Софийски университет
"Св. Климент Охридски"
Факултет по математика и информатка

Тема: Синтактичен и семантичен анализатор на свободно изписани пощенски адреси

Дипломант: Даниела Василева Славкова
Магистърска програма: Разпределени системи и мобилни технологии
Катедра: Информационни технологии
Фак.№: M21301

Дипломен ръководител: доц. д-р. Тинко Тинчев

Дата на защита:29.06.2005г.

Анотация:

Българските интернет карти са уникална услуга. Те предлагат географски, административни и транспортни карти на България и големите градове. Търсене по име или пощенски код е една ценна възможност, която спестява много време и усилия. Намирането на всяка една улица или квартал, или по-общо – на всеки един адрес в София – е голямо удобство.

Намирането на информация за заявен от потребител адрес е един специален случай на географски анализ. В практиката се наблюдава, че при въвеждането в компютър на пощенски адрес не се спазва реда на попълване и прецизно изписване на имената на селища, кварали, улици и другите съставни части. Това пречи на по-нататъшното им анализиране и разпознаване. От друга страна анализът на съставните му части е препятстван от размера на базата от данни за претърсване с номенклатурите на действителните адреси. Следователно, система, която от една страна потвърждава дали даден адрес съществува, а от друга, ако не съществува, какви са най-добрите негови приближения, има практически смисъл.

Дипломната работа разработва алгоритъм и реализираща го програмна система, такива че при направен вече анализ на съставните части на изписания адрес, го нормализира и предлага правилно изписване на същия. Това означава, че при направена заявка за улица, например, системата връща най-подходящите кандидати за тази улица, съществуващи в налична база от данни, един от които

вероятно е имал предвид потребителят. Реализацията има за цел да нормализира отделните части на адреса като самостоятелни единици. Процесът на програмиране стартира с проучване и анализиране на потребителските желания. След формиране на потребителските желания преминава към избор на технологиите, използвани в реализацията.

Разработката представлява сървър приложение, което осъществява достъпа до базата от данни с действителните адреси. То изпълнява задачата за бързо обхождане и претърсване на данните. Първоначално се решава проблема за намирането на алгоритъм за бързо обхождане и претърсване на базата данни с действителните адреси. Решението на тази задача, е чрез еднократно построяване на речник от всички пощенски адреси в базата от данни (по-точно няколко речника за всяка от частите на пощенския адрес като улица, квартал и други) във вид на минимален детерминиран автомат и зареждането му в паметта на компютъра, което позволява разумно бързо обхождане на базата от данни. Второ, едновременно с обхождането се колекционират подходящите кандидати в случай на недействителен адрес. Под подходящи кандидати се разбира близки до заявката записи от базата. Това е проблем, който в практиката е познат като "Altarnatte Pattern matching" – приближено съвпадане на шаблони. За мярка на близост между две думи се използва растояние на Левенщайн.

Когато действителната реализация е завършена, е извършено системно тестване и отстраняване на грешките. След приключването на тази фаза проекът придобива форма, подходяща за използване.