Enlarging Master Data Management with Machine Learning to Provide Solutions for Healthcare Industry

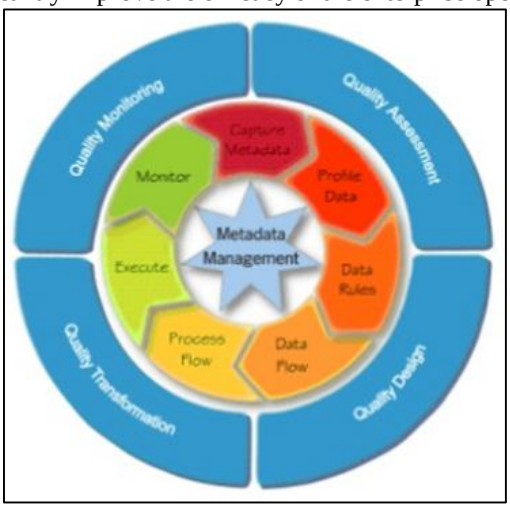
Abhishek Choudhury
Software Development Engineer in CSG R&D
Cisco Systems India Pvt. Ltd, Bengaluru, Karnataka, India

Abstract— With the evolvement of technology in the field of computer science such as Digital media, Artificial intelligence, data science, etc. in various fields, which has helped many industries to flourish over the period. However, with the emergence of copious challenges in the healthcare industry, it leaves lots of gaps, hence a lot of space for further improvement. MDM (Master data management)- Master data is a type of data that describes subjects related to the 'who', 'what' and 'where' in business transactions communications and events. The concept of MDM can effectively be used in the field of healthcare; stakeholders like pharmaceutical organizations, qualified doctors and medical representatives can be hugely benefited. In this paper, we aim to discuss the existing system in places, and how machine learning and artificial intelligence can contribute to take MDM into the next level.

Keywords: MDM, Healthcare Industry, Machine Learning

I. Introduction

Master data are mainstream business objects apportioned in the middle of operational and analytical systems that incorporate particulars about customers, suppliers, organizational units, as well as regulatory reference information (RRI) that is necessary for the enterprise functioning. The foremost challenge in MDM is to ensure its probity, coherence, and steadiness. Large enterprise vendors proffer their solutions for MDM. The best-known products in this area are Informatics MDM, SAP NetWeaver MDM, Oracle MDM suite, etc. These solutions make it possible to significantly improve the efficacy of the enterprise operation.



However, many (especially small and medium) enterprises undermine the prodigious benefits of the use of master data management systems allowing a significant reduction of transaction costs. A major obstacle to their use is that such systems are high-priced and require an utmost qualified specialist. Nevertheless, the potential positive economic effect, which may achieve several orders of magnitude over expenses, stimulates the search for new ways to solve this problem.

II. Background

In today's practice, MDM is used in healthcare in a very ineffective way. Trivia- Pharmaceutical companies received doctors profile through multiple data sources; Each data source provides various information of doctors such as their personal information, specialization, educational qualifications, primary address, secondary address etc. The data will have the same doctor's information through various data sources yet, the quality and the accuracy of the information may be different. Example: Consider there is a pharmaceutical company A receives doctor X's profile through three different data sources: P, Q, and R. Furthermore, data source P, Q provides the same information but, the quality of data is more than 90%, and the data provided by data source R has some additional information about the doctor such as his secondary address as well as his qualifications. Now, the primary task is to collect all the data from each data source and manually identifies and match the similar profiles across all the data sources and upload them into their portal.

A. Goal:

Further, Marketing representatives access their portal, specific to the location and approach the doctors; based on the medicine catalogue they possess. This process helps them to disseminate their product globally, in return it benefits their overall revenue.

Below are the following challenges in the existing design:

1) Accuracy: Since this process is done manually, therefore its accuracy gets compromised.
2) Time: The overall process is time-consuming, and the entire procedure becomes cumbersome when the data received is huge.
3) Data corruption: If the quality of data provided by the data source is inaccurate then the profile will not be considered for matching. Likewise, if a certain amount of information is not matched then also the record will not be considered to store in the portal.

III. Proposed Solution

A. Data Loading

Firstly, the raw data needs to be loaded in the staging tables, which we have received from multiple data sources. There are various tools available in the existing market which can assist us to perform the same. e.g. Oracle data loader.

B. Data Integrity and Normalization

Secondly, we need to process the data and improve the quality of it. For instance, if there are unwanted special characters in name or address then those characters need to be clean off.

Then we need to check the integrity of the data based on its nature and create integrity wherever it is needed. (United States Patent No. US 8341131 B2, 2010)

C. Data Quality and Scoring

Further each record should be passed through the spectrum of rules to gauge the quality of the data; based on the observation every record should be marked with a quantitative score. This step aims to choose the best data among various item in the set.

D. Matching

This is the most critical step in the entire process because it provides the authority to decide whether the profile is a candidate for master data management, needs to be discarded or to archive it form future use. The following are the possible outcome-based on the matching rules:

1) Consider we have set A= {a, b, c, d}- Set A has a list of doctors received from data source A. Likewise, set B= {p, a, b, r} has a list of doctor's profile received from data source B. Now, the matching process will match the elements of dataset A with respect to the elements of dataset B. (Christen, 200)
2) After the above process is finished, elements- c and d from set A will be discarded for the master data selection. Similarly, elements p and r from set B will also be discarded as a potential candidate for master data. Now, we have an amalgamated set; which is generated post matching process and it has elements 'a' and 'b'. Whereas, elements 'a' and 'b' is selected from the respective Set A and Set B based on the data quality i.e. scoring given in the "data quality and scoring". (Ravikumar & Fienberg, 2003)

E. Matching Algorithm

Since the nature of the data will mostly be a type of string or set of characters, therefore, string distance algorithms are the best approach to calculate the match percentage. Following are the algorithms which can be used:

1) Jaro-Winkler algorithm- The Jaro-Winkler distance is a string metric measuring an edit distance between two sequences. It is a variant proposed in 1990 by William E. Winkler of the Jaro distance metric.

The Jaro Similarity Simj of two given strings S1 and S2 is

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

Where:

- |Si| is the length of the string Si.
- m is the number of matching characters.
- t is half the number of transpositions.

Two characters from S1 and S2 respectively, are considered matching only if they are the same and farther than

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1.$$

2) Levenshtein algorithm- The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

Mathematically, the Levenshtein distance between two strings, a and b (of length |a| and |b| respectively), is given by lev a,b(|a|,|b|) where:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Here, 1(ai≠bi) is the indicator function equal to 0 when ai≠bi and equal to 1 otherwise, and leva, b(i,j) is the distance between the first i characters of a and the first j characters of b.

3) Deciding which to use is not just a matter of performance. It's important to pick a method that is suited to the nature of the strings you are comparing. In general, though, both of the algorithms you mentioned can be expensive, because each string must be compared to every other string, and with millions of strings in your data set, that is a tremendous number of comparisons. That is much more expensive than something like computing a phonetic encoding for each string, and then simply grouping strings sharing identical encodings.

IV. ENHANCEMENTS

A. Approach to Machine Learning

The crux of the proposed solution has the biggest challenge in achieving data quality. Also, it may become vague for the organization to simply discard data if it is not in the proper format or prohibiting the MDM data rules. In the same way, over a while, it may become onerous for the data source providers to adhere to the quality rules.

1) Assume that your MDM accepts a street address in the following format- 29A, Redmond Road, Bengaluru. However, your data source provided you data in the following format: Redmond Road, 29A, Bengaluru. In this case, the traditional MDM application is bound to discard the data.
2) Furthermore, in this kind of situations, Deep learning will take the MDM into the next level. The ability to make the decisions and judgements based on the nature of data will vastly improve the data quality without compromising the loss or corruption of data. Machine learning can make MDM loosely coupled with the data sources i.e. organizations will no more be dependent on the quality of data received from the data sources. Indeed, they will be capable of rectifying the errors, standardize if required, and make judgements accordingly.
3) ML permits enterprises to unearth patterns in data, as well as propose associations, correlations, and adaptation. As the system learns more about data, it eclipses traditional extract-transform-load (ETL) approaches making it a thing of the past.

V. CONCLUSION

It is true that the existing healthcare industry has many gaps that need innovation to fill the gaps. Pharmaceutical companies face many challenges in disseminating their

product, Likewise, many qualified doctors are not reachable to the marketing teams of the famous pharmaceutical companies due to which practitioners are getting away from the latest and trending medicines or the technology. This gap is the crux of modern health problems, which is impacting our society. Therefore, after using MDM effectively, and integrating machine learning, we can close the gaps.

REFERENCES

[1] Christen, P. (200). A comparison of Personal Name Matching: Techniques and Practical Issues. Canberra: September 2006.
[2] Cohen, R. (2010). United States Patent No. US 8341131 B2.
[3] L, G., P, I. G., H, J., N, K., S, M., & Shrivastava, D. (2001). Approximate database joins in a database (almost) for free. 491-500.
[4] Ravikumar, P., & Fienberg, E. (2003). A comparison of string distance metrics for name-matching tasks. 73-78.

