МАТЕМАТИКА

УДК 519.71, 612.82

SEMANTIC RETRIEVAL AND CLASSIFICATION OF WEB PAGES USING FRAMES AND NEURAL NETWORKS

A. LINKEVICH
(Polotsk State University)

Vectors of Hilbert space are used to represent (hyper) texts. Instead of a basis of the space, a frame is constructed so as to represent particular semantic units that could be relevant for a given specific domain. It is suggested to use Kohonen's neural networks for unsupervised classification of texts. A special emphasis is put on handling information saved in the WWW.

1. Introduction

It is a growing tendency that knowledge is saved on the Internet. However this information is computer-retrievable but hardly accessible without human interpretation. It is therefore of great importance to design computer systems able not only to provide a user with information from the Internet, but also to extract and employ the knowledge represented in the WWW.

Currently, retrieval of information saved on the Internet is accomplished with search engines that performs spidering the Web. High coverage of Web pages over the WWW as a whole is the main goal of general-purpose Web search engines, i.e. the aim is to find as many distinct Web pages as possible. In contrast, domain-specific search engines pursue the objective to find Web pages of a particular kind or on a particular topic.

Retrieval of specified information is also required for the automated creation of Web-based knowledge bases. In order to automatically populate such a knowledge base with information available on the WWW, it is necessary to accomplish semantic retrieval and classification of Web pages. There is also a variety of other applications of trainable text classifiers such as browsing assistants, information filtering systems, etc.

The problem of text (document) classification can generally be posed as follows (for an introduction into the subject see, e.g. [1; 2]). Given a set of classes $C = c_1,\ldots,c_L$ and a document d consisting of words $w_1,\ldots,w_n$, it is required to assign the document to a certain class $c^* \in C$ that is most probable. To this end, some features are used as the ground of classification. In principle, one can admit any function of the document and the class, $f_i(d,c)$, to be a feature.

The classification of Web pages mainly rests on the so-called *bag-of-words* or *unigram* document representation when each word in the document is regarded as a separate feature, the classifier ignores the sequence of words in the document and relies on statistics about single words [1]. Most of learning algorithms rely on the *multi-variate Bernoulli* model which uses a Boolean variable to encode the presense of a particular word in the document. The *multinomial* model deals with integer word counts, words are treated as "events", and the document is viewed as a collection of these events. The second method outperforms the first one on several data sets [2].

There are a number of supervised learning algorithms used for text classification, to wit: naive Bayes, k-nearest neighbor, support vector machines, boosting, rule learning algorithms, maximum entropy. (For an introduction into a general framework of learning algorithms for classification see, e.g. [1; 3].) However, not a single technique consistently outperforms the others throughout the variety of particular domains.

The aim of this paper is twofold. The first objective we pursue is to show (in Section 3) that semantic retrieval and classification of texts can be treated with the help of frames outlined in the next section. The general idea of analysis of the meaning of information in terms of some semantic factors represented in Hilbert space by vectors of a frame has been put forward in [4], while here we focus on ways of semantic treatment of texts with special emphasis on handling Web pages. Further, in Section 4, we pose the problem of exploring an unknown Web server as a task of the unsupervised classification that can be accomplished with the help of Kohonen's neural networks.

2. Frames

A natural approach to semantic treatment of a text would be to make a kind of analysis of its meaning in terms of some semantic factors (units, elements, or something the like). To this end, it could be relevant to represent (encode) words, phrases, sentences, texts as well as the above semantic factors by vectors of some Hilbert space H called the *semantic space* [4]. However, it is hardly possible to form such a set of semantic factors in which all the elements are independent of one another. Accordingly, vectors assigned to represent such

semantic factors will appear linearly dependent, and they cannot constitute a basis of the space H. A possible way out is to employ a more general construction, *frame*, instead of *basis*. (More precisely, we use generalized frames as they are defined in [5 – 7]).

A *frame* in Hilbert space H is a set of vectors $|h_a>\in H,\ a\in A$ such that the *metric operator*

$$G=\int_A d\mu\ a\ |h_a><h_a|$$

with a Borel measure $\mu\ a$ is invertible, i.e. there exists an operator T such that $TG=I$. Symbolically, $T=G^{-1}$. Vectors of a frame $|h_a>\in H,\ a\in A$ and vectors of the *reciprocal (dual) frame* $|h^a>\in H,\ a\in A$, where

$$|h_a>=T|h_a>,\qquad \forall a\in A,$$

jointly provide *resolution of unity*:

$$I=\int_A d\mu\ a\ |h_a><h^a|=\int_a d\mu\ a\ |h^a><h_a|. \tag{1}$$

The two forms of the latter decomposition (1) enable one to introduce the two kinds of the transform of vectors. So, if we expand a vector $|u>\in H$ over the reciprocal frame $|h^a>\in H,\ a\in A$ so that

$$|u>=\int_A d\mu\ a\ u_a|h^a>$$

then the components $u_a=<h_a|u>$ are expressed through the vectors of the original frame $|h_a>\in H,\ a\in A$. This transformation performs *analysis* of the vector $|u>$ in terms of the frame vectors so that the inner product $u_a=<h_a|u>$ separates from the vector $|u>$ such a "detail" that is described by the frame vector $|h_a>$. Namely, the quantity $|u_a|^2$ measures how much of the element $|h_a>$ is contained in the vector $|u>$ [5; 8]. Similarly, *synthesis*, or reconstruction, of vectors represents the inverse map $u_a,\ a\in A\mapsto|u>$.

Obviously, the second kind of the transform is to take the representation

$$|u>=\int_A d\mu\ a\ u^a|h_a>.$$

Then the coefficients are $u^a=<h^a|u>$.

The prevailing case is when the lebel variable a can take on only discrete values from some set A and vectors $|h_a>$ are determined by their coordinates $h_{ai}=<e_i|h_a>$ with respect to a countble orthonormal basis $|e_i>,\ i=1,2,\ldots$. Then the problem of construction of frames is posed as follows. *Given* components $h_{ai},\ a\in A,\ i=1,2,\ldots$ of vectors $|h_a>\in H,\ a\in A$ it is *required* to calculate components $h_i^a,\ a\in A,\ i=1,2,\ldots$ of vectors $|h^a>\in H,\ a\in A$ so as to provide the resolution of unity. An *algorithm* of direct computing these h_i^a consists of the following steps: (1) compute matrix elements of the metric operator G_{jk}; (2) find matrix elements T_{ij} of the operator T; (3) find components h_i^a.

3. Using Frames for Semantic Retrieval and Classification of Web Pages

For *semantic retrieval* of information pertaining to a given specific subject, we employ a set of some *semantic units* $\sigma_a, a\in A$ that represent all relevant particular meanings of information in all feasible forms and manners. For humans, any semantic unit can be explained, e.g., with the help of definitions and example sentences. For computer systems, special ontologies are produced.

Each semantic unit σ_a is associated with a vector $|\sigma_a>\in H$ of a Hilbert space H called the *semantic space*. One can compose vector $|\sigma_a>$ so as

$$|\sigma_a>=\ p\ \tilde{w}_1|a\ ,\ldots,p\ \tilde{w}_M|a\ \ ^T,$$

where $p\ \tilde{w}_i|a$ is the probability for word $\tilde{w}_i$ to occur in a “typical” text corresponding to the unit σ_a, and $\tilde{w}_1,\ldots,\tilde{w}_M$ are words from a special vocabulary $\tilde{W}$. (More precisely, $p\ \tilde{w}_i|a$ is not a probability, but its

estimate determined by the average frequency of appearance of the word in appropriate texts. Generally, by word $\tilde{w}_i$ one can imply a proper word of any language and any word group that has a certain particular meaning.)

Any document d is represented in the same way by a vector $|d>\in H$.

Generally, the vectors $|\sigma_a>\in H, a\in A$ can occur linearly dependent. Accordingly, they can not be taken as basis vector of the space. However, under proper conditions we can treat them as elements of a frame and construct the reciprocal frame $|\sigma_a>\in H,\ a\in A$, which allows to accomplish analysis and synthesis of meanings.

Namely, one can expand vector $|d>\in H$ over the reciprocal frame so as

$$|d>=\int_A d\mu\ a\ s_a\ d\ |\sigma_a>.$$

Here the coefficient $s_a\ d$ measures how much of the semantic unit σ_a is represented in the document d, and this quantity will be referred to as the (semantic) *capacity* of the (semantic) unit σ_a in the document d.

The task of a particular semantic retrieval of information is specified by indicating desired values for the semantic capacities as follows: find the set $D_a=\ \forall d\in D: s_a\ d\ \in S_a, \forall a\in A$ of all documents d from a repository D such that their semantic capacities $s_a\ d$ take values from given prescribed sets S_a.

For ordinary text documents, the ground of classification is the content of the document (e.g., the set of words it contains). In contrast, the WWW provides diverse sources of information: (1) the full text of pages, (2) the text in titles and headings, (3) the text associated with the hyperlinks, (4) the text in neighboring pages, (5) the file organization provided by URLs. Accordingly, different kinds of classifiers can be designed. However, not a means appears able to cope with Web pages with sufficient accuracy [2], and combining different approaches could be promising. Obverse that the vector associated in our approach with a document can incorporate data obtained from any source of information about the contents of the document (words in the main body of a hypertext, words in its title, heading, hyperlinks, keywords, abstract, annotations, etc.). In particular, the power of the method increases by combining it with annotating the contents of Web pages supported by SHOE and XML.

A peculiarity of information structured in the Web is that "hyperlinks encode a considerable amount of latent human judgement" [9]. So, the creator of a Web page p, by including a hyperlink to a page q, has a certain *authority* on q, whereas a *hub* page points to multiple relevant authorities. It seems perfectly reasonable to associate each of authorities with a vector included in a frame of the semantic space. Then the user is provided with facilities to search for Web pages that are close, to a certain prescribed degree, to particular chosen authorities.

4. Unsupervised Learning for Agents Exploring Web Servers

Let us suppose that a user encounters a Web server and wonders what kind of information the server can provide. The straightforward wandering along hyperlink paths and browsing all the hypertexts can appear too time consuming, exhausting and does not ensure reaching satisfactorily complete and reliable conclusions. How to classify Web pages when there is ignorance what kinds of information can be represented on them? In this section this issue of efficient exploring an unknown Web server is addressed and treated as the task of unsupervised classification of hypertexts saved on the server.

We consider employment of Kohonen's self-organizing topology-preserving neural networks that are based on the cluster analysis and deal with sequences of statistical samples as follows (see [10] and references therein).

Every input signal $x\ t\ \in R^M,\ t=1,2,\ldots$ is fed to all neurons. Each neuron is characterized by a reference (or "codebook") vector $r_i\ t\ \in R^M$ connected with its states. The input is compared with all $r_i\ t$ and the best-matching neuron *m*, the "winner", is determined so that $d\left[x\ t\ , r_m\ t\ \right]=\min_{i=1,\ldots,N} d\left[x\ t\ , r_i\ t\ \right]$ where $d\ \cdot,\cdot$ is some metric. Further, some neighborhood of neurons $N_m\ t$ is determined around the winner. The states of the neurons from the neighborhood $N_m\ t$ are updated so as to decrease the distance between $x\ t$ and $r_i\ t$. The Kohonen's network is specified by the iterated map

$$r_i\ t+1\ =r_i\ t\ +\alpha\ t\ \cdot\left[x\ t\ -r_i\ t\ \right],\qquad \forall i\in N_m\ t\ ,$$

where $\alpha\ t$ is monotonically decreased during the learning. All the other neurons are left intact, i.e. $r_i\ t+1\ =r_i\ t\ ,$ $\forall i\notin N_m\ t\ .$

The learning process produces an ordered map of inputs to states of the network such hat the reference vectors $r_i\ t$ approximate the probability density function of the entering signals $p\ x$ in the sense of a

minimal residual error. In particular, if the distribution $p\ x$ is clustered then the reference vectors $r_i\ t$ describe these clusters.

If the network is used for classification of some patterns (samples) then each class is normally represented by several reference vectors. Moreover, the borders between the classes turn out to be optimized so as to minimize the errors of misclassification. As a result, the majority of the reference vectors appear placed at the class borders where probability distributions of adjacent classes can intersect one another. In contrast, the number of the reference vectors within a class depends crucially on the variance of samples in the class so that a very dense class can be represented by a single vector, whereas a class with a large variance requires more vectors to settle inside the domain in order to accurately delineate the border.

Heterogeneous vectors can also be handled. So, one part could constitute the pattern (signal) by itself, while the second part would furnish some additional, e.g., symbolic, information pertaining to the signal.

Based on our above consideration, one can suggest a procedure of text classification that includes the following steps. (1) Each document $d \in D$ is represented by a vector $|d>= d_1, d_2, \ldots^T \in H$ composed of, e. g., estimates of the probability distribution function $p\ \tilde{w}_i \,|\, d$ for word $\tilde{w}_i \in \tilde{W}$ to appear in the document d, i. e. $d_i = p\ \tilde{w}_i \,|\, d$. (2) The components of each of these vectors $|d>$ are put into a machine for the unsupervised classification (like the Kohonen's neural network) as the input signal $x\ t$, $t = 1, 2, \ldots$. (3) Operation of the classifier results in clusters of documents that are close to one another in the sense of the metric incorporated in the classifier. (4) The vector corresponding to the center of a claster can naturally be included in a frame of the space H.

Conclusions

In this paper we addressed some issues of semantic treatment of (hyper) texts that appears of great concern as the WWW becomes the information source of paramount importance. Our results might be useful for development of Web spiders and search engines, Web-based knowledge bases, interface assistants and so on. To implement such systems, one needs a special vocabulary similar to computer ontologies (such as CYC), while to combine them with AI systems based on the ontologies, just the same vocabularies should be shared. Yet another appoach to joining different systems is to produce particular appropriate ontologies and/or vocabularies and converters (translators) between them. In this connection, it may occur beneficial to employ the so-called "Regularized" English-Like Language (RELL) [4] which enables one to convey the meaning of information including situations when there is no ontology. RELL can also appear fruitful to handle the grammatical organization of texts, which is, generally speaking, essential to cope with their semantic treatment because the representation of a document as a bag of words is a highly impoverished vision of texts.

In a sense, we made in this paper two steps: the first one deals with the supervised classification of texts, while the second pertains to the unsupervised learning. From another point of view, the first task amounts, in essense, to analysis of information in terms of predefined semantic factors (meanings), while the second issue constitutes, in fact, categorization of information and elaboration of meanings. What is worth apparently doing further is to try to join and integrate these two stages.

REFERENCES

1. Mitchell, T. Machine Learning (Wiley, New York, 1997).
2. Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K., Slattery S. In: Proc. of the Fifteenth National Conf. on Artificial Intelligence (AAAI-98), 1998. – P. 509 – 516 (AAAI Press, Madison, WI); Artificial Intelligence, 2000.
3. Duda R.O., Hart P.E. Pattern Clasification and Scene Analysis (Wiley, New York, 1973).
4. Linkevich A.D. Nonlinear Phenomena in Complex Systems, 2000, 3, 253.
5. Kaiser G. A Friendly Guide to Wavelets (Birkhauser, Boston, 1994).
6. Linkevich A.D. Nonlinear Phenomena in Complex Systems, 2000, 3, 135.
7. Linkevich A.D. Mathematical Methods and Models for Investigation of Neurodynamical Mechanisms of Cognitive Processes (IEC/PSU, Minsk/Novopolotsk, 2001).
8. Daubechies I. Ten Lectures on Wavelets (SIAM, Philadelphia, 1992).
9. Kleinberg J. Proc of the Ninth ACM-SIAM Symposium on Discrete Algorithms, San Francisco, 1998, P. 25 – 27.
10. Kangas J.A., Kohonen T.K., Laaksonen J.T. IEEE Trans. Neural Networks, 1990, 1, 93.

Received 28.06.2012