The Social Spiders in the Clustering of Texts: Towards an Aspect of Visual Classification

Reda Mohamed Hamou, Department of Mathematics and Computer Science, Dr Moulay Tahar University of Saïda, Saïda, Algeria
Abdelmalek Amine, Department of Mathematics and Computer Science, Dr Moulay Tahar University of Saïda, Saïda, Algeria
Ahmed Chaouki Lokbani, Department of Mathematics and Computer Science, Dr Moulay Tahar University of Saïda, Saïda, Algeria

ABSTRACT

In this paper the authors experiment and test a new biomimetic approach based on social spiders to solve a combinatorial problem ie the automatic classification of texts because a very large data stream flows and particularly on the web. Representation of textual data was performed by a method independent of the language ie n-gram characters and words because there is currently no method of learning that can directly represent unstructured data (text). To validate the classification, the authors used a measure of evaluation based on recall and precision (F-measure). During the experiment, the authors found a powerful visualization tool in social spiders that they exploit to make visual classification.

Keyword: Biomimetic Methods, Clustering, Reuters 21578, Social Spiders, Text Mining

INTRODUCTION AND BACKGROUND

Given the problems faced by the supervised classification such as the need of many human, the initial classification should be reviewed when the number of documents increases, according to (Tan, 1999) where about 80% of documents are in text format. This huge volume of unstructured or semi structured gives rise to an act to find relevant information more difficult to achieve, that creates a problem known as the problem of information overload (Chen, 1997). The techniques and tools for knowledge discovery in texts (KDT; Feldman, 1995) or simply text mining (Tan, 1999) are being developed to address this problem. One such technique is clustering, a technique for grouping similar documents of a given collection by helping to understand its contents (Jain, 1999; Willet, 1988). One of his goals is to similar documents in the same group and placing documents in various different groups. The assumption is that through a process of clustering, similar

DOI: 10.4018/jalr.2012070101
objects remain in the same group based on the attributes they have in common. This assumption is known as the hypothesis of cluster described by Rijsbergen (1979).

In this paper, we introduced a new model from nature in this case the social spiders to solve a critical problem of data mining or text mining more precisely what the clustering of data by which we hope to contribute about solving this problem because the grouping or clustering of textual documents, especially Web pages, is one of the challenges of current research.

State Of The Art

A well designed clustering algorithm generally follows the four phases of design: data representation, modeling, optimization, and validation (Buhmann, 2003). Phase representation of the data structures predetermine what kind of cluster can be found in the data. Based on data representation, the modeling phase defines clusters and the criteria that separate the desired group structures of these unwanted or unfavorable. In this phase, a measure of quality, which can be either optimized or approximate when searching hidden structures in the data is produced (see Figure 1).

Data Representation

Learning algorithms cannot deal directly with unstructured data such as text for this you have to go through a stage of indexing which is simply a representation of the text in numeric vector. But this step is very delicate and very important because a poor representation of data results in a misclassification of such data. The digital components of the vector associated with the text are the words of the text document that is associated with a weight. In this way learning algorithms directly uses the vectors that represent text (Hamou, 2010).

The choice of terms is to convert each document into a vector $v_{di} (W_{1i}, W_{2i}, ..., W_{Tj})$ where $T$ is the number of all terms that appear at least once in the corpus. $W_{ki}$ weight indicates the importance of the term $t_k$ in document $d_i$.

In our study the representations used are those bags of words that represent the simplest representation of textual records and that was introduced in the vector model and the n-grams that operate independently of the language of the corpus used. The principle of representation of the first approach is simply to turn each text into a vector where each component represents a word. With this approach, documents are represented by vectors of very large which makes the learning algorithms very difficult to use. For this we had to make a reduction in size as we will explain in the next section.

Once the choice of components of the vector representing a text $d_i$ is made, it must be consolidated for each coordinate of the vector $v_i$ by calculating the weight of each term $W_{ki}$ as follows:

- **Binary**: “0” to indicate that the term does not exist in the document and “1” to indicate that the term exists in the document.
- **For Instance**: $W_{ki}$ equal to the number of times the word appears in the document.

Coding methods are based on two observations:

- Plus the word is common in a text; it is more related to the topic of this text.
- Plus the word is common in a corpus, unless it will be used to differentiate between documents.

TF-IDF weighting (“Term Frequency-Inverse Document Frequency”) is used to calculate the weight of a term. This gives more weight to words that appear often within the same text, which corresponds to the first observation. But its peculiarity is that it also gives less weight to words that belong to several documents and corresponding to the second observation.

Encoding TF-IDF does not correct the document length, coding for this TFC is similar to that of TF * IDF, but it fixes the lengths of the texts by cosine normalization, not favor longer documents. It is formalized as follows:
Related Content

On the Ordering Property and Law of Importation in Fuzzy Logic
[www.igi-global.com/article/ordering-property-law-importation-fuzzy/46025?camid=4v1a](www.igi-global.com/article/ordering-property-law-importation-fuzzy/46025?camid=4v1a)

TCP Symbiosis: Bio-Inspired Congestion Control Mechanism for TCP
[www.igi-global.com/chapter/tcp-symbiosis-bio-inspired-congestion/58303?camid=4v1a](www.igi-global.com/chapter/tcp-symbiosis-bio-inspired-congestion/58303?camid=4v1a)

Competitive Advantage of Geographical Clusters
[www.igi-global.com/chapter/competitive-advantage-geographical-clusters/21137?camid=4v1a](www.igi-global.com/chapter/competitive-advantage-geographical-clusters/21137?camid=4v1a)
The Clonal Selection Principle for In Silico and In Vitro Computing
Vincenzo Cutello and Giuseppe Nicosia (2005). Recent Developments in Biologically Inspired Computing (pp. 140-147).
www.igi-global.com/chapter/clonal-selection-principle-silico-vitro/28326?camid=4v1a