

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
3 August 2006 (03.08.2006)

PCT

(10) International Publication Number
WO 2006/079892 A2

(51) International Patent Classification:
G06F 17/30 (2006.01) **G06F 17/20** (2006.01)

(21) International Application Number:
PCT/IB2006/000104

(22) International Filing Date: 23 January 2006 (23.01.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/042,445 25 January 2005 (25.01.2005) US

(71) Applicant (for all designated States except US): **NOKIA CORPORATION** [FI/FI]; Keilalahdentie 4, FIN-02150 Espoo (FI).

(71) Applicant (for LC only): **NOKIA, INC.** [US/US]; 6000 Connection Drive, Irving, TX 75039 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **TAIN, Jilei** [FI/FI]; Tohtorinkatu 32 B 9, FIN-33720 Tampere (FI). **NURMI-NEN, Jani** [FI/FI]; Koskelantie 9, FIN-37560 Lempaala (FI).

(74) Agents: **ALBERT, Peter, G., Jr.** et al.; Foley & Lardner LLP, 321 N. Clark Street, Suite 2800, Chicago, IL 60610 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, 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, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with declaration under Article 17(2)(a); without abstract; title not checked by the International Searching Authority

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD FOR OPTIMIZING RUN-TIME MEMORY USAGE FOR A LEXICON

(57) Abstract:



WO 2006/079892 A2

SYSTEM AND METHOD FOR OPTIMIZING RUN-TIME MEMORY USAGE FOR A LEXICON

FIELD OF THE INVENTION

[0001] The present invention relates generally to speech and language processing techniques. More particularly, the present invention relates to systems that require lexicon minimization and fast search capabilities when the language at issue has a relatively large character set, such as the Chinese language.

BACKGROUND OF THE INVENTION

[0002] Speech and language processing techniques, such as automatic speech recognition (ASR) and text-to-speech (TTS) synthesis, are constantly becoming increasingly important in multimedia systems. Many of these multimedia systems require a lexicon or dictionary for particular languages. A lexicon typically contains a great deal of information, including words, pronunciations, part-of-speech (POS), and other syntactic and semantic information. POS is a primitive form of linguistic theory that posits a restricted inventory of word-type categories such as nouns, verbs, etc. Such lexicons therefore normally require a large amount of memory.

[0003] Because of these memory issues, when a system processes text by using lexicon data, the lexicon data requires a very high run-time memory footprint and a time-consuming search. This particularly applies to the Chinese language and other languages with a large character set. For the Chinese language, for example, this is due to the fact that there are at least 20,901 Chinese characters (in a closed set) and more than 100,000 Chinese words (in an open set). A Chinese character is a basic written unit and is denoted by two bytes in Unicode. Unicode provides a unique number for every character, regardless of the platform, the program, and the language. The whole set of Chinese characters contains 20,901 characters, including the simplified and the traditional character sets. In the Unicode chart, Chinese characters are represented using the range from 4E00 to 9FA5. A Chinese character has ambiguous meanings and may have multiple pronunciations. Pronunciation of

Chinese character is presented by monosyllable pinyin. A Chinese word is a sequence of Chinese characters without separators. For example, for a given word consisting of N characters, can denoted as an array word[N]. In this particular arrangement, word[0] stands for the Unicode value of the first character in the word.

[0004] Until now, conventional systems having lexicon data from a similarly complex language typically have been realized on other platforms than embedded systems, where the memory and processing power were not major implementation bottlenecks. For this reason, memory and speed optimization issues for handling lexicon data have not been seriously addressed in the past.

[0005] When processing Chinese lexicon data, the system normally downloads the whole data into a memory unit and assigns the data to predefined data structures. Alternatively, the whole data can be split into N parts and loaded one small part at a time. Although run-time memory can be reduced to some extent when more data partitions are enabled, the system still needs to load the whole lexicon data (by distributing or balancing the load over time), and either repeatedly scan the given text with each partition of the data, or load the partitioned data repeatedly. This leads to the very frequent data transfer from file to memory and increases the text processing time.

SUMMARY OF THE INVENTION

[0006] The present invention addresses the issues identified above by introducing an intermediate location variable. The intermediate location variable serves as a bridge between the lexicon database and a given text. Only location data is extracted and loaded from the lexicon data and stored in the run-time memory. Therefore, the amount of data in run-time memory is significantly reduced. Furthermore, given a word, the corresponding location value can be obtained directly without a time-consuming search. This results in a significant reduction in search time as well.

[0007] The present invention can be applied to virtually any system that requires a Chinese lexicon or a lexicon for other languages with large character sets such as Japanese. In particular, the present invention can be an integral part of Chinese ASR and TTS systems, and the present invention enables a system to achieve a very low

run-time memory, fast search and reduced lexicon data size without any loss of text processing accuracy. For example and in one reference implementation, in a Chinese TTS system, the run-time memory footprint for a lexicon was reduced from 4.5MB to 63kB with the present invention, i.e. the size was reduced to 1/76 of the original size. The present invention can be implemented in conjunction with a wide variety of voice user interface software programs.

[0008] These and other objects, advantages and features of the invention, together with the organization and manner of operation thereof, will become apparent from the following detailed description when taken in conjunction with the accompanying drawings, wherein like elements have like numerals throughout the several drawings described below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] Figure 1 is a flow chart showing a process for the implementation of one embodiment of the present invention;

[0010] Figure 2 is a perspective view of a mobile telephone that can be used in the implementation of the present invention; and

[0011] Figure 3 is a schematic representation of the telephone circuitry of the mobile telephone of Figure 2.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0012] The present invention involves the introduction of an intermediate location variable. The intermediate location variable serves as a bridge between the lexicon database and a given text. Only location data is extracted and loaded from the lexicon database and stored in the run-time memory. Therefore, the amount of data in run-time memory is significantly reduced. Furthermore, given a word, the corresponding location value can be obtained directly without a time-consuming search. This results in a significant reduction in search time as well.

[0013] Figure 1 is a flow chart showing the implementation of one embodiment of the present invention. In the case of the Chinese language, the Chinese lexicon data is arranged in ascending order by Unicode values for the words. It should be noted that,

although the examples provided herein discusses the use of Unicode values, the present invention can use virtually any set of uniquely defined code values and is not limited to Unicode values. This is represented at step 100. For example, the data can be stored in the stream of:

Basic_info → *item*[0] → ... → *item*[*Word_Number*-1]

[0014] *Basic_info* contains information such as the version, the number of words, the number of POS, etc. Additional pieces of information can also be included in various embodiments of the present invention. Each *item* is presented in the stream of:

word[*Character_Number*] → *pronunciation* → *pos* → *multiple_pronunciations* → *multiple_pos*;

[0015] Ascending order means that for any *i* between 0 and a number of words, there is

$$item[i].word[0] \geq item[i-1].word[0]$$

[0016] where *item*[*i*].*word* is the Unicode sequence and *word*[*i*] is the Unicode value of the *i*+1-th character in the *word*.

[0017] When processing the lexicon data, instead of loading the whole lexicon data into the run-time memory, a location array *location*[*Character_Number*] is established at step 110. Location variable *location*[*i*] indicates the length from the beginning of the lexicon data file to the first *item* whose first character has a Unicode value derived from *i*. Known Chinese characters are defined in the range of 4E00 and 9FA5 in the Unicode chart. Since the size of lexicon data is usually about 1MB, 3 Bytes (0-4MB) are more than enough to represent one entry in the location array.

[0018] The location array can be defined as:

```
#define CHARACTER_NUMBER 20,901
```

```
#define LOCATION_SIZE 3
```

```
#define UNICODE_START 4E00
```

```
location = malloc(CHARACTER_NUMBER * LOCATION_SIZE);
```

Location values can be extracted from lexicon data. The reference algorithm for extracting location information is shown below:

For *i*=0

```

For i = 0 to CHARACTER_NUMBER - 1{
  location[i] = -1
}

```

```

unicode_value = item[0].word[0]
location[unicode_value - UNICODE_START] = location of item[0]

```

```

For i = 1 to CHARACTER_NUMBER - 1{
  if item[i].word[0] and ≠ item[i-1].word[0]
  then{ unicode_value = item[i].word[0]
        location[unicode_value - UNICODE_START] = location of item[i] }
}

```

[0019] Searching any given *word* from the lexicon file can be accomplished very quickly. The first step, which is represented at step 120, is to locate the lexicon entries that contain a matching initial character:

```

unicode_value = word[0];
start = location[unicode_value-UNICODE_START];
length = location[next_unicode_value-UNICODE_START]-location[unicode_value-
UNICODE_START];

```

[0020] In the above pseudo code, *next_unicode_value* is the next *unicode_value* for which *location[next_unicode_value]* is not -1.

[0021] The next step, represented at step 130, is to load the *length* size data from *start* in the lexicon data file, and save it into the run-time memory. Loaded data is usually very small (on the order of less than 1KB) and a binary search, represented at step 140, can be applied to find matching *item* to given *word* within a very small range. Alternatively, the search can be performed during the loading process. With this technique, it is possible to limit the size of the memory block needed for the loaded data to the maximum size of a single lexicon entry.

[0022] The location data can be extracted on-line during the initialization phase. However, the location data is preferably extracted during the off-line processing of lexicon data, and then stored as part of the lexicon data. This process does not necessarily increase the lexicon data size. If the location data is stored in the lexicon

data, the first character of every word is already known, so the first character becomes redundant and can be omitted. For example, in the case of given *text[]*, the code value of the first character in the text string (*text[]*) is *unicode_value=text[0]*. The *start* and *length* can then be used as mentioned above. The first character of all words between *start* and *start+length* in the lexicon data has the same Unicode value of *unicode_value*; there is therefore no need to store that information.

[0023] In order to better understand the outcome of the memory optimization system and method of the present invention, the following analysis is shown being carried out on the Chinese lexicon used in one high quality text-to-speech system as a reference.

[0024] The lexicon data of this example contains 20,901 Chinese characters (full Unicode set), 92,901 words and 68 POS. The size of the lexicon data in the file is 1,119,707 bytes. After loading the lexicon data into data structures in the run-time memory, the size is 4,771,860 bytes. The maximum memory used in this particular system is 8,859,922 bytes. This indicates that the run-time memory usage must be reduced for embedded platforms.

[0025] Because the size of the lexicon data is about 1MB, 3 Bytes (0-4MB) are sufficient to represent the location information. The number of different characters is 20,901, so the size of the location array is $20,901 \times 3 = 62,703$ Bytes = 61.2 KB, which represents about 5% of the total lexicon data.

[0026] Since the loaded lexicon data in the run-time memory takes 4,771,860 bytes and the number of characters is 20,901, the average size of each character in run-time memory is 228 bytes. In total, the average run-time memory required by the system and method of the present invention is $62,703 + 228 = 62,931 = 61.5$ KB. The gain on memory usage between a conventional system and a system incorporating the present invention is $4,771,860/62,931 =$ a factor of 76. The search complexity is also reduced because each search is only conducted on a small amount of data, rather than on the whole database as was required in a conventional system.

[0027] Regarding the lexicon data, the location information takes a 61.2KB overhead, but all of the first characters can be removed with size of 92,901 (number of words) $\times 2$ (two bytes of each character) = 185,802 Bytes. In this case, the lexicon

data savings is $185,802 - 62,931 = 122,871$ Bytes = 120 KB. Thus, in addition to other advantages, about 10% a size reduction of the lexicon data file is also obtained with the present invention.

[0028] Figures 2 and 3 show one representative mobile telephone 12 within which the present invention may be implemented. It should be understood, however, that the present invention is not intended to be limited to one particular type of mobile telephone 12 or other electronic device. For example, the present invention can be incorporated into personal digital assistants (PDAs), integrated messaging devices (IMDs), notebook computers, handheld computers, and other devices. The mobile telephone 12 of Figures 2 and 3 includes a housing 30, a display 32 in the form of a liquid crystal display, a keypad 34, a microphone 36, an ear-piece 38, a battery 40, an infrared port 42, an antenna 44, a smart card 46 in the form of a UICC according to one embodiment of the invention, a card reader 48, radio interface circuitry 52, codec circuitry 54, a controller 56 and a memory 58. Individual circuits and elements are all of a type well known in the art, for example in the Nokia range of mobile telephones.

[0029] The present invention is described in the general context of method steps, which may be implemented in one embodiment by a program product including computer-executable instructions, such as program code, executed by computers in networked environments.

[0030] Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Computer-executable instructions, associated data structures, and program modules represent examples of program code for executing steps of the methods disclosed herein. The particular sequence of such executable instructions or associated data structures represents examples of corresponding acts for implementing the functions described in such steps.

[0031] Software and web implementations of the present invention could be accomplished with standard programming techniques with rule based logic and other logic to accomplish the various database searching steps, correlation steps, comparison steps and decision steps. It should also be noted that the words "component" and "module" as used herein, and in the claims, is intended to

encompass implementations using one or more lines of software code, and/or hardware implementations, and/or equipment for receiving manual inputs.

[0032] The foregoing description of embodiments of the present invention have been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the present invention to the precise form disclosed, and modifications and variations are possible in light of the above teachings or may be acquired from practice of the present invention. The embodiments were chosen and described in order to explain the principles of the present invention and its practical application to enable one skilled in the art to utilize the present invention in various embodiments and with various modifications as are suited to the particular use contemplated.

WHAT IS CLAIMED IS:

1 1. A method of extracting information from a lexicon and using the
2 information with a computer software program, comprising the steps of:
3 arranging lexicon data for a language using uniquely defined code
4 values for characters of words included in the lexicon;
5 creating a location array for the lexicon data arranged by uniquely
6 defined code value;
7 upon a request to search for a word, identifying words having a
8 matching initial character as the searched-for word using the location array; and
9 searching through the identified words for an identified word that
10 matches the searched-for word.

1 2. The method of claim 1, further comprising the step of, before or during
2 searching through the identified words for an identified word that matches the
3 searched-for word, loading one or more words having the same initial character as the
4 searched-for-word.

1 3. The method of claim 1, wherein the lexicon data is arranged in
2 ascending order by uniquely defined code value.

1 4. The method of claim 1, wherein the language is Chinese.

1 5. The method of claim 1, wherein the language is Japanese.

1 6. The method of claim 1, wherein the computer software program
2 comprises a speech recognition program.

1 7. The method of claim 1, wherein the computer software program
2 comprises a text-to-speech synthesis program.

1 8. A computer program product for extracting information from a lexicon
2 and using the information with a computer software program, comprising:

3 computer code for arranging lexicon data for the language using
4 uniquely defined code values for characters of words included in the lexicon;
5 computer code for creating a location array for the lexicon data
6 arranged by uniquely defined code value;
7 computer code for, upon a request to search for a word, identifying
8 words having a matching initial character as the searched-for word; and
9 computer code for searching through the identified words for an
10 identified word that matches the searched-for word.

1 9. The computer program product of claim 8, further comprising
2 computer code for, before or during searching through the identified words for an
3 identified word that matches the searched-for word, loading one or more words
4 having the same initial character as the searched-for-word.

1 10. The computer program product of claim 8, wherein the lexicon data is
2 arranged in ascending order by uniquely defined code value.

1 11. The computer program product of claim 8, wherein the language is
2 Chinese.

1 12. The computer program product of claim 8, wherein the language is
2 Japanese.

1 13. The computer program product of claim 8, wherein the computer
2 software program comprises a speech recognition program.

1 14. The computer program product of claim 8, wherein the computer
2 software program comprises a text-to-speech synthesis program.

1 15. An electronic device, comprising:
2 a processor and
3 a memory unit operatively connected to the processor,

4 wherein the memory unit and the processor cooperate to extract
5 information from a lexicon and use the information with a computer software
6 program, the extraction and use comprising the steps of:
7 arranging lexicon data for the language using uniquely defined code
8 values for characters of words included in the lexicon;
9 creating a location array for the lexicon data arranged by uniquely
10 defined code value;
11 upon a request to search for a word, identifying words having a
12 matching initial character as the searched-for word; and
13 searching through the identified words for an identified word that
14 matches the searched-for word.

1 16. The electronic device of claim 15, wherein the extraction and use
2 further comprises for the step of, before or during searching through the identified
3 words for an identified word that matches the searched-for word, loading one or more
4 words having the same initial character as the searched-for-word.

1 17. The electronic device of claim 15, wherein the lexicon data is arranged
2 in ascending order by uniquely defined code value.

1 18. The electronic device of claim 15, wherein the language is selected
2 from the group consisting of Chinese and Japanese.

1 19. The electronic device of claim 15, wherein the computer software
2 program comprises a speech recognition program.

3 20. The electronic device of claim 15, wherein the computer software
4 program comprises a text-to-speech synthesis program.

1 / 2

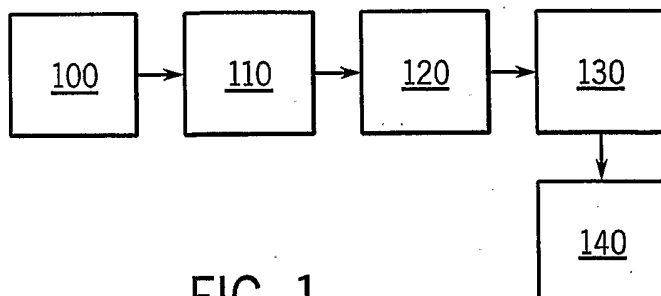


FIG. 1

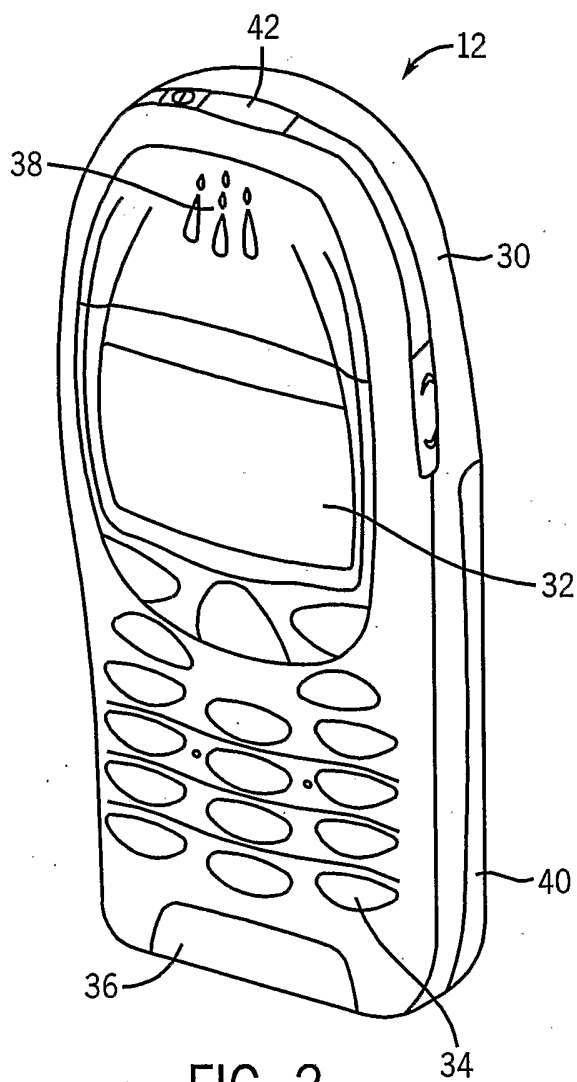


FIG. 2

2 / 2

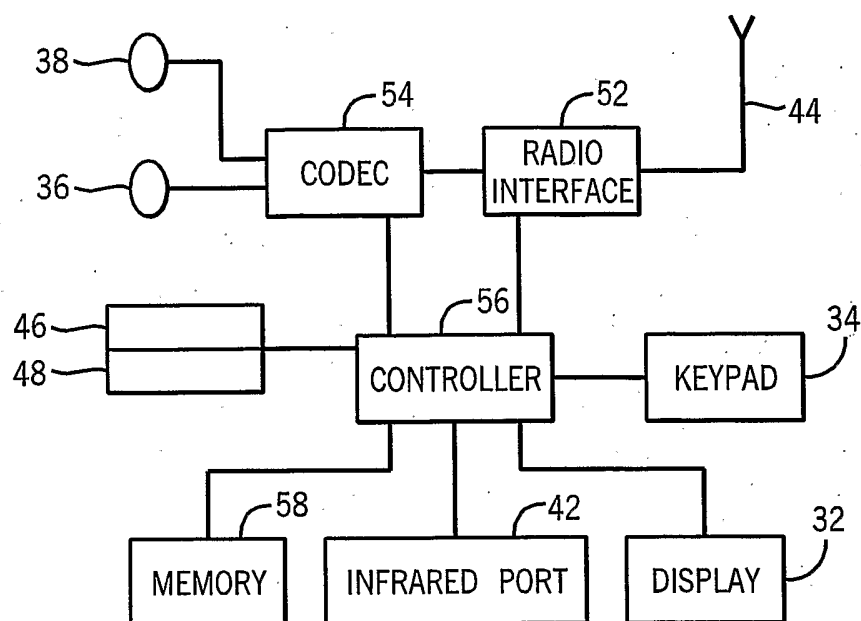


FIG. 3

PATENT COOPERATION TREATY

PCT

DECLARATION OF NON-ESTABLISHMENT OF INTERNATIONAL SEARCH REPORT

(PCT Article 17(2)(a), Rule 13ter.1(c) and 39)

Applicant's or agent's file reference 037145-0826	IMPORTANT DECLARATION	Date of mailing (day/month/year) 05-05-2006
International application No. PCT/IB2006/000104	International filing date (day/month/year) 23-01-2006	(Earliest) Priority Date (day/month/year) 25-01-2005
International Patent Classification (IPC) or both national classification and IPC See extra sheet		
Applicant Nokia Corporation et al		

This International Searching Authority hereby declares, according to Article 17(2)(a), that **no international search report will be established** on the international application for the reasons indicated below.

1. ☒ The subject matter of the international application relates to:
 - a. ☐ scientific theories.
 - b. ☐ mathematical theories.
 - c. ☐ plant varieties.
 - d. ☐ animal varieties.
 - e. ☐ essentially biological processes for the production of plants and animals, other than microbiological processes and the products of such processes.
 - f. ☐ schemes, rules or methods of doing business.
 - g. ☒ schemes, rules or methods of performing purely mental acts.
 - h. ☐ schemes, rules or methods of playing games.
 - i. ☐ methods for treatment of the human body by surgery or therapy.
 - j. ☐ methods for treatment of the animal body by surgery or therapy.
 - k. ☐ diagnostic methods practised on the human or animal body.
 - l. ☐ mere presentations of information.
 - m. ☐ computer programs for which this International Searching Authority is not equipped to search prior art.
2. ☐ The failure of the following parts of the international application to comply with prescribed requirements prevents a meaningful search from being carried out:

☐ the description
 ☐ the claims
 ☐ the drawings
3. ☐ A meaningful search could not be carried out without the sequence listing; the applicant did not, within the prescribed time limit:

☐ furnish a sequence listing on paper complying with the standard provided for in Annex C of the Administrative Instructions, and such listing was not available to the International Searching Authority in a form and manner acceptable to it.

☐ furnish a sequence listing in electronic form complying with the standard provided for in Annex C of the Administrative Instructions, and such listing was not available to the International Searching Authority in a form and manner acceptable to it.

☐ pay the required late furnishing fee for the furnishing of a sequence listing in response to an invitation under Rule 13ter.1(a) or (b).
4. ☐ A meaningful search could not be carried out without the tables related to the sequence listings; the applicant did not, within the prescribed time limit, furnish such tables in electronic form complying with the technical requirements provided for in Annex C-bis of the Administrative Instructions, and such tables were not available to the International Searching Authority in a form and manner acceptable to it.
5. Further comments: See extra sheet

Patent- och registreringsverket

Box 5055

S- 102 42 Stockholm

Tel: + 46 8 782 25 00 Fax: + 46 8 666 02 86

Authorized officer

Oskar Pihlgren /itw

Telephone no. 08 782 2666

**DECLARATION OF NON-ESTABLISHMENT
OF INTERNATIONAL SEARCH REPORT**

International application No.
PCT/IB2006/000104

Continuation of cover sheet

G06F 17/30 (2006.01)

G06F 17/20 (2006.01)

**DECLARATION OF NON-ESTABLISHMENT
OF INTERNATIONAL SEARCH REPORT**

International application No.
PCT/IB2006/000104

The claims relate to subject matter for which no search is required according to Rule 39 PCT. Given that the claims are formulated in terms of such subject matter or merely specify commonplace features relating to its technological implementation, the search examiner could not establish any technical problem which might potentially have required an inventive step to overcome. Hence it was not possible to carry out a meaningful search into the state of the art (Art. 17(2)(a)(i) and (ii) PCT).

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 subject of an international preliminary examination (Rule 66.1 (e) PCT). The applicant is advised that the policy of International Preliminary Examining Authorities is 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 the receipt of the "Declaration of non-establishment..." or during any Chapter II procedure. If the application proceeds into the national/regional phase before a national/regional office, the applicant is reminded that a search may be carried out during examination of the national/regional office, should the problems which led to the Article 17(2) declaration be overcome.