Classifying texts through time : a complexity science approach to dynamic web page filtering

Paolo Allegrini, Simonetta Montemagni, Vito Pirrelli

Classifying texts through time

A complexity science approach to dynamic web page filtering

Paolo Allegrini, Simonetta Montemagni et Vito Pirrelli
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche,
Via Moruzzi 1 56124 Pisa Italy.
{allegrip,simo,vito}@ilc.pi.cnr.it

RÉSUMÉ. *Cet article présente une contribution du projet POESIA, consacré au filtrage des contenus illicites sur Internet. Dans ce contexte, nous présentons une méthode originale, utilisant des principes de la théorie de l'information, pour identifier automatiquement des contenus érotiques.*

ABSTRACT. *This paper reports our contribution to the POESIA project, aimed at the automatic filtering of illicit contents on the Internet. In this context, we present an original text classification system, based on information theory principles, which is able to detect and filter with a high accuracy pornographic texts.*

MOTS-CLÉS :*filtrage de texte, classification automatique*

KEYWORDS: *text filtering, classification*

Actes de CIFT 2004, pages 3 à 5

4 Actes de CIFT 2004, pages 3 à 5

1. Summary

POESIA (Public, Open-source Environment for Safer Internet Access), a European project funded under the EU Internet Action Plan, developed an advanced internet filtering system, with the aim of providing safe and educationally appropriate internet access for young people. The POESIA filtering system is Open-Source and combines standard filtering methods, such as positive/negative URL lists, with more advanced techniques, such as image processing and NLP-enhanced text filtering (for English, Italian and Spanish). All POESIA text filters offer both ``light'' and ``heavy'' filtering modes, where the former is simpler and faster, and the latter is computationally more expensive (Hepple et al., 2004). Light filtering, which uses little NLP, provides rapid accept/reject decisions for straightforwardly classifiable pages. For other pages, heavy filtering, making greater use of NLP, is invoked to provide more sensitive detection of content indicators. In the present talk, we shall focus on heavy filtering for Italian only, where the filtering domain is pornography.

For the purposes of Italian heavy filtering of erotic web pages, we suggested tackling a classical text classification task through the comparatively novel methodology developed for the scientific study of complex systems. A complex system consists of a typically very large number of parts, presenting an equally large number of mutual nonlinear interactions. In some cases, elementary units cannot even be identified, and the definition of a constituent part may vary depending on the specific scale of observation. Such a complex system is said to show scaling relations, namely a layering of time and space scales, with interactions spanning the whole range of scales. This kind of complexity has strictly been related to evolution. The pioneering work of Nobel laureate Ilya Prigogine shed light on the role of time in causing a negative entropy in what he called dissipative structures, namely quasi-static systems, far from equilibrium (Prigogine, 1984). The prototype of this kind of systems is a living being, with its inextricable metabolic network (Kaufmann, 1995). The intuition that human language is a complex (and “auto-catalytic”) system is gaining consensus in the scientific literature. An associative network of semantically cognate words (Allegrini et al. 2003), to give but one example, tends to exhibit the well known Zipf’s law, with a large number of isolated words, and a small number of densely interconnected words. In between, inverse-power laws are witnessed, that can be seen as the mathematical fingerprint of “scale invariance”.

For our present concerns, we operated on the morpho-syntactically tagged and lemmatized text of erotic web pages. Filtering is based on recognition of any of about 2400 domain-relevant lemmata in the texts to be filtered. In the pornographic domain, many domain-relevant lemmata are “ordinary” words, that can however be used with sexual innuendoes. Category learning uses an entropy-based classifier, CASSANDRA (Complex Analysis of Sequences via Scaling AND Randomness Assessment), which computes the rate of information increase generated by salient

lemmata occurring in an input document. Shannon's information functional S for the probability P(x;l) of finding a fixed number x of lemmata in a sliding window of length l was recently shown to give a maximum entropy change dS/d(log l) when domain-salient lemmata are selected (Allegrini et al. 2004). The agreement between the model and the data is so remarkable that we can take a time-bound abrupt information change in a comparatively short text span as a reliable indicator that the document in question is pornographic.

On a theoretical level, this seems to suggest the operation of a generative device which, when running in text production, performs a random walk on a tightly interconnected word network. The network mostly consists of common-or-garden lemmata. However, from time to time, the random walker finds itself in a sub-network of sexual expressions, which are thus repeatedly generated at a close distance in a comparatively short text span. The originality of our dynamic text classifier thus rests on the use of time as a significant classificatory dimension, typically neglected by more popular bag-of-words approaches to the problem. Our classifier is OpenSource and freely downloadable from the POESIA website (http://www.poesia-filter.org).

2. Bibliography

Allegrini P., Girgolini P., Palatella L., 2004, Intermittency and scale-free networks: a dynamical model for human language complexity, *Chaos, Solitons and Fractals*, vol. 20, p.95-105.

Hepple M., N. Ireson, P. Allegrini, S. Marchi, S. Montemagni, J. M. Gomez Hidalgo, 2004, NLP-enhanced Content Filtering within the POESIA Project, in Proceedings of LREC 2004, Lisbon, Portugal.

Kaufmann, Suart A., 1995, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*, Oxford University Press

Prigogine I, Stenger I. *Order out of chaos: man 's new dialogue with nature*. New York: Bantam Books;1984.