Assessing Documents' Credibility with Genetic Programming

João Palotti, Thiago Salles, Gisele L. Pappa, Marcos A. Gonçalves and Wagner Meira Jr.
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email: {palotti,tsalles,glpappa,mgoncalv,meira}@dcc.ufmg.br

***Abstract*—The concept of example credibility evaluates how much a classifier can trust an example when building a classification model. It is given by a credibility function, which is application dependent and estimated according to a series of factors that influence the credibility of the examples. Here we deal with automatic document classification and study the credibility of a document according to three factors: content, authorship and citations. We propose a genetic programming algorithm to estimate the credibility of training examples, and then add this estimation to a credibility-aware classifier. For that, we model the authorship and citation data as a complex network, and select a set of structural metrics that can be used to estimate credibility. These metrics are then merged with other content-related ones, and used as terminals for the GP. The GP was tested in a subset of the ACM-DL, and results showed that the credibility-aware classifier obtained results of micro and macroF_1 from 5% to 8% better than the traditional classifiers.**

I. Introduction

Evolutionary algorithms have been extensively used to solve classification problems [1], [2], [3]. As most non-evolutionary classification methods, evolutionary-based ones assume that all training examples should be equally trusted, i.e., all examples should contribute equally to the classification model being generated. Nevertheless, in many application domains, different examples present different levels of what we name credibility.

Consider, for instance, a dataset of documents (e.g., research-related articles) which should be automatically assigned to a set of classes. Suppose a training document was published by an anonymous author on the Web, and another was peer-reviewed and published in a scientific journal. Should they contribute equally to the classification process? Or should the peer-reviewed document, which comes from a more reliable source, contribute more?

Classifiers such as the K-nearest neighbor implicitly deal with this problem by considering only the closest examples to the test instance as credible ones. Nevertheless, most classifiers ignore this problem. Aiming at tackling this issue, in this paper genetic programming changes its role in classification: instead of being used to generate classification functions [3], it is modeled to determine the credibility of an example.

The concept of example credibility, first proposed in [4], is given by a credibility function generated for a specific dataset. In essence, the credibility function is responsible for a transformation in the examples data space, where this transformation is performed according to a series of factors (or credibility dimensions) that influence the credibility of the examples. As these factors are application dependent, here we deal with automatic document classification (ADC), where the credibility of a document relates to factors such as its terms, authors, citations, venues, time of publication, among others.

Defining the list of all factors which might influence an example credibility might be an impossible task. Hence, here we focus on three factors: the **content** (terms), the **authors** and the **citations** of the documents. Note that this is only one instance of the problem, and many other combination of factors could be considered. Previously, we used genetic programming to evolve a credibility function related to the terms of the documents [4], and then simply combine it with well-known metrics of citation and authorship to test a monotonicity hypothesis, i.e., a definition of a credibility function that takes into account many factors should produce more accurate scores than a credibility score referring to an isolated factor.

We chose to first evolve the terms credibility because it can be easily found in any text collection, while authorship and citation are commonly found in scientific documents. In this paper, however, we develop an approach that allows defining credibility functions from any dataset that can be represented as a graph, by extracting from these graphs features commonly used to characterize complex networks, such as clustering coefficient, betweenness centrality, etc [5]. Specifically, here we work with an authorship and a citation graph.

We estimate the credibility function for terms, authors and citations using two different approaches, and then use the scores provided by the functions to a modified version of the Naïve Bayes algorithm, which takes the evolved function into account when generating classification models. This modified version basically weights the documents and its content according to their associated credibility, in order to increase its classification effectiveness. Experiments were performed in the ACM-DL, and showed improvements in micro and macroF_1 of up to 8%.

The remainder of this paper is organized as follows: Section II discusses some related work. Section III introduces the concepts of credibility in document classification and how content, authorship and citation functions can be used to generate credibility functions. Section IV describes the

modeled GP, exploring how to modify a classifier in order to account for credibility. Finally, Section V shows experimental results and Section VI draws conclusions and discusses future work.

II. Related Work

The term credibility first appeared in computer science as a synonym of believability in order to evaluate computer systems, and was defined as a *perceived quality* that results from evaluating multiple dimensions (factors) of an entity simultaneously [6]. This definition suffered some adaptations when it started being applied to evaluate information from the Web [7], specially related to the dimensions which should be addressed.

Regardless of the dimensions evaluated, it is a consensus in the literature that credibility is mainly a subjective matter, but also depends on some objective measures. The most common objective measures proposed so far are strongly based on trust and reputation and citation networks [8], and also credibility rankings that take into account mainly the source of information [9] and content [10].

In contrast with the previously cited works, here we consider the credibility of a document from the **perspective of a classifier**. When building a classification model, the classifier, in the same way as the user, may consider some documents more credible for classifying unseen documents than others. It is important to note that some traditional classifiers implicitly take the credibility of a document into account when classifying new documents. For instance, a weighted voting K-nearest neighbor (KNN), which decides the class of a new document based on a weighted majority voting, considers that closer documents to the one being classified should be more credible than those a bit further in the sample space. Hence, their vote receives a higher weight. Unlike the weighted voting KNN, the proposed credibility function performs a series of transformations on the sample space in order to reflect its associated credibility.

Using a completely different framework, we previously exploited documents credibility based on one factor: its time of creation [11]. We considered that documents which are temporally close to the one being classified are more credible. In [11], documents were weighted using a function following a lognormal distribution, based on the terms' dominance (which captures the strength of the term-class relationship) of a document over time. Based on the good results obtained by this first work, in [4] we proposed a finer-grained approach, and introduced a genetic programming methodology to evolve a credibility function related to the examples' content.

As the definition of credibility may take several dimensions into account, we hypothesize that credibility functions have an asymptotic behavior, and the more the dimensions (or factors) analyzed, the better the scores it produces. Hence, in [4], after evolving credibility functions based only in one dimension, we combined the evolved content-credibility function with a well known bibliometric measure named *amsler* [12], to account for citation-based credibility, and defined an authorship score that takes into account the major publication areas of the authors to assess authorship-based credibility.

With this simple approach, we did not explore the search space to model the credibility based on factors other than terms, leaving unanswered an important question: *can we find a better credibility function for the other dimensions of this problem, namely citation and authorship dimensions ?* To answer this question, in this paper we extend our previous approach to evolve credibility functions associated not only with content, but also to virtually any credibility dimension that can be modeled as a graph, ultimately going one step further towards a more general credibility-aware framework.

Finally, recall that the credibility function defines a transformation in the examples space. By transformation we mean that the examples are moved around the feature space according to the credibility score they receive, and that works as a weight. Previous work have already studied the impact of term weighting schemes in the accuracy of automatic document classification (ADC) [13], [15], [16]. A deep analysis regarding global weighting methods (variants of TF-IDF schema) is presented in [13]. In [15] the authors proposed to use what they call Supervised Term Weighting, which exploits information regarding the distribution of training examples among categories to generate more robust term weighting schemas (local metrics). Their work advocates the use of typical feature selection metrics for term weighting, namely Information Gain (IG), χ^2 and Gain Ratio, reporting significant gains over the simple TF-IDF schema. Indeed, in [16] the authors reported significant improvements in KNN accuracy using Supervised Term Weighting. However, how to *combine* term evaluation functions for ADC continues to be an open challenge, specially if we want to use different credibility dimensions.

III. Modeling credibility in ADC

As previously explained, this work proposes to model a GP for generating a credibility function based on a set of factors that helps the classifier to better discriminate examples from different classes. Hence, when proposing new credibility functions, we first have to identify these factors and then find ways to extract good discriminative properties from each factor representation.

A. Choosing Important Factors

This paper explores document classification, where the most straightforward factor one can consider is content: examples where the majority of terms are highly discriminative may be the most useful to create (learn) accurate models, as they provide important information on how to identify the examples' classes. However, this is just a small part of the information that can be used to assess documents' credibility. Other factors, such as the documents' authors, citations, timeliness—among others—may also significantly improve classification results.

We selected three factors to work with: content, authorship and citations. After selecting them, we looked for ways of measuring how the information provided by each of these factors can help improve classification. In [4] we concluded

that metrics regarding terms distribution, which ultimately explore term-class relationships, are good to assess the content-based credibility of examples. The complete set of 28 explored metrics can be found in [4]. Here we only define the metric which will be used as a baseline for the GP, namely the ambiguity measure (AM), which performed better in previous experiments. The AM [14] measures the strength of term-class relationships, indicating how much confidence one should give to a particular term t as a strong class discriminator, and is given by the ratio of the occurrences of t in a specific class c to the occurrences of t in the entire dataset.

Regarding authorship and citation, we modeled both as networks, which are becoming popular ways for representing data. We then extracted from the networks some of their structural properties. Our basic assumption is that the structural properties of the nodes in a complex network may offer useful information regarding its credibility. As an example, consider a document d assigned to a class c. Let a be its author. If a is an authoritative author regarding class c and has written many documents related to c, then there is a higher probability that d also belongs to c. However, if a is neither an authoritative author in c nor publishes in several distinct classes, then the probability of d belong to c is smaller than in the previous case. The same rationale can be applied to the underlying citation network.

B. Extracting Relevant Credibility Metrics from Complex Networks

If a dataset is modeled as a graph, with non-trivial topological properties (i.e., neither purely regular as a lattice nor purely random) it might be considered a complex network. Formally, consider that a dataset is modeled as a (maybe directed and/or weighted) graph $G^k = (V, E^k)$, where V denotes a set of vertices v_i representing the input examples, and E^k denotes a set of edges $e^{ij} = (v_j, v_k)$, meaning that v_i is connected with v_j (with $v_i \neq v_j$—self loops are not allowed). In our case, we represent the citation network as a directed graph G_{CIT}, where an edge $e_{ij} = (v_j, v_k)$ represents that v_i cites v_j. Similarly, we represent the authorship network as a weighted graph G_{AUT}, where an edge $e_{ij} = (v_j, v_k)$ indicates that v_j and v_k share common authors, and are labeled according to the number of them.

There are several metrics already proposed in the literature to characterize complex networks represented by these graphs [5]. We chose a set of 14 metrics to represent the credibility provided by the examples modeled in the graph, summarized in Table I. Here, we briefly describe the intuition behind the most used in the literature. A detailed definition of these metrics can be found in [5].

In networks, it is quite intuitive to assume that the greater the number of paths in which a vertex or edge participates, the higher its importance. A widely used metric to quantify such importance is the *betweenness centrality*. The betweenness centrality of a vertex v_i (or an edge e_j) is defined as the fraction of the existing shortest paths between two distinct vertices that pass through v_i (or e_j). Considering our case, an influential work in a determined area will tend to be cited by several other works belonging to that same area, ultimately increasing its probability to reside on the shortest paths between these works.

TABLE I
COMPLEX NETWORK METRICS USED AS TERMINALS.

Neighborhood_1	Closeness
Neighborhood_2	Betweenness
Neighborhood_3	Strength
Degree	Burt's constraint
Eigenvector Centrality	Page Rank
Hub Score	AuthScore
Authority Score	Amsler

The *closeness centrality* of a vertex v_i, in contrast, measures how easily other vertices v_k can be reached from it. If an author that publishes in a certain area can be easily reached from another author in the network, the probability that the documents written by the two authors belong to the same class increases as closeness increases. This happens because the probability of co-authors working in the same area is higher than the probability of working in different areas. We realize there are exceptions, such as multidisciplinary areas (e.g., bioinformatics). But we leave for the GP to combine those metrics, weighting them according to their discriminative power.

The *AuthScore* [4] is a metric first defined for characterizing authorship networks, but that can be easily extended to citation. The AuthScore of v_i with relation to a class c is defined as the fraction of authors $\{v_j\}$ that have collaborated with v_i in documents belonging to class c. We assume that there is a positive correlation between the probability of a document to be assigned to class c and the frequency of publication done by its authors regarding documents from class c. To reflect this, we define $d.\textit{authors}$ as the set of authors of d and $Adj(d)$ as the set of documents d_j such that $d.\textit{authors} \cap d_j.\textit{authors} \neq \emptyset$. Let $\mathbb{D}_c$ be the set of documents assigned to class c. The *AuthScore* of a document d' with relation to class c is thus given by

$$\text{AUTHSCORE}(d', c) = \frac{\sum_{d \in Adj(d')} \text{I}(d, d', c)}{|Adj(d')|}, \text{where}$$

$$\text{I}(d, d') = \begin{cases} \frac{|d.\textit{authors} \cap d'.\textit{authors}|}{|d.\textit{authors} \cup d'.\textit{authors}|} & \text{if } d \in \mathbb{D}_c, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the *Amsler* measure [12] reflects the following assumption: two papers d_i and d_j are related if (i) both are cited by the same paper, (ii) both cite the same paper, or (iii) d_i cites a third paper d_k that also cites d_j. As in all cases, we assume that if they are related, they tend to share the same class. Let P_{d_i} be the set of documents that cite d_i and C_{d_i} be the set of documents cited by d. Also, let $Adj(d)$ be the set

of documents that d cites. The Amsler measure is defined as

$$\text{AMSLER}(d', c) = \sum_{\substack{d \in Adj(d') \\ d \in \mathbb{D}_c}} \frac{|(P_{d_i} \cup C_{d_i}) \cap (P_{d_j} \cup C_{d_j})|}{|(P_{d_i} \cup C_{d_i}) \cup (P_{d_j} \cup C_{d_j})|}.$$

In order to calculate these metrics, classes of examples have to be considered. Hence, the networks are built as follows. Let $\mathbb{C}$ be the set of observed classes. When evolving credibility functions, we use a series of projected networks $G^c_{\text{CIT}} \subset G_{\text{CIT}}$ and $G^c_{\text{AUT}} \subset G_{\text{AUT}}$, for each class $c \in \mathbb{C}$, in order to quantify the networks' structural properties for each class. This is important since we aim to establish the examples' credibility scores for a specific class. Given the metrics in Table I, their values are computed for each node in the projected networks. After that, a major challenge remains: how to combine these metrics into a credibility score. This is the problem discussed in the Section IV.

IV. Evolving Credibility Functions

In Section III, we showed how credibility factors can be expressed using metrics of complex networks when a dataset can be modeled as a graph. We also described how citation/authorship networks can be built from the ACM collection, and selected a set of metrics potentially good at discriminating examples from different classes. We also review some content-based metrics, previously studied in [4]. We end up with a set of metrics, which combined can generate effective credibility functions.

Genetic programming [17] appears as a strong candidate method for combining these 60 metrics (28 metrics related to terms, 16 related to authorship and 16 related to citation), as it was previously used with success to combine a large set of metrics related to the term-class relationships (that is, content-based metrics). In [4] we presented the first version of the algorithm to combine only metrics related to content. Here, we end up with three different versions of the algorithm: the first evolves only authorship credibility functions, the second focuses on citation credibility functions, and the third deals with the three factors being studied in this paper altogether.

In all versions, individuals represent a credibility function based on a predefined (set of) factor(s). The terminals are those defined in Table I, and the function set has four arithmetic operators: $+$ (addition), $\times$ (multiplication), Pow (exponentiation), and $\%$ (protected division[1]). In the third version of the algorithm, however, we force the individual to use at least one terminal related to each factor.

At each generation, the fitness of an individual is assessed by modifying the way a classifier creates and evaluates its model, as described in Section IV-A. The best individuals are then selected using tournament selection, and undergo crossover, mutation and reproduction operations according to user defined probabilities. We use a conventional sub-tree swap crossover operation, and a replacement mutation operator. This process goes on until a maximum number of generations is reached.

[1]The division is called protected because when the divisor is zero, it returns zero instead of an error.

A. Fitness Evaluation

Let f_t be the credibility function associated to the examples' content, f_c and f_a denote the credibility functions derived from the citation and authorship networks, respectively, defined as

$$\begin{aligned} f_t &: \mathbb{V} \times \mathbb{C} \mapsto \mathbb{R}^+, \\ f_c &: \mathbb{D} \times \mathbb{C} \mapsto \mathbb{R}^+, \\ f_a &: \mathbb{D} \times \mathbb{C} \mapsto \mathbb{R}^+, \end{aligned}$$

where $\mathbb{V}$ is a set of attributes, $\mathbb{D}$ denotes the set of examples, $\mathbb{C}$ denotes the set of classes and $\mathbb{R}^+$ is the set of non-negative real numbers denoting the credibility scores. Thus, the credibility function associates to each example $d \in \mathbb{D}$ a credibility score for each class $c \in \mathbb{C}$, indicating to what extent d is a credible source of information to build an accurate classification model when considering the class c.

Regardless of the version of the GP being evolved, when evaluating an individual, it is first mapped into the associated credibility functions f_t, f_c and f_a, accounting for the content-based, the citation-based and the authorship-based credibility functions, respectively. Then, the fitness is calculated according to Algorithm 1.

Algorithm 1 Fitness evaluation.

```
1:  function EvaluateFitness(individual)
2:      Content-based Credibility:
3:      for each t ∈ 𝕍 do
4:          for each c ∈ ℂ do
5:              f_t(t, c) ← eval(individual_term, t, c)
6:          end for
7:      end for
8:      Citation/Authorship-based Credibility:
9:      for each d ∈ 𝔻 do
10:         for each c ∈ ℂ do
11:             f_c(d, c) ← eval(individual_cit, d, c)
12:             f_a(d, c) ← eval(individual_auth, d, c)
13:         end for
14:     end for
15:     Fitness Assessment:
16:     fitness ← MicF1(Classifier(d', 𝔻, ℂ, f_t, f_c, f_a))
17:     return fitness
18: end function
```

Algorithm 1 emphasizes the fact that, when the content-based credibility function is used, the credibility is calculated for each pair $\langle$*term,class*$\rangle \subset \mathbb{V} \times \mathbb{C}$ in the dataset. The authorship and citation credibilities, in contrast, are calculated for each $\langle$*document,class*$\rangle \subset \mathbb{D} \times \mathbb{C}$ pair. After these values are calculated, following the credibility function defined by the individual, the terms and documents weights are used to modify a classifier, as the one described in Section IV-B.

Each test document d' is classified using a credibility-aware text classifier, and the achieved microF$_1$ is used as the final fitness function. The microF$_1$ measures the fraction of correct decisions made by the classifier, and is given by $\frac{2\times P\times R}{(P+R)}$, where P (precision) is defined as the fraction of documents assigned to class c that were correctly classified and R (recall) denotes the fraction of documents of class c that were correctly classified. We shall stress here that one could use another classifier effectiveness measure for fitness evaluation. For example, the macroF$_1$ measure (given by the average F$_1$ computed for each class in isolation) could be used for this purpose, ultimately devising credibility functions better suitable to classify minority class examples. However, as reported in Section V, the use of microF$_1$ lead to significant improvements in macroF$_1$, which means that the classification of examples belonging to the minority classes was indeed improved.

B. Credibility-Aware Naïve Bayes

As previously said, a credibility-aware classifier is employed to assess the individuals' fitness, according to the three evolved credibility functions. We modified a Naïve Bayesian classifier to take into account the credibility of the examples. The Naïve Bayes classifier was chosen because in [11] it was showed that it easily outperforms other well-known ADC classifiers, such as Rocchio and KNN. Although SVM can outperform Naïve Bayes in some text classification scenarios, the cost of running SVM in a multi-class problem is very high. Based on this cost/effectiveness trade-off, we chose to perform our first experiments using Naïve Bayes, leaving experiments with SVM for future work.

Considering the original formulation of the Naïve Bayes classifier, a test example d' is classified to the class with maximum a posteriori probability $P(d'|c)$. That is, the decision function is given by:

$$\arg\max_c P(d'|c) = \eta \cdot \frac{N_c}{N} \cdot \prod_{t\in d'} \frac{f_{tc}}{\sum\limits_{t'\in\mathbb{V}} f_{t'c}},$$

where η denotes a normalizing factor, N_c denotes the number of examples previously assigned to class c, and f_{tc} accounts for the frequency of occurrence of term t in class c. Notice that, when considering such formulation, a rather intuitive modification to the Naïve Bayes classifier is to include the content-based credibility function at the term level, that is, into the conditional probabilities

$$P(t|c) = \frac{f_{tc}}{\sum\limits_{t'\in\mathbb{V}} f_{t'c}},$$

since it depends on both the term and class being considered.

Moreover, as both citation and authorship based credibilities depend on the class of the example, it is intuitive to embed their associated function at the example level, that is, into the final a posteriori computation. Thus, we combine these functions in a straightforward manner, using a conjunction operator, where each credibility score is multiplied according to the following credibility-aware decision rule:

$$\arg\max_c P(d'|c) = \eta \cdot \frac{N_c}{N} \cdot \prod_{t\in d'} (P(t|c)\cdot f_t(t,c)) \cdot f_c(d',c) \cdot f_a(d',c).$$

V. Experiments

Experiments were performed on the ACM-DL collection, a sub-collection of the ACM Digital Library which contains 24.897 computer science articles, distributed in 11 classes from the first level of the taxonomy adopted by the ACM-DL Computing Classification System (CCS). All documents in the collection belong to only one class. A preprocess step was used to remove stop words, as well as documents with multiple categories.

In order to generate the credibility functions based on citations and authorship, we generated graphs that represent them from the available documents. The citation graph G_{CIT} has 31.482 vertices connected by 95.812 edges, while the authorship graph G_{AUT} has 16.005 vertices connected by 72.645 edges. Note that the number of documents (vertices) in the citation graph is greater than the number of documents in ACM-DL. This is possible because some of the cited documents are not within the ACM-DL collection, but they are considered for graph construction. More precisely, 5.305 cited documents are in the collection, while 26.176 are not. For documents not in ACM-DL we do not have information about their classes, and thus they are not included in the projected networks G^c_{CIT} and G^c_{AUT} when evaluating the individuals, as described in Section III.

It is important to stress that the citation network is typically more sparse then the authorship network. This reveals some clues regarding promising metrics. As described, the number of vertices with class information in the citation network is typically smaller then the dataset size, and it becomes even sparser since the network is built considering just the documents from the training set (the test document is projected into this network when testing). As a matter of fact, metrics that exploit the neighborhood of vertices may be compromised by the smaller amount of information, whereas metrics that effectively exploit the correlation between nodes and classes (that is, local metrics) tend to perform better.

Indeed, as we shall see, the most promising metrics, selected by the GP framework, typically exhibit such behavior. Considering the authorship network, as it showed up to be less sparse than the citation network, we claim that metrics that explore the neighborhood of vertices may perform quite well. This was observed in our experiments, as will be described soon.

After a series of preliminary experiments, where the GP parameters were set (see Table III), two sets of experiments were run. First, the GP was used to evolve credibility functions for all possible combinations of the three factors, that it, the set of terminals was changed to consider, for instance, only authorship, a combination of the authorship and terms, or all

TABLE II
MACRO-F1 AND MICRO-F_1 OBTAINED BY NAÏVE BAYES WHEN USING DIFFERENT CREDIBILITY FUNCTIONS FOR ACM-DL.

Baseline			GP functions				GP functions linearly combined			
Terms	**Citation**	**Authors**	**MicroF$_1$**	**Gain**	**MacroF$_1$**	**Gain**	**MicroF$_1$**	**Gain**	**MacroF$_1$**	**Gain**
Baseline			73.63 ± 0.90	-	57.26 ± 0.93	-	73.63 ± 0.90	-	57.26 ± 0.93	-
X			73.97 ± 0.66	0.47 ●	58.91 ± 0.77	2.89 ▲	73.97 ± 0.66	0.47 ●	58.91 ± 0.77	2.89 ▲
	X		75.60 ± 0.79	2.67 ▲	59.01 ± 0.83	3.07 ▲	75.60 ± 0.79	2.67 ▲	59.01 ± 0.83	3.07 ▲
		X	76.06 ± 0.76	3.30 ▲	60.35 ± 0.79	5.40 ▲	76.06 ± 0.76	3.30 ▲	60.35 ± 0.79	5.40 ▲
X	**X**		72.83 ± 1.22	-1.08 ●	59.00 ± 1.65	3.04 ●	75.79 ± 0.60	2.94 ▲	60.50 ± 0.92	5.66 ▲
X		**X**	76.13 ± 0.77	3.40 ▲	61.60 ± 0.79	7.57 ▲	76.00 ± 0.58	3.49 ▲	$60.84 \pm .0.32$	6.88 ▲
	X	**X**	77.41 ± 0.83	5.14 ▲	61.68 ± 1.06	7.73 ▲	77.36 ± 0.83	5.07 ▲	61.52 ± 0.79	7.44 ▲
X	**X**	**X**	73.86 ± 1.30	0.31 ●	60.00 ± 1.54	4.78 ▲	77.44 ± 0.72	5.18 ▲	62.19 ± 0.60	8.61 ▲

TABLE III
CONFIGURATION OF GP PARAMETERS.

Parameter	**Value**
Population size	200
Number of generations	50
Crossover probability	0.90
Reproduction probability	0.10
Mutation probability	0.15
Maximum Depth of tree	8
Tournament Size	2

TABLE IV
CREDIBILITY FUNCTION USING DIFFERENT COMBINATIONS OF THREE FACTORS IN ACM-DL.

f_t	f_c	f_a	**MicF$_1$**	**Gain**	**MacF$_1$**	**Gain**
bl.			73.63 ± 0.90	-	57.26 ± 0.93	-
X			72.43 ± 0.98	-1.66 ●	57.39 ± 1.01	0.26 ●
	X		72.46 ± 0.97	-1.61 ●	57.42 ± 0.99	0.28 ●
		X	74.63 ± 1.01	1.36 ●	59.26 ± 0.74	**3.49** ▲
X	**X**		72.46 ± 0.97	-1.61 ●	57.42 ± 0.99	0.28 ●
X		**X**	74.63 ± 1.01	1.36 ●	59.26 ± 0.74	**3.49** ▲
	X	**X**	74.64 ± 1.00	1.37 ●	59.28 ± 0.73	**3.53** ▲
X	**X**	**X**	74.64 ± 1.00	1.37 ●	59.28 ± 0.73	**3.53** ▲

terminals related to the three factors altogether. Since the ADC task is inherently a stochastic process, it is fundamental to adopt some evaluation strategies that guarantee the statistical validity of the obtained classification results, which is achieved by replicating the experiments using different training sets to learn a classification model. For this purpose, we conducted our experiments using a 5-fold cross validation procedure [18]. The results are reported in Table II.

The first three columns in Table II show the types of factors included in the search space, while the next four columns show the values of microF$_1$ (which measures the classification effectiveness over all decisions made by the classifier) and macroF$_1$ (which measures the classification effectiveness for each individual class, averaging them) followed by their standard deviations for the GP evolved functions. The percentage gains provided by the GP-evolved credibility function when compared to the standard Naïve Bayes (baseline) are also reported, and followed by a symbol that indicates whether the variations are statistically significant according to a 2-tailed paired t-test, given a 99% confidence level. ▲ denotes a significant positive variation and ● a non significant variation.

Observe that, apart from the term+citation combination, which obtained the same results as the baseline for microF$_1$ and macroF$_1$, all other results improved macroF$_1$ or both metrics. It is interesting to notice that, when evolving functions for each factor independently, both citation and authorship information led to gains in both metrics. Only the results obtained when using just terms-credibility were statistically equal to those of the baseline.

Analyzing the combinations, we can see that citations and authorship together obtained, for instance, improvements of up to 5.14% in microF$_1$ and 7.73% in macroF$_1$. However, when adding terms to this combination, the gains in microF$_1$ were substantially reduced. This might be explained by the growth in the search space, as we have now 28 terminals related to terms along with the terminals for the other two factors (thus, a total of 60 terminals). We believe that an optimized parameter setting could increase these values substantially.

Overall, the results show that adding credibility to the Naïve Bayes algorithm leads to improvements in microF$_1$ and macroF$_1$. However, based on the hypothesis that a growth in the search space might have confused GP when combining all factors, in a second experiment we combined the three functions evolved independently (that is, in a factor basis) by GP using simply a multiplication factor. The results are reported in the last four columns of Table II.

Notice that, in this case, all combinations improved the results of both micro and macroF$_1$, including the combination of the three factors. Combining content, authorship and citation obtained results of microF$_1$ and macroF$_1$ statistically better than Naïve Bayes, with gains of $+5.18$ and $+8.61$, respectively. It is important to point out that combining the three factors is statistically equivalent to combining only citation and authors.

We performed exhaustive combinations of the three factors because, in theory, we believe the credibility function should be monotonic, that is, adding new factors could improve the values of the scores obtained. Notice that this monotonicity hypothesis is only true when considering the evolutionary process, since it is responsible for generating the improved credibility functions by combing the "best" metrics and discarding the "not so good" ones. The introduction of new

single credibility metrics without evolution may not guarantee monotonicity.

Turning back to our results, they showed that combining citation and authorship is statistically better than using each of them in isolation. At the same time, combining the three factors resulted in credibility scores statistically equivalent to the combination of the two aforementioned factors, what, by the way, does not affect our monotonicity hypothesis. Again, we emphasize that this may be the result of the large growth of the search space (it almost doubled). At the same time, when combining the three functions evolved independently by the GP, and combining them with a simple multiplication operation, our best results were obtained. In any case, our results also indicate that the evolution of credibility functions using only terminals related to one factor along with a simple combination of the generated functions may also be a good compromise between effectiveness and the difficulty related to the growth in the search space.

To further illustrate these arguments, we also compare our results with a set of baselines in which we considered as representatives of the credibility functions for each factor a single metric (i.e., ambiguity measure (AM) for terms, Amsler for citations and AuthScore for authorship) that produced the highest gains. The results are reported in Table IV. Again, the combination of metrics was performed using a simple multiplication operation. Note that none of the baseline metrics obtained gain in microF_1, and four others (including the combination of all metrics) obtained gains around 3.5% when compared to the standard Naïve Bayes classifier. When comparing these results with the best ones reported in Table II, which were obtained by the multiplication of the GP-evolved function independently, our gains are of $+3.75$ and $+4.91$ for microF_1 and macroF_1, respectively.

Finally, it is also interesting to analyze which terminals were used in the evolved functions generated for each factor. As a 5-fold cross-validation procedure was performed, by analyzing the five individuals for each factor, we notice that the most common terminals regarding the metrics of complex networks, and used to describe citation and authorship, were the *Degree*(d,c) and *Neighborhood*$_n$(d,c), followed by the *AuthScore*(d,c). For content, the ambiguity measure and χ^2 were the most common terminals. In Table V, we show three credibility functions evolved for the three credibility factors addressed in this paper. Recall that one of the challenges regarding the use of GP, regardless of the application domain, refers to the intrinsic difficulty of interpreting the semantic meaning of the evolved individuals. We leave a detailed study regarding the behavior and characteristics of the evolved credibility functions as future work.

TABLE V
EXAMPLE OF INDIVIDUALS.

Metric	Individual
Terms	$\left(AM(t,c)^{CHI(t,c)}\right)^{MaxCHI(t)}$
Authorship	$2 * AuthScore(d,c)$
Citation	$Neighbor3(d,c)^2 + Degree(d,c)$

VI. CONCLUSION

This paper proposed to model a genetic programming algorithm to combine different credibility dimensions (factors) in ADC which can improve classifiers results significantly. We focused on three factors: documents content, authorship and citations. The two latter were modeled using a complex network, from which we extracted a set of metrics which give us indications of class separation. This approach can be easily extended to virtually any other application domain where data can be modeled as a graph.

One of the most important parts of this research was to identify which metrics should be combined in order to represent the credibility function. Once these metrics were identified, they were provided to a genetic programming algorithm, which evolved the credibility function. In order to evaluate the function being evolved, we modified the Naïve Bayes algorithm to take into account credibility.

The first set of results showed that the citation and authorship views by themselves can improve results of micro and macroF_1 from 3% to 5%. However, the combination of the three factors did not improve the results from the standard Naïve Bayes in a monotonical way. For instance, combining all the three factors altogether was not better than combining two of them (namely, citation + authors). We believe this happened because the size of the search space grew significantly when all factors were combined. In the future, a better study of the parameters of the algorithms will be necessary, and might change this results. Proof of this are the results obtained in the second set of experiments, where the functions evolved independently were combined using a simple multiplication operation. In this case, the obtained gains increased as the number of factors increased, with gains up to 8% in macroF_1.

The next step is to test the developed approach to other contexts and application domains, such as bioinformatics, where the variety of the examples may demand even more challenging aspects to be tackled by the proposed approach to determine the credibility function. We also want to explore how the SVM classifier can be modified to take credibility into account.

ACKNOWLEDGMENTS

This work was partially supported by CNPq, CAPES, Fapemig, FINEP, and InWeb—Brazilian National Institute of Science and Technology for the Web.

REFERENCES

[1] M. J. Cavaretta and K. Chellapilla, "Data mining using genetic programming: The implications of parsimony on generalization error," in *Proceedings of the Congress on Evolutionary Computation*, vol. 2. IEEE Press, 2009, pp. 1330–1337.

[2] A. A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, 2002.

[3] J. K. Kishore, L. M. Patnaik, V. Mani, and V. K. Agrawal, "Application of genetic programming for multicategory pattern classification," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 242–258, 2000.

[4] J. R. de M. Palotti, T. Salles, G. L. Pappa, F. Arcanjo, M. A. Gonçalves, and W. M. Jr., "Estimating the credibility of examples in automatic document classification," *Journal of Information and Data Management*, vol. 1, no. 3, pp. 439–454, 2010.

[5] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, no. 1, 2005.

[6] S. Tseng and B. J. Fogg, "Credibility and computing technology," *Communications of the ACM*, vol. 42, no. 5, 1999.

[7] M. J. Metzger, A. J. Flanagin, K. Eyal, D. R. Lemus, and R. M. McCann, "Bringing the concept of credibility into the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment," *Communication yearbook*, vol. 27, 2003.

[8] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proceedings of the 13th international conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 403–412.

[9] A. Amin, J. Zhang, H. Cramer, L. Hardman, and V. Evers, "The effects of source credibility ratings in a cultural heritage information aggregator," in *Proceedings of the 3rd workshop on information credibility on the web*. New York, NY, USA: ACM, 2009, pp. 35–42.

[10] A. Juffinger, M. Granitzer, and E. Lex, "Blog credibility ranking by exploiting verified content," in *Proceedings of the 3rd workshop on Information credibility on the web*, ser. WICOW '09. New York, NY, USA: ACM, 2009, pp. 51–58.

[11] T. Salles, L. Rocha, G. L. Pappa, F. Mourão, M. A. Gonçalves, and W. M. Jr., "Temporally-aware algorithms for document classification," in *Proceedings of the International ACM SIGIR Conference on Research & Development of Information Retrieval*, Genebra, Switzerland, 2010, pp. 307–314.

[12] R. Amsler, "Application of citation-based automatic classification," The University of Texas at Austin, Linguistics Research Center, Austin, USA, Technical Report, 1972.

[13] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Cornell University, Ithaca, USA, Tech. Rep., 1987.

[14] S. S. R. Mengle and N. Goharian, "Using ambiguity measure feature selection algorithm for support vector machine classifier," in *Proceedings of the 2008 ACM symposium on Applied computing*, ser. SAC '08. New York, NY, USA: ACM, 2008, pp. 916–920.

[15] F. Debole and F. Sebastiani, "Supervised term weighting for automated text categorization," in *Proceedings of the ACM symposium on Applied computing*, ser. SAC '03. New York, NY, USA: ACM, 2003, pp. 784–788.

[16] I. Batal and M. Hauskrecht, "Boosting knn text classification accuracy by using supervised term weighting schemes," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2009.

[17] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)*. Cambridge, MA, USA: The MIT Press, 1992.

[18] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.