

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
29 January 2004 (29.01.2004)

PCT

(10) International Publication Number
WO 2004/010337 A2

(51) International Patent Classification⁷: **G06F 17/30**

(21) International Application Number:
PCT/US2003/021621

(22) International Filing Date: 11 July 2003 (11.07.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/201,061 22 July 2002 (22.07.2002) US

(71) Applicant (*for all designated States except US*): **CON-
TIVO, INC.** [US/US]; 640A Clyde Court, Mountain View,
CA 94043 (US).

(72) Inventor; and

(75) Inventor/Applicant (*for US only*): **DUMAS, François,
P.** [FR/US]; 130 Rinconada Avenue, Palo Alto, CA 94301
(US).

(74) Agents: **MALLIE, Michael. J.** et al.; Blakely, Sokoloff,
Taylor & Zafman LLP, 12400 Wilshire Boulevard, 7th
Floor, Los Angeles, CA 90025 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU,
AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU,
CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH,
GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC,
LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW,
MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG,
SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN,
YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM,
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),
Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,
ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,
SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM,
GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished
upon receipt of that report*

*For two-letter codes and other abbreviations, refer to the "Guid-
ance Notes on Codes and Abbreviations" appearing at the begin-
ning of each regular issue of the PCT Gazette.*

(54) Title: A METHOD AND SYSTEM FOR MODELING COMPONENTS IN DOCUMENTS

(57) Abstract: The invention provides a system and method for transforming data between a source document in a source standard and a target document in a target standard. According to one embodiment, the method comprises deriving a semantic model for semantically equivalent concepts that can occur in documents in the source and target standards, wherein a semantic meaning comprising a generic meaning and a specific meaning is assigned the each semantically equivalent concept, and modeling the source and target documents based on the semantic model.



WO 2004/010337 A2

A METHOD AND SYSTEM FOR MODELING COMPONENTS IN DOCUMENTS

FIELD OF THE INVENTION

[0001] This invention relates to the processing of electronic documents. In particular, it relates to the integration of electronic documents from diverse sources.

BACKGROUND

[0002] Modern business enterprises today face the challenge of integrating business documents from diverse document sources or systems. For example, an enterprise may wish to access data in a legacy system, or a system of a previous competitor that has been acquired, for example, through merger. Integrating the data from these various systems can be beneficial in achieving operational competitiveness by integrating internal systems within the enterprise to obtain a single view of the enterprise's data. Alternatively, such integration can be beneficial in achieving collaborative competitiveness by integrating the enterprise's systems with those of strategic trading partners.

[0003] One aspect of integration involves the use of maps or transformations which map or transform data from one source or system representation to another. Such maps can be automatically generated if the source and target documents are modeled in terms of a common vocabulary of semantic concepts. This involves deriving a semantic model for the source and target documents wherein equivalent components comprising a set of semantically related fields in the source and target documents are related using the common vocabulary. In the known prior art, to derive such a semantic model, each component in the source and target document must be manually related to an equivalent concept in the semantic model. This can be a tedious process, especially in cases where a document has a large number of components.

SUMMARY OF THE INVENTION

[0004] According to one aspect of the invention there is provided a method for transforming a source document in a source format into a target document in

a target format. The method comprises, for each of the source and target formats, determining all species of a concept that can occur in a document in that format, a concept being a set of semantically related fields and deriving a semantic model for transforming documents between the source and target formats. The semantic model contains generic information common to all species to allow automatic recognition of the concept, and specific information unique to each species to allow differentiation of the concept into a particular species thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] Figure 1 shows a source document which has to be transformed into a target document and a semantic model which will be used in generating appropriate transformations;

[0006] Figure 2 shows a flowchart of operations that are performed in transforming the source document of Figure 1 into the target document of Figure 1;

[0007] Figure 3 shows a table of all possible fields for an "address" concept;

[0008] Figure 4 shows how the components within a semantic model are structured in accordance with one embodiment of the invention; and

[0009] Figure 5 shows a block diagram of a system in accordance with one embodiment of the present invention.

DETAILED DESCRIPTION

[0010] In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one skilled in the art that the invention can be practiced without these specific details. In other instances, structures and devices are shown in block diagram form in order to avoid obscuring the invention.

[0011] Reference in the specification to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one

embodiment of the invention. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment, nor are separate or alternative embodiments mutually exclusive of other embodiments. Moreover, various features are described which may be exhibited by some embodiments and not by others. Similarly, various requirements are described which may be requirements for some embodiments but not other embodiments.

[0012] Embodiments of the present invention relate to a method for transforming data from a source document in a source format to a target document in a target format. More particularly, embodiments of the present invention allow components comprising a set of semantically related fields within a document to be modeled in a fashion that reduces the amount of manual mapping required when transforming data from a source document in a source format to a target document in a target format.

[0013] According to embodiments of the present invention a semantic concept that can occur within documents of a source or target standard (format) is modeled by assigning to it a generic meaning as well as a specific meaning. Thereafter, whenever the semantic concept appears within an actual document in the source or target standard, it is automatically assigned its generic meaning and a user is prompted to select its specific meaning from a list of specific meanings. Thus, each concept is modeled once per document standard and thereafter each occurrence of the concept within an actual document of the standard is assigned a generic and specific meaning based on the model. According to embodiments of the present invention, the specific meaning is assigned based on a location of the component within a document. Thus, embodiments of the present invention serve to reduce the degree of manual mapping required when modeling source and target documents in terms of a semantic model.

[0014] Figure 1 of the drawings shows a source document 100 for example, an invoice, a sales order, etc. which has to be transformed into a target document 102, such as for purposes of document integration. The source document 100 includes four components designated 100A-100D, respectively,

and the target document 102 includes four components designated 102A-102D, respectively. The semantic meaning of each of the components in source document 100 and target document 102 is shown in semantic model 104 wherein each of the components of the source and target document are related to a semantic component designated 104A-104D, respectively. Thus, for example, component 100A and component 102C are related and will contain information relating to a semantic concept which provides a corporate head office address. It will be seen that semantic model 104 specifies the relationship between components in source document 100 and target document 102 in terms of a common semantic concept, which allows for a mapping program to use this information to automatically generate mappings for transforming data from source document 100 into target document 102.

[0015] In generating the semantic model 104 each of the components in the source and target documents will have to be manually modeled. However, it will be appreciated that because each of the components in the source and target documents is related to a generic semantic concept which provides address information, a certain portion of information contained in each of the components will overlap. For example, each of the components may have address line, city and zip code fields which overlap. One of the advantages of the present invention is that it reduces the degree of manual mapping required to produce a semantic model such as semantic model 104, by taking advantage of the information which overlaps or is common to various components within the source and target documents, as will be described below.

[0016] Figure 2 of the drawings shows a flow chart of operations performed in accordance with one embodiment of the present invention, when transforming source document 100 into target document 102. Referring to Figure 2, at block 200 a semantic model is derived for documents of the source and target data standards or formats. Block 200 involves examining allowable concepts in each of the source and target document formats and identifying a generic form for each of the concepts as well as specific forms or species of the generic form. For example, Figure 3 shows a table 300 which includes all possible fields for the "address" concept, which fields can occur or are allowable in a valid address

in the source or target document of a source or target format respectively. By examining all allowable types of addresses, it may be determined that fields containing address, city, state, and zip code information are generic in a sense that they are common to all valid types of addresses, whereas the remaining fields occur additionally in only some types of addresses. In deriving the semantic model, at block 200 each semantic concept is structured into a generic concept and specific or species concepts. Each generic concept comprises fields which are common to all particular species of the concept, whereas each specific or species concept includes particular fields in addition to the common fields which are shared with the generic concept. This is shown in Figure 4 of the drawings where a component 400 comprising fields 1 to n is structured so as to have a generic meaning 402 and a specific meaning 404. By way of example, the generic meaning 402 relates to an address and the specific meaning could be a particular kind or type of address such as a "shipping address". At block 202 the semantic model is stored within a system, such as is shown in Figure 5 of the drawings, which is used to perform the transformation. At block 204 actual source and target documents that are of the source target standards, respectively, are loaded into the system. At block 206 using the previously stored semantic model, components in the source and target documents are automatically recognized based on the generic meanings of concepts from the semantic model. At block 208, a prompt is displayed to a user, prompting the user to select a specific meaning for a particular concept. This is done based on a location of a component within a document. Thus, using document 100 as an example, components 100A – 100D will automatically be recognized as an address concept based on its generic meaning and thereafter a user will be prompted to select a particular meaning for each address thus specifying whether the recognized address concept is a corporate head office address, a shipping address, a billing address, or a branch address based on a location of the component within the document. Once the user has made the selection, the selected specific meaning is assigned to the component. Thereafter, at block 212 a mapping based on the assigned meaning of concepts in source document 100 and 102 is created.

[0017] Referring now to Figure 5 of the drawings, reference numeral 500 generally indicates an example of a system which may be used to implement to perform embodiments of the invention described above. The system 500 includes a memory 504, which may represent one or more physical memory devices, which may include any type of random access memory (RAM), read only memory (ROM) which may be programmable, flash memory, non-volatile mass storage device, or a combination of such memory devices. The memory 504 is connected via a system bus 512 to a processor 502. The memory 504 includes instructions 506 which when executed by the processor 502 cause the processor to perform the methodology of the invention as discussed above. Additionally, the system 500 includes a disk drive 508 and a CD ROM drive 510 each of which is coupled to a peripheral-device and user-interface 514 via bus 512. Processor 502, memory 504, disk drive 508 and CD ROM 510 are generally known in the art. Peripheral-device and user-interface 514 provides an interface between system bus 512 and various optional components connected to a peripheral bus 516 as well as to user interface components, such as a display, mouse and other user interface devices. A network interface 518 is coupled to peripheral bus 516 and provides network connectivity to system 500.

[0018] Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that the various modification and changes can be made to these embodiments without departing from the broader spirit of the invention as set forth in the claims. Accordingly, the specification and drawings are to be regarded in an illustrative sense rather than in a restrictive sense.

CLAIMS

What is claimed is:

1. A method for transforming data between a source document in a source standard and a target document in a target standard, the method comprising:
deriving a semantic model for semantically equivalent concepts that can occur in documents in the source and target standards, wherein a semantic meaning comprising a generic meaning and a specific meaning is assigned to each semantically equivalent concept; and
modeling the source and target documents based on the semantic model.
2. The method of claim 1, wherein the generic meaning of a semantic concept is independent of a context of the semantic concept within a document.
3. The method of claim 1, wherein the specific meaning of a semantic concept is dependent on a context of the semantic concept within a document.
4. The method of claim 1, wherein modeling the source and target documents comprises recognizing a component therein; and assigning a generic meaning to the recognized component based on the semantic model.
5. The method of claim 4, further comprising assigning a specific meaning to the recognized component based on its context within the source document or target document.
6. The method of claim 5, wherein assigning the specific meaning comprises displaying a list of possible specific meanings for the component to a user; and receiving the user's selection of a specific meaning in response.
7. A method for transforming a source document in a source format into a target document in a target format, the method comprising:

for each of the source and target formats, determining all species of a concept that can occur in a document in that format, a concept being a set of semantically related fields; and

deriving a semantic model for transforming documents between the source and target formats wherein the semantic model contains generic information common to all species to allow automatic recognition of the concept, and specific information unique to each species to allow differentiation of the concept into a particular species thereof.

8. The method of claim 7, further comprising building a map to transform between the source document and the target document based on the semantic model, wherein the generic information is used to automatically recognize equivalent concepts in the two documents, and the specific information is used together with user input information to differentiate species of the equivalent concept.

9. The method of claim 8, further comprising prompting the user to input information to facilitate differentiation between the species.

10. The method of claim 9, wherein prompting a user comprises presenting a list of all species of the concept to the user; and receiving the user selection of a particular species from the list.

11. A system for transforming data between a source document in a source standard and a target document in a target standard, the system comprising a processor and a memory coupled thereto, the memory storing instructions which when executed by the processor cause the processor to perform a method comprising:

deriving a semantic model for semantically equivalent concepts that can occur in documents in the source and target standards, wherein a semantic meaning comprising a generic meaning and a specific meaning is assigned to each semantically equivalent concept; and

modeling the source and target documents based on the semantic model.

12. The system of claim 11, wherein the generic meaning of a semantic concept is independent of a context of the semantic concept within a document.

13. The system of claim 11, wherein the specific meaning of a semantic concept is dependent on a context of the semantic concept within a document.

14. The system of claim 11, wherein modeling the source and target documents comprises recognizing a component therein; and assigning a generic meaning to the recognized component based on the semantic model.

15. The system of claim 14, wherein the method further comprises assigning a specific meaning to the recognized component based on its context within the source document or target document.

16. The system of claim 15, wherein assigning the specific meaning comprises displaying a list of possible specific meanings for component to a user; and receiving the user's selection of a specific meaning in response.

17. A system for transforming a source document in a source format into a target document in a target format, the system comprising a memory and a processor coupled thereto, the memory storing instructions which when executed by the processor cause the processor to perform a method comprising:

for each of the source and target formats, determining all species of a concept that can occur in a document in that format, a concept being a set of semantically related fields; and

deriving a semantic model for transforming documents between the source and target formats wherein the semantic model contains generic information common to all species to allow automatic recognition of the concept, and specific information unique to each species to allow differentiation of the concept into a particular species thereof.

18. The system of claim 14, wherein the method further comprises building a map to transform between the source document and the target document based on the semantic model, wherein the generic information is used to automatically recognize equivalent concepts in the two documents, and the specific information is used together with user input information to differentiate species of the equivalent concept.

19. The system of claim 18, wherein the method further comprises prompting the user to input information to facilitate differentiation between the species.

20. The system of claim 19, wherein prompting a user comprises presenting a list of all species of the concept to the user; and receiving the user selection of a particular species from the list.

21. A computer-readable medium having stored thereon a sequence of instructions which when executed by a computer cause the computer to perform a method for transforming data between a source document in a source standard and a target document in a target standard, the method comprising:

deriving a semantic model for semantically equivalent concepts that can occur in documents in the source and target standards, wherein a semantic meaning comprising a generic meaning and a specific meaning is assigned to each semantically equivalent concept; and

modeling the source and target documents based on the semantic model.

22. The computer-readable medium of claim 21, wherein the generic meaning of a semantic concept is independent of a context of the semantic concept within a document.

23. The computer-readable medium of claim 21, wherein the specific meaning of a semantic concept is dependent on a context of the semantic concept within a document.

24. The computer-readable medium of claim 21, wherein modeling the source and target documents comprises recognizing a component therein; and assigning a generic meaning to the recognized component based on the semantic model.

25. The computer-readable medium of claim 24, wherein the method further comprises assigning a specific meaning to the recognized component based on its context within the source document or target document.

26. The computer-readable medium of claim 25, wherein assigning the specific meaning comprises displaying a list of possible specific meanings for component to a user; and receiving the user's selection of a specific meaning in response.

27. A computer-readable medium having stored thereon a sequence of instructions which when executed by a computer cause the computer to perform a method for transforming a source document in a source format into a target document in a target format, the method comprising:

for each of the source and target formats, determining all species of a concept that can occur in a document in that format, a concept being a set of semantically related fields; and

deriving a semantic model for transforming documents between the source and target formats wherein the semantic model contains generic information common to all species to allow automatic recognition of the concept, and specific information unique to each species to allow differentiation of the concept into a particular species thereof.

28. The computer-readable medium of claim 27, wherein the method further comprises building a map to transform between the source document and the target document based on the semantic model, wherein the generic information is used to automatically recognize equivalent concepts in the two documents,

and the specific information is used together with user input information to differentiate species of the equivalent concept.

29. The computer-readable medium of claim 28, wherein the method further comprises prompting the user to input information to facilitate differentiation between the species.

30. The computer-readable medium of claim 29, wherein prompting a user comprises presenting a list of all species of the concept to the user; and receiving the user selection of a particular species from the list.

31. A system for transforming data between a source document in a source standard and a target document in a target standard, the system comprising:
means for deriving a semantic model for semantically equivalent concepts that occur in documents in the source and target standards, wherein a semantic meaning comprising a generic meaning and a specific meaning is assigned to each semantically equivalent concept; and
means for modeling the source and target documents based on the semantic model.

32. A system for transforming a source document in a source format into a target document in a target format, the system comprising:
means for determining, each of the source and target formats, all species of a concept that can occur in a document in that format, a concept being a set of semantically related fields; and
means for deriving a semantic model for transforming documents between the source and target formats wherein the semantic model contains generic information common to all species to allow automatic recognition of the concept, and specific information unique to each species to allow differentiation of the concept into a particular species thereof.

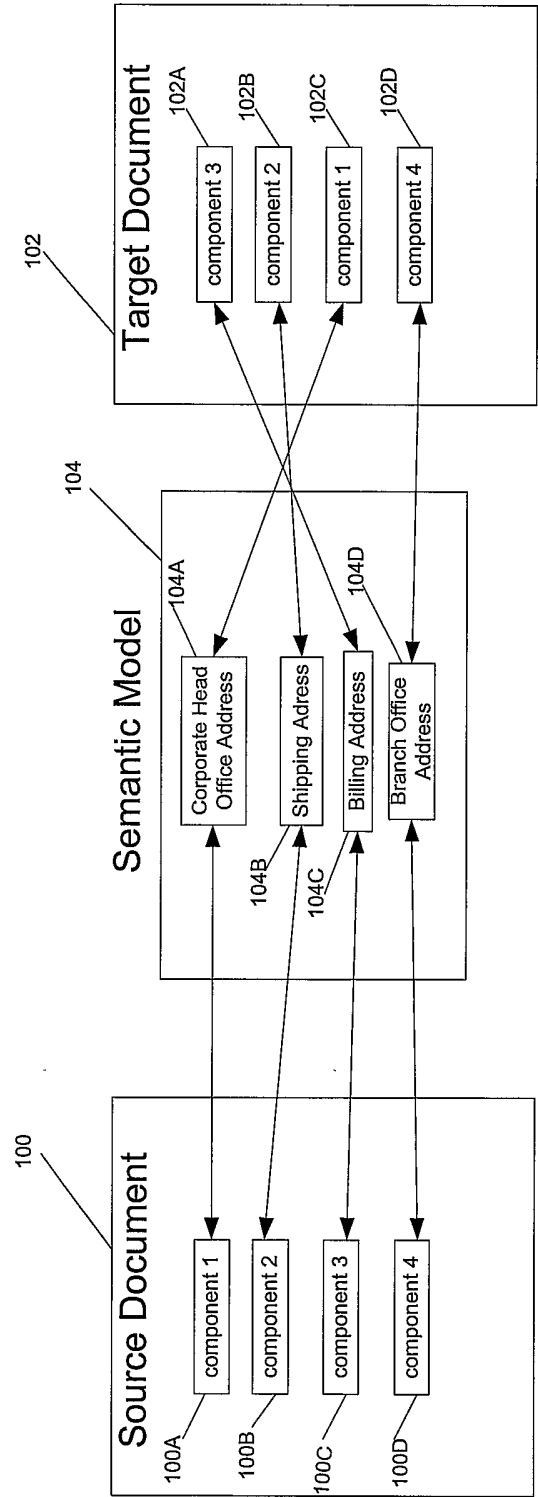


FIGURE 1

2/5

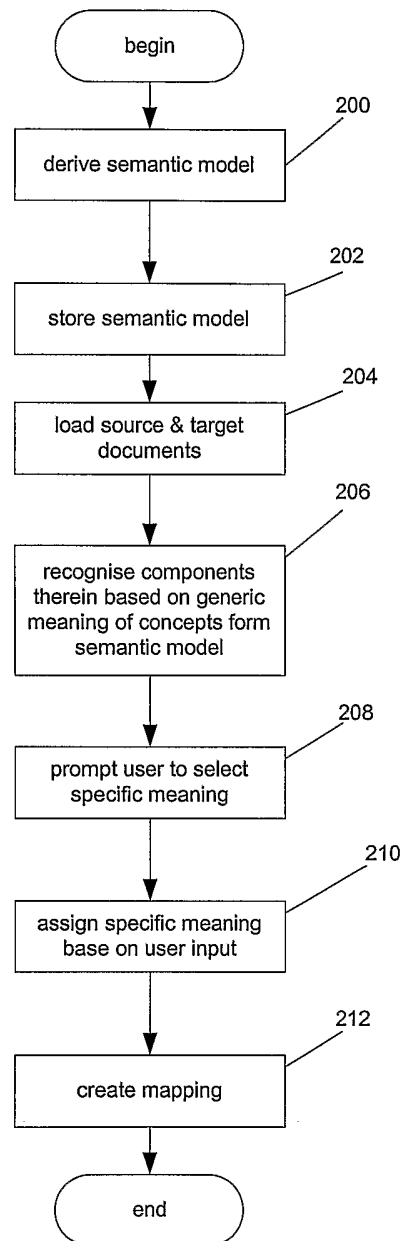


FIGURE 2

3/5



AddressLines 1..n
Apt No
City
State
Zip
Country
County
Region
PostOfficeBox
HouseNumber
Building
Floor
RoomNumber
Department

FIGURE 3

4/5

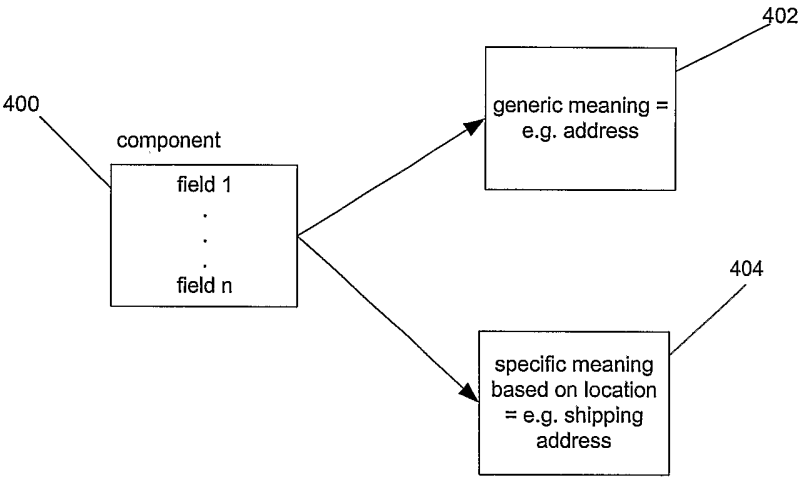


FIGURE 4

5/5

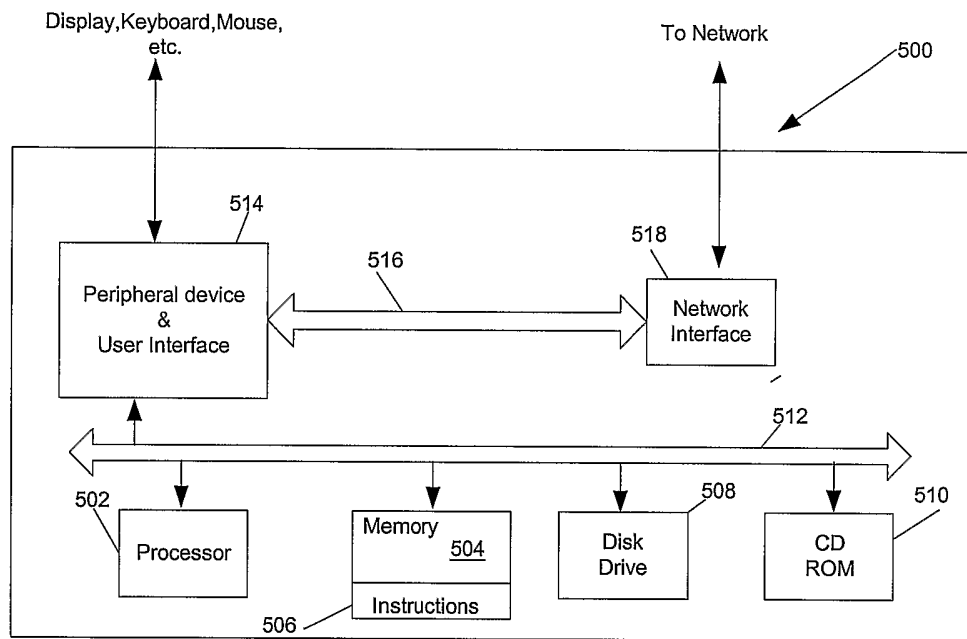


FIGURE 5