Sofia University
St. Kliment Ohridski
Faculty of Mathematics and Informatics

Subject: Syntactical and semantic analyser of free written post addresses.

Student: Daniela Vasileva Slavkova
Master program: Distributed Systems and Mobile Technologies
Department: Information Technologies
Fac.№: M21301

Counselor : reader. doc. Tinko Tinchev

Defence of graduation date:29.06.2005

Annotation:

Bulgarian internet cards are unique service. They offer geographic, administrative and transport cards of Bulgaria and large bulgarian towns. Searching for name or post code is an esteemed possibility; it economizes much time and efforts. Finding each one street or quarter or whole address in Sofia is big convenience.

Finding information for requested user address is special case of geographic analysis. In practice it is observed when some post address is written in computer, the sequence and writing of name of towns and villages, quarter and street are not correct and precise. This is problem for theirs analyzing and identification. Also the analysis is hindered from size of data base with all interesting addresses. Therefore system which confirms real existing address or finds its best approximation has practical meaning.

The purpose of this Diploma paper is to develop a program system which normalizes the free written post addresses and suggests a correct writing. When the request for street is made for example, the system returns the most suitable candidates for this street, existing in available data base. Some of them are probably the requested ones. The realization has to normalize the separate part of the address like independent parts.

The development process starts with examining and analyzing of the users' demands. After the forming of the users' demands, moves towards choosing the technologies used in its realization.

The system is a server application, which realizes the access to the data base with real addresses. It has the task to seek fast in the data base with real addresses. This is realized by single time building of a dictionary from all post addresses from data base (exactly, several dictionaries for every part of the post address like streets, district and so on). The dictionary is realized like minimal deterministic automata. It is loaded in computer memory which allows a sensibly fast seeking in the data base. The problem for finding the appropriate candidates which are close to the requested is known like "Altarnatte Pattern Matching". For criteria for proximity of the two words is used Levenshtein distance.

When the factual realization is completed there follows system testing and mistake elimination. At completion of this stage the developing acquires a form eligible for utilization.