

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/232656356>

Malaysian Address Semantic: The Process of Standardization

Article · January 2010

DOI: 10.1109/IICRD.2010.30

CITATIONS

0

READS

565

2 authors, including:



Mohamad Noorman Masrek

Universiti Teknologi MARA

181 PUBLICATIONS 945 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Project Library Usage Survey among Medical Students [View project](#)



Project Investigating the relationships between elearning platforms usability and student's motivation [View project](#)

Malaysian Address Semantic : The Process of Standardization

Mohamad Noorman Masrek

Faculty of Information Management
University of Technology MARA
Shah Alam, Malaysia
mnoormanm@gmail.com

Zakiah A. Razak

Faculty of Information Management
University of Technology MARA
Shah Alam, Malaysia
zack@pos.com.my

Abstract— Mail delivery has always been problematic to the Malaysian postal services due to incomplete or incorrect postal addresses. To this effect, the Malaysian Standard Body known as SIRIM has developed the Malaysian Standard Address with the purpose of standardizing Malaysian postal addresses and hence reducing the problem of undelivered mail. This article presents the findings of a study aimed at investigating the various postal-related problems in the Malaysian context. 275 domestic mails, randomly picked from Pos Malaysia Mail Processing Centre were examined and compared against the Malaysian Standard Address. The findings suggest that 58% of the addresses were non-conforming to the SIRIM address standard. Based on this finding, an automated system based on the semantic approach is proposed to address the issue of incomplete and inaccurate address in the context of Malaysia.

Keywords- Postal address; address semantic;Information Extraction

I. INTRODUCTION

An address can be defined as a collection of information, which is presented in a mostly fixed format and used for describing the location of a building, apartment, or other structure. An address generally uses political boundaries and street names as references, along with other identifiers such as house or apartment numbers [9]. Among the functions of addresses are (i) to provide a means of physically locating a building, especially in a city where there are many buildings and streets; (ii) to identify buildings as the end points of a postal system; (iii) a social function i.e. someone's address can have a deep effect on their social standing and (iv) as parameters in statistics collection, especially in census-taking or the insurance industry [9].

Given that address has major roles and functions, writing complete and accurate addresses are therefore very crucial especially for the purpose successful mail delivery. Thus, due to the aforesaid reasons, mail delivery has always been problematic to the Malaysian postal services. To address the problem, in 2006, the Malaysian Standard Body known as SIRIM has developed the Malaysian Standard Address aimed at standardizing Malaysian postal addresses and thus reducing the problem of undelivered mail [4]. However, after four years since its inception, the problem of undelivered mails is still plaguing the Malaysian postal services. Against

this background, a study was conducted with the purpose of investigating to what extent do Malaysian adhered to the Malaysian Standard Address in writing their postal mail addresses. In addition, the study is aimed at finding alternative solution to the incomplete or incorrect Malaysian postal addresses.

RELATED WORKS

Mining the extant literature unveiled that studies focusing on postal address related problems are still very scarce. The few related studies that were found conducted in countries other than Malaysia ([1], [10], [11]). Considering the fact that different country has different addressing need and format, previous findings may not be appropriate or relevant in the context of Malaysia.

In Australia, Christen and Belacic [1] developed an automated probabilistic approach based on Hidden Markov Model. The system was capable to correctly standardize even complex and unusual address. On the other hand, Sargur et al. [10], developed an automatic address interpretation systems for US postal address based on the interaction content components characterized in terms of Shannon's entropy. Nagabushan et al. [11], developed a system that employed a structured knowledge base devised to model the pattern of mail dispatch and the type of addresses sorting for the sorting office in South India. The system was reported to achieve 90% efficiency in validation and sorting. Nagabushan et al.[12], carried out another method namely Symbolic knowledge base. The accuracy and efficiency was improved to 95.60%; compared to Nagabushan et al [11].

RESEARCH METHODOLOGY

275 of domestic mail pieces were picked randomly from the Pos Malaysia mail processing centre located in Kuala Lumpur. The written address for each and every mail was examined and analyzed by comparing against the Malaysian Standard Address format issued by SIRIM [4]. Based on the findings, an automated systems based on the semantic approach is proposed to address the problem of incomplete and inaccurate postal address, which will be discussed in subsequent section.

FINDINGS

Figure 1 presents the findings of the study. Based on the analysis, it was found that 58% of the addresses were non-conforming to the SIRIM address standard while 31% of the postcodes were wrongly stated. The findings also unveiled that 23% of the investigated addresses had dual addresses and 16% had wrong postcodes.

Figure 1.0 List of the non-standard address elements

Problem of addresses	SIRIM Standard	%
Wrong Postcode	Ref 12: Postcode is to identify the various processing/delivery facilities and post offices.	0.31
Dual address	Ref 2.6: An Address that is formed by two types of address information.	0.23
Wrong Postcode Position	Ref 4.2: Postcode should be placed before cityname.	0.16
Missing Prefixes	Ref 4.6: All single information in address should use prefix	0.13
No Street and Locality Name	(Incomplete information category)	0.12
Wrong element position	(Create ambiguity)	0.11
Wrong building address format	Ref 4.7: Address for parcel or unit in multi-storey building	0.11
Missing house no	(Incomplete information category)	0.08
No building name	(Incomplete information category)	0.06
Unnecessary information (eg: Off street)	(Create ambiguity)	0.06
Wrong Street name	(Incomplete information category)	0.04
Wrong numeric streetname	(Incomplete information category)	0.02
Wrong Locality name	(Incomplete information category)	0.02

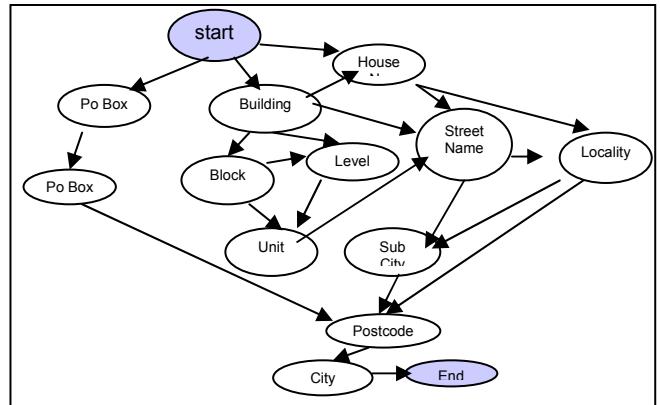
MALAYSIAN ADDRESS CHARACTERISTICS

Several characteristics of Malaysian address were recognized in order to choose the relevant methodology. The characteristics are: (i) The address guideline was published in 2006[4]. (ii) The national reference address does not exist. (iii) The postal codes could not indicate any street name or locality. (iv) Building name is more popularly apply in comparison to the building number. This will lead to a variety of the way to identify the name.

While the standard address documentation exists, it will provide the information that can be utilized in the rule standardization as well as tagging purposes. Based on the guidelines provided by SIRIM, the Malaysian Standard Address topology which is manually constructed is presented in Figure 2.0. The Malaysian address topology consists of 4 types of main input which are postbox, building, street address and non-street address. Elements in the postbox type consist of Postbox number, postcode and city. Whereas elements in the building address type consist of variety combination. The elements in the street address type will embed the street name together with house number while the non-street address consists of house number and locality as a basic reference. This topology shall be utilized in the phase of rules development [1]. In the absence of national reference address, the knowledge-base approach is appropriate. This approach is the adaptation from the ([11], [12], [2]).

Since there is no indicator of street name or locality name in the postcode; in addition to the 31% unreliable postcode stated in the mail pieces address based on the findings, the knowledge repository of the correct relationship among the address elements is required. The numerous abbreviations of the address elements are expected and shall be embedded into the knowledge-base.

MALAYSIAN ADDRESS TOPOLOGY



PROPOSED SOLUTION

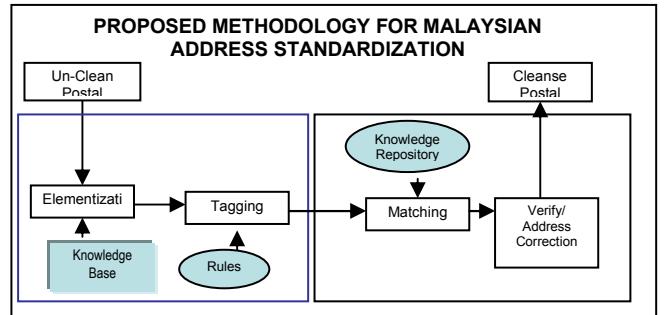


Figure 3.0 Proposed methodology for Malaysian Address Standardization.

Prototype system was developed in order to demonstrate the cleansing process for Malaysian address, which is based on the proposed methodology as shown in Figure 3.0. The prototype system consists of database, process of elementization, tagging, dictionary look-up and finally matching with the knowledge repository.

- Database: consists of 2 main objects which are dictionary and reference address. Dictionary will determine each item names which bind to the specific tag. As for example, the item name is 'Taman', the tag is locality prefix and street name (can be more than one). Reference Address consists of full standard addresses.

- Tagging: is a process of comparing the input versus dictionary for each item and determine the category of the item. As for example, ‘Jalan’ is a street prefix and the tag for it is ‘PS’. Other sample of tag category are ‘PL’ for Prefix Locality, ‘ST’ for street name, ‘CT’ for city name, ‘PH’ for house number prefix, ‘SL’ for locality.
- Parsing: is a process of splitting the phrases into a meaningful manner. The rules will apply here to identify the specific address elements. Using the clue of prefix, the tag starting with ‘P’ will be followed by address element. As for example, ‘PS’ is a prefix for street. So, the next phrases will consider as street name and usually be ended with another prefix or postcode or city name.
- Matching: is to extract a very accurate and official address from the Reference Address. Using the result of parsing process, the comparison will take place stages by stages, start with postcode, city name and street name until it will narrow down up to house number.

The detail process is shown in Figure 3.0. Starting from the input known as unclean address, the input will be parsed. Parsing is a process of splitting the word and classification into the correct category. In determining the correct arrangement of address component, the process tagging is embedded. Matching will be comparing the input address and the knowledge repository. The comparison will be done in stages as to ensure the correct match had been performed. The snapshot of the prototype has presented in Figure 4.0.

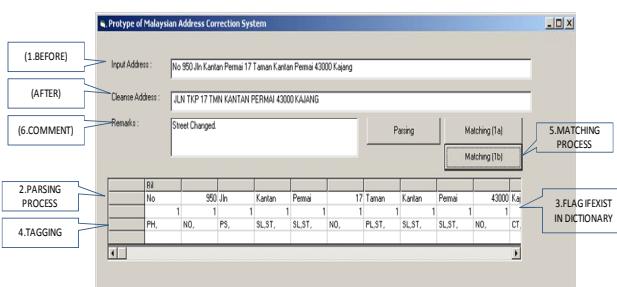


Figure 4.0 Snapshot of prototyping for Malaysian Address Cleansing.

EXPECTED CONTRIBUTION

The main aim of this paper is to address the gap between standard address format defined by Malaysia Standard department and the real practice in address writing. Hence, this paper will propose address standardization.

The process of address standardization will have a positive impact on many areas. Postal department shall have a direct impact due to address standardization. Mail can be delivered timely and accurately. In the perspective of Geographical Information System (GIS), standard address will be an advantage in Location Based Services using GPS devices. By entering the standard addresses in the GPS device, the specific location can be traced accurately. In business point of view, the address standardization is a primary step in data cleansing of a data warehouse [13]. A result shall produce more accurate on market segmentation information.

In the context of enriching the body of knowledge, very few studies found to embark in Malaysian address standardization. This methodological approach adapts the Malaysian culture, language and rules.

CONCLUSION

The advancement of ICT and the Internet in particular has led to the emergence and proliferation of electronic mailing or email and text messaging service. This however, has not affected the growing trend of postal mail delivery services. Hence, in line with the sophistication of the ICT, continuous studies have to be conducted to enhance the existing postal mail services.

REFERENCES

- [1] P.Christen and D.Belacic. “Automatic Probabilistic Address Standardisation and Verification,” Proc. Data Mining Conference(AUSDM’05), 2005.
- [2] G.Kim and S.Lee, “Analysis of postal address fields for efficient encoding of Korean mail pieces.
- [3] P.Jorcin, A.F.Hamzah, H.S.Norhisham, H.Daud and S.Sarip. “Geocoding and Reverse Geocoding Application. Case studies of implementation in Malaysia for Municipalities and Real Estate agencies.” Proc. MAP ASIA 2008. Aug 2008.
- [4] SIRIM.”Malaysian Standard: Addresses-Standard Format – Requirements.” Department of standards Malaysia, MS2039-2006. (2006).
- [5] A.R.D Prasad and Nabonita Guha.Concept naming vs concept categorisation: a faceted approach to semantic annotation.” Online Information Review, vol. 32 No. 4, pp.500-510, Dec 2007.
- [6] J.Evermann, and W.Yair. “Ontology Based Object-Oriented Domain Modeling”. Journal of Database Management, vol.20(1),pp 48-75,2009.
- [7] L.B.David ,W.S.Edmund,J.A.Stuart and K.Pinaki. “An Introduction to Semantic Modeling For Logistical Systems.” Journal of Business Logistics, vol.26- 2,pp 97-117, 2005.

- [8] G.Avigador, M.Giovanni,H.Jamil and E.Ami. "Automatic Ontology Matching Using Application Semantics." AI Magazine, Vol.26(1),pp 21-31, Spring 2005.
- [9] Wikipedia "Address (Geography)" [http://en.wikipedia.org/wiki/Address_\(geography\)](http://en.wikipedia.org/wiki/Address_(geography))
- [10] Sargur N.Srihari, Wen-jann Yang and Venugopal Govindaraju."Information Theoretic Analysis of Postal Address Fields for Automatic Address Interpretation," Proc. International Conference on Document Analysis and Recognition (ICDAR99),1999.
- [11] P.Nagabushan,Angadi,S.A. and B.S.Anami, "A Knowledge based Fast PIN code Validation System for Dispatch Sorting of Postal Mail," Proc. International Conference on Cognitive Systems(ICCS2004),2004.
- [12] P.Nagabushan,Angadi,S.A. and B.S.Anami, "Symbolic Data Structure for Postal Address Representation and Address Validation Through Symbolic Knowledge Base," PReMI2005, LNCS 3776,pp388-394,2005.
- [13] Rabeeh Ayaz Abbasi,"Information Extraction Techniques for Postal Address Standardization",Proc. International Multitopic Conference, IEEE INMIC 2005.