(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 214 643 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
24.11.2010 Bulletin 2010/47

(21) Application number: **00923179.6**

(22) Date of filing: **06.04.2000**

(51) Int Cl.:
G06F 17/27 *(2006.01)* *G06F 17/30* *(2006.01)*

(86) International application number:
PCT/US2000/009357

(87) International publication number:
WO 2000/062155 (19.10.2000 Gazette 2000/42)

(54) SYSTEM AND METHOD FOR PARSING A DOCUMENT

SYSTEM UND VERFAHREN ZUM PARSEN EINES DOKUMENTES

SYSTEME ET PROCEDE SERVANT A ANALYSER UN DOCUMENT

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE

(30) Priority: **09.04.1999 US 288994**

(43) Date of publication of application:
19.06.2002 Bulletin 2002/25

(73) Proprietor: **Entrieva, Inc.**
Reston, VA 20191 (US)

(72) Inventor: **VOGEL, Claude**
Foster City, CA 94404 (US)

(74) Representative: **HOFFMANN EITLE**
Patent- und Rechtsanwälte
Arabellastrasse 4
81925 München (DE)

(56) References cited:
WO-A-98/38560 **US-A- 5 371 807**
US-A- 5 819 260 **US-A- 6 014 680**
US-A- 6 070 158

- **IAN H. WITTEN, GORDON W. PAYNTER, EIBE FRANK, CARL GUTWIN, CRAIG G. NEVILL-MANNING: "KEA: Practical Automatic Keyphrase Extraction" ARXIV.ORG, [Online] 5 February 1999 (1999-02-05), XP002511489 Cornell University Library, Ithaca, NY, USA Retrieved from the Internet: URL:http://arxiv.org/abs/cs/9902007v1> [retrieved on 2009-01-22]**
- **P. TURNEY: "Learning to extract key phrases from text" NRC-CNRC NATIONAL RESEARCH COUNCIL OF CANADA, [Online] vol. NRC/ERB-1057, 12 February 1999 (1999-02-12), XP002511490 Retrieved from the Internet: URL: http://www.iit-iti.nrc-cnrc.gc.ca/iit- publications-iti/docs/NRC-41622.pdf> [retrieved on 2009-01-21]**

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

EP 1 214 643 B1

Description

Background of the Invention

[0001] This invention relates generally to a system and method for processing a document and in particular to a system and method for identifying a plurality of phrases within the document which indicate the context of the document.

[0002] Various factors have contributed to the extensive storage and retrieval of textual data information using computer databases. A dramatic increase in the storage capacity of hard drives coupled with a decrease in the cost of computer hard drives, and increases in the transmission speed of computer communications have been factors. In addition, the increased processing speed of computers and the expansion of computer communications networks, such as a bulletin board or the Internet, have been factors. People therefore have access to the large amounts of textual data stored in these databases. However, although the technology facilitates the storage of and the access to the large amounts of textual data, there are new problems that have been created by the large amount of textual data that is now available.

[0003] In particular, a person trying to access textual data in a computer database having a large amount of data needs a system for analyzing the data in order to retrieve the desired information quickly and efficiently without retrieving extraneous information. In addition, the user of the system needs an efficient system for condensing each large document into a plurality of phrases (one or more words) which characterize the document so that the user of the system can understand the document without actually viewing the entire document. A system for condensing each document into a plurality of key phrases is known as a parsing system or a parser.

[0004] In one typical parser, the parser attempts to identify phrases which are repeated often within the document and identifies those phrases as being key phrases which characterize the document. The problem with such a system is that it is very slow since it must count the repetitions of each phrase in the document. It also requires a large amount of memory. As the amount of data to be parsed increases, the slow speed of this parser becomes unacceptable. Another typical parser performs a three step process to identify the key phrases. First, each word in the document is assigned a tag based on the part of speech of the word (i.e., noun, adjective, adverb, verb, etc.) and certain parts of speech, such as an article or an adjective, may be removed from the list of phrases which characterizes the document. Next, one or more sequences of words (templates) may be used to identify and remove phrases which do not add any understanding to the document. Finally, any phrase which is an appropriate part of speech and does not fall within one of the templates is accepted as a key phrase which characterizes the document. This conventional parser, however, is also slow which is unacceptable as the amount of data to be parsed increases.

[0005] In all of these conventional parser systems, the parser attempts to break the document down into smaller pieces based on the characteristics (frequency of repetition or part of speech) of the particular words in the document. The problem is that language generally is not that easily classified and therefore the conventional parser does not accurately parse the document or requires a large amount of time to parse the document. In addition, the conventional parser systems are very slow because they all attempt to use complex characteristics of the language as a method for parsing the key phrases out of the document. These problems with conventional parsers becomes more severe as the number of documents which must be parsed increases. Today, the number of documents which must be parsed is steadily increasing at a tremendous rate due to, among other things, the Internet and the World Wide Web. Therefore, these conventional parsers are not acceptable. Thus, it is desirable to provide a parsing system and method which solves the above problems and limitations with conventional parsing systems and it is to this end that the present invention is directed.

[0006] US 5,819,260 relates to a phrase recognition method based on breaking streams of text into text chunks and selecting phrases from the chunks according to frequency based definitions.

"KEA: Practical Automatic Keyphrase Extraction", by H. Witten et al., provides a prior art algorithm for identifying keyphrases using lexical method and comprises two stages of training and extraction.

"Learning to Extract Keyphrases from Text", by P. Turkey, discloses an algorithm named GenEx based on the combination of a Genitor steady-state genetic algorithm and an Extractor parameterized key phrase extraction algorithm according to the prior art.

The above prior art techniques also suffer from the problems and disadvantages above described.

Summary of the Invention

[0007] The object is achieved by the subject-matter of the independent claims. Advantageous embodiments are defined in the dependent claims. Further examples are provided for facilitating the understanding of the invention.

[0008] A parser system and method in accordance with an example is provided in which the break characters within a sentence or a paragraph are used to parse the document into a plurality of key phrases. The parser system in accordance with the invention is very fast and does not sacrifice much accuracy for the speed. The break characters within the document may include punctuation marks, certain stop words and certain types of words such as verbs and articles. The parser system may include a buffer which receives one or more words before it receives a break character. When the buffer receives

a break character, the parser may determine whether the phrase before the break character is saved based on the type of break character. In particular, if the break character is a punctuation mark, the parser may keep the one or more words before the break character as a key phrase. If the break character is another type of character, the phrase before the break character may or may not be saved. Once the fate of the phrase has been determined, the buffer is flushed and the next sequence of one or more words is read into the buffer so that it may also be parsed. In this manner, a plurality of phrases in the document may be rapidly extracted from the document based on the break characters within the sentences and paragraphs of the document.

[0009] Thus, in accordance with an example, a system for parsing a piece of text into one or more phrases which characterize the document is provided. The system comprises a buffer for reading one or more words from the piece of text into the buffer and a parser for identifying a phrase contained in the buffer, the phrase being a sequence of two or more words in between break characters. The parser further comprises means for determining the type of break character that follows the identified phrase and means for saving a key phrase from the buffer based on the determined type of break character. The key phrases are stored in a database.

Brief Description of the Drawings

[0010]

Figure 1 is a block diagram of a text processing system;

Figure 2 is a block diagram of a parsing system in accordance with an arrangement of the invention;

Figure 3 is a flowchart illustrating a method for parsing a document in accordance with an arrangement of the invention;

Figure 4 is an example of a document to be parsed by the parsing system in accordance with an arrangement of the invention; and

Figures 5A - 5L are diagrams illustrating the operation of the parsing buffer in accordance with an arrangement of the invention on the document shown in Figure 4.

Detailed Description of a Preferred Embodiment

[0011] The invention is particularly applicable to a system for parsing English language documents and it is in this context that the invention will be described. It will be appreciated, however, that the system and method in accordance with the invention has greater utility, such as to other languages and to various different pieces of textual data. To better understand the invention, a text processing system will now be described.

[0012] Figure 1 is a block diagram of a text processing system 10. The text processing system 10 may include a parser system 12, a clusterizer 14, a map generator 16 and a database (DB) 18. The text processing system may receive one or more pieces of text , such as stories, press releases or documents, and may generate a map graphically showing the relationships between the key phrases in the document. Each piece of text may be received by the parser system 12 which processes each piece of incoming text and generates one or more key phrases for each piece of text which characterizes the piece of text. The key phrases may be stored in the database 18. The details about the parser system will be described below with reference to Figures 2- 5. Once the key phrases are extracted from each piece of text, the clusterizer 14 may generate one or more clusters of the key phrases based on the relationships between the phrases. The clusters generated may also be stored in the database 18. The map generator 16 may use the generated clusters for the pieces of text in the database in order to generate a graphical map showing the relationships of the key phrases within the various pieces of text in the database to each other so that a user of the system may easily search through the database by viewing the key phrases of the pieces of text. More details about the clusterizer and map generator are disclosed in WO 98/38560 which is owned by the assignee of the present invention. The text processing system may be implemented in a variety of manners including a client/server type computer system in which the client computers access the server via a public computer network, such as the Internet. The parser, the clusterizer and the map generator may be software applications being executed by a central processing unit (not shown) of the text processing system 10. Now, the parser system 12 in accordance with the invention will be described in more detail.

[0013] Figure 2 is a block diagram of the parsing system 12 in accordance with an arrangement of the invention. The parsing system 12 may include a buffer 20, a parser 22 and a rules database (rules DB) 24. The buffer may store one or more words of the incoming piece of text, which may be a document, which are analyzed by the parser 22 using the rules contained in the rules DB 24. The output of the parser system 12 is one or more phrases (each phrase containing one or more words) which characterize the document being parsed. In particular, the parser may separate phrases in the document based on break characters within the document in accordance with the invention. In more detail, one or more words may be read into the buffer from the document until a break character is identified. Thus, the parser system 12 identifies phrases which are between break characters. Then, based on the type of break character, the phrase may be saved as a key phrase or deleted. The parser system 12, for example, may be implemented as one or more pieces of software being executed by a mi-

croprocessor (not shown) of a server computer which may be accessed by a plurality of client computers over a computer network, such as the Internet, a local area network or a wide area network. The parser 22 advantageously rapidly extracts key phrases from a piece of text using break characters. The break characters in accordance with the invention will now be described. The break characters may include an explicit break, such as a punctuation mark, numbers, words containing numbers, and stop words. The stop words may be further classified as soft stop words or a hard stop words. Each of these different break characters will now be described. The explicit break characters may include various punctuation symbols, such as a period, a comma, a semicolon, a colon, an exclamation point, right or left parenthesis, left or right square brackets, left or right curly braces, a return character or a line feed character. The stop characters may be a generated list or it may include a slash (/) and an ampersand symbol (@). A separator may be defined as digits, letters, foreign characters, break characters, apostrophes, dashes and other stop characters. The various words in a piece of text may be categorized as articles, connectors, hard and soft stop characters, linguistic indicators, a syntactic categories such as nouns, verbs, irregular verbs, adjectives and adverbs.

[0014] In parsing the characters in the piece of text, separators may always be added to a phrase. A apostrophe or dash at the beginning of a word is treated as a break character (see below), an apostrophe or dash at the end of a word is also treated as a break character and a word with an apostrophe or dash in the middle of the word is added to the phrase in the buffer. All stop characters and breaks are treated as stop characters and breaks as described below. At the word level of parsing, proper nouns are retained by testing for an upper case letter at the first character of the word. In addition, all words with only upper case letters and numeric words are kept in the buffer. Optionally, a numeric string may be classified and treated as a stop character. The following are mandatory word level parsing rules. First, the word following as possessive "s" may be deleted. For example, as the sentence "The cat's paw is wet." is parsed in accordance with the invention, "the" is deleted and "cat" is put into the buffer and then deleted when the break character (the apostrophe) is detected. The apostrophe is deleted because it is punctuation and then the next character to parse is the possessive "s" after the apostrophe which is deleted along with the word "paw" since it follows the possessive "s". Connector words appearing at the beginning of a phrase are also deleted although a connector word followed by "The" is kept in the buffer. For a hard stop character, the last phrase connected to the hard stop character is deleted and the remaining buffer is processed. A soft stop character may be treated as a break character. A repeated character is treated as a stop character.

[0015] To further remove unwanted words for parsing, some optional phrase level parsing rules may be used. In particular, phrases longer than a predetermined length, such as six words, may be deleted, a phrase with all upper case words may be deleted and a phrase with all numeric words may be deleted. All of the above parsing rules may be stored in the parsing rules database 24 shown in Figure 2. Now, the details of the parser system 12 will now be described with reference to Figure 3.

[0016] Figure 3 is a flowchart illustrating a method 40 for parsing a document in accordance with an arrangement of the invention. The method begins as a first word of the document is loaded into the buffer from a document database or a memory of the server in step 42. Next, the parser determines if the word is a break character in step 44. The parser may also delete certain characters or words at this stage of the parsing process. If the word is not a break character, the method loops back to step 42 and the next word of the document is read into the buffer. This process of reading a word into the buffer is repeated until a break character is encountered so that the buffer contains a sequence of words (a phrase) which has a break character before the sequence of words and a break character after the sequence of words. In this manner, the document is parsed into phrases which are separated from one another by break characters.

[0017] If a break character is encountered, the parser may determine if the break character is an explicit break character in step 46, delete the break character and extract the phrase contained in the buffer if an explicit break character exists in step 48. The phrase extracted from the buffer may be stored in a database for future use. Next, in step 50, the buffer may be flushed to empty the words from the buffer and the buffer may begin loading new words into the buffer in steps 42 and 44 until another break character is identified. Returning to step 46, if the break character is not an explicit break character, the parser determines if the break character is a soft stop word in step 52. If the break character is a soft stop word, then the soft stop word is deleted and the phrase in the buffer is stored in the database in step 54, the buffer is flushed in step 50 and the buffer is refilled with new words from the document. If the break character is not a soft stop word (i.e., the break character is a hard stop word), the hard stop word and the phrase in the buffer are deleted in step 56, the buffer is flushed in step 50 and refilled with new words from the document in steps 42 and 44. In this manner, phrases from the document are extracted in accordance with the invention using the break characters and the type of break character to separate the phrases from each other and determine which phrases are going to be saved in the database. The parser in accordance with an arrangement of the invention does not attempt to analyze each word of the document to identify key phrases as with conventional systems, but does extract phrases from the document more quickly than conventional parsers and with as much accuracy as the conventional parsers. Now, an example of the operation of the parser in accordance with an arrangement of the invention will be described with reference to Fig-

ures 4 and 5A - 5L.

[0018] Figure 4 is an example of a document 60 to be parsed by the parsing system in accordance with an arrangement of the invention while Figures 5A - 5L illustrate the operation of the buffer during the parsing of the document 60 shown in Figure 4. In this example, the document is a short electronic news story, but the parser may also extract phrases from any other piece of text. In fact, the parser in accordance with an arrangement of the invention may be able to extract phrases from various types of documents at speeds of up to 1 MByte of data per second. The particular story shown describes a new "snake-like" robot developed by NEC. Figures 5A - 5L illustrate, in a table 68, the operation of the buffer in accordance with an arrangement of the invention on the above story. In particular, a first column 70 of the table contains the current word being read into the buffer, a second column 72 contains the determination of the type of word by the parser in accordance with an arrangement of the invention, a third column 74 contains the contents of the buffer at the particular time, a fourth column 76 contains the word index (i.e., the phrases which are being extracted from the document) and a fifth column 78 contains comments about the parsing process.

[0019] As shown in Figure 5A, the first word read into the buffer is a sequence of asterisks at the beginning of the story which are classified by the parser as a break word (punctuation) and deleted from the buffer. The next word is "Computer" which is entered into the buffer since it is not a break word and the next word, which is "Select" is also entered into the buffer since it is also not a break word. Thus, the buffer contains the phrase "Computer Select" as shown in a cell 80. The next word in the document is a comma which is classified as a break character by the parser. Because the break character is punctuation (an explicit break), the words in the buffer are saved in the database as shown in the Word Index column 76 and the buffer is flushed. Now, new words are read into the buffer and parsed. The next word into the buffer is "October" which is a hard stop word because it relates to a date and it is deleted. The next word received by the buffer is "1995" which is a break character since it is a number and it is also deleted. The next word received by the buffer is "COPYRIGHT" which is identified as a stop word because it is all capital letters and it is deleted. The next word is "Newsbytes" which is not a break character and is therefore stored into the buffer. The next word is "Inc." which is also stored in the buffer. The next word is a period which is a break character so that the buffer contents "Newsbytes Inc." are saved into a database as shown in the Word Index column, the break character is deleted and the buffer is flushed.

[0020] The next two word received by the buffer, which are "1995" and a sequence of asterisks, are both break words which are deleted. The next two words received by the buffer are "Newsbytes" and "Newsbytes" which are both stored in the buffer. The next word received is "August" which is a hard stop word so that the contents of the buffer and the hard stop word is deleted. The next three words received by the buffer are all break characters (i.e., numbers or punctuation) which are deleted. The next word is a word containing a number in a cell 82 which is stored in the buffer, but then deleted when the next character is a break character because the buffer only contains a single word. As can be seen in Figures 5B - 5L, the parsing process continues for the entire document so that a list of key phrases, as shown in the Word Index column 76, are extracted from the document and saved in a database.

[0021] In summary, phrases which characterize the document or piece of text may be rapidly extracted from the document in accordance with the invention. The invention uses the break characters in the document or the piece of text to separate the phrases from each other and to extract the key phrases for a document. In the example above, the extracted phrases, such as "Newsbytes Inc.", "snake-like robot", "NEC Corporation", "robotically controlled electronic snake", "disaster relief work" and "world's first active universal joint" permit a person reviewing only the key phrases to understand the context of the document without reviewing the entire document. The parsing system in accordance with an arrangement of the invention performs the extraction of the key phrases more rapidly than any other conventional parsing systems which is important as the total amount of textual data and documents available for parsing increases at an exponential rate due, in part, to the explosion of the user of the Internet.

[0022] While the foregoing has been with reference to a particular arrangement of the invention, it will be appreciated by those skilled in the art that changes in this arrangement may be made without departing from the principles of the invention, the scope of which is defined by the appended claims.

Claims

1. A system (10) for parsing a piece of text into one or more phrases which characterize the document, the system (10) comprising:

   a buffer (20) for reading (42, 44) one or more words from the piece of text into the buffer (20) until a break character is identified;
   a parser (21) for identifying (44) a phrase contained in the buffer, the phrase being a sequence of two or more words in between break characters;
   the parser (21) further comprising means for determining (46, 52) the type of break character that follows the identified phrase and means for saving (48, 54, 50) the identified phrase as a key phrase from the buffer (20) or deleting the identified phrase based on the determined type of break character;

a database (18) for storing the key phrases.

2. The system (10) of Claim 1, wherein the buffer (20) further comprises means for flushing the buffer (20) when the key phrase is stored in the database (18) or the phrase in the buffer (20) is deleted.

3. A method for parsing a piece of text into one or more phrases which characterize the document, the method comprising:

reading (42, 44) one or more words from the piece of text into a buffer (20) until a break character is identified;
identifying (44) a phrase contained in the buffer, the phrase being a sequence of two or more words in between break characters;
determining (46, 52) the type of break character that follows the identified phrase; and
saving (48, 54, 50) the identified phrase as a key phrase from the buffer (20) into a database (18) or deleting (56) the identified phrase based on the determined type of break character.

4. The method of Claim 3 further comprising flushing the buffer (20) when the key phrase is stored in the database (18) or the phrase in the buffer (20) is deleted.

Patentansprüche

1. System (10) zum Gliedern eines Textstücks in einen oder mehrere Sätze, welche das Dokument kennzeichnen, das System (10) umfassend:

einen Puffer (20) zum Lesen (42, 44) eines oder mehrerer Wörter von dem Teststück in den Puffer (20) bis ein Abbruchzeichen identifiziert ist;
einen Gliederer (21) zum Identifizieren (44) eines Satzes, der in dem Puffer beinhaltet ist, wobei der Satz eine Sequenz von zwei oder mehr Wörtern zwischen Abbruchzeichen ist;
wobei der Gliederer (21) ferner umfasst Mittel zum Bestimmen (46, 52) des Typs von Abbruchzeichen, welches dem identifizierten Satz folgt und Mittel zum Speichern (48, 54, 50) des identifizierten Satzes als ein Schlüsselsatz von dem Speicher (20) oder Löschen des identifizierten Satzes basierend auf dem bestimmten Typ von Abbruchzeichen;
eine Datenbank (80) zum Speichern der Schlüsselsätze.

2. System (10) nach Anspruch 1, wobei der Puffer (20) ferner umfasst Mittel zum Leeren des Speichers (20), wenn der Schlüsselsatz in der Datenbank (80) gespeichert ist oder der Satz in dem Puffer (20) gelöscht ist.

3. Verfahren zum Gliedern eines Textstücks in einen oder mehrere Sätze, welche das Dokument kennzeichnen, das Verfahren umfassend:

Lesen (42, 44) eines oder mehrerer Worte von dem Textstück in einen Puffer (20) bis ein Abbruchzeichen identifiziert wird;
Identifizieren (44) eines Satzes beinhaltet in dem Speicher, wobei der Satz eine Sequenz von zwei oder mehr Wörtern zwischen Abbruchzeichen ist;
Bestimmen (46, 52) des Typs von Abbruchzeichen, der dem identifizierten Satz folgt; und
Speichern (48, 54, 50) des identifizierten Satzes als ein Schlüsselsatz von dem Puffer (20) in eine Datenbank (18) oder Löschen (56) des identifizierten Satzes basierend auf dem bestimmten Typ von Abbruchzeichen.

4. Verfahren nach Anspruch 3, ferner umfassend Leeren des Speichers (20), wenn der Schlüsselsatz in der Datenbank (18) gespeichert ist oder der Satz in dem Puffer (20) gelöscht ist.

Revendications

1. Système (10) servant à réaliser une analyse syntaxique d'un texte en une ou plusieurs phrases qui caractérisent le document, le système (10) comprenant :

une mémoire tampon (20) pour lire (42, 44) un ou plusieurs mots à partir du texte dans la mémoire tampon (20) jusqu'à ce qu'un caractère d'interruption soit identifié ;
un analyseur syntaxique (21) servant à identifier (44) une phrase contenue dans la mémoire tampon, la phrase étant une séquence de deux ou de plusieurs mots entre des caractères d'interruption ;
l'analyseur syntaxique (21) comprenant en outre un moyen destiné à déterminer (46, 52) le type de caractère d'interruption qui suit la phrase identifiée et un moyen destiné à sauvegarder (48, 54, 50) la phrase identifiée comme une phrase clé de la mémoire tampon (20) ou à supprimer la phrase identifiée sur la base du type déterminé de caractère d'interruption ;
une base de données (18) destinée à stocker les phrases clés.

2. Système (10) de la revendication 1, dans lequel la mémoire tampon (20) comprend en outre un moyen destiné à vider la mémoire tampon (20) lorsque la phrase clé est stockée dans la base de données (18)

ou lorsque la phrase dans la mémoire tampon (20) est supprimée.

3. Procédé visant à faire une analyse syntaxique d'un texte en une ou plusieurs phrases qui caractérisent le document, le procédé comprenant le fait :

de lire (42, 44) un ou plusieurs mots à partir du texte dans une mémoire tampon (20) jusqu'à ce qu'un caractère d'interruption soit identifié ;
d'identifier (44) une phrase contenue dans la mémoire tampon, la phrase étant une séquence de deux ou de plusieurs mots entre des caractères d'interruption ;
de déterminer (46, 52) le type de caractère d'interruption qui suit la phrase identifiée ; et
de sauvegarder (48, 54, 50) la phrase identifiée comme une phrase clé à partir de la mémoire tampon (20) dans une base de données (18) ou de supprimer (56) la phrase identifiée sur la base du type déterminé de caractère d'interruption.

4. Procédé de la revendication 3 comprenant en outre le fait de vider la mémoire tampon (20) lorsque la phrase clé est stockée dans la base de données (18) ou lorsque la phrase dans la mémoire tampon (20) est supprimée.

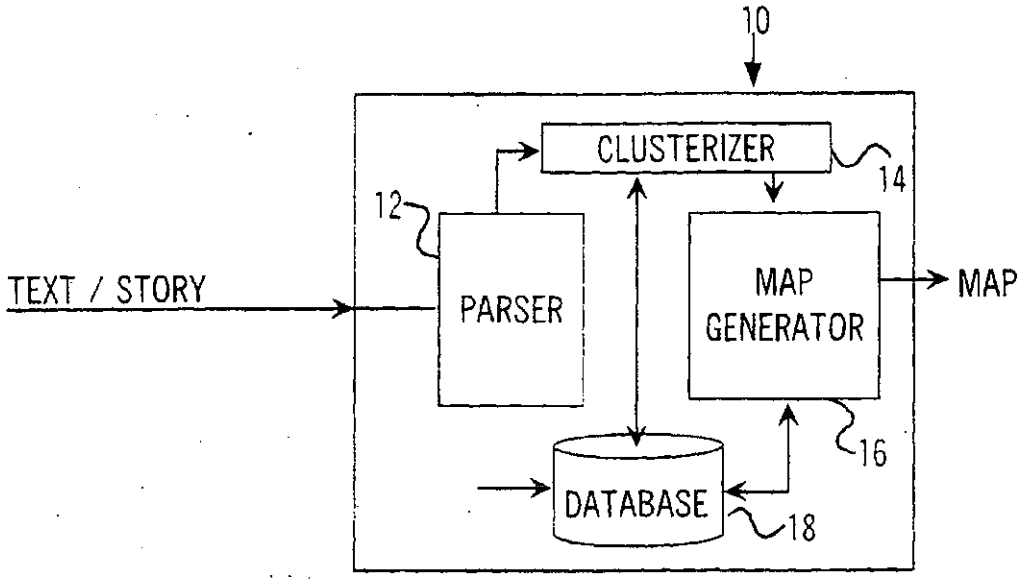


FIGURE 1

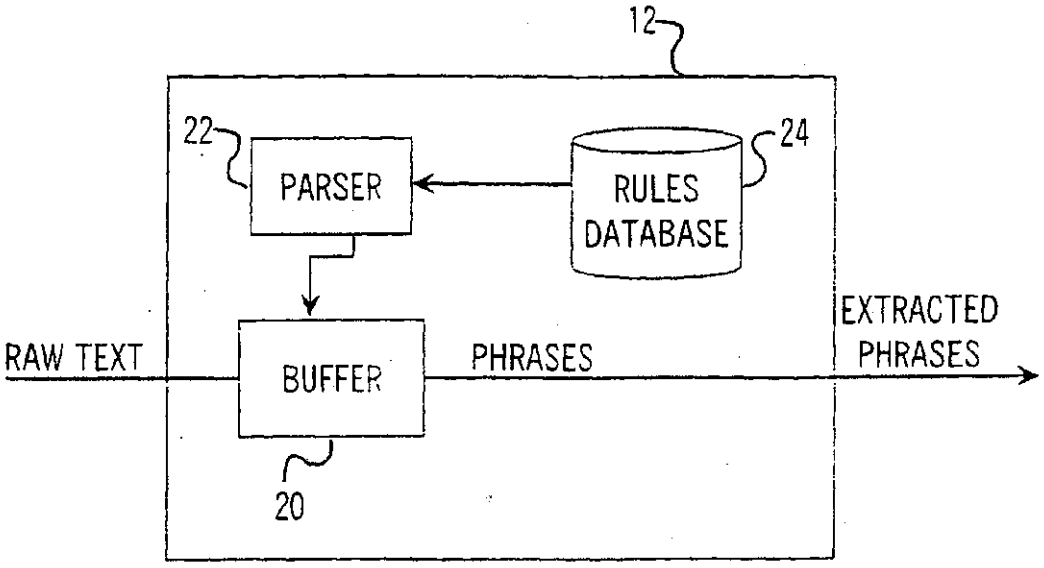


FIGURE 2

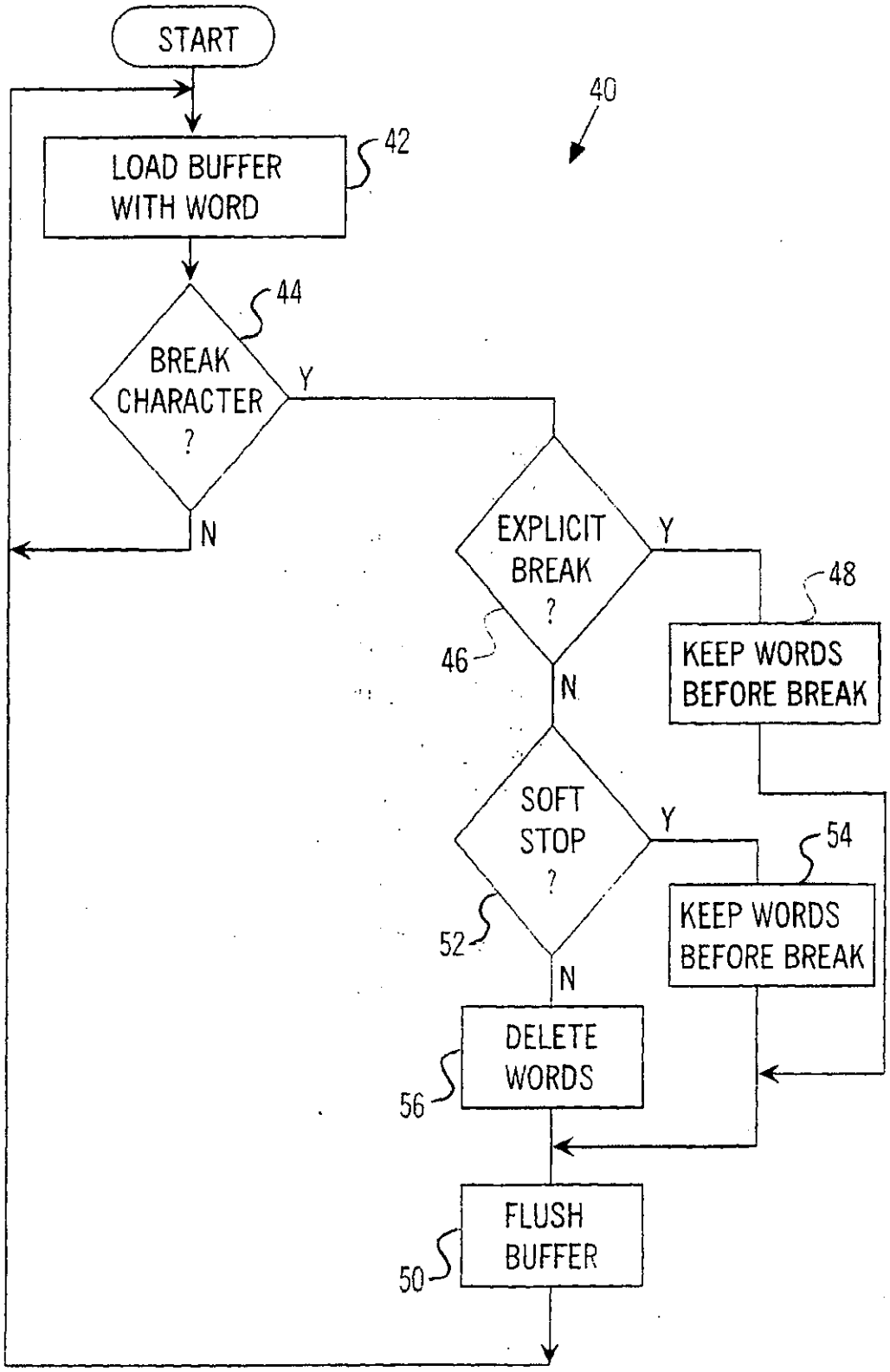

START
40
LOAD BUFFER WITH WORD
42
BREAK CHARACTER ?
44
Y
N
EXPLICIT BREAK ?
Y
48
46
KEEP WORDS BEFORE BREAK
N
SOFT STOP ?
Y
54
52
KEEP WORDS BEFORE BREAK
N
DELETE WORDS
56
FLUSH BUFFER
50

FIGURE 3

60

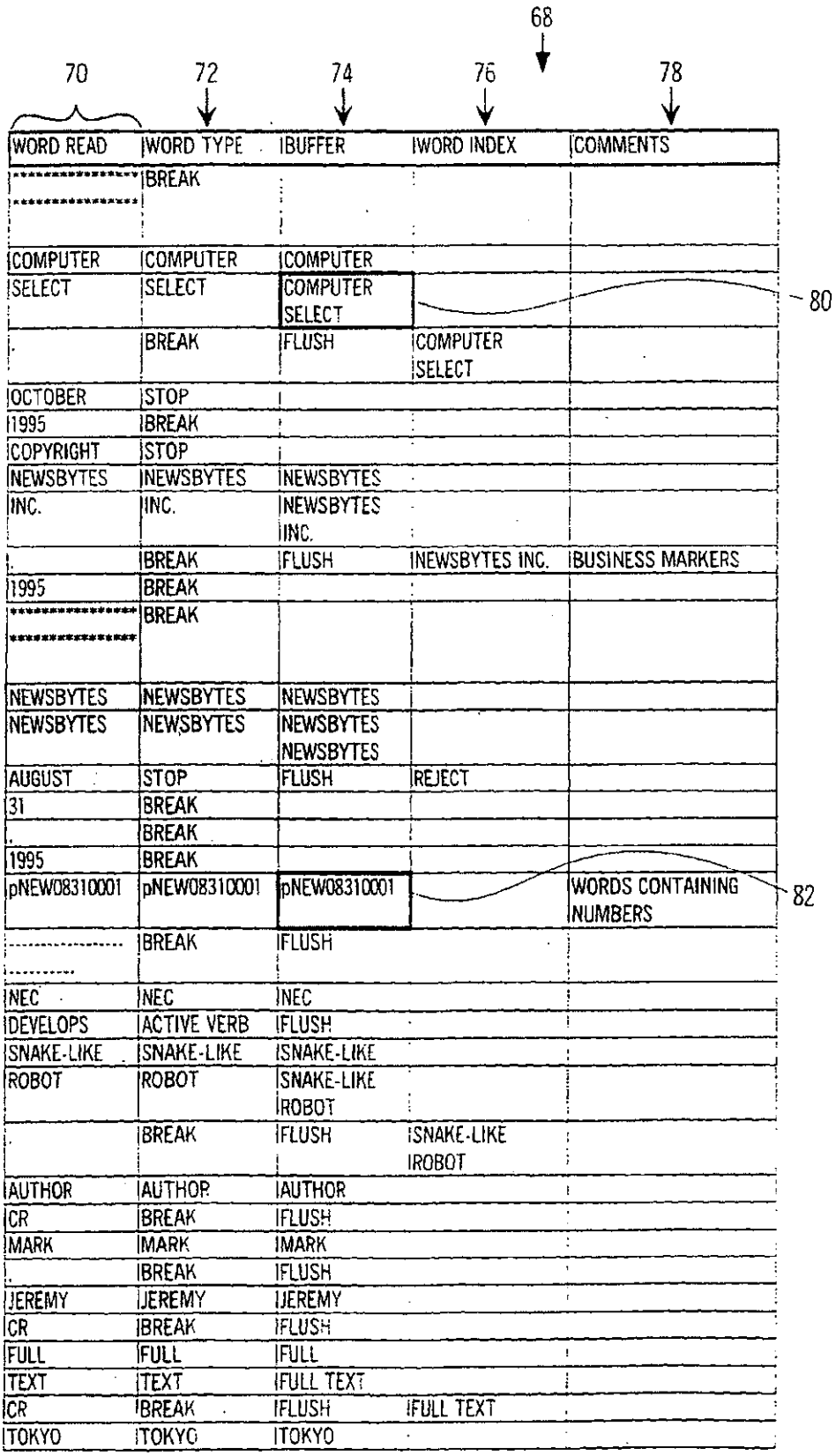
****
COMPUTER SELECT. OCTOBER 1995
COPYRIGHT NEWSBYTES INC. 1995
****

NEWSBYTES
NEWSBYTES AUGUST 31. 1995 PNEW08310001

NEC DEVELOPS "SNAKE-LIKE" ROBOT.
AUTHOR
MARK. JEREMY
FULL TEXT
TOKYO. JAPAN. 1995 AUG 31 (NB) - - NEC CORPORATION HAS DEVELOPED A ROBOTICALLY CONTROLLED ELECTRIC SNAKE THAT OFFERS FAR MORE MOVEMENT THAN PREVIOUSLY DESIGNED ROBOTS. THE COMPANY SAYS THAT THE MUCH GREATER FREEDOM OF MOVEMENT "MAKES IT PERFECT" FOR EVERYTHING FROM INDUSTRIAL TO DISASTER RELIEF WORK.
THE SECRET OF THE NEW DEVICE LIES IN A REVOLUTIONARY NEW TYPE OF UNIVERSAL JOINT. PREVIOUSLY JOINTS IN THE BODY OF ROBOT HAVE BEEN RESTRICTED TO MOVEMENT IN JUST ONE PLANE, EITHER LEFT AND RIGHT OR UP AND DOWN. BUT THE TOKYO-BASED COMPANY SAYS IT HAS SUCCEEDED IN DEVELOPING THE WORLD'S FIRST ACTIVE UNIVERSAL JOINT.
CONTROLLED BY TWO MOTORS, THE JOINT ALLOWS FULL FREEDOM OF MOVEMENT IN ALL PLANES AT EACH OF THE SIX JOINTS ALONG THE ROBOT'S BODY, ALLOWING THE UNIT TO CRAWL INTO PLACES PREVIOUSLY INACCESSIBLE.
AT THE HEART OF THE ROBOT IS A COMPUTER PROCESSING UNIT THAT RECEIVES SIGNALS FROM THE OPERATOR'S HANDSET AND CONTROLS MOVEMENT. THE CONTROLLER CAN INSTRUCT THE COMPUTER TO CONTROL ALL THE JOINTS IN HARMONY OR SPECIFY INDIVIDUAL CONTROL OF EACH JOINT IF NECESSARY.
A VIDEO CAMERA AT THE HEAD OF THE ROBOT SENDS SIGNALS BACK TO THE OPERATOR WHO CAN USE THEM TO STEER THE UNIT AND ALSO TO EXAMINE PLACES INACCESSIBLE TO HUMANS.
THE ENTIRE DEVICE IS 1.4 METERS LONG AND MEASURES 42 MILLIMETERS IN DIAMETER. IT WEIGHS 4.6 KILOGRAMS.
THE AS-YET UNNAMED DEVICE IS NOT YET COMMERCIALLY AVAILABLE, NEC'S MARK PEARCE TOLD NEWSBYTES. "IT WILL BE A COUPLE OF YEARS BEFORE EVERYTHING IS SORTED OUT AND IT'S READY TO BE SOLD. WE HAVE TO INCREASE THE SPEED AMONGST OTHER THINGS," HE SAID.
NEC SAYS TYPICAL APPLICATION FOR SUCH A ROBOT COULD BE INVESTIGATION OF COMPLEX PIPEWORK OR AS AN AID TO SEARCH TEAMS IN DISASTER HIT AREAS WHERE THE DEVICE COULD CRAWL THROUGH THE RUBBLE OF COLLAPSED BUILDINGS.
(MARTYN WILLIAMS/19950830/PRESS CONTACT: MARK PEARCE. NEC CORPORATION. TEL +81-3-3457-7249. INTERNET E-MAIL MAKU_10-22150@ALADDIN.NEC.CO.JP/NEC950831/PHOTO)

RECORD #
17 337 353

FIGURE 4



WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
**********	BREAK			
COMPUTER	COMPUTER	COMPUTER		
SELECT	SELECT	COMPUTER SELECT		
.	BREAK	FLUSH	COMPUTER SELECT	
OCTOBER	STOP			
1995	BREAK			
COPYRIGHT	STOP			
NEWSBYTES	NEWSBYTES	NEWSBYTES		
INC.	INC.	NEWSBYTES INC.		
.	BREAK	FLUSH	NEWSBYTES INC.	BUSINESS MARKERS
1995	BREAK			
**********	BREAK			
NEWSBYTES	NEWSBYTES	NEWSBYTES		
NEWSBYTES	NEWSBYTES	NEWSBYTES NEWSBYTES		
AUGUST	STOP	FLUSH	REJECT	
31	BREAK			
,	BREAK			
1995	BREAK			
pNEW08310001	pNEW08310001	pNEW08310001		WORDS CONTAINING NUMBERS
----------	BREAK	FLUSH		
NEC	NEC	NEC		
DEVELOPS	ACTIVE VERB	FLUSH		
SNAKE-LIKE	SNAKE-LIKE	SNAKE-LIKE		
ROBOT	ROBOT	SNAKE-LIKE ROBOT		
.	BREAK	FLUSH	SNAKE-LIKE ROBOT	
AUTHOR	AUTHOR	AUTHOR		
CR	BREAK	FLUSH		
MARK	MARK	MARK		
.	BREAK	FLUSH		
JEREMY	JEREMY	JEREMY		
CR	BREAK	FLUSH		
FULL	FULL	FULL		
TEXT	TEXT	FULL TEXT		
CR	BREAK	FLUSH	FULL TEXT	
TOKYO	TOKYO	TOKYO		

FIGURE 5A

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
.	BREAK	FLUSH		
JAPAN	JAPAN	JAPAN		
,	BREAK	FLUSH		
1995	BREAK			
AUG	STOP			
31	BREAK			
(	BREAK			
NB	STOP			
)	BREAK			
--	BREAK			
NEC	NEC	NEC		
CORPORATION	CORPORATION	NEC CORPORATION		
HAS	STOP	FLUSH	NEC CORPORATION	
DEVELOPED	DEVELOPED	DEVELOPED		
A	STOP	FLUSH		
ROBOTICALLY	ROBOTICALLY	ROBOTICALLY		
CONTROLLED	CONTROLLED	ROBOTICALLY CONTROLLED		
ELECTRONIC	ELECTRONIC	ROBOTICALLY CONTROLLED ELECTRONIC		
SNAKE	SNAKE	ROBOTICALLY CONTROLLED ELECTRONIC SNAKE	ROBOTICALLY CONTROLLED ELECTRONIC SNAKE	
THAT	STOP	FLUSH		
OFFERS	ACTIVE VERB			
FAR	STOP			
MORE	STOP			
MOVEMENT	MOVEMENT	MOVEMENT		
THAT	STOP	FLUSH		
PREVIOUSLY	PREVIOUSLY	PREVIOUSLY		
DESIGNED	DESIGNED	PREVIOUSLY DESIGNED		
ROBOTS	ROBOTS	PREVIOUSLY DESIGNED ROBOTS		
.	BREAK	FLUSH	PREVIOUSLY DESIGNED ROBOTS	
THE	STOP			
COMPANY	COMPANY	COMPANY		
SAYS	STOP	FLUSH		
THAT	STOP			
THE	STOP			

FIGURE 5B

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
MUCH	STOP			
GREATER	STOP			
FREEDOM	FREEDOM	FREEDOM		
OF	CONNECTOR	FREEDOM OF		
MOVEMENT	MOVEMENT	FREEDOM OF MOVEMENT		
MAKES	STOP	FLUSH	FREEDOM OF MOVEMENT	
IT	STOP			
PERFECT	PERFECT	PERFECT		
FOR	STOP	FLUSH		
EVERYTHING	STOP			
FROM	STOP			
INDUSTRIAL	INDUSTRIAL	INDUSTRIAL		
TO	STOP	FLUSH		
DISASTER	DISASTER	DISASTER		
RELIEF	RELIEF	DISASTER RELIEF		
WORK	WORK	DISASTER RELIEF WORK		
.	BREAK	FLUSH	DISASTER RELIEF WORK	
THE	STOP			
SECRET	SECRET	SECRET		
OF	CONNECTOR	SECRET OF		
THE	THE	SECRET OF THE		
NEW	NEW	SECRET OF THE NEW		
DEVICE	DEVICE	SECRET OF THE NEW DEVICE		
LIES	ACTIVE VERB	FLUSH	SECRET OF THE NEW DEVICE	
IN	STOP			
A	STOP			
REVOLUTIONARY	REVOLUTIONARY	REVOLUTIONARY		
NEW	NEW	REVOLUTIONARY NEW		
TYPE	TYPE	REVOLUTIONARY NEW TYPE		
OF	CONNECTOR	REVOLUTIONARY NEW TYPE OF		
UNIVERSAL	UNIVERSAL	REVOLUTIONARY NEW TYPE OF UNIVERSAL		
JOINT	JOINT	REVOLUTIONARY NEW TYPE OF UNIVERSAL JOINT		

FIGURE 5C

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
	BREAK	FLUSH	REVOLUTIONARY NEW TYPE OF UNIVERSAL JOINT	
PREVIOUSLY	PREVIOUSLY	PREVIOUSLY		
JOINTS	ACTIVE VERB	FLUSH		
IN	STOP			
THE	STOP			
BODY	BODY	BODY		
OF	CONNECTOR	OF		
ROBOT	ROBOT	BODY OF ROBOT		
HAVE	STOP	FLUSH	BODY OF ROBOT	
BEEN	STOP			
RESTRICTED	RESTRICTED	RESTRICTED		
TO	STOP	FLUSH		
MOVEMENT	MOVEMENT	MOVEMENT		
IN	STOP	FLUSH		
JUST	STOP			
ONE	STOP			
PLANE	PLANE	PLANE		
.	BREAK	FLUSH		
EITHER	STOP			
LEFT	LEFT	LEFT		
AND	CONNECTOR	LEFT AND		
RIGHT	RIGHT	LEFT AND RIGHT		
OR	STOP	FLUSH	LEFT AND RIGHT	
UP	STOP			
AND	CONNECTOR			
DOWN	STOP			
.	BREAK			
BUT	STOP			
THE	STOP			
TOKYO-BASED	TOKYO-BASED	TOKYO-BASED		
COMPANY	COMPANY	FLUSH COMPANY		
SAYS	STOP	FLUSH		
IT	STOP			
HAS	STOP			
SUCCEEDED	SUCCEEDED	SUCCEEDED		
IN	STOP	FLUSH		
DEVELOPING	DEVELOPING	DEVELOPING		
THE	STOP	FLUSH		
WORLD'S	WORLD'S	WORLD'S		
FIRST	FIRST	WORLD'S FIRST		

FIGURE 5D

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
ACTIVE	ACTIVE	WORLD'S FIRST ACTIVE		
UNIVERSAL	UNIVERSAL	WORLD'S FIRST ACTIVE UNIVERSAL		
JOINT	JOINT	WORLD'S FIRST ACTIVE UNIVERSAL JOINT		
.	BREAK	FLUSH	WORLD'S FIRST ACTIVE UNIVERSAL JOINT	
CONTROLLED	CONTROLLED	CONTROLLED		
BY	STOP	FLUSH		
TWO	STOP			
MOTORS	MOTORS	MOTORS		
.	BREAK	FLUSH		
THE	STOP			
JOINT	JOINT	JOINT		
ALLOWS	ACTIVE VERB	FLUSH		
FULL	FULL	FULL		
FREEDOM	FREEDOM	FULL FREEDOM		
OF	CONNECTOR	FULL FREEDOM OF		
MOVEMENT	MOVEMENT	FULL FREEDOM OF MOVEMENT		
IN	STOP	FLUSH	FULL FREEDOM OF MOVEMENT	
ALL	STOP			
PLANES	ACTIVE VERB			
AT	STOP			
EACH	STOP			
OF	CONNECTOR			
THE	THE			
SIX	STOP			
JOINTS	ACTIVE VERB			
ALONG	STOP			
THE	STOP			
ROBOT'S	ROBOT'S	ROBOT'S		
BODY	BODY	ROBOT'S BODY		
.	BREAK	FLUSH	ROBOT'S BODY	
ALLOWING	ALLOWING	ALLOWING		
THE	STOP	FLUSH		
UNIT	UNIT	UNIT		
TO	STOP	FLUSH		
CRAWL	CRAWL	CRAWL		

FIGURE 5E

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
INTO	STOP	FLUSH		
PLACES	ACTIVE VERB			
PREVIOUSLY	PREVIOUSLY	PREVIOUSLY		
INACCESSIBLE	INACCESSIBLE	PREVIOUSLY INACCESSIBLE		
.	BREAK	FLUSH	PREVIOUSLY INACCESSIBLE	
AT	STOP			
THE	STOP			
HEART	HEART	HEART		
OF	CONNECTOR	HEART OF		
THE	THE	HEART OF THE		
ROBOT	ROBOT	HEART OF THE ROBOT		
IS	STOP	FLUSH	HEART OF THE ROBOT	
A	STOP			
COMPUTER	COMPUTER	COMPUTER		
PROCESSING	PROCESSING	COMPUTER PROCESSING		
UNIT	UNIT	COMPUTER PROCESSING UNIT		
THAT	STOP	FLUSH	COMPUTER PROCESSING UNIT	
RECEIVES	ACTIVE VERB			
SIGNALS	SIGNALS	SIGNALS		
FROM	STOP	FLUSH		
THE	STOP			
OPERATOR'S	OPERATOR'S	OPERATOR'S		
HANDSET	HANDSET	OPERATOR'S HANDSET		
AND	CONNECTOR	OPERATOR'S HANDSET AND		
CONTROLS	ACTIVE VERB	FLUSH	OPERATOR'S HANDSET	
MOVEMENT	MOVEMENT	MOVEMENT		
.	BREAK	FLUSH		
THE	STOP			
CONTROLLER	CONTROLLER	CONTROLLER		
CAN	STOP	FLUSH		
INSTRUCT	INSTRUCT	INSTRUCT		
THE	STOP	FLUSH		
COMPUTER	COMPUTER	COMPUTER		
TO	STOP	FLUSH		
CONTROL	CONTROL	CONTROL		

FIGURE 5F

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
ALL	STOP	FLUSH		
THE	STOP			
JOINTS	ACTIVE VERB			
IN	STOP			
HARMONY	HARMONY	HARMONY		
OR	STOP	FLUSH		
SPECIFY	SPECIFY	SPECIFY		
INDIVIDUAL	INDIVIDUAL	INDIVIDUAL		
CONTROL	CONTROL	INDIVIDUAL CONTROL		
OF	CONNECTOR	INDIVIDUAL CONTROL OF		
EACH	STOP	FLUSH	INDIVIDUAL CONTROL	
JOINT	JOINT	JOINT		
IF	STOP	FLUSH		
NECESSARY	NECESSARY	NECESSARY		
.	BREAK	FLUSH		
A	STOP			
VIDEO	VIDEO	VIDEO		
CAMERA	CAMERA	VIDEO CAMERA		
AT	STOP	FLUSH	VIDEO CAMERA	
THE	STOP			
HEAD	HEAD	HEAD		
OF	CONNECTOR	HEAD OF		
THE	THE	HEAD OF THE		
ROBOT	ROBOT	HEAD OF THE ROBOT		
SENDS	ACTIVE VERB	FLUSH	HEAD OF THE ROBOT	
SIGNALS	SIGNALS	SIGNALS		
BACK	STOP	FLUSH		
TO	STOP			
THE	STOP			
OPERATOR	OPERATOR	OPERATOR		
WHO	STOP	FLUSH		
CAN	STOP			
USE	STOP			
THEM	STOP			
TO	STOP			
STEER	STEER	STEER		
THE	STOP	FLUSH		
UNIT	UNIT	UNIT		
AND	CONNECTOR	UNIT AND		
ALSO	STOP	FLUSH		
TO	STOP	CONTROL		

FIGURE 5G

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
EXAMINE	EXAMINE	EXAMINE		
PLACES	ACTIVE VERB	FLUSH		
INACCESSIBLE	INACCESSIBLE	INACCESSIBLE		
TO	STOP	FLUSH		
HUMANS	HUMANS	HUMANS		
.	BREAK	FLUSH		
THE	STOP			
ENTIRE	ENTIRE	ENTIRE		
DEVICE	DEVICE	ENTIRE DEVICE		
IS	STOP	FLUSH	ENTIRE DEVICE	
1	BREAK			
.	BREAK			
A	STOP			
METERS	ACTIVE VERB			
LONG	LONG	LONG		
AND	CONNECTOR	LONG AND		
MEASURES	ACTIVE VERB	FLUSH		
42	BREAK			
MILLIMETERS	MILLIMETERS	MILLIMETERS		
IN	STOP	FLUSH		
DIAMETER	DIAMETER	DIAMETER		
.	BREAK	FLUSH		
IT	STOP			
WEIGHS	WEIGHS	WEIGHS		
4	BREAK	FLUSH		
.	BREAK			
6	BREAK			
KILOGRAMS	KILOGRAMS	KILOGRAMS		
.	BREAK	FLUSH		
THE	STOP			
AS-YET	AS-YET	AS-YET		
UNNAMED	UNNAMED	AS-YET UNNAMED		
DEVICE	DEVICE	AS-YET UNNAMED DEVICE		
IS	STOP	FLUSH	AS-YET UNNAMED DEVICE	
NOT	STOP			
YET	STOP			
COMMERCIALLY	COMMERCIALLY	COMMERCIALLY		
AVAILABLE	AVAILABLE	COMMERCIALLY AVAILABLE		
.	BREAK	FLUSH	COMMERCIALLY AVAILABLE	

FIGURE 5H

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
NEC'S	NEC'S	NEC'S		
MARK	MARK	NEC'S MARK		
PEARCE	PEARCE	NEC'S MARK PEARCE	NEC'S MARK PEARCE	
TOLD	ACTIVE VERHB (PAST TENSE)	FLUSH TOLD		
NEWSBYTES	NEWSBYTES	FLUSH NEWSBYTES		
.	BREAK	FLUSH		
IT	STOP			
WILL	STOP			
BE	STOP			
A	STOP			
COUPLE	COUPLE	COUPLE		
OF	CONNECTOR	COUPLE OF		
YEARS	YEARS	COUPLE OF YEARS		
BEFORE	STOP	FLUSH	COUPLE OF YEARS	
EVERYTHING	STOP			
IS	STOP			
SORTED	SORTED	SORTED		
OUT	STOP	FLUSH		
AND	CONNECTOR			
IT'S	STOP			
READY	READY	READY		
TO	STOP	FLUSH		
BE	STOP			
SOLD	SOLD	SOLD		
.	BREAK	FLUSH		
WE	STOP			
HAVE	STOP			
TO	STOP			
INCREASE	INCREASE	INCREASE		
THE	STOP	FLUSH		
SPEED	SPEED	SPEED		
AMONGST	STOP	FLUSH		
OTHER	STOP			
THINGS	THINGS	THINGS		
.	BREAK	FLUSH		
HE	STOP			
SAID	STOP			
.	BREAK			
NEC	NEC	NEC		
SAYS	STOP	FLUSH		
TYPICAL	TYPICAL	TYPICAL		

FIGURE 5I

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
APPLICATIONS	APPLICATIONS	TYPICAL APPLICATIONS		
FOR	STOP	FLUSH	TYPICAL APPLICATIONS	
SUCH	STOP			
A	STOP			
ROBOT	ROBOT	ROBOT		
COULD	STOP	FLUSH		
BE	STOP			
INVESTIGATION	INVESTIGATION	INVESTIGATION		
OF	CONNECTOR	INVESTIGATION OF		
COMPLEX	COMPLEX	INVESTIGATION OF COMPLEX		
PIPEWORK	PIPEWORK	INVESTIGATION OF COMPLEX PIPEWORK		
OR	STOP	FLUSH	INVESTIGATION OF COMPLEX PIPEWORK	
AS	STOP			
AN	STOP			
AID	AID	AID		
TO	STOP	FLUSH		
SEARCH	SEARCH	SEARCH		
TEAMS	ACTIVE VERB	FLUSH		
IN	STOP			
DISASTER	DISASTER	DISASTER		
HIT	HIT	DISASTER HIT		
AREAS	AREAS	DISASTER HIT AREAS		
WHERE	STOP	FLUSH	DISASTER HIT AREAS	
THE	STOP			
DEVICE	DEVICE	DEVICE		
COULD	STOP	FLUSH		
CRAWL	CRAWL	CRAWL		
THROUGH	STOP	FLUSH		
THE	STOP			
RUBBLE	RUBBLE	RUBBLE		
OF	CONNECTOR	RUBBLE OF		
COLLAPSED	COLLAPSED	RUBBLE OF COLLAPSED		

FIGURE 5J

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
BUILDINGS	BUILDINGS	RUBBLE OF COLLAPSED BUILDINGS		
.	BREAK	FLUSH	RUBBLE OF COLLAPSED BUILDINGS	
(	BREAK			
MARTYN	MARTYN	MARTYN		
WILLIAMS/19950830/PRESS	WILLIAMS/19950830/PRESS	MARTYN WILLIAMS/19950830/PRESS	MARTYN WILLIAMS/19950830/PRESS	WORDS CONTAINING NUMBERS
CONTACT	CONTACT	FLUSH CONTACT		
:	BREAK	FLUSH		
MARK	MARK	MARK		
PEARCE	PEARCE	MARK PEARCE		
,	BREAK	FLUSH	MARK PEARCE	
NEC	NEC	NEC		
CORPORATION	CORPORATION	NEC CORPORATION		
,	BREAK	FLUSH	NEC CORPORATION	
TEL	STOP			
+81-3-3798-6511	BREAK			
,	BREAK			
FAX	STOP			
+81-3-3457-7249	BREAK			
.	BREAK			
INTERNET	INTERNET	INTERNET		
E-MAIL	STOP	FLUSH		
MAKU_10-22150@ALADDIN	MAKU_10-22150@ALADDIN	MAKU_10-22150@ALADDIN		WORDS CONTAINING NUMBERS WORDS CONTAINING INTERNET CODES
.	BREAK	FLUSH		
NEC	NEC	NEC		
.	BREAK	FLUSH		
CO	CO	CO		
.	BREAK	FLUSH		
JP/NEC950831/PHOTO	JP/NEC950831/PHOTO	JP/NEC950831/PHOTO		WORDS CONTAINING NUMBERS
)	BREAK	FLUSH		
----------	BREAK			

FIGURE 5K

WORD READ	WORD TYPE	BUFFER	WORD INDEX	COMMENTS
RECORD	RECORD	RECORD		
#	BREAK	FLUSH		
17	BREAK			
337	BREAK			
353	BREAK			

FIGURE 5L

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 5819260 A **[0006]**
- WO 9838560 A **[0012]**

Non-patent literature cited in the description

- **H. Witten.** *KEA: Practical Automatic Keyphrase Extraction* **[0006]**
- **P. Turkey.** *Learning to Extract Keyphrases from Text* **[0006]**