistic ...	vibration ...
electromagnetic ...	intersystem correlation ...	potential ...	wave ...
electron ...	isotropic ...	propagating wave ...	weak ...
electron thermal ...	kinetic ...	pseudo...	wind ...
electrons with ...		quantum vacuum ...	zero ...
electrostatic ...		radiant ...	
...		radiated ...	
equipartition law of ...		radiates ...	

segmentation, and found 1,364,638 unique collocation chains. The list size of bigrams is twice the size that of collocation chains. List size grows quickly with the length of n-grams. Although collocation chains reduce the number of unique items, nevertheless n-gram features are preserved and many noisy bigrams and trigrams that occur only once are often omitted. The maximal length of collocation chains is 6 tokens and, which is similar as into [6], where the maximal chain length was 7 tokens. The average collocation chain length in the dictionary is 1.68 tokens. Collocation chains with 1, 2, or 3 tokens cover 99.5 percent of the corpus. In Table 3 we show collocation chains ending with the token *energy*. Chains starting or ending with verbs, determiners, numbers, or prepositions were removed from this list. The list exposes many different energy types and con-

ceptions of energy, which givinge us a good sense of the variety and use of the term.

3.4 Keyphrase weighting with TF-IDF and NTF-PIDF

Since 1972, when the inverse document frequency measure was introduced [17], the TF-IDF weighting method has been very successfully used in information retrieval and other natural language processing tasks. We use TF-IDF, which is defined as follows:

$$\text{TF-IDF}(x) = \text{TF}(x) \cdot \ln\left(\frac{N}{\text{D}(x)}\right),$$

where $\text{TF}(x)$ is the raw frequency of a term x in the data-set, N is the total number of articles in the data-set, and $\text{D}(x)$ is the number of articles where the term x occurs.

We also use the normalized probabilistic TF-IDF variation. We make normalize term frequency normalization against the article length and the average article length. The length of articles in the data-set varies from 246 to 88642 tokens. The weight of a term TF-IDF in a shorter article is much lower than the its weight of ain a longer article, even if the two articles are about the same topic. The term-occurrence counts in articles are proportional to the article length, therefore, we normalize frequencies to make them comparable. Such normalization is important when we compare keyphrases in separate articles, but though not within each article itselfindividually. The term-frequency normalization is as follows:

$$\text{NTF}(x) = \text{TF}(x) \cdot \frac{\text{avgDocLen}}{\text{length}(D_x)},$$

where avgDocLen is the average article length in the data-set, and $\text{length}(D_x)$ is the length of an article with term x. The average article length in the data-set is 9090 unigram tokens and 5655 collocation chain tokens. In general, the average article length is constant and is not necessary in calculation, i.e., the constant becomes equal to 1.

IDF probabilistic variation is discussed in [14]. The probabilistic variation of IDF we use is defined as follows:

$$\text{PIDF}(x) = \ln\left(\frac{N - \text{D}(x) + 1}{\text{D}(x) + 1}\right).$$

In the case of a term which occurs in more than half of the articles in the data-set, the formula defines a negative weight. This is a somewhat odd prediction for a term. In practice, all terms with negative weight are stop-words or function words[3]. The normalized probabilistic TF-IDF is defined as follows:

$$\text{NTF-PIDF} = \text{NTF} \cdot \text{PIDF}.$$

[3] Function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence (https://en.wikipedia.org/wiki/Function_word). For instance, *and*, *or*, *the*, and *a* are all function words.

TF-IDF and NTF-PIDF are applied to each term of each article in the data-set. We take the top 100 most significant terms of each article for the next steps described in the following sections.

3.5 Journal keyphrases

We have described keyphrase extraction from articles in the sections above. In this section we extend keyphrase extraction from a single article to subgroups of larger sets of articles. Articles can be grouped by year and/or by journal, and/or by other categories. One straightforward way to extract keyphrases from groups of large data-sets is to calculate the term weights for each group separately. While the term frequency of a group is easy to calculate, it is not clear what should be the number of articles for each term occurrence. Typically, a journal issue contains from 10 to 50 articles. Such a small subgroup is not sufficient for TF-IDF weighting, and we will not be able to extract keyphrases properly for each separate journal, or yearly journal article groups. To tackle the problem of document count, we calculate the average of TF-IDF of article terms in each subgroup as follows:

$$\text{TF-IDF}_{\text{AVG}}(x|g) = \frac{\sum \text{TF-IDF}(x,g)}{\text{D}(x,g)},$$

where $\sum \text{TF-IDF}(x,g)$ is the sum of TF-IDF(x) terms of articles in a subgroup g, and $\text{D}(x,g)$ is the number of articles in the subgroup g where the term x occurs. The average of NTF-PIDF is calculated similarly:

$$\text{NTF-PIDF}_{\text{AVG}}(x|g) = \frac{\sum \text{NTF-PIDF}(x,g)}{\text{D}(x,g)}.$$

An example of the top 50 extracted keyphrases from ASTR is shown in Columns 3 and 4 of Table 4. The most significant keyphrase is *gsxr flares*, using both weighting measures. GSXR is an acronym of the term *great soft x-ray*. The extracted keyphrase lists with the two weighting measures do not significantly correlate. However, the keyphrase *gsxr flares* is used frequently in only one article. An expanded *gsxr flares* keyphrase is used only once, and it occurs in the title of that same article. What are the true keyphrases for the ASTR journal? Can we accept the *gsxr flares* keyphrase, which is used frequently in only one article, as a descriptor of the entire ASTR journal for the 2005–2015 period? In the following section we present a solution to this problem using *additive smoothing*.

We also noticed that keyphrases extracted with TF-IDF $_{\text{AVG}}$ weighting are much longer than the keyphrases extracted with NTF-PIDF $_{\text{AVG}}$. This is due to a high correlation between term frequency and term length (see [18]). NTF-PIDF $_{\text{AVG}}$ uses term frequency normalization to reduce the impact of term frequency to the term significance value. Short and frequent terms are more abstract than longer and less frequent ones. Therefore, this property can be used to extract either more abstract or more detailed terms, which might depend on the task.

Table 4. Top keyphrase lists of ASTR. Grey highlights keyphrases that occur in both top lists of two different measures.

Sigmoid additive smoothing		**No smoothing**	
NTF-PIDF $_{\text{ADD}}$	TF-IDF $_{\text{ADD}}$	TF-IDF $_{\text{AVG}}$	NTF-PIDF $_{\text{AVG}}$
periodic orbits	periodic orbits	gsxr flares	gsxr flares
black hole	equilibrium points	cemp	proxima
equilibrium points	black hole	outgrowths	bal quasars
brane	sn ia	bexb	emp
bulk viscosity	brane	fz ori	pbps
chaplygin gas	primaries	spin spacecraft	center manifold
bulk viscous	triangular points	pbps	cemp
dusty plasma	jet	proxima	smf
dark energy	cluster	finger	comet holmes
primaries	body problem	nutation damper	flyer
jet	co	lp uma	tv columbae
triangular points	disk	gclf	string ball
positrons	bulk viscosity	ba stars	spin spacecraft
shock waves	dark energy	magnetosonic critical curve	dsphs
scalar field	dusty plasma	issv solutions	ba stars
co	scalar field	stereo pairs	strong spes
cluster	oblateness	shaped fingers	nutation damper
solitary waves	planet	bal quasars	u geminorum
bianchi type	shock waves	center manifold	information bits
body problem	disc	bf	rgda
apparent horizon	star formation	cehe	wz sge
alma	bl lacs	pywifes	soft excess
oblateness	globular clusters	rgda	triple asteroids
gamma	asteroid	cyclical universe	ao psc
de sitter	positrons	smf	diamond detector
entropy	chaplygin gas	fingers	fec
growth rate	de sitter	periodic modes	electron hole
universe	growth rate	fyris alpha	submm ebl
neutron star	equilibrium point	attitude stability	gm cep
event horizon	solitary waves	expansive homogeneous	bf
star formation	moon	matter objects	qq vul
restricted three	neutron star	u geminorum	lmt
black holes	bulk viscous	isotropic relativistic universe	rv psc
holographic dark energy	shell	preferred alignment	primary cmes
mach number	positron	outgrowth	ulp cepheids
negative ions	light curves	protons flares	w ser
well behaved	mass loss	dibs	narrow cmes
positron	clusters	information bits	umfs
ion acoustic	galaxies	barium stars	ux mon
double layers	holographic dark energy	sxr emissions	bellert
hot electrons	magnetic field	edod	sterile neutrinos
disk	eos parameter	koi	capella
clusters	restricted three	dsphs	gclf
wso	apparent horizon	lunar wake	nqda
hawking radiation	double layers	w ser	giant pulses
dust grains	blazars	smgs	aloh
radiation pressure	universe	issv	nel rates
thermodynamics	mach number	neutral formaldehyde	tev j
blazars	metallicity	ux mon	ircss
disc	h	polarization force	bok globule

3.6 Additive smoothing of TF-IDF

Smoothing is widely used to reduce noise and irregularities in data series, or to smooth categorical data. Additive smoothing is a common approach in statistical language modeling. The aim of applying smoothing is to reduce probabilistic irregularities. In Section 3.5 we showed that the average of TF-IDF is not sufficient for the extraction of keyphrases for larger sets of articles. If a term is used

in a few articles only but its significance is high, then its average significance will also be high. Our goal is to extract keyphrases that represent entire set of articles. Therefore, we need some balance between article term significance and the number of articles in which this term occurs. The larger the number of articles with a term x and the higher the significance of the term x in these articles, the more this term is significant for the whole subset of articles. We implement this approach by applying additive smoothing to the TF-IDF $_{\mathrm{AVG}}$ calculation. This approach reduces the average of TF-IDF $_{\mathrm{AVG}}$ regarding the number of articles in which a keyphrase is used. Thus, the fewer the number of articles, the greater the amount of additive smoothing. After some manual experimentation, we adjusted the following s*igmoid* function:

$$\mathrm{sigmoid}(i) = \frac{200}{1 + e^{-(-0.05)\cdot i}},$$

where i is the number of articles in which a term x occurs.

This gives us a high smoothing value to terms that occur in a few articles only, and a lower smoothing value to terms that occur in many articles. For instance, if a term occurs in only one article then the smoothing parameter is equal to 97; for 50 or 100 articles, respectively, the parameter drops to 15 and 1.3. We use the following additive smoothing formulas:

$$\text{TF-IDF}_{\mathrm{ADD}}(x) = \frac{\mathrm{sigmoid}(\mathrm{D}(x)) + \sum \text{TF-IDF}(x)}{\mathrm{sigmoid}(\mathrm{D}(x)) + \mathrm{D}(x)},$$

$$\text{NTF-PIDF}_{\mathrm{ADD}}(x) = \frac{\mathrm{sigmoid}(\mathrm{D}(x)) + \sum \text{NTF-PIDF}(x)}{\mathrm{sigmoid}(\mathrm{D}(x)) + \mathrm{D}(x)}.$$

The results of additive smoothing (see Table 4) show high correlation for the top 50 extracted keyphrases using both weighting measures. For instance, the keyphrase *periodic orbits* occurs in 94 articles of ASTR, and its frequency is 1397; *black hole*: 562 and 6184; *dark energy*: 544 and 5073; and *equilibrium points*: 134 and 1684. We do not count instances when a keyphrase is nested in another term; for example, if nesting is considered, then the *black hole* term counts are 637 and 8063. In Table 4 we see *co* as a keyphrase, which is ambiguous in ASTR, as it can stand for either *cobalt* or *carbon monoxide*. Both meanings are extensively used in articles. Two-word keyphrases are the most common in journals, except SOLA and ANOR (see Table 5). The top keyphrase list of SOLA is full of abbreviations and acronyms. The top list of ANOR contains many single-word keyphrases. Keyphrase lists extracted with (TF-IDF $_{\mathrm{ADD}}$ and NTF-PIDF $_{\mathrm{ADD}}$) correlate. The question of how this correlation can help get even higher keyphrase extraction accuracy could be answered in a following study.

At this point, we have extracted keyphrases from a set of articles. In the following sections, these keyphrases will be used to analyze the main research trends in particular sets of articles.

Table 5. The top 50 list of extracted keyphrases of different journals. *Keyphrases with strikeouts cannot really be used as keyphrases. We show all automatically extracted keyphrases in order of significance. We omit significance values here and later to save space.*

AIMA	convex body, dg, convex bodies, hopf algebra, operad, polytope, weak equivalence, cell, cells, stack, groupoid, intersection body, ample, hausdorff dimension, tree, weak equivalences, spectral sequence, functor, cocycle, homotopy, moduli space, intersection bodies, quiver, line bundle, category, geodesic, scalar curvature, vector bundle, symmetric monoidal, sheaf, simplex, initial data, block, von neumann, edges, graph, triangulated category, monad, orbifold, simplices, bundle, volume, graded, hypersurface, gorenstein, simplicial, finite type, morphism, simplicial complex, semistable
JABR	vertex operator, lie superalgebra, vertex algebra, superalgebra, hopf algebra, crystal, defect group, fusion system, grobner basis, koszul, jordan algebra, definable, numerical semigroup, permutable, hilbert function, locally nilpotent, braided, representation type, bialgebra, lie superalgebras, soluble, engel, subnormal, leavitt path, almost split, del pezzo, tableau, monomial ideal, hopf algebras, dimension vector, central simple, gorenstein, betti numbers, cofinite, complete intersection, inner ideal, ample, quiver, clean, comodule, pi, local cohomology, polycyclic, artin algebra, semiprime, lie algebra, supersolvable, block, line bundle, supersoluble
ACAP	random variables, differential equations, initial data, boundary conditions, vector field, weak solution, vector fields, periodic solutions, system, hamiltonian, boundary value, positive solution, positive solutions, differential equation, h, nonlinear, operator, estimator, solutions, symmetries, stokes equations, pdes, ~~solution of~~, graph, distribution, functional, initial conditions, hopf bifurcation, conservation laws, species, fixed point, periodic solution, hpm, curve, estimators, fluid, asymptotically stable, global existence, symmetry, lie algebra, positive constant, differential invariants, equilibrium point, ham, prolongation, reaction, operators, banach space, zeros, population
ANOR	job, supply chain, jobs, machine, queue, game, retailer, customers, customer, dmu, machines, supplier, coalition, players, local search, player, server, policy, operations research, schedule, dea, portfolio, processing times, manufacturer, items, servers, risk, inventory level, column generation, criteria, markov chain, facility, agent, makespan, inventory, processing time, service time, master problem, benders decomposition, stage, node, patients, dmus, period, resource, special volume, completion time, buffer, graph, network
ASTR	periodic orbits, black hole, equilibrium points, brane, bulk viscosity, chaplygin gas, bulk viscous, dusty plasma, dark energy, primaries, jet, triangular points, positrons, shock waves, scalar field, co, cluster, solitary waves, bianchi type, body problem, apparent horizon, alma, oblateness, gamma, de sitter, entropy, growth rate, universe, neutron star, event horizon, star formation, restricted three, black holes, holographic dark energy, mach number, negative ions, well behaved, positron, ion acoustic, double layers, hot electrons, disk, clusters, wso, hawking radiation, dust grains, radiation pressure, thermodynamics, blazars, disc
JOTA	~~ref~~, optimal control, game, nash equilibrium, player, value function, maximal monotone, quasiconvex, valued mapping, constraint qualification, delay, primal, variational inequality, pseudomonotone, p, lsc, optimality conditions, strongly monotone, efficient solution, optimal solution, central path, algorithm, stationary point, global minimizer, variational inequalities, players, stochastic, multifunction, firm, augmented lagrangian, inequality constraints, lower semicontinuous, cost function, control, usc, upper semicontinuous, locally lipschitz, subdifferential, differential game, lmi, euclidean jordan, convex subset, duality gap, convex, normal cone, lmis, banach space, necessary optimality conditions, newton method, proposed method
SOLA	hi, filament, mc, flux rope, mcs, ca ii, solar wind, cme, type ii, icmes, icme, sep events, meridional flow, type iii, ar, prominence, flux tube, rotation rate, sunspot groups, gcr intensity, sunspot number, sunspot, type iii bursts, filaments, cmes, burst, hard x, coronal holes, he ii, hcs, hmi, ips, flare, radio bursts, fe i, coronal hole, loop, sunspot numbers, time series, eis, flux emergence, shock, cor, ars, solar activity, eruption, current sheet, tsi, solar cycle, sunspots
JFAN	toeplitz operators, completely positive, von neumann, dirichlet form, heat kernel, composition operator, composition operators, ground state, semigroup, completely bounded, poincare inequality, banach space, brownian motion, banach algebra, h, amenable, ricci curvature, invariant subspace, approximation property, almost surely, weak solution, initial data, locally convex, nuclear, critical point, posedness, unital, fixed point, lipschitz domain, cocycle, inner function, carleson measure, locally compact, metric space, toeplitz operator, weakly compact, graph, sobolev inequality, hilbert, representing measure, positive solution, invariant subspaces, positive definite, crossed product, operator space, fredholm, banach, strongly continuous, stokes equations, ergodic

4 Journal keyphrases

In this and the following sections we discuss possible use cases for extracted keyphrases. In this section we analyze the differences among journals in our data-set.

We can find the publishing topics of journals on the Internet. Often these descriptions overlap considerably. We found the following research subjects of journals:

- *Journal of Functional Analysis* (JFAN) related subjects[4]: Significant applications of functional analysis, including those to other areas of mathematics; New developments in functional analysis; Contributions to important problems in and challenges to functional analysis.
- *Journal of Algebra* (JABR) related subjects[5]: Results obtained by computer calculations; Classifications of specific algebraic structures; Description and outcome of experiments; Papers emphasizing the constructive aspects of algebra, such as description and analysis of new algorithms; Interactions between algebra and computer science, such as automatic structures, word problems, and other decision problems in groups and semigroups; Contributions are welcome from all areas of algebra, including algebraic geometry or algebraic number theory.
- *Advances in Mathematics* (AIMA) related subjects[6]: Emphasizing contributions that represent significant advances in all areas of pure mathematics.
- *Acta Applicandae Mathematicae* (ACAP) related subjects[7]: Classical Continuum Physics, Complexity, Computer Science, Mathematics, Theoretical, Mathematical, and Computational Physics.

The differences among the listings of related subjects are fuzzy. The topics covered by JFAN, JABR, AIMA, and ACAP overlap. Whether or not a submitted article is accepted depends on how successfully it conforms to the journal research subjects. There is a high chance that a submitted article will be rejected if the article is not in line with the journal research subjects. Journal keyphrases from at least the past five years would definitely help authors to choose the most appropriate journal for submission. Table 5 shows the top keyphrases for each journal in the data-set. The lists of keyphrases show that, in fact, all four journals focus on different topics, and there is no intersection among the main terms of each journal, making it easier to get a sense of the specificities of these journals.

5 Research dynamics

In this section, we analyze the *Astrophysics and Space Science* (ASTR) journal. ASTR is an international journal of astronomy, astrophysics, and space science. ASTR related subjects are as follows: Astrobiology; Astronomy, Observations, and Techniques; Astrophysics and Astroparticles; Cosmology; Extraterrestrial Physics, Space Sciences. The description of the journal is as follows [8]:

[4] http://www.journals.elsevier.com/journal-of-functional-analysis/
[5] http://www.journals.elsevier.com/journal-of-algebra/
[6] http://www.journals.elsevier.com/advances-in-mathematics/
[7] http://www.springer.com/mathematics/journal/10440
[8] http://www.springer.com/astronomy/astrophysics+and+astroparticles/journal/10509

Table 6. The top 50 extracted keyphrases from ASTR each year, 2007–2015.

2007 jet, gamma, neutron star, magnetic field, pulsar, ray emission, rx j, neutron stars, rays, crust, black hole, star, field equations, sources, pulsars, source, tev, glast, ray sources, psr b, disk, blazars, ray, energy density, high energy, axps, radio pulsars, emission, universe, hard x, radio emission, radio, kev, accretion rate, magnetic fields, g, flux, vhe, polar cap, accretion disk, energy, psr j, cosmic rays, egret, wind, grb, agn, angular momentum, light curve, scalar field

2008 alma, universe, co, field equations, star formation, bianchi type, mass, energy density, molecular gas, einstein, galaxies, cosmological models, scalar field, hcn, perfect fluid, h, magnetic field, gas, equilibrium points, star, metric, code, cosmological model, general relativity, dust, sma, molecules, ~~km s~~, stars, molecular clouds, disks, angular momentum, primaries, chemistry, emission, galaxy, stellar models, black hole, periodic orbits, gravitation, radiation pressure, cn, dark energy, ch, aca, disk, deceleration parameter, body problem, light curve, cm

2009 cluster, clusters, stars, universe, star clusters, star formation, galaxies, field equations, energy density, magnetic field, galaxy, o vi, dark energy, mass, uv, massive stars, ~~km s~~, cosmological constant, gas, star, mass segregation, cm, black hole, age, plasma, m, hii regions, h, metallicity, ly, cluster mass, experiment, ages, stellar, jet, mag, chaplygin gas, bianchi type, experiments, ns, shock, orbital period, disk, virial equilibrium, hst, flyer, antennae, ism, perfect fluid, uvot

2010 black hole, universe, stars, star, hot subdwarfs, sdb stars, dark energy, magnetic field, field equations, energy density, scalar field, cosmological constant, hot subdwarf, helium, mass, quasinormal frequencies, metallicity, solar, orbital period, black holes, solitary waves, einstein, modes, white dwarf, mass transfer, age, sun, main sequence, sdbs, general relativity, convective core, galaxies, binaries, galaxy, scale factor, kinetic energy, sdo stars, convection, mass loss, dh cmes, sdb, angular momentum, accelerated expansion, binary, globular clusters, bulk viscosity, cosmological models, event horizon, white dwarfs, models

2011 universe, black hole, dark energy, magnetic field, energy density, field equations, neutron star, star, event horizon, chaplygin gas, solitary waves, scalar field, periodic orbits, stars, cosmological constant, pressure, perfect fluid, galaxies, plasma, mass, wso, apparent horizon, maximum mass, dust, well behaved, thermodynamics, deceleration parameter, positrons, dark matter, scale factor, general relativity, temperature, red shift, star formation, sound, dust grains, uv, galaxy, galex, gsxr flares, particle, ~~equation of~~, quintessence, electron, metric, gas, energy, hubble parameter, angular momentum, positron

2012 black hole, dark energy, universe, magnetic field, solitary waves, scalar field, gravity, entropy, field equations, dusty plasma, energy density, cosmological constant, mass, ions, eos parameter, dark matter, plasma, dust, electron, soliton, positrons, dust grains, ion, general relativity, restricted three, stars, black holes, brane, body problem, cepheids, ~~scale factor~~, horizon, periodic orbits, positron, electrons, deceleration parameter, dls, shock waves, star, hot electrons, ion acoustic, number density, perfect fluid, amplitude, bianchi type, electric field, thermodynamics, event horizon, galaxies, ~~well behaved~~

2013 black hole, dark energy, universe, solitary waves, dusty plasma, magnetic field, black holes, entropy, energy density, scalar field, event horizon, field equations, dust, cosmological constant, horizon, gravity, ~~scale factor~~, periodic orbits, dispersion relation, deceleration parameter, ion acoustic, ions, electron, shock waves, dark matter, amplitude, positrons, mass, equilibrium points, plasma, ~~eqs~~, bulk viscous, positron, bulk viscosity, body problem, double layers, primaries, solitons, brane, solitary wave, de sitter, eos parameter, dust grains, hubble parameter, electrons, reductive perturbation, negative ions, growth rate, spectral index, da

2014 black hole, dark energy, equilibrium points, universe, body problem, primaries, scalar field, magnetic field, solitary waves, dusty plasma, dust, field equations, gravity, dark matter, oblateness, ~~restricted three~~, energy density, triangular points, ~~scale factor~~, motion, plasma, deceleration parameter, spectral index, holographic dark energy, ion acoustic, ions, hubble parameter, equilibrium point, radiation pressure, mass, chaplygin gas, electron, bianchi type, wave, black holes, infinitesimal mass, electrons, ~~eqs~~, positrons, growth rate, cosmological constant, ~~number density~~, dispersion relation, ~~equation of~~, eos parameter, redshift, ion, galaxies, solitary wave, dust grains

2015 black hole, dark energy, gravity, universe, scalar field, field equations, energy density, star, magnetic field, hubble parameter, eos parameter, mass, de sitter, dark matter, ~~scale factor~~, bianchi type, cosmological constant, equilibrium points, gr, inflation, asteroid, ~~equation of~~, deceleration parameter, body problem, ~~eqs~~, black holes, radial pressure, primaries, stars, einstein, psr j, motion, modified gravity, momentum tensor, fluid, general relativity, compact stars, perfect fluid, event horizon, quintessence, model, metric, electric field, eos, spacetime, spherically symmetric, anisotropic, gravitational collapse, disk, accelerated expansion

"*Astrophysics and Space Science* publishes original contributions and invited reviews covering the entire range of astronomy, astrophysics, astrophysical cosmology, planetary and space science and the astrophysical aspects of astrobiology. [...] We particularly welcome papers in the general fields of high-energy astrophysics,

astrophysical and astrochemical studies of the interstellar medium including star formation, planetary astrophysics, the formation and evolution of galaxies and the evolution of large scale structure in the Universe. [...]"

In Table 4 in Column 1, we show the top 50 most significant keyphrases of ASTR journal. The keyphrases describe the main terms of the journal for the past 10 years. As we expected, the terms *periodic orbits*, *black hole*, *equilibrium points*, and *dark energy* are on the top 10 list. The top 10 list shows the recent research trends of ASTR.

The top keyphrases of ASTR for each year from 2007 to 2015 are shown in Table 6. The term *star formation* is significant up to 2011, while the term *magnetic field* is used constantly over many years. *Black hole* becomes one of the most significant terms after 2010. In the Wikipedia article about black holes, we find the following description, which explains the rise of *black hole*[9]:

"... black holes do not directly emit any signals other than the hypothetical Hawking radiation; [...] A possible exception to the Hawking radiation being weak is the last stage of the evaporation of light (primordial) black holes; searches for such flashes in the past have proven unsuccessful [...]. NASA's Fermi Gamma-ray Space Telescope launched in 2008 will continue the search for these flashes."

It is likely that this new gamma-ray telescope had a significant impact on research into black holes. Evidence for this conclusion can be found in the trending significance of the keyphrase *black hole* across many years of ASTR. In less than two years, researchers began collecting data from the new telescope and publishing their new discoveries.

This year-by-year comparison of the keyphrases of ASTR shows research trends, dynamics, and breakthroughs in astrophysics. Journals could publish such keyphrase lists with significance values as supplementary data either as an appendix in each volume, or yearly, significantly assisting potential contributors in assessing where best to submit their research for publication.

6 Discussion and future work

We have presented our framework for keyphrase extraction from sets of scientific articles. The framework differs from similar commercial tools, e.g., JANE[10] and HelioBLAST[11]. The HelioBLAST text similarity engine finds text records that are similar to the submitted query. JANE (Journal/Author Name Estimator) uses the short text of an article (e.g., the title and/or abstract) and searches for journals, authors, or articles. JANE compares the document to millions of documents in the MEDLINE database. Both tools are article-based approaches. The similarity distance between query and journal is based on the accumulated similarity between query and articles.

The properties of our framework are the following:

[9] https://en.wikipedia.org/wiki/Black_hole
[10] http://jane.biosemantics.org/
[11] http://helioblast.heliotext.com/

– The extracted keyphrases are journal-dependent and the direct connections to articles are dropped. It lifts up keyphrase lists to more general journal representation.
– The extracted keyphrases allow queries to be compared to journals instead of only articles. It allows query processing to be sped up considerably.
– The extracted keyphrases expose more general representations of journals than sets of articles. The framework is simple to implement and can be adopted and used independently for separate journals. It is flexible enough that centralized databases are not substantial.

A list of keyphrases can be useful as supplementary material for the following (albeit incomplete) list of tasks:

– It can support describing the main research objectives of a journal and can help the research community to follow the main research trends and changes.
– It can support deciding whether a newly submitted article conforms with a journals topics and can help researchers to choose the most appropriate journal to which to submit a new article.
– It can help to evaluate whether a topic has been extensively studied, is a new trend, or is the revival of an old topic.
– It can help libraries and search systems to index journals, and make queries more accurate.

The lack of manually annotated data is a formidable barrier for a thorough evaluation of the quality of extracted journal keyphrases. While we can evaluate the precision of extracted keyphrases (i.e., how accurately we selected them), we cannot easily evaluate recall (i.e., have we selected all of the important ones). In the near future, we plan to involve several editorial boards to evaluate the framework's quality and relevance.

The proposed framework only uses statistics, and no language-dependent tools are necessary to apply this framework. Therefore, the framework can be applied to new journals and new languages without specifically requiring language-dependent tools.

References

1. Baldwin, T., Lui, M.: Language identification: The long and the short of the matter. In: Human Language Technologies: The 2010 Annual Conference of the NAACL. pp. 229–237. Los Angeles, CA (June 2010)
2. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.Y., Lee, D., Powley, B., Radev, D., Tan, Y.F.: The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In: Proceedings of the Language Resources and Evaluation Conference (LREC 08). Marrakesh, Morocco (May 2008)
3. Choueka, Y.: Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In: Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling. pp. 21–24. Cambridge, MA (1988)
4. Daudaravicius, V., Marcinkeviciene, R.: Gravity counts for the boundaries of collocations. International Journal of Corpus Linguistics 9(2), 321–348 (2004)

5. Daudaravicius, V.: The influence of collocation segmentation and top 10 items to keyword assignment performance. In: CICLing. pp. 648–660 (2010)
6. Daudaravicius, V.: Applying collocation segmentation to the ACL anthology reference corpus. In: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. pp. 66–75. Jeju Island, Korea (July 2012)
7. Daudaravicius, V.: Collocation segmentation for text chunking. Ph.D. thesis, Vytautas Magnus University (Jan 2013)
8. Gollapalli, D.S., Caragea, C., Li, X., Giles, L.C.: Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction (2015)
9. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: A survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 1262–1273. Baltimore, Maryland (June 2014)
10. Kilgarriff, A., Rychly, P., Kovar, V., Baisa, V.: Finding multiwords of more than two words. In: Proceedings of the 15th EURALEX International Congress. pp. 693–700. Oslo (2012)
11. Kim, N.S., Medelyan, O., Kan, M.Y., Baldwin, T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 21–26 (2010)
12. Lin, D.: Extracting collocations from text corpora. In: First Workshop on Computational Terminology. Montreal (1998)
13. Lopez, P., Romary, L.: HUMB: Automatic key term extraction from scientific articles in GROBID. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 248–251. Uppsala, Sweden (July 2010)
14. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation (60), 503–520 (2004)
15. Seretan, V.: Syntax-Based Collocation Extraction, Text, Speech and Language Technology, vol. 44. Springer (2011)
16. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics 19, 143–177 (1993)
17. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation (28), 11–21 (1972)
18. Strauss, U., Grzybek, P., Altmann, G.: Contributions to the Science of Text and Language: Word Length Studies and Related Issues, chap. Word Length and Word Frequency, pp. 277–294. Springer Netherlands, Dordrecht (2006)
19. Tjong Kim Sang, E.F., Buchholz, S.: Introduction to the CoNLL-2000 shared task: Chunking. In: Proc. of CoNLL-2000 and LLL-2000. pp. 127–132. Lisbon, Portugal (2000)
20. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval 2(4), 303–336