Symbolic Data Structure for Postal Address Representation and Address Validation Through Symbolic Knowledge Base

P. Nagabhushan[1], S.A. Angadi[2], and B.S. Anami[3]

[1] Department of Studies in Computer Science, University of Mysore, Mysore*
[2] Department of Studies in Computer Science, University of Mysore, Mysore
and Department of Computer Applications BEC, Bagalkot
[3] Department of Computer Science & Engineering, BEC, Bagalkot
{pnagabhushan, anami_basu}@hotmail.com,
vinay_angadi@yahoo.com

Abstract. The postal address data and the domain information for address validation contain qualitative, numeric, interval and other types of data. The efficient processing of such data required for postal automation needs a robust data structure that facilitates their storage and access. A symbolic data structure is proposed to represent the postal address and the information relevant for validating the postal address is stored in a newly devised symbolic knowledge base. The symbolic representation gives a formal structure to the information and hence is more beneficial than other representations such as frames, which do not reflect the structure inherent in the domain knowledge. The process of postal address validation checks the different components of the postal address for consistency before using it for further processing. In the present work a symbolic knowledge base supported address validation system is developed and tested for about 500 addresses. The system efficiency is observed to be 95.6% in validating the addresses automatically.

Keywords: Postal Address validation, Symbolic object, knowledge base, Frames.

1 Introduction

Postal automation aims at rendering the postal service that delivers mail to the doorstep of the addressee, efficient. There is a spurt of activity in postal automation area in recent times [1-4]. The different aspects of postal services that need to be automated are identified in [3]. The most important step in postal automation is interpretation of the destination address. Apart from the pattern recognition and image processing activities for reading the postal address, one of the major tasks is to find a generic data structure to store all types of addresses. The data structure is to be further employed for various sub tasks of postal automation, particularly mail sorting, such as address validation, address component identification, etc [9,11]. Some works in this direction are reported. An algorithmic prototype for automatic validation of postal

* Work carried out during sabbatical at Amrita Vishwa Vidyapeetam , Coimbatore.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 388 – 394, 2005.

addresses is presented in [5]. A knowledge based approach to generation of destination postal codes from the addresses is presented in [6]. A truthing, testing and evaluation mechanism for postal addresses is proposed in [7].

Most of the postal automation efforts are seen in the countries that have standard address formats [8], where the correctness and consistency of the address is not in question. The same is not true in India, where the destination addresses are written using any known information about the geographical location of the addressee. Typical unstructured descriptions of the addresses include, Near Playground, Besides City Hospital etc. Also the destination place might be specified by any of the alias names a place has, for example, Madras/ Chennai, Calcutta/ Kolkata etc. Some times even the place name might be mis-spelt like, Bangalore as Bangloor, Mysore as Mysooru and the like. Symbolic data objects are very much suitable to model the concepts of the real world [12] such as those described by the postal address.

The components of the postal address (especially in the Indian context) may not be consistent especially with respect to the Postal Index Number (PIN), hence there is a need to validate the address for its correctness before it is used for sorting and distribution. This paper presents a symbolic knowledge base supported methodology for automatically validating the postal address and sorting of postal mail using postal address as a symbolic object.

The paper is organized into 5 sections. The section 2 describes postal address as a symbolic data object. Section 3 presents the knowledge base supported automated solution to postal address validation and sorting. It also describes the symbolic object knowledge base. Section 4 presents the experimental results and their analysis. Section 5 gives the conclusion.

2 The Postal Address as a Symbolic Object

The postal address contains many fields, all of which may not be present in every postal address. Some of these fields are qualitative such as Addressee name, Care of name and other fields may be numeric such as house number, road number and PIN. Although these data are numeric, most often their role in the address could be non-numeric in nature. Symbolic objects offer a formal methodology to represent such type of Information. Symbolic objects are extensions of classical data types. Symbolic objects can be of three different types, Assertion Object, Hoard Object and Synthetic Object. An assertion object is a conjunction of events pertaining to a given object. An event is a pair which links feature variables and feature values. A Hoard object is a collection of one or more assertion objects, whereas a synthetic object is a collection of one or more hoard objects [10,12,13]. The postal address object is described as a hoard object consisting of three assertion type objects as described in (1).

POSTAL ADDRESS OBJECT= {[Addressee],[Location],[Place]} (1)

Each of the assertion objects describes an important component of the destination address. The Addressee specifies the name and personal details of the mail recipient; the Location specifies the geographical position of the mail delivery point and Place specifies the city/ town or village where the mail recipient resides. Each of these assertion objects is defined as a collection of many events. The features (address fields) of the different assertion objects are listed in (2),(3) and (4). Each of the feature describes some aspect of the object and all the features together completely specify the assertions objects namely, Addressee, Location and Place. However, certain features remain missing in a typical postal address because they are not available.

[Addressee=(Addressee Name)(Care of Name)(Qualification)(Profession)
(Salutation)(Designation)] (2)
[Location=(House Number)(House Name)(Road)(Area)(LandMark)
(PBNo)(Firm)] (3)
[Place=(Post)(Taluk)(District)(State)(Place)(PIN)(Via)] (4)

A typical postal address and its symbolic object representation is given in Table 1.

Table 1. A Typical Postal Address Object

Postal Address	Symbolic Representation
Shri M.M.Patil Lakshmi Extension Gokak-591307 Belgaum Karnataka	PostalAddressObject={[Addressee=(Salutation=Shri), (AddresseeName=MMPatil)],[Location=(Area=LaxmiExtension)], [Place=(place=Gokak),(PIN=591307),(District=Belgaum), (State=Karnataka)]

The symbolic object defined for representing a postal address can be further used to perform various sub tasks of postal automation such as address component identification, address validation etc.

3 The Knowledge Base Supported Symbolic Data Analysis Approach to Postal Address Validation

The proposed system employs symbolic data analysis techniques for address validation and a predefined procedure for mail forwarding in the dispatch sorting office. Every mail specifies the destination location (area) by providing area and place names as well as PIN code. The validation process checks for the correctness of the place names and area names and their mapping to PIN code. The system further corrects the PIN code of the address if there is any inconsistency in the information conveyed by the area/place name and PIN code. If the mail does not contain PIN code, and the other address components are validated, then the PIN is generated to the extent possible depending on the information available at the sorting office. The system works on the premise that the destination place and geographical area are probably more correct than the PIN code. The validated/corrected address is then employed for mail sorting at the dispatch sorting office.

The symbolic knowledge base supported address validation and sorting system processes the input postal address object and validates/ corrects the PIN code for further sorting of postal mail, Figure-1 gives the block diagram of the proposed system. The system comprises of an inference engine consisting of two processes namely address validation and mail sorting. The address validation module checks for consistency among the various address components which are part of the Place and Location objects, whereas the mail sorting module sorts the mail using the rule base employed by the mail sorting office.

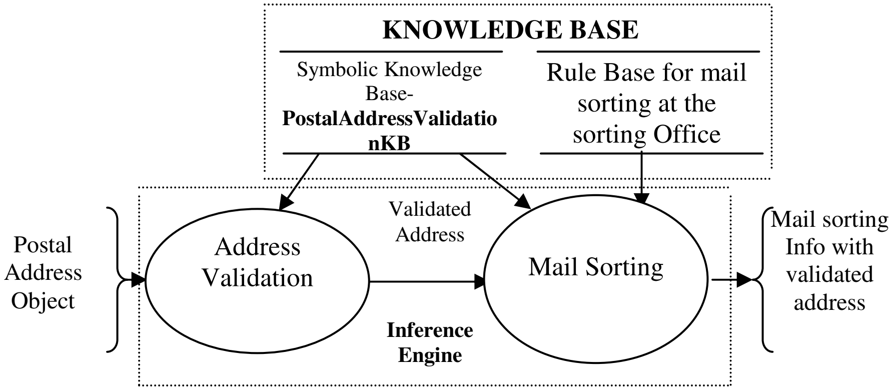

Fig. 1. Symbolic Knowledge base supported Address Validation and Sorting System

Further for validating the mail address the required local knowledge is stored in a knowledge base of symbolic objects. The rule base used for sorting is based on the mail sorting rules of India Posts and has been implemented as a set of external rules separate from the validating knowledge base. The structure of the knowledge base devised is presented in section 3.1.

3.1 The Symbolic Postal Knowledge Base for Dispatch Sorting

The information required for address validation such as various forms of place names, area names, PIN codes etc. and their relationships are to be stored in the knowledge base. A frame structured knowledge base for address validation is proposed in [11]. Though the frame structure helps in storing the variety of information required for postal address validation it does not explicitly bring out the relationships between different pieces of knowledge and hence the inherent structure in the domain information is lost. Also the frame-structured knowledge base has resulted in complex validation procedure because of the complexity in data. The symbolic knowledge base offers a more systematic knowledge representation technique for storing the domain knowledge. The validation procedure is also made easier because of the structure of the knowledge base. A knowledge base of symbolic synthetic object PostalAddressValidationKB has been devised for the purpose. The synthetic object comprises of two hoard objects namely, PLACE_DETAILS and AREA_DETAILS. The knowledge embedded in the two hoard objects is employed for validation depending on the context. The developed knowledge base stores information about various forms of place names, area names, PIN codes, places to which direct bags are closed and the like which are needed for address validation. The symbolic data object that represents the structure of the knowledge base is given in Figure-2. The content of the knowledge base is sorting office dependant, but the proposed structure can be employed by any sorting office.

PostalAddressValidationKB={ {Place_Details=[Icity=((Name)(Length)] [States=((Name)(2digpin))] [Nplaceid=((Placename)(Length))(Npid))] [Nplace=((Npid)(Name)(Pin)(Forwardingplace) (Forwadingplacepin))] [Distid=((Distname)(Length))(Distid))] [Dist=((Distid)(Name)(Pin)(Forwardingplace) (Forwadingplacepin))]	[Dbplaceid=((Placename)(Length))(Placeid))] [Dbplace=((Placeid)(Name)(Pin)(Sdflag) (Bagtype))] [Oplace=((Place_Id)(3digitpin))] [Coveredplace=((Place_Name) (Covering_Place)(Coveringplaceid))]} {Area_Details=[Areainfo=((Areaname) (Dpoid))] [Dpo=((Name)(Dpoid)(Pin) (Placeid))] [Plpin=((Plid)(Plpins))}]}

Fig. 2. Symbolic Data Object Knowledge Base

3.2 Symbolic Data Analysis Technique for Postal Address Validation and Sorting

The process of address validation and sorting employs the symbolic knowledge base and rule base. The process involves systematic comparison of the information about the place in the postal address object with the feature values in place_details object of symbolic knowledge base, generating the PIN to the extent possible. The generated PIN is compared with input PIN if it exists and is corrected if necessary. If the PIN does not exist in the input, the generated PIN is added to the postal address. Further if the mail is destined to a place within the sorting district, then the input information about the location of the addressee is used to validate the PIN upto 6 digits using the information stored in the location_details object of the knowledge base.

4 Experimental Results and Analysis

The knowledge base for Bagalkot (a district place in Karnataka State of India) sorting office is built. The validation and sorting system is implemented in C language and is tested on a P-III machine with 128 MB RAM. The symbolic knowledge base is stored as a separate sorted flat file for every assertion object. The text strings forming the postal address object are input to the system. The system outputs the validated addresses after correction, if necessary. Some of the mails may require manual intervention for validation and sorting.

The system is exhaustively tested for sorting and validation for dispatch at Bagalkot sorting office and the results are summarized in Table-2. The results show that about 95.60% of the mail is either fully or partially validated by the proposed system, which is better than 90.80%[11] of the frame based system. Hence the methodology developed is sufficiently robust and can be used by any sorting office in India with appropriate knowledge base.

The algorithm failed to properly validate the mail destined to a place outside the sorting district, but had a name similar to the name of a place within the sorting district and had incorrect pin code, such a mail required manual intervention.

Table 2. Implementation results

Sl No.	Particulars	Observed Value	% of total test mails
1	The number of mails tested	500	100
2	The number of mails completely validated	390	78.00
3	The number of mails partially validated	88	17.60
4	The number of mails that required human intervention	20	4.00
5	The number of mails that cold not be resolved	2	0.40

5 Conclusions

The symbolic knowledge base supported method for address validation and sorting of mail for dispatch proposed in this paper is robust and takes care of mis-spelt, correctly spelt, abbreviated names and also alias names of places and geographical areas. The system validates or generates, as many PIN digits as are possible using the information extracted from the mail destination address, with the help of symbolic data analysis. The success of the address validation and sorting system is largely dependant on the efficiency of the address reading system. The address validation technique also finds application for bulk mailers in customer relationship management context and even to private courier service providers. The symbolic data analysis approach proposed here can be further used in automation of other sub tasks of mail handling.

Acknowledgement

The work is supported by grant from AICTE, New Delhi under the RPS scheme for the project titled “Development of Techniques for optimal postal mail distribution in the Indian context”, vide FNO. 8022/RID/NPROJ/RPS-60/2003-04 dated 22/3/04.

References

1. Kanehiro Kubota and Kazunari Egami, (1999), “ Technology trend of postal automation”, NEC Research and Development, Vol 40, No. 2, Spl Issue on Postal Technology, pp 127-136
2. Giovani Garibotto, 2002, “*Computer Vision in Postal Automation*” Elsag Bailey-TELEROBOT, 2002.
3. P.Nagabhushan, 1998, “ Towards Automation in Indian Postal Services : A Loud Thinking”, Technovision , Spl Volume, pp 128-139
4. Rangachar Kasturi, Lawrence O’Gorman and Venu Govindaraju,2002, “Document Image Analysis: A Primer”, Sadhana, Vol 27, Part 1, Febraury 2002, pp 3-22.
5. M.R.Premalatha and P. Nagabhushan, 2001, “*An algorithmic prototype for automatic verification and validation of PIN code: A step towards Postal Automation*”, NCDAR-2001, 13th and 14th July 2001, pp 225-233

6. M.R.Nagamani and P.Nagabhushan, 2003, "*Knowledge based approach to Determine the Destination Postal Code Through Address Block Extraction : A case study towards Postal Automation*", NCDAR-2003 held at PESCE, Mandya, 11th and 12th July 2003, pp 152-163
7. Srirangaraj Setlur, A Lawson, Venu Govindaraju and Sargur N Srihari,, 2001," Truthing, Testing and Evaluation Issues in Complex Systems", Sixth IAPR International Conference on Document Analysis and Recognition, Seattle, WA, pp 1205-1214
8. Universal Postal Union Address Standard, "FGDC Address Standard Version 2".
9. P.Nagabhushan,S.A.Angadi and B.S.Anami, 2005"A Knowledge base supported Inferencing of Address Components in Postal Mail" NVGIP-2005, Shimoga, 2nd and 3rd March 2005
10. Lecture Notes of short term course on symbolic and fuzzy approaches to data analysis, 21-26 April 1997
11. P.Nagabhushan,S.A.Angadi and B.S.Anami, 2004, "A Knowledge based Fast PIN code Validation System for Dispatch Sorting of Postal Mail", International Conference on Cognitive systems New Delhi, 14th and 15th December 2004
12. Edwin Diday,2000, "Knowledge Discovery from the Symbolic Data and the SODAS Software", PKDD 2000 workshop on Symbolic data Analysis, Lyon, 12th September 2000.
13. Bock H.-H. ,Diday E.,2000, "Analysis of symbolic Data", Heidelberg 2000