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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 November 2004 (18.11.2004)

PCT

(10) International Publication Number
WO 2004/100016 A1

(51) International Patent Classification[7]: G06F 17/28, 17/27

(21) International Application Number: PCT/US2004/014313

(22) International Filing Date: 6 May 2004 (06.05.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/468,573 6 May 2003 (06.05.2003) US

(71) Applicant: **AMERICA ONLINE INCORPORATED** [US/US]; 22000 AOL Way, Dulles, VA 20166 (US).

(72) Inventor: **TEI, Takayuki**; 360 West Caribbean Drive, Sunnycale, CA 94089 (US).

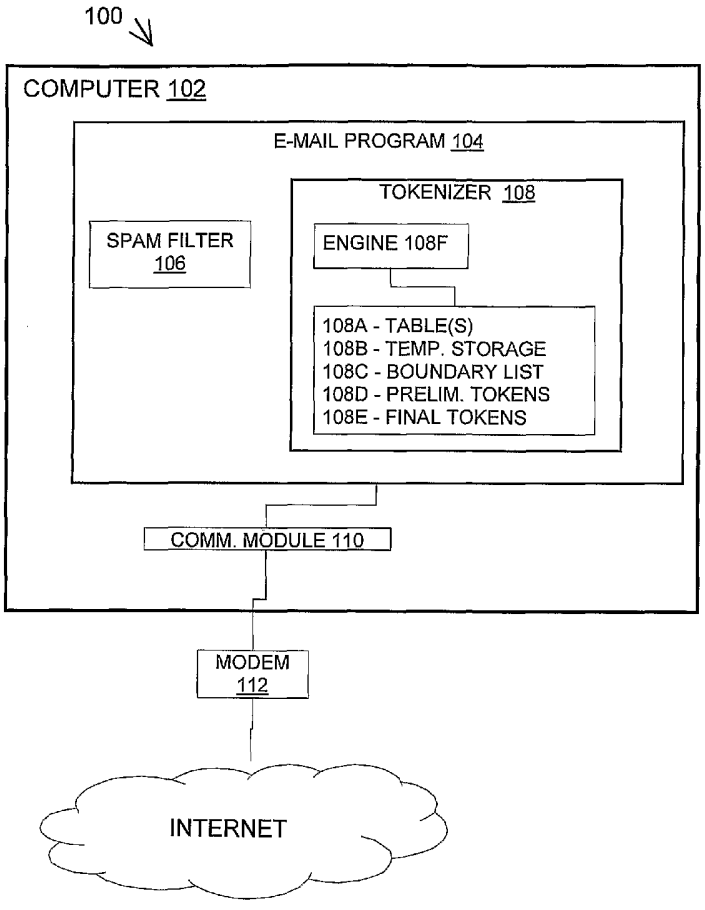
(74) Agents: **GLENN, Michael, A.** et al.; Glenn Patent Group, 3475 Edison Way, Suite L, Menlo Park, CA 94025 (US).

(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, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, SC, SD, SE, SG, SK, SL, 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, IT, LU, 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).

[Continued on next page]

(54) Title: NON-DICTIONARY BASED JAPANESE LANGUAGE TOKENIZER



(57) Abstract: In a method for parsing Japanese language script, an electronic representation of a series of Japanese characters is received, and the characters are reviewed to identify each character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, or punctuation mark. Boundaries (108C) occur between characters of different predetermined types. Each group of one or more characters between adjacent boundaries is designated as a preliminary token (108D). Predetermined context rules are applied to selectively redefine the preliminary tokens. An output of final tokens (108E) is provided, comprising the redefined preliminary tokens.

WO 2004/100016 A1

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

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.

NON-DICTIONARY BASED JAPANESE LANGUAGE TOKENIZER

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a tokenizer for analyzing character-based languages such as Japanese. Unlike previous Japanese language tokenizers, the subject tokenizer does not require the use of cumbersome dictionaries.

Description of the Related Art

Spam is the pernicious plague of the Internet. Most spam constitutes unsolicited commercial e-mail sent en masse. To combat spam, there are numerous e-mail filters on the market today. Most of these are designed for English and other Western languages. Few if any are designed for Eastern languages. This is because Eastern languages can be significantly more difficult to process and analyze. Japanese, for example, is a language of script, which does not use spaces as boundaries between words. Of course, when Japanese script is printed on paper or displayed onscreen, the characters are spaced in order to promote readability. However, when Japanese script is entered into a computer, the characters are entered one after another, without any intervening spaces. In Japanese, some words are formed by one character and other words are formed by multiple characters. Without having a space or other marker between character groups that form

words, it is up to the reader to mentally assemble characters into appropriate words. This works fine for human beings, but presents a challenge for computers. And for this reason, it is especially difficult for computer programs to recognize when Japanese language script contains spam.

To address this issue, the present inventor considered the novel approach of employing a Japanese language "tokenizer" to preliminarily organize Japanese characters into the intended words, before applying a spam filter. A tokenizer is a specialized program that extracts words from a document or other source. Tokenizers work by comparing the source script to words in a sizeable, predefined dictionary. Although these tokenizers are satisfactory for some applications, such as web-based cataloguing servers, the present inventor found that they are not well suited for use as spam filters. Due to the size of the dictionary, such tokenizers consume a substantial amount of disk space and memory that can make them impractical to download and operate on a personal computer. Therefore, known Japanese language tokenizers would be impractical for use in client-side anti-spam products.

Consequently, the state of the art is still inadequate when it comes to Japanese language spam filters.

SUMMARY OF THE INVENTION

Broadly, one aspect of this disclosure concerns a method for parsing Japanese language script. After receiving an electronic representation of a series of Japanese characters, the characters are reviewed to identify each

character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, or punctuation mark. Boundaries occur between characters of different predetermined types. Each group of one or more characters between adjacent boundaries is designated as a preliminary token. Predetermined context rules are applied to selectively redefine the preliminary tokens. An output of final tokens is provided, comprising the redefined preliminary tokens.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a block diagram of a computer that employs a non-dictionary based Japanese language tokenizer.

FIGURE 2 is a block diagram of a digital data processing apparatus.

FIGURE 3 shows an exemplary signal-bearing medium.

FIGURE 4 is a flowchart of an operational sequence of operations to parse Japanese language script.

DETAILED DESCRIPTION

The nature, objectives, and advantages of the invention will become more apparent to those skilled in the art after considering the following detailed description in connection with the accompanying drawings.

HARDWARE COMPONENTS & INTERCONNECTIONS

As mentioned above, one aspect of the present disclosure is a tokenizer that works to parse Japanese language script. Although such a tokenizer may be used for many different purposes, one of the chief implementations concerns an anti-spam filter, such as an e-mail filter. The subject tokenizer may be applied in a variety of environments, such as a personal computer, mainframe computer, computer workstation, computer network, personal data assistant (PDA), etc. The tokenizer may also be implemented by a centralized facility, such as an Internet service provider (ISP) or another online service provider. FIGURE 1 shows an example of the subject tokenizer in the environment 100 of simplified functional components of a personal computer 102.

Namely, the computer 102 includes an e-mail program 104 and a communications module 110. Although FIGURE 1 omits some components whose illustration is not necessary to explain the function of the tokenizer, ordinarily skilled artisans (having the benefit of this disclosure) will be familiar with the necessary components to properly implement a personal computer. The communications module 110 encodes digital data for transmission to the Internet 150 via the modem 112, and also decodes digital data arriving from the Internet via the modem 112. The modem 112 may comprise a device for wired or wireless communications, such as satellite, telephone line, DSL line, cable connection, public Internet, private Intranet, Ethernet, wide or local area network, or any other communications suitable for the purposes described herein. The modem 112 includes appropriate equipment to modulate/demodulate, encode/decode, multiplex, and perform other processing to connect many users to the Internet 150. Although the public

Internet 150 is shown as one example of a potential spam source, this disclosure contemplates other sources such as a private Internet (Intranet), local or wide area network, floppy diskette, CD-ROM, DVD, etc.

The e-mail program 104 includes a spam filter 106 and a tokenizer 108, among other components that will be apparent to the ordinarily skilled artisan. The spam filter 106 and tokenizer 108 may be part of the e-mail program 104 (as shown), or they may be programs that are separately loaded and installed onto the computer 102 and pre-designed to work with the e-mail program 104.

The tokenizer 108 comprises a software engine to parse Japanese script into appropriate words, as described in greater detail below. Unlike traditional tokenizers, the tokenizer 108 is not primarily dictionary based and therefore provides unforeseen efficiency in disk space and memory usage.

The spam filter 106 is a software program that receives output words from the tokenizer 108 and analyzes them for the presence of spam. The spam filter 106 may serve to block spam, alert the e-mail program of the spam, or perform another related task. Suitable spam filters may be found in various technical publications as well as various commercially available products. As a particular example, the spam filter 106 may comprise a Bayesian-rule mathematical processor, which analyzes the tokenizer 108's token-based representation of the script for the presence of spam.

Returning to the tokenizer 108, the illustrated example of this component utilizes an engine 108f along with several storage locations such a registers,

bytes, bits, addresses, extents, ranges, relational databases, hard disk drives, optical storage, or other appropriate storage sites/devices/media. The storage locations, in this example, include storage for tables 108a, temporary storage 108b, boundary list(s) 108c, preliminary tokens 108d, and final tokens 108e. The tokenizer 108 may be integrated with the spam filter 106 or a separate component.

Exemplary Digital Data Processing Apparatus

As mentioned above, the computer 102 or any of its functional components may be implemented in various forms. One example is a digital data processing apparatus, as exemplified by the hardware components and interconnections of the digital data processing apparatus 200 of FIGURE 2.

The apparatus 200 includes a processor 202, such as a microprocessor, personal computer, workstation, controller, microcontroller, state machine, or other processing machine, coupled to a storage 204. In the present example, the storage 204 includes a fast-access storage 206, as well as nonvolatile storage 208. The fast-access storage 206 may comprise random access memory ("RAM"), and may be used to store the programming instructions executed by the processor 202. The nonvolatile storage 208 may comprise, for example, battery backup RAM, EEPROM, flash PROM, one or more magnetic data storage disks such as a "hard drive", a tape drive, or any other suitable storage device. The apparatus 200 also includes an input/output 210, such as a line, bus, cable, electromagnetic link, or other means for the processor 202 to exchange data with other hardware external to the apparatus 200.

Despite the specific foregoing description, ordinarily skilled artisans (having the benefit of this disclosure) will recognize that the apparatus discussed above may be implemented in a machine of different construction, without departing from the scope of the invention. As a specific example, one of the components 206, 208 may be eliminated; furthermore, the storage 204, 206, and/or 208 may be provided on-board the processor 202, or even provided externally to the apparatus 200.

Logic Circuitry

In contrast to the digital data processing apparatus discussed above, a different embodiment of the present disclosure uses logic circuitry instead of computer-executed instructions to implement some or all of the processing entities in the system 100. Depending upon the particular requirements of the application in the areas of speed, expense, tooling costs, and the like, this logic may be implemented by constructing an application-specific integrated circuit (ASIC) having thousands of tiny integrated transistors. Such an ASIC may be implemented with CMOS, TTL, VLSI, or another suitable construction. Other alternatives include a digital signal processing chip (DSP), discrete circuitry (such as resistors, capacitors, diodes, inductors, and transistors), field programmable gate array (FPGA), programmable logic array (PLA), programmable logic device (PLD), and the like.

OPERATION

Having described the structural features of the present disclosure, the operational aspect of the disclosure will now be described.

Signal-Bearing Media

Wherever the functionality of a component of this disclosure is implemented using one or more machine-executed program sequences, these sequences may be embodied in various forms of signal-bearing media. In the context of FIGURE 2, such a signal-bearing media may comprise, for example, the storage 204 or another signal-bearing media, such as a magnetic data storage diskette 300 (FIGURE 3), directly or indirectly accessible by a processor 202. Whether contained in the storage 206, diskette 300, or elsewhere, the instructions may be stored on a variety of machine-readable data storage media. Some examples include direct access storage (e.g., a conventional "hard drive", redundant array of inexpensive disks ("RAID"), or another direct access storage device ("DASD")), serial-access storage such as magnetic or optical tape, electronic non-volatile memory (e.g., ROM, EPROM, flash PROM, or EEPROM), battery backup RAM, optical storage (e.g., CD-ROM, WORM, DVD, digital optical tape), paper "punch" cards, or other suitable signal-bearing media including analog or digital transmission media and analog and communication links and wireless communications. In an illustrative embodiment of the present disclosure, the machine-readable instructions may comprise software object code, compiled from a language such as assembly language, C, etc.

Logic Circuitry

In contrast to the signal-bearing medium discussed above, some or all of a component's functionality may be implemented using logic circuitry, instead of using a processor to execute instructions. Such logic circuitry is therefore

configured to perform operations to carry out the method of the present disclosure. The logic circuitry may be implemented using many different types of circuitry, as discussed above.

Operational Sequence

FIGURE 4 shows a sequence 400 of operations to parse and process Japanese "script," which may also be referred to as text, characters, writing, or another such term. For ease of illustration, without any intended limitation, the operations 400 are explained in the context of the system 100 described above, with particular reference to the tokenizer 108 and spam filter 106.

In step 402, the tokenizer 108 receives an electronic representation of an input series of Japanese characters, which may be represented in UNICODE or other digital data representation of each character. UNICODE is provides a global product similar to ASCII, with the added ability to represent most of major characters around the world. The series of characters arriving in step 402 may come from the spam filter 106, the e-mail program 104, communications module, or any another source. Next, in steps 404-416, the engine 108f processes the input characters to identify each character's type. As explained below, predefined types include Kanji, Hiragana, Katakana, Arabic numerals, Arabic letters, and punctuation marks.

More specifically, step 404 starts with the first character in the first sentence of the input series, and step 406 notes whether this character is a Kanji character, Hiragana character, Katakana character, Arabic numeral, Arabic letter, or punctuation mark. In one example, step 406 is performed by

consulting a table 108a or other data construct that maps between each predefined character type and its corresponding UNICODE values. Thus, in this embodiment, step 404 references each input character against the table 108a to identify the predefined type to which each character belongs. The engine 108f may store a record of the character type in a suitable location such as 108b.

After step 406, the engine 108f advances (step 408) to the next character in the input string, namely, the character that a reader would encounter next. Step 410 then determines the character's type in similar fashion as step 406. If the characters are of different types, the engine 108f notes (step 416) an imaginary boundary between the current character and the previous character, since they are of different types. The current boundary may be recorded, for example, in 108c. Step 416 also classifies all characters between the current boundary and the last boundary (or all characters if the current sentence does not have any boundaries yet) as being a preliminary "token." This preliminary token might represent a word, but further processing will be performed (as described below) to determine whether two or more preliminary tokens will actually form a final token. Also in step 416, the preliminary token may be recorded in 108d, for example. After step 416, the routine proceeds to consider the next character in step 408.

In contrast to the preceding description, if step 410 finds that the current and previous characters are the same type, step 410 returns to step 408. Step 416 is not performed in this case, since no boundary has occurred.

On the other hand, if step 410 finds that the current character is an appropriate punctuation mark, then the end of the current sentence has been reached. Having finished analyzing the individual characters in the current sentence, identified boundaries, and designated preliminary tokens where applicable, step 420 analyzes the context of the current sentence. Namely, step 420 applies predetermined context rules to identify and redefine the extent of selected preliminary tokens. As a more particular example, step 420 first identifies any preliminary tokens that meet certain criteria, and then redefines the identified preliminary tokens by combining or eliminating them. The redefined tokens, along with the remaining untouched tokens, constitute "final" tokens, which may be stored in 108e. The following are some examples of context rules.

Rule 1. Two adjacent preliminary tokens will be redefined as one token when the first is a numerical string and the second is a Kanji string. As an additional context rule, or as a more detailed implementation of Rule 1, two adjacent preliminary tokens will be consolidated when the first is a so-called "full width" digit and the second is a Kanji character. In UNICODE, Arabic digits are represented by ASCII digits as well as full width digits. Full width digits include UNICODE code points between 0xff10 and 0xff19.

Rule 2. Any preliminary token that only contains only or more Hiragana characters will be ignored, or in other words, removed from designation as a preliminary token (and deleted from 108d). These are most likely a part of verb, adjective or adverb used with nearby Kanji characters that provide the

more meaningful content. Under this rule, then, preliminary tokens of Hiragana-only characters are ignored.

Rule 3. "Isolated" tokens are ignored. This rule ignores (i.e., disqualifies from preliminary token status) any token that differs in character type from its preceding and trailing neighbors. The isolated token is likely to be a common word, or a part of verb, adjective, adverb, punctuation mark, and the like, which is therefore less likely meaningful for spam recognition. Instead of this broad rule, one or more detailed rules may be used, such as the following: only isolated Kanji tokens are ignored. This rule is helpful since most Japanese Kanji words utilize two or more characters.

The foregoing list of context rules is merely exemplary, and definitely not exclusive. With the benefit of this disclosure, ordinarily skilled artisans may develop, refine, or add context rules. At any rate, step 420 serves to redefine the list 108d of preliminary tokens, with the redefined tokens stored in 108d as a set of "final tokens. The redefined tokens include the tokens consolidated by the context rules, and those tokens left untouched by the context rules.

After step 420 applies the context rules, step 424 asks whether there any more sentences remaining to process. This includes any remaining sentences in the original input script, as well as any script that has newly arrived during ongoing performance of the routine 400. If any sentences remain to be processed, the routine 400 returns to step 408, which focuses on the next character of the next sentence.

Otherwise, if there are no more sentences to process, then 424 advances to step 426, which outputs a list of final tokens. The final tokens include the preliminary tokens as redefined by step 420. Thus, the list includes all preliminary tokens that were consolidated in steps 420, and all untouched preliminary tokens, but none of the preliminary tokens that were eliminated.

After step 426, downstream processing is performed in step 428. For example, this step may involve the tokenizer outputting the final tokens to the spam filter 106, and the spam filter applying Bayesian or other applicable processing rules to the final tokens in order to analyze them for spam.

As an alternative to the foregoing description, steps 426, 428 may be initiated after each sentence is processed (i.e., between steps 420, 424), in order to provide the spam filter 106 with a more continuous stream of output rather than waiting until step 424 concludes processing of all sentences in the script.

OTHER EMBODIMENTS

While the foregoing disclosure shows a number of illustrative embodiments of the invention, it will be apparent to those skilled in the art that various changes and modifications can be made herein without departing from the scope of the invention as defined by the appended claims. Furthermore, although elements of the invention may be described or claimed in the singular, the plural is contemplated unless limitation to the singular is explicitly stated. Additionally, ordinarily skilled artisans will recognize that operational sequences must be set forth in some specific order for the purpose of

explanation and claiming, but the present invention contemplates various changes beyond such specific order.

In addition, those of ordinary skill in the relevant art will understand that information and signals may be represented using a variety of different technologies and techniques. For example, any data, instructions, commands, information, signals, bits, symbols, and chips referenced herein may be represented by voltages, currents, electromagnetic waves, magnetic fields or particles, optical fields or particles, other items, or a combination of the foregoing.

Moreover, ordinarily skilled artisans will appreciate that any illustrative logical blocks, modules, circuits, and process steps described herein may be implemented as electronic hardware, computer software, or combinations of both. To illustrate one exemplary embodiment, various functional aspects of the invention have been described in terns of illustrative components, blocks, modules, circuit, and steps. Whether such functionality is implemented as hardware, software, or both depends upon the particular application and design constraints imposed on the overall system. Skilled artisans may implement the described functionality in varying ways for each particular application without such implementation decisions departing from the scope of the present invention.

CLAIMS

1. A method for parsing Japanese language script, comprising operations of:

receiving an electronic representation of a series of Japanese characters;

reviewing the series of characters to identify each character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, punctuation mark;

recognizing boundaries between adjacent characters of different predetermined types, and designating each group of one or more characters between adjacent ones of the boundaries as being a preliminary token;

applying predetermined context rules to selectively redefine the preliminary tokens;

providing an output of final tokens comprising the redefined preliminary tokens.

2. The method of claim 1, the operation of applying the predetermined context rules comprising:

separately applying the context rules to each sentence of the series.

3. The method of claim 1, the operation of selectively redefining the preliminary tokens comprising:

responsive to presence of any preliminary tokens whose characters meet a first predetermined criteria, withdrawing preliminary token designation for each of the identified preliminary tokens;

responsive to presence of any group of two or more adjacent preliminary tokens meeting second predetermined criteria, redefining the preliminary tokens of each said group to form a single corresponding preliminary token.

4. The method of claim 3, the operation of providing an output of final tokens comprising:

providing an output of tokens comprising all consolidated preliminary tokens and all remaining preliminary tokens that have remained free of withdrawal and consolidation.

5. The method of claim 1, the operations further comprising:

providing the final tokens as input to a Bayesian rule based spam filter.

6. The method of claim 1, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of adjacent preliminary tokens representing one or more Arabic numbers followed by one or more Kanji characters, redefining the identified preliminary tokens as being a single preliminary token.

7. The method of claim 1, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of adjacent preliminary tokens representing a full-width UNICODE digit followed by a Kanji string, redefining the identified preliminary tokens as being a single preliminary token.

8. The method of claim 1, where the operation of selectively redefining the preliminary tokens comprises:

withdrawing preliminary token status of each preliminary token made up of one or more Hiragana characters only.

9. The method of claim 1, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of any preliminary token differing in character type from both preceding and following preliminary tokens, withdrawing preliminary token status of the identified preliminary token.

10. The method of claim 1, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of any preliminary token representing a single Kanji character, withdrawing preliminary token status of the identified preliminary token.

11. A signal-bearing medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform operations to parse Japanese language script, the operations comprising:

receiving an electronic representation of a series of Japanese characters;

reviewing the series of characters to identify each character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, punctuation mark;

recognizing boundaries between adjacent characters of different predetermined types, and designating each group of one or more characters between adjacent ones of the boundaries as being a preliminary token;

applying predetermined context rules to selectively redefine the preliminary tokens;

providing an output of final tokens comprising the redefined preliminary tokens.

12. The medium of claim 11, the operation of applying the predetermined context rules comprising:

separately applying the context rules to each sentence of the series.

13. The medium of claim 11, the operation of selectively redefining the preliminary tokens comprising:

responsive to presence of any preliminary tokens whose characters meet a first predetermined criteria, withdrawing preliminary token designation for each of the identified preliminary tokens;

responsive to presence of any group of two or more adjacent preliminary tokens meeting second predetermined criteria,

consolidating the preliminary tokens of each said group to form a single corresponding preliminary token.

14. The method of claim 13, the operation of providing an output of final tokens comprising:

providing an output of tokens comprising all consolidated preliminary tokens and all remaining preliminary tokens that have remained free of withdrawal and consolidation.

15. The medium of claim 11, the operations further comprising:

providing the final tokens as input to a Bayesian rule based spam filter.

16. The medium of claim 11, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of adjacent preliminary tokens representing one or more Arabic numbers followed by one or more Kanji characters, redefining the identified preliminary tokens as being a single preliminary token.

17. The medium of claim 11, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of adjacent preliminary tokens representing a full-width UNICODE digit followed by a Kanji string, redefining the identified preliminary tokens as being a single preliminary token.

18. The medium of claim 11, where the operation of selectively redefining the preliminary tokens comprises:

withdrawing preliminary token status of each preliminary token made up of one or more Hiragana characters only.

19. The medium of claim 11, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of any preliminary token differing in character type from both preceding and following preliminary tokens, withdrawing preliminary token status of the identified preliminary token.

20. The medium of claim 11, where the operation of selectively redefining the preliminary tokens comprises:

responsive to presence of any preliminary token representing a single Kanji character, withdrawing preliminary token status of the identified preliminary token.

21. A logic circuit of multiple interconnected electrically conductive elements configured to perform operations to parse Japanese language script, the operations comprising:

receiving an electronic representation of a series of Japanese characters;

reviewing the series of characters to identify each character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, punctuation mark;

recognizing boundaries between adjacent characters of different predetermined types, and designating each group of one or more characters between adjacent ones of the boundaries as being a preliminary token;

applying predetermined context rules to selectively redefine the preliminary tokens;

providing an output of final tokens comprising the redefined preliminary tokens.

22. A computer with Japanese language spam protection, comprising:

nonvolatile digital data storage;

volatile digital data storage;

digital data input/output;

a digital data processor, coupled to the nonvolatile programmed to perform operations comprising:

receiving an electronic representation of a series of Japanese characters;

reviewing the series of characters to identify each character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, punctuation mark;

recognizing boundaries between adjacent characters of different predetermined types, and designating each group of one

or more characters between adjacent ones of the boundaries as being a preliminary token;

applying predetermined context rules to selectively redefine the preliminary tokens;

providing an output of final tokens comprising the redefined preliminary tokens.

23. A computer with Japanese language spam protection, comprising:

means for nonvolatile digital data;

means for volatile digital data;

means for digital data input/output;

means for processing digital data by performing operations comprising:

receiving an electronic representation of a series of Japanese characters;

reviewing the series of characters to identify each character as being one of the following predefined types: Kanji, Hiragana, Katakana, Arabic numeral, Arabic letter, punctuation mark;

recognizing boundaries between adjacent characters of different predetermined types, and designating each group of one or more characters between adjacent ones of the boundaries as being a preliminary token;

applying predetermined context rules to selectively redefine the preliminary tokens;

providing an output of final tokens, which comprise the redefined preliminary tokens.

FIG. 1

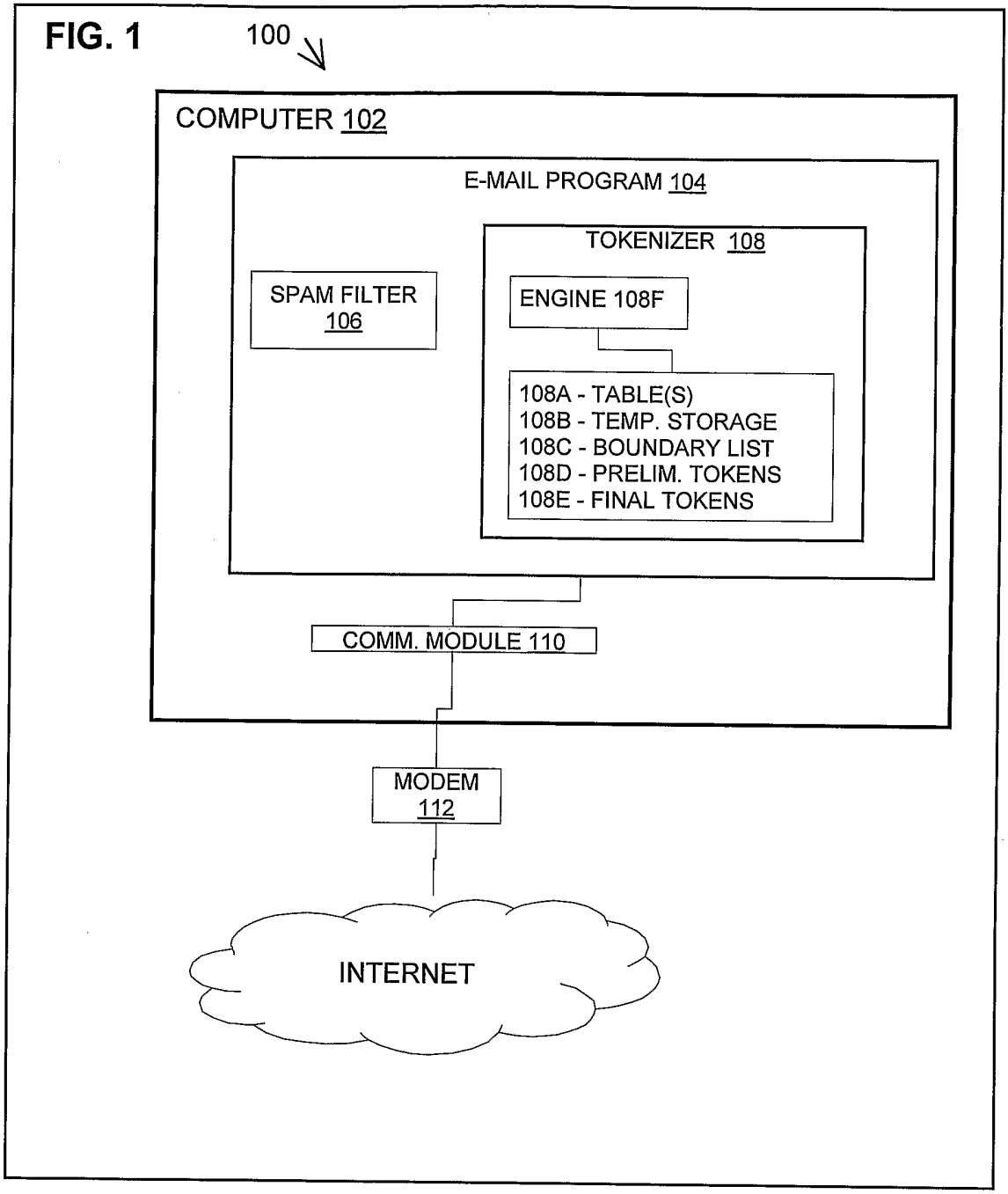
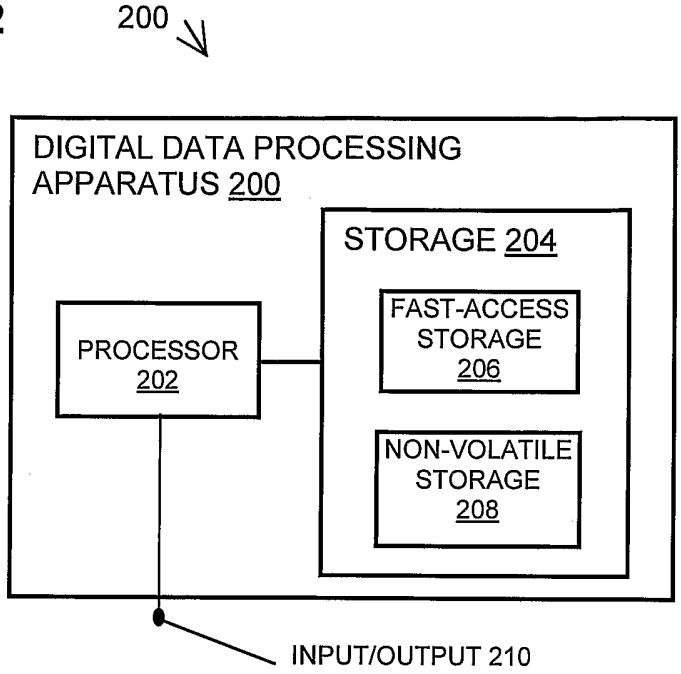

100
COMPUTER 102
E-MAIL PROGRAM 104
TOKENIZER 108
SPAM FILTER 106
ENGINE 108F
108A - TABLE(S)
108B - TEMP. STORAGE
108C - BOUNDARY LIST
108D - PRELIM. TOKENS
108E - FINAL TOKENS
COMM. MODULE 110
MODEM 112
INTERNET

FIG. 2


200
DIGITAL DATA PROCESSING APPARATUS 200
STORAGE 204
FAST-ACCESS STORAGE 206
PROCESSOR 202
NON-VOLATILE STORAGE 208
INPUT/OUTPUT 210

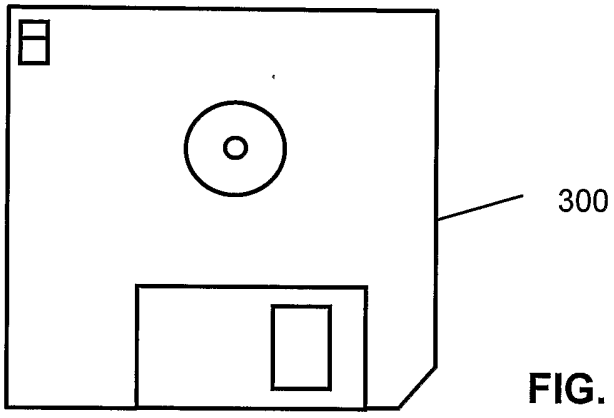

300

FIG. 3

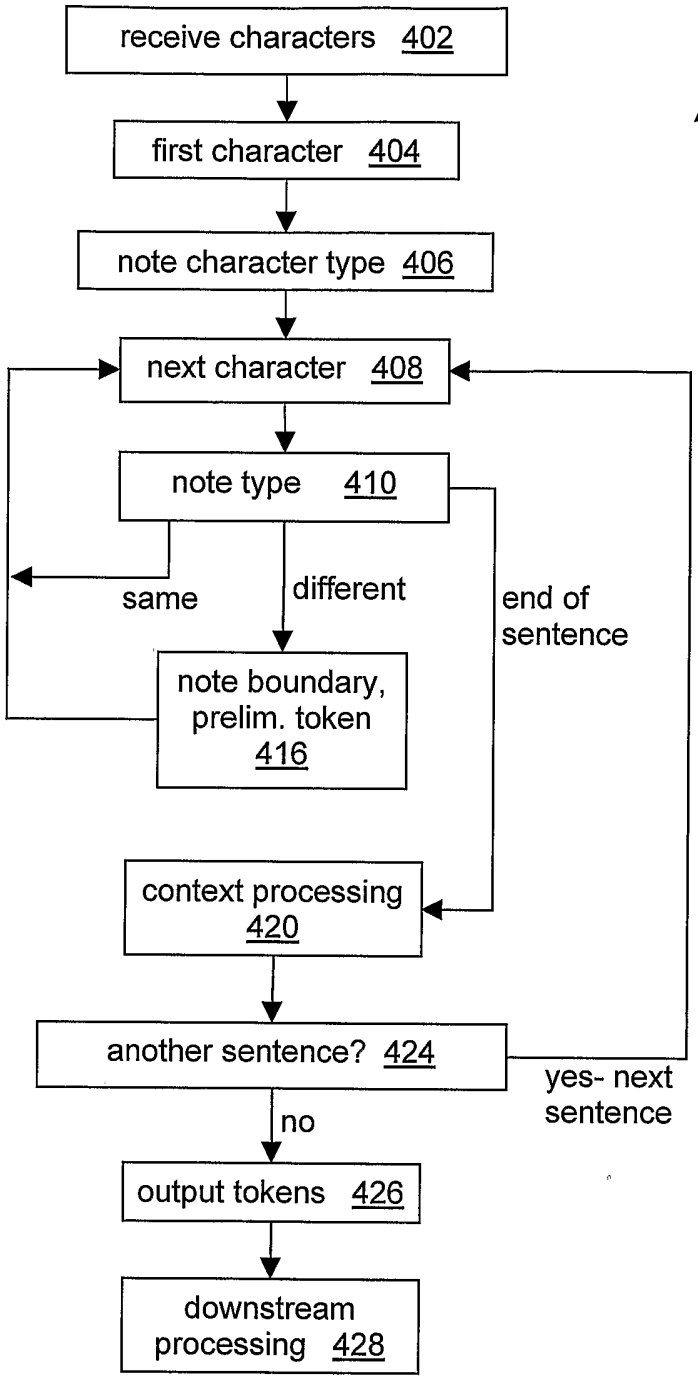

receive characters 402
400
first character 404
note character type 406
next character 408
note type 410
same
different
end of sentence
note boundary, prelim. token 416
context processing 420
another sentence? 424
yes- next sentence
no
output tokens 426
downstream processing 428

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US04/14313

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : G06F 17/28, 17/27
US CL : 704/2, 8
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)
U.S. : 704/2, 8

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
704/4,6

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EAST search (USPAT, US-PGPUB, EPO, JPO, DERWENT, IBM_TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 6,470,306 B1 (PRINGLE et al.) 22 October 2002 (22.10.2002), column 6, lines 43-54, column 10, lines 49-61, column 11, lines 10 and 48-55, column 12, lines 1-5, and 49-60.	1-23
Y	US 5,432,948 A (DAVIS et al.) 11 July 1995 (11.07.1995), column 3, lines 36-42, and column 4, lines 27-32.	1-23
Y	Graham, Paul. Better Bayesian Filtering, 2003 Spam Conference, January 2003.	5,15,22,23

☐ Further documents are listed in the continuation of Box C. ☐ 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 application or patent 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
03 September 2004 (03.09.2004)

Date of mailing of the international search report
29 SEP 2004

Name and mailing address of the ISA/US
Mail Stop PCT, Attn: ISA/US
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
Facsimile No. (703) 305-3230

Authorized officer
Talivaldis Smits
Telephone No. (703)306-3011

Form PCT/ISA/210 (second sheet) (January 2004)