Abstract

This book presents results of over two years of joint research by linguists, programmers and scientists of Fido Intelligence, a hi-tech company located in the Pomeranian Science and Technology Park in Gdynia, and the Department of Knowledge Engineering, at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, in the R&D project "ISPAD – and intelligent system for acquiring and analyzing data of social networking Internet services" funded in 2009–2011 by the Polish Ministry of Higher Education and Science, and the national Center for Research and Development. The first ISPAD prototype has been deployed successfully in a target business environment of Fido, a leading Polish company developing advanced technologies for natural language processing.

The underlying concept of ISPAD is a knowledge base containing language patterns, characterizing specific language forms used by various groups of Web authors publishing texts in Polish. Such patterns are essential for efficient information search and retrieval, as the texts are written as collections of colloquial phrases and expressions in a rapidly evolving language. Depending on the terminology, specific word formation, subjects covered, types of events commented, etc., one may find there hobbyists or professionals from various disciplines (eg., rock musicians), celebrities or subcultures (or even sects), private citizens or government agencies, and so on. Generic sentence patterns for each of their specific languages, supported by specialized dictionaries, and built-in machine learning mechanisms, allow ISPAD to effectively localize Web pages of interest and to analyze their textual content. Definition of language patterns, and selection of machine learning rules, requires, of course, serious research and preparatory works by linguists; for the first prototype of ISPAD they described law offensive contents, especially promoting violence, racism, acts of conspiracy and terrorism. Upon loading the system knowledge base with relevant patterns and configuring it, ISPAD is able to explore the Web, recognize the desired content, and make it progressively faster and more effective, owing to its embedded machine learning capability.

The book is carefully organized to take the Reader through language analysis, Web search engines, pattern recognition and machine learning in a how-to-do fashion that lays the ground for acquisition of information from the Internet. Chapter 1 introduces the concept of grammars, formal languages and their automatic analysis. Chapter 2 concentrates on the notion of finite state automata, their design principles and transformations that are necessary for creating and maintaining dictionaries. Chapter 3 builds upon these concepts and introduces a multi-layered model of natural language processing, which goes beyond the limits of the formal language analysis described in previous chapters. Chapter 4 takes a concrete view of the Polish language and defines the principles for designing language patterns, essential for the ISPAD functionality. Chapter 5 introduces the concept of Web search engines. A common feature of their query languages is the lack of in-depth analysis of language phrases, however they provide a useful vehicle for supporting users of practically any Web browser. Chapter 6 complements development of preceding chapters with graphical content recognition and analysis methods. Although the current ISPAD prototype in not yet able to analyze texts represented in a graphical form, eg. logos or Flash animations, such a functionality is planned for its future versions. Chapters 7 and 8 introduce relevant models and algorithms that are necessary to make the system intelligent and self-learning. Notions of feature selection and extraction, their related quality metrics and criteria are introduced, to the extent required to represent the analyzed page content for machine learning algorithms, explained and recommended to speed-up and increase accuracy of page content retrieval with language patterns. Chapter 9 wraps-up the book with a case study - the first ISPAD prototype; it describes its architecture, principal components and associated tools, along with results collected during experiments with the prototype in its target environment.