80 New Packages to Mine Database Query Logs[*] [Vision]

Thibault Sellam
CWI, the Netherlands
thibault.sellam@cwi.nl

Martin Kersten
CWI, the Netherlands
martin.kersten@cwi.nl

ABSTRACT

The query log of a DBMS is a powerful resource. It enables many practical applications, including query optimization and user experience enhancement. And yet, mining SQL queries is a difficult task. The fundamental problem is that queries are symbolic objects, not vectors of numbers. Therefore, many popular statistical concepts, such as means, regression, or decision trees do not apply. Most authors limit themselves to ad hoc algorithms or approaches based on neighborhoods, such as k Nearest Neighbors. Our project is to challenge this limitation. We introduce methods to manipulate SQL queries as if they were vectors, thereby unlocking the whole statistical toolbox. We present three families of methods: feature maps, kernel methods, and Bayesian models. The first technique directly encodes queries into vectors. The second one transforms the queries implicitly. The last one exploits probabilistic graphical models as an alternative to vector spaces. We present the benefits and drawbacks of each solution, highlight how they relate to each other, and make the case for future investigation.

1. INTRODUCTION

The query log of a SQL database gives us precious hints about what its users are interested in. From this dataset, we can infer query auto-completions [2, 15, 19]. We can simulate realistic queries, for testing purposes [22]. We can even reduce the latency of the queries, thanks to speculative execution [18]. Furthermore, the log describes the database itself: it describes which queries succeeded or failed, how long they took, and how many tuples they returned. Combined with predictive algorithms, this information could help us emit warnings, chose efficient query plans and build more robust engines.

Yet, mining query logs is subject to a fundamental problem: SQL queries do not live in a vector space. In their

---

[*]Number of R packages for machine learning recommended by the official website (December 2015)
`cran.r-project.org/web/views/MachineLearning.html`

natural form, queries are structured, symbolic objects - not vectors of real numbers. Hence, the vast majority of statistical concepts are undefined. Elementary methods such as means, correlations or regression do not apply. The same problem arises with advanced methods such as neural networks or SVMs. Consequently, most authors resort to *application-specific frameworks* [1, 12, 13, 24, 25, 26]: they devise some encoding specifically for the problem at hand, and feed it to a custom algorithm. This approach is neither practical nor efficient, because each use case requires a complete new representation system and a new algorithm.

A few authors have developed more general, application-independent solutions: *neighborhood-based algorithms* [2, 3, 7, 17]. These algorithms are popular because they require no encoding. Instead, they rely on a *pairwise dissimilarity function*, which quantifies the similarity or difference between two queries. Once the authors have defined such a function, they apply it to all the pairs of queries in the log. They obtain a neighborhood graph, in which they detect discrete patterns. But these methods are limited: we observed that few papers, if any, venture beyond the strict realm of clustering and k Nearest Neighbors (kNN). One explanation is that statistical textbooks and software provide little support for other tasks. To illustrate, the official R Website does not even mention NN-regression on its machine learning page (cf. footnote). Besides, these approaches suffer from qualitative drawbacks. They cannot interpolate between training examples, e.g., to compute centroids. They have little to no notion of prediction confidence. Finally, they are very sensitive to small training sets, local sparsity, and class imbalance. Several empirical studies reveal cases where they are under-optimal [9, 16].

Our ambition is to unlock the rest of the statistical toolbox. We want to perform kNN and clustering, but also density estimation, sampling, regression, classification, dimensionality reduction, reinforcement learning and visualization, directly over SQL queries. To do so, we develop methods to encode the query log in such a way that it becomes subject to these tasks. We envision software "converters", to process query logs in R, Weka or Matlab as if they were classic tables of numbers. Thus, database designers will benefit from the rich libraries offered by these platforms. They will be able to focus on insights and functionalities rather than implementation.
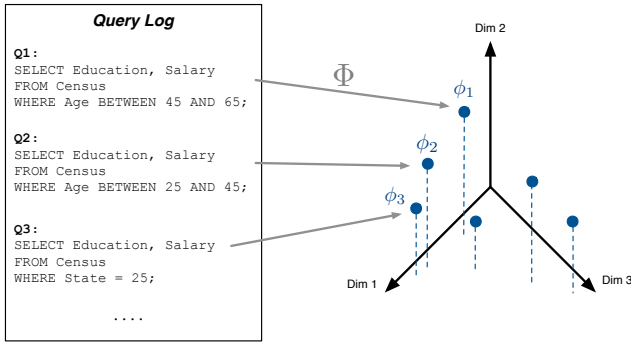
Figure 1: Example of feature map Φ.

Figure 2: Heatmap of the log's density. The dashed rectangle represents the recommendation window, defined by the user's input.

In this paper, we describe promising methods to represent query logs in an application-independent fashion. We present three families of encodings:

- *Feature maps* directly transform queries into vectors.

- *Kernel methods* manipulate queries as if they were vectors, but without actually transforming them.

- *Bayesian methods* rely on probabilistic graphical models rather than vector spaces.

We highlight the advantages and drawbacks of each solution, and present mathematical transformations to switch from one representation to the other. For all three families, we make the case for longer term investigations.

The rest of the paper is organized as follows. In Section 2, we motivate our work and we present our requirements. In Sections 3, 4 and 5, we introduce our solutions. We highlight their relationships in Section 6. We discuss related work in Section 7, and conclude in Section 8.

2. OVERVIEW

We established that queries do not live in a vector space. But what if we could devise a function Φ to transform SQL statements into vectors? In this section, we present the immense range of practical applications which would follow. We then discuss how realistic this vision is.

2.1 Visions for Query Log Mining

From Queries to Vectors. Suppose that we could access a function Φ, to map any SQL query $Q \in \mathbb{Q}_{SQL}$ to a vector $\phi \in \mathbb{R}^D$. We illustrate it in Figure 1. To be consistent with the machine learning literature, we name it *feature map* [5], and we suppose that it is one-to-one. How could this function be useful?

First, we could perform *density estimation*: for each query Q, we could estimate the probability function $p(\Phi(Q))$, as illustrated in Figure 2. The density function is a powerful tool, because it lets us perform many classic tasks from the log mining literature. For instance, we could detect "hot zones" in the query log (i.e. clusters). We could also recommend queries: when users start typing SQL statements, they implicitly define a window of interest, as shown in Figure 2. To help them, we could highlight the most popular queries in this window.

More importantly, a function Φ would allow us to perform *regression* and *classification*. In regression, we infer
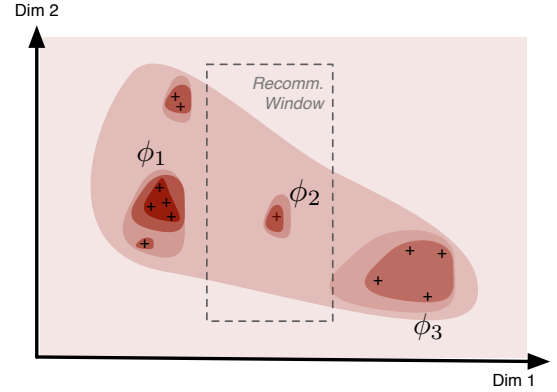
quantities from SQL statements, based on past observations. Thanks to this method, we could estimate the runtime of a query, the cardinality of its output, or or the number of machines involved in a cluster. In classification, we predict a discrete variable. Thus, we could detect which user is currently querying the database, and pre-fetch some data accordingly. We could also emit warnings, if the user's query is dangerously close to one that failed previously. Finally, we could machine-learn tasks which were previously coded by hand: among others, we could train a neural network to associate SQL queries with visualizations.

To conclude, the combination of the function Φ and statistical algorithms would lead to dozens of applications. A few of them have been proposed in the literature before (those related to density estimation), others are new. In any case, they would all run on top of a unified, complete formalism.

From Vectors to Queries. We now go one step further: what if we had access to an *inverse feature map* Φ^{-1} to reconstitute queries from vectors?

The function Φ^{-1} would have a dramatic effect: it would let us create new queries from scratch. Observe the density function pictured in Figure 2. By *sampling* from this distribution, we could produce queries that have never been written before, but which are likely to occur. Thus, we could generate artificial, but realistic workloads. This technique could be useful for testing and exploration. Combined with adaptive indexing mechanisms such as database cracking [14], it could also help us build more efficient indices.

Another application of this idea is *query regression*: we could extrapolate SQL queries from other SQL queries. Consequently, we could detect usage patterns, and exploit those to predict which query will come next, using time series models. Figure 3 provides an example. This scenario is fictive, and we suspect that real workloads are more chaotic in practice. But we do not need to predict precise queries. Predicting general areas of interest would already be helpful, and probabilistic methods excel at that.

Finally, more applications could come from *active learning*. In particular, we envision adaptive DBMS benchmarks. Such systems would pose queries, observe how the database reacts and adapt their behavior accordingly. Thus, they would automatically identify performance bottlenecks, and report them to DBMS designers.
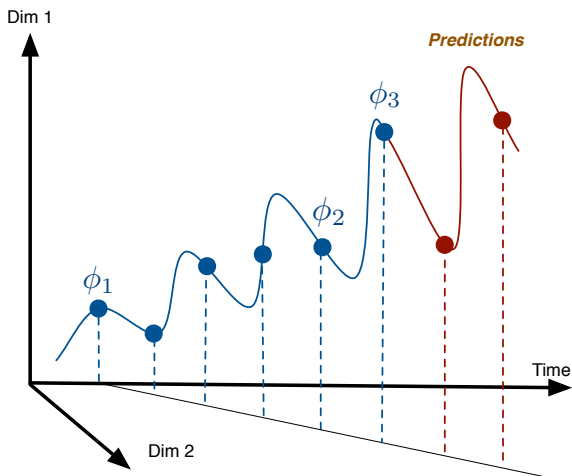
Figure 3: **Example of querying pattern based on time.**


```
Q1:SELECT Education, Salary FROM Census
WHERE Age BETWEEN 45 AND 65;

Q2: SELECT Education, Diploma FROM Census
WHERE Age BETWEEN 25 AND 45;
```

Φ

ID	SELECT			FROM	WHERE	
	Education	Salary	Diploma	Census	Age-Min	Age-Max
Q1	1	1	0	1	45	65
Q2	1	0	1	1	25	45

Figure 4: **Example of dummy coding.**

2.2 How Far Are We?

In fact, constructing a function to map queries to vectors is not a difficult task. For example, we could count n-grams, as in information retrieval. The whole challenge is to build an application-independent transformation. Such a transformation should be *lossless*, that is, non destructive. The vector representation of a query should convey all the information contained in its SQL form. It should contain lexical and grammatical information: which keywords are used, and what are their roles. But it should also convey the *set relationships* between the queries. By nature, queries represent sets of tuples, which can be disjoint, overlapping, or nested. With continuous variables, they can even be ordered. These properties should be preserved in the encoding. The actual feature selection, which depends on the use case, should be left to the user.

Unfortunately, we suspect that if such a mapping Φ exists, then the vector space it yields will have a huge, unpractical dimensionality. We discuss this point further in Section 3. In the rest of this paper, we present several restricted versions of the function Φ. Two of these methods are lossless: dummy coding and Bayesian modeling. However, their scope is limited: we have not yet found any practical way to process all the possible queries from SQL. The remaining approaches are more flexible, but they are lossy. The users must specify the properties of interest in advance. For instance, they may focus on the syntactical structure of the queries, or their extent. The encoding will reflect these attributes, and destroy the remaining information. Consequently, two distinct queries can have the same encoding, and the inverse mapping Φ^{-1} is undefined.

3. FEATURE MAPS

We now present two methods to build feature maps, *dummy coding* and *dissimilarity-based feature maps* (DBFMs).

3.1 Dummy Coding

Method. The idea behind dummy coding is to represent queries with vectors of binary variables, where each component represents a degree of freedom offered by SQL. For example, a variable could signal the presence or absence of a certain table in the WHERE clause, or an aggregation in the SELECT section. Additionally, we include continuous columns to deal with numeric selection predicates. Figure 4 illustrates this method with a fictive example.

In fact, dummy coding has a fundamental flaw: to support all of SQL, it requires vectors of infinite length. In consequence, we must limit its scope. One option is to represent only the queries in the log, as we did in Figure 4. Another is to specify a subset of SQL a priori. For example, we can restrict the encoding to Select-Project-Join queries with a limited number of components. Additionally, we can compress the resulting vectors with dimensionality reduction methods, such as factor analysis or autoencoders [5].

Discussion. Dummy coding is the naive approach. It is straightforward and lossless. It produces flat tables, which effectively make it possible to mine query logs with mainstream statistical tools. But we foresee that it will return huge, sparse vector spaces with complex queries. The subsequent vectors will be costly to store, to process, and statistical methods will be prey to overfitting (as per the curse of dimensionality [5]). Dimensionality reduction algorithms can help, but they are lossy, expensive, and they require careful tuning. Besides, binary variables are not real numbers, thus not all statistical methods can cope with them (for example, k-means is excluded). For all these reasons, we need alternative encoding schemes.

3.2 Dissimilarity-Based Mapping

We now present dissimilarity-based feature maps (DBFMs), which generalize of existing work on query log mining.

Method. To build a DBFM, we operate in three steps. First, we chose one or several pairwise dissimilarity measures from the literature. Second, we embed them into an encoding function. Thanks to this function, we can represent the log with a large matrix. In the last step, we compress it.

Defining the dissimilarity between two queries is subject to all the problems presented in Section 2.2. Currently, we know no perfect measure of dissimilarity. However, several authors have already proposed specific functions, in the context of neighborhood-based approaches. Chatzopoulou et al. [7] have reported a measure based on query results: two queries are similar if they involve the same tuples. Akbarnejad et al. [2] have used fragments of text. More recently, Nguyen et al. [17] have developed a method to exploit the results of queries without running them. In a recent paper, Aligon et al. review 14 of these functions [3]. Collectively, those cover a wide range of use cases. Our idea is to embed them in an encoding Φ.

For a given dissimilarity measure, the square matrix $\mathbf{D}$ represents the *dissimilarity matrix* of the query log. This matrix contains the pairwise dissimilarities between all the couples (Q_i, Q_j) in the log, as follows:

$$\mathbf{D} \equiv \begin{bmatrix} d(Q_1, Q_1) & d(Q_1, Q_2) & \ldots & d(Q_1, Q_N) \\ d(Q_2, Q_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ d(Q_N, Q_1) & \ldots & \ldots & d(Q_N, Q_N) \end{bmatrix} \quad (1)$$

It turns out that we can derive a trivial feature map from this representation: we map each query Q_i to the vector $\phi_i = [d(Q_i, Q_1), \ldots, d(Q_i, Q_N)]^\top$. In other words, we associate each query to its corresponding line in $\mathbf{D}$. Hence, DBFMs represent queries by their difference with regards to the other queries in the log. The resulting space is called *dissimilarity space*, and its theoretical properties were described by Pekalska and Duin [10]. Observe that this method lets us combine several dissimilarity measures: we simply concatenate the resulting dissimilarity matrices. To deal with the dimensions of the result, we apply dimensionality reduction. Specifically, we can use PCA, or we can cluster the columns and pick a few representative dimensions.

Discussion. The advantage of the DBFM method is its flexibility. In comparison with dummy coding, DBFMs can deal with complex queries. Also, they generate continuous variables, which involves a broader class of algorithms. However, these functions are lossy: the user must specify the properties of interest. Also, the compression step is costly and it requires tuning, as discussed in Section 3.1. Finally, DBFMs are by definition sensitive to the queries in the log. If those are similar to each other, then the columns of the dissimilarity matrix $\mathbf{D}$ will be highly correlated. Therefore this matrix will contain little information. The physical dimensionality of the dissimilarity space will be high, but its intrinsic dimensionality will be low. In conclusion, DBFMs appear as viable substitutes for dummy coding in cases where the log is small and the queries diverse. But we need more general methods for larger and sparser data sets.

Multidimensional Scaling. An alternative approach is Multidimensional Scaling [6]. This method takes the dissimilarity matrix $\mathbf{D}$ as input, and generates a vector space in which the pairwise distances between the objects are preserved. Multidimensional scaling is relevant, but it suffers from the exact same problems as DBFMs: it is costly, it requires tuning and it depends crucially on the queries in the log.

4. KERNEL FUNCTIONS

In the previous section, we presented two general classes of feature maps. We now discuss *implicit* alternatives: kernel approaches.

4.1 Introducing Kernel Functions

The aim of this section is to communicate the intuition behind kernels. We refer the reader to Bishop [5] for a more rigorous introduction.

In this paper, we mention a number of statistical methods applicable to vectors, such as regression, classification and clustering. In fact, we do not need all of algebra to perform them. We need only one fundamental operation: *the dot-product*. If we can compute the dot-product $\phi_i \cdot \phi_j$ between

two vectors ϕ_i and ϕ_j, then we can run linear regression, Support Vector Machines, K-means, PCA and many others. The process of rewriting a statistical method in terms of dot-products is known as *kernelization* [5].

At this point, computing the dot-product $\phi_i \cdot \phi_j$ is problematic because we need to compute the vectors $\phi_i = \Phi(Q_i)$ and $\phi_i = \Phi(Q_j)$. To do so, we need the mapping function Φ. Kernel functions let us bypass this operation. A kernel function $K(Q_i, Q_j)$ is analog to a dissimilarity measure: it has a low value if Q_i and Q_j are similar, and it has a high value otherwise. But kernels have a convenient mathematical property: for every such function K, there exists a feature map Φ such that:

$$K(Q_i, Q_j) = \Phi(Q_i) \cdot \Phi(Q_j) \quad (2)$$

In plain words, computing the similarity between two queries according to K is equivalent to mapping them to some feature space and applying the dot-product. Therefore, each kernel defines an *implicit* feature map. This property is powerful: we can manipulate SQL queries as if they lived in a vector space, but without actually materializing the space. In essence, kernel methods offer a middle way between neighborhood-based approaches and feature mapping.

4.2 Kernels for the Query Log

In the past, authors have successfully built kernel functions for complex objects, such as texts, DNA strings, images or even videos [11]. Our task is now to design a kernel function for SQL queries.

Dissimilarity-Based Kernels. Not all dissimilarity measures are kernel functions. To qualify, a measure must obey *Mercer's conditions* [5]. Those imply that the eigenvalues of the dissimilarity matrix are positive. We know no function that guarantees these conditions. However, authors have presented methods to turn arbitrary dissimilarity measures into kernels, such as *spectral shifting* or *spectral clipping* [8, 23]. These methods compute the spectrum of the dissimilarity matrix, and correct the eigenvalues to meet Mercer's conditions. In effect, they let us reuse the dissimilarity measures from the literature, similarly to DBFMs. But they are costly, i.e., cubic with the number of items. Also, it is not clear how to maintain their results as new queries come in.

Custom Kernels. An alternative approach is to engineer new kernels from scratch. Authors have developed such functions for graphs, sets, and even logic programs [11]. We could extend those to SQL queries. To tackle different use cases, we could generate several kernels. For example, we envision a function to describe the syntax of the queries, and another to describe their set properties. We could easily aggregate them, because the weighted sum of two kernels is itself a kernel. But we could also attempt to design a lossless solution. Indeed, kernels can encode infinite dimension spaces. The Gaussian dissimilarity is a popular illustration of this property [5]. Therefore, we do not exclude the existence of a "perfect" kernel function for SQL queries.

Discussion. Compared to feature maps, kernel methods have many advantages. They are possibly more space efficient, because they do not materialize the vectors. The underlying encoding Φ is robust: it does not involve arbitrary restrictions, and it is independent from the other queries in the log. Lastly, kernels bypass the costly compression operations of feature maps: the whole space is embedded in the dissimilarity function.
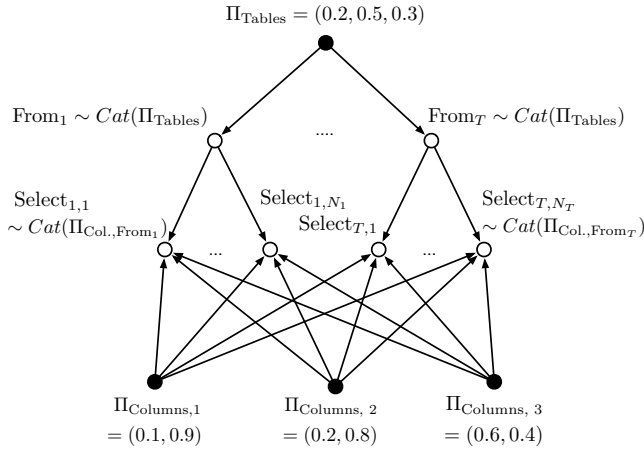
Figure 5: Simple Bayesian model to describe the distribution of SELECT-FROM queries, in a database made of three tables with two columns each. The full circles represent constants, the empty circles represent random variables.

Nevertheless, our quest for a transformation Φ does not stop here. Even if we had access to a perfect kernel, it is likely that its implicit feature space would remain theoretical: we would know that the inverse feature map Φ^{-1} exists, but we could not access it. Also, not all statistical methods were kernelized, hence kernel approaches are less general than explicit methods. Finally, their accuracy for SQL log mining remains to be studied. In particular, we must evaluate their sensitivity to the curse of dimensionality.

5. GRAPHICAL MODELS

So far, we have only considered methods related to vector spaces. But there exists an alternative conceptual framework for which many statistical methods were developed: *probabilistic graphical models*, also called *Bayesian networks*.

Presentation. The aim of graphical models is to decompose complex probability distributions into elementary, low-dimension components. Let us introduce an example. We wish to describe the distribution of all the SELECT-FROM queries from the log of a DBMS. In other words, we want to estimate the function $p_{SF} : \mathbb{Q}_{\texttt{SELECT-FROM}} \to [0, 1]$, which maps each query to its probability. Finding a closed mathematical form for this function is difficult: it involves complex operations, many parameters, and the number of these parameters is variable. Bayesian networks give us a mean to express p_{SF} in a graphical way. Figure 7 displays an example of model. This graph can be understood as an algorithm to generate new queries. We read it as follows:

- Set the constant vectors Π_{Tables}, $\Pi_{\text{Columns},1}$, $\Pi_{\text{Columns},2}$, and $\Pi_{\text{Columns},3}$. The vector Π_{Tables} describes the probability of occurrence of all the tables. The vectors $\Pi_{\text{Columns},t}$ describes the probability of occurrence of the columns in each table t.

- Chose T random tables $\{\text{From}_1, \ldots, \text{From}_T\}$, picking them randomly with probabilities Π_{Tables}

- For each table $t \in \{\text{From}_1, \ldots, \text{From}_T\}$, chose N_t random columns $\{\text{Select}_{t,1}, \ldots, \text{Select}_{t,N_t}\}$, picking randomly with probabilities $\Pi_{\text{Columns},t}$.
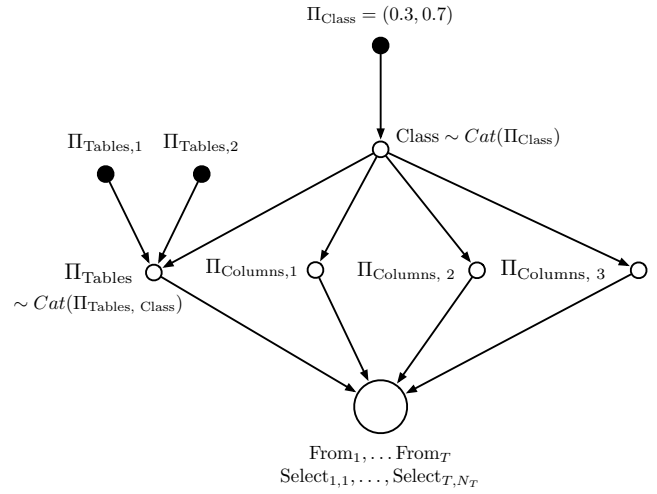

Figure 6: Extension of the SELECT-FROM model to support clusters. The latent variable $Class$ influences the distribution of the parameters Π_{Tables}, $\Pi_{\text{Columns},1}$, $\Pi_{\text{Columns},2}$, and $\Pi_{\text{Columns},3}$, which themselves influence the query, as illustrated in Figure 7. The distribution of the variables $\Pi_{\text{Columns},t}$ have form as that of Π_{Tables}, for $t \in \{1, 2, 3\}$.

Thus, the network describes a method to sample from the distribution p_{SF}. In fact, it also gives us a tractable way to compute the probability $p_{SF}(Q)$ for any given query Q. Here again, we refer readers to Bishop [5] for more details.

Extensions. With graphical models, we can compute complex probability functions and generate samples. Accordingly, if we had a complete model for SQL queries, we could detect "typical" or "outlying" queries, and we could generate realistic SQL statements. But we could also extend the model to cover more complex tasks. In the machine learning literature, authors have described dozens of statistical methods with Bayesian networks, including all those that interest us [5]. We could exploit them, by "plugging in" our own SQL network. As an illustration, we present an elementary clustering model in Figure 6. To build this model, we plugged our SELECT-FROM model into a mixture of distributions. In Section 6, we will introduce more general methods, to support all types of machine learning algorithms.

Discussion. Aside from dummy coding, Bayesian modeling is the only method which provides both the mapping Φ and its inverse Φ^{-1}. To obtain the image $\Phi(Q)$ of a given query Q, we instantiate the variables in the network. To obtain its inverse $\Phi^{-1}(Q)$, we execute the generative process. Additionally, graphical models are more flexible than vectors. For instance, they support variable numbers of parameters and recursivity. Besides, they are interpretable, and they have convenient statistical properties: among others, Bayesian methods natively incorporate regularization and adaptivity (cf. empirical Bayes [5]).

Yet, producing a complete Bayesian network for SQL queries remains a challenge. Also, adapting its parameters to the log may involve costly computation methods, such as Monte-Carlo simulations. Finally, as with feature maps and kernel functions, the empirical performance of this method remains to be studied. At this point, we do not know how accurate it is for log mining.
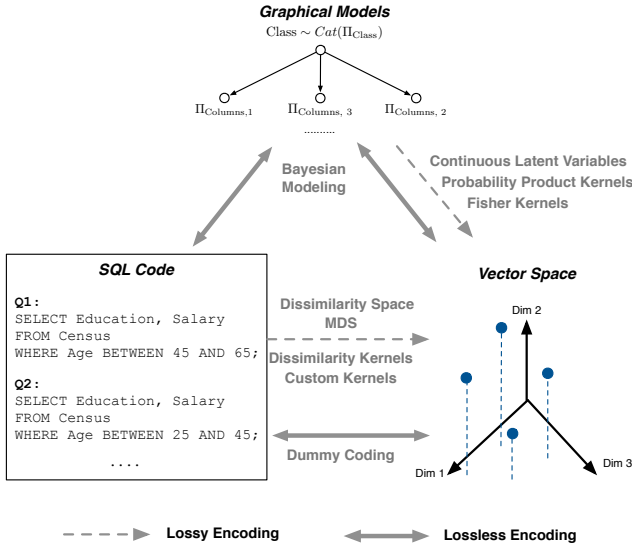
Figure 7: Summary of the methods discussed in this paper.

6. BRIDGING GRAPHICAL MODELS AND VECTOR SPACES

To close our presentation, we highlight a powerful feature of probabilistic graphical models: they can yield vector spaces, both implicitly and explicitly.

For a start we can embed graphical models into kernel functions. We know at least two methods to do so, *probability product kernels* and *Fisher kernels* [5]. Thanks to these solutions, we can benefit from both the generative features of graphical models and the libraries of kernel methods.

Furthermore, we conjecture that we can generate vectors directly from graphical models. In Figure 6, we show an example of latent variable model, where the discrete variable *Cluster* influences the distribution of the query's components. We could generalize this model to *continuous latent variables*. In this case, a fixed-size random vector would condition the distribution of the parameters Π_{Tables} and $\Pi_{Columns,t}$. The exact parametric form of the dependency has yet to be determined.

Finally, observe that we can operate in the opposite direction, and convert query-vectors ϕ_i into instances of a Bayesian network. Several methods exist to learn such models automatically from matrices. Nevertheless, their practical interest is limited: we have no guarantee that the generated graphical models will be complete, or interpretable. And they have no way to recover the information destroyed by the feature maps.

We summarize all the methods in this paper and their relationships in Figure 7. Bayesian models seem to offer the "best of all worlds": they are lossless, reversible, and they can yield vector spaces. For this reason, we chose to place them on top of our agenda. But we should not underestimate their competitors. Even dummy coding may come in handy, in conjunction with advanced compression algorithms such as autoencoders. Now, our task is to implement these ideas and conduct extensive benchmarks. Eventually, only practice and experiments will reveal which of these solutions truly fulfills our vision.

7. RELATED WORK

Several authors have developed methods to infer knowledge from the query log, either to improve the performance of the database or to help users write queries.

Application-Specific Methods. On the performance side, Ghosh et al. [12] associate each query from the log with a vector of predefined scores (e.g., number of tables mentioned, number of joins, presence of index) to recommend query plans. Aouiche and Darmont [4] mine the column names mentioned in the log to chose materialized views and indices. The optimizer LEO [21] monitors the execution of queries to predict cardinalities. On the user side, Agrawal et al. [1] have presented a method to recommend individual tuples. Yang et al. [24] mine the log for join predicates. SnipSuggest [15] suggests context-sensitive snippets. Zhang has developed an interface to explore the Sloan Digital Sky Survey database [26]. Giacometti et al. [13] present a method to detect unexpected patterns. Finally, Yao et al. [25] exploit cluster analysis to detect so-called query sessions.

Each of these papers use a different, task-specific encoding. Our ambition is to develop one framework to encompass all those cases.

Neighborhood-Based Methods. We discuss these methods in detail in our introduction. We generalize them with DBFMs, in Section 3.

Hierarchical Modelling of Queries. In Section 5, we present generative approaches. In fact, the early system PROMISE [18], based on Markov Models, is remarkably close to our vision. However, it targets very specific OLAP workloads. SnipSuggest also represents the queries with a tree [15], but the leaves represent fragments of plain text. Finally, the Oracle Workload Intelligence also uses a Bayesian model [22], but it operates at the user session level: each node represents a complete query.

Log Analysis in Information Retrieval. Authors have developed many methods to mine search engine query logs [20]. In principle, we could use those, exploiting natural language models such as n-grams or tf-idf. But these methods incur a major loss of information. First, they neglect the grammar of SQL. This is wasteful, because the language is simple, highly structured, and well-known. Second, they neglect the set relationships between the queries, such as inclusion, overlap or order. Those are crucial for many of the applications we target.

8. CONCLUSION

Too many methods to mine SQL query logs are isolated. They are isolated from each other: each paper uses its own conventions and its own algorithms. They are also isolated from the rest of machine learning research: they only exploit a narrow subset of its literature. In this paper, we presented three research directions to unify and broaden the scope of DBMS log mining. We purposely stepped out of specific applications, and presented frameworks to apply general statistical inference on SQL queries.

We now envision two lines of research. First, we will implement all the methods discussed in this paper, compare them, and understand which one performs best and why. Once we have solid tools to encode SQL queries, we will experiment with new machine learning algorithms. Given the recent advances in this field, with e.g. deep learning, we are convinced that this agenda holds a bright future.

9. ACKNOWLEDGMENTS

10. REFERENCES

[1] R. Agrawal, R. Rantzau, and E. Terzi. Context-sensitive ranking. In *Proc. SIGMOD*, pages 383–394, 2006.

[2] J. Akbarnejad, M. Eirinaki, S. Koshy, D. On, and N. Polyzotis. Sql querie recommendations: a query fragment-based approach. *Proc. VLDB*, 2010.

[3] J. Aligon, M. Golfarelli, P. Marcel, S. Rizzi, and E. Turricchia. Similarity measures for olap sessions. *Knowledge and Information Systems*, pages 463–489, 2014.

[4] K. Aouiche and J. Darmont. Data mining-based materialized view and index selection in data warehouses. *Journal of Intelligent Information Systems*, pages 65–93, 2009.

[5] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[6] I. Borg and P. J. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[7] G. Chatzopoulou, M. Eirinaki, and N. Polyzotis. Query recommendations for interactive database exploration. In *SSDBM*, 2009.

[8] Y. Chen, M. R. Gupta, and B. Recht. Learning kernels from indefinite similarities. In *Proc. ICML*, pages 145–152, 2009.

[9] C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. *Recommender Systems Handbook*, pages 107–144, 2011.

[10] R. P. Duin and E. Pekalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, pages 826–832, 2012.

[11] T. Gärtner. A survey of kernels for structured data. *SIGKDD Explorations*, pages 49–58, 2003.

[12] A. Ghosh, J. Parikh, V. S. Sengar, and J. R. Haritsa. Plan selection based on query clustering. In *Proc. VLDB*, pages 179–190, 2002.

[13] A. Giacometti, P. Marcel, E. Negre, and A. Soulet. Query recommendations for olap discovery driven analysis. In *Proc. DOLAP*, pages 81–88, 2009.

[14] F. Halim, S. Idreos, P. Karras, and R. H. Yap. Stochastic database cracking: Towards robust adaptive indexing in main-memory column-stores. *Proc. VLDB*, pages 502–513, 2012.

[15] N. Khoussainova, Y. Kwon, M. Balazinska, and D. Suciu. Snipsuggest: context-aware autocompletion for sql. *Proc. VLDB*, 2010.

[16] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proc. SIGKDD*, pages 426–434, 2008.

[17] H. V. Nguyen, K. Böhm, F. Becker, B. Goldman, G. Hinkel, and E. Müller. Identifying user interests within the data space - a case study with skyserver. In *Proc. EDBT*, pages 641–652, 2015.

[18] C. Sapia. Promise: Predicting query behavior to enable predictive caching strategies for olap systems. In *Proc. DaWaK*, pages 224–233, 2000.

[19] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of olap data cubes. *Proc. EDBT*, 1998.

[20] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, pages 1–174, 2010.

[21] M. Stillger, G. M. Lohman, V. Markl, and M. Kandil. Leo-db2's learning optimizer. In *Proc. VLDB*, pages 19–28, 2001.

[22] Q. T. Tran, K. Morfonios, and N. Polyzotis. Oracle workload intelligence. In *Proc. SIGMOD*, pages 1669–1681, 2015.

[23] G. Wu, E. Y. Chang, and Z. Zhang. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In *Proc. ICML*, 2005.

[24] X. Yang, C. M. Procopiuc, and D. Srivastava. Recommending join queries via query log analysis. In *Proc. ICDE*, pages 964–975. IEEE, 2009.

[25] Q. Yao, A. An, and X. Huang. Finding and analyzing database user sessions. In *Proc. DASFAA*, pages 851–862, 2005.

[26] J. Zhang. *Data Use and Access Behavior in eScience—Exploring data practices in the new data-intensive science paradigm*. PhD thesis, Drexel University, 2011.