Automatic classification of journalistic documents on the Internet[1]

Classificação automática de documentos jornalísticos na Internet

Elias OLIVEIRA[1]
Delermando BRANQUINHO FILHO[2]

Abstract

Online journalism is increasing every day. There are many news agencies, newspapers, and magazines using digital publication in the global network. Documents published online are available to users, who use search engines to find them. In order to deliver documents that are relevant to the search, they must be indexed and classified. Due to the vast number of documents published online every day, a lot of research has been carried out to find ways to facilitate automatic document classification. The objective of the present study is to describe an experimental approach for the automatic classification of journalistic documents published on the Internet using the Vector Space Model for document representation. The model was tested based on a real journalism database, using algorithms that have been widely reported in the literature. This article also describes the metrics used to assess the performance of these algorithms and their required configurations. The results obtained show the efficiency of the method used and justify further research to find ways to facilitate the automatic classification of documents.

Keywords: Automatic classification. Internet. Vector model.

Resumo

As publicações de textos jornalísticos na Internet *aumentam a cada dia. São muitas as agências de notícias, jornais e revistas com publicações digitais na grande rede. Os documentos publicados ficam disponíveis aos usuários, que, por sua vez, usam buscadores para encontrá-los. Para se encontrarem os documentos mais próximos da busca realizada, estes devem estar previamente indexados e classificados. Com o volume gigantesco de materiais publicados a cada dia, muitas pesquisas estão sendo realizadas para encontrar facilitadores para a classificação automática de documentos. Este artigo objetiva demonstrar uma experiência de classificação automática de documentos jornalísticos publicados na* Internet*, usando o Modelo Vetorial de representação. A partir de uma base de dados jornalística real, o modelo é testado por meio de algoritmos largamente utilizados na literatura. O artigo descreve ainda as métricas de avaliação de desempenho desses algoritmos e as configurações necessárias para a sua reprodução. Os resultados mostram a eficiência do método e justifica as pesquisas acerca de meios facilitadores para a classificação automática de documentos.*

Palavras-chave: Classificação automática. Internet. *Modelo vetorial.*

[1] Universidade Federal do Espírito Santo, Programa de Pós-Graduação em Informática, Laboratório de Computação de Alto Desempenho. Av. Fernando Ferrari, 514, Goiabeiras, 29075-910, Vitória, ES, Brasil. *Correspondência para*/Correspondence to: E. OLIVEIRA. *E-mail*: <elias@lcad.inf.ufes.br>.

[2] Faculdade Católica Salesiana do Espírito Santo, Cursos de Informática. Vitória, ES, Brasil.

Received in 26/1/2016, resubmitted on 7/11/2016 and approved in 9/2/2017.

Introduction

With the advent of the Internet and the proliferation of documents on the World Wide Web, finding documents quickly and effectively is a recurring problem. Search tools such as Google, Yahoo, and Bing, among others, help users accomplish this task. Robredo (2013) carried out a study addressing the concept of indexing and information retrieval in the era of electronic publication. On the other hand, in our computers there is an overload of folders and files, including text documents, photos, and videos. Thus, when a new file is received or posted by the news agency, we identify it according to a classification, which can be a simple text, a video, or an image. This file is then stored in one of these classes, which had been previously separated into folders.

The variety of classes or subclasses can increase at such a fast rate that it would be difficult to manage. In a library, for example, there are many classes predetermined by naturally grouping characteristics, such as books, magazines, and dissertations, among others. At the same time, it can be observed that within the magazine category, there may be dozens of other classes, such as clothing magazines, without separating men's and women's fashion magazines, scientific magazines, etc. (BERGMAN *et al.*, 2012; KWAZACUTE, 1991).

Focusing exclusively on digital documents, separating or classifying a new document becomes an arduous task, and due to the considerable amount of documents, manual classification can be infeasible. As a result, many solutions for automatic document classification have been studied by researchers all over the world (AGGARWAL; ZHAI, 2012; BAEZA-YATES; RIBEIRO-NETO, 2013; OLIVEIRA *et al*., 2014).

Millions of documents are generated on the Internet daily. Social networks are inexhaustible document sources. Thus, it is possible to consider each post on a social network as a document, and each one expresses an idea or opinion. Some authors consider these postings as sentiment expression (LIU, 2012; LIU; ZHANG, 2012; PAK; PAROUBEK, 2010; TANG *et al*., 2015).

Some methods for grouping documents are based on the similarity between their elements. In information retrieval, documents are elements and their attributes are words (terms). According to Dattola (2013), the classification of document files can be divided into two categories. The first one is *a priori* classification, in which the class already exists and each new document is placed into the cluster whose centroid is most similar to that document. The second one, a no *a priori* classification, is specified, and clusters are formed only on the basis of similarities between documents.

Classification schemes that fall into the first class are very common and often involve manual work. Those in the second type of category are usually more difficult to handle, and automatic or semi-automatic methods are often used. This type of methods are widely used in statistical programs, and in information retrieval applications, the number of elements may reach millions of documents (BAEZA-YATES; RIBEIRO-NETO, 2013; DATTOLA, 2013).

The present study employed a widely used methodology for automatic classification of a large number of documents. The methodology used is associated to the vector space model for representation of document features, and it allows comparing the distance between the new incoming document and the existing documents that had been previously separated into classes by humans. By comparing the results obtained, it was possible to measure the degree of efficiency of the automatic classification method. The *A Tribuna* database, used in this experiment, is an online database for journalism, a factor that may be important for professionals in this field and may contribute to the understanding of other similar databases.

This study is organized as follows: Section 2 presents the technologies and studies on document classification. Section 3 introduces the literature review that will help readers understand how the classification of document using vector space model works and its implications regarding the *a priori* classification of documents. Section 4 describes the experiment carried out and discusses the results obtained. Section 5 presents the conclusion and a perspective of future research.

https://doi.org/10.1590/2318-08892017000300003

Methodological procedures

Classification of Digital Documents

Different kinds of files (text documents, videos, images, music, etc.) are distributed across the Internet. Most of these files are not correctly cataloged, which may be due to the fact that the people responsible for it do not know or are not concerned about it. This apparent lack of concern leads to searches that will often return irrelevant documents to the user (ALVES, 2005; CHEN *et al*., 2012).

Seeking to improve the performance of digital library searches, some initiatives have been developed, such as Google Book Search (COYLE, 2006; HAMILTON, 2014; OLIVEIRA *et al*., 2007; SAMUELSON, 2009). Moreover, companies in the information industry (library and publishers) have invested in digital books such as eBooks. Therefore, in addition to books, digital collections include dissertations, theses, and scientific articles, which often involve cataloging (LOURENÇO, 2007), in which online databases, such as Online Public Access Catalogs, are commonly used resources in libraries (ALVARENGA, 2003).

The *a priori* classification requires time and professional effort of human specialists. Therefore, Alvarenga (2003, p.19) adds that:

> In the new context of production, organization, and retrieval of digital objects, the goals are not restricted to the creation of symbolic representations of the documents in a collection; they also include the creation of new ways of writing for hypertexts and the creation of the so-called metadata, many of which can be extracted directly from the objects themselves. Therefore, they are access keys for Internet users.

Thus, based on the understanding that the representation of documents in a collection can be extracted from the documents themselves, there is a window for documents that have not been classified by human specialists (*a priori* classification) (ALVARENGA, 2003; DATTOLA, 2013; OLIVEIRA *et al*., 2007).

Digital document indexing

Documents must be treated to ensure efficient information retrieval (ALVES, 2005; CASTRO *et al*., 2007). This important step, such as cataloging performed by a human specialist during technical document treatment, consists of document indexing. Like the manual process of extracting terms or words that can represent the document, automatic indexing is based on word frequency - the number of times terms occur in the document itself and in the collection. Moreover, it can also be based on the presence of words in a dictionary or thesaurus. Therefore, automatic indexing is the extraction of meaningful terms for document representation (BÜTTCHER *et al*., 2010).

Despite research efforts and computational resources, the number of text documents on the Internet makes manual catalog infeasible, which has motivated the launch of initiatives for the automatic classification of documents on the Internet (SOUZA *et al*., 1997).

Indexing Models

There are many procedures for identification and selection of terms that can represent a document. The present study focuses on the Vector Space Model (BAEZA-YATES; RIBEIRO-NETO, 2013; SALTON *et al*., 1975), but there are others, such as the Probabilistic, the Boolean, the LSI (Latent Semantic Index), and Neural Network models (BAEZA-YATES; RIBEIRO-NETO, 2013).

Automatic indexing takes into account the frequency with which a term occurs in each document. Another factor is the frequency with which a term appears in the whole document collection (BÜTTCHER *et al*., 2010; BAEZA-YATES; RIBEIRO-NETO, 2013). As part of the document treatment process, some terms can be removed if they do not act as meaningful criteria in a query; these words are called Stopwords. On the other hand, relevant terms are weighted, which reflects their significance in terms of representativeness.

The exclusion of terms should be done with caution and depends on some variables, such as the documents' domain and goals of future search since in a certain context verbs can be considered stopwords, whereas in another context they are considered relevant (LO *et al.,* 2005; BAEZA-YATES; RIBEIRO-NETO, 2013).

There are several techniques for the treatment of documents in terms of Stopwords, such as Term Frequency-Inverse Document Frequency (TF-IDF) (which will be defined in the next section), the Genetic Algorithm, lists of prohibited words for a certain language, and other automatic construction algorithms. According to the literature, this is the pre-processing stage, and it can be used combining lexical analysis, Stopwords removal, and Stemming (extract the root of the word). In practice, stemming is the procedure used to extract syntactic variations of words, such as the plural and the affixes represented by prefixes and suffixes. In addition to these procedures, there is also the selection of terms to be indexed or categorical structures such as the thesaurus, which is an instrument that gathers terms chosen from a previously established conceptual structure and are intended for indexing and retrieving of documents and information in a certain field of knowledge. It is not just a dictionary but an instrument that serves to guide the indexer and the researcher in the processing and search for information (http://portal.inep.gov.br/o-que-e-o-thesaurus, available on Aug. 21, 2015) (BAEZA-YATES; RIBEIRO-NETO, 2013; BLANCHARD, 2007; DOLAMIC; SAVOY, 2010; PORTER, 1980; WILBUR; SIROTKIN, 1992).

In the present study, we used lemmatization to improve the representativeness of the terms. Accordingly, Lucca and Nunes (2002, p.4) argue that "lemmatization is the act of representing words by the infinitive form of verbs and the singular masculine form of nouns and adjectives"[3]. Thus, we can exemplify the process used in this study based on one of the terms of the document, such the following words in Portuguese: *abala, abalada, abaladas, abalado, abalados, abalam, abala-me, abalam-me, abalamos, abalando, abalar, abalara, abalará, abalara-lhe, abalaram, abalarem, abalariam, abala-se, abalasse, abalassem, abalava, abalavam, abalava-nos, abalava-se, abalei, abalem, abalo, abalou, abalou-a, abalou-me, abalou-o, abalou-se*; and they were all replaced with *abalar* (infinitive form of the verb).

It is important to remember that the main objective of automatic document indexing is to reduce human involvement, *i.e.*, considering the volume of documents published on the Internet, the automation process helps human specialists in the technical treatment, i.e., they will not have to read all documents to choose their representative terms (ALVES, 2005).

Documents on the Internet, as well as in any other domain, can be divided into two types. The first type refers to structured texts, in which the choice of terms may be done based on titles among other predefinitions, such as police reports, newspapers, and magazines whose textual bases remain the same although their content is different in each publication. The second type refers to the unstructured texts, which are documents found mostly in online texts (ARAÚJO JÚNIOR; TARAPANOFF, 2006; TRYBULA, 1999). For the purpose of this study, the documents in our database, extracted from the online version of the newspaper *A Tribuna*, were considered unstructured texts.

The methodology and algebra involved in the indexing process will be presented below. As for the methodology, the documents will be represented as vectors, and statistical methods were used for data analysis.

Vector space representation

The vector space model used represents the documents as vectors in a multidimensional space R^n, where n is the number of terms (words) found in the collection (*D*) or the set of all documents considered in the classification (BAEZA-YATES; RIBEIRO-NETO, 2013).

Therefore, each document becomes a term vector. Using Linear Algebra, let $D=\{d_1,d_2,;....d_j,...,d_n\}$ be the set of documents, where d_i are the documents of the collection D. In the space model vector representation, a word or term is replaced with a number whose value is the frequency of that term in each document. However, these values may be weighted due to their importance or other criteria established by the researcher. When it is assigned values instead of words, the weight vector is built based on each document $d_i=\{w_1, w_2,...w_k\ w_{k+1},w_{k+2},...,w_n\}$, where k is the number of different terms $\{t_1, t_2,...,t_k\}$ in the document d_i. The other terms $\{t_{k+1}, t_{k+2},...,t_n\}$ and their respective weights $\{... w_{k+1}, w_{k+2}, ... ,w_n\}$ belong to the other documents. Therefore, $\{t_1, t_2,...,it_k\ t_{k+1}, t_{k+2},...,t_n\}$ are the terms found in the document d_i.

[3] "A lematização é o ato de representar as palavras através do infinitivo dos verbos e masculino singular dos substantivos e adjetivos".

https://doi.org/10.1590/2318-08892017000300003

This type of representation allows one to conclude that a term can appear in one document or in several documents, as well as in all documents in the collection. Terms may be rare, appearing in only one document or may occur very frequently in most or all documents in the collection and are assigned different weights (BAEZA-YATES; RIBEIRO-NETO, 2013). For each term, a weight w_j is assigned according to two aspects, as previously mentioned: the first is the frequency with which the term appears in the analyzed document Term Frequency (TF); the second is the frequency of the term in the other documents of the collection Inverse Document Frequency (IDF). This is one of the simplest proposals available in the literature, which according to Baeza-Yates and Ribeiro-Neto (2013, p.68):

$$w_{i,j} = \begin{cases} \left(1 + log\left(f_{i,j}\right)\right) \times log\left(\frac{n}{n_i}\right) & if\ f_{i,j} > 0\ ;\ n_i > 0 \\ 0 & otherwise \end{cases}$$

Definition: For the vector space model, the weight $w_{i,j}$ associated with the document/term pair (k_i, d_j) is positive and non-binary. The index terms are assumed to be mutually independent and are represented as vector units of a t-dimensional space, where t is the total number of terms. The representation of the document d_j and the query q are t-dimensional vectors given by:

$$\overrightarrow{d_j} = \left(w_{1,j}, w_{2,j}, \ldots . w_{t,j}\right)$$

$$\overrightarrow{q} = \left(w_{1,q}, w_{2,q}, \ldots . w_{t,q}\right)$$

Where $w_{i,q}$ is the weight associated with the index term-query pair (k_i,$_q$), with $w_{i,q} \geq 0$. The weights in the vector space model are basically TF-IDF weights. In the present study, the third weighting scheme was adopted, according to "table 3.6" in (BAEZA-YATES; RIBEIRO-NETO, 2013, p.74), as follows:

$$w_{i,j} = \left(1 + \log f_{i,q}\right) \times \log\left(\frac{N}{n_i}\right)$$

$$w_{i,q} = \left(1 + \log f_{i,j}\right) \times \log\left(\frac{N}{n_i}\right)$$

A word in the document is a term associated with a weight w_i. The weight is the frequency with which the term occurs in a given document (TF) multiplied by the weight of the document in the collection. That is, the variable N is the number of documents in the collection, and n_i is the number of times the term appears in the collection documents (IDF). Thus, it can be concluded that if a term occurs in all documents or only in one of them, its IDF weight will be low and tend to zero, which means the term representativeness in the document will be low or zero. However, there are other weighting factors in the literature, such as the Inverse Category Frequency (ICF), which is similar to the IDF but takes into account the frequency of the term in the classes. Among other weighting factors, another example is the *Chi-Square*, which measures the independence of the term in relation to a certain class (SOUZA *et al*., 2014).

To exemplify the model proposed, the following text was used to stratify the terms and their respective frequencies. In this example, the IDF relative value considered for all terms is one [IDF=1], using only the absolute frequency (TF). The document content is the publication of a fictitious news story from a news agency, i.e., an online newspaper: - d_1: As 'crianças' querem ir para a 'escola', mas a 'professora' está 'doente'. Então as 'crianças' não poderão assistir a 'aula' hoje, pois a escola está 'fechada' e também porque 'tem' muitos 'alunos doentes' (the children want to go to school, but the teacher is sick. Thus, the children will not be able to attend class today because the school is closed and there are a lot of sick students). For illustrative purpose, Table 1 was created to show the terms and their frequencies. It is worth noting that the Stopwords were removed, i.e., they were not counted.

Table 1. Term frequency in the document. Terms given in Portuguese language.

Index i	Weight w_i	Term (t_i, d_j)
1	2	Criança
2	3	Escola
3	1	Professora
4	2	Doente
5	1	Aula
6	1	Hoje
7	1	Fechada
8	1	Aluno
9	1	Muito

Source: Prepared by the authors (2015).

For the purpose of this illustration, the following Stopwords in Portuguese were randomly chosen: *As*

querem ir para a mas está então as não poderão assistir hoje pois está e porque também tem; the lemmatized words were: *crianças, doentes, alunos*. In Table 1, the term with the highest value is *escola*, with frequency equal to three. To improve the visual representation of the model, we will use the terms with a frequency ≥50% of the highest frequency found, *i.e.*, the terms with values ≥1.5 since this value was found by dividing the highest frequency of the term *escola* by two (3÷2=1.5). In the example, the terms that represent the document in the vector space model lead to the weight vector, which in this case are the frequencies only, as follows:

$$\overrightarrow{d_1} = (2,3,2)$$

In order to demonstrate how computing can facilitate the work of human specialists in terms of document classification, let us suppose that two other documents were treated the same way. Thus, their vectors will be represented as:

$$\overrightarrow{d_2} = (3,5,3)$$

$$\overrightarrow{d_3} = (2,1,3)$$

Therefore, it is possible to visualize the three documents graphically. Figure 1 shows the graph drawn based on the vectors representing the documents d_1, d_2, and d_3. The axes, where T_1=criança, T_2=escola, and T_3=doente, show that the term t_1 has weight 2 for the document d_1, weight 3 for the document d_2, and weight 2 for d_3. Whereas the term t_2 has weight 3 for the document d_1, weight 5 for d_2, and weight 1 for d_3. The term t_3 has weight 2 for the document d_1, weight 3 for d_1 and weight 3 for d_3.

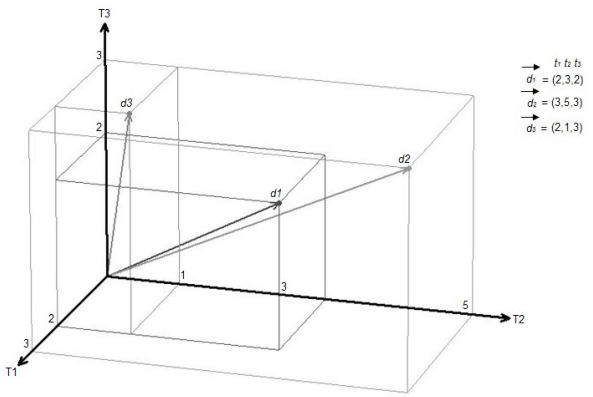

Figure 1. Graphical representation of the documents in the vector space model.

Source: Created by the authors (2015).

This example shows that there is a distance between the points that represent the documents $D=\{d_1, d_2, d_3\}$. In the present study, we will use the cosine of the angle between the pairs of straight lines joining the origin to the points that represent the documents. The calculation of the cosine, described below, is more appropriate since it is normalized (MOITA NETO; MOITA, 1998), *i.e.*, the values vary between zero (those with the lowest similarity) and one (documents with similar terms) (BAEZA-YATES; RIBEIRO-NETO, 2013). The normalization of the distance values is important when there are very different values; the values are then kept between zero and one.

Based on the angular distance, obtained with the calculation of the cosine, it is possible to introduce the concept of similarity sim (d_i, d_j) between pairs of documents since using this technique the distance between them can be verified. In the vector representation of the example above, the calculation is as follows (BAEZA-YATES; RIBEIRO-NETO, 2013, p.78):

$$sim(d_i, d_j) = \frac{\vec{d}_i \circ \vec{d}_j}{\left|\vec{d}_i \times \vec{d}_j\right|} = \frac{\sum_{k=1}^{n} w_k^i \times w_k^j}{\sqrt{\sum_{k=1}^{n}\{w_k^i\}^2 \times \sum_{k=1}^{n}\{w_k^j\}^2}} = \cos(\theta) \quad (1)$$

Where, w_k^i is the weight of the term t_k of the document d_i, and w_k^j is the weight of the term t_k of the document d_j; k is the index or length of the weight vector. The calculation can now be performed between the documents d_1 and d_2 and between d_2 and d_3, and finally between the documents d_1 and d_3, as shown below:

$$sim(d_1, d_2) = \frac{2\times3 + 3\times5 + 2\times3}{\sqrt{2^2+3^2+2^2}\times\sqrt{3^2+5^2+3^2}} = \frac{27}{27{,}04} = 0{,}998 = \cos(\theta)$$

$$sim(d_2, d_3) = \frac{3\times2 + 5\times1 + 3\times3}{\sqrt{3^2+5^2+3^2}\times\sqrt{2^2+1^2+3^2}} = \frac{20}{24{,}53} = 0{,}815 = \cos(\theta)$$

$$sim(d_1, d_3) = \frac{2\times2 + 3\times1 + 2\times3}{\sqrt{2^2+3^2+2^2}\times\sqrt{2^2+1^2+3^2}} = \frac{13}{15{,}42} = 0{,}842 = \cos(\theta)$$

The highest similarity value, 0.998, was found between documents d_1 and d_2, followed by that between documents d_1 and d_3, 0.842, and lastly between the documents d_2 and d_3, 0.815, which was the lowest similarity value. Therefore, it can be seen from Figure 1, that these values correspond to the distances found.

https://doi.org/10.1590/2318-08892017000300003

The example given above was used as an illustration of the vector space model of the representation of the documents; the three terms occur in each one of these documents. In the real world, as previously mentioned, automatic document classification involves, whenever possible, hundreds of thousands of documents, and each document has hundreds or thousands of terms. The experiment carried out in this study, described in the next section, included 45,908 journalistic documents provided by the newspaper *A Tribuna* (online Tribuna is a portal and an Internet provider; it belongs to *Rede Tribuna de Comunicação* (Tribuna communication network), a company based in the city of *Vitória,* in the state of *Espírito Santo* (ES), Brazil: http://www.redetribuna.com.br/online/). The documents had more than 1,600K terms; therefore, it was not possible to represent the vectors graphically.

Applying the model in the real world

For a better understanding of the validation of the vector space model used for document representation, this section will be divided into two parts. The first one shows the necessary adjustments to configure the automatic classification using the model presented here. As previously mentioned, document classification requires the work of a human specialist, who will initially determine the class of the documents that will be used for comparison when an unknown document is presented to the classifier. The second part of this section shows the model validation with the introduction of new documents. The values will be compared to verify the degree to which the model is an accurate representation of the documents.

Model configurations

The main objective of this study is to show how text documents can be classified automatically or semi-automatically, using a vector space model for document representation. The possibility of developing tools that can reduce the time and effort required for human specialists is one of the specific objectives of the present study, but it should be remembered that in many cases. computers will not be able to properly select a document class, thus requiring human involvement (ALVARENGA, 2003; DATTOLA, 2013; OLIVEIRA *et al.*, 2007).

In order to carry out this experiment, the *A Tribuna* database was used, whose main characteristics were previously described showing that this collection is a good source for research since it has already been classified by human specialists and it has 21 classes, as shown in Table 2. The results of the experiments can therefore be compared.

Table 2. Distribution of the documents into the "A Tribune" classes (classes' titles are given in Portuguese language).

Class	Documents (n)	Class	Documents (n)	Class	Nº documents
Atualidades	5,617	Especial	1 470	Opinião	1,634
Qual a Bronca?	346	Família	442	Polícia	4,671
Cidades	5,234	Imóveis	124	Política	5,918
Ciência e Tecnologia	470	Informática	1 506	Regional	1,802
Concursos	309	Internacional	2 187	Sobre Rodas	352
Economia	6,558	Minha Casa	37	Tudo a Ver	30
Esporte	6,657	Mulher	103	TV Tudo	440

Source: Prepared by the authors (2015).

To reduce the dimension of the characteristic vector with 1600K terms in the collection, lemmatization and a set of Stopwords formed by prepositions, articles, numerals, pronouns, and conjunctions were used. As a result, a reduction of 92.31% was achieved, remaining 123K terms. The metrics used to evaluate the results and the performance of the algorithms used are known in the literature as F1, Micro-average, and Macro-average. As suggested by Forman (2003), the metrics used in the present study are as follows:

Precision (p) is the fraction of retrieved documents (Set A) that are relevant (BAEZA-YATES; RIBEIRO-NETO, 2013); in other words, it refers to the number of selected items that are relevant.

$$Precision = p = \frac{|R \cap A|}{|A|} \quad (2)$$

Recall (r) is the fraction of relevant documents (Set R) that are retrieved (BAEZA-YATES; RIBEIRO-NETO, 2013) or the number of relevant items that were selected.

$$Recall = r = \frac{|R \cap A|}{|R|} \quad (3)$$

F1: is the harmonic mean of Precision (p) and Recall (r). This metric is a particular case of a family of similar metrics called F-measure. In the case of F1, it is the ultimate measure of performance of the classifier (FORMAN, 2003).

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4)$$

The basic measures for performance evaluation shown in Table 3 illustrate these metrics (FORMAN; SCHOLZ, 2010). Based on Table 3, we can rewrite equations three and four, respectively, as:

Table 3. Performance of the document x for a class Ci.

	True lable Ci	True not Ci
Predicted lable Ci	True positive (tp)	False positive (fp)
Predicted not Ci	False negative (fn)	True negative (tn)

Source: Adapted from Van Asch (2013, p.2).
Note: *Ci*: Of the Class C indexed by i.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positives}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negatives}$$

And accuracy is defined as:

$$accuracy = \sum_{c}^{C} \frac{tp_c}{N} \quad (5)$$

Where,

tp_c = true positive for class C_i

C = number of different classes

N = total number of C_i being tested

Equation five can be compared to the micro-averaging of Recall and Precision, which can be summarized as:

$$accuracy = \frac{tp + tn}{N} \quad (6)$$

Where,

tp = true positive for class C_i

tn = true negative for class C_i

N = total number of C_i being tested

Found that macroveraging gives equal weights for each class, whereas microaveraging gives equal weight to each per-document classification decision. This is due to the fact that the F1 measure ignores true negatives, and its magnitude is mostly determined by the number of true positives. Thus, large classes dominate small classes in microaveraging (BAEZA-YATES; RIBIERO-NETO, 2013).

The database (*A Tribuna*) was divided by the human specialists into 21 different classes. These data were used to verify how many documents were correctly classified in the automatic model and, consequently, to evaluate the performance of the algorithms, as discussed further ahead.

Knowing which files belong to which class, every class was divided, and consequently the whole database was divided into two parts; after preparation (lemmatization, and removal of Stopwords), two thirds of the classes were divided again for training, and one third was divided for testing.

The part used in training identified the characteristics of the class in the automatic classifier. The part used in the test determines whether the classifier is assigning the document to the correct classes according to their characteristics (LERTNATTEE; THEERAMUNKONG, 2004; SHANKAR; KARYPIS, 2000). Thus, the document class is given by:

$$\underset{j = 1, \ldots k}{arg\ max}\left(\cos\left(\vec{x}, \overrightarrow{C_j}\right)\right) \quad (4)$$

The classification algorithms used, k-Nearest Neighbor (k-NN) and Centroid-Based Classifier (CBC), were chosen to test the model, and the Python (https://www.python.org - Programming language) language and the scientific libraries Scikit-learn [(http://www.scikit-learn.org) Tools for data mining and data analysis],

https://doi.org/10.1590/2318-08892017000300003

NumPy [(http://www.numpy.org) Package for scientific computing with Python], and Scipy [(http://www.scipy.org) Open-source software for mathematics, science, and engineering] were used in the model construction. Moreover, the Shell Script language on Fedora Scientific (Linux operating system with scientific software packages, such as Octave, programming languages, and Integrated Development Environment [IDE]) Linux distribution was also used in the process automation.

The k-NN algorithm is a classifier based on the nearest neighbor, where *k* is the only free parameter that indicates the number of neighbors to be compared. Therefore, the value of *k* was randomly chosen for the experiments. The classification chosen uses some n-dimensional vectors of the vector space model as a training set, in which each element represents a point in the n-dimensional space. This is an exhaustive process that requires a lot of computating power (BEYER *et al*., 1999; SONG; ROUSSOPOULOS, 2001).

The CBC (Centroid-Based Classifier) algorithm uses barycenter calculation, or centroid *Ci* for an ith class, because the centroid vector of documents belonging to the same class is computed. If there are *k* classes {C1, C2, ..., Ck}, there will be one centroid for each class. Thus, for a new document *x*, the similarity between *x* to all *k* centroids will be computed using the cosine measure that will be represented by TF-IDF weights.

In the next sections, we will present the results of the application of the model and its configurations and the conclusions of the experiment.

Results and discussion

Table 4 shows the results of Micro-F1 and Macro-F1 for each algorithm applied to the vector space model used for document representation. Value analysis demonstrated that the micro-averaged value, which shows the performance of the classifier on the set of test documents, was the same in the two algorithms; the break-even point was reached, i.e., precision and recall are equal and the break-even point is close to the optimal value of F1, but they are not necessarily equivalent (YANG, 1999). In other words, the break-even score of a system is always equal or less than the optimal value of F1 of that system. Therefore, the break-even point of a system should not be compared directly with the optimal F1 value of another system. It was also found that classification accuracy of the algorithm k-NN is 7.3% higher than that of the CBC. Macro-averaged scores reflect the average performance of each individual class. In this specific case, different values in relation to the algorithms show that the CBC classification accuracy is 13.4% higher than that of the k-NN algorithm in terms of the values of F1.

Table 4. Comparison between Micro- and Macro-averaged F1 scores.

Micro-averaged scores			
Micro	Precision	Recall	F1
k-NN	0.7787	0.7787	0.7787
CBC	0.7271	0.7271	0.7271
Micro-averaged scores			
Macro	Precision	Recall	F1
k-NN	0.6806	0.5479	0.5715
CBC	0.6274	0.7436	0.6482

Neighbor; CBC: Centroid-Based Classifier.
Source: Prepared by the authors (2015).
Note: F1: is the harmonic mean of Precision (p) and Recall (r); k-NN: k-Nearest

The analysis of the 20 documents classified into categories (classes) other than those randomly selected for the class *Família* (FAM) by the human specialists indicated that the similarity values between these documents suggest that they may be occur simultaneously in more than one class; the similarity value of 0.332303 between the documents a0330072006fam.txt and c3010012004esp.txt is higher than the similarity between the documents a0330072006fam.txt and 0628082006fam.txt 0628082006fam.txt, (0.1755296); both occur in the same class (FAM), which also concerns health. Other documents, such as those in the special class were correctly classified in the experiment; the average similarity between the same documents mentioned above was less than 0.00001.

Discussion and conclusion

Two algorithms were used in the experiment carried out in the present study, and the vector space model for document representation was used to classify the documents based on previously established classes (a priori classification). The results obtained allow us to draw some conclusions about the use of this model: TF-IDF weighting scheme is efficient, but there are other term weighting schemes that should also be considered, such as the ICF, which is similar to the IDF, but the former is related with the total number of occurrences of a term in the classes.

The results also indicate that it is possible to identify the characteristics of the documents, index them, and, with human intervention, classify them for later retrieval thus facilitating the work of human specialists in large volumes of documents since they will not have to evaluate all documents to be classified.

We suggest that future studies should test the *A Tribuna* database using other algorithms in order to verify whether the classification performance can be improved and to evaluate the database response, when using a multi-label classification, for example, i.e., when a document belongs to more than one class simultaneously.

Another aspect that deserves consideration is that a journalistic database for Portuguese language was used; therefore, it can be used to establish standards and configurations for extraction and cataloging of other similar data sources.

Contributions

All authors contributed equally to the conception and design of this study, to data analysis, to manuscript writing, and to critical revision of the final version.

References

AGGARWAL, C. C.; ZHAI, C-X. A survey of text classification algorithms. In: AGGARWAL, C.C.; ZHAI, C-X (Ed.). *Mining text data*. New York: Springer US, 2012. p. 163-222.

ALVARENGA, L. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais 10.5007/1518-2924.2003. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, v. 8, n. 15, p. 18-40, 2003.

ALVES, R. C. V. *Web semântica: uma análise focada no uso de metadados*. 2005. 180f. Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual Paulista, 2005. Disponível em: <http://hdl.handle.net/11449/93690>. Acesso em: 15 out. 2015.

ARAÚJO JÚNIOR, R. H.; TARAPANOFF, K. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. *Ciência da Informação*, v. 35, n. 3, p. 236-247, 2006.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*: The concepts and technology behind search. New York: ACM Press, 2013.

BERGMAN, O. *et al.* How do we find personal files? The effect of OS, presentation & depth on file navigation. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2012, Austin. *Proceedings*… Austin: ACM, 2012. p. 2977-2980.

BEYER, K. *et al*. When is "nearest neighbor" meaningful? In: BEERI, C.; BUNEMAN, P. (Ed.). *Database theory*: ICDT'99. Berlin: Springer Berlin Heidelberg, 1999. p. 217-235.

BLANCHARD, A. Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information,* v. 29, n. 4, p. 308-316, 2007.

BÜTTCHER, S.; CLARKE, C. L. A.; CORMACK, G. V. *Information retrieval*: Implementing and evaluating search engines. Cambridge: Mit Press, 2010.

CASTRO, F. F. *et al*. Os metadados como instrumentos tecnológicos na padronização e potencialização dos recursos informacionais no âmbito das bibliotecas digitais na era da Web semântica. *Informação & Sociedade*: Estudos, v. 17, n. 2, p. 13-19, 2007.

CHEN, W. et al. A noise-aware click model for Web search. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, 5., 2012, Washington. *Proceedings*… Washington: ACM, 2012. p. 313-322.

COYLE, K. Mass digitization of books. *The Journal of Academic Librarianship*, v. 32, n. 6, p. 641-645, 2006.

DATTOLA, R. T. A Fast algorithm for automatic classification. *Information Technology and Libraries*, v. 2, n. 1, p. 31-48, 2013.

DOLAMIC, L.; SAVOY, J. When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*, v. 61, n. 1, p. 200-203, 2010.

FORMAN, G. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, v. 3, p. 1289-1305, 2003.

FORMAN, G.; SCHOLZ, M. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, v. 12, n. 1, p. 49-57, 2010.

HAMILTON, S. The Google Book Settlement: An international library view. *Against the Grain*, v. 22, n. 3, p. 10, 2014.

KWAZACUTE, B. H. The importance of factors that are not

document attributes in the Organisation of Personal Documents. *Journal of documentation*, v. 47, n. 4, p. 389-398, 1991.

LERTNATTEE, V.; THEERAMUNKONG, T. Effect of term distributions on centroid-based text categorization. *Information Sciences*, v. 158, p. 89-115, 2004.

LIU, B. Sentiment analysis and opinion mining. S*ynthesis Lectures on Human Language Technologies*, v. 5, n. 1, p. 1-167, 2012.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: AGGARWAL C. C.; ZHAI CX. *Mining text data*. New York: Springer US, 2012. p. 415-463.

LO, R. T-W; HE, B.; OUNIS, I. Automatically building a stopword list for an information retrieval system. ISSUE ON THE 5TH DUTCH-BELGIAN INFORMATION RETRIEVAL WORKSHOP (DIR), 5., 2005, Utrecht. *Proceedings*… Utrecht: Utrecht University, 2005. p. 17-24.

LOURENÇO, C.A. Metadados: o grande desafio na organização da Web. *Informação & Sociedade*: Estudos, v. 17, n. 1, p. 65-72. 2007. Disponível em: <http://periodicos.ufpb.br/ojs/index.php/ies/article/viewFile/466/1466> Acesso em: 13 jul. 2015.

LUCCA, J. L.; Nunes, M. G. V. *Lematização versus stemming*. São Paulo: USP, 2002. (Série de Relatórios Técnicos do NILC-ICM-USP). Disponível em: <http://www.nilc.icmc.usp.br/nilc/download/lematizacao_versus_steming.pdf>. Acesso em: 12 out. 2015.

MOITA NETO, J. M.; MOITA, G. C. Uma introdução à análise exploratória de dados multivariados. *Química Nova*, v. 21, n. 4, p. 467-469, 1998.

OLIVEIRA, E. *et al*. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. *Revista Brasileira de Biblioteconomia e Documentação*, v. 3, n. 1, 2007, p. 73-98, 2007.

OLIVEIRA, E. *et al*. Combining clustering and classification approaches for reducing the effort of automatic tweets classification. INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND INFORMATION RETRIEVAL, 6., 2014, Rome. *Proceedings*… Rome: KDIR, 2014. p. 465-472.

PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 7., 2010, Valletta. *Proceedings*… LREC: Valletta, 2010. p. 1320-1326.

PORTER, M. F. An algorithm for suffix stripping. *Program*, v. 14, n. 3, p. 130-137, 1980.

ROBREDO, J. Indexação e recuperação da informação na era das publicações virtuais *Comunicação & Informação*, v. 2, n. 1, p. 83-97, 2013. http://dx.doi.org/10.5216/cei.v2i1.22847

SALTON, G.; WONG, A.; YANG, C-S. A vector space model for automatic indexing. *Communications of the ACM*, v. 18, n. 11, p. 613-620, 1975.

SAMUELSON, P. Google Book search and the future of books in cyberspace. *Minnesota Law Review*, v. 94, n. 5, p. 1308-1374, 2009.

SHANKAR, S.; KARYPIS, G. *Weight adjustment schemes for a centroid based classifier*. Minneapolis: University of Minnesota, 2000.

SONG, Z.; ROUSSOPOULOS, N. K-nearest neighbor search for moving query point. In: JENSEN, C. S. *et al*. (Ed.). Advances in Spatial and Temporal Databases. Berlin: Springer Berlin Heidelberg, 2001. p. 79-96.

SOUZA, F. P.; CIARELLI, P. M.; OLIVEIRA, E. Combinando fatores de ponderação para melhorar a classificação de Textos. COMPUTER ON THE BEACH: ANAIS DO COMPUTER ON THE BEACH, 2014, São José. *Anais*... São José: UNIVALI, 2014. p. 32-41.

SOUZA, T.B.; CATARINO, M.E.; SANTOS, P.C. Metadados: catalogando dados na Internet. *Transinformação*, v. 9, n. 2, p.93-105, 1997.

TANG, D.; QIN, B.; LIU, T. Learning semantic representations of users and products for document level sentiment classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 53., INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 7., 2015, Beijing. *Proceedings*… Beijing: ACL, 2015.

TRYBULA, W. J. Text mining. *Annual Review of Information Science and Technology*, v. 34, p. 385-419, 1999.

VAN ASCH, V. *Macro-and micro-averaged evaluation measures basic draft*. Belgium: CLiPS, 2013.

WILBUR, W. J.; SIROTKIN, K. The automatic identification of stop words. *Journal of Information Science*, v. 18, n. 1, p.45-55, 1992.

YANG, Y. An evaluation of statistical approaches to text categorization. *Information Retrieval*, v. 1, n. 1-2, p. 69-90, 1999.