CHAPTER 8

SOFT COMPUTING MODEL FOR MAIL DELIVERY POINT MAPPING

8.0 INTRODUCTION

Sorting of mail for delivery to the addressee is referred to as delivery or distribution sorting. The delivery sorting involves mapping the destination postal addresses to the mail delivery points and sorting the corresponding mail to the beats of the postmen. The mapping of the postal address requires interpretation of the different components of postal address and their comparison with a knowledge base of delivery points to identify the mail delivery point. The comparison of the input address components with the knowledge base should ignore small errors in the writing of addresses for use in practical scenario. Very few works that refer to errors in postal address and methods for their processing are found in literature. Buckley et al (2000) present a multi-paradigm approach for exploiting knowledge about the structure for the purpose of extracting information from noisy / erroneous textual data. This approach combines aspects of database organization, clustering of records, fuzzy parsing and fuzzy retrieval to extract information from structured and labeled noisy data. The developed system is tested with erroneous postal addresses of United States of America (USA) for retrieving the matching component. The system specifies the degree of match of the noisy labeled data for each probable component, which aids in interpretation of the address.

The task of interpretation of address in presence of errors for delivery sorting and mapping to mail delivery point is more complex in countries like India where the postal addresses are not structured and are more often incomplete and approximate. A study of Indian postal addresses reveals that a mail delivery point can be specified using different descriptions;

Some parts of the material in this chapter appear in the following research paper

also two almost similar addresses may be referring to completely different mail delivery points. A few such typical cases of ambiguous postal addresses are depicted in Figure 8.1.

<table>
<thead>
<tr>
<th>Number</th>
<th>Addressee Name</th>
<th>House Number, Road Number</th>
<th>Landmark, Area</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Mr Shanmukhappa A Angadi</td>
<td>H No 8, 12th Main Road</td>
<td>Near Basava vana, Vidyagiri</td>
<td>Bagalkot</td>
</tr>
<tr>
<td>2.</td>
<td>Shri S A Angadi</td>
<td>H No 8, 12th Main Road</td>
<td>Near Gangoor Hospital</td>
<td>Bagalkot-587102</td>
</tr>
<tr>
<td>3.</td>
<td>Mr S A Angadi,</td>
<td>H No 10, 12th Main Road</td>
<td>Vidyagiri, Bagalkot-587102</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 8.1: Ambiguity in Destination Postal Addresses**

Addresses 1 and 2 of Figure 8.1 refer to the same person though the descriptions found in the address are slightly different. The addresses 1 and 3 are almost similar with one component missing in address 3. This may seem to represent the same addressee (if error is assumed to be present) but actually they refer to two different persons and delivery points. This shows that Indian postal addresses along with being incorrect/ inaccurate are also ambiguous and imprecise hence the processing of postal address to interpret and identify the mail delivery point is a more complex problem.

Human mail sorters identify the delivery point referred to by the ambiguous and inaccurate addresses making use of the knowledge gained by experience and juxtaposition of the information conveyed by individual components of the address. The automation of the process of identifying the mail delivery point from the deciphered address (in text form) will make the postal service more efficient, as the mail can be automatically sorted and can be further used to generate optimal mail distribution route on a daily basis. To the best of our knowledge there is no work reported on identification of the mail delivery points, especially with approximate/incomplete postal addresses in the Indian context. Development of computational strategies for mapping such ambiguous postal addresses to mail delivery points requires tools that can work with imprecise and inaccurate
information. As brought out in the previous chapter soft computing techniques are better suited to deal with such applications.

In this chapter a novel soft computing model for mapping the incomplete/approximate postal address to mail delivery point is presented. The model takes the destination postal address in symbolic object form (generated by the fuzzy symbolic address component labeling and postal address object generation system described in Chapter 7) as input and maps it to one or more mail delivery points as depicted in Figure 8.2. If the addresses are ambiguous then the system will identify all the probable delivery points for the address. Further the system also computes the relative confidence in each of the identified delivery point.

![Figure 8.2: Block Schematic of the Soft Computing Model](image)

A symbolic knowledge base for representing mail delivery points in a locality is devised to utilize the advantageous aspects of symbolic object representation of the postal address (PAO). A symbolic similarity measure is developed for mapping the entire address to a mail delivery point. In sequel to the similarity computations, which are viewed as fuzzy membership values, a $\alpha$-cut de-fuzzification methodology (Ross, 1997) is employed to evaluate the confidence in the mapping of the postal address to mail delivery point. The system is tested exhaustively and an efficiency of 86% is obtained in mail delivery point mapping.

The remaining part of the chapter is organized into six sections. The overall approach for mapping incomplete/imprecise addresses to mail delivery points using soft computing techniques is described in section 8.1. Section 8.2 describes the symbolic knowledge base used for mapping addresses to mail delivery points. A symbolic similarity measure employed in mapping of addresses to mail delivery points is presented in section 8.3.
Section 8.4 elaborates the fuzzy symbolic methodology for resolving ambiguity in mail delivery point mapping. The results and discussions are presented in section 8.5. Section 8.6 summarizes the work.

8.1 SOFT COMPUTING TECHNIQUES FOR MAIL DELIVERY POINT MAPPING

The automatic mapping of postal addresses to mail delivery points requires imitating the operation of the skilled mail sorter/postman. Emulating human behavior requires appropriate storage of knowledge base and an efficient inference mechanism. In this work knowledge based system employing soft computing techniques for inference is presented. The block schematic of the knowledge based system is illustrated in Figure 8.3.

![Knowledge Based System for Mail Delivery Point Mapping](image)

**Figure 8.3:** Knowledge Based System for Mail Delivery Point Mapping

The symbolic knowledge based system maps the input postal hoard object (generated using the soft computing techniques described in Chapter 7) to one or more mail delivery points. The similarity measure defined in section 8.3 is used to find the symbolic distance of the input address to all the mail delivery points in the knowledge base. The input address is mapped to the mail delivery point with highest similarity. Sometimes the input address may have comparable amount of similarity with more than one mail delivery point. This leads to fuzziness in deciding the mail delivery point represented by the address. Hence the
symbolic similarity of the input address with a mail delivery point is treated as a fuzzy membership function. Further the membership information is de-fuzzified using a fuzzy \( \alpha \)-cut technique. The relative confidence in the identified mail delivery points is computed using the fuzzy \( \alpha \)-cut set. The symbolic knowledge base employed in this work is described in section 8.2.

8.2 SYMBOLIC KNOWLEDGE BASE FOR MAIL DELIVERY POINT MAPPING

The components of the postal address object (PAO), describing a postal address are compared with knowledge base of a place (PL_MDP_KB) to identify the mail delivery point represented by the address. The knowledge base has been designed by studying a large number and variants of postal addresses. The symbolic knowledge base consists of a synthetic object: PL_MDP_KB, consisting of three hoard objects describing the three aspects of a postal address namely addresssee, location and place. The three hoard objects are, mail delivery point addresssee knowledge base: mdp_addressseeKB, mail delivery point location knowledge base: mdp_locationKB, and mail delivery point place knowledge base: mdp_placeKB as depicted in equation 8.1.

\[
\text{PL}_\text{MDP}_\text{KB} = \{[\text{mdp}_{\text{addresseeKB}},[\text{mdp}_{\text{locationKB}}],[\text{mdp}_{\text{placekb}}]) \} \quad (8.1)
\]

The assertion objects that constitute the hoard objects are depicted in Figure 8.4. The hoard object mdp_addressseeKB has instances consisting of same Addresssee\_name with different Id\_No to characterize same person having different delivery points or different persons with same name. Similarly instances of mdp_addressseeKB containing different Addresssee\_name with the same Id\_No may also be present to characterize the scenario, where different persons reside in the same delivery point. Hence there is many to many mapping between Id\_No and other assertion objects of mdp_addressseeKB giving multiple entries for each Id\_No. The Id\_No object refers to the delivery point in the locality and is used to map the addresssee information to a delivery point. The hoard object mdp_locationKB has only one entry for each mail delivery point represented by Id\_No, giving the various descriptions of the mail delivery point. The other hoard object mdp_placeKB gives the various descriptions of the place where the mail is to be
distributed. This organization of the symbolic knowledge base helps in symbolic analysis of the input postal address object.

```
PL_MDP_KB={
    //Synthetic Object
    {mdp_addressKB=
        //Hoard Object
        [Salutation]
        //Assertion Object
        [Addresse_Name]
        //Assertion Object
        [Alias_Name]
        //Assertion Object
        [Qualification]
        //Assertion Object
        [Profession]
        //Assertion Object
        [Designation]
        //Assertion Object
        [Id No]
        //Assertion Object
    }

    //Hoard Object
    {mdp_locationKB=
        [Id_No]
        //Assertion Object
        [House NO]
        //Assertion Object
        [House_Name]
        //Assertion Object
        [House_Name_alias]
        //Assertion Object
        [Road_No1]
        //Assertion Object
        [Road_Name1]
        //Assertion Object
        [Road1_Alias]
        //Assertion Object
        [Road_No2]
        //Assertion Object
        [Road_Name2]
        //Assertion Object
        [Road2_Alias]
        //Assertion Object
        [Area_Name]
        //Assertion Object
        [Area_Alias]
        //Assertion Object
        [Land_Mark1]
        //Assertion Object
        [Land_Mark1_Alias]
        //Assertion Object
        [Land_Mark2]
        //Assertion Object
        [Land_Mark2_Alias]
        //Assertion Object
        [POST_BOX]
        //Assertion Object
        [Firm_Name]
        //Assertion Object
        [PIN_Code]
        //Assertion Object
        [POST]
        //Assertion Object
    }

    //Hoard Object
    {mdp_placeKB=
        [Place]
        //Assertion Object
        [Taluk]
        //Assertion Object
        [District]
        //Assertion Object
        [VIA]
        //Assertion Object
        [State]
        //Assertion Object
        [Country]
        //Assertion Object
    }
}
```

**Figure 8.4: Structure of Symbolic Mail Delivery Point Knowledge Base**

A snapshot of the knowledge base **PL_MDP_KB** comprising information about the mail delivery points of Bagalkot, a district place in Karnataka state of south India is depicted in Figure 8.5. The knowledge base is employed for symbolic data analysis to map the input
postal address object to mail delivery point. The similarity measure used in symbolic data analysis is described in section 8.3.

\[
\text{PL\_MDP\_KB=} \\
\{\text{mdp\_adresseKB=} \\
\{\text{Salutation},<\text{DR,MR}>\} \{\text{Addresssee Name}, <\text{PATIL CHANDRASHEKHAR P}>\}; \{\text{Alias Name}, <\text{CHANDRU}>\}; \{\text{Qualification}, <\text{MBBS,MD}>\}; \{\text{Profession}, <\text{DOCTOR}>\}; \{\text{Designation}, <\%>\} \\
\{\text{Id No},<000001>\} \\
\{\text{Salutation},<\text{MR,ER}>\} \{\text{Addresssee Name}, <\text{ARAVIND C PATIL}>\}; \{\text{Alias Name}, <\text{BABU}>\} \\
\{\text{Qualification}, <\text{BE}>\}; \{\text{Profession}, <\%>\}; \{\text{Designation}, <\text{ENGINEER}>\}; \{\text{Id No},<000001>\} \\
\} \\
\} \\
\{\text{indp\_locationKB=} \\
\{\text{Id_No},<000001>\}; \{\text{House NO},<\%>\}; \{\text{House Name},<\%>\}; \{\text{House Name alias},<\%>\}; \{\text{Road No1},15\} \\
\{\text{Road Name1},<\text{CROSS}>\}; \{\text{Road1 Alias},<\text{MAIN}>\}; \{\text{Road No2},<\%>\}; \{\text{Road Name2},<\%>\} \\
\{\text{Area Name},<\text{VIDYAGIRI}>\}; \{\text{Area Alias},<\text{SHIVAGIRI}>\} \\
\{\text{Land Mark1},<\text{KENCHAMMA TEMPLE}>\}; \{\text{Land Mark1 Alias},<\text{KENCHAMMADEVI TEMPLE}>\} \\
\{\text{Land Mark2},<\text{ALADA MARA}>\}; \{\text{Land Mark2 Alias},<\%>\}; \{\text{Post BOX},<\%>\}; \{\text{Firm Name},<\%>\} \\
\{\text{PIN code},<587102>\}; \{\text{POST},<\%>\} \\
\{\text{Id_No},<000002>\}; \{\text{House NO},<\%>\}; \{\text{House Name},<\%>\}; \{\text{House Name alias},<\%>\}; \{\text{Road No1},18\} \\
\{\text{Road Name1},<\text{CROSS}>\}; \{\text{Road1 Alias},<\text{MAIN}>\}; \{\text{Road No2},<\%>\}; \{\text{Road Name2},<\%>\} \\
\{\text{Area Name},<\text{VIDYAGIRI}>\}; \{\text{Area Alias},<\text{VIIYANAGAR}>\} \\
\{\text{Land Mark1},<\text{BSNL OFFICE}>\}; \{\text{Land Mark1 Alias},<\%>\}; \{\text{Land Mark2},<\%>\} \\
\{\text{Land Mark2 Alias},<\%>\}; \{\text{POST BOX},<\%>\}; \{\text{Firm Name},<\%>\}; \{\text{PIN code},<587102>\}; \{\text{POST},<\%>\} \\
\} \\
\} \\
\} \\
\{\text{mdp\_placeKB=} \\
\{\text{Place},<\text{BAGALKOT}>\}; \{\text{Taluk},<\text{BAGALKOT}>\}; \{\text{District},<\text{BAGALKOT}>\}; \{\text{VIA},<\%>\} \\
\{\text{State},<\text{KARNATAKA}>\}; \{\text{Country},<\text{INDIA}>\} \\
\} \\
\}
\]

\text{Note:} \% \text{indicates that the corresponding assertion object has null values}

\text{Figure 8.5: Sample Values of Symbolic Knowledge Base}

\section*{8.3 Symbolic Similarity Measure for Mail Delivery Point Mapping}

The symbolic data analysis for mail delivery point mapping employs a similarity measure to map the input postal address object to mail delivery points. The similarity measure quantifies the similarity of the input postal address object to the description of the mail delivery points in the symbolic knowledge base. The input postal address is mapped to the mail delivery point having the highest similarity.
The similarity measure defined in (Gowda, 2004) is modified and adapted for mapping the postal address to the mail delivery point. The symbolic similarity measure defined in (Gowda, 2004) consists of three components, namely similarity due to position, similarity due to content and similarity due to span of the two objects being compared. The position similarity is defined only for interval type of data and describes the distance of one object to the initial position of other object. The span similarity is defined for both interval and absolute type of data and describes the range/fraction of similarity between the objects. The content similarity describes the similarity between the contents of the two objects. As the postal address object, which is being processed in this work consists of only absolute values the span and content components are defined for measuring the similarity of the input postal address object with the symbolic representation of mail delivery points. The similarity measure has been designed to process postal addresses written in various forms. In particular use of initials to represent the first / second / last names, errors in spelling when writing proper nouns describing addressee or location object of the postal address are some of the specific issues considered.

The symbolic processing of the address for mapping to mail delivery point is made robust by performing the word/ token comparison at three different levels namely full word match, first character match and partial word match. The partial word match is used to handle wrong spellings and involves the match of first, last and at least one other character in the corresponding position. The symbolic processing employs full word match / first character match or partial word match depending on the type of input component. For all numeric fields only full word match is resorted to, whereas for descriptive text components full word, first character and partial word matches are combined to generate the similarity values. The similarity formulation employed for mapping postal address to mail delivery point is presented in the following.

Let the Input Symbolic Postal Address Object be represented by $\tilde{I}$. The symbolic object has three assertion object components (refer equation 7.1) represented as $\tilde{I}_A$, $\tilde{I}_L$ and $\tilde{I}_P$ as depicted in equation 8.2. The assertion objects $\tilde{I}_A$, $\tilde{I}_L$ and $\tilde{I}_P$ describe the addressee, location and place aspects of the postal address.
The assertion object $\hat{I}_p$ is used to verify whether the input mail address belongs to the place
where mail sorting is done, whereas the $\hat{I}_A$ and $\hat{I}_L$ are used to compute the similarity
measure.

Let each instance of hoard object $mdp\text{-}addressKB$ (addressee knowledge base) be
represented by $Ak_i$ and each instance of the hoard object $mdp\text{-}locationKB$ (location
knowledge base) be represented by $Lk_j$, for all $1 \leq i \leq m$ and all $1 \leq j \leq n$, where $m$ is the
number of instances of $mdp\text{-}addressKB$ and $n$ is the number of instances of $mdp\text{-}location
KB$. In general the number of instance of $mdp\text{-}addressKB$ is greater than or equal to the
number of instances of $mdp\text{-}location KB$ $(m = n)$.

The similarity measure for addressee information gives the nearness of the input addressee
object with each instance in the symbolic hoard object $mdp\text{-}addresseeKB$. The similarity
measure between input addressee information and an instance of addressee object in the
knowledge base is computed as in equation 8.3.

$$S^A_{\hat{I}_A, Ak_1} = \frac{1}{EC_A} \sum_{p=1}^{EC_A} netsim_p \quad \ldots(8.3)$$

Where,

$EC_A$ is the number of components/events of postal addressee object $\hat{I}_A$

The value of $netsim_p$ (net similarity measure) for every component / event of $\hat{I}_A$ is
calculated as in equation 8.4.

$$netsim_p = \frac{scanSim_p + contentSim_p}{2.00} \quad \ldots(8.4)$$
The parameters, scanSim and contentSim, represent the span and content similarity for the p\textsuperscript{th} event in I\textsubscript{A} to the corresponding assertion object in the instance A\textsubscript{ki} of mdp_addresseeKB. They are computed as depicted in equations 8.5 and 8.6.

\[
\text{scanSim} = \frac{\text{Comp}_PO + \text{Comp}_KB}{2 \times \text{Sum}_PO \_ KB} \quad \text{...(8.5)}
\]

\[
\text{contentSim} = \frac{\text{interse}}{\text{Sum}_PO \_ KB} \quad \text{...(8.6)}
\]

where,

- Comp\_PO is the number of words/tokens in the p\textsuperscript{th} event in I\textsubscript{A}
- Comp\_KB is the number of words/tokens in the corresponding assertion object of knowledge base entry A\textsubscript{ki}
- Sum\_PO\_KB is evaluated as in equation 8.7.

\[
\text{Sum}_PO \_ KB = \text{Comp}_PO + \text{Comp}_KB - \text{interse} \quad \text{...(8.7)}
\]

The value of interse is computed using equation 8.8.

\[
\text{interse} = \text{COM} \times \text{MF} \quad \text{...(8.8)}
\]

The values of COM (common elements between the objects being compared) and MF (the match factor) are computed differently for different types of events/ components in the input postal address object (PAO). If the event of PAO is numeric then full word match between input event and corresponding assertion object in the knowledge base is taken up. If the input event is descriptive text but contains at least one word with a single character then first character match is employed along with the full word match. If the input event is descriptive and contains more than one character for each token / word then partial word matching is resorted to along with full word match for overcoming small errors in spelling. The parameters COM and MF are computed as depicted in Table 8.1 for the three different cases of input address.
### Table 8.1: Computation of Parameters COM and MF for different Input Events

<table>
<thead>
<tr>
<th>SI No</th>
<th>Description of Input Event</th>
<th>Computation of COM</th>
<th>Computation of MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Numeric Type-Requires Full word Match</td>
<td>$k$</td>
<td>$\frac{k}{n}$</td>
</tr>
<tr>
<td>2</td>
<td>Descriptive Type, with Single Character in a word/ token-Requires Match of First Characters</td>
<td>$k+k1$</td>
<td>$\frac{0.25*k1+k}{n}$</td>
</tr>
<tr>
<td>3</td>
<td>Descriptive Type, more than one character in all words/tokens-Requires Partial Word Match</td>
<td>$k+k2$</td>
<td>$\frac{\left(\sum_{a=1}^{k2} (2 * f_l_a + o_c_a) + k\right)}{n}$</td>
</tr>
</tbody>
</table>

Where,

- $n$ is the total number of words in the $p^{th}$ event of $\hat{I}_A$
- $k$ is the number of full word matches between $p^{th}$ event of $\hat{I}_A$ and the corresponding assertion object of $A_k_i$
- $k1$ is the number of first character matches between words in $p^{th}$ event of $\hat{I}_A$ and the corresponding assertion object of $A_k_i$
- $k2$ is the number of partial matches between words in $p^{th}$ event of $\hat{I}_A$ and the corresponding assertion object of $A_k_i$. The partial word match is defined as match at first, last and at least one other character positions in the word.
- $f_l_a$ is the weight assigned in evaluating match factor for the match of first and last characters of $a^{th}$ word, when partial word match is considered for $p^{th}$ event of $\hat{I}_A$
- $o_c_a$ is the weight assigned in evaluating match factor for the match of a character of $a^{th}$ word, other than the first and last, when partial word match is considered for $p^{th}$ event of $\hat{I}_A$, when partial word match is considered
- $\text{length}_\text{of}_\text{input}_\text{word}_a$ is the length of the $a^{th}$ word considered for partial match

The similarity measure between input location object $\hat{l}_L$ and an instance of location object of the knowledge base $L_K_j$ is found as given in equation 8.9.
Mail Delivery Point Mapping

\[ S_{L_{1L, LK_j}}^L = \frac{1}{EC_L} \sum_{p=1}^{netsim_p} \]  

...(8.9)

Where,

\( EC_L \) is the number of components of postal location object existing in the input \( \hat{L}_L \).

The \( netsim_p \) values are calculated in a manner similar to the addressee object using \( \hat{L}_L \) and \( LK_j \).

The similarity measure for a given mail delivery point, including addressee and location information is then calculated by taking the average of both the similarities as given in equation 8.10.

\[ S_{KB_{dp}} = \frac{S_{L_{1L, LK_j}}^L + S_{L_{1L, LK_j}}^A}{2.0} \quad \text{such that } (Id\_No(Aki)=Id\_No(LK_j)) \]  

...(8.10)

Where,

\( dp=Id\_No(Aki) \)

The similarity values are computed for all the \( n \) delivery points and each of the \( m \) addressee information in the symbolic knowledge base. Thus yielding a total of \( m \) (as \( m > n \)) similarity values. The mail delivery point with the highest similarity is identified as the mail delivery point represented by the postal address.

The similarity measure devised for mapping addresses to mail delivery points is similar in structure to the similarity measure for address component labeling described in Chapter 7. The similarity measure for mail delivery point mapping uses only the span and content similarity, as is the case with the similarity measure for address component labeling. The similarity value for address component labeling (described in Chapter 7) is computed using the features extracted from a set of words/tokens treated as an address component. These features are compared with seven events of the assertion objects in the knowledge base for similarity computation and labeling of address components. In contrast the similarity value for mail delivery point mapping, is computed by comparing the labeled components (events of addressee and location object) with the corresponding assertion.
objects of the knowledge base `mdp_addresseeKB` / `mdp_locationKB`. This computation employs comparison of words/ tokens to evaluate the similarity. The computation is made robust by using three different types of word match depending on the context for evaluating the similarity.

The postal addresses in general are imprecise and ambiguous; hence the similarity formulation for mapping the postal address will yield comparable values for more than one distinct delivery point leading to ambiguity in mapping. This ambiguity in mapping of the postal address to mail delivery point is resolved by using the fuzzy technique described in section 8.4.

### 8.4 FUZZY SYMBOLIC METHODOLOGY FOR AMBIGUITY RESOLUTION

The symbolic similarity measure defined in section 8.3 gives the approximate nearness of the input postal address to all the delivery points in the knowledge base. To handle the imprecision and ambiguity in addresses the similarity value is treated as a fuzzy membership function in different mail delivery points. The de-fuzzification is carried out using fuzzy α- cut methodology (Ross, 1997). The fuzzy α- cut set so obtained is employed in deriving the confidence value in the decision made by the system. The fuzzy symbolic methodology employed for de-fuzzification and decision making is similar to the one used for labeling of address components and uses the symbolic similarity measure defined in section 8.3.

The symbolic similarity measure is computed for the input postal address object with all the mail delivery points, in sequel the mail delivery points are arranged in decreasing order of similarity. This list gives the fuzzy membership of the input address object in the mail delivery points and is depicted in the similarity array of Figure 8.6.

<table>
<thead>
<tr>
<th>index</th>
<th>I₀</th>
<th>I₁</th>
<th>I₂</th>
<th>I₃</th>
<th>Iₙ⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim</td>
<td>S₀</td>
<td>S₁</td>
<td>S₂</td>
<td>S₃</td>
<td>Sₙ⁻¹</td>
</tr>
<tr>
<td>MDP</td>
<td>dp₀</td>
<td>dp₁</td>
<td>dp₂</td>
<td>dp₃</td>
<td>dpₙ⁻¹</td>
</tr>
</tbody>
</table>

**Figure 8.6: The Similarity Array Sorted in Descending Order (`S₀ > S₁ > S₂`...)**
A fuzzy α-cut is defined using equation 8.11. The α-cut set is constructed as the set of all delivery points having similarity value greater than α, as illustrated in Figure 8.7.

\[ \alpha = S_0 - DFC \cdot S_0 \]  

\[ \cdots (8.11) \]

Where,

- \( S_0 \) is the maximum similarity value obtained for the input postal address object
- \( DFC \) is the de-fuzzification constant and is taken as 0.1, based on the experimentation with postal address components.

\[ 0.1 \cdot M \cdot \{ \]

\[ \text{Alpha (α) Cut} \]

\[ \alpha \text{-cut set}=\{dp_0,dp_1,dp_2\} \]

\[ \text{Figure 8.7: De-fuzzification Process and the α cut set} \]

If the α- cut set has only one mail delivery point, then it is identified as the mail delivery point represented by the input postal address with 100% confidence. If the α- cut set has more than one mail delivery points then the probable mail delivery points are output with the decreasing order of confidence. The confidence of the system in a given mail delivery point identification is evaluated using equation (8.12).

\[ C_j = \frac{S_j}{\sum_{k=1}^{p} S_k} \cdot 100 \]

\[ \text{for } 1 \leq j \leq p \]

\[ \cdots (8.12) \]
Where,

$C_j$ is Confidence of assigning $j^{th}$ mail delivery point to input address object

$p$ is the number of mail delivery points in a-cut set

$S_j$ is the similarity of input address with $j^{th}$ mail delivery point $dp_j$ in similarity array

$S_k$ is the similarity of input address with $k^{th}$ mail delivery point $dp_k$ in similarity array

If the number of addresses in the $\alpha$- cut set is more than four, then historical evidence of similar postal addresses are used to make the decision as described in algorithm 8.1. The complete procedure is presented in algorithm 8.1.

**Algorithm 8.1: Identification of the Mail Delivery Point (MDP) from address components**

**Input:** The identified postal address components (Postal Address Object)

**Output:** The identified mail delivery point/s with similarity measure/s and confidence/s

1. Find the similarity measure of the input address object $IA$ with $A_{ki}$ the $i^{th}$ instance of address information present in the knowledge base $mdp\_addresseeKB$. For all $i=1:m$, $m$ is the number of instances of $mdp\_addresseeKB$.

2. Find the similarity measure of the input Location object $IL$ with $L_{Kj}$ the $j^{th}$ instance of location information present in the knowledge base $mdp\_locationKB$. For all $j=1:n$, $n$ is the number of instances of $mdp\_locationKB$.

3. Combine the similarity measures for location and address matches with similarity measure as the average of the two and sort them in descending order to form a similarity array consisting of $m$ entries.

4. Find $\alpha$ and generate $\alpha$- cut set, if only one MDP lies in the $\alpha$- cut set, then output the mail delivery point id and address with 100% confidence measure and stop.

5. If more than one and less than 5 matches lie in the range then calculate the confidence value and output the mail delivery points with confidence values.

6. If more than 4 matches lie in the $\alpha$- cut set the history information is used to check whether the mail address as input is available in the history file, if so depending on the score the confidence values are altered by adding some constant value, and the probable delivery points with confidence values are output, otherwise manual resolution is needed.

The algorithm 8.1 is computationally intensive. It generates $m$ similarity values as depicted in equation (8.10), where $m$ is the number of instances of address information in the knowledge base $mdp\_addresseeKB$. Further these $m$ similarity values are to be sorted in
descending order. A simple bubble sort is employed in this work, hence the worst case time complexity is equal to $O(m+m^2+k)=O(m^2)$ (Horowitz et al, 1998). The usage of an efficient sorting procedure such as quick sort will give a worst case time complexity of $O(m*\ln(m))$.

This exhaustive search of knowledge base was purposely designed to cater to all types of errors in writing addresses and also errors in spellings. In particular the exhaustive search helps in correct mapping of postal address even in the presence of wrong information in one or two components. The knowledge base is to be built only for the locality served by the delivery post office. This limits the size of the knowledge base to about 1000 mail delivery points and about 2000/3000 prominent persons in the maximum. The increased availability of computing power makes exhaustive search in such a knowledge base feasible, and provides more efficiency in mapping of addresses to mail delivery points with an associated cost of more computing time. However the knowledge about the place and the different valid combinations that exist can be utilized to implement a directed search in the knowledge base thus reducing time complexity. The development of such a strategy utilizing local knowledge to direct the search has been kept outside the scope of this research to concentrate on development of other tools required for integrated postal automation and can be explored further.

8.5 RESULTS AND DISCUSSIONS

The knowledge based system using soft computing methodology for mapping postal addresses to mail delivery points is thoroughly tested using a knowledge base developed for sorting office of Bagalkot. The knowledge base is populated with addressee information of 500 persons and mail delivery point information for 300 delivery points. The soft computing methodology for Mail Delivery Point identification was tested considering the various issues that may occur in practice. A few typical cases are listed and explained in Table 8.2. The output indicated in the table does not show the addressee name, which is same as given in the input.
### Table 8.2: The results of Soft Computing Approach to Mail Delivery Point Identification

<table>
<thead>
<tr>
<th>SI No</th>
<th>Input Address</th>
<th>Location Output</th>
<th>Similarity Measure (SM)/ Percentage Confidence (PC)/ MDP</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mr Shanmukappa A Angadi H No 8, 12th Main Road Near Basava Vana, Vidyagiri Bagalkot</td>
<td>H No. 8, 12th Main Road, Near Basava Vana, Vidyagiri, Bagalkot 587102, Karnataka State</td>
<td>SM=0.825, PC=100%, MDP=000003</td>
<td>The expected complete address was given as input.</td>
</tr>
<tr>
<td>2</td>
<td>Prof S A Angadi H No 8, 12th Main Road Near Gangoor Hospital Bagalkot-587102</td>
<td>H No. 8, 12th Main Road, Near Basava Vana, Vidyagiri, Bagalkot 587102, Karnataka State</td>
<td>SM=0.5, PC=100%, MDP=000003</td>
<td>The input address did not contain the area name but the system could still identify the mail delivery point with 100% confidence.</td>
</tr>
<tr>
<td>3</td>
<td>Kumar Manoj S Biradar c/o S A Angadi H No 8, 12th Main Road Vidyagiri, Bagalkot</td>
<td>H No. 8, 12th Main Road, Near Basava Vana, Vidyagiri, Bagalkot 587102, Karnataka State</td>
<td>SM=0.626, PC=100%, MDP=000003</td>
<td>The addressee name is not present in the knowledge base and PIN is not present but c/o information is used here.</td>
</tr>
<tr>
<td>4</td>
<td>Mr S A Angadi c/o M A Angadi “Gurukrupa”, Near Old IB, Extension, Bagalkot-587101</td>
<td>“Gurukrupa”, Near Old IB, Extension, Bagalkot 587101, Karnataka State</td>
<td>SM=0.687, PC=100%, MDP=0000029</td>
<td>The addressee has more than one location information, still the correct address was generated.</td>
</tr>
<tr>
<td>5</td>
<td>Mr Suresh Basappa Angadi “Shivakrupa”, H No 10, 12th Main Road Vidyagiri Bagalkot-587102</td>
<td>H No 1, “Shiva Krupa”, 12th Main Road, Vidyagiri, Bagalkot 587102, Karnataka State</td>
<td>SM=0.7, PC=100%, MDP=000028</td>
<td>The input address contained wrong house number and was corrected and the decision was correct.</td>
</tr>
<tr>
<td>6</td>
<td>Mr Suresh Appasaheb Angadi, “Shivakrupa, H No 10, 12th Main Road Vidyagiri, Bagalkot-587102</td>
<td>This has resulted in three probable outputs: i) H No 1, “Shivakrupa”, 12th Main Road, Vidyagiri, Bagalkot 587102, Karnataka State</td>
<td>i) SM=0.566, PC=33.65%, MDP=000028</td>
<td>i) The addressee name matches partially and house number for this match is wrong.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>i)SM=0.566, PC=33.65%, MDP=000003</td>
<td>i) The addressee information matches partially and road number and area information match in the location part.</td>
</tr>
</tbody>
</table>
The outputs of the knowledge based system, listed in Table 8.2 illustrates the power of the soft computing model employed for mapping the incomplete/ approximate postal addresses to the mail delivery points. The table indicates that even though the similarity to an address is 0.5 the confidence of the system in identification is 100% as in case 2. Also in case 6 though the similarity is 0.566 the confidence is only 33.65%, both of which replicate the human expert behavior. For the ninth entry the similarity is only 0.343 but still the confidence in the decision is 100%. So the knowledge based soft computing model for mail delivery point mapping has been successful in emulating the human behavior.

The overall results of the thorough testing using addresses of different nature are enlisted in Table 8.3. The system could unambiguously decide on the mail delivery point for about
86% of addresses (Corresponding to SI Nos 1, 2 & 3) and for remaining addresses either manual resolution was necessary or the addresses were not deliverable. The results are pictorially depicted in Figure 8.8.

### Table 8.3: Test results of Mail Delivery Point (MDP) Mapping

<table>
<thead>
<tr>
<th>SI No</th>
<th>Particulars</th>
<th>Number of Addresses*</th>
<th>Percentage of total addresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Addresses mapping to single MDP with a confidence of 100% in decision</td>
<td>255</td>
<td>72.86</td>
</tr>
<tr>
<td>2</td>
<td>Addresses mapping to single MDP with a confidence greater than 75% in decision</td>
<td>33</td>
<td>9.43</td>
</tr>
<tr>
<td>3</td>
<td>Addresses mapping to single MDP with a confidence less than 75% in decision</td>
<td>13</td>
<td>3.71</td>
</tr>
<tr>
<td>4</td>
<td>Addresses mapping to multiple MDP’s; one MDP is mapped with a percentage confidence greater than 50 and others less than 50 (Manual resolution is needed)</td>
<td>38</td>
<td>10.86</td>
</tr>
<tr>
<td>5</td>
<td>Addresses mapping to multiple MDP’s: all MDP’s are mapped with confidence less than 50 (Undeliverable addresses, needs verification)</td>
<td>11</td>
<td>3.14</td>
</tr>
</tbody>
</table>

*Note: The total number of addresses used for testing is 350

---

**LEGEND**

1: Addresses mapping to single MDP with a confidence of 100%
2: Addresses mapping to single MDP with a confidence greater than 75%
3: Addresses mapping to single MDP with a confidence less than 75%
4: Addresses mapping to multiple MDP’s; one MDP is mapped with percentage confidence greater than 50 and others less than 50 (Manual resolution is needed)
5: Addresses mapping to multiple MDP’s: all MDP’s are mapped with confidence less than 50 (Undeliverable addresses, needs verification)

Note: Total Number of addresses tested: 350

**Figure 8.8: Overall Results of Soft Computing Model for MDP Mapping**
8.6 SUMMARY

The knowledge based system employing soft computing techniques presented in this chapter focuses on one of the very important tasks of integrated postal automation, namely mapping an incomplete/ approximate/ imprecise postal address to a mail delivery point. The system takes the symbolic representation of postal address namely Postal Address Object (PAO), generated using the soft computing techniques as input. It employs a newly devised symbolic similarity measure for mail delivery point identification. The similarity measure is treated as fuzzy membership function and a fuzzy $\alpha$-cut method is employed for de-fuzzification and deciding on the mail delivery point. The soft computing methodology also computes the confidence in the mail delivery points identified. The confidence values help in sequencing and deciding on the delivery of the mail.

The symbolic data analysis employs an exhaustive search in the knowledge base to identify the mail delivery point. The exhaustive search helps in interpreting erroneous addresses and makes the methodology more robust. The soft computing approach has efficiently modeled the human behavior while dealing with inaccuracies and imprecisions of postal addresses. The soft computing methodology can be used in a post office to map the addresses of the mail to be distributed on a given day. The list of such mail delivery points can be further used for generating optimal route to be utilized by the postman for delivery of mail. The part III of the thesis presents various methodologies for generating optimal route for mail distribution.

The methodology presented here can be extended to identification and interpretation of any structured text such as in automatic analysis of resumes. The soft computing methodology used in this work can be made more efficient by making use of the local knowledge and defining a directed search to identify the probable mail delivery points.
Part III

OPTIMIZATION ISSUES WITH DISTRIBUTION OF POSTAL MAIL

Mail Collection → Mail Sorting → Document Image Analysis Tasks → Knowledge Based Tasks →
Overview of Part III

The postal mail is delivered to the addressee by a postman/delivery agent, who is appointed by the Mail Distribution Agency/Delivery Post Office (MDPO). Every MDPO caters to a pre-assigned area/locality, which is divided into different contiguous parts called as 'beats'. The MDPO assigns a beat to a postman who delivers the mail to the delivery points in the beat on a daily basis. The postman seeks to use a minimal distance (optimal) path for traversal in the beat. The optimal path is dependant on the delivery points, which receive mail on a given day and is to be found on a daily basis. Manual estimation of the optimal path will quickly lead to suboptimal routes because of the many geographic and mail delivery conditions that are to be considered. Consequently, there is a need for automatic tools for generation of the optimal path for mail distribution.

Intelligent computational strategies for finding the optimal path for distribution of postal mail to the delivery points on a daily basis are devised and presented in Part III of the thesis. The mail delivery points, especially in India, are situated in unstructured localities and there is a need for systematic representation of these delivery points. Geographical Information System (GIS) technology provides an efficient tool for representation of mail delivery points. A new GIS data model is devised for representing the mail delivery points and is further used in generation of the optimal path for mail distribution. Novel computational strategies based on Genetic Algorithms, Cluster Analysis and Graph Theoretic techniques are devised for finding the optimal path for distribution of mail.

The Part III of the thesis is divided into three chapters. Chapter 9 presents the various issues in finding the optimal route for distribution of postal mail and introduces a GIS data model for representing mail delivery points. An efficient data structure for organizing the spatial information of the beat is also devised. Chapter 10 presents intelligent computational strategies for generation of the optimal route for distribution of mail using GIS, genetic algorithm, and graph theoretic techniques. Chapter 11 describes a GLOCAL (Global + Local) strategy using genetic algorithm, cluster analysis, and GIS techniques to find the best path for distribution of mail using the traverse-park-distribute strategy.