

# JAVA-SERVLET TECHNOLOGY FOR BUILDING NEW WEB DOCUMENT CLASSIFIERS

Maya Dimitrova<sup>1</sup>, Ivan Terziev<sup>1</sup>, Plamena Andreeva<sup>1</sup>, Petia Radeva<sup>2</sup> and Joan Jose Villanueva<sup>2</sup>

<sup>1</sup>Bulgarian Academy of Sciences

[dimitrova@icsr.bas.bg](mailto:dimitrova@icsr.bas.bg)

<sup>2</sup>Autonomous University Barcelona

[petia@cvc.uab.es](mailto:petia@cvc.uab.es)

**Abstract:** The paper presents the working architecture of an existing module for web search, classification and presentation of documents with different lexical features. Unlike the existing search engines, the module discriminates aspects of the style of the document – readability, explanation, illustrations and summarization. The visualization component enhances perceptual support of these features when retrieval of searched documents take place. We present the current state of research on the classifier and give examples from the educational, medical and environmental domains.

## 1. INTRODUCTION

Servlet technology has become the leading web development technology in the recent years due to its safety, portability, efficiency and elegance in design. The servlets are specialized in browsing the web (web documents and various databases) for data mining, information extraction and meaningful presentation inside the web browser. More important, their power is in online analysis, filtering and structuring of the retrieved information. The servlets are object-oriented and therefore highly modular and multifunctional. We are currently investigating their efficiency for classification and meaningful representation of various aspects of web documents, including text style and web genre, which have not been implemented in web document search and retrieval systems. The main application is general and specialized search engine design, summarization systems on the web and decision support system design.

## 2. TASK DEFINITION

The research task is to propose *linguistic* features of web documents which can be discriminative for the specific style or genre in which the web page is designed. In particular, we are interested in features which can account for readability, explanation, illustrations and summarization. These 4 aspects of web genre are our long-term research goal. Currently, we have dealt with features that can be extracted from vectors of *long* English words contained in the document. The second group of features are related to the natural language frequency of these words. For this purpose a lexicon/thesaurus was included, which contains 490 high frequency *long* words (above 60 per 1 million). Other discriminative features are – relative presence of adjectives, gif files and technical elements e.g. HTML tags, special math symbols, etc. The motivation for our *lexical* approach to the analysis of web documents is twofold: first, by using only part of the English language lexicon the document feature vector is about twice shorter than the standard “bag-of-words” (BOW) approach, massively used by the search engines, where *all* of the words are elements of the feature vector, including common *short* words (like can, must, will, etc.), pronouns (there, them, etc.), articles (a, the, etc.) and prepositions (in, on, at, etc.). The second reason is that the frequency aspect of the text has not been rigorously explored to date and is appropriate for our aim to diagnose readability, explanation, illustration and summarization. These implemented aspects can help the user find the web document that is most appropriate to his knowledge and preferences/style.

### 3. SERVLET ARCHITECTURE WITH AN IMPLEMENTED WEB DOCUMENT CLASSIFIER

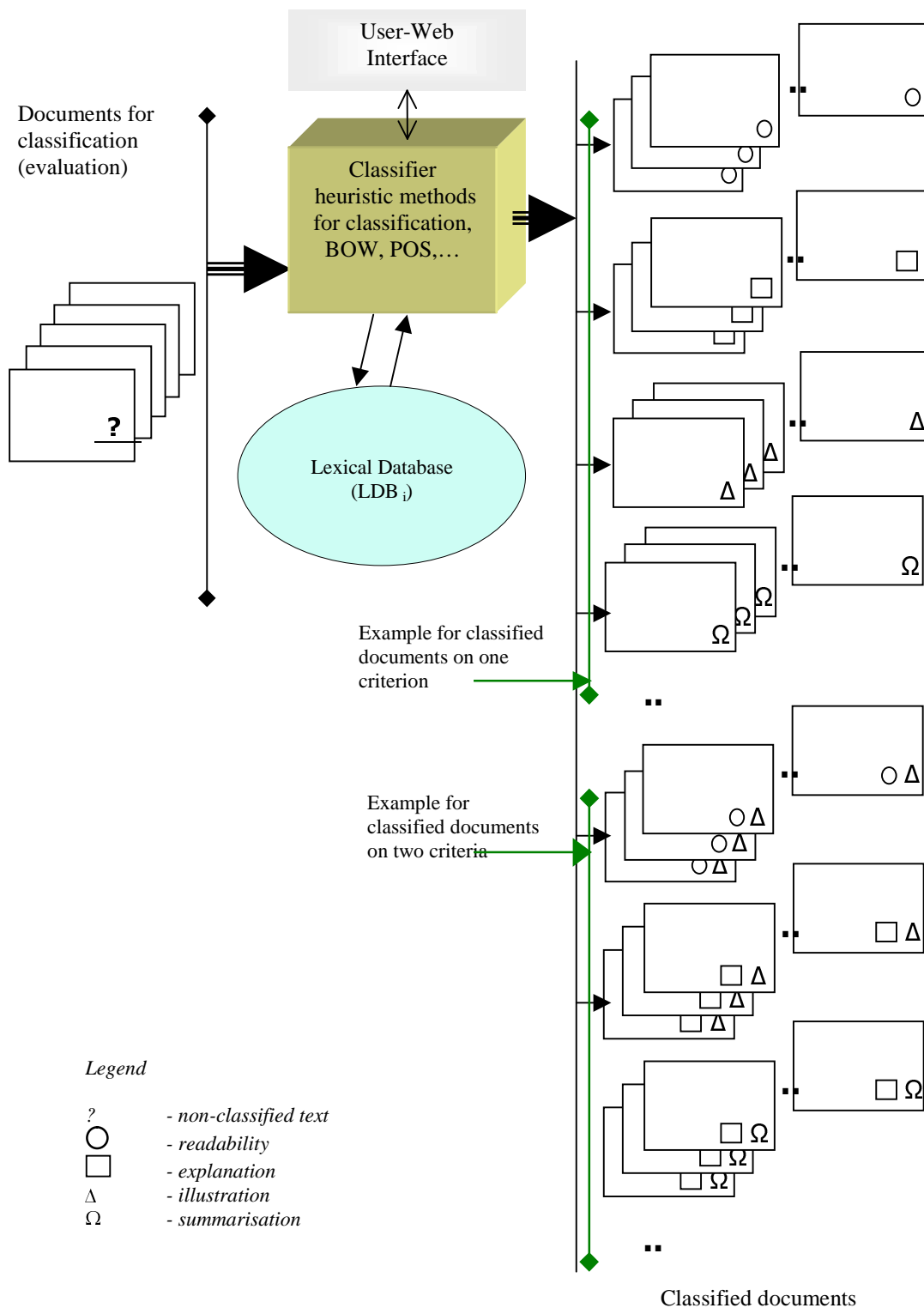


Figure 1. Main servlet architecture with implemented classifier and lexical database

The main servlet architecture with implemented classifier of web documents is given in figure 1. The servlet is browsing the web in response to the user query. It extracts the Urls of the gathered documents, opens them one by one and classifies them according to the pre-specified criteria. One of the criteria is the lexicon check for the presence of *long* English words (4). For this, first it tokenizes the text into separate words/tokens and compares all the words from the document (which are the elements of the document feature vector) with those from the lexicon.

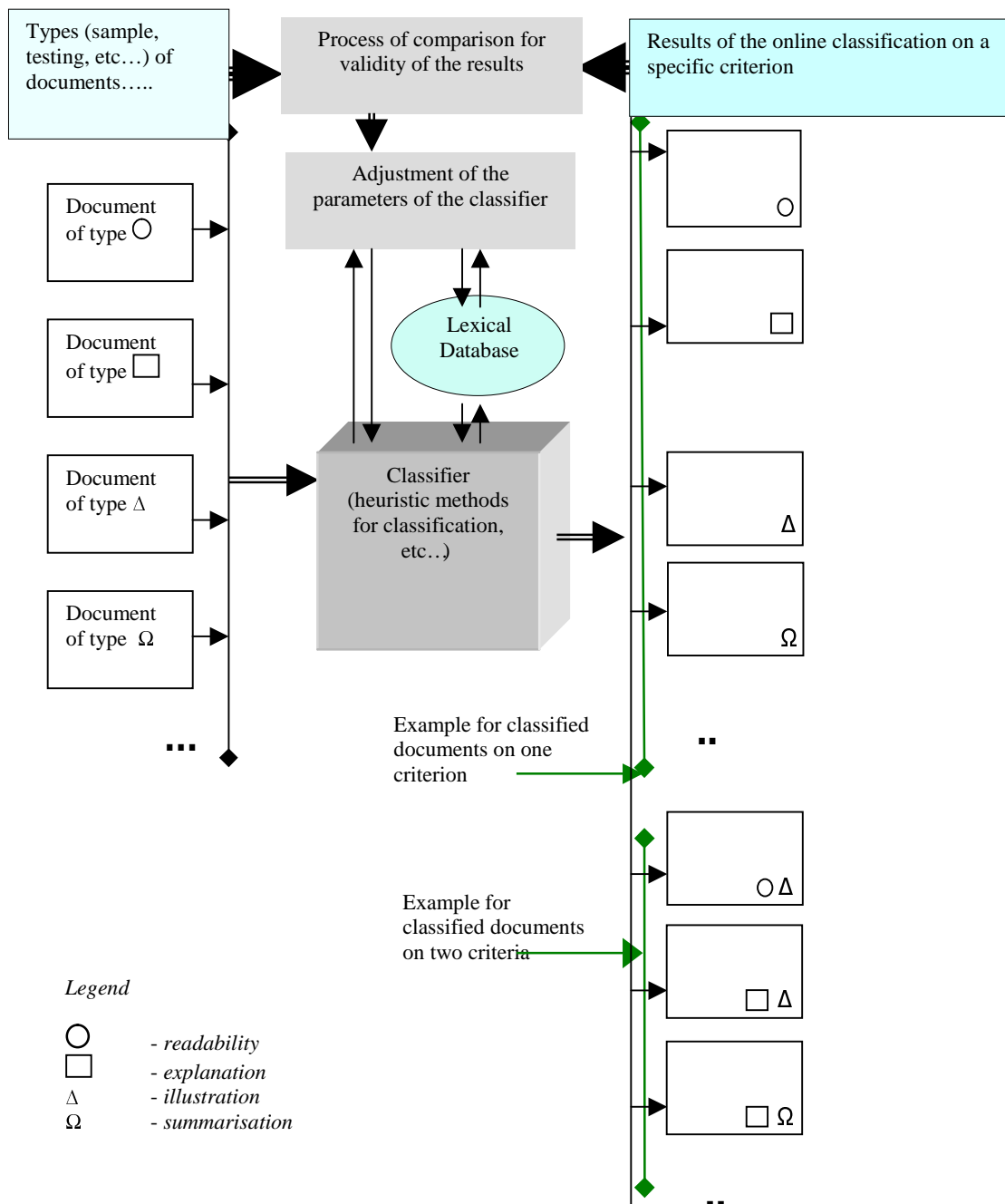


Figure 2. Servlet architecture with an on-line evaluation module

The amount of *long* words within a given frequency range is one of the criteria for document classification. Another criteria are the presence of gif files, the length of the text itself, the presence

of HTML tags, mathematical symbols and so on. The online performance of our servlet is with good speed due to the simplified form of the feature vector. At the moment, we are working on the development and the precision of the classifier as well as to test it in various thematic domains.

#### **4. CLASSIFIER EVALUATION AND ADAPTATION**

The next step in the implementation of the classifier is to include a module for self- adaptation as shown in figure 2. This feature will logically follow the extensive tests we are currently performing on the work of the classifier. The tests include explicit and implicit evaluation of the work of the classifier by web users as well as formalization of the way the servlet-based classifier adapts its own parameters (1,2,3). An important aspect of the evaluation phase is to make sure the classifier performs well across different domains. It was tested on the following queries: “Java Servlets”, “Implicit Memory”, “Pearson Correlation” and “Neural Networks”. The study has shown that the classifier can be useful in general university education. Next, we investigated its utility in translator work in medicine, e.g. orthopedics and cardiology. We are currently investigating its work in environmental and ecological domains like assessment of documents containing information about pollution indexes or general geographic/demographic information.

#### **5. CONCLUSIONS AND FUTURE WORK**

The complexity of the task of building new web document classifier comes from the need to fine-tune our classifier parameters as well as to define the common contribution of several parameters simultaneously. We are currently working on implementation of advanced and sensitive formal modeling components for improvement and self- adjustment of the classifier. This task is promising and feasible, because it relies on optimized number of elements of the document feature vector, which is justified by the fast on-line servlet implementation. This work is also related to the current research on user adaptive web interface and to modern fast and user friendly decision support systems in complex areas like education, medical diagnosis and environmental protection.

#### **6. Acknowledgement**

This work was partially supported by grants SFI/01/F.1/C015 from Science Foundation Ireland, and N00014-00-1-0021 from the US Office of Naval Research, by a scholarship granted to the first author by the Universitat Autònoma de Barcelona, 2004, the Joint research project of CVC-UAB and the ICSR-BAS “Intelligent interface to a cardiological diagnosis (CAD) system” and a Fifth Framework Project on trans-border decision support for environmental problems. We would like to thank Dr. Nick Kushmerick from UCD, Ireland for his supervision and help with this work.

#### **REFERENCES**

1. Dimitrova, M., Kushmerick, N., Radeva, P. and Villanueva, J. J. (2003). User assessment of a visual Web genre classifier, Proc. 3rd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP), Benalmadena, Spain, 886-889
2. Dimitrova, M. (2003). Cognitive modelling and Web search: Some heuristics and insights, Journal "Cognition, Brain, Behaviour" Vol. VII, No 3, Pp. 251-258.
3. Dimitrova, M., Kushmerick, N., Terziev, I. and Gegov, A. (2004) Web Users and Web Document Classifiers: Emergent Cognitive Phenomena, To appear in Proc. 8<sup>th</sup> International ISKO Conference on Knowledge Organization and the Global Information Society, ISKO2004, London, UK
4. Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English*, Brown University Press. [[www.psy.uwa.edu.au/MRCDatBase/uwa\\_mrc.htm](http://www.psy.uwa.edu.au/MRCDatBase/uwa_mrc.htm)]

