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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 April 2006 (13.04.2006)

PCT

(10) International Publication Number
WO 2006/038008 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number:
PCT/GB2005/003839

(22) International Filing Date: 6 October 2005 (06.10.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0422196.6 6 October 2004 (06.10.2004) GB

(71) Applicant *(for all designated States except US)*: **IMPERIAL COLLEGE INNOVATIONS LTD** [GB/GB]; Electrical and Electronic Engineering Building, Level 12, Imperial College, Exhibition Road, London SW7 2AZ (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RUEGER, Stefan, M.** [DE/GB]; Department of Computing, South Kensington Campus, Imperial College, London SW7 2AZ (GB). **PICKERING, Marcus, Jerome** [GB/GB]; 5 Peacock Way, Histon, Cambridge CB4 9XZ (GB).

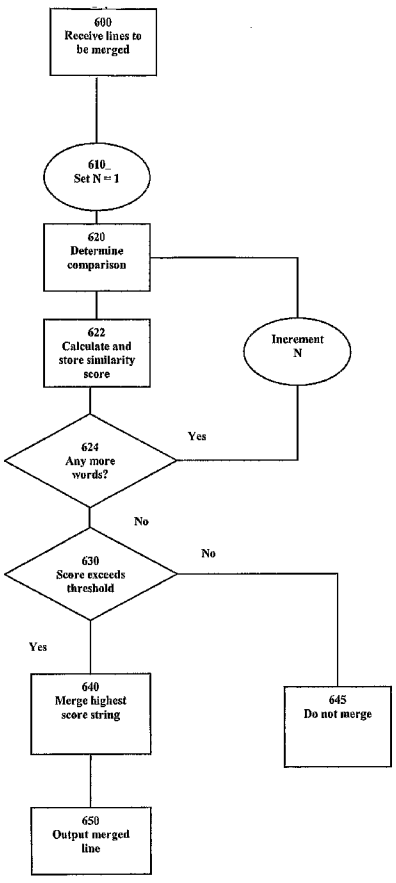
(74) Agents: **ASHMEAD, Richard, John** et al.; Kilburn & Strode, 20 Red Lion Street, London WC1R 4PJ (GB).

(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, 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).

[Continued on next page]

(54) Title: METHOD FOR MERGING SUBTITLES



(57) Abstract: A method of electronically processing a broadcast subtitle signal so as to merge lines of subtitles and correct possible transmission errors allows subtitles to be rendered and/or stored in a form suitable for further text processing and, in particular, keyword searches. Applications of the method or computer program or system implementing the method include the creation, maintenance, indexing or searching of a multimedia library or database, in particular when the literary comprises broadcast television programmes and the corresponding subtitles.

WO 2006/038008 A1

Declaration under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report*

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.

METHOD FOR MERGING SUBTITLES

The invention relates to data processing.

The growing amount of data available in general requires ever more sophisticated techniques to allow this data to be searched. Increasing amounts of data is available in multimedia databases, that is collections of interrelated data of different media types, for example images, movies, sounds or text. Under this broad definition, examples of a multimedia database are a DVD disk or the Internet. In the absence of effective search tools, the ever increasing amounts of data may become difficult to locate or recover.

Where such multimedia databases contain text, powerful search tools based on textual keywords may be applied. Examples of this are a web search using a search engine on the Internet or searching for items on a DVD by looking for certain keywords in the subtitles stored along the moving images.

However, problems arise in relation to certain types of content, for example the subtitles supplied by broadcasters with live television programmes, such as a news programme. In particular, the way in which these subtitles are transmitted gradually, and the fact that transmission errors occur, limit the usefulness of subtitles for keyword searches that rely on exact textual matches.

The invention is set out in the independent claims and further, optional, features are defined in the dependent claims.

Because successive lines of subtitles are merged in such a way as to form a single searchable text string and also to correct errors, the utility of subtitles in

indexing and searching a database of recorded television programmes, such as a news broadcast, is greatly enhanced.

A specific embodiment of the invention will now be described by way of example only and with reference to the accompanying drawings, in which:
Figures 1 and 2 illustrate a transmission scheme for subtitles;
Figures 3 and 4 illustrate the occurrence of transmission errors;
Figure 5 illustrates a method of merging subtitle lines according to an embodiment of the invention; and
Figure 6 is a flow chart representing a method according to an embodiment of the invention.

Poor quality subtitles often contain insufficient or unreliable information to provide the necessary keywords to facilitate acceptable information retrieval performance for a search algorithm, especially as known language processing techniques rely on having good quality text in complete sentences. This is not the case for subtitles transmitted with TV broadcast programmes and therefore some processing of title text captured together with the television signal is necessary to render the subtitles in a form that is suitable for further language processing and searching.

As well as imperfection caused by interference in the broadcast, a number of problems are caused by the way that live subtitles are transmitted. In the case for example, of British Broadcasting Corporation news programmes, many duplicate phrases and lines are transmitted although these appear seamlessly when displayed on a TV screen. These problems are illustrated in figures 1, 2, 3 and 4.

With reference to figures 1 and 2, normal error frequency transmission of subtitles in the blanking interval of the TV signal is distinguished from subtitles as provided with, for example, a movie, in that a new line of subtitles is transmitted approximately every second. This can be seen from the timestamp shown in the left hand column of the figures, each line potentially being an extension of, or overlapping with, the previous line. This repetitive retransmission of the subtitle signal is irrespective of the actual content of the subtitles, such that two distinct situations may arise.

The first, illustrated in figure 1, occurs in the situation where the spoken words of the television programme resume after a pause. In this case, the subtitles grow by zero or more words from one line to the next, thus resulting in a set of subtitle lines which are partially overlapping and aligned on the first word of the subtitle. In the second case, in the middle of a sequence of spoken words, subsequent subtitle lines will also be partially overlapping, but not necessarily aligned with the first word of the subtitle, as shown in figure 2, such that a window or mask of varying length effectively moves along the sequence with each pass.

The problem of matching and merging the line fragments shown in figures 1 and 2 is exacerbated by the further problem of transmission errors occurring in the subtitles. Figures 3 and 4 illustrate the occurrence of such transmission errors (highlighted), for the cases corresponding to figures 1 and 2, respectively. As a result of the transmission errors, adjacent lines of subtitles cannot be merged directly and further processing is necessary.

The method according to the present invention implements a similarity measure between successive lines of subtitles in an iterative approach as discussed in more detail below. An appropriate similarity measure will first be discussed.

In order to correct transmission errors and allow the subtitles to be merged, a function that returns a measure of the similarity of two strings, for example Marc Lehmann's Perl module String::Similarity, is employed to detect lines which are duplicates or partial duplicates of preceding and succeeding lines even if they may contain errors. The similarity function returns, for example, a score of one if two strings are identical and zero is the strings are entirely different, with all other values in between zero and one based on the edit distance between the two strings, that is the number of single letter edits required to change one string into the other. Of course, any other robust measure of similarity may also be used.

To find out whether two lines have indeed any overlap and should be merged, an initial comparison is carried out to determine the possibility of a match. . This initial comparison is done by directly comparing consecutive lines using String::Similarity. If the similarity score is greater than a high threshold, the lines are considered exact duplicates and only the earlier one is kept.

Consecutive lines which display the possibility of a match, i.e. where the similarity score obtained above is greater than zero, , are compared in detail, to find a whether a point exists in the subtitles at which they are appropriate to be matched. This is described in detail below, with reference to figure 5.

As shown in figure 5, two, usually subsequent, lines of subtitles are "zipped across each other" in a number of iterations by defining a comparison window that grows from one iteration to the next. With each iteration, the window size (which starts with one word) is increased by one word until the window size is equal to the number of words in the shorter of the two lines. The window is arranged such that it contains an overlap of words of the two strings which are being compared, for example the last word of the first line and the first word of

the second line on the first iteration, then the last two words of the first line and the first two words of the second line on the second iteration and so on. The two strings constituted by the words within the comparison window are examined on each iteration using the similarity function and the score returned by the similarity function is stored for further analysis. Once the last comparison has been carried out, that is when the entire shorter string has been compared to the corresponding words of the longer string, the similarity scores are compared and the iteration with the highest similarity score is used to indicate the merge point, providing the score exceeds a threshold, for example 0.8. If the threshold is not exceeded on any iteration, no merge is made between the two subtitle lines in question.

The iteration resulting in the highest similarity score, which is used for subsequent merging, is shown in italics in figure 5. In order to form a new merged line, the string to the left of the comparison window is taken as the start of the line. If the similarity score for the window is 1 (i.e. an identical match) the strings inside the comparison window (one from each subtitle line) are identical, and it does not matter which one is chosen to continue the line. For example, the string in the comparison window coming from the first line may be used to continue the merged line. The merged line is then completed by concatenating the remaining words of the second line. However, if the strings inside the comparison window do not match exactly (a score lower than 1 but larger than the threshold), it becomes necessary to select which of the two strings in the comparison window is used in the concatenation to form the merged strings. In this case, the string inside the window with the most charactersis used to complete the lines, since subtitle errors are usually characterised by omission of characters. In the example shown in figure 6, "blood donations" has becomes "blod doatins" through the omissions of characters. The correct string is the longer one. It will be seen that by repeating

this process for each successive pair of lines a single optimum line can be constructed from the multiple concatenations.

An algorithm for merging two lines of subtitles is now described in more detail with reference to figure 6. At step 600, a computer program implementing the algorithm receives two subtitle lines which are to be merged. This may be two adjacent or successive lines of subtitles which are received in real time from a broadcast or from a recording of a broadcast for example from a television tuner (terrestrial, satellite or cable) or network connection. The two received subtitle lines may alternatively be pre-selected according to the likelihood of giving a successful merge and may or may not be temporally subsequent to each other.

At step 610, a counter determining the size of the comparison window is initiated to a value of N=1 and at step 620 a comparison window is determined which comprises the last N word or words of the first line and the first N word or words of the second, possible subsequent line. At step 622, a similarity score is calculated, as described above as a measure of the similarity between the first and second line inside the comparison window and the score is stored in memory. At decision node 624, if the counter is smaller than the number of words in the shorter of the two strings, the counter N is increased by one at step 626 and the algorithm returns to step 620, repeating the comparison for a window of increased size.

If at decision node 624 it is determined that N is equal to the number of words in the smaller of the two lines, then at decision node 630, if any of the scores exceeds a threshold value, the comparison window corresponding to the iteration highest score having the scores is selected for merging at step 640. The two subtitle lines are merged by selecting one of the two strings inside the

comparison window and concatenating it to the left with the substring of the first, possibly longest or temporally first, line and concatenating the result with the substring on the second, possibly shorter or temporally subsequent, line to the right of the comparison window. If the score of the selected comparison window is equal to one, the substrings inside the comparison window are identical and any of the two substrings can be chosen for the above concatenation. If, on the other hand, the score if lower than one, presumably due to a mismatch resulting from a transmission error, the substring inside the selected comparison window that has a larger number of non blank characters is selected for the concatenation.

Following a successful merge operation, the new, merged, line of subtitles is outputted at step 650, either for further textual processing or to be stored on a storage medium.

If, on the other hand it is determined at decision node 630 that none of the scores calculated for any of the comparison windows exceeds the threshold, the pair of subtitle lines is considered to be distinct from each other and no merge is made (step 645).

It will be understood that the invention described with reference to the embodiment above can be implemented using any suitable hardware and software. The subtitles may be received from a live feed, such as a television signal or may be stored in a storage medium. For example, subtitles may be captured by a script on a Linux based PC fitted with a TV PCI card, for example a Hauppauge WIN TV PCI card. In the set up, video recording may be carried out using the "streamer" application from the XAWTV suite (see http://bytesex.org/xawtv). The subtitles may be captured using a modified version of the ALEVT software (see http://www.goron.de/~froese/).

Although the specific embodiment of the invention has been described with reference to the subtitles of a television broadcast, it is understood that the invention may be applied to any other form of subtitles or other text-strings which contain redundant information and may be in need of error-correction. For example, the inventions may be applied to text-strings which are the result of electronic voice recognition.

Claims

1. A method of electronically processing a broadcast subtitle signal comprising a plurality of lines of subtitles, the method comprising merging first and second lines of subtitles.

2. A method as claimed in claim 1 comprising selecting a portion of each of the first and second lines, and calculating a similarity measure between the portions.

3. A method as claimed in claim 2 further comprising successively iteratively increasing the portion and merging the portion having the greatest similarity measure.

4. A method as claimed in claim 1, the method further comprising;

a. receiving a first and second subtitle line;

b. setting a counter to a value N=1;

c. determining a comparison window comprising as respective portions the last N word or words of the first line and the first N word or words of the second line;

d. calculating a score which is a measure of the similarity between a substring inside the comparison window of the first line and a substring inside the comparison window of the second line and storing the score;

e. if the counter is smaller than the number of words in the shorter of the of the first and second strings, increasing the counter N by one and repeating steps c to e; and

f. if the counter N is equal to the number of words in the shorter of the first and second string, merging the first and second string according to the stored scores.

5. A method as claimed in any preceding claim wherein the first and second lines of subtitles are received from a storage device, television tuner or internet connection and/or are subsequent with respect to each others.

6. A method as devised in any of the preceding claims in which the first and second lines comprise successively received lines.

7. A method as claimed in any preceding claims wherein the first and second subtitle lines are selected from a plurality of subtitle lines based on the likelihood of representing overlapping subtitle lines.

8. A method as claimed in claim 4 further including:
g. if any of the scores exceeds a threshold value, selecting a window corresponding to the highest score, concatenating a string selected from the two strings out of the strings from each of the first and second line inside the selected window to the left with the sub-string of the first line to the left of the selected window and concatenating the selected string to the right with the sub-string of the second line to the right of the selected window.

9. A method as claimed in claim 6, wherein the selected string is selected such that it has a larger number of characters than the other string inside the comparison windows.

10. A method of computing a searchable text-string comprising receiving as respective data input first and second text lines, computing merged text lines according to a method as claimed in any claims 1 to 9 and

providing as a data output the merged text lines as a searchable text string.

11. A computer system implementing a method as claimed in any of the preceding claims.

12. A computer program comprising code instruction implementing a method as claimed in any of claims 1 to 10.

13. A computer readable medium or physical signal carrying a computer program as claimed in claim 12.

14. A method of creating, maintaining, indexing or searching a multimedia library comprising a method as claimed in any of claims 1 to 10.

91.913086 An attack
92.912983 An attack on the UK is
93.912868 An attack on the UK is now, it
94.912876 An attack on the UK is now, it seems,
95.995791 An attack on the UK is now, it seems, inevitable. That

Figure 1

60.433061 Good evening. Britain's most senior policeman,
60.995228 Britain's most senior policeman, Sir John
62.073082 Britain's most senior policeman, Sir John Stevens, has spelled out
62.953060 Sir John Stevens, has spelled out the nature of
63.953063 Sir John Stevens, has spelled out the nature of the terrorist threat.
66.392871 the nature of the terrorist threat. The Metropolitan Police

Figure 2

108.313087 Counter-terrorist officers on the strets of London
109.112912 Counter-terrorist officers on the streets of London today, are we

Figure 3

80.912854 London's Mayor, Ken Livingstone, speaking at the same
81.912859 speaking at the same news conference
83.153141 nes conference said it would be "miraculous"

Figure 4

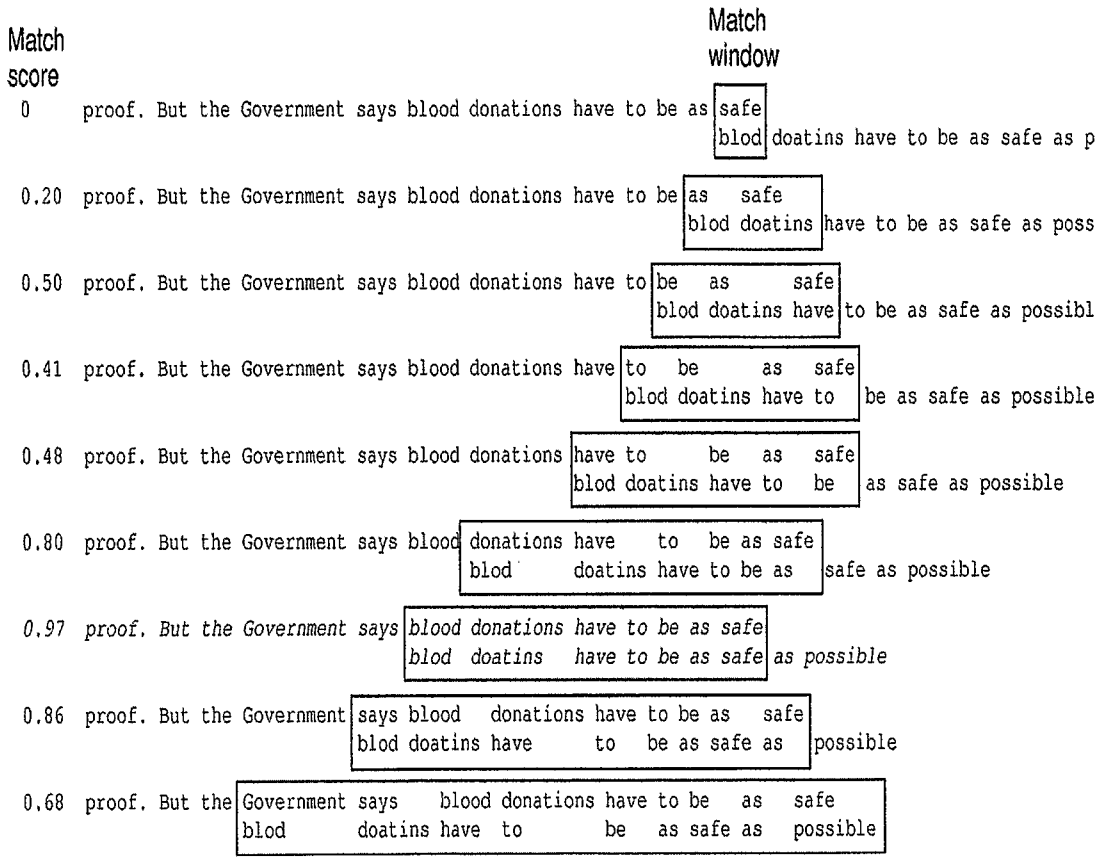

Match score
Match window
0 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as p
0.20 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as poss
0.50 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possibl
0.41 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possible
0.48 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possible
0.80 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possible
0.97 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possible
0.86 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possible
0.68 proof. But the Government says blood donations have to be as safe
blod doatins have to be as safe as possible

Figure 5

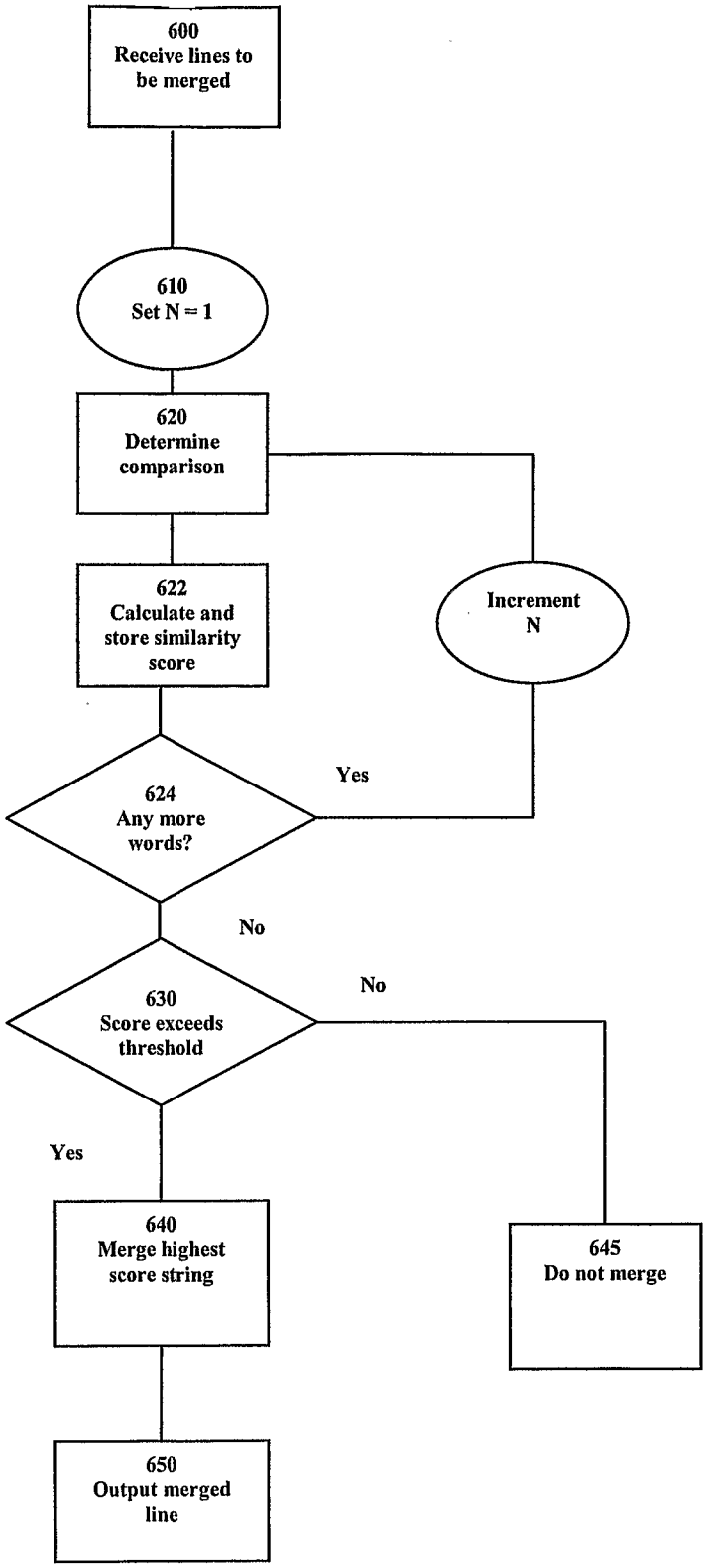

600
Receive lines to be merged
610
Set N = 1
620
Determine comparison
622
Calculate and store similarity score
Increment N
624
Any more words?
Yes
No
630
Score exceeds threshold
No
Yes
640
Merge highest score string
645
Do not merge
650
Output merged line

Figure 6

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PC1/GB2005/003839

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30

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

B. FIELDS SEARCHED

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

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	NEWMAN E ET AL: "Comparing Redundancy Removal Techniques for Multi-document Summarisation" IN THE PROCEEDINGS OF STAIRS, 'Online! August 2004 (2004-08), pages 223-228, XP002358715 Retrieved from the Internet: URL:http://citeseer.ist.psu.edu/newman04comparing.html> 'retrieved on 2005-12-12!	1-3,5-7, 10-14
A	page 223, paragraph 4 - page 224, paragraph 4 page 224, last paragraph - page 225, last paragraph	4,8,9

-/--

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

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 December 2005	02/01/2006

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Konak, E

1

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

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PC'1 / uB2005/003839

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BROWN M G ET AL: "Automatic Content-Based Retrieval of Broadcast News" ACM MULTIMEDIA, 'Online! 1995, pages 35-43, XP002358716 Retrieved from the Internet: URL:http://citeseer.ist.psu.edu/brown95automatic.html> 'retrieved on 2005-12-12!	1,2,5-7, 10-14
A	page 38, left-hand column, paragraph 2 figure 6	3,4,8,9
X	KENT W J ET AL: "Assembly of the Working Draft of the Human Genome with GigAssembler" COLD SPRING HARBOR LABORATORY PRESS, 2001, pages 1541-1548, XP002358717	1,2,5-7, 10-14
A	page 1543, right-hand column - page 1544, left-hand column	3,4,8,9

1