International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 180-184
Published Online May 2022 in IJEAST (http://www.ijeast.com)

MATCHING UNSTRUCTURED WEB DATA WITH STRUCTURED MATERIAL DATA FOR PRICE REFERENCE

John Bhaaswanth D
Manager
Centre of Excellence –Advanced Analytics
L&T Construction,
Chennai, Tamil Nadu, India

Chandrabose M
Senior Manager
Centre of Excellence – Advanced Analytics
L&T Construction,
Chennai, Tamil Nadu, India

Balaji B
Lead
Centre of Excellence – Advanced Analytics
L&T Construction,
Chennai, Tamil Nadu, India

***Abstract*—To leverage the power of e-commerce web data for procurement with the material prices across the spectrum of categories, an additional reference tool was developed based on NLP (Natural Language Processing), Elastic Search etc. This recommender mechanism aids the procurement team to efficiently negotiate the prices and analyze new supplier options for an organization. This paper covers a crucial step of the process which is a methodology to match external material data from web with internal material data of an organization and how it helps business to achieve its objective.**

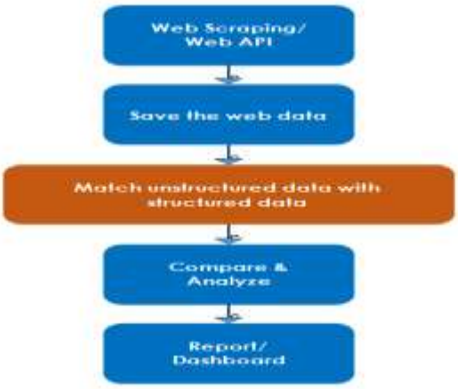
***Keywords*—Text Matching, Text Analytics, Natural Language Processing, Machine Learning, Elastic Search, Regular Expressions, Web Data, Unstructured Data, Structured Data, Boosting, Synonyms, Multi match, e-commerce Data, Construction, Material, Procurement.**

I. INTRODUCTION

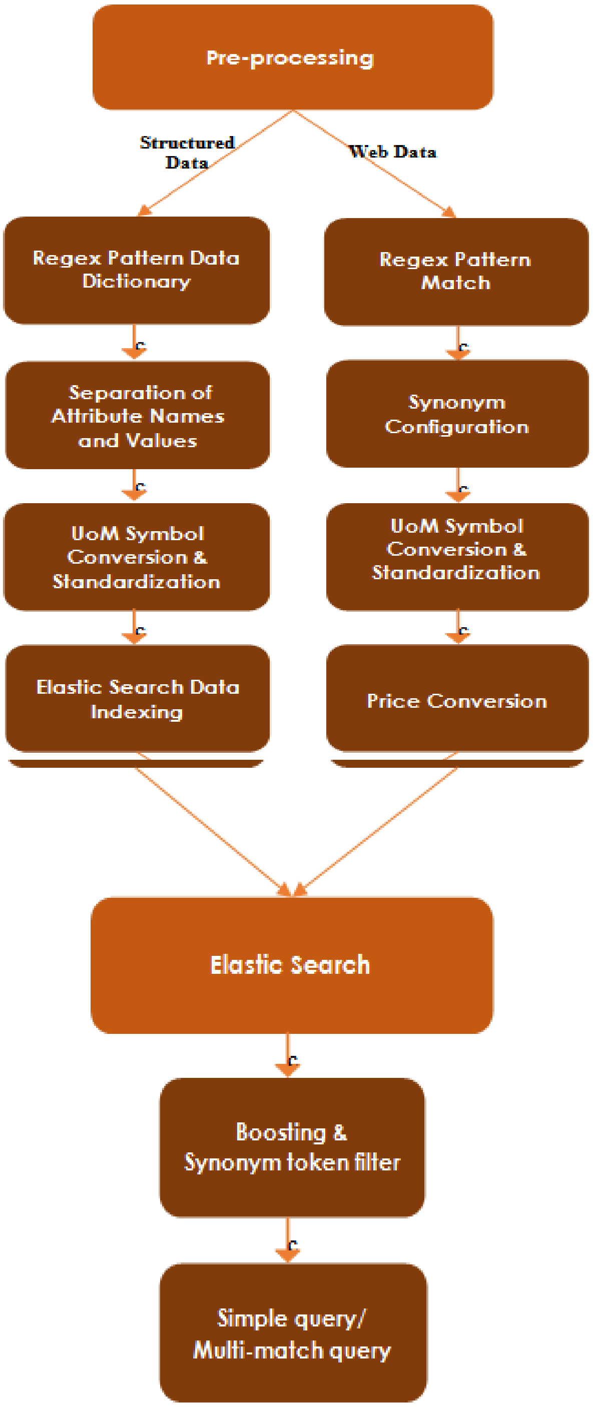
One of the key strategy of materials procurement team in construction firms is to negotiate hard and procure materials at an optimum price possible. While they negotiate harder based on existing available data, this data driven approach by leveraging e-commerce data would validate the claim of procuring at best discounted price. So, mapping of material purchased price with the existing price available online through web data is done by web scraping or web API (Application Programming Interface) and NLP (Natural Language Processing) analytics. This will help minimize the cost spent on material by leveraging existing data in the e-commerce. It is an intelligent process which gives a consolidated matched list of existing procured materials and e-commerce materials with comparison of current purchase rate with e-commerce price from various e-commerce web-sites (that allows scraping).

Here, the main focus is on solving a major pain point in this process that is matching unstructured web data with structured or semi-structured data which are of different taxonomies. The detailed approach on how this problem is addressed is explained further in detailed steps.

A. Overview of Process Flow



B. Matching Layer - Detailed Flow


Pre-processing
Structured Data
Web Data
Regex Pattern Data Dictionary
Regex Pattern Match
Separation of Attribute Names and Values
Synonym Configuration
UoM Symbol Conversion & Standardization
UoM Symbol Conversion & Standardization
Elastic Search Data Indexing
Price Conversion
Elastic Search
Boosting & Synonym token filter
Simple query/ Multi-match query

II.PRE PROCESSING – STRUCTURED DATA

A.Material Name – Regex Pattern Data Dictionary

Distinct material names are taken and probable regex (Regular Expression) patterns with jumbled words, plural scenarios are prepared and stored as regex patterns for each distinct material name. These regex patterns are further used to extract material names from web data.

Example:
Material Name: UPVC PIPE
Regex Pattern 1: \\bUPVC\\b.*PIPE
Regex Pattern 2: \\bPIPE .*\\bUPVC\\b

B.Attribute Names and Values

Attributes in structured data are separated as Attribute Names and Attribute Values for better matching. Tags section is used for values without attribute names.

Example:
"attribute_keys": ["CLASS", "TYPE OF JOINT", "DIAMETER", "RING STIFFNESS"],
"attribute_vals": ["CLASS 6 (12.5 KG/CM2)", "RUBBER GASKET TYPE", "DN 300 MM", "SN6"],
"tags": []

C.Unit of Measurement Symbol Conversion

In attribute values, unit of measurement symbols are converted to their respective names by replacing symbol with its relevant name.

Example:
2 “ => 2 INCH
30 ° => 30 DEGREE

D.Unit of Measurement Standardization

Observe units of measurement of structured data and arrive at a common/company’s standard unit for better comparison of attribute values of similar measurements.

Example:
2 INCH => 50.8 MM

E.Attribute Names and Values

Structured data with material names, attribute names, attribute values and tags after the mentioned steps of pre-processing are indexed using Elastic Search indexing. Each data point needs to have a unique id. This structure can be modified or customized according to the requirement.

Example:
"Id": "6CXXMXX7***",
"material_name": "UPVC PIPE",
"attribute_keys": ["CLASS", "TYPE OF JOINT", "DIAMETER", "RING STIFFNESS"],
"attribute_vals": ["CLASS 6 (12.5 KG/CM2)", "RUBBER GASKET TYPE", "DN 300 MM", "SN6"],
"tags": []

III. PRE PROCESSING – WEB DATA

A.Regex Pattern Match

Regex (Regular Expression) patterns from the data dictionary created using structured data are used to match and extract material name from web data using Regex pattern match. Some texts can have multiple regex pattern matches which would be stored corresponding to each data point. The

material names corresponding to matched regex patterns are stored and further used as filters. This is the primary step to remove irrelevant web data.
Example:
Web Data: “Finolex 4 inch UPVC Pipes, 6m”
Matched Regex Pattern: \\bUPVC\\b.*PIPE
Corresponding Material Name for this RegexPattern: UPVC PIPE

B.Synonym Configuration

To address matching for acronyms and their abbreviations, synonym configuration is defined for all the acronyms and their respective abbreviations. Plural words can be included in the synonym configuration for better matching.
Example:
STAINLESS STEEL => SS
UNPLASTICIZED POLYVINYL CHLORIDE => UPVC
BEARING, BEARINGS
SLEEVE, SLEEVES

C.Unit of Measurement Symbol Conversion

Unit of measurement symbols are converted to their respective names based on the observation of web data and structured data. This can be achieved by replacing symbol with its relevant name.
Example:
2 “ => 2 INCH
30 ° => 30 DEGREE
60 DEG => 60 DEGREE

D.Unit of Measurement Standardization

Based on the structured data, arrive at standard unit of measurement at material or material group level. Attribute values of web data are then converted to standard unit of measurement.
Example:
2 INCH => 50.8 MM

E.Price Conversion

Depending on the type of currency, price value is converted to one currency for better comparison.
Example:
2 USD => □ 144
2 DOLLARS => □ 144
$ 2 => □ 144
US $ 2 => □ 144
Rs. 144 => □ 144

IV. ELASTIC SEARCH

A.Index Creation

Index in Elastic Search is used to store documents (structured data in our scenario) in dedicated data structures which are defined as per the requirement. While loading documents to the index, a non-existing name has to be given to the index which is referred as index name. There is no need to specifically create index separately in Elastic Search as this would be created automatically while loading documents just by giving a non-existing index name.

B.Indexing Documents

As Elastic Search is a document oriented framework, data is indexed in the form of documents. In this context, the structured data is stored and indexed as documents. Through indexing, documents can be created or updated. After indexing, many actions can be performed on index documents like search, filter and sort complete documents.
Example:
Structured data with material IDs, material names, attribute names, attribute values and tags after the steps of pre-processing are indexed using Elastic Search indexing.

C.Boosting

While querying, boosting feature can be used to enhance and give more importance to material name or required attribute during matching for better matching results according to the requirement.
Example:
In the below example, “material_name” is boosted by 10 points, “attribute_keys” is boosted by 2 points and “attribute_vals” is boosted by 5 points to the overall score.

```
{"query": {
    “bool": {
            “must": [
{"match": { "material_name":   "UPVC PIPE"       }},
{"multi_match": { "query" : " Finolex 100 MM UPVC Pipes,
6 M ", "type" : "best_fields",
"fields":["material_name^10","attribute_keys^2",
“attribute_vals^5” ]
}}]
             }
         }
}
```

D.Synonym Token Filter

While querying, synonym token filter can be used along with boosting feature to address matching for acronyms and their abbreviations there by enhancing overall matching logic scenario.
Example:
In the below example, “STAINLESS STEEL” is considered as “SS” while matching. Similarly, “BEARING” and “BEARINGS” are considered as same while matching.

```
PUT /index_name
{
  "settings": {
    "index": {
      "analysis": {
        "analyzer": {
          "synonym_analyzer": {
            "tokenizer": "whitespace",
          "filter": ["uppercase", "my_synonyms"]
         }
       },
       "filter": {
         "my_synonyms": {
          "type": "synonym",
          "synonyms": [
"STAINLESS STEEL => SS”,
“UNPLASTICIZED POLYVINYL CHLORIDE => UPVC”,
“BEARING, BEARINGS”,
“SLEEVE, SLEEVES”
] }
   }}
  }
}
```

E.Querying

Full text search with multi match can be performed by using Elastic Search query. Input to query would be processed web data (after completion of all steps of pre-processing) along with matched material name based on Regex pattern match (III-A step). Matched name processed from III-A step is input to “match” section in the query and processed web data is input to “multi_match” query. Boosting and Synonym filter can be added in the “multi_match” query for better matching results.

Example:

```
{"query": {
    “bool": {
            “must": [
{"match": { "material_name":   "UPVC PIPE"       }},
{"multi_match": { "query" : "Finolex 100 MM UPVC Pipes, 6
M ", "type" : "best_fields",
"fields":["material_name^10","attribute_keys^2",
“attribute_vals^5” ],
"analyzer": "synonym_analyzer"}}]
            }
}
}
```

F.Results

After querying, matched results are received in descending order of their Elastic Search score. Higher scored results are better matched. If nothing is matched, there will be no results. Approximately, 80-85% of accuracy can be achieved with this approach.

The best matched results for the processed web data can be stored in a dataframe for further analysis.

Example:

```
"hits": [
  {
    "_index": "index_name ",
    "_id": "6CxxxxUM0xxx0",
    "_score": 1.3278645,
    "_source":
          "product_name": "UPVC PIPE",
          "attribute_keys": [
             "STANDARD",
             "TYPE OF JOINT",
             "DIAMETER",
             "LENGTH"
          ],
          "attribute_vals": [
             "IS 13592",
             "PLAIN",
             "DN 100 MM",
             "6 M"
          ],
          "tags": []
  },

  {
    "_index": "index_name ",
    "_id": "6CxxxxUM2xxx1",
    "_score": 1.1298637,
    "_source":
          "product_name": "UPVC PIPE",
          "attribute_keys": [
             "STANDARD",
             "TYPE OF JOINT",
             "DIAMETER",
             "LENGTH"
          ],
          "attribute_vals": [
             "IS 13592",
             "PLAIN",
             "DN 100 MM",
             "10 M"
          ],
          "tags": []
  }
]
```

V.COMPARE & ANALYZE

The consolidated matched results of processed web data and the relevant structured data are used to compare purchase price of the firm with e-commerce price. Material wise minimum, median, maximum prices of e-commerce can be calculated and analyzed with firm’s existing purchase price.

This gives insights on areas where the firm is lacking and can re-negotiate with vendors or re-strategize their procurement plans to optimize the cost spent on materials. Also, this analysis can provide new list of vendors/suppliers who are offering the product/services at lowest price possible. It also helps to validate and confirm whether the existing purchase price is better than available e-commerce price, which would be a vital information for the purchase team.

This analysis also provides continuous month-on-month savings (calculated by computing the difference between the offered price and negotiated price multiplied by the purchase quantity by referring to this platform). There is a scope to update internal price reference database for future reference.

VI.CONCLUSION

The major pain point in this process of comparing various material descriptions that is matching unstructured web data with structured or semi-structured data which are of different taxonomies is now addressed with this approach. This methodology can be used for purchase requirement across various companies to check the existing market price where there is a perfect match available between the company's material taxonomy and e-commerce material taxonomy.

The same approach can be reused in various other domains by tweaking processes in pre-processing and querying phases according to the domain requirement. Also, this approach can be further modified by adding certain steps in pre-processing phase and also by applying other matching frameworks like Azure Search or any other methodologies instead of Elastic Search based on the requirement and ease of use. Future studies should look at the possibility of implementing robust reinforcement learning to improve the accuracy.

VII.ACKNOWLEDGMENT

The authors wish to thank Mr. R Ganesan (Head - Corporate Centre, L&T Construction) and Mr. Binaya Bhusan Mishra (Head – Supply Chain Management, L&T Construction) for their continuous support and encouragement.

VII. REFERENCE

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-match-query.html

[2] https://www.elastic.co/blog/boosting-the-power-of-elasticsearch-with-synonyms

[3] https://www.elastic.co/guide/en/elasticsearch/reference/current/full-text-queries.html

[4] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-multi-match-query.html

[5] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-boosting-query.html

[6] https://www.javatpoint.com/regex

[7] https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html