(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 July 2010 (29.07.2010)

PCT

(10) International Publication Number
WO 2010/084344 A1

(51) International Patent Classification:
G06F 11/30 (2006.01) *G06F 21/00* (2006.01)

(21) International Application Number:
PCT/GB2010/050074

(22) International Filing Date:
19 January 2010 (19.01.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/145,881 20 January 2009 (20.01.2009) US

(71) Applicant *(for all designated States except US)*: **SECER-NO LTD** [GB/GB]; Seacourt Tower, West Way, Oxford, Oxfordshire OX2 0JJ (GB).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **MOYLE, Stephen, Anthony** [GB/GB]; Highview, Vernon Avenue, Oxford, Oxfordshire OX2 9AU (GB).

(74) Agent: **FREEMAN, Avi**; BECK GREENER, Fulwood House, London WC1V 6HR (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

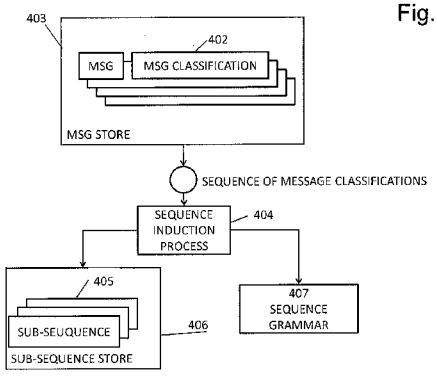
— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

WO 2010/084344 A1

(54) Title: METHOD, COMPUTER PROGRAM AND APPARATUS FOR ANALYSING SYMBOLS IN A COMPUTER SYSTEM

Fig. 5



(57) Abstract: The present invention provides a computer-implemented method of analysing messages in a computer system to allow workflows constituted by the messages to be identified, the method comprising: analysing a sequence of messages in a computer system in order to classify the messages, thereby producing a corresponding sequence of classifications of the messages; and, applying sequence induction to the sequence of classifications of the messages to produce (i) a set of sub-sequences of the classifications of the messages and (ii) a sequence grammar for the sub-sequences, from which a workflow constituted by the sequence of messages can be identified.

METHOD, COMPUTER PROGRAM AND APPARATUS FOR ANALYSING SYMBOLS IN A COMPUTER SYSTEM

The present invention relates to a method, a computer program and apparatus for analysing symbols in a computer system.

There are many examples of computer systems in which it is useful to be able to analyse symbols passing through or stored in the computer system. As will be appreciated from the following, the term "symbols" in this context is to be construed broadly. In general, the term "symbols" is used herein in the broad sense as used in the field of Universal Turing Machines. For example, "symbols" includes computer messages, which term is also to be construed broadly and includes for example computer messages in a computer language (including computer instructions, such as executable programs), natural languages in computer-readable form (such as in documents, emails, etc.). "Symbols" also includes computer data in the conventional sense, i.e., typically, abstractions of real world artefacts, etc.

Individual computer messages entering into or passing through a computer system are often one element of a more complicated transaction that it is intended to be processed by the computer system. Such a transaction is also known as a "workflow". The individual computer messages can be generated by for example an application (i.e. some computer software), typically at the instigation of a user, as part of an overall process or "workflow". Such workflows are often business related.

The discovery, monitoring and control of workflow is beneficial in a number of domains, including but not limited to: security, operational effectiveness, metering and quality assessment. In the security domain, determining what are normal and expected workflows with sufficient accuracy allows the automated detection of anomalies. Anomalous workflows can be controlled in a number of ways. For example, they can be logged for later forensic analysis; they can trigger an alarm or alert so that immediate attention is given to them; they can be controlled by preventing them from being committed or terminated early; they can be redirected to alternative resources (e.g. as in

load balancing); and they can be dynamically restructured to provide a more acceptable outcome. In the operational effectiveness domain, accurate determination of appropriate workflow can ensure that only the correct resources are used (e.g. workflows that do not interact with the correct information can be prevented from or redirected the appropriate resources); workflows that consume an undesirable proportion of resources (e.g. time, memory and/or power) can be identified and remediated; and workflows that include redundant operations can be recast into more effective workflows. In the metering domain, the accurate measurement of workflows provides non-repudiable evidence for accountancy (e.g. for charging in a pay-as-you-use manner). Last, workflows that have been effectively identified can be subjected to some quality criteria so that the developers of such workflows can understand and improve the workflows to a higher level of quality.

The present invention is principally concerned with the extraction of plausible sub-sequences of messages from sequences of messages. The sub-sequences can then be used to identify workflows, which can then be monitored, controlled, etc., as desired and as described in general terms above.

According to a first aspect of the present invention, there is provided a computer-implemented method of analysing messages in a computer system to allow workflows constituted by the messages to be indentified, the method comprising:

analysing a sequence of messages in a computer system in order to classify the messages, thereby producing a corresponding sequence of classifications of the messages; and,

applying sequence induction to the sequence of classifications of the messages to produce (i) a set of sub-sequences of the classifications of the messages and (ii) a sequence grammar for the sub-sequences, from which a workflow constituted by the sequence of messages can be identified.

The messages, which typically are entering or attempting to enter a computer system and which comprise the sequence of elements that make up a workflow, can be observed or collected by standard eaves-dropping techniques. When considering

computing resources and computer networks, the messages can for example be observed by "sniffing" the traffic passing through the computer system or carried by the network. Having collected sequences of messages, the extraction of plausible sub-sequences of classifications of the messages and their grammar can then be used to identify complete and partial workflows. Initially, in a "set-up" phase, a human analyst will identify workflows from the sub-sequences and their grammar. These workflows can be described in a policy workflow description. The policy workflow description can then be monitored and enforced in a running system in which the complete and partial workflows are extracted automatically from the sub-sequences and their grammar and compared to workflows as described in the policy workflow description. The workflows that have been identified can then be used for a number of purposes, including for example securing workflows, metering workflows for accountancy purposes, improving operational effectiveness of process relying on workflows, determining inappropriate workflows and reforming and redirecting inappropriate workflows.

In an embodiment, the sequence of messages is analysed and the messages are classified by clustering the messages according to the semantic intent of the messages. Whilst this classification by clustering the messages according to the semantic intent of the messages is the most preferred method, principally for reasons of efficiency (i.e. speed), other approaches may be used, such as clustering the messages according to their syntax.

In an embodiment, the sequence of messages is analysed and the messages are classified by clustering the messages according to similarity of the messages.

In an embodiment, the classification of a message is the numbers returned in the path sequence of a successful derivation path taken by the message when a stochastic logic program is fitted to the message.

These embodiments make use of the "clustering" techniques disclosed in our EP-A-1830253, US-A-2007/0185703 and US patent application no. 12/187104 and

discussed in more detail below. This provides a very efficient and quick way of appropriately analysing the messages.

In an embodiment, the method comprises storing a copy of each of the messages to allow a representation of the messages and the corresponding sub-sequences of the classifications of the messages and sequence grammars for the sub-sequences to be displayed to a user. This provides the user with a way of easily identifying workflows corresponding to the messages. As well as the copy of each message, other attributes about the message can be stored, including for example: the date and time the message was received; the username or application name that sent the message; network addressing information about the source and destination of the message; and such like.

In an embodiment, the method comprises automatically identifying a workflow constituted by the messages from the set of sub-sequences of the classifications of the messages and the sequence grammar for the sub-sequences. In an embodiment, the method comprises comparing the automatically identified workflow with previously stored workflow descriptions. As noted above, a policy workflow description can then be monitored and enforced in a running system in which the complete and partial workflows are extracted automatically. The workflows that have been identified can then be used for a number of purposes, including for example securing workflows, metering workflows for accountancy purposes, improving operational effectiveness of process relying on workflows, determining inappropriate workflows and reforming and redirecting inappropriate workflows.

In an embodiment, the set of sub-sequences of the classifications of the messages and the sequence grammar for the sub-sequences are obtained by building rules that describe bigrams formed between classifications of the messages in the sequence of classifications of the messages.

In an embodiment, the choice to build a new rule is based on all the bigrams that can be formed given the most recent classification in the input sequence of classifications

of the messages and each of the classifications in the sequence of classifications of the messages falling within a window.

In an embodiment, the choice to build a new rule is based on considering each of the classifications in the sequence of classifications of the messages falling within a window as a single set.

In an embodiment, the choice to build a new rule takes into account the temporal proximity of the classifications in the sequence of classifications.

These embodiments are particularly helpful in providing a level of robustness against noise.

A further aspect of the present invention includes a computer arranged, configured and/or controlled to perform the method of the first aspect of the present invention

Embodiments of the present invention will now be described by way of example with reference to the accompanying drawings, in which:

Fig. 1 shows schematically a computer system interacting with applications;

Fig. 2 shows schematically the collection and clustering of message sequences;

Fig. 3 shows an example of message classification for four specific messages;

Fig. 4 shows an example of a sequence of message classifications;

Fig. 5 shows schematically an example of sequence induction of messages that have been classified;

Fig. 6 shows schematically an example of the identification of the workflows and definition of the workflow descriptions;

Fig. 7 shows an example of a visualization of a sequence grammar which can be presented to a user;

Fig. 8 shows an example of a visualization of a sub-sequence which can be presented to a user; and,

Fig. 9 shows schematically an example of the enforcement of semantic sequences.

In broad terms, embodiments of the present invention operate as follows. A stream of computer messages is captured. The stream of messages is analysed into its constituent components, including in particular the semantic structures to determine precisely the intent of the stream of messages. This provides a classification of each of the constituent messages. Sequence induction methods are then applied to produce a description of the sequences observed based on their classification. The description is further generalized. The underlying sequence elements are then grouped and analysed so that a unified policy can be assigned to the entire group of sequences. This allows for example an operator to approve the elements of the generalized sub-sequences by assessing their appropriateness in the context of the behaviour of the computer system. It also allows the installation of an approved baseline of sub-sequences (or workflows) that can be enforced as a control policy. It also allows a determination, in real time, of correct, incorrect, and novel sequences.

Referring now to Figure 1, there is shown schematically a user 101 using (computer software) applications 102 to interact with a computer resource 103. The interaction with the computer resource 103 is mediated through some computer language via the transmission of messages (MSG) 104 within the computer language. The computer resource 103 may be made available to the applications 102 either directly, for example in the case that the applications 102 and computer resource 103 are part of the

same computer system 106, or indirectly, for example via a computer network 105 in the case that the applications 102 are outside the computer system 106. In either case, the messages 104 can be observed by some observation process 202 and the intent of the message can be determined via a determination process 201.

As mentioned, the interactions that users 101 have with the applications 102 make up the workflows of the computer system 106. These workflows can be effectively determined by the observation process 202 observing the sequences of messages 104 arriving at the computer system 106 which are then subject to the determination process 201. A preferred determination process 201 is shown schematically in Figure 2.

Reference is made here to our EP-A-1830253, US-A-2007/0185703 and US patent application no. 12/187104, the entire contents of which are incorporated herein by reference. In these patent applications, there are disclosed methods for analysing symbols in a computer system. These methods, which we refer to as "Efficient Grammatical Clustering" (EGC), provide a mechanism to understand usage patterns based on the semantics of messages entering (or leaving) computer systems. This allows for example the different database commands entering a relational database system to be recognised so that a baseline of normal behaviour can be determined. EGC enables all new commands/messages (i.e. those that have not been seen previously by the system) to be recognised so that a proactive device can determine whether the message should be allowed to pass to the database or not. The preferred embodiments of the present invention make use of the ECG techniques disclosed in our EP-A-1830253, US-A-2007/0185703 and US patent application no. 12/187104 as part of the determination process 201. These techniques can be fully understood from a review of these patent applications and will only be described relatively briefly here. It should be noted that whilst it is preferred to analyse and classify the messages according to their semantic intent, other approaches may be used, such as analysing and classifying the messages according to their syntax.

Referring again to Figure 2, the sequence of messages 104 are "clustered" by the clustering process 401 disclosed in our EP-A-1830253, US-A-2007/0185703 and US

patent application no. 12/187104. This produces a classification 402 of each message (MSG CLASSIFICATION) which are stored along with a copy of the respective messages 104 in a message store (MSG STORE) 403. (Referring again to our EP-A-1830253, US-A-2007/0185703 and US patent application no. 12/187104, the classification of a message in this context is the numbers returned in the path sequence of a successful derivation path taken by the message when the corresponding stochastic logic program is fitted to the message, i.e. the clause identifiers for the relevant instrumented predicate (given in reverse order).) As well as the copy of each message 104, other attributes about the message 104 can be included in the message store 403. This can include for example: the date and time the message 104 was received; the username or application name that sent the message 104; network addressing information about the source and destination of the message 104; and such like.

For each message 104, the clustering process 401 provides a unique classification of the semantic intent of the message 104 (this being the MSG CLASSIFICATION 402 that is stored in the message store 403). This uniqueness of classification allows messages 104 that are syntactically different to be classified in the same way (i.e. to be classified as being of the same "type" or class) on the basis that their class of semantic intent is identical.

In a particular example, in the context of a computer resource 103 that is a relational database, the messages 104 may be received at the computer resource 103 in the language of Structured Query Language (SQL). An example of the unique message classification 402 for four specific SQL messages, in the sequence in which they arrived, is shown in Figure 3. An example of a sequence of message classifications 402 stored in the message store 403 is shown in Figure 4. It should be noted that, in general, the messages can be a sequence of "tokens" forming a "sentence" in any language. This applies in general to all computer languages, including those languages for inter-communication (e.g. XML, SOAP), protocol languages (e.g. SIP) and scripting languages (e.g. JavaScript), and even to natural (human) languages.

The preferred embodiment then makes use of the message classifications 402, stored in the message store 403, to automatically induce sub-sequences of the message classifications and their descriptions so that the sub-sequences of the messages can be identified with actual workflows, thus allowing the underlying workflow process to be generated or identified. The workflow discovery process of the preferred embodiment is based on a sequence induction process and shown schematically in Figure 5. It may be noted that, in general, it is not always possible to obtain a specification for the process that is generating workflows, for example because the documentation is no longer available or is not accurate.

Thus, referring to Figure 5, the sequence of message classifications 402 is input to a sequence induction process 404 which produces (i) a set of sub-sequences 405 of the of message classifications 402 which are held in a sub-sequence store 406 and (ii) a sequence grammar 407 which provides a generative description of the sub-sequences 405, i.e. the syntax to which the sub-sequences conform.

A number of techniques for the sequence induction process 404 are possible. The preferred embodiment makes use of grammatical induction to efficiently detect sub-sequences from high rate streams. A modified form of the known SEQUITUR algorithm may be employed for this. SEQUITUR is a recursive algorithm that was designed for use as a lossless compression/decompression technique. As described further below, SEQUITUR can be used for inferring a compositional hierarchical structure from strings, i.e. sequences of discrete symbols. It detects repetition and factors it out of the string by forming rules in a grammar. The rules can be composed of "non-terminals", giving rise to a hierarchy. It is useful for recognizing lexical structure in strings, and excels at very long sequences.

In general, one method for sequence induction, i.e. extracting useful recurring sequences from (typically much) larger sequences, is to use or modify techniques used for performing data compression. Many of these techniques stem from information theory principles (e.g. those described by Claude Shannon). Many compression techniques that are good at data compression do not extract valuable structural

information from the data to be compressed. One compression technique that provides good compression properties as well as generating a comprehensible and useful internal structure is SEQUITUR. SEQUITUR, in its original form, is focused on data compression rather than extracting sequence information from temporal data, and has significant limitations. The following description will describe the basic SEQUITUR technique and then preferred modifications that allow sequences to be extracted from noisy temporal sequences.

SEQUITUR was primarily developed as a lossless data compression technique. The preferred embodiments of the present invention utilises a by-product of the SEQUITUR technique to induce sequences.

To achieve compression of data, SEQUITUR takes as an input a sequence of tokens (also known as an input stream) and processes the sequence sequentially from the first element to the last element. The output of the process is a rewritten input sequence (known as rule 0) and a set of rewrite rules in the form of a context-free grammar. Rule 0 is typically much shorter than the input sequence. Rule 0 may contain either the original element or a rule number. To uncompress Rule 0, every rule number mentioned in Rule 0 is replaced with the rule's rewrite symbols (which may themselves contain the original elements from the input sequence (known as "terminals") and/or rule numbers). If the rewrite symbol is another rule, the process is recursively continued until the "terminals" are reached. This guarantees that the original input sequence is precisely recovered from the compressed Rule 0 and all of the descendent rules.

As an illustrative example, consider the sequence of characters "abcabdabcabd". The compressed output produced by SEQUITUR is the set of rewrite rules as follows:

0 -> 1 1

1 -> 2 c 2 d

2 -> a b

To reconstruct the original input from the rules, start with Rule 0 and rewrite the rule with the right hand side of the rule. This process is recursively repeated until no rule

numbers exist. This ensures that the original input is recovered. The following four steps illustrate this:

0 → 1 1 - replace Rule 0 with two occurrences of Rule 1

1 1 → 2 c 2 d 2 c 2 d – replace each occurrence of Rule 1 with Rule 1's right hand side

2 c 2 d 2 c 2 d → a b c a b d a b c a b d – replace each occurrence of Rule 2 with Rule 2's right hand side

a b c a b d a b c a b d – Stop – the is the complete uncompressed version of Rule 0 as no rule numbers exist in the sequence.

SEQUITUR builds rules by considering bigrams formed between two consecutive tokens of the input stream. When a bigram occurs more times than some pre-set threshold in any rule (including the Rule 0 input stream), then the bigram is assigned the next unused rule number, and every occurrence of the bigram in any existing rule (including Rule 0) is replaced with the new rule number. Also, a new rule with this new rule number is added to the list of rules with the new rule number on the left hand side and the bigram tokens on the right hand side. If the result of the creation of a new rule leaves an existing rule no longer participating in the hierarchy, then that existing rule is deleted from the rule set. The result of the process is a hierarchical rule set where each rule has some utility when decompressing Rule 0 back into the original input stream of tokens.

It is possible to work out the uncompressed sequences that relate to any rule. In the context of the preferred embodiments of the present invention, these induced sequences can be used to identify workflows. For example, the full expansion of Rule 1 in the example above is "abcabd" which occurs twice in the original input stream and therefore may be worthy of being assigned some level of importance. If the input stream consists of transaction components, then SEQUITUR's output will contain sequences that may relate to workflows.

A key deficiency of SEQUITUR for present purposes is that it assumes a total ordering of the sequence. When extracting potentially useful sub-sequences from the original input stream, SEQUITUR fails to take into account temporal proximity between the arrival of successive input tokens. SEQUITUR considers the bigram "a b" in the same way whether "b" follows "a" by a millisecond or whether "b" follows "a" an hour later. For the discovery of transaction workflows this is important. Furthermore SEQUITUR is not robust against a noisy input. Consider two inputs "abababababa" and "ababadbaba" where in fact the second input is the same as the first, but is contaminated with a spurious "d". SEQUITUR is poor at ignoring the spurious "d". In monitoring workflows in real computer systems, it often happens that something abnormal is interspersed with the underlying workflow.

The preferred embodiments of the present invention make use of a number of improvements or modifications of SEQUITUR to improve the robustness of inducing sequences and sequence grammars.

First, the preferred embodiments of the present invention may make use of a sliding window/multiple bigram approach during the rule-building phase. A sliding window of a fixed size is now used when considering bigrams for the decision on building rules. The choice to build a new rule is based on all the bigrams that can be formed given the most recent token in the input and each of the tokens in the window of size W in the input.

Consider the following illustrative example. Assume a window size of 3. With the input seen so far being "abababc" and the next input token being "a", then the window contains the subsequence "abc" and the possible bigrams are: "ca" (which is the same as in the normal SEQUITUR case), "ba" and "aa". The rest of the rule building proceeds as with the normal SEQUITUR. It may be noted that it is not possible to use this as a compression/decompression technique. However, it is useful in the present context as part of a sequence induction technique whilst providing a level of robustness against noise.

Secondly, the preferred embodiments of the present invention may make use of a generalised sliding window multi-gram approach during the rule-building phase. This is similar to the approach just described. Instead of forming multiple bigrams, it considers the entire window contents with the newly arrived token as a single set. This set then provides mapping to a new rule number if required.

Consider the following illustrative example. Again, assume a window size of 3. With the input seen so far being "abababc" and the next input token being "a", then the window contains the subsequence "abc". The sequence including the arriving token "a" becomes "abca", which as a set (sorted alphabetically) is {a, b, c}. This unique set can then be used to determine a unique rule number. The rest of the rule building proceeds as with the normal SEQUITUR. Again, it may be noted that it is not possible to use this as a compression/decompression technique. However, it is useful in the present context as part of a sequence induction technique whilst providing a level of robustness against noise.

A further modification is to consider the building of a bigram taking into account the temporal proximity of the tokens in the input stream during the rule-building phase. This can be applied to normal SEQUITUR or either of the two modifications mentioned immediately above.

Consider an input stream "a-1-b-5-a-1-b-5-a-1-b" where the digit "N" (1 and 5 being the only digits present in this example) separating two tokens represents the time between the arrivals of the tokens. A maximum time delay can be added as a threshold to control whether a bigram is built. If, in this example, the maximum time considered is 2 then the bigram "ab" is considered ($1 < 2$), whilst the bigram "ba" is not considered ($5 > 2$). The rest of the rule building proceeds as with the normal SEQUITUR. Again, it may be noted that it is not possible to use this as a compression/decompression technique. However, it is useful in the present context as part of a sequence induction technique whilst providing a level of robustness against noise.

It will be understood that many further variations are possible. Bigram and rule building could be triggered by a form of average of the marginal arrival times of a sequence of tokens, such as the geometric mean. Bigram and rule building could be triggered by some form of spectral analysis of the arrival time differences.

Having obtained the sub-sequences, analysis of the utility of the extracted sub-sequences can then be carried out by a (human) analyst. The analyst can be provided with a visualisation of the original sequence with overlays of the extracted sub-sequences and the sequence grammar. This is shown schematically in Figure 6. In the workflow identification process 501, the sub-sequences 405 and the sequence grammar 407 are presented visually to a workflow user or analyst 103 in the form of a sequence grammar visualization 502 and a sub-sequence visualization 503. This allows the workflow user 103 to identify workflows that correspond to the sub-sequences and therefore to specify workflows from the sub-sequences and make corresponding workflow descriptions 507 which are kept in a workflow description store 508 for use as described further below.

Examples of a sequence grammar visualization 502 and a sub-sequence visualization 503 are shown in Figures 7 and 8 respectively. As shown in Figure 7, the familiar computer file system browser metaphor is used to browse the sequence grammar. The "folders" are rules that describe sub-sequences 405 of message classifications 402 while the "files" are individual messages 104 in the language. For the example in Figure 7, the messages 104 are SQL statements. In Figure 8, a map is produced and shown whereby each message classification is indicated by a unique colour code (seen as the shaded square dots in the black and white figure). The original sequence is presented by the square dots from left-to-right, top-to-bottom. For square dots (i.e. the message classifications) that have been identified by the sequence induction process 404 to belong to a sub-sequence 405 from the sub-sequence store 406, the sub-sequence number is given a unique colour and the presented as a rectangle overlaying the message classifications. This allows the workflow user 103 to identify common sequences and describe them for the workflow description store 508.

Having identified workflows and having stored descriptions of them in the workflow description store 508, further sequences of messages 104 entering or attempting to enter the computer system 106 can be automatically analysed in the same manner and compared to the stored workflow descriptions. Thus, the workflows of the further messages 104 can be discovered, monitored and/or controlled as described in general terms above.

For example, having decided what are acceptable workflows in particular situations, and indeed having built a policy of acceptable workflows, an enforcement process can be put in place. An example of this is shown in Figure 9. The policy of workflows described as workflow descriptions 507 are held in the workflow description store 508. An input message sequence 602 consisting of a sequence of messages 104 is intercepted by the policy comparison process 601. The policy comparison process 601 compares the arriving sequences of messages 104 with the workflow descriptions 507 stored in the workflow description store 508. An input message sequence 602 that is determined to be acceptable by the policy comparison process 601 will be passed as an output message sequence 603 to the destination computer resource 103 without further processing. On the other hand, an input message sequence 602 determined by the policy comparison process 601 as requiring some other action will result in an alternative output message sequence 603 to be sent to the destination computer resource 103. Such an alternative may be to initiate and terminate the sequence with begin and commit commands (e.g. as in the SQL way); may be to send an empty sequence of messages; may be to send a rollback message (e.g. as in the SQL way). A further result of the policy comparison process 601 may be to produce other outputs 604. For example, the other outputs 604 may be, but not limited to: send an alert; cause the sequence of messages to be recorded for further process; etc.

The preferred embodiments of the present invention provided for automatic and efficient extraction of plausible sub-sequences of messages from sequences of messages passing into or across a computer system. The sub-sequences can then be used to identify workflows, which can then be monitored, controlled, etc., as desired.

It will be understood that the methods described herein will typically be carried out by appropriate software running on appropriate computer equipment. The term "computer" is to be construed broadly. The term "a computer" or similar may include several distributed discrete computing devices or components thereof. The computer program may be in the form of source code, object code, a code intermediate source and object code such as in partially compiled form, or in any other form suitable for use in the implementation of the processes according to the invention. The software may be recorded on a carrier, which may be any entity or device capable of carrying the program. For example, the carrier may comprise a storage medium, such as a ROM, for example a CD ROM or a semiconductor ROM, or a magnetic recording medium, for example a floppy disk or hard disk. Further, the carrier may be a transmissible carrier such as an electrical or optical signal which may be conveyed via electrical or optical cable or by radio or other means.

Embodiments of the present invention have been described with particular reference to the examples illustrated. However, it will be appreciated that variations and modifications may be made to the examples described within the scope of the present invention.

CLAIMS

1. A computer-implemented method of analysing messages in a computer system to allow workflows constituted by the messages to be identified, the method comprising:

analysing a sequence of messages in a computer system in order to classify the messages, thereby producing a corresponding sequence of classifications of the messages; and,

applying sequence induction to the sequence of classifications of the messages to produce (i) a set of sub-sequences of the classifications of the messages and (ii) a sequence grammar for the sub-sequences, from which a workflow constituted by the sequence of messages can be identified.

2. A method according to claim 1, wherein the sequence of messages is analysed and the messages are classified by clustering the messages according to the semantic intent of the messages

3. A method according to claim 1 or claim 2, wherein the sequence of messages is analysed and the messages are classified by clustering the messages according to similarity of the messages.

4. A method according to any of claims 1 to 3, wherein the classification of a message is the numbers returned in the path sequence of a successful derivation path taken by the message when a stochastic logic program is fitted to the message.

5. A method according to any of claims 1 to 4, comprising storing a copy of each of the messages to allow a representation of the messages and the corresponding sub-sequences of the classifications of the messages and sequence grammars for the sub-sequences to be displayed to a user.

6. A method according to any of claims 1 to 5, comprising automatically identifying a workflow constituted by the messages from the set of sub-sequences of the classifications of the messages and the sequence grammar for the sub-sequences.

7. A method according to claim 6, comprising comparing the automatically identified workflow with previously stored workflow descriptions.

8. A method according to any of claims 1 to 7, wherein the set of sub-sequences of the classifications of the messages and the sequence grammar for the sub-sequences are obtained by building rules that describe bigrams formed between classifications of the messages in the sequence of classifications of the messages.

9. A method according to claim 8, wherein the choice to build a new rule is based on all the bigrams that can be formed given the most recent classification in the input sequence of classifications of the messages and each of the classifications in the sequence of classifications of the messages falling within a window.

10. A method according to claim 9, wherein the choice to build a new rule is based on considering each of the classifications in the sequence of classifications of the messages falling within a window as a single set.

11. A method according to any of claims 8 to 10, wherein the choice to build a new rule takes into account the temporal proximity of the classifications in the sequence of classifications.

12. A computer program comprising program instructions for causing a computer to perform a method according to any of claims 1 to 11.

13. A computer programmed to carry out a method according to any of claims 1 to 11.

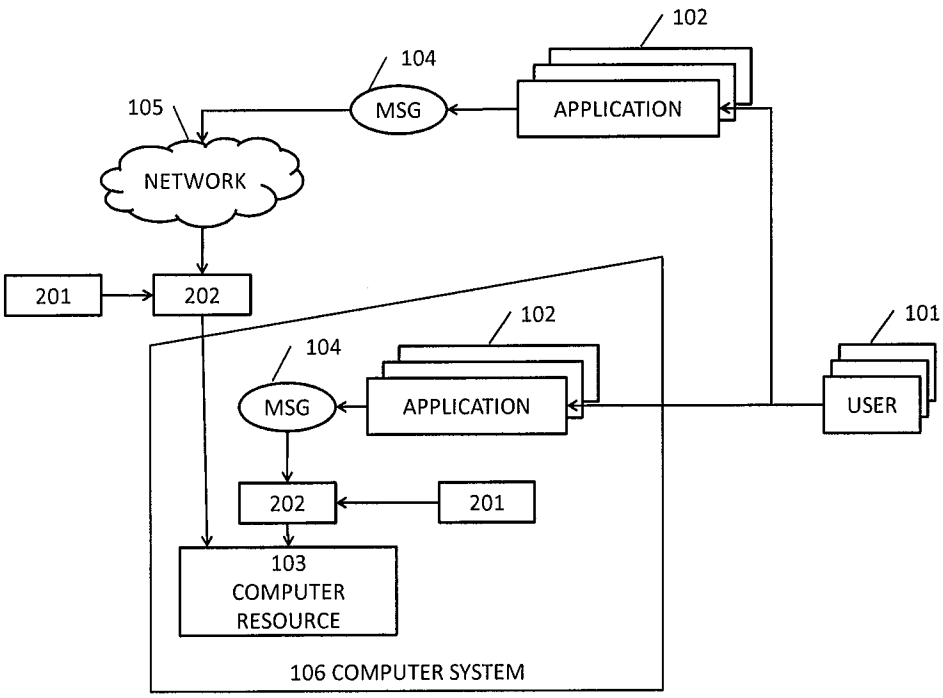

102
104
MSG
APPLICATION
105
NETWORK
201
202
102
101
104
MSG
APPLICATION
USER
202
201
103
COMPUTER
RESOURCE
106 COMPUTER SYSTEM

Fig. 1

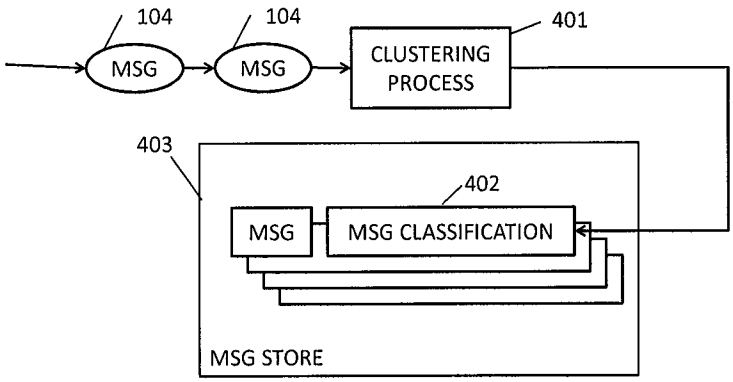

104
104
401
MSG
MSG
CLUSTERING
PROCESS
403
402
MSG
MSG CLASSIFICATION
MSG STORE

Fig. 2

Message Classification	Message
13428614	select max(Sno), min(Sno) from tbl_Syslog
440536710	select 100
376125926	SELECT t3.Sno, t3.MasterIP, t3.MasterPort, t3.SlaveIP, t3.SlavePort, t3.Mode, t3.Enable, t3.Status, t3.Remark, t3.FilterSeverity, t3.FilterString, t3.ForwardNoForNMS, t3.DeleteNoForNMS, t3.ForwardNoForNE, t3.DeleteNoForNE, t3.AlarmState FROM tbl_SyslogServer t3
440536710	select 100

Fig. 3

13428614→440536710 → 376125926 → 440536710 → 13428614 → 440536710 → 376125926 → 440536710 → 13428614 → 440536710 → 376125926 → 440536710 → ... <100's more> → 376125926 → 440536710 → 13428614 → 440536710 → 376125926 → 440536710

Fig. 4

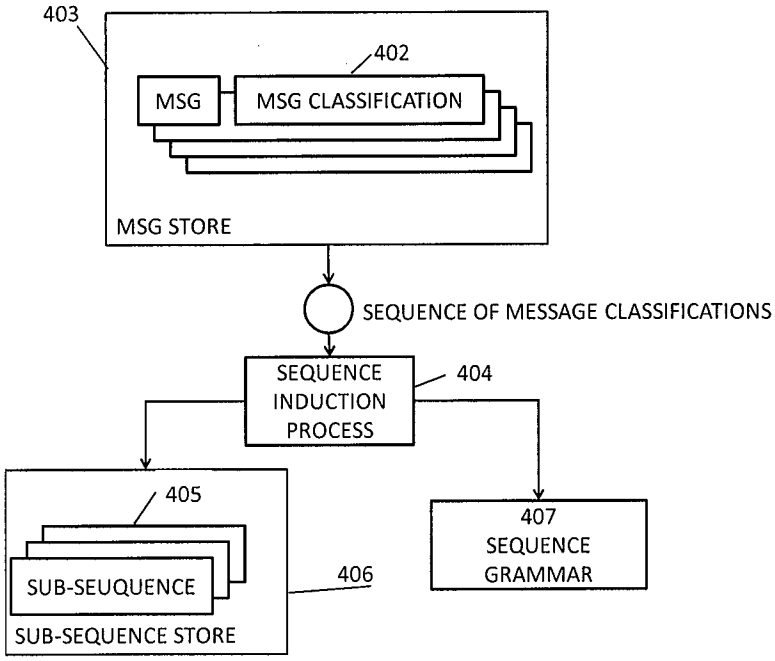

403
402
MSG
MSG CLASSIFICATION
MSG STORE
SEQUENCE OF MESSAGE CLASSIFICATIONS
SEQUENCE INDUCTION PROCESS
404
405
SUB-SEUQUENCE
406
SUB-SEQUENCE STORE
407
SEQUENCE GRAMMAR

Fig. 5

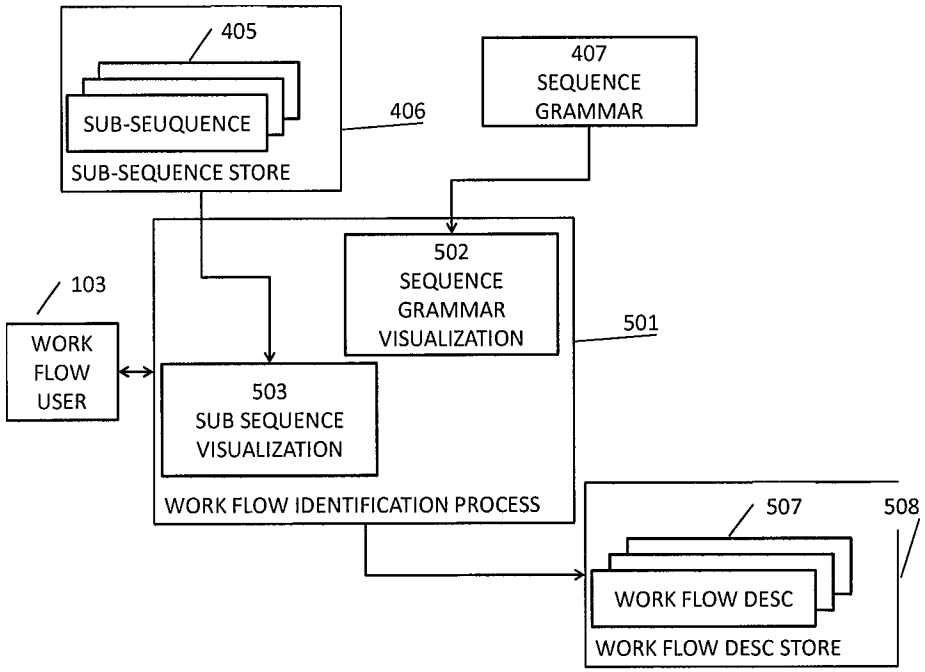

405
SUB-SEUQUENCE
406
SUB-SEQUENCE STORE
407
SEQUENCE GRAMMAR
502
SEQUENCE GRAMMAR VISUALIZATION
501
103
WORK FLOW USER
503
SUB SEQUENCE VISUALIZATION
WORK FLOW IDENTIFICATION PROCESS
507
508
WORK FLOW DESC
WORK FLOW DESC STORE

Fig. 6

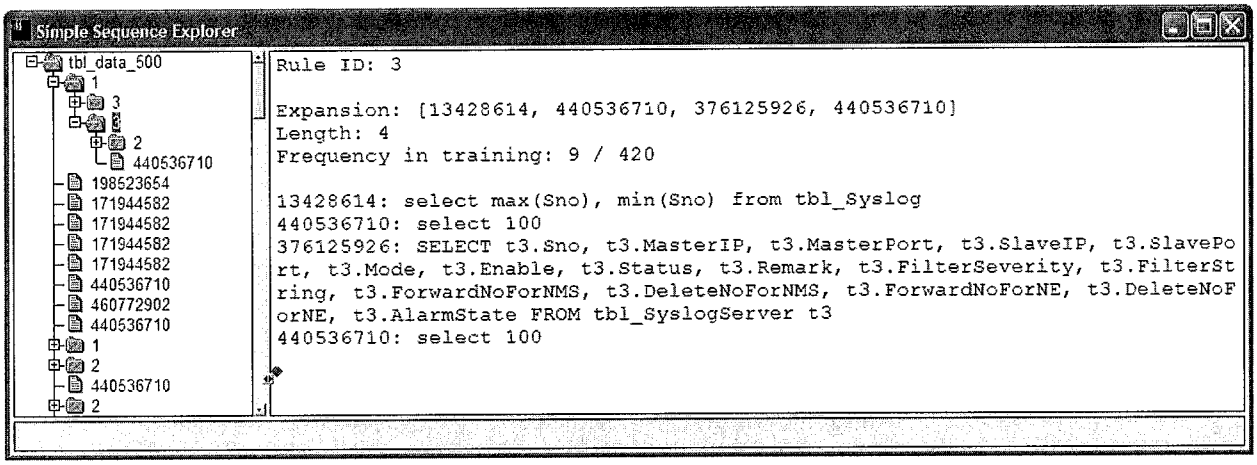

Simple Sequence Explorer
tbl_data_500
1
3
2
440536710
198523654
171944582
171944582
171944582
171944582
440536710
460772902
440536710
1
2
440536710
2
Rule ID: 3
Expansion: [13428614, 440536710, 376125926, 440536710]
Length: 4
Frequency in training: 9 / 420
13428614: select max(Sno), min(Sno) from tbl_Syslog
440536710: select 100
376125926: SELECT t3.Sno, t3.MasterIP, t3.MasterPort, t3.SlaveIP, t3.SlavePo
rt, t3.Mode, t3.Enable, t3.Status, t3.Remark, t3.FilterSeverity, t3.FilterSt
ring, t3.ForwardNoForNMS, t3.DeleteNoForNMS, t3.ForwardNoForNE, t3.DeleteNoF
orNE, t3.AlarmState FROM tbl_SyslogServer t3
440536710: select 100

Fig. 7

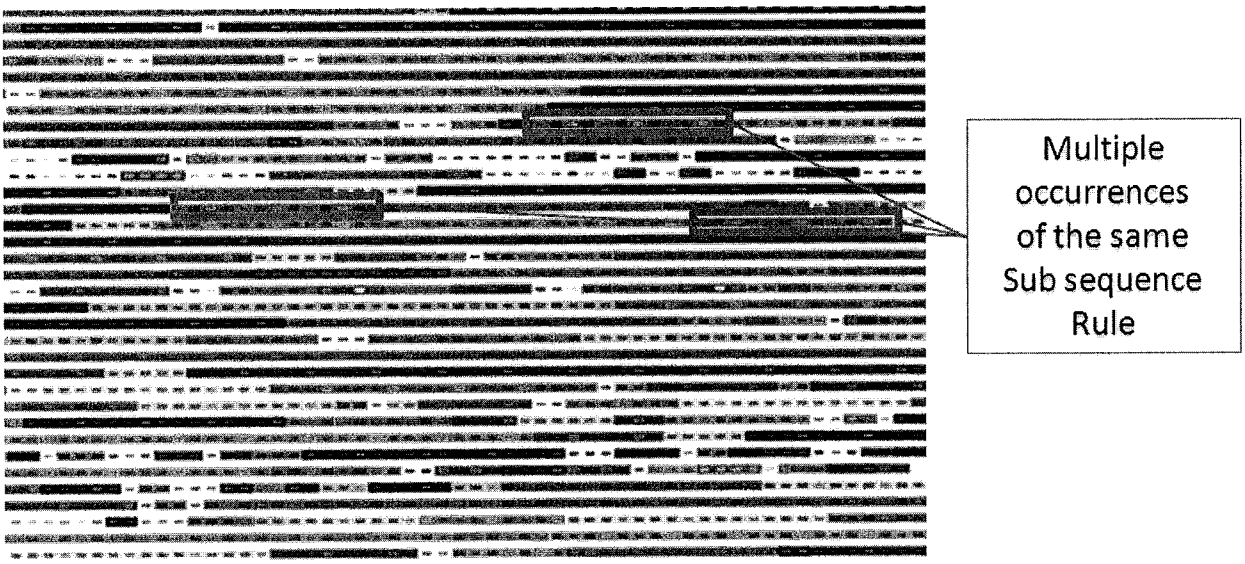

Multiple occurrences of the same Sub sequence Rule

Fig. 8

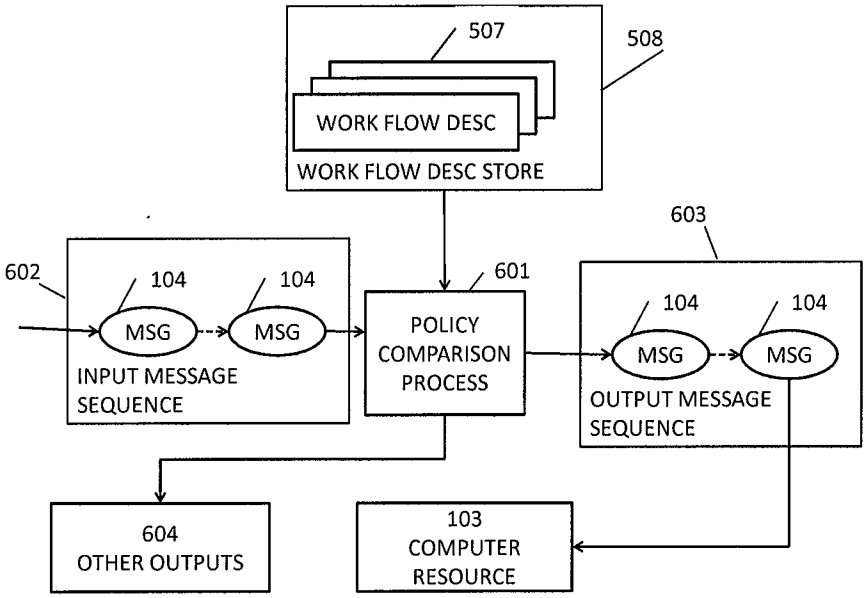

507
508
WORK FLOW DESC
WORK FLOW DESC STORE
603
602
104
104
601
104
104
MSG
MSG
POLICY
COMPARISON
PROCESS
MSG
MSG
INPUT MESSAGE
SEQUENCE
OUTPUT MESSAGE
SEQUENCE
604
OTHER OUTPUTS
103
COMPUTER
RESOURCE

Fig. 9

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2010/050074

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F11/30 G06F21/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 1 830 253 A2 (SECERNO LTD [GB]) 5 September 2007 (2007-09-05) the whole document paragraph [0001] -----	1-7,12, 13
A	WO 02/09339 A2 (PERIBIT NETWORKS INC [US]) 31 January 2002 (2002-01-31) page 2, line 7 - page 4, line 8 page 11, line 1 - line 18 page 14, line 13 - page 17, line 17 page 18, line 25 - page 19, line 5 page 22, line 10 - line 17 page 29, line 24 - page 30, line 14 ----- -/--	1-7,12, 13

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 April 2010	18/05/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Kielhöfer, Patrick

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2010/050074

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
T	NEVILL-MANNING C G ET AL: "Compression and explanation using hierarchical grammars" COMPUTER JOURNAL OXFORD UNIVERSITY PRESS FOR BRITISH COMPUT. SOC UK, vol. 40, no. 2-3, 1997, pages 103-116, XP002579593 ISSN: 0010-4620 Retrieved from the Internet: URL:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.1150> the whole document	

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2010/050074

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 8-11(completely); 1-7, 12, 13(partially)
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 8-11(completely); 1-7, 12, 13(partially)

Claims 1-12 are drafted in such a way that they are not in compliance with Art.6 PCT which makes it particularly burdensome for a skilled person to establish the subject-matter for which protection is sought. The claims use terms without a unique and/or well-defined (technical) meaning, e.g. "sequence induction" (claim 1), "sequence grammar for sub-sequences" (claim 1, 5, 6, 8), "semantic intent" (claim 2), "similarity of messages" (claim 3), "successful derivation path" (claim 4), "stochastic logic program" (claim 4), "bigrams" (claim 8, 9). Furthermore, claims 2, 4, 6, 7, 8 do not meet the requirements of Article 6 PCT because the matter for which protection is sought is not clearly defined. The claims attempt to define the subject-matter in terms of the result to be achieved (clustering messages according to semantic intent, a stochastic logic program is fitted to the message, automatically identifying a workflow, comparing a workflow with previously stored workflow descriptions, building rules that describe bigrams), which merely amounts to a statement of the underlying problem, without providing the technical features necessary for achieving this result. Claims 4 uses a completely incomprehensible wording: "the classification of a message is the numbers returned in the path sequence of a successful derivation path taken by the message when a stochastic logic program is fitted to the message". The term "bigrams" used in claim 8 refers to "groups of two written letters, two syllables, or two words, and are very commonly used as the basis for simple statistical analysis of text" (wikipedia: bigram). According to the description p.12 l.14-23, "the invention make use of a number of improvements or modifications of SEQUITUR", one being "a sliding window of a fixed sized [...] when considering bigrams for the decision". These features therefore relate to scientific or mathematical theories which is not required to be searched under Rule 39.1(i) PCT. As claims 9-11 depend from claim 8, the same reasoning applies to these claims. The search has been carried out on the definition of the invention, given in broad terms, on p.6 l.13-25 of the description.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.2), should the problems which led to the Article 17(2) declaration be overcome.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2010/050074

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1830253 A2	05-09-2007	NONE	
WO 0209339 A2	31-01-2002	AT 446612 T	15-11-2009
		AU 7791401 A	05-02-2002
		AU 2001277914 B2	06-07-2006
		CA 2418314 A1	31-01-2002
		CN 1630984 A	22-06-2005
		EP 1307967 A2	07-05-2003
		IL 153957 A	31-10-2007
		JP 2004511928 T	15-04-2004
		NZ 523657 A	26-11-2004
		US 2005169364 A1	04-08-2005
		US 2009022413 A1	22-01-2009
		US 2002037035 A1	28-03-2002