(12)Indian Patent Application

(21) Application Number: 202341080898

(22) Filing Date: 28/11/2023 (43) Publication Date: 30/05/2025

(71) Applicant(s): L&T TECHNOLOGY SERVICES LIMITED

(72) Inventor(s): Gogoi, Pinak Pani
Kanth, Ashwin
Srivastava, Prakhar
Singh, Madhusudan

(51) International Classifications: G06N 20/00 G06F 16/9535 G06F 16/2457 G06F 16/957 G06F 16/93

(54) Title: METHOD AND SYSTEM FOR ANALYSING DOMAIN SPECIFIC DOCUMENTS

(57) Abstract: A method (600) and system (100) of analysing domain-specific documents is disclosed. A processor (104) determines one or more attributes in a first text data extracted from the plurality of domain-specific documents based on a first set of queries (400A) using a first generative machine learning (ML) model. A second set of queries (400B) are determined based portions of the first text data corresponding to the attributes using a second generative ML model. A third set of queries (400C) are determined based on domain type of the domain-specific documents and a second text data extracted for a co0rresponding domain-specific document from the domain-specific documents. A fourth set of queries (400D) are determined based on an iterative comparison between the second set of queries (400B) and the third set of queries (400C) until the fourth set of queries (400D) are determined as about same as second set of queries (400B).

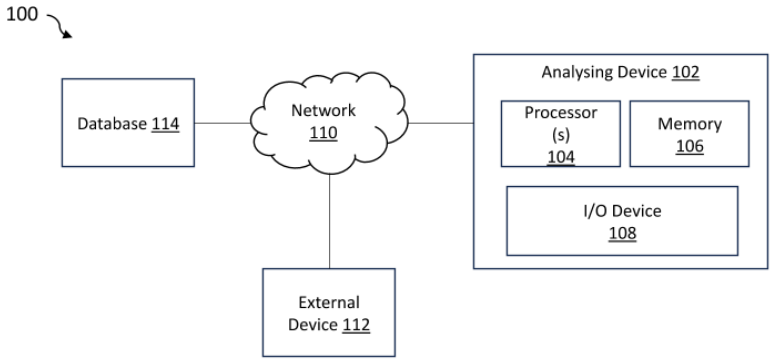



FIG. 1

# FORM 2

THE PATENTS ACT 1970
(39 OF 1970)
&
The Patent Rules, 2003

# Complete Specification
(See Section 10 and Rule 13)

**1. TITLE OF THE INVENTION**

**METHOD AND SYSTEM FOR ANALYSING DOMAIN SPECIFIC DOCUMENTS**

**2. APPLICANT(S)**

(a) NAME : **L&T TECHNOLOGY SERVICES LIMITED**

(b) NATIONALITY : **INDIAN**

(c) ADDRESS : DLF IT SEZ Park, 2nd Floor – Block 3
1/124, Mount Poonamallee Road,
Ramapuram, Chennai – 600 089,
INDIA.

**3. PREAMBLE TO THE DESCRIPTION**

**COMPLETE**

The following specification particularly describes the invention and the manner in which it is to be performed

## DESCRIPTION

### Technical Field

**[001]** This disclosure relates generally to data extraction and analysis, and more particularly to a method and a system for analysing domain specific document.

## BACKGROUND

**[002]** Many organizations, each with their own unique goals and purposes, frequently find themselves in the position of having to review a multitude of documents. However, the traditional method of manually inspecting these documents and extracting information from them can be an incredibly laborious and time-consuming task. This challenge becomes even more pronounced when dealing with domain-specific documents like resumes, medical records, legal contracts, research papers, financial statements, etc. as their content and format can vary significantly from one document to another. Conventional methodology uses rule-based approach for data extraction from domain-specific documents. However, domain specific application and non-scalability of rule-based approach makes it not suitable for documents having variable content and format.

**[003]** Therefore, there is a requirement for an efficient and effective methodology to analyze domain specific documents.

## SUMMARY OF THE INVENTION

In an embodiment, a method of analysing a plurality of domain-specific documents is disclosed. The method may include, determining by a processor, one or more attributes in a first text data extracted from the plurality of domain-specific documents based on a first set of queries using a first generative machine learning model. The method may further include determining, by the processor, a second set of queries based on one or more portions of the first text data corresponding to the one or more attributes using a second generative machine learning model. The method may further include determining, by the processor, for each of a portion the plurality of domain-specific documents, a third set of queries based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document from the portion of the plurality of domain-specific documents. In an embodiment, the domain type may be determined based on a subject matter analysis of the plurality of domain-specific documents. The method may further include, determining by the processor, a fourth set of queries based on a comparison between the second set of queries and the third set of queries. In an embodiment the first set of queries may be iteratively updated based on the corresponding fourth set of queries determined for each of the

corresponding domain-specific document from the portion of the plurality of domain-specific documents. The method may further include determining by the processor, a final set of queries corresponding to the plurality of domain-specific documents based on the fourth set of queries determined as about same as the second set of queries.

**[004]** In another embodiment, a system of analysing a domain specific document is disclosed. The system may include a processor, a memory communicatively coupled to the processor, wherein the memory may store processor-executable instructions, which when executed by the processor may cause the processor to determine a first set of queries based on determination of one or more attributes in a first text data extracted from the plurality of domain-specific documents based on a first set of queries. The processor may further determine a second set of queries based on one or more portions of the first text data corresponding to the one or more attributes using a second generative machine learning model. The processor may further determine, for each of a portion of the plurality of domain-specific documents, a third set of queries, based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document from the portion of the plurality of domain-specific documents. In an embodiment the processor may determine the domain type based on a subject matter analysis of the plurality of domain-specific documents. The processor may further determine a fourth set of queries based on a comparison between the second set of queries and the third set of queries. In an embodiment, the first set of queries may be iteratively updated based on the corresponding fourth set of queries determined for each of the corresponding domain-specific document from the portion of the plurality of domain-specific documents. Further, the processor may determine a final set of queries corresponding to the plurality of domain-specific documents based on the fourth set of queries determined as about same as the first set of queries.

**[005]** Various objects, features, aspects, and advantages of the inventive subject matter will become more apparent from the following detailed description of preferred embodiments, along with the accompanying drawing figures in which like numerals represent like components.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[006]** The accompanying drawings, which are incorporated in and constitute a part of this disclosure, illustrate exemplary embodiments and, together with the description, serve to explain the disclosed principles.

**[007]** **FIG. 1** illustrates a block diagram of an exemplary document analysis system for analysing domain-specific documents, in accordance with an embodiment of the present disclosure.

**[008]** **FIG. 2** illustrates a functional block diagram of the analysing device, in accordance with an embodiment of the present disclosure.

**[009]** **FIG. 3** illustrates an exemplary domain-specific document, in accordance with an embodiment of the present disclosure.

**[010]** **FIG. 4A, FIG. 4B**, **FIG. 4C** and **FIG. 4D** illustrate exemplary first set of queries, second set of queries, third set of queries and fourth set of queries respectively, in accordance with an embodiment of the present disclosure.

**[011]** **FIG. 5** illustrates a flowchart of a methodology to determine a final set of queries for analysing a plurality of domain-specific document, in accordance with an embodiment of the present disclosure.

**[012]** **FIG. 6** illustrates a flowchart of a methodology of analyzing a plurality of domain-specific documents, in accordance with an embodiment of the present disclosure.

## DETAILED DESCRIPTION OF THE DRAWINGS

**[013]** Exemplary embodiments are described with reference to the accompanying drawings. Wherever convenient, the same reference numbers are used throughout the drawings to refer to the same or like parts. While examples and features of disclosed principles are described herein, modifications, adaptations, and other implementations are possible without departing from the scope of the disclosed embodiments. It is intended that the following detailed description be considered exemplary only, with the true scope being indicated by the following claims. Additional illustrative embodiments are listed.

**[014]** Further, the phrases “in some embodiments”, “in accordance with some embodiments”, “in the embodiments shown”, “in other embodiments”, and the like mean a particular feature, structure, or characteristic following the phrase is included in at least one embodiment of the present disclosure and may be included in more than one embodiment. In addition, such phrases do not necessarily refer to the same embodiments or different embodiments. It is intended that

the following detailed description be considered exemplary only, with the true scope being indicated by the following claims.

**[015]** Given the complexity of extracting accurate information from domain-specific documents through conventional or manual methods. The present disclosure provides a method and system for analysing domain-specific documents for an accurate extraction of required information.

**[016]** Referring now to **FIG. 1**, a block diagram of an exemplary document analysis system 100 for analysing domain-specific documents, in accordance with some embodiments of the present disclosure is illustrated. The document analysis system 100 may include an analysing device 102, an external device 112, and a database 114 communicably connected to each other through a wired or a wireless communication network 110. The analysing device 102 may include a processor 104, a memory 106 and an input/output (I/O) device 108. In an embodiment, examples of processor(s) 104 may include, but are not limited to, an Intel® Itanium® or Itanium 2 processor(s), or AMD® Opteron® or Athlon MP® processor(s), Motorola® lines of processors, Nvidia®, FortiSOC™ system on a chip processors or other future processors. In an embodiment, the memory 106 may store instructions that, when executed by the processor 104, and cause the processor 104 to analyze domain-specific documents, as discussed in more detail below. In an embodiment, the memory 106 may be a non-volatile memory or a volatile memory. Examples of non-volatile memory may include but are not limited to, a flash memory, a Read Only Memory (ROM), a Programmable ROM (PROM), Erasable PROM (EPROM), and Electrically EPROM (EEPROM) memory, etc. Further, examples of volatile memory may include but are not limited to, Dynamic Random Access Memory (DRAM), and Static Random-Access memory (SRAM), etc. In an embodiment, the I/O device 108 may comprise of variety of interface(s), for example, interfaces for data input and data output, and the like. The I/O device 108 may facilitate inputting of instructions by a user communicating with the analyzing device 102. In an embodiment, the I/O device 108 may be wirelessly connected to the analysing device 102 through wireless network interfaces such as Bluetooth®**,** infrared**,** or any other wireless radio communication known in the art. In an embodiment, the I/O device 108 may be connected to a communication pathway for one or more components of the analyzing unit 102 to facilitate the transmission of inputted instructions and output results generated by various components such as, but not limited to, the processor(s) 104 and the memory 106.

**[017]** In an embodiment, the database 114 may be enabled in a cloud or a physical database and may store domain-specific documents and an initial first set of queries for each domain types of documents. In an embodiment a final set of queries may be output by analysing device 102. In an embodiment, the database 114 may store data output by an external device 112 or output generated by the analysing device 102. In an embodiment, the domain-specific document may include, but not limited to documents including text data related to a domain type from a plurality of domain types such as, but not limited to, marksheets, resumes, academic books, etc. In an embodiment, the domain-specific documents may be in one or more document formats. In an embodiment, the domain-specific documents may be in any universal file formats such as, but not limited to, jpeg, png, Portable Document Format (PDF), etc.

**[018]** In an embodiment, the communication network 110 may be a wired or a wireless network or a combination thereof. The network 110 can be implemented as one of the different types of networks, such as but not limited to, ethernet IP network, intranet, local area network (LAN), wide area network (WAN), the internet, Wi-Fi, LTE network, CDMA network, 5G and the like. Further, network 110 can either be a dedicated network or a shared network. The shared network represents an association of the different types of networks that use a variety of protocols, for example, Hypertext Transfer Protocol (HTTP), Transmission Control Protocol/Internet Protocol (TCP/IP), Wireless Application Protocol (WAP), and the like, to communicate with one another. Further network 110 can include a variety of network devices, including routers, bridges, servers, computing devices, storage devices, and the like.

**[019]** In an embodiment, the analysing device 102 may receive a request for analysing a plurality of domain-specific documents from the external device 112 through the network 110. In an embodiment, the analysing device 102 and the external device 112 may be a computing device, including but not limited to, a smart phone, a laptop computer, a desktop computer, a notebook, a workstation, a portable computer, a personal digital assistant, a handheld, a scanner, or a mobile device. In an embodiment, the analysing device 102 may be, but not limited to, in-built into the external device 112 or may be a standalone computing device.

**[020]** In an embodiment, the analysing device 102 may perform various processing for analysing the domain-specific document. By way of an example, the analysing device 102 may receive a plurality of domain-specific documents and a an initiated first set of queries. Examples of the plurality of domain-specific documents may be but are not limited to resumes, medical records, legal contracts, research papers, financial statements, etc. The analysing

device 102 may extract first text data from the plurality of domain-specific documents using a text extraction technique such as, but not limited to, optical character recognition, etc. The analysing device 102 may then extract one or more attributes from the first text data by querying a first generative machine learning model based on the first set of queries. Accordingly, the analysing device 102 may determine one or more attributes in the first text data extracted from the plurality of domain-specific documents based on the first set of queries using the first generative machine learning model. Based on the one or more attributes determined, the analysing device 102 may determine a second set of queries based on portions of the first text data corresponding to the one or more attributes using a second generative machine learning model. In an embodiment, the second set of queries may be determined based on validation of the one or more portions of the first text data corresponding to the one or more attributes based on a user input. Further, the analysing device 102 may determine, for each of a portion of the plurality of domain-specific documents, a third set of queries based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document from the portion of the plurality of domain-specific documents using a third machine learning model. In an embodiment, the second text data may be extracted using a text extraction technique such as, but not limited to, OCR, etc. Further, the analysing device 102 may determine the domain type based on a subject matter analysis of the plurality of domain-specific documents.

**[021]** Further, the analysing device 102 may determine a fourth set of queries based on a comparison between the second set of queries and the third set of queries. In an embodiment, the comparison between the second set of queries and the third set of queries may be based on a comparison of character embeddings and position vector of each character in a query text data of each query of the second set of queries and the third set of queries. In an embodiment, the first set of queries may be iteratively updated based on the corresponding fourth set of queries that may be determined for each corresponding domain-specific document from the portion of the plurality of domain-specific documents.

**[022]** Further, the analysing device 102 may determine a final set of queries corresponding to the plurality of domain-specific documents based on the fourth set of queries determined as about same as the second set of queries. Further, the analysing device 102 may analyze the plurality of domain-specific documents by querying a fourth generative machine learning model using the final set of queries. Further, the analysing device 102 may receive an answer

text data output by the fourth generative machine learning model for each query of the final set of queries based on the first text data. The analysing device 102 may display the answer text data received as output from the fourth generative machine learning model on a display of the I/O device 108.

**[023]** Referring now to **FIG. 2,** a functional block diagram of the analysing device 102 is illustrated, in accordance with an embodiment of the present disclosure. In an embodiment, the analysing device 102 may include a text extraction module 202, a first set of queries determination module 204, a second set of queries determination module 207, a third set of queries determination module 208, a fourth set of queries determination module 210, a final set of queries determination module 212, a query module 214 and an output module 216.

**[024]** The analysing device 102 may receive a plurality of domain-specific documents. In an embodiment, the domain-specific documents may be documents such as, but not limited to, resume, financial documents, marksheets, academic books, employee data, purchase orders of any product-based company, invoices of various types, technical documents, etc. Referring now to **FIG. 3** an exemplary domain-specific document 300 is illustrated, in accordance with some embodiments of the present disclosure. The domain-specific document 300 illustrates a resume that may include text that may include details related to a candidate's personal profile, academic experience, professional experience, etc. The analysing device 102 may receive multiple domain-specific documents 300 as resumes for analysis. Conventional methods of analysing information from a plurality of domain-specific documents may include determination of keywords in the text data based on predefined rules. The predefined rules may list keywords that may be detected based on which domain type may be determined and data may be extracted. However, not all documents belonging to a single domain may include same keywords and accordingly the rule-based approach may fail to extract the required information accurately.

**[025]** Referring back to **FIG**. **2**, the text extraction module 202 may use one or more text extraction techniques to extract first text data from the plurality of domain specific documents 300 received by the analysing device 102. The one or more text extraction techniques that may be, but not limited to, NLP, OCR, etc.

**[026]** Accordingly, the first set of queries determination module 204 may further sub-includes an attributes determination module 206. The attribute determination module 206 may

determine one or more attributes in the first text data extracted from the plurality of domain-specific documents 300 based on an initialized first set of queries using a first generative machine learning model. The second set of queries determination module 207 may determine a second set of queries based on one or more portions of the first text data corresponding to the one or more attributes using a second generative machine learning model. The first generative machine learning model and the second generative machine learning model may be pretrained text generation models such as, but not limited to, GPT -3.5, LLAMA, etc.

**[027]** Referring now to **FIG. 4A, FIG. 4B**, **FIG. 4C** and **FIG. 4D**, exemplary first set of queries, second set of queries, third set of queries and fourth set of queries respectively are illustrated, in accordance with an embodiment of the present disclosure. In an embodiment, the first generative machine learning model may be queried based on the exemplary first set of queries 400A to determine text output based on the one or more attributes extracted from the first text data. The text data output generated by the first generative machine learning model in response to the exemplary first set of queries 400A may be displayed by the I/O device 108 and verified by a user (not shown). In an embodiment, the user may validate if the text data output by the first generative machine learning model is correct. The user may provide an input by selecting portions of the first text data in one of the plurality of domain specific documents 300, in case the text data output by the first generative machine learning model is incorrect.

**[028]** Accordingly, based on the user validation, a correct context may be defined by the user corresponding to the first set of queries 400A and the correct context may be used by the second generative machine learning model for generation of the second set of queries by the second set of queries determination module 207. Further, the correct context may be used by the first generative machine learning model of the first set of queries determination module 204 to determine text output in subsequent iterations. **FIG. 4B** depicts exemplary second set of queries 400B generated by the second generative machine learning model enabled by the second set of queries determination module 207. Accordingly, the exemplary second set of queries 400B may be determined based on validation of the one or more portions of the first text data corresponding to the one or more attributes based on the user input.

**[029]** Referring back to **FIG. 2,** the third set of queries determination module 208 may determine a third set of queries based on a domain type of the plurality of domain-specific documents 300 and a second text data extracted for a corresponding domain-specific document 300 from a portion of the plurality of domain-specific documents. In an embodiment, the

second text data may be extracted from the corresponding domain-specific document 300 using text extraction techniques. Further, the third set of queries determination module 208 may determine a third set of queries for the corresponding domain-specific document 300. In an embodiment, the domain-type may be determined based on a subject matter analysis of the plurality of domain-specific documents 300. In an embodiment, the domain type may be determined based on an analysis of the plurality of domain-specific documents 300 by the first generative machine learning model.

**[030]** Referring now to **FIG. 4C,** exemplary third set of queries are illustrated, in accordance with an embodiment of the present disclosure is illustrated, in accordance with an embodiment of the present disclosure. The third set of queries 400C may include the set of questions extracted based on the domain type of the domain-specific documents 300 and the second text data extracted for the corresponding domain-specific document 300 from the plurality of domain-specific documents.

**[031]** Referring back to **FIG. 2,** the fourth set of queries determination module 210 may determine a fourth set of queries based on a comparison between the second set of queries 400B and the third set of queries 400C for each of the corresponding domain-specific document 300 from the portion of the plurality of domain-specific documents. In an embodiment, the comparison between the second set of queries 400B and the third set of queries 400C may be based on a comparison of character embeddings and position vector of each character in a query text data of each query of the second set of queries 400B and the fourth set of queries 400C. In an embodiment, the comparison may be based on a similarity of the character embeddings and the position vector of each character in the query text data of each of the second set of queries 400B and the third set of queries 400C.

**[032]** In an embodiment, the character embeddings of query text data of each query of the second set of queries 400B and the fourth set of queries 400C may include a character embedding vector of ‘n’ bits and the position embeddings may include a position embedding vector of ‘m’ bits representing coordinate position of each character, a font type, a font size, case information such as (small case, upper case, mixed case, camel case, etc.), a typography information such as (underline, bold, italic, etc.), etc. In an embodiment, the character embeddings may be determined using an encoder-decoder deep learning (DL) model.

**[033]** Referring now to **FIG. 4D**, the exemplary fourth set of queries 400D are illustrated which includes queries that are common in both the second set of queries 400B and the third set of queries 400C. In case the fourth set of queries 400D are determined to be about same as the second set of queries 400B, the context of the first generative machine learning model and the second generative machine learning model may be determined as accurate for the analysis of the plurality of the plurality of domain-specific documents 300. However, as can be seen from the exemplary fourth set of queries 400D, some of the second set of queries 400B and the third set of queries 400C are about same, hence, the based on the first set of queries 400A may be updated based on the fourth set of queries 400D.

**[034]** In an embodiment, the first set of the queries 400A may be replaced by the fourth set of queries 400D. In a second iteration, the process of determination of attributes in the first text data based on the updated first set of the queries using the first generative machine learning model may be repeated. The second set of query Accordingly, the first set of queries 400A may be iteratively updated based on the corresponding fourth set of queries determined, based on the comparison between the second set of queries 400B and the third set of queries 400C for each corresponding domain-specific document 300 from the portion of the plurality of domain-specific documents. Accordingly, the iteration may continue until the third set of queries 400C may be determined as about equal to the second set of queries 400B.

**[035]** Referring back to **FIG. 2**, the final set of queries determination module 212 may determine a final set of queries corresponding to the plurality of domain-specific documents based on the fourth set of queries 400D determined as about same as the second set of queries 400B. The analysing device 102 may analyze the plurality of domain-specific documents by the query module 214 that may query a fourth generative machine learning model using the final set of queries. Further, the output module 216 may receive an answer text data output by the fourth generative machine learning model for each query of the final set of queries from the first text data. In an embodiment, the answer text data output by the fourth generative machine learning model may be based on the first text data extracted from the plurality of domain-specific documents 300 based on the correct context provided based on user input. Further, the answer text data for each of the query of the final set of queries may be displayed on a display interface of the I/O device 108.

**[036]** Referring to **FIG. 5,** a flowchart 500 of a methodology to determine a final set of queries for analyzing a plurality of domain-specific documents 300 is illustrated, in accordance with

an embodiment of the present disclosure. In an embodiment, method 500 may include a plurality of steps that may be performed by the processor 104 to determine the final set of queries to analyze the plurality of domain specific documents 300.

**[037]** At step 502, one or more attributes may be determined in a first text data extracted from the plurality of domain-specific documents based on a first set of queries using a first generative machine learning model. The text data output generated by the first generative machine learning model in response to the first set of queries may be based on the one or more attributes determined. A user may validate the text data output generated by the first generative machine learning model. In case, on review of the text data output, if the text data output is determined to be incorrect, the user may input one or more portions from the first text data that may be correct output for the first set of queries in order to provide a context to the first generative machine learning model. Accordingly, at step 504, a second set of queries may be determined based on one or more portions selected of the first text data by the user corresponding to the one or more attributes using a second generative machine learning model. At step 506, for each domain-specific document from a portion of the plurality of domain-specific documents, a third set of queries may be determined based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document from the portion of the plurality of domain-specific documents. In an embodiment the domain type may be determined based on a subject matter analysis of the plurality of domain-specific documents.

**[038]** Further, at step 508, a fourth set of queries may be determined based on a comparison between the second set of queries and the third set of queries determined for each domain-specific document from the portion of the plurality of domain-specific documents. In an embodiment, the fourth set of queries may be determined to include one or more queries that are common or same in the second set of queries and the third set of queries.

**[039]** At step 510, the processor 104 may determine if fourth set of queries is about equal to the second set of queries for the corresponding domain-specific document. In case, at step 508, the fourth set of queries is determined to be about equal to the second set of queries for the corresponding domain-specific document, a final set of queries is determined at step 514 as the fourth set of queries. However, in case at step 508, the fourth set of queries is not determined to be equal to the second set of queries, then the first set of queries are updated as the fourth set of queries at step 512. The steps of 502 to 510 are repeated for each of the portion of the

plurality of domain-specific documents until the fourth set of queries are determined to be equal to the second set of queries at step 510. Accordingly, the final set of queries may be determined to be same as the fourth set of queries at step 514.

**[040]** Referring to **FIG. 6,** a flowchart 600 of a methodology of analyzing a plurality of domain-specific documents is illustrated, in accordance with an embodiment of the present disclosure. In an embodiment, method 600 may include a plurality of steps that may be performed by the processor 104 to analyze a plurality of domain specific documents.

**[041]** At step 602, one or more attributes may be determined in a first text data extracted from the plurality of domain-specific documents based on a first set of queries 400A using a first generative machine learning model. At step 604, a second set of queries 400B may be determined based on one or more portions of the first text data corresponding to the one or more attributes using a second generative machine learning model.

**[042]** Further, at step 606, a third set of queries 400C may be determined, for each of a portion of the plurality of domain-specific documents, based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document 300 from the portion of the plurality of domain-specific documents. In an embodiment the domain type may be determined based on a subject matter analysis of the plurality of domain-specific documents.

**[043]** Further, at step 608, a fourth set of queries may be determined based on a comparison between the second set of queries 400B and the third set of queries 400C determined for each of a portion of the plurality of domain-specific documents. In an embodiment, the fourth set of queries 400D may be determined to include one or more queries that are common or same in the second set of queries 400B and the third set of queries 400C. In an embodiment, the first set of queries 400A may be iteratively updated based on the corresponding fourth set of queries 400D determined for each of the corresponding domain-specific document from the portion of the plurality of domain-specific documents.

**[044]** Further at step 610, a final set of queries may be determined corresponding to the plurality of domain-specific document based on the fourth set of queries 400D determined as about same as the second set of queries 400B.

**[045]** It is intended that the disclosure and examples be considered as exemplary only, with a true scope of disclosed embodiments being indicated by the following claims.

**WE CLAIM:**

1. A method (600) of analysing a plurality of domain-specific documents, the method (600) comprising:

determining (602), by a processor (104), one or more attributes in a first text data extracted from the plurality of domain-specific documents based on a first set of queries (400A) using a first generative machine learning model,

determining (604), by the processor (104), a second set of queries (400B) based on one or more portions of the first text data corresponding to the one or more attributes using a second generative machine learning model;

for each of a portion of the plurality of domain-specific documents:

determining (606), by the processor (104), a third set of queries (400C) based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document from the portion of the plurality of domain-specific documents, using a third generative machine learning model,

wherein the domain type is determined based on a subject matter analysis of the plurality of domain-specific documents; and

determining (608), by the processor (104), a fourth set of queries (400D) based on a comparison between the second set of queries (400B) and the third set of queries (400C),

wherein the first set of queries (400A) are iteratively updated based on the corresponding fourth set of queries (400D) determined for each of the corresponding domain-specific document from the portion of the plurality of domain-specific documents; and

determining (610), by the processor (104), a final set of queries corresponding to the plurality of domain-specific documents based on the fourth set of queries (400D) determined as about same as the second set of queries (400B).

2. The method (600) as claimed in claim 1, comprising:

analysing, by the processor (104), the plurality of domain-specific documents by:

querying, by the processor (104), a fourth generative machine learning model using the final set of queries; and

displaying, by the processor (104), an answer text data output by the fourth generative machine learning model for each query of the final set of queries based on the first text data.

3. The method (600) as claimed in claim 1, wherein the second set of queries (400B) are determined based on validation of the one or more portions of the first text data corresponding to the one or more attributes based on a user input.

4. The method (600) as claimed in claim 1, wherein the first text data and the second text data are extracted using a text extraction technique.

5. The method (600) as claimed in claim 1, wherein the comparison between the second set of queries (400B) and the third set of queries (400C) is based on a comparison of character embeddings and position vector of each character in a query text data of each query of the second set of queries (400B) and the third set of queries (400C).

6. A system (100) of analysing a plurality of domain-specific documents, comprising:

a processor (104); and

a memory (106) communicably coupled to the processor, wherein the memory (106) stores processor-executable instructions, which when executed by the processor (104), cause the processor (104) to:

determine a first set of queries (400A) based on one or more attributes in a first text data extracted from the plurality of domain-specific documents based on a first set of queries (400A) using a first generative machine learning model,

determine a second set of queries (400B) based on one or more portions of the first text data corresponding to the one or more attributes using a second generative machine learning model;

for each of a portion of the plurality of domain-specific documents:

determine a third set of queries (400C) based on a domain type of the plurality of domain-specific documents and a second text data extracted for a corresponding domain-specific document from the portion of the plurality of domain-specific documents,

wherein the domain type is determined based on a subject matter analysis of the plurality of domain-specific documents; and

determine a fourth set of queries (400D) based on a comparison between the second set of queries (400B) and the third set of queries (400C),

wherein the first set of queries (400A) are iteratively updated based on the corresponding fourth set of queries (400D) determined for each of the corresponding domain-specific document from the portion of the plurality of domain-specific documents; and

determine a final set of queries corresponding to the plurality of domain-specific documents based on the fourth set of queries (400D) determined as about same as the second set of queries (400B).

7. The system (100) as claimed in claim 6, wherein the processor (104) is configured to:

analyze the plurality of domain-specific documents based on:

query a fourth generative machine learning model using the final set of queries; and

display an answer text data output by the fourth generative machine learning model for each query of the final set of queries based on the first text data.

8. The system (100) as claimed in claim 6, wherein the second set of queries (400B) are determined based on validation of the one or more portions of the first text data corresponding to the one or more attributes based on a user input.

9. The system (100) as claimed in claim 6, wherein the first text data and the second text data are extracted using a text extraction technique.

10. The system (100) as claimed in claim 6, wherein the comparison between the second set of queries (400B) and the third set of queries (400C) is based on a comparison of character embeddings and position vector of each character in a query text data of each query of the second set of queries (400B) and the third set of queries (400C).

Dated this 28th day of November 2023

***--Digitally Signed--***

Bhanu Prasad (INPA No: 3253)
Head, IPR Dept.,
L&T Technology Services Limited,
DLF 3rd Block, 2nd Floor,
Manapakkam, Chennai, TN, 600089.

## ABSTRACT

**METHOD AND SYSTEM FOR ANALYSING DOMAIN SPECIFIC DOCUMENTS**

A method (600) and system (100) of analysing domain-specific documents is disclosed. A processor (104) determines one or more attributes in a first text data extracted from the plurality of domain-specific documents based on a first set of queries (400A) using a first generative machine learning (ML) model. A second set of queries (400B) are determined based portions of the first text data corresponding to the attributes using a second generative ML model. A third set of queries (400C) are determined based on domain type of the domain-specific documents and a second text data extracted for a co0rresponding domain-specific document from the domain-specific documents. A fourth set of queries (400D) are determined based on an iterative comparison between the second set of queries (400B) and the third set of queries (400C) until the fourth set of queries (400D) are determined as about same as second set of queries (400B).

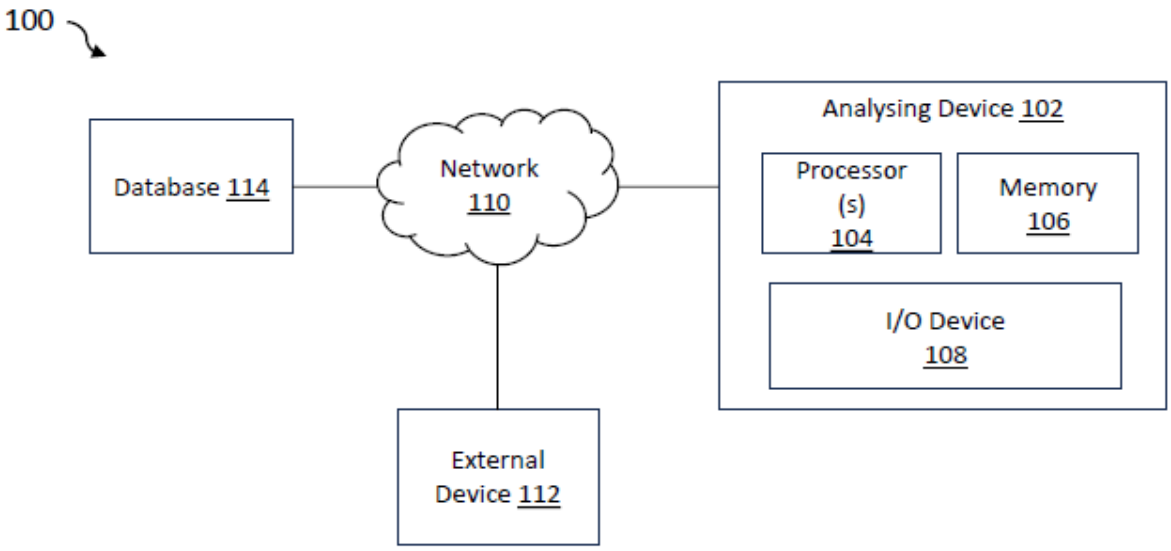




**FIG. 1**

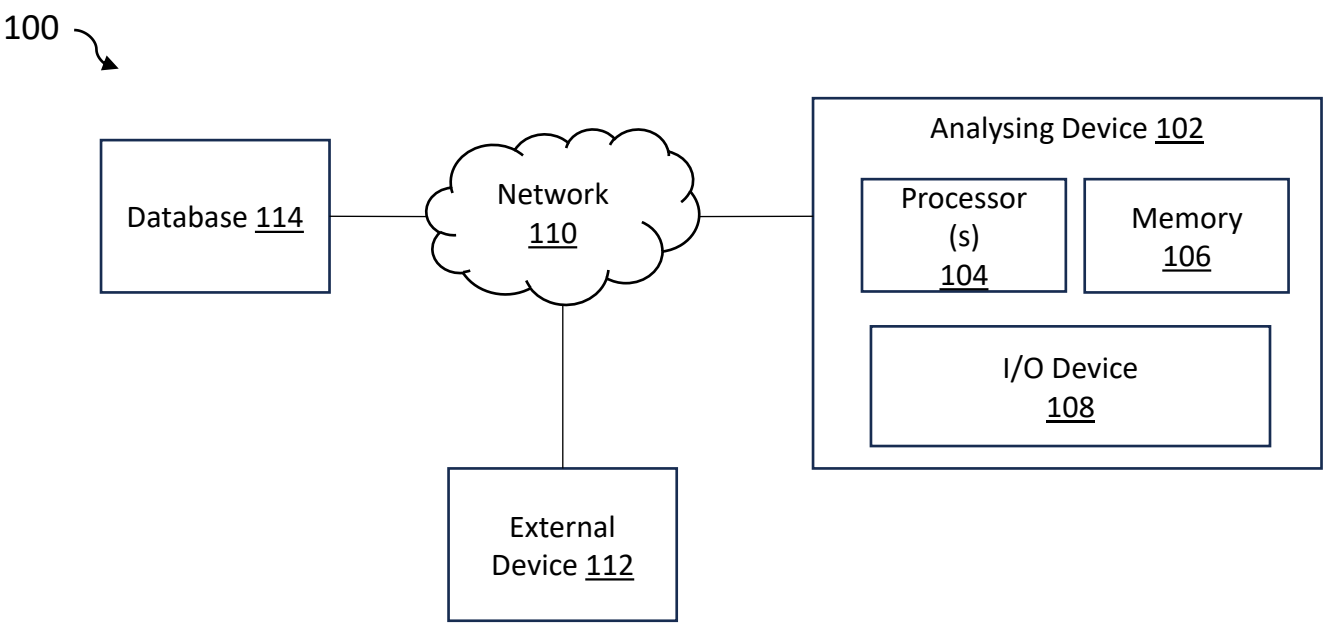



**FIG. 1**

**--Digitally Signed--**
**Bhanu Prasad (INPA No: 3253)**
**Head, IPR Dept.,**
**L&T Technology Services Limited,**
**DLF 3rd Block, 2nd Floor,**
**Manapakkam, TN, Chennai - 600089.**

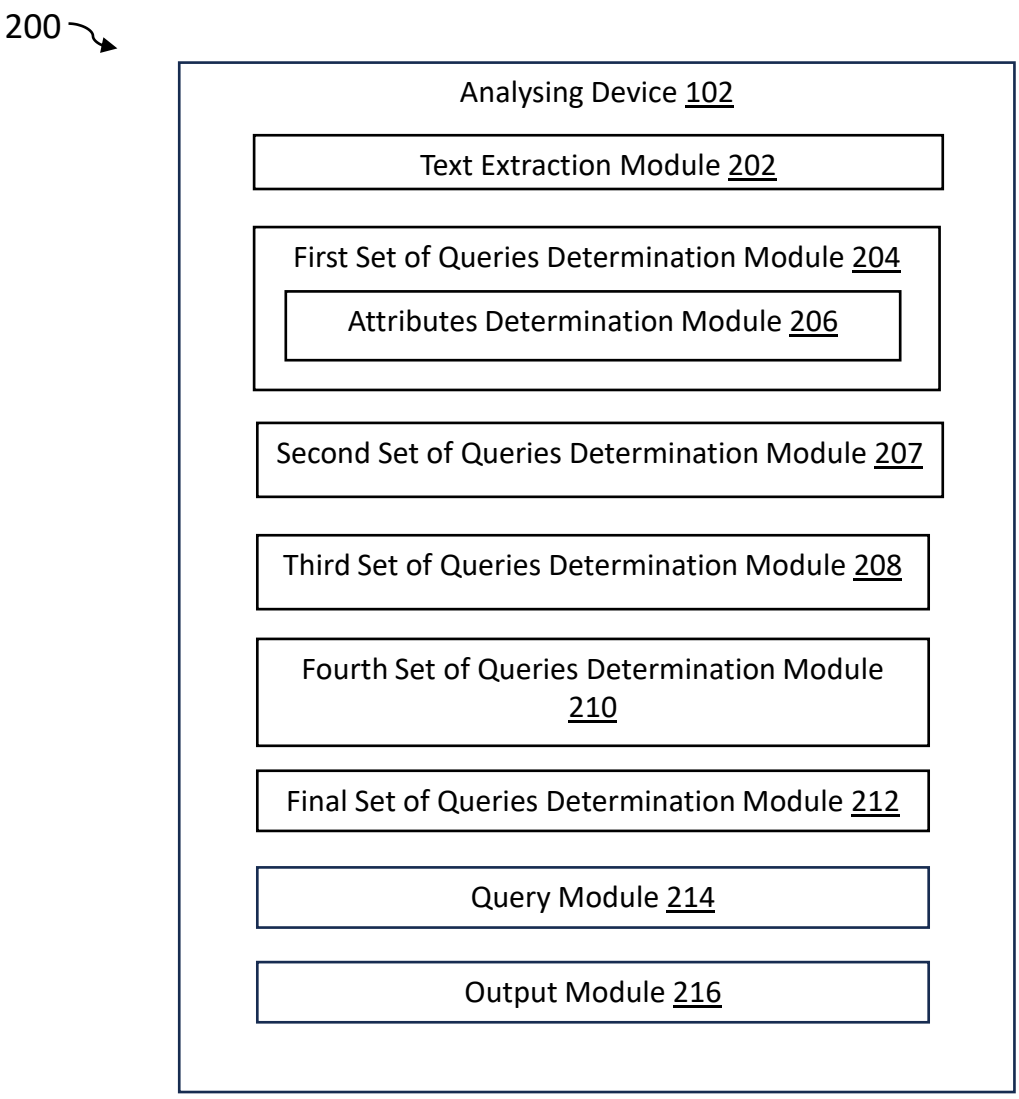
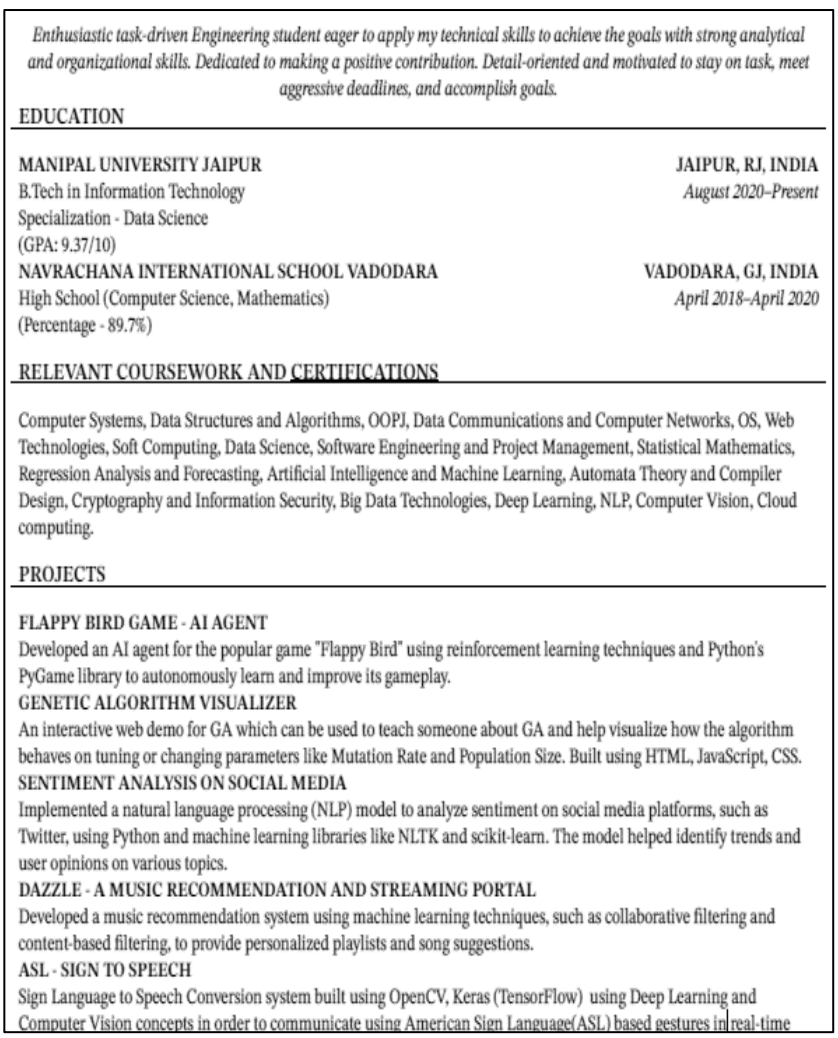



**FIG. 2**

**--Digitally Signed--**
**Bhanu Prasad (INPA No: 3253)**
**Head, IPR Dept.,**
**L&T Technology Services Limited,**
**DLF 3rd Block, 2nd Floor,**
**Manapakkam, TN, Chennai - 600089.**

300



*Enthusiastic task-driven Engineering student eager to apply my technical skills to achieve the goals with strong analytical and organizational skills. Dedicated to making a positive contribution. Detail-oriented and motivated to stay on task, meet aggressive deadlines, and accomplish goals.*

EDUCATION

**MANIPAL UNIVERSITY JAIPUR** — JAIPUR, RJ, INDIA
B.Tech in Information Technology — *August 2020–Present*
Specialization - Data Science
(GPA: 9.37/10)

**NAVRACHANA INTERNATIONAL SCHOOL VADODARA** — VADODARA, GJ, INDIA
High School (Computer Science, Mathematics) — *April 2018–April 2020*
(Percentage - 89.7%)

RELEVANT COURSEWORK AND CERTIFICATIONS

Computer Systems, Data Structures and Algorithms, OOPJ, Data Communications and Computer Networks, OS, Web Technologies, Soft Computing, Data Science, Software Engineering and Project Management, Statistical Mathematics, Regression Analysis and Forecasting, Artificial Intelligence and Machine Learning, Automata Theory and Compiler Design, Cryptography and Information Security, Big Data Technologies, Deep Learning, NLP, Computer Vision, Cloud computing.

PROJECTS

**FLAPPY BIRD GAME - AI AGENT**
Developed an AI agent for the popular game "Flappy Bird" using reinforcement learning techniques and Python's PyGame library to autonomously learn and improve its gameplay.

**GENETIC ALGORITHM VISUALIZER**
An interactive web demo for GA which can be used to teach someone about GA and help visualize how the algorithm behaves on tuning or changing parameters like Mutation Rate and Population Size. Built using HTML, JavaScript, CSS.

**SENTIMENT ANALYSIS ON SOCIAL MEDIA**
Implemented a natural language processing (NLP) model to analyze sentiment on social media platforms, such as Twitter, using Python and machine learning libraries like NLTK and scikit-learn. The model helped identify trends and user opinions on various topics.

**DAZZLE - A MUSIC RECOMMENDATION AND STREAMING PORTAL**
Developed a music recommendation system using machine learning techniques, such as collaborative filtering and content-based filtering, to provide personalized playlists and song suggestions.

**ASL - SIGN TO SPEECH**
Sign Language to Speech Conversion system built using OpenCV, Keras (TensorFlow) using Deep Learning and Computer Vision concepts in order to communicate using American Sign Language(ASL) based gestures in real-time

**FIG. 3**

**--Digitally Signed--**
**Bhanu Prasad (INPA No: 3253)**
**Head, IPR Dept.,**
**L&T Technology Services Limited,**
**DLF 3rd Block, 2nd Floor,**
**Manapakkam, TN, Chennai - 600089.**

400A

First Set of Queries

Initial pre-defined Query List,

1. What is the total years of experience?
2. What is his/her core expertise ?
3. Has he/she done any certification ?

**FIG. 4A**

400B

Second Set of Queries

1. When was the graduation done ?
2. From where was the graduation done ?
3. Does he know AI related skills ?
4. Any certification he has done ?
5. ……

**FIG. 4B**

400C

Third Set of Queries

1. Where is he located ?
2. From where was the graduation done ?
3. How many years of work experience ?
4. Any certification he has done ?
5. ……

**FIG. 4C**

400D

Fourth Set of Queries

1. From where was the graduation done ?
2. Any certification he has done ?

**FIG. 4D**

**--Digitally Signed--**
**Bhanu Prasad (INPA No: 3253)**
**Head, IPR Dept.,**
**L&T Technology Services Limited,**
**DLF 3rd Block, 2nd Floor,**
**Manapakkam, TN, Chennai - 600089.**

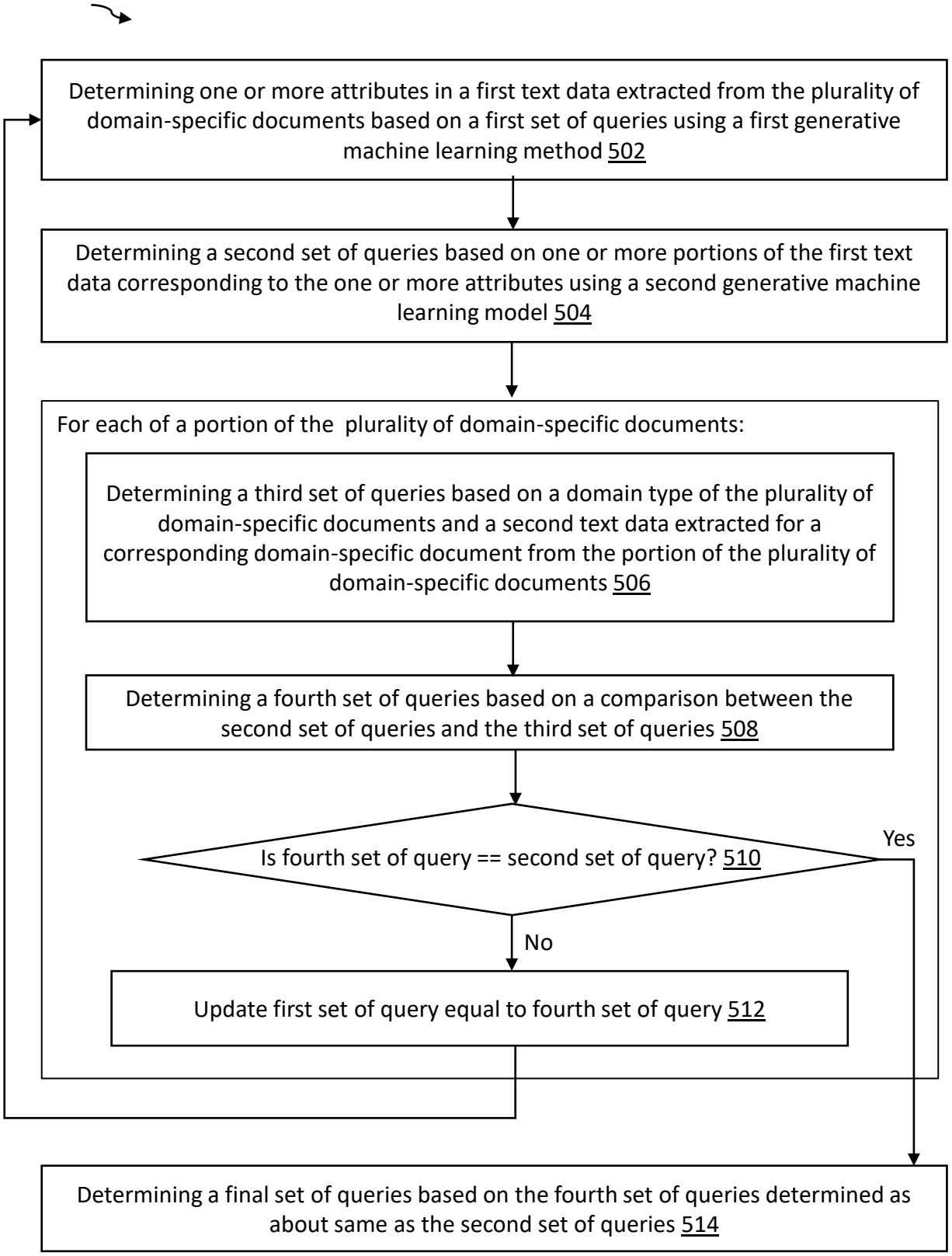



**FIG. 5**

**--Digitally Signed--**
**Bhanu Prasad (INPA No: 3253)**
**Head, IPR Dept.,**
**L&T Technology Services Limited,**
**DLF 3rd Block, 2nd Floor,**
**Manapakkam, TN, Chennai - 600089.**

600

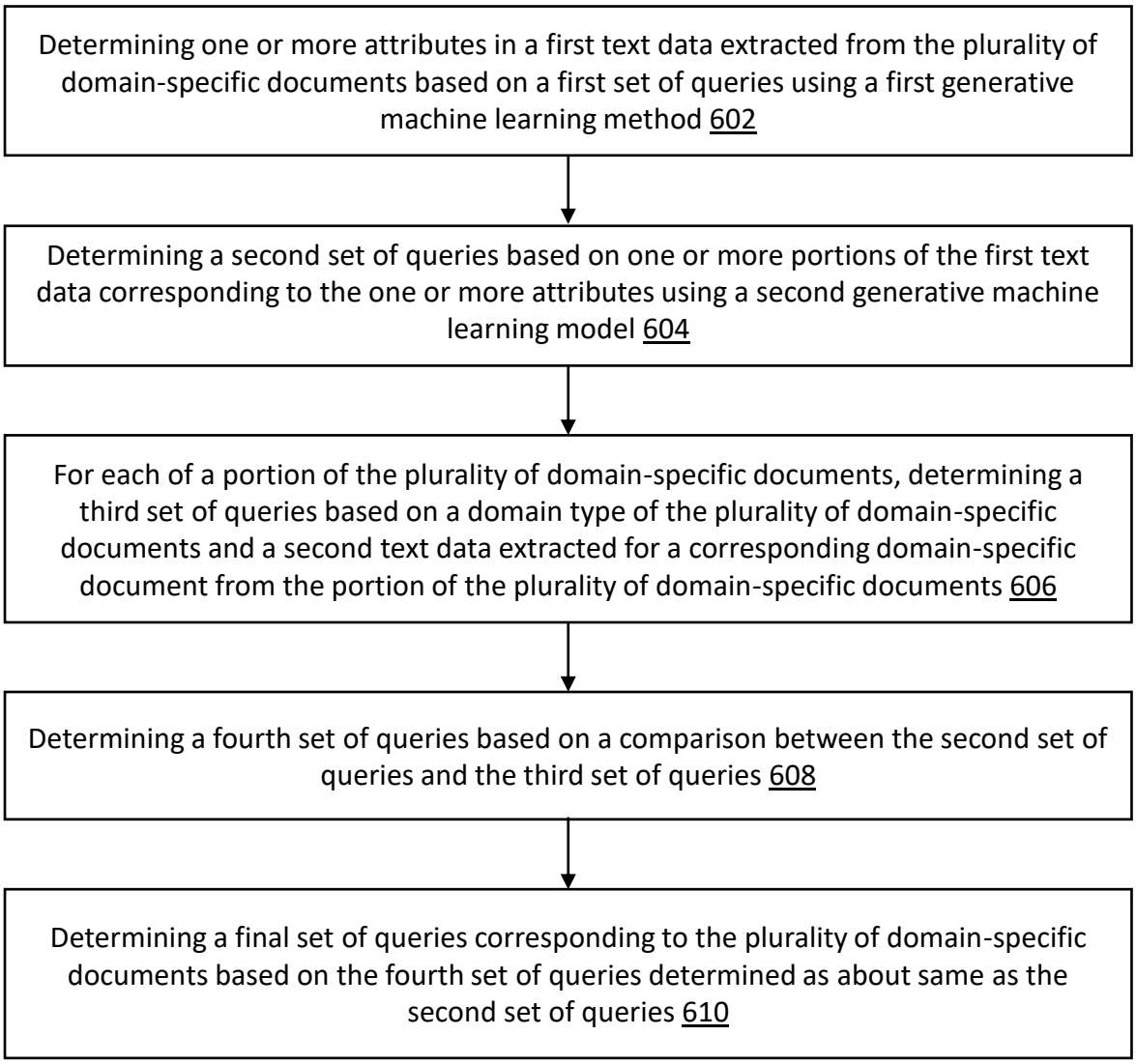



**FIG. 6**

**--Digitally Signed--**
**Bhanu Prasad (INPA No: 3253)**
**Head, IPR Dept.,**
**L&T Technology Services Limited,**
**DLF 3rd Block, 2nd Floor,**
**Manapakkam, TN, Chennai - 600089.**