20 Years of Pattern Mining: a Bibliometric Survey

Arnaud Giacometti, Dominique H. Li, Patrick Marcel, Arnaud Soulet
Université François-Rabelais de Tours, LI EA 6300
3 place Jean Jaurès
F-41029 Blois France
firstname.lastname@univ-tours.fr

ABSTRACT

In 1993, Rakesh Agrawal, Tomasz Imielinski and Arun N. Swami published one of the founding papers of Pattern Mining: "Mining Association Rules between Sets of Items in Large Databases". Beyond the introduction to a new problem, it introduced a new methodology in terms of resolution and evaluation. For two decades, Pattern Mining has been one of the most active fields in Knowledge Discovery in Databases. This paper provides a bibliometric survey of the literature relying on 1,087 publications from five major international conferences: KDD, PKDD, PAKDD, ICDM and SDM. We first measured a slowdown of research dedicated to Pattern Mining while the KDD field continues to grow. Then, we quantified the main contributions with respect to languages, constraints and condensed representations to outline the current directions. We observe a sophistication of languages over the last 20 years, although association rules and itemsets are so far the most studied ones. As expected, the minimal support constraint predominates the extraction of patterns with approximately 50% of the publications. Finally, condensed representations used in 10% of the papers had relative success particularly between 2005 and 2008.

1. INTRODUCTION

In 1993, Rakesh Agrawal, Tomasz Imielinski and Arun N. Swami published one of the seminal papers of Pattern Mining [1]: "Mining Association Rules between Sets of Items in Large Databases" in the proceedings of the ACM SIGMOD International Conference on Management of Data by introducing the problem of extracting interesting association rules. Formally, this problem is to enumerate all the rules of type $X \rightarrow I$ where X is a set of items and I an item not found in X such that the probabilities $P(X \wedge I)$ and $P(I|X)$, respectively estimated by *support* and *confidence*, are sufficiently high. Agrawal et al. [1] has mostly replaced the traditional heuristic search by a complete and consistent one. Indeed, the problem of discovering classification rules (where I is a class value) was already a topic of active research in the field of artificial intelligence but the existing algorithms were not exhaustive [5; 23; 21]. These notions of completeness and consistency are crucial features of methods published in the database field, and this may partly explain its publication in ACM SIGMOD. Similarly, the generalization proposed the next year [2] (where the conclusion of the rule is now a set of items) was published in Very Large Data Bases Conference[1] (VLDB).
For 20 years, the community of *Pattern Mining* has continued to draw inspiration from this seminal paper [1] as shown by numerous citations:

- It is the 28th most cited paper in Computer Science according to CiteSeer[2],
- the 7th most cited paper in the data mining field according to Microsoft Academic Research[3] and,
- more than 12,000 citations according to Google Scholar[4].

Consequently, this paper received the ACM SIGMOD Test of Time Award in 2003. Clearly, this work has not only introduced a problem but also a new methodology at the core of Pattern Mining. Let us detail it by focusing on the original problem which is divided into two subproblems:

1. find all patterns (itemsets) present in at least s% of transactions and,
2. generate from these patterns all interesting association rules.

This division shows the two major issues that have animated the Pattern Mining community the last 20 years: the *extraction* and the *use* of patterns. First, Pattern Mining aims at enumerating all the patterns of a *language* (e.g., itemsets or sequences [3]) which satisfy a *constraint* (e.g., a minimal support). In addition, it is possible to compress the result by means of a *condensed representation* of these patterns i.e., a fraction of the patterns that guarantees the total regeneration of rules [7]. These three dimensions (i.e., language, constraint and representation) proposed by Mannila and Toivonen [18] lead to a large number of problems. Second, the use of patterns is to combine several patterns for building more complex/global models [11]. Most often the completeness of the extraction phase is a key point for this second phase. Typically, associative classifiers such as

[1]Rakesh Agrawal himself claimed his affiliation to the field of databases in an interview [26]:"I'm a database person, so my view of data mining has been that it is essentially a richer form of querying. We want to be able to ask richer questions than we could conveniently ask earlier."
[2]`citeseerx.ist.psu.edu/stats/articles`, January 2013
[3]`academic.research.microsoft.com`, March 2013
[4]`scholar.google.com`, March 2013

CBA [17] are built by combining classification rules, themselves derived from itemsets.

This paper aims to study the work related to pattern discovery published from 1995 to 2012. Rather than proposing a literature review based on a few dozen papers and necessarily partial, we opted for a bibliometric survey based on a thousand papers. We selected 1,087 papers devoted to Pattern Mining from the 6,888 papers published in five major conferences on Knowledge Discovery in Databases: KDD, PKDD, PAKDD, ICDM and SDM. Our corpus provides an overall vision and is therefore sufficient to quantify phenomena during the last two decades. At the same time, targeting certain conferences avoids the inherent latency in the settlement of very large databases as it is the case with Thomson Web of Science database [9]. The automated processings exclusively focus on the titles and the authors of publications while in addition, we rely on summaries to manually remove certain ambiguities. To the best of our knowledge, only one bibliometric study [9] has been conducted on the field of Knowledge Discovery in Databases but with a coarse granularity. In particular, this study does not particularly focus on Pattern Mining.

The main contributions of this paper are:

- **Position Pattern Mining comparatively to the rest of KDD:** Clearly the last 20 years have been marked by the development of Knowledge Discovery in Databases. In particular, the five major conferences listed above have successively appeared between 1995 and 2001. We measure the overall activity of these conferences and compare it with that of Pattern Mining.
- **Analyze the main trends in Pattern Mining:** We also want to better understand the domain of Pattern Mining in the light of the three dimensions discussed above: language, constraint and condensed representation. For each of these dimensions, we list the different categories (e.g., itemsets or sequences for the language). Then, we quantify the importance and the evolution of each category using the number of publications and their freshness.

The rest of this paper is organized as follows. Section 2 describes the scope of the study and introduces the methodology applied to the five selected conferences. Section 3 shows that Pattern Mining really is a subfield of KDD and compares their progressions. Section 4 details the distribution of papers according to language, constraint and condensed representation.

2. MATERIALS AND METHODS

2.1 Conferences on Knowledge Discovery

This study focuses on the proceedings of all the conferences whose title contains "data mining" and ranked A by The Computing Research and Education Association of Australia[5]: KDD (ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[6]), PKDD[7] (European Conference on Principles of Data Mining and Knowledge Discovery[8]), PAKDD (Pacific Asia Knowledge Discovery and Data Mining[9]), ICDM (IEEE International Conference on Data Mining[10]) and SDM (SIAM International Conference on Data Mining[11]). Table 1 indicates for each conference the year of the first edition, its h5-index, its h5-median and its rank. The first edition of all the conferences took place several years after the paper of Agrawal et al. [1] (details of the advent of KDD conference are given by Piatetsky-Shapiro [22]). h5-index is the h-index considering papers published in the last 5 complete years (2008-2012). h5-median for a publication is the median number of citations for the papers that make up its h5-index. Note that h5-index and h5-median were computed with Google Scholar (July, 2013). The rank was computed with Microsoft Academic Search using Top Conferences in data mining[12] (March 13, 2013). All these indicators underline the significance of the selected conferences.

Table 1: Selected "data mining" conferences

Conf.	First edition	h5-index	h5-median	Rank
KDD	1995	67	106	1
PKDD	1997	32	42	6
PAKDD	1997	23	32	7
ICDM	2001	37	54	4
SDM	2001	33	46	5

Unlike the bibliometric study performed by Deng et al. [9] that uses a large database (like Thomson Web of Science), we have chosen a sample of the KDD publications. First, our choice may tend to exclude more mature work published in journals and more prospective work published in workshops. We think that conferences reproduce the activity of field better thanks to their annual organization and their short submission process. Second, our work misses publications in related conferences in the field of Databases (e.g., Very Large Data Bases Conference, VLDB) and Information Retrieval (e.g., International Conference on Information and Knowledge Management, CIKM). However, integrating these conferences in the study would have diluted the essence of KDD (and therefore that of Pattern Mining). Likewise, selected events are "non specialized" conferences related to data mining unlike other more specific as Data Warehousing and Knowledge Discovery or ACM International Conference on Web Search and Data Mining.

In the end, we estimate that the average annual 383 publications from the 5 conferences is a significant and consistent sample for a statistical study of the entire world production. Our approach is independent of the latency in the settlement of the database as it is the case with Thomson Web of Science database [9]. Furthermore, the reasonable number of papers allows the use of manual approach to increase the

[5] www.core.edu.au, 2010

[6] www.kdd.org

[7] PKDD was attached in 2001 to ECML (European Conference on Machine Learning) then two conferences merged in 2008. Since 2008, PKDD corresponds to ECML/PKDD.

[8] www.ecmlpkdd.org

[9] www.pakdd.org

[10] www.cs.uvm.edu/~icdm

[11] www.siam.org/meetings/archives.php#sdm

[12] We think that ICDE and CIKM respectively ranked as the second and the third conferences are not specially dedicated to KDD. In particular, their names do not contain "data mining".

accuracy of results.

Table 2: Papers indexed by DBLP

Conf.	First indexing	# of indexed editions	Total # of papers	Average annual # of papers
KDD	1995	18	1,905	105.8
PKDD	1997	16	1,295	80.9
PAKDD	1998	15	1,277	85.1
ICDM	2001	12	1,598	133.2
SDM	2002	11	813	73.2
	Average:	14.4	1,377.6	95.8

This study focuses on the titles of publications indexed in The DBLP Computer Science Bibliography[13] for the five conferences. Although the volume and type of publications (e.g., papers, posters, tutorials, panels) vary according to the conference and the year, the vast majority are long and short papers. Only the first edition of the PAKDD and SDM conferences were not indexed by DBLP. Finally, Table 2 summarizes for each conference: the year of the first indexing by DBLP and the total number of indexed publications. If the automated processes focus exclusively on titles[14], manual validation of these treatments was based on summaries when titles were insufficient.

2.2 Semi-Automated Topic Assignment

One of the main challenges of our study is to determine what are the papers in relation with Pattern Mining. Then it is also necessary to categorize these papers according to the language, the constraint and the condensed representation. Usually, for classical surveys relying on a few dozen papers, a completely manual approach is used to analyze the content. In contrast, for bibliometric surveys based on dozen thousands of papers, keyword filtering identifies topics [9] or unsupervised methods directly learn topics [8]. Since the size of our corpus is between these two extremes, we propose to apply an intermediate solution that avoids reading all the titles of the 6,888 publications, but where assigned topics are validated by domain experts.

We now describe this semi-automatic process for assigning a topic:

1. **Keyword filtering:** The first step aims at keeping all papers which may concern the targeted topic. We automatically select all publications whose title refers to at least one keyword stemming from this topic. This list of keywords is handcrafted by domain experts. It must be open enough to avoid too many false negatives (i.e., relevant but not selected papers).

2. **Manual filtering:** The second step eliminates all false positives, i.e., publications not related to the targeted topic but retained among the publications obtained from Step 1. To do so, we read each title one by one. Whenever it is necessary, we also consult the abstract of the paper and even its entirety to remove any ambiguity.

[13] `www.informatik.uni-trier.de/~ley/db`

[14] The use of abstracts or papers in their entirety would probably improve the automatic filtering (if using NLP methods to accurately identify the contributions of the article).

Clearly our approach is not completely objective because of the choice of keywords used in Step 1 and the elimination of false positives in Step 2. Nevertheless, we think that this protocol does not introduce more subjectivity than traditional literature review. In addition, as with any automatic process, errors may occur even if Step 2 eliminates some.

2.3 Measuring Activity

The metrics most commonly used in bibliometric studies are those based on the number of publications (to measure the activity) and the number of citations (to measure the popularity). We exclusively focus on the number of publications because citation indexes (like citation number or h-index [15]) increase over time and would need to take into account other data sources. Additionally, we introduce a new indicator to estimate the dynamism of a topic through the publications devoted to it. The *freshness* measures whether a paper p is recent compared to the period covered by a reference set of publications $\mathcal{P}$ (containing at least two distinct years):

$$freshness(p, \mathcal{P}) = \frac{p.year - \min(\mathcal{P}.year)}{\max(\mathcal{P}.year) - \min(\mathcal{P}.year)}$$

where $p.year$ is the publication year of p, $\min(\mathcal{P}.year)$ is the oldest year and $\max(\mathcal{P}.year)$ is the most recent one.

For instance, in this study, the reference set of publications $\mathcal{P}$ always corresponds to all the publications of the five selected conferences. Then, a freshness of a publication close to 1 means that this publication is recent i.e., close to 2012. Conversely, a freshness of 0 means that the publication dates of 1995. We then extend this measure to a set P of publications by calculating its average value:

$$freshness(P, \mathcal{P}) = \frac{1}{|P|} \times \sum_{p \in P} freshness(p, \mathcal{P})$$

This metric gives a rough trend of the dynamism of a domain through its publications P compared to a reference set of publications $\mathcal{P}$. When the freshness of a set of publications reaches 1, this means that the publications focus on the last years of the period 1995-2012. As baseline, the freshness of the 6,888 publications selected for the study is 0.657 (and not 0.5) due to the increase of the number of annual publications. We will also use the freshness to observe languages, constraints or condensed representations. In the following, we consider that a topic is dynamic if it exceeds the freshness of KDD (i.e., 0.657).

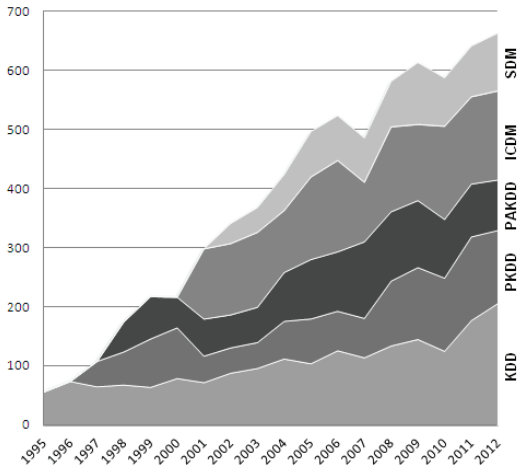
3. PATTERN MINING IN KDD

The purpose of this section is to delimit the field of Pattern Mining inside Knowledge Discovery and to compare their evolution.

3.1 The Growth of KDD

Figure 1 shows the number of publications in KDD for each conference over the years between 1995 and 2012. This number has steadily increased over the 18 years (except in 2000, 2007 and 2010), reflecting the growth of the field of knowledge discovery. This overall increase is due to both the creation of new conferences (until 2002) and the increase of the number of publications for each conference (e.g., more than 94% of publications since 2002). Piatetsky-Shapiro also notes the development of the field with the increase in

Figure 1: Number of publications in KDD



the number of companies providing data mining tools and the increase of KDNuggets subscribers [22]. In recent years, a stabilization of the number of publications has emerged. Indeed the increase rate since 2009 has been below average and even gradually slowing down.

For having a recent overview of the field, Figure 2 depicts the word cloud of KDD giving greater prominence to fresh words that appear more frequently in the titles. More precisely, the 100 freshest words appearing at least 5 times in titles were selected based on the measure introduced in Section 2.3. Font size is proportional to frequency (using a log scale) and grayscale represents freshness.

Strong issues clearly emerge from Figure 2:

- **Recent (or rediscovered) methods:** MapReduce, multi-task learning, hashing, matrix factorization, gradient descent
- **Specific types of data:** uncertainty, mobility (location), sparsity
- **Application domains:** advertising, social media (twitter, opinion mining, tag), malware detection, privacy (publishing, anonymizing)

Interestingly, KDD issues highlighted in the summary of Fayyad et al. [10] were especially observed in recent years like privacy concerns and the importance of web content mining (amplified by social networks).

3.2 Local Patterns vs Global Models

The major difficulty is to determine the publications concerning Pattern Mining. By pattern, we mean *local* pattern in the strict sense of the word [14], i.e., that describes a portion of the database. For this reason, we consider that decision trees, Bayesian networks, neural networks or support vector machines are not local patterns, but global models. However, we do not limit ourselves to the works dedicated to the extraction of patterns but we also consider the works benefiting from local patterns (e.g., to build global models like classifiers).

We use the semi-automatic process described in Section 2.2. A list of words was compiled by relying on the three dimensions mentioned in the introduction:

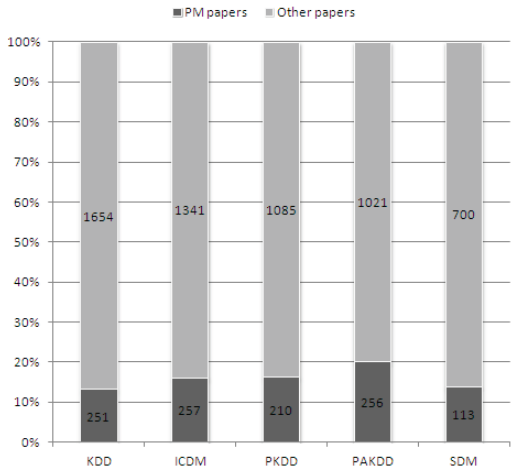
1. *Language:* pattern, item, sequence, rule, tree, graph, string, association, stream, subgroup, episode
2. *Constraint:* support, frequent, monotone, anti-monotone, constraint, contrast
3. *Condensed representation:* free, generator, closed, condensed, concise

Of course, this list of terms was extended to their variations (e.g., the term string leads to substring, strings and so on). Thus, 1,732 publications were initially identified as Pattern Mining papers representing about a quarter of the database after this first step. The manual filtering excluded all publications not related to Pattern Mining among this collection. Finally, 1,087 publications were identified as Pattern Mining papers knowing that summaries were consulted in 148 cases. As expected, the number of false positive was high because we did not want to miss too many relevant papers.

As mentioned above relevant papers were missed. For estimating this number of false negatives, we randomly selected 100 publications from the 5,156 excluded in Step 1. 5 of these publications are dedicated to Pattern Mining. Therefore, we estimate that 258 ($= 0.05 \times (6,888 - 1,732)$) publications relating to Pattern Mining have been missed by our approach. Following this estimation, we assume that the 1,087 selected publications are a representative sample of Pattern Mining.

3.3 The Slowdown of Pattern Mining

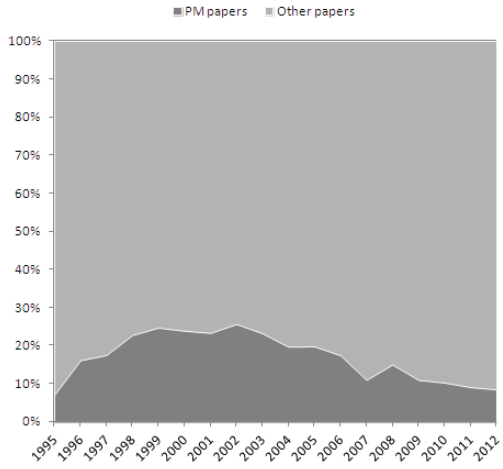
Figure 3: Portion of publications in PM per conference



Pattern Mining (denoted by PM on figures) is really a subfield of KDD since about 1 paper out of 6 concerns it (1,087 out of 6,888). More precisely, Figure 3 shows the number of publications in Pattern Mining by conference and calculates their proportions. We note that the papers are fairly distributed throughout the conferences. Among the 9,368 authors who contributed to the 5 conferences, 1,789 of them (19.09%) participated in at least one publication in Pattern

Figure 2: Word cloud of Knowledge Discovery in Databases

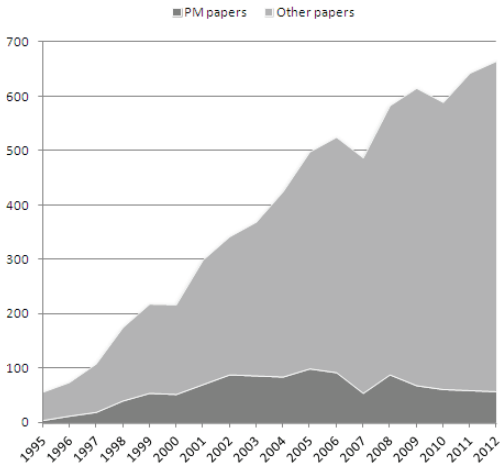
Figure 4: Portion of publications in PM per year



Mining. 890 authors (10.02%) were confined to publications in Pattern Mining.

Figures 4 and 5 show respectively the relative and absolute evolution of Pattern Mining compared to KDD. We see a flourishing decade between 1997 and 2006 where the portion of Pattern Mining is higher than its average portion of 16%. During the period between 1998 and 2003, 1 paper out of 5 is even devoted to Pattern Mining, its golden age somehow. Until 1999, the development of Pattern Mining is more important than that of KDD (see Figure 4). It is sure that KDD, PKDD and PAKDD were privileged conferences to disseminate the work related to the emerging field of Pattern Mining [22]. In contrast, other works like decision trees had already recognized forums with conferences in machine learning and artificial intelligence. The recognition of "data mining" conference might have attracted more diverse work from 2002. Anyway, since 2002, the portion of Pattern Mining has gradually decreased to less than 8.6% in 2012 (close to 7% in 1995). This relative decline is also reflected by an absolute considerable drop from 99 papers in 2005 to 57 papers in 2012 (see Figure 5). Finally, the freshness of the 1,087 publications about Pattern Mining is only 0.579 while the freshness of KDD is 0.657. The difference therefore reflects a weakening of Pattern Mining compared to the whole of KDD.

Figure 5: Number of publications in PM per year



4. TYPOLOGY OF PUBLICATIONS

The three dimensions discussed in this section are clearly distinguishable in Figure 6 that depicts the word cloud of Pattern Mining. This figure gives a greater prominence to fresh words that appear more frequently in the titles. *1) Language* A large number of words are related to the nature of the input data and extract patterns: "stream", "dynamic", "spatio-temporal" and so on. *2) Constraint* The expected semantic of patterns imposed by the constraint is particularly emphasized with "contrast", "utility" or "correlated". *3) Condensed representation* Finally, the method

Figure 6: Word cloud of Pattern Mining

to condense the extracted collection is modestly visible (e.g., "closed" or "generator").

Each dimension has been studied in isolation because the intersection of two dimensions concerns few papers in general (if closed frequent itemsets are ignored). To give an overview, among the papers that use a constraint or a condensed representation, 31% consider either the frequent itemsets, the closed itemsets or the closed frequent patterns.

4.1 Language

Language assignment

A language gathers all properties or all possible subgroups of the data [18]. Originally, the first language consisted of all sets of items due to the context of basket market analysis [1]. We have compiled a list of languages in leveraging our domain knowledge and the words that are the most frequently used in titles (see Figure 6). The topic assignment method is used for each type of languages as described in Table 3. "generic" language denotes an approach dedicated to multiple languages [18][15]. In this study, all papers (including unclassified ones) are scanned manually for final classification (using the title and the abstract). During this phase, it was observed that the word "pattern" implicitly refers to "itemset" as 148 papers containing this word corresponds to itemsets. The last column of Table 3 reports the freshness of publications associated with each language by highlighting in bold those with positive dynamic compared to KDD (≥ 0.657).

As expected, association rules and itemsets which are at the origin of Pattern Mining, are the most studied up to approximately 2/3 of the whole. About a quarter of papers concerns sequences and graphs. The discovery of patterns in spatio-temporal data and relational data remains quite marginal. More surprisingly, we find that very few studies have addressed generic approaches in terms of language. A probable explanation is the difficulty to propose a general framework both theoretically and in terms of implementation[16]. The high freshness of this topic (0.683) indicates, however, a rather recent interest for this type of work. Furthermore, Figure 7 depicts the evolution of the four most representative languages during the past two decades. To smooth the results and make them more readable, we divided the period into 5 slices of four years. The plots report the average results given in absolute (left) and in percentage (right).

Table 3: List of languages

Language	Keywords	Nbr.	Prop.	fresh.
rule	association	345	0.32	0.445
itemset	set	340	0.31	0.624
sequence	episode, string, stream, protein, periodic, temporal	190	0.17	0.632
graph	molecular, structure, network	107	0.10	**0.712**
tree	xml	49	0.05	0.610
spatial	spatio-temporal	30	0.03	**0.688**
generic	–	18	0.02	**0.683**
relational	–	8	0.01	0.588

Total number: 1,087 ; Average freshness: 0.579

Language sophistication

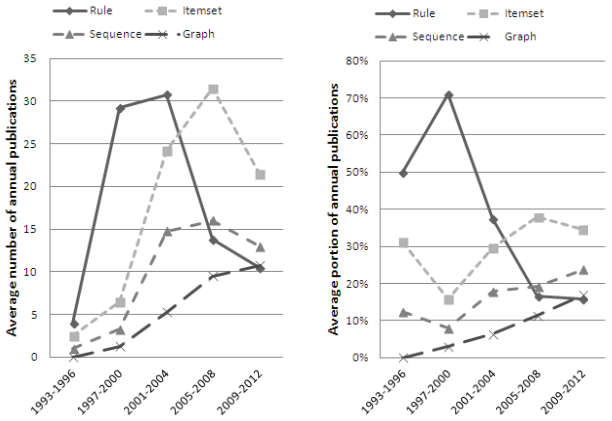
Table 3 shows that the more complex a language, the fewer papers dedicated to it. First, the intrinsic complexity related to the combinatorial problem makes it difficult to exhaustively extract patterns when sophisticated languages are involved. For example, with three items, it is possible to form 80 distinct sequences against only 8 itemsets. Second, the evolution of this sophistication of language was gradual as

[15]For this language, there are no keywords in addition to "generic".

[16]However, journals (absent from our data) may be more appropriate for generic approaches often making the synthesis of previously published work for distinct languages. For instance, the pattern-growth method [19] is at the core of several publications focusing on either itemsets or sequences.

described in Figure 7: itemsets, sequences and then, graphs. In fact, the knowledge gained with the first languages have reduced the number of scientifical challenges for the next languages. Typically, pruning methods of the search space for itemsets (based on anti-monotonicity for instance) are transferable to other languages.

Figure 7: Evolution of the number of publications per language



Nevertheless, we observe two exceptions with trees and itemsets which are respectively less studied than graphs and rules. Trees are sometimes simplified to be treated as variants of sequences or as special cases of graphs. The most notable exception are itemsets that are simpler than association rules, and yet less studied. The fact that rules were particularly studied prior to 2000 can be explained historically, since the extraction of classification rules was already an important research topic in artificial intelligence before 1993. In addition, the seminal paper [1] designated association rules as the ultimate goal while itemsets are considered as intermediate tools (although technically, obtaining itemsets is the most difficult phase). The low freshness of the papers about rules (0.455) reinforces the hypothesis of these historic roots.

Limit of the sophistication

While the proportion of publications concerning rules and itemsets decreases, the most sophisticated languages continue to progress in Pattern Mining (see Figure 7) with +4.5% for sequences and +5.5% for graphs[17]. For example, over the last 4 years, the proportion of papers devoted to graphs exceeds those devoted to rules. However, this sophistication reaches its limit because no language (even spatio-temporal or relational patterns) seems to succeed to graphs significantly. These data may not be available in sufficient quantity while those available are reduced to simpler languages such as graphs. Given the significant freshness of the "spatial" language, this tentative conclusion could be revised soon (especially as mobility is a hot topic, see Section 3.1).

[17] A recent survey [16] confirms the craze for subgraph mining between 1994 and 2007 through bibliometric information.

Finally, the features of the input data such as incompleteness or scalability could become an issue more important than the nature of language (e.g., itemset or sequence). Indeed, we observed that some keywords appear more in the titles: uncertain data ("uncertain" with a freshness of 0.852), heterogeneous data ("heterogeneous" with 0.823), massive data ("massive" with 0.686) or dynamic data ("dynamic" with 0.684).

4.2 Constraint

Constraint assignment

Mannila and Toivonen [18] define a constraint as a selection predicate. Its goal is to restrict the mining of patterns to those useful according to the domain or task. For instance, the construction of a classifier may require contrast patterns. The topic assignment method is used with the list of constraints described in Table 4. This list was compiled using the words that are the most frequently used in titles (see Figure 6). The topic "significant" includes pattern selection based on statistical validity while "interesting" refers to more varied approaches often based on subjective knowledge. Note that the term "generic" corresponds to approaches dedicated to a class of constraints (e.g., anti-monotone constraints [18], convertible constraints [20]). As for languages (and probably for the same reasons), few publications are devoted to generic constraints.
In addition, Figure 8 depicts the evolution of the five most important types of constraints during the past two decades.

Table 4: List of constraints

Constraint	Keywords	Nbr.	Prop.	fresh.
regularity	frequent, support	263	0.48	0.608
contrast	emerging, discriminative	72	0.13	0.573
significant	chi-square, correlated	57	0.11	0.647
interesting	relevant	50	0.09	0.547
generic	monotone, anti-monone, constraint	42	0.08	0.609
exception	abnormal, surprising, anomaly, unexpected	32	0.06	0.551
utility	–	22	0.04	**0.754**

Total number: 538 ; Average freshness: 0.604

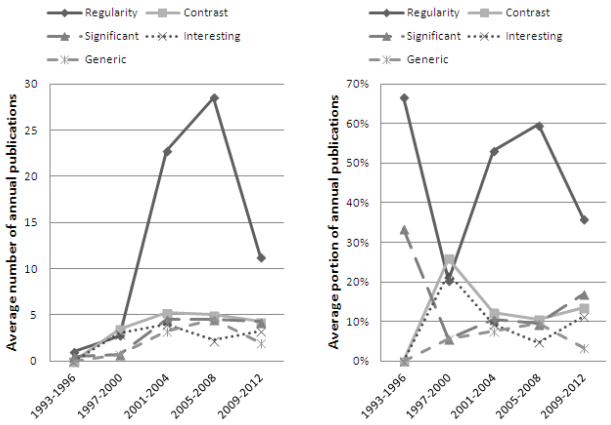
The obsession of frequency

Overall, the minimal frequency constraint with 50% of publications is by far the most used. Indeed, many papers tackles Subproblem 1 (presented in the introduction) so as to provide a new or more effective algorithm by variying either the language in input or the condensed representation in output. Rather than an application need, we are convinced that the recurrent use of frequency constraint stems from the paradigm imposed by the seminal paper [1] as explained below.
First, replacing the frequency constraint by a more selective one prevents the regeneration of *all* interesting rules (Subproblem 2). However, this completeness is a pillar of the paradigm that leads to an overabundance: "When we

started doing data mining, we were concerned that we were generating too many rules, but the companies we worked with said, 'this is great, this is what exactly what we want!' " said Agrawal [26]. Therefore, the constraint may eliminate some patterns that could be essential for the domain expert. This fear is also illustrated by the obsession to reduce the minimal support threshold even if the vast majority of extracted patterns become spurious.

Second, the evaluation of the approach proposed by Agrawal et al. [1] is not based on the evaluation of the quality of extracted patterns as it is the case with classifiers when cross-validation is performed. More generally, most papers about Pattern Mining do not evaluate the quality of extracted patterns but the efficiency of algorithms in terms of running time and amount of required memory. From this point of view, improving the process of extracting patterns means reducing the cost of time and/or space, but above all leaves the result unchanged, i.e., frequent patterns. In addition, the minimal frequency constraint has interesting properties (due to anti-monotonicity) that facilitate extraction. Evaluation of a method based on another constraint is doubly disadvantageous. Indeed, if a relevant constraint does not satisfy the anti-monotone property, the stemming mining algorithm will be less efficient than the one dedicated to extraction of frequent patterns. Moreover, it is difficult to demonstrate that the extracted patterns according to a new constraint are better than those extracted with the minimal frequency constraint because there is no objective validation protocol.

Figure 8: Evolution of the number of publications per constraint



Toward better quality of mined patterns

Now, whatever the language, the extraction of frequent patterns is a well-mastered task. For this reason, the number of publications on frequent patterns have plunged since 2005 (see Figure 8). This fall partly explains the decrease in Pattern Mining. The combinatorial challenge due to the large search space of patterns gives way to the quality of extracted patterns. Thus, the use of a constraint to refine the filtering gains legitimacy following the perspective proposed by Agrawal: "we need work to bring in some notion of 'here is my idea of what is interesting,' and pruning the generated rules based on that input." [26]. However, the definition of such constraints remains a complex issue. The proposal of a general theory of *Interestingness* was already indicated as a challenge for the past decade by Fayyad et al. in 2003 [10]. Later, Han et al. [13] follow the same idea: "it is still not clear what kind of patterns will give us satisfactory pattern sets in both compactness and representative quality".

Despite this difficulty, the freshness of certain topics shows a renewal of constraint-based pattern mining. Even if the freshness of the "contrast" topic is low (only 0.573, see Table 4), there is a high freshness 0.840 for "discriminative", 0.709 for "subgroup" and 0.764 for the word "contrast". This renewed interest is also marked for significant patterns (with a freshness of 0.647) and especially for utility constraints (with a freshness of 0.754). This dynamic is also visible on the right graph in Figure 8. Finally, instead of using a filtering based on thresholds, another way to point out relevant patterns is the ranking of patterns using a measure as illustrated by the words "ranking" and "top-k" with a freshness of 0.774 and 0.764 respectively.

4.3 Condensed Representation

Representation assignment

The topic assignment method is used with the list of condensed representations described in Table 5. As a reminder, the purpose of condensed representations is to reduce redundancies between patterns [7]. The notion of borders relies on the most general/specific patterns with respect to the inclusion. The closed patterns and generators (free or keys) operate on the same principle but with equal frequency. Some generalizations of free patterns based on minimality (e.g., non-derivable itemset [6]) are counted with the "free" topic. Note that the topic "other" includes mainly articles focusing on the generic bases of association rules [4].

As done for the two other dimensions, Figure 9 reports the evolution of the different kinds of condensed representations during the past two decades.

Table 5: List of condensed representations

CR	Keywords	Nbr.	Prop.	fresh.
closed	closure	73	0.59	0.666
border	maximal, minimal	25	0.20	0.581
free	generator, non-derivable, NDI	16	0.13	0.636
other	–	9	0.07	0.523

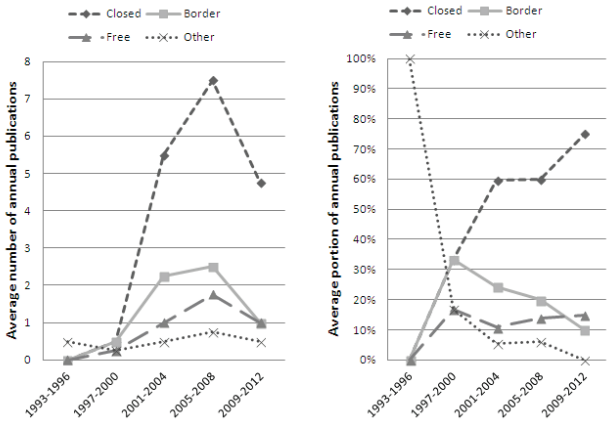
Total number: 123 ; Average freshness: 0.634

The success of condensed representations: maximal patterns and closed patterns

11.31% of the publications about the discovery of patterns exploit the concept of condensed representation. This relative success stems from their undeniable benefit and easier validation (i.e., a methodological context opposite to that of constraints). The concept of condensed representation quickly became indispensable because it combines the reduction of the number of patterns and the conservation of completeness through regeneration. From this point of view, it fits perfectly in the context of Subproblem 2. In addition, the works on condensed representations are easy to evalu-

ate. On the one hand, the validity of regeneration can be formally demonstrated. On the other hand, the quality of the reduction can be empirically estimated by calculating the ratio of compression. Usually, this compression gain is accompanied by a gain in speed and reduced consumed memory resources.

Figure 9: Evolution of the number of publications per condensed representation



Among the different representations, borders are the first successful representation (see Figure 9) even if the number of papers on borders decreases steadily from 1997-2000. By nature, these maximal/minimal patterns have extreme properties (e.g., very low frequency for maximal patterns) and do not allow to infer the properties of other patterns (e.g., infer the frequency of a smaller set). The free and closed patterns by coping with these limits have been widely adopted. Finally, the overwhelming success of the closed patterns compared to generators can be explained by a combination of factors: fewer, easier to extract and higher statistical validity (for instance, p-value is maximized by closed patterns [12]).

Good representation, but bad model

Now, techniques for condensed representations are well mastered (especially those based on closure) for most languages. The number of publications on this topic peaked between 2005 and 2008. Only the publications dedicated to generators and closed patterns persist in the landscape of Pattern Mining. However, the size of condensed representations (exact or approximate) is still too large to allow a comprehensive analysis of patterns. It is hence necessary to use other mechanisms to reduce their size either by individually filtering each pattern (using constraints) or by collectively filtering patterns (building model). The latter option is similar to the original purpose of condensed representations but it does not guarantee a perfect regeneration of patterns. This direction can be seen as a rich use of patterns: Han et al. [13] underlined that "to make frequent pattern mining an essential task in data mining, much research is needed to further develop pattern-based mining methods". The keywords "collection" and "pattern-based" with a freshness of 0.739 and 0.680 respectively show a recent interest on this topic.

5. CONCLUDING REMARKS

The seminal paper [1] has initiated a school of thought strongly influenced by the field of databases. In contrast to the field of Machine Learning, particular attention is paid to complete and consistent extractions while the evaluation is mainly based on the speed and the required memory. Following this paradigm, a community of thousands of scientists contributed to the development of incredibly efficient algorithms whatever the constraint or the language. This study has highlighted and confirmed some insights about Pattern Mining:

- The flexibility of the language is an essential feature of Pattern Mining. It enables to handle complex data that cannot be considered by many methods. This explains the sophistication of languages during the last two decades. In addition, changing the language has also offered a wide variety of challenges in the community of Pattern Mining. If the current trend continues the pattern languages related to spatial-temporal data and multi-relational data should occupy a prominent place.

- The craze for the minimal frequency constraint is again the fingerprint of databases where the speed of execution of a query is essential and the large size of the response is not a problem. Yet too large answers become a problem in knowledge discovery. For that reason, it is better to use more selective constraints while maintaining the original notion of completeness. But the lack of objective gold standards makes it difficult to assess patterns extracted under such constraints. Furthermore, the ranking of patterns (like top-k approach [24]) is now a popular method which has the advantage of directly controlling the number of patterns unlike filtering based on threshold.

- Beyond the benefit of the reduction (not enough to make collections of patterns readable), we think that the gain in speed and memory of condensed representations ensured their success. This explains why the condensed representations that have the highest compression rate (including those based on generators) are not the most popular ones. On the other hand, these works can also be seen as a promising and rich use of local patterns to build global models. Also, more recent works [25] release the completeness constraint to provide pattern-based models.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *SIGMOD Conference*, pages 207–216. ACM Press, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB*, pages 487–499, 1994.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. L. P. Chen, editors, *ICDE*, pages 3–14. IEEE Computer Society, 1995.

[4] J. L. Balcázar. Minimum-size bases of association rules. In W. Daelemans, B. Goethals, and K. Morik, editors, *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science*, pages 86–101. Springer, 2008.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[6] T. Calders and B. Goethals. Non-derivable itemset mining. *Data Min. Knowl. Discov.*, 14(1):171–206, 2007.

[7] T. Calders, C. Rigotti, and J.-F. Boulicaut. A survey on condensed representations for frequent sets. In *Constraint-Based Mining and Inductive Databases*, volume 3848 of *LNCS*, pages 64–80. Springer, 2004.

[8] D. Chavalarias and J.-P. Cointet. Phylomemetic patterns in science evolutionthe rise and fall of scientific fields. *PLoS ONE*, 8(2):e54847, 02 2013.

[9] S. Deng, Y. Tian, and H. Zhang. Using the bibliometric analysis to evaluate global scientific production of data mining papers. In *DBTA*, pages 233–238. IEEE Computer Society, 2009.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and R. Uthurusamy. Summary from the KDD-03 panel: data mining: the next 10 years. *SIGKDD Explor. Newsl.*, 5(2):191–196, Dec. 2003.

[11] J. Fürnkranz and A. J. Knobbe. Guest editorial: Global modeling using local patterns. *Data Min. Knowl. Discov.*, 21(1):1–8, 2010.

[12] A. Gallo, T. D. Bie, and N. Cristianini. MINI: mining informative non-redundant itemsets. In J. N. Kok, J. Koronacki, R. L. de Mántaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *PKDD*, volume 4702 of *Lecture Notes in Computer Science*, pages 438–445. Springer, 2007.

[13] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86, 2007.

[14] D. J. Hand. Pattern detection and discovery. In *Pattern Detection and Discovery*, volume 2447 of *LNCS*, pages 1–12. Springer, 2002.

[15] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, Nov. 2005.

[16] C. Jiang, F. Coenen, and M. Zito. A survey of frequent subgraph mining algorithms. *Knowledge Eng. Review*, 28(1):75–105, 2013.

[17] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In R. Agrawal, P. E. Stolorz, and G. Piatetsky-Shapiro, editors, *KDD*, pages 80–86. AAAI Press, 1998.

[18] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov.*, 1(3):241–258, 1997.

[19] J. Pei and J. Han. Constrained frequent pattern mining: a pattern-growth view. *SIGKDD Explorations*, 4(1):31–39, 2002.

[20] J. Pei, J. Han, and L. V. S. Lakshmanan. Pushing convertible constraints in frequent itemset mining. *Data Min. Knowl. Discov.*, 8(3):227–252, 2004.

[21] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[22] G. Piatetsky-Shapiro. Knowledge discovery in databases: 10 years after. *SIGKDD Explor. Newsl.*, 1(2):59–61, Jan. 2000.

[23] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[24] A. Salam and M. S. H. Khayal. Mining top-k frequent patterns without minimum support threshold. *Knowl. Inf. Syst.*, 30(1):57–86, 2012.

[25] J. Vreeken. Making pattern mining useful. *SIGKDD Explorations*, 12(1):75–76, 2010.

[26] M. Winslett. Interview with Rakesh Agrawal. *SIGMOD Record*, 32(3):83–90, 2003.