Towards Robust Entity Matching: Cases Studies for the Pharmaceutical Databases

Aleksei Zhukov   Adelina Gamaiunova   Denis Sidorov   Aleksander Novolotsky

Aptekar IC and Irkutsk National Research Technical University

Abstract

Entity matching is one of the principal challenges for success of the **e-commerce development**. It makes it possible to identify whether two entity records refer to the same real-world entity, i.e. matching the product offers from different vendors with the clients requests.

- The overview of the state of the art methods for attribute-based matching methods and systems is discussed.
- The approach to entity matching is proposed based on the fuzzy logic, machine learning and using the cosine distance.
- The efficiency of the proposed system is demonstrated on the pharmaceutical databases provided by the pharmaceutical aggregator.

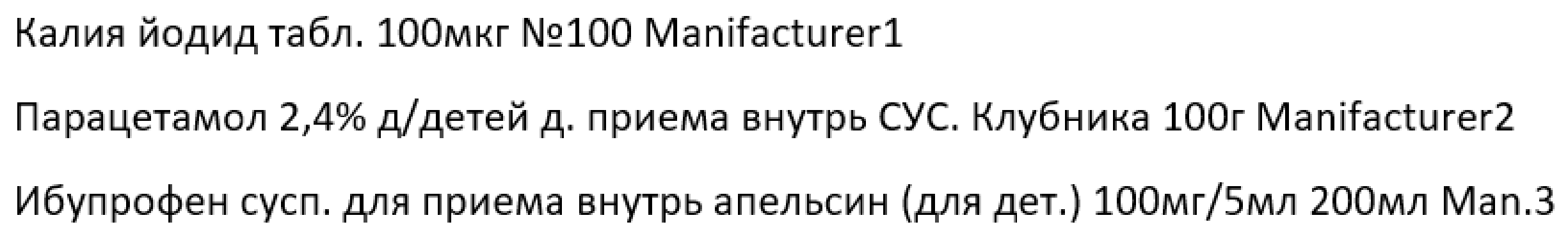
Motivation

Entity matching (EM) is the well known challenge in the database community. Its objective is to identifying which records refer to the same real-world entity (e.g., stock keeping units, SKU) among heterogeneous data sources. It is also known as **"entity alignment"** to make it possible to judge whether entities refer to the same object in the real world. This field attracts a lot of attention due to the rise of electronic trading worldwide.

Despite the high practical value, the systematic approach and technologies overview is mainly missing in the literature.

We need to compare nomenclatures of medical products, medicines from different sources, such as pharmacies, hospitals, and others. This is a complex and time-consuming task.

Pharma Data.



Калия йодид табл. 100мкг №100 Manifacturer1

Парацетамол 2,4% д/детей д. приема внутрь СУС. Клубника 100г Manifacturer2

Ибупрофен сусп. для приема внутрь апельсин (для дет.) 100мг/5мл 200мл Man.3

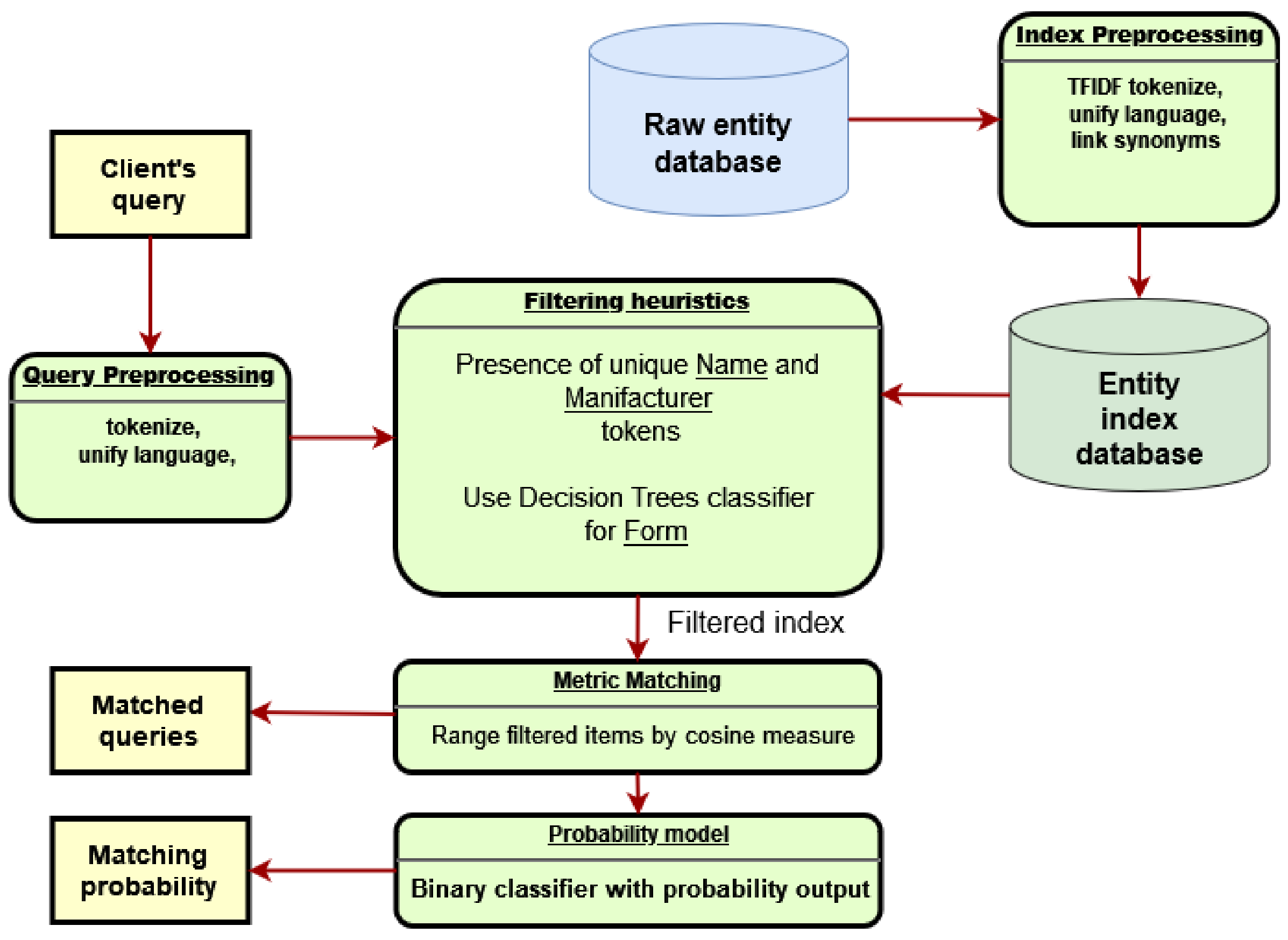
Methodologically EM problem can be soved by applying many heuristics during preprocessing stage with classical text similarity methods or using metric learning-based approaches.

The Heuristic-based Approach

The *heuristic-based approach* could seem straightforward but combination of heuristics is found to be efficient. Additionally, fuzzy matching is applied in the matching process during tokenization. This allows to take into account possible misprints, spelling errors and spelling variations in nomenclatures.

The encoding of each nomenclature string relies on a bag-of-words vector representing. To determine the similarity between the query and the nomenclature dictionary, cosine distance.

For matching of nomenclature name types, the TFIDF (Term Frequency-Inverse Document Frequency) method proves to be effective due to the fact that many item names are almost unique and occur only a few times.



Raw entity database
Index Preprocessing
TFIDF tokenize, unify language, link synonyms
Client's query
Query Preprocessing
tokenize, unify language,
Filtering heuristics
Presence of unique Name and Manifacturer tokens
Use Decision Trees classifier for Form
Entity index database
Filtered index
Matched queries
Metric Matching
Range filtered items by cosine measure
Matching probability
Probability model
Binary classifier with probability output

Figure 1: Heuristic-based process flow architecture

Prior to vectorization, each string undergoes a preprocessing stage that involves the removal of special characters, punctuation, and stop words. Abbreviations and synonyms are standardized for consistent representation, and numerical designations for dosage and volume are uniformly written. Additionally, English names are transcribed into phonetic-like representation to enhance language consistency.

Seamise network-based approach

Despite of the fact that heuristic based approach can be effective on semi-standartized data, it has lack of robustness. That's why we have tried another method. The Siamese neural network was trained on a marked dataset containing of triplets (query, positive matching sample, negative sample). The network exploits **triplet loss with construction of hard triplets**.

$$\mathcal{L}\left(A,P,N\right)=\max\left(\|\,\mathrm{f}\left(A\right)-\mathrm{f}\left(P\right)\|_{2}-\|\,\mathrm{f}\left(A\right)-\mathrm{f}\left(N\right)\|_{2}+\alpha,0\right)$$

We find negative samples that have minimal distance to positive to build dataset.

Each nomenclature passes through the network and gets a vector representation. Then, the vector representations compared using cosine distance. We use **mean reciprocal rank** as quality metric.

Experiment results & Conclusions

To evaluate the effectiveness of the proposed method, we conducted experiment on a dictionary of medical products. In the dataset there were 600,000 items that needed to be matched between a pharmacy and a hospital. Each item description contains item name, manufacturer, number (for example, number of pills), dosage, from of manufacturing and other characteristics. Usually each hospital query looks like: Name, Dosage, Form, Number, etc. But sometimes it highly varies from institution to institution. We managed to achieve **92.542% accuracy on institution-specific data**, indicating the high effectiveness of the proposed method. However method can provide lower accuracy on **unspecified data which is about 68%**.

It makes the system non-robust to format changes, because of that we try fine-tune pre-trained transformer model on our data. Our results is still intermediate, but we got up to **70% quality on our data with BERT**. It can be much better with specific tokenizer which is aware or medicines terms. Except this we going to consider numeric content separately.

The proposed method can be applied in practice to the automatic comparison of nomenclature of medical goods, for example, between a pharmacy and a hospital. This will simplify and speed up the process comparison, reduce the number of errors and increase the efficiency of inventory management and search for alternative items.

Contact Information

- Web: IC Aptekar
- aptekar@irk.ru

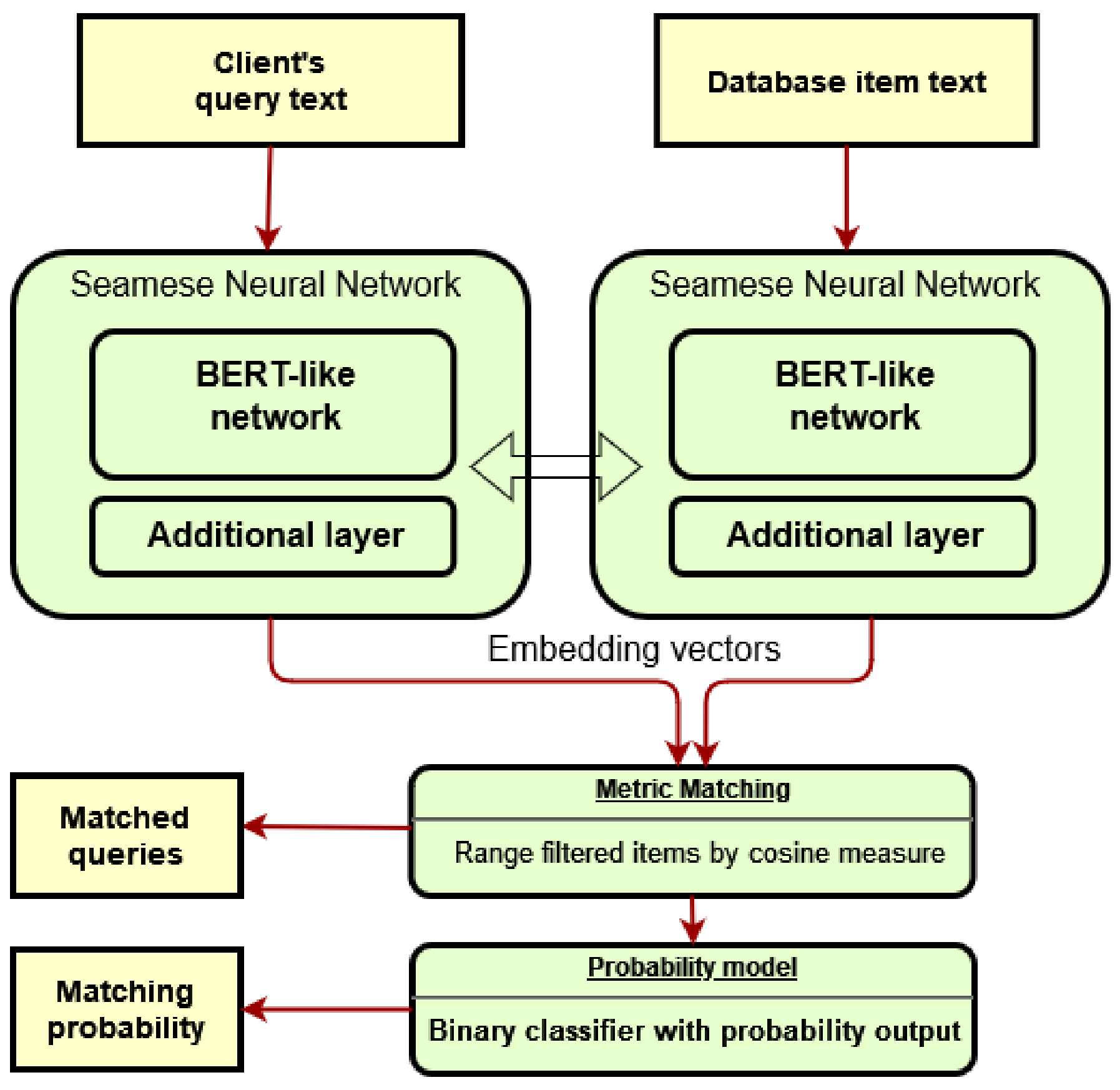


Figure 2: Seamise fine-tuned BERT network-based process architecture