Auto-Categorization Methods For Digital Archives

Nathaniel Payne
School of Library, Archives and Information Science
University of British Columbia
Vancouver, Canada
nathaniel.joseph.payne@gmail.com

Jason R. Baron
Of Counsel. Drinker Biddle & Reath LLP
Washington, D.C.
jason.baron@dbr.com

***Abstract*—Archivists and records managers would benefit from a greater understanding of the use and effectiveness of various machine learning methods, especially in the related context of electronic discovery. However, the binary classification methods used in advance search techniques in the e-discovery space may or may not prove efficacious where the information task involves sorting records into multiple categories. A survey of the landscape of machine learning methods reveals areas of potential weakness, which in turn serve as a starting point for future research in the computational archives space.**

Keywords—auto-categorization, auto-classification, binary classification, e-discovery, machine learning, coverage, detail

I. Introduction

Over the last decade, much research has been done and new promises made around the use and effectiveness of various machine learning methods in electronic discovery (e-discovery) and information governance [1, 2, 3, 4, 5]. The informal retrieval task involved in e-discovery is, however, at bottom an exercise in *binary* classification for relevance or privilege [6]. For retrieval tasks involving auto-classification for relevance, what has been learned in e-discovery arguably has direct applicability to the world of searching vast archival and records repositories in digital form [7]. For other sorts of tasks that records managers in particular are interested in, including the proper classification of records into granular record series set out in records schedules, the degree of difficulty for the learning algorithm involved is much higher, due to the presence of multiple classification possibilities [8, 9, 10, 11]. This paper aims to survey the landscape of machine learning methods used in e-discovery to assess their possible utility to archivists charged with carrying out various auto-classification tasks, of both the binary and the multi-categorical kind. Our aim is that the paper may serve as one point of departure for future research into the efficacy of these methods in the computational archives space.

E-discovery is especially important within the review phase of the legal process, when, during a typical review task, a "requesting party" makes a demand for relevant electronically stored information (ESI) pertaining to the claims or causes of action at issue in the lawsuit. The party does so by way of serving a "request for production" on its adversary, the "responding party," triggering a duty to conduct a review for the production of responsive, non-privileged documents identified as a result of a reasonable search [6]. Failure to effectively complete this task can result in case failures for both parties within the legal process, and in particular, sanctions on responding parties whose search efforts have fallen short in either failing to identify important sources of evidence, or in extreme cases, losing or destroying such evidence [12].

In tackling their information retrieval obligations, in large, complex cases involving up to terabytes of potentially responsive ESI, lawyers have in the main abandoned *sole* reliance on manual review processes, in favor of carrying out keyword or Boolean searches -- and increasingly machine learning methods -- before proceeding with quality checks using manual (eyes-on) review [13]. Regardless of methodology, the information retrieval task is one involving classification of documents either as "relevant" or "non-relevant" (R/NR), with relevant documents parsed further into categories of "privileged" and "not privileged" (P/NP). Where machine learning methods are used, relevance rankings result in the generation of a rank-ordered listing of most relevant (0.999) to least relevant (0.001) with a given universe of documents (U), where U is typically greater than 50,000 and may approach or exceed 1 billion.

In contrast, consider the state of recordkeeping today in public and private realms. One estimate of the complexity of records schedules found that the U.S. Department of Defense has created on the order of 17,000 separate records schedules for use by staff [14]. From the authors' experience, a typical corporate records schedule, created in the era of paper – and presumptively applied to the world of ESI in the firm's custody and control -- sorts records into several hundred separate categories, either by function or by organizational component. These in turn typically involve 25 to 50 unique records retention periods, which can include conditional, event-based triggers that assume human intervention at some point in the future [10, 15, 16].

More recently, there has been movement towards creating "big bucket" approaches to simplify recordkeeping burdens on staff, i.e., merging traditional record series into far fewer categories, coupled with retaining documents based on the longest retention period found in each bucket [11]. In the public sector, a grand experiment is underway with much of the U.S. federal government's voluntary adoption of the "Capstone" approach to e-mail preservation, a policy which collapses the classification of e-mail essentially into a binary exercise based on the role individual content creators play in the organization. That is, under Capstone, either email records are deemed to be permanent or temporary, depending on whether particular

users are designated as "Capstone" account holders, which in most cases involve those individuals occupying the most senior positions at an agency [17, 18]. Capstone, however, remains the exception to the status quo in records management, which involves dividing the world into hundreds or thousands of different records series and/or retention periods, rather than simply two.

Focusing solely on the "search" task for relevant documents, the methods referred to in the legal marketplace as "predictive coding" or "technology assisted review" tools, have achieved strong results in matching or improving on measures of "recall" (the percentage of responsive documents retrieved out of the total number of responsive documents in U) when employed to carry out especially the task of distinguishing R from NR documents [2, 12, 19]. Such advanced search techniques recently have been "blessed" by numerous judicial authorities in reported court decisions [20].

In this paper, we provide a taxonomy of sorts both for (i) exploring various types of auto-categorization algorithms that exist within the machine learning context, and (ii) reviewing the challenges that the various methods entail when it comes to categorization in the archives and records management environment. In particular, we are interested in how fast the methods degrade, especially when the number of categories grows significantly. We will discuss some of the considerations that must be addressed within the context of each method, and general steps that must be followed when analyzing an underlying method in order to understand its suitability for the auto-classification task for archives and records management. We will also seek to present a framework that can be used when evaluating and building auto-categorization systems, in connection with articulating the benefits of auto-categorization as part of further research. Through this work, we hope to develop a framework which can be used as a basis for discussion and thought in various areas of practice and industry contexts, including the information governance realm.

II. Understanding the Argument for Auto-Categorization

In nearly every community, including the archival science community, timely access to information is a critical concern [21, 22, 23]. From email to collections of documents, archivists and records managers are continually faced with the challenge of not only storing information received in digital and non-digital forms, but also building systems to categorize and manage the growing volumes of information that exist [24]. Organizations are faced today with the challenge of reacting to urgent access demands for records and ESI, not only as part of litigation but also in the form of audits and investigations [25]. Tasks like document categorization, which previously were done "offline" in a sequential process, are no longer limited to being solely offline requests that precede traditional archiving activities [10, 26, 27]. Rather, categorization methods need to adapt so that they could be applied in real time to enable requests and documents to be processed [28]. This need is further complicated by the multi-language environmental support challenges which many organizations face [29].

In previous times, the manual capturing, indexing, and categorization of records worked well. This categorization was often done by records managers, archivists, and legal professionals who were able to "read" documents as they were received, categorize those documents by distinct attributes, and index document attributes accordingly [10, 30]. They were also able to manually capture detailed information from various records and documents which could be used to build indexes which supported the efficient storage and retrieval of information and which often leveraged the power of created meta-data [31]. Unfortunately, while this method worked well for cases where one archivist had a few hundred documents to process, the situation has changed dramatically today, with incoming information volumes rendering manual methods infeasible [3, 28, 32, 33, 34].

In the era of Big Data, organizations simply are ingesting orders of magnitude more information from both traditional and non-traditional sources (e.g., the Internet of Things), in many cases leaving unattended traditional records management controls over their ESI [35]. Yet we know from the e-discovery context that there is an enormous cost to devoting time to searching large collections of uncategorized ESI [12] (RAND study finding that 25% of total cost of litigation or more is due to complying with discovery demands). For the reasons stated above, the legal profession has adopted more advanced methods for search purposes in light of the new reality of the volume and complexity of information [36, 37, 84].

Accepting that categorization is an important component of information retrieval, and assuming for our purposes here that large volumes of ESI coupled with a large number of record categories make the categorization activity more difficult to complete, one then turns to the question of automating categorization [38]. Today, within the data storage realm, the majority of information is organized into index-sequential forms [39]. In order to put information into this format, categorization methods must be developed which preserve as much context as possible, with index attributes being updatable, reflecting the content within the records. In an optimal world, index attributes for any given record would reflect all possible content-related combinations. Unfortunately, without the use of more abstract approaches, content, which is often not representable within an index, cannot adequately be captured, logically structured within databases, or mapped into hierarchies in a timely feasible manner [15, 40].

As a result, new automated categorization tools are needed which can overcome the limitations of traditional categorization and maintain the strength of the archival bond, enabling the automated identification and storage of document or record attributes, as well as the comparison of record attributes using machines [41, 42, 43, 44]. Such methods, when properly created, offer many benefits to the archival science community, as well as the communities that leverage this technology. Arguably, the most important of these benefits include replication, reliability, stability from an operational perspective, language independence, and the ability to enable fault tolerant information processing which enables multi-dimensional categorization [45, 46]. Referred to from this point on as "auto-categorization," these computationally based approaches, which often leverage machine learning, are designed to offer significant advantages, including:

- The ability to automatically index a single record, or a group of records, so as to optimize the proportion of responsive documents within a single pass and effectively deal with large numbers of categories. Such a process would reduce the need for manual intervention, reduce error, improve reproducibility, and can have significant time, cost, and resource savings.

- The ability to enable mass automated updating of document relationships. This avoids the use of manual linking tasks which preserve data set integrity, and can enable the creation of a linked set of documents which can be searched efficiently.

- The ability to develop consistent natural language queries which can improve user access to data and information repositories post-process. Access is an important concern within the archival science community. Typical discovery stage queries are very complex and the development of a streamlined approach to querying new information can be significant.

- The ability to handle multiple languages effectively by creating a consistent standard for language independence.

- The ability to be accessed platform independently by diverse users for diverse purposes.

Importantly, as noted, the use of auto-categorization methods which leverage machine intelligence is designed to significantly reduce the costs that organizations carry relative to their information management budgets. These cost improvements, due often to significant improvements in efficiency, reduce the overall cost of storage and archiving by reducing the time required for such tasks as document preparation, data capture, and training times.

III. REVIEWING METHODS AND THEIR CHALLENGES

Accepting the potential benefits that auto-categorization methods carry, in this section we focus on reviewing the major classes of auto-categorization. These methods, which are generally machine based, are designed to meet two key principles, *coverage* and *detail*. Coverage, in this sense, ensures that every document or record that the system interacts with is categorized. Detail, on the other hand, is the principle that seeks to extract and capture every available piece of information from a document, record, or corpus. These principles form one of the foundations for evaluating the accuracy and efficacy of auto-categorization methods in the archival environment, with the optimal method providing wide, reliable, and repeatable coverage with low variance and low bias after all possible information is extracted and captured from the information source [47, 48, 49, 50, 51, 52, 83].

While coverage and detail are important, our review and critique of these methods, especially with knowledge gained from the e-discovery realm, will focus on their ability to deal with data that is of a high dimensional nature, i.e., where the number of categories is large. As we will see, significant work and opportunities are needed to build accurate classification systems when the data dimensions are high. To be clear, if, informally, we let q denote the dimension of what is 'unknown' and let m denote the cardinality of what is 'known', then traditional theory, and most practice, has until recently been largely limited to the 'small q, large m' scenario. Let us consider m as corresponding to the number of experimental units on which data are available. On the other hand, q, can be considered a measure of complexity of the model to be fitted to the data. This measurement is often determined by the dimension of the data as given by the number of items (variables) recorded for each experimental unit (or, in our case categories) [53]. Over the last 20 years, the practical environment has changed dramatically, with the spectacular evolution of data acquisition technologies and computing facilities. That said, as multiple researchers have noted, gaps remain, especially as it relates to model theory, with significant gaps in two key scenarios from a theoretical and practical approach: 1) 'large q, small m' 2) 'large q, large m'. Both of these theories operate in the situation where q is growing faster than m [53]. What's more, with the acknowledgement historically that the best outcomes seem to be achieve in analysis when the number of data points m exceeds the number of parameters q to be estimated by some solid margin, maintaining a $m/q \geqslant 5$ ratio has been cited as a critical rule of thumb that is not always possible [54, 55].

A. Rule-Based Classification - Decision Rules

The basic foundation and most simple approach for auto-categorization focuses on decision rule or rule-based systems. Decision rule-based systems use rule-based

inference to classify documents to their annotated categories [56]. Algorithms for this type of auto-classification usually construct a rule set that describes the profile for each category. Specific rules are then constructed in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category, and the conclusion portion is represented with the category's name or another rule to be tested. The rule set for a particular category is then constructed by combining every separate rule from the same category with logical operator, typically use "and" and "or". During the classification tasks, not necessarily every rule in the rule set needs to be satisfied [57].

In general, the rule-based approach provides significant coverage but lacks detail. The primary advantage of this approach is that the implementation of a decision rule method for classification tasks requires the construction of a local dictionary for each individual category [56]. Local dictionaries often aid in the method's ability to distinguish the meaning of a particular word for different categories. That said, local dictionaries also render the method less generalizable. Moreover, with this approach, it can be difficult to assign a document to a category exclusively -- especially if the rules from different rule sets are dually applicable to each other. This can create significant problems within records management systems, including requiring records managers and archivists to spend significant time updating decision rule methods and creating and maintaining rule sets. This method also fails when the number of distinguishing features or attributes of a record or document is large [57].

Overall, when looking at the use of rule based classification methods from an archival perspective, we see a number of issues and opportunities for research especially at high dimensions. Importantly, association rule algorithms can be formulated to look for sequential patterns which meet the need of users conducting searches in large collections. That said, at high dimensions with large numbers of categories, association rules do not show reasonable patterns and are not able to do variable or categorical selection at scale. The association rule method also fails to produce useful results if the information it is built around does not provide clear support or is easily distinguishable. As a result, as the number of information categories within a repository grows, and the complexity of the queries needed to be used expands, rule based methods typically fall short and are not adequate. Thus, tools that leverage this approach are not currently adequate and need further review, especially when the number of dimensions within the data is high [58].

B. Fuzzy Correlation

Fuzzy correlation or phrase analysis is a a technique that can be used to allow a system to either precisely or fuzzily match single words or multi-word phrases together. It is especially useful when rule-based classification is challenging due to the attribute size growing, and the technique can help be a bridge between rule-based matching and more robust machine leaning based tools. The advantage of fuzzy correlation is that it can often deal with fuzzy information or incomplete data, allowing for document classification [59]. In recent years, fuzzy classification has been used to augment traditional methods of linguistic analysis, including neural networks and similarity based methods. For example, in Wang & Huei-Min [60] the authors explore the challenges of multi-class text categorization using a one-against-one fuzzy support vector machine with Reuters news as the example data set. Their work shows better results using a one-against-one fuzzy support vector machine as a new technique when compared with a traditional support vector machine. Cohen & Singer [61] presented the improvement of a decision rule and the design of a new algorithm of f-k-NN (fuzzy k-NN) to improve categorization performance when the class distribution is uneven, and showed that the new method is more effective. The research recently has shown great promise in using fuzzy rules and sets to improve classification accuracy, by incorporating the fuzzy correlation or fuzzy logic with the machine learning algorithm and the feature selection methods to improve the classification process [57].

C. Vector Space Methods

While rule based classification and fuzzy correlation offer two interesting solutions to the problem at hand, the following methods leverage the power of machine learning to improve both the coverage and detail of the methods in question. From a machine learning approach, a document can be classified using one of three methods: unsupervised, supervised and semi-supervised learning [19, 62]. Supervised learning is the machine learning task of inferring a function from labeled training data [62]. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples [62]. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. Classification, or auto-classification, is a supervised learning approach which uses training data to produce an inferred function that is used for classification and mapping of new examples [62].

The first machine learning methods that we will consider are vector space methods. Rocchio's Algorithm [63] is a vector space method for document routing or filtering in informational retrieval, where the algorithm builds a prototype vector for each class using a training set

of documents (i.e. the average vector over all training document vectors that belong to a document class), and calculates similarity between test documents and each of the prototype vectors which assign test documents to the class with maximum similarity. This algorithm is easy to implement, efficient in computation, is a fast learner and employs relevance feedback. Unfortunately, the algorithm has low classification accuracy and is low in detail [57].

Overall, when looking at the use of vector space methods from the archival perspective, we see a number of issues and opportunities for research especially at high dimensions. While vector space methods are easy to implement and efficient in computation, they can have low classification accuracy especially as the number of dimensions grows [64]. Within text, vector space methods often have difficulty dealing with lexical ambiguity and variability, and do not deal well with feature sparseness. Thus, as the number of categories grows within a set, vector space methods will often struggle if the categories present occur at infrequent intervals -- assigning an irrelevant score to texts which may be relevant simply due to sparseness. This can reduce the quality of output, and lead one to advocate to reduce the number of categories that exist within the records system if vector space methods are to be used to analyze the system.

A particular vector space application is the popular use of support vector machines (SVMs) to automatically and discriminatively classify documents and records. SVM's have a good reputation from an efficacy approach and are thus used in many auto-classification applications. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory [62, 65]. The idea of this principle is to find a hypothesis to guarantee the lowest true error. From an action perspective, the SVM needs both positive and negative training sets (here, both R and NR documents), a requirement which is uncommon for other classification methods. These positive and negative training sets are needed for the SVM to seek the decision surface that best separates the positive from the negative data in the n-dimensional space, also known as the hyper plane.

The SVM classification method is often highly regarded due to its classification effectiveness on various tests [66, 67, 68]. Furthermore, it can handle documents with high-dimensional input space, and culls out most of the irrelevant features. However, the major drawback of SVMs is their relatively complex training and categorizing algorithms, and also the resource-intensive time and memory consumptions during the training and classifying stages. Besides, confusion can occur during classification tasks due to documents notated in several categories, because similarity is typically calculated individually for each category [66].

Overall, when looking at the use of support vector machines from an archival perspective, we see a number of issues and opportunities for research especially at high dimensions. First, these methods produce very accurate classifiers on good data, and are robust to noise. That said, this suggests that the organization must bear some responsibility for ensuring high data quality, which here means especially from a categorization perspective. And, while this method is popular in text classification problems where very high-dimensional spaces are the norm, the method is memory-intensive [69]. What's more, SVM as a technique is a binary classifier. This presents significant challenges if the number of categories that a document has to be classified into is greater than two. In such a case, multi-class classification can be used, where pair-wise classifications can run one against each other. This is, however, computationally expensive, especially when the number of documents is large, and is an area where future research is needed to support expanded use outside the e-discovery domain.

D. Prototype Methods & Nearest Neighbor / Similarity Methods

The next method to consider includes the similarity class of methods, for which k-nearest neighbor algorithm is the most well-known (k-NN). k-NN is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents [70]. This method is an instant-based learning algorithm that categorizes objects based on closest feature space in the training set [71]. The training sets are mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Euclidean Distance is typically used in computing the distance between the vectors. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document [70]. This method is effective, non-parametric and easy to implement [57]. That said, when compared with fuzzy matching based approaches, the classification time for similarity methods can be much longer, with software applications using this method having difficulty finding optimal values that are reliable especially if the data is high dimensional data with many attributes [57].

Overall, when looking at the use of nearest neighbor or similarity methods, we see a number of issues and opportunities for research especially at high dimensions. In general, the algorithm for the nearest neighbor methods is shown below [69].

K ← number of nearest neighbors
    For each object X in the test set do

calculate the distance D(X,Y) between X and every object Y in the training set neighborhood ← the k neighbors in the training set closest to X

X.class ← SelectClass (neighborhood)
End for

The algorithm is easy to understand with a training process that is fast, robust to noisy data, is particularly well suited for multimodal classes, and has other advantages. On the other hand, within the higher dimensional environment, the algorithm is very sensitive to the local structure of the data, can have significant memory limitations as implemented in its most basic form, runs slowly which can be a significant problem as the number of documents grows, and is sensitive to initialization. Also, with the requirement that we need to specify the number of clusters in advance, similarity methods like nearest neighbor may be completely infeasible if the data sources are dark in nature, are not well understood, have noisy data, or unusual features.

E. Tree Based Methods

While similarity measures improve in detail over vector space methods, they are often most similarly compared with tree based methods which seek to support wide coverage in classification tasks but lack detail. The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure [62]. In a decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories. Using this logic, organized decision trees can easily classify a document by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf [57, 62].

The decision tree classification method has many advantages which make it widely used in many applications. The main advantage of a decision tree is its simplicity in understanding and interpreting, even for non-expert users. What's more, the explanation of a given result can be easily replicated by using simple mathematical algorithms, and so can provide a consolidated view of the classification logic. While this is positive, the major risk of implementing a decision tree is that a selected model may overfit the training data [72]. Some of these issues are addressed by using ensembles or random forests. Random forests [73] are a substantial modification of bagging that build a large collection of de-correlated trees, and then averages them [57].

Overall, when looking at the use of decision tree based methods, we see a number of issues and opportunities for research especially at high dimensions within the archival realm. While decision trees take less memory time to execute and have a short searching time [69], the reliability of the information within a decision tree, especially at high dimensions, depends on the tree accessing precise internal and external information at the onset. If this is not possible, i.e., if the input data is messy, or the input data from the repository of legal documents changes, this can cause significant changes in the outcome of the tree and the accuracy of the relevance or non-relevance classification. Moreover, the excluding of duplicate information, change in the sequence of analysis of the documents, and/or change in the number of categories (as well as the categories' relationships to each other), can result in either empty branches within a tree, or insignificant branches, or overfitting, or other challenges. Adding more categories can create problems at high dimensions, although new techniques have been developed which seek to leverage the power of ensembles. That said, more work is needed here for decision trees to be a viable alternative for archival tasks. As with the other methods recounted here, organizations are recommended to focus on streamlining and standardizing their document categorization schemes to ensure efficient information retrieval and categorization.

F. Neural Networks & Content Based Matching

The next auto-categorization methods relate to the use of artificial neural networks (ANN's), which continue to get significant press due to the focus on "Deep Learning" [82]. Artificial neural networks are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional architectures [74]. These elements, namely artificial neurons, are interconnected into groups using a mathematical model for information processing based on a connectionist approach to computation.

Different types of neural network approaches have been implemented for document classification tasks. Some of the research uses the single-layer perceptron, which contains only an input layer and an output layer due to its simplicity of implementing. The main advantage of the implementation of an artificial neural network in classification tasks is its ability in handling documents with high-dimensional features and/ornoisy and contradictory data. Furthermore, linear speed up in the matching process with respect of the large number of computational elements is provided by a computing architecture which is inherently parallel, where each element can compare its input value against the value of stored cases independently from others. The drawback of artificial neural networks is their high computing cost, which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to understand for average users. This may negatively influence the acceptance of these methods [57].

Overall, when looking at the use of neural networks from an archival perspective, we see a number of issues

and opportunities for research especially at high dimensions. In general, neural networks are hard to tune. As a number of authors have pointed out, at high dimensions, neural networks that use probability estimation for posterior can lack rigor in producing reliable results [75]. What's more, the back propagation algorithm for neural networks often struggles when the number of documents is large [76], with maximum likelihood classification based approaches -- the most common case -- also struggling [77]. These issues create significant opportunities for further research.

IV. DESIGNING AN AUTO-CLASSIFICATION SYSTEM – STEPS AND ISSUES

While the aforementioned classes of auto-categorization methods provide a great starting point, a closer analysis of the methods themselves and their use in both commercial software and viability in the archival domain necessitates the development of a standardized design template that can be used to both (i) evaluate any new method that is considered, while (ii) supporting the design of applications and recordkeeping systems which a search process would run against. Below are some key steps and considerations that we recommend be considered when analyzing any auto-categorization system, when analyzing the record systems on which the technology will run against, and when deciding on the number of categories that exists within the system itself.

- **Step 1:** Understand the training requirements. Supervised learning methods require training data. This data must be carefully considered particularly in the context of the bias-variance trade off (see IV.A, below) in which bias can impact the outcome of a model. Thus, for any auto-categorization method, one must determine the type of training data to be used, volume, its completeness, accuracy, repeatability, and the population from which that training volume is generated. This training set should be representative of the solution space which the algorithm or method will be used in.

- **Step 2:** Understand how input features will be represented within the learned function. Within the coverage-detail paradigm, detail assumes that a large volume of features can be collected accurately. For every method, one must review the state of input objects, how they are stored, whether and how they are transformed into a feature vector, and how descriptive those feature vectors are. To avoid the curse of dimensionality the number of features should be reduced. That said, enough features must be captured to enable adequate coverage during the analysis step.

- **Step 3:** Determine the structure of the learned function and corresponding learning algorithm. This step is important, and requires the users to understand the situations in which the various methods will fail, the limitations or biases of the various methods, as well as the optimal data structure needed to prepare data for use. Whether one chooses a rule based approach or something more challenging, the proper consideration of the algorithm prior to the analysis and testing phase must be completed, especially if results are to be repeated.

Once these three items have been considered, normal testing and implementation can occur. This includes splitting the data into a test, training, and validation step, completing the tactical design, running the learning algorithm on a sample of training data with various parameters tuned in a logical manner, and then completing the parameter optimization against a subset of the training data, or validation set, along with cross-validation. After this is complete, the measurement of the accuracy of the new learned function on a set of data, record, or document set that is independent from the training set should be undertaken.

G. Bias-Variance Trade-off

While the design of the auto-categorization system needs to be considered, there are a number of critical theoretical considerations that need to be considered especially when it comes to standard setting and the empirical performance of any method. The first issue relates to the bias-variance trade-off, i.e.,the trade-off between the bias and variance within a method. For all the supervised learning methods above, one must take care to ensure that the algorithm itself is not biased for a particular input, especially if, when trained on various data sets, it is systematically incorrect when predicting the correct output. If, on the other hand, a supervised learning algorithm predicts different output when trained on different training sets, then this is a concern. Thus, for any method that is considered, one must take care to ensure that the method created or selected has low bias -- enabling it to fit the data well, but is not too flexible so that it fits every training set differently [78, 79, 80, 81].

H. Functional Complexity & Training Data Volumes

The second critical issue that must be addressed within the context of any empirical method relates to the functional complexity and volume of training data. Based on empirical tests, if the true function is simple, then an "inflexible" learning algorithm with high bias and low variance will be able to learn it from a small amount of data. But if the true function is highly complex (e.g., because it involves complex interactions among many different input features and behaves differently in different parts of the input space), then the function will only be learnable from a very large amount of training data, using a "flexible" learning algorithm with low bias and high variance. This is something that must be considered especially when one is considering a method.

I. Dimensionality

A third important issue relates to the dimensionality of the input space. As noted before, if the input feature vectors have very high dimension, theory tells us that the learning problem can be difficult even if the true function only depends on a small number of those features. This is because the many "extra" dimensions can confuse the learning algorithm and cause it to have high variance. Hence, high input dimensionality typically requires tuning the classifier to have low variance and high bias. If this is the case, one may want to seek to understand or implement feature selection or use a larger dimensionality reduction strategy, reduce the number of dimensions that any algorithm considers, or reduce the number of categories present within an organization's record keeping system.

J. Other Issues Including Noise, Heterogeneity, Redundancy & Interaction

Another critical issue that must be addressed is the degree of noise in the response variables. If the response variables are frequently incorrect because of human or sensor errors, then one must take care to ensure the software used does not attempt to find a function that exactly matches the training data. This is because this will lead to overfitting, which will render the method useless. Moreover, heterogeneity of the data must also be considered, particularly of the training data. This is because many algorithms which form the basis of the previously mentioned methods require that their data be scaled prior to use (support vector machines, linear regression, logistic regression, neural networks, and many similarity measures). Moreover, these methods often use means which are highly sensitive to heterogeneity. Fortunately, tree models are able to handle heterogeneous models despite their other short falls.

Two other remaining issues relate to the redundancy of the data and the presence of interactions. In general, if the input features contain redundant information, some of the underlying methods above will perform poorly due to redundant data. This issue can be fixed by regularization. At the same time, the presence of interactions and non-linearity can cause issues that continue to be addressed by ongoing research.

K. The Need For Quality Standards

While the open questions around the various methods referenced here create research opportunities, the lack of quality standards that govern the application of new categorization methods is potentially a pressing concern. As new tools have evolved which can fill the void where traditional categorization is challenged, this need also gives rise to the need for new quality standards to be developed which can be used to evaluate and govern auto-categorization methods based on artificial intelligent software applications responsible for completing categorization, data capture, and indexing tasks. These standards, which do not exist, are key, as they will allow organizations in similar professions to collaborate and share information both internationally and inter-departmentally, using the same classification standards which will enable unified storage and access. Such standards will ensure that the rules and regulations applied to various document collections and records lead to the same categorization result independent of the specific operating conditions that an organization faces.

V. CONCLUSIONS

As we hope to have shown in this paper, over the last decade, much research has been done and new promises made around the use of and effectiveness of various auto-categorization methods for recordkeeping and document management. That said, as has been also shown, there are many open issues to be considered when selecting and using an auto-categorization method in the archival environment, as well as much work to be done to understand the situations that will suit each method optimally. This work is critically important in light of the vastness and complexity of the new forms of records which archival science researchers, records managers, and organizations in every discipline all face. Moreover, focused work must be undertaken to standardize the measurement of effectiveness of the various methods, especially considering the variability in which the methods perform from a situation perspective.

As one looks forward, working to proactively address the gaps in auto-categorization that exist, there are a number of areas of focus that immediately come to mind that can be addressed. Firstly, more work needs to be done to improve the various methods' ability to deal with *detail*. All the methods above have challenges in extracting and capturing every available piece of information from a document, record, or corpus. To date, these issues have not been sufficiently addressed. Similarly, text representation continues to be an issue that needs to be explored and discussed especially in the legal and financial context. Most of the research today that has been done gives the statistical of syntactic solution for the text representation. However, the representation model for the optimal methods going forward depends on the applicable informational context. Concept based or semantic representations of documents require more attention.

Moreover, more work needs to be done to understand how to deal with low quality data sources. Too often in practice, irrelevant and redundant features of data increase the cost of training and either bias results from analysis providing misleading categorizations, or break altogether. This must be addressed in a tactical way. As this is being addressed, one then needs to build a proper standard on which to compare the different methods. A survey of various methods that exist shows over a dozen various

algorithmic measurements in use today validating various auto-classification research. This issue must be addressed if researchers and practitioners can hope to accurately and openly compare software and research outcomes in practice, rather than relying on highly controlled experimental environments to validate a new method. This will also prevent background conditions from influencing results as well as various governance standards being set in error.

Finally, important technical work needs to be completed that will assist in our making progress in the advancement of auto-categorization work. This includes the identification and exploration of new feature selection methods, with the goal of improving the classification process, reducing the training and testing time of classifiers, while also improving classification accuracy, precision, and recall. The use of semantics and ontology also needs to be considered, especially when going through the standard setting process.

Even within the expanding realm of what artificial intelligence can accomplish, as of this date the newer forms of auto-categorization methods seem to continue to fall short, failing to deliver against the promise of a robust, standards-based auto-categorization method that would succeed in environments more complex than those involving binary classification. In this respect, research in computational archives holds the promise of advancing the state of knowledge already gained during the past decade in the field of e-discovery.

REFERENCES

[1] TREC Legal Track Overview Papers (2006-2010), https://trec-legal.umiacs.umd.edu/.

[2] Grossman, M.R. and Cormack, G.V. (2013). "The Grossman-Cormack glossary of technology-assisted review with Foreword by John M. Facciola, U.S. Magistrate Judge," Federal Courts Law Review, 7 (1):1–34.

[3] Sedona Conference, The (2013). Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, Sedona Conference Journal 15:217.

[4] Borden, B. and Baron, J.R. (2014). "Finding The Signal in the Noise: Information Governance, Analytics, and The Future of Legal Practice," 20 Richmond J. of Law and Technology Art. 7, http://jolt.richmond.edu/v20i2/article7.pdf.

[5] Baron, J.R., Losey, R. Berman, M., eds. (2016). PERSPECTIVES ON PREDICTIVE CODING AND OTHER ADVANCED SEARCH TECHNIQUES FOR THE LEGAL PRACTITIONER (ABA) ("PERSPECTIVES ON PREDICTIVE CODING").

[6] Oard, D. and Webber, W. (2013). "Information Retrieval for E-Discovery," Foundations and Trends in Information Retrieval, 7 (2-3), pp. 99–237.

[7] Underwood, W. et al. (2010). "Advanced Language Processing Technology Applied To Digital Records: Annual Status Report," Georgia Tech Research Institute. Technical Report, ITTL/CSITD 10-04.

[8] Man, E. (2010). "A Functional Approach to Appraisal and Records Scheduling," Records Management Journal 20(1), pp. 104-116.

[9] McDonald, J. and Léveillé, V. (2014). "Whither the retention schedule in the era of big data and open data?," Records Management Journal, 24(2), pp. 99–121.

[10] Franks, P.C. (2013). "Records and information management," American Library Association, Chap. 4, pp. 85-111.

[11] Fischer, L. (2006). "Condition Critical: Developing Records Retention Schedules," Information Management Journal 40 (1), pp.26-34.

[12] Pace, N., and Zakaras, L. (2012). "Where The Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery," RAND Institute for Civil Justice.

[13] Drynan, T.D.; Baron, J.R.. "The Road to Predictive Coding: Limitations on the Defensibility of Manual and Keyword Searching," chap. 1 in Baron, et al., eds., PERSPECTIVES ON PREDICTIVE CODING.

[14] Baron, J.R. (2003). "The PROFS Decade: NARA, Email, and the Courts," chap. 6 in Ambacher, B., ed., THIRTY YEARS OF ELECTRONIC RECORDS (Scarecrow Press, Lanham, MD).

[15] Shepherd, E.; Yeo, G. (2003). "Chapter 4: Creating and Capturing Records," in MANAGING RECORDS: A HANDBOOK OF PRINCIPLES AND PRACTICE, pp. 101-145. London, UK: Facet.

[16] Yakel, E. (1996). "The Way Things Work: Procedures, Processes, and Institutional Records," American Archivist 59(4), pp. 454-464.

[17] Baron, J.R. (2016)."Opening up Dark Archives Through the Use of Analytics to Identify Sensitive Content," 1st Computational Archival Workshop, IEEE Big Data 2016 IEEE Workshop, Washington D.C.

[18] Baron, J.R., Payne, N. (2017). "Dark Archives and eDemocracy: Strategies for Overcoming Access Barriers to the Public Record Archives of the Future," 7th International Conference for E-Democracy and Open Government, Krems, Austria.

[19] Grossman, M.R., Cormack, G.V. (2016). "A Tour of TAR," chap. 3 in Baron et al., eds., PERSPECTIVES ON PREDICTING CODING.

[20] Shelton, A.L. and Berman, M.D., (2016). "The Emerging Acceptance of Technology-Assisted Review in Civil Litigation, chap. 2 in Baron et al., eds., PERSPECTIVES ON PREDICTIVE CODING.

[21] Trace, C.B. (2002). "What is Recorded is Never Simply 'What Happened': Record Keeping in Modern Organizational Culture." Archival Science 2 (1-2), pp.137-59.

[22] Tough, A. (2011). "Accountability, Open Government and Record Keeping: Time to Think Again?," Records Management Journal 21(3), pp. 225–236.

[23] Andolsen, A. (2006). "Will Your Records Be There When You Need Them?," The Information Management Journal 40(3) (May/June), pp. 56-61.

[24] Maguire, R. (2005.) "Lessons Learned from Implementing an Electronic Records Management System," Records Management Journal 15(3), pp.150-157.

[25] McLeod, J. and Childs, S. (2013). "A Strategic Approach to Making Sense of the 'Wicked' Problem of ERM," Records Management Journal 23(2), pp. 104-135.

[26] Bak, G. (2012). "Continuous Classification: Capturing Dynamic Relationships among Information Resources," Archival Science 12 (3), pp/ 287-318.

[27] Duranti, L.; Thibodeau. K. (2006). The concept of record in interactive, experiential and dynamic environments: the view of interpares," Archival Science 6 (1), pp.13-68.

[28] Foscarini, F. (2009). "Chapter. 2: Literature Review," in Functional-based Records Classification Systems. An Exploratory Study of Records Management Practices in Central Banks.

[29] Rotiblat, H., Kershaw, A. and Oot, P. (2010). "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review," Journal of the American Society For Information Science. 61(1), pp. 70–80

[30] Duranti, L.(2015). "Records Lifecycle," in Duranti, L. & Franks, P.C., eds., ENCYCLOPEDIA OF ARCHIVAL SCIENCE (Rowman & Littlefield), pp. 342-346.

[31] Duranti, L. (1989). "The Odyssey of Records Managers Part I: From the Dawn of Civilization to the Fall of the Roman Empire," ARMA Records Management Quarterly 23(3), pp. 3-11.

[32] Lemieux, V. (2001). "Let the Ghosts Speak: An Empirical Exploration of the 'Nature' of the Record," Archivaria (51), pp. 81-111.

[33] Gunnlaugsdottir, J. (2012). "Functional Classification Scheme for Records," Records Management Journal 22(2), pp. 116-129.

[34] Singh, P., Klobas, J.E. and Anderson, K. (2007). "Information Seeking Behaviour of ERMS Users. Implications for Records Management Practices," Human IT 9(1), pp. 135-81.

[35] Lemieux, V., Gormly, B. and Rowledge, L. (2014). "Meeting Big Data challenges with visual analytics: The role of records management," Records Management Journal, 24(2), pp. 122-141.

[36] Meehan, J. 2009. "Towards an Archival Concept of Evidence," Archivaria 61, pp. 127-146.

[37] Lemieux V. and Baron, J.R. (2011). "Overcoming The Digital Tsunami in e-discovery: Is Visual Analysis The Answer?," The Canadian Journal Of Law And Technology, 9 (1-2).

[38] Golub, K. (2006), "Automated subject classification of textual web documents," Journal of Documentation, 62(3), pp. 350-71.

[39] Henttonen, P., and Kettunen, K. (2011). "Functional Classification of Records and Organisational structure," Records Management Journal 21(2), pp. 86-103.

[40] Lemieux, V. (1996). "The Use of Total Quality Management in a Records Management Environment," ARMA Records Management Quarterly, 30(3) (July): 28.

[41] Lemieux, V. (2016). "Trusting Records: Is Blockchain Technology the Answer?," Records Management Journal 26(2), pp. 110-139.

[42] Loehrlein, A., Lemieux, V. and Bennett, M. (2013). "The Classification Of Financial Products," Journal of The Association For Information Science and Technolog, 65(2), pp. 263-280.

[43] McDonald, J. (2010). "Records management and data management: Closing the gap," Records Management Journal, 20(1), pp. 53–60.

[44] Duranti, L. (1997). "The Archival Bond," Archives and Museum Informatics 11(3-4), pp. 213-18.

[45] Andolsen, A. (2008). "The Pillars of Vital Records Protection," Information Management Journal 42(2), pp. 28-32.

[46] Gable, J. (2015). "The Principles and External Audits," Information Management Journal 49(4), pp. 24-27.

[47] Hammouda, K., Matute, D.N., and Kamel, M.S. (2005). "CorePhrase: Keyphrase Extraction for Document Clustering," MLDM: 265-274.

[48] Halkidi, M., Nguyen, B., Varlamis, I., and Vazirgiannis, M. (2003)."THESUS: Organizing Web document collections based on link semantics," VLDB J. 12(4), pp. 320-332

[49] Maqbool, O. and Babri, H.A. (2005) "Interpreting clustering results through cluster labeling," in Proceedings of the IEEE Intl. Conference on Emerging Technologies, 429-434. (September).

[50] Tonella, P., Ricca, F., Pianta, E. and Girardi, C. (2003). "Using Keyword Extraction for Web Site Clustering," in Proc. of WSE 2003, 5th International Workshop on Web Site Evolution, Amsterdam, The Netherlands, (September).

[51] Glover, E. J. Tsioutsiouliklis, K. Lawrence, S, Pennock, D.M. and Flake, G.W. (2002). "Using Web structure for classifying and describing Web pages," in Proceedings of WWW-02, International Conference on the World Wide Web. ACM Press, New York, US, Honolulu, US, 562–569.

[52] Sebastiani, S. (1999). "A Tutorial on Automated Text Categorisation," in Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Analia Amandi, y Ricardo Zunino, eds., Buenos Aires, AR, 7–35.

[53] Johnstone, I. M., & Titterington, D. M. (2009). "Statistical challenges of high-dimensional data," Philosophical Transactions, Series A, Mathematical, Physical, and Engineering Sciences, 367, pp. 4237–4253. http://doi.org/10.1098/rsta.2009.0159.

[54] Hamilton,W.C.(1970). "The revolution in crystallography," Science 169, pp. 133–141

[55] Huber P. J. (1981). ROBUST STATISTICS. New York, NY: Wiley.

[56] Chidanand A., Damerau, F., and Weiss, S.M.. (1994). "Automated Learning of Decision Rules for Text Categorization," ACM Transactions on Information Systems (TOIS), 12 (3), pp. 233 – 251.

[57] Khan, A., Baharudin, B., Lee, L.L., and Khan, K. (2010). "A Review Of Machine Learning Algorithms For Text-Document Classification. Journal Of Advances In Information Technology." 1(1), pp. 4-21.

[58] Sundaram, S. (2012) Comparative Study of Data Mining Algorithms For High Dimensional Analysis. International Journal of Advances in Engineering & Technology, 4(2), pp. 173-8.

[59] Que, H. -E. (2000). "Applications of fuzzy correlation on multiple document classification.Unpublished master thesis", Information Engineering epartment, Tamkang University, Taipei, Taiwan.

[60] Wang, T-W, and Chiang, H-M (2009). "One-Against-One Fuzzy Support Vector Machine Classifier: An Approach to Text Categorization," Expert Systems with Applications.

[61] Cohen W. and Singer, Y. (1996). "Context-sensitive learning method for text categorization," SIGIR' 96, 19th International Conference on Research and Development in Informational Retrieval, pp-307-315.

[62] Hastie, T.; Tibshirani, R.; Friedman, J. (2017). THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE AND PREDICTION. Springer. New York.

[63] Rocchio, J. (1965) "Relevance Feedback in Information Retrieval", in G. Salton (ed.), The SMART System, pp.67-88.

[64] Cohen, W. and Singer, Y. (1999). "Context-sensitive learning method for text categorization", Proc. of SIGIR 96, 19th International Conference on Research and Development in Informational Retrieval, 17(2) (April), pp. 307-315.

[65] Vapnik, V.N. (1995) THE NATURE OF STATISTICAL LEARNING THEORY. Springer, New York.

[66] Brücher, H., Knolmayer, G., Mittermayer, M-A. (2002). "Document Classification Methods for Organizing Explicit Knowledge," Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland.

[67] Joachims, T. (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features" ECML-98, 10th European Conference on Machine Learning, pp. 137-142. .

[68] Chakrabarti, S., Roy, S., Soundalgekar, M.V. (2003). " Fast and Accurate Text Classification via Multiple Linear Discriminant Projection," The International Journal on Very Large Data Bases (VLDB), pp. 170-185.

[69] Archana, S., Elangovan, K. (2014). "Survey of Classification Techniques in Data Mining International," Journal of Computer Science and Mobile Applications, 2 (2) (Feb.), pp. 65-71

[70] Tam, V., Santoso, A., and Setiono, R (2002). "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization," Proceedings of the 16th International Conference on Pattern Recognition, pp.235–238.

[71] Eui-Hong, S.H., Karypis, G, Kumar, V. (1999). "Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification," Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA.

[72] Greiner, R. and Schaffer, J. (2001). AIxploratorium – Decision Trees, Department of Computing Science, University of Alberta, Edmonton, ABT6G2H1, Canada, http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees.

[73] Breiman, L. (2001). "Random forests," Machine Learning 45, pp. 5–32.

[74] Myllymaki, P. and Tirri, H. (1993). "Bayesian Case-Based Reasoning with Neural Network," in Proceeding of the IEEE International Conference on Neural Network, 93, Vol. 1, pp. 422-427.

[75] Guoqiang, P.Z. (2014). "Classification of Breast Cancer Data with Harmony Search and Back Propagation Based Artificial Neural Network," IEEE 22nd Signal Processing and Communications Applications Conference.

[76] Priyadarshini, R. (2010). "Functional Analysis of Artificial Neural Network for Dataset Classification," Special Issue of IJCCT Vol. 1 Issue 2, 3, 4, 2010 for International Conference [ACCTA-2010], 3-5 August.

[77] Hosseini, E., Amini,A.J., Saradjian, M.R. (2003). "Back Propagation Neural Network for Classification of IRS-1D Satellite Images," 1, (2)

[78] Mohri ,M., Rostamizadeh, A., and Talwalkar, A. (2012). FOUNDATIONS OF MACHINE LEARNING, The MIT Press.

[79] James, G. (2003) Variance and Bias for General Loss Functions, Machine Learning 51, 115-135, http://www-bcf.usc.edu/~gareth/research/bv.pdf.

[80] Geman, S. Bienenstock, E. and Doursat, R. (1992). "Neural networks and the bias/variance dilemma," Neural Computation 4, pp. 1–58.

[81] Brodely, C.E. and Friedl, M.A. (1999). "Identifying and Eliminating Mislabeled Training Instans," Journal of Artificial Intelligence Research 11, pp. 131-16, http://jair.org/media/606/live-606-1803-jair.pdf.

[82] Guoqiang, P.Z. (2000). "Neural Networks for Classification: A Survey," IEEE transactions on systems, man, and cybernetics , Vol. 30 (4).

[83] Hastie T. and Tibshirani R. 1990. GENERALIZED ADDITIVE MODELS. London, UK: Chapman and Hall/CR.

[84] Paul. G.L., and Baron, J.R. (2007). "Information inflation: Can the legal system adapt?" Richmond Journal of Law and Technology, 13:art. 10, pp. 1-41.