Spectrochimica Acta Part B 65 (2010) 664–670

Contents lists available at ScienceDirect

Spectrochimica Acta Part B

journal homepage: www.elsevier.com/locate/sab

# Progress towards an unassisted element identification from Laser Induced Breakdown Spectra with automatic ranking techniques inspired by text retrieval ☆

G. Amato [a], G. Cristoforetti [b], S. Legnaioli [b], G. Lorenzetti [b], V. Palleschi [b], F. Sorrentino [c,d,e,*], E. Tognoni [f]

[a] ISTI-CNR, Area della Ricerca, Via Moruzzi 1, 56124, Pisa, Italy
[b] IPCF-CNR, Area della Ricerca, Via Moruzzi 1, 56124, Pisa, Italy
[c] Dipartimento di Fisica e astronomia, Università di Firenze, Polo Scientifico, via Sansone 1, 50019 Sesto Fiorentino (FI), Italy
[d] Istituto di Cibernetica CNR, via Campi Flegrei 34, 80078 Pozzuoli (NA), Italy
[e] Marwan Technology, c/o Dipartimento di Fisica "E. Fermi", Largo Pontecorvo 3, 56127 Pisa, Italy
[f] INO-CNR, Area della Ricerca, Via Moruzzi 1, 56124 Pisa, Italy

ARTICLE INFO

*Article history:*
Received 23 November 2009
Accepted 25 April 2010
Available online 17 June 2010

*Keywords:*
Automatic processing
Element identification
Ranking techniques
LIBS
Spectral analysis

ABSTRACT

In this communication, we will illustrate an algorithm for automatic element identification in LIBS spectra which takes inspiration from the vector space model applied to text retrieval techniques. The vector space model prescribes that text documents and text queries are represented as vectors of weighted terms (words). Document ranking, with respect to relevance to a query, is obtained by comparing the vectors representing the documents with the vector representing the query.

In our case, we represent elements and samples as vectors of weighted peaks, obtained from their spectra. The likelihood of the presence of an element in a sample is computed by comparing the corresponding vectors of weighted peaks. The weight of a peak is proportional to its intensity and to the inverse of the number of peaks, in the database, in its wavelength neighboring.

We suppose to have a database containing the peaks of all elements we want to recognize, where each peak is represented by a wavelength and it is associated with its expected relative intensity and the corresponding element.

Detection of elements in a sample is obtained by ranking the elements according to the distance of the associated vectors from the vector representing the sample.

The application of this approach to elements identification using LIBS spectra obtained from several kinds of metallic alloys will be also illustrated. The possible extension of this technique towards an algorithm for fully automated LIBS analysis will be discussed.



## 1. Introduction

In Laser-Induced Breakdown Spectroscopy (LIBS) the traditional approach for deciding about the presence of one element in the sample involves the measure of the emission intensity of the expected most intense line from the element in object. Typically, the matrix of the sample investigated is known, so that the choice of the line to be measured can be changed according to the possible interfering elements. A threshold is set for the intensity (or the signal-to-noise ratio) to decide whether the element is detected or undetected within the actual limits of detection [1]. Multi-element detection based on multivariate analysis methods was also reported [2]. Some methods for classification of materials by comparison of spectra with spectral libraries representing different groups were also proposed [3–7].

However, LIBS applications to the recognition of materials of a priori unknown composition are constantly growing [8–10], as a consequence of the increasing availability of broadband spectrometers [11]. Unknown materials do not allow the preliminary decision on interfered lines and large amounts of spectral data must be processed. Reliable automatic algorithms able to identify elements from the characteristic spectral peaks detected in LIBS spectra are therefore needed, to support the analyst in the processing of this massive amount of data. A semiautomatic software for processing LIBS spectra has been described in [12]. However, to our knowledge no methods for fully unassisted LIBS analysis have been proposed so far.

A typical broadband LIBS spectrum may include more than thirty thousand wavelength values and corresponding intensities [13]. Although not all of those data carry significant information, in many cases the analyst has nevertheless to cope with LIBS spectra characterized by thousands of emission peaks, affected by a certain

☆ This paper was presented at the 5th Euro-Mediterranean Symposium on Laser Induced Breakdown Spectroscopy, held in Tivoli Terme (Rome), Italy, 28 September–1 October 2009, and is published in the Special Issue of Spectrochimica Acta Part B, dedicated to that Symposium.

* Corresponding author. Dipartimento di Fisica e astronomia, Università di Firenze, Polo Scientifico, via Sansone 1, 50019 Sesto Fiorentino (FI), Italy.
*E-mail address:* sorrentino@fi.infn.it (F. Sorrentino).

0584-8547/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sab.2010.04.019

amount of noise. The origin of this noise can be instrumental, and then in principle reducible through a proper optimization of the experimental equipment. Among this category we could mention the electronic noise of the optical detector, which may partially or totally hinder the emission of weak peaks, the effect of a poor wavelength calibration of the spectrometer which would apparently shift the peak from its predicted spectral position, or the result of a poor spectral resolution which would broaden the spectral features of the elements in the plasma [14]. On the other hand, there are sources of "noise" arising directly from the physical properties of the laser induced plasma, such as the presence of a background emission due to continuum-continuum and continuum-bond transitions in the laser-induced plasma [15], or the physical shift and broadening of the spectral peaks due to the Stark and Doppler effects [16]. Those effects can be reduced with a careful choice of the experimental conditions but, as well as the instrumental noise, they could never be completely eliminated from the spectrum, so that the identification of the spectral peaks cannot be simply done through a comparison of the measured line emission wavelength with the theoretical one, corresponding to the observed transition. In fact, the experimental uncertainties in wavelength calibration, the finite resolution of the spectral detectors and the physical shift and broadening of the emission peaks, would eventually result in the association of several possible transitions characteristics of different parent elements to a given peak in the spectrum.

The purpose of this work is to develop a method for automatic determination of the most probable qualitative composition of the sample under study compatible with the spectral features determined by the LIBS analysis. This is obviously a pre-requisite towards the development of systems for automatic *quantitative* analysis of LIBS spectra capable of operating without the assistance of the analyst.

Assuming to have a list of experimental peaks, i.e. wavelength/intensity pairs, our algorithm compares the observed peaks to those expected from a database for each element of the periodic table. The result is a number (score) associated to each element, representing an estimate of the likelihood of the presence of the element in the sample.

The ideas presented in this paper were implemented in two versions. We produced both a Java command line implementation and a Matlab implementation.

## 2. Principles of the method

The basic idea of the approach that we propose is to estimate the likelihood of the presence of an element in a sample by automatically comparing the peaks of the spectra of elements and those of the sample. To obtain that we perform two steps: 1) extracting the peaks from spectral data, and 2) matching the peaks of the sample against the peaks that are supposed to be present in an element.

### 2.1. Peak detection

The first step in the analysis consists in the extraction of the relevant information from the spectral data. The problem of automatically determining number and position of significant emission peaks is common to many spectroscopic techniques. For obtaining the wavelength and associated intensity values corresponding to the observed emission lines, the peak can be for example identified as the occurrence of a local minimum of the second derivative of the intensity curve [17]. A threshold in the second derivative value can be defined for rejecting the instrumental noise of the spectrum background and retaining only the meaningful information. The selected pairs of wavelength-intensity values corresponding to the peaks are then stored in a vector (vector of experimental peaks).

This method is particularly powerful for locating the maxima of partially superimposed peaks; however, it is also quite sensitive to noise. A smoothing of the LIBS spectrum with, for example, a Savitzky–Golay filter [18], would help in reducing the noise and, thus, in improving the effectiveness of the peak finding algorithm.

Since the purpose of this paper is the description of an automatic method for unassisted element detection from a list of given experimental peaks, the optimization of the peak finding algorithm will not be discussed further. In the following we will assume that a series of significant peaks is available, and we will present a method for the determination of the most probable set of elements that could be associated to such peaks. In Section 4 we discuss possible algorithms for automatic choice of the parameters used for peak finding (i.e. threshold and smoothing degree).

### 2.2. Use of a database

For the second step, we suppose to have a database containing the peaks of all elements we want to recognize. In the database, each peak would be represented by its wavelength, the corresponding emitting species and its intensity (relative to the other peaks of the same species). The peak database used in this work is derived from the NIST Atomic Spectra Database [19]; it contains, among other information not used for the purpose of this paper, the main emission peaks of the elements, the corresponding emitting species and an estimation of the intensity of the spectral peak. This latter parameter is important for the implementation of the method proposed; however, the theoretical peak intensities reported in the NIST database may refer to different temperature conditions compared to those typical in LIBS plasmas. Moreover, the theoretical intensity values cannot be used for comparing the intensities of the emissions of different species. One indeed could consider the possibility of recalculating the (relative) values of line intensities for each element, by measuring the plasma temperature from the LIBS spectrum and using the parameters for the transition probabilities and level degeneracies available in the literature [20]. However, it should be considered that the spectral parameters needed for such kind of calculations are available only for a limited number of lines and elements; moreover, the accuracy of the transition probabilities is extremely variable (ranging from 5 to 50% and over, for some line) and the theoretical calculation which allows to obtain the plasma temperature from the LIBS spectrum is valid only in the approximation of optically thin, homogeneous plasma in Local Thermal Equilibrium [21]. Last but not least, the determination of the plasma temperature that can be obtained from the LIBS spectrum is based on the identification of at least one element in the plasma, while for the application of the proposed method the composition of the sample is considered a priori unknown.

The theoretical intensities in the database must thus be normalized in such a way that for each species the sum of the intensities of all the peaks in the database would be equal to unit.

For each element of the periodic table, the pairs of wavelength-intensity values contained in the database are then stored in a vector (vector of the element).

### 2.3. Full description of the method

As anticipated before, our objective is to obtain an estimation about the presence of an element in a sample by matching the peaks that are supposed to be seen if the element is present and those that are actually detected in the sample.

In this paper we propose to automatically perform this matching taking inspirations from text retrieval techniques [22], which are widely used in text retrieval search engines, like for instance web search engines. More technically, we will build our technique using the *vector space model*, which, in case of text retrieval, prescribes that text documents and text queries are represented as vectors of weighted terms (words). The weight of a term is proportional to the importance of the term for the query or document. Document ranking