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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 February 2009 (26.02.2009)

PCT

(10) International Publication Number
WO 2009/026271 A1

(54) Title: CONSISTENT USER EXPERIENCE IN INFORMATION RETRIEVAL SYSTEMS

(57) Abstract: An information retrieval system for searching a corpus is config-ured to operate in a manner that optimizes the consistency of a user experience given a subset of a corpus and a search query.

FIG. 1

ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL,
NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG,
CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

CONSISTENT USER EXPERIENCE IN
INFORMATION RETRIEVAL SYSTEMS

Cross-Reference to Related Applications

[001] This application claims the benefit of U.S. Provisional Application No. 60/956,764 filed August 20, 2007, the contents of which are incorporated herein in its entirety.

Background

[002] This description relates to techniques for providing a consistent user experience in information retrieval systems.

[003] In an information retrieval system (e.g., a wordspotting system), to search a corpus of documents (e.g., set of audio files) one enters a query, typically in the form of a search term or Boolean expression of search terms. Then the system returns a set of putative hits (search results) based on the search query. Each putative hit is usually associated with a score, or confidence measure, (e.g., a score ranging from 0 to 1, with 1 being "better" than 0) which allows the set of putative hits to be ranked (e.g., a list of putative hits, sorted by descending score, can be constructed) and it also allows a threshold to be applied to the set, i.e., the set is partitioned in two disjoint subsets: one subset containing putative hits equal or above a particular threshold, and the other subset containing the putative hits below that threshold. In some systems, the threshold is fixed at a particular value for all queries, whereas in other systems, the threshold is adjusted for each query.

Summary

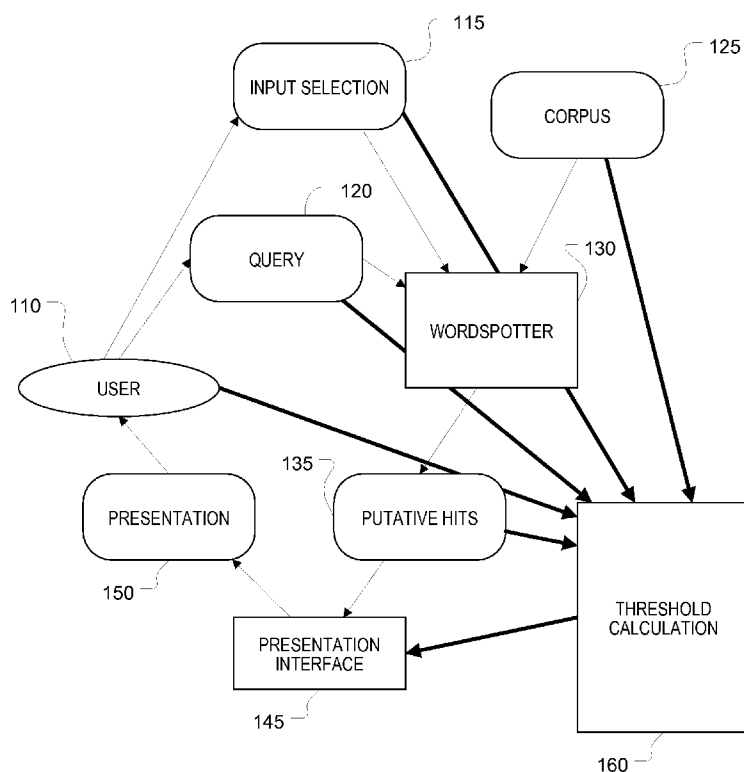[004]   In one aspect, in general, an information retrieval system for searching a corpus is configured to operate in a manner that optimizes the consistency of a user experience given a subset of a corpus and a search query.

[005] Aspects may include one or more of the following.

[006] The system is further configured to generate a confidence score for putative hits from the retrieval system, and display results above a threshold to maintain consistency by some criterion.

[007]   The consistency criterion comprises a constant number of false alarms, a constant false alarm rate, a constant number of true positives, a constant precision value, a constant rate of true positives, a constant number of misses, a constant number of false negatives, and/or a constant recall value.

[008]   The information retrieval system comprises a wordspotter.

[009]   The active corpus comprises a corpus defined by metadata.

[010]   The retrieval system is further configured to modify the search algorithm to optimize the consistency criterion.

[011]   The system is configured to account for an audio quality of the active corpus in optimizing the consistency criterion.

[012]   The consistency criterion comprises a constant computation time, a constant amount of computation, or a constant amount memory used.

[013]   In another aspect, in general, a system is configured to receive a search of a corpus from a user, present the user with a list of putative hits above a certain threshold, receive a modification to the active subset of the corpus from the user, dynamically adjust the threshold, and present to the user a new list of putative hits based on the adjusted threshold.

[014]   Other features and advantages of the invention are apparent from the following description, and from the claims.

Description of Drawings

[015]   FIG. 1 is a block diagram of a word spotting system.

Description

[016]   In an information retrieval system such as a wordspotting system, a user enters a query to search a corpus of documents such as a set of audio files. The system returns a set of putative hits (search results) where each putative hit has an associated score, or confidence measure, which allows the set of putative hits to be ranked. A threshold based on the score can be used to partition the results and/or to determine which of the results to display.

[017]   In general, scores for putative hits can depend on a number of factors, including without limitation the nature of the query (e.g., the phoneme length, how typical the phoneme patterns are in the language), the nature of the audio source (e.g., the degree to which they match the training material for the

wordspotting system in terms of acoustic channel, such as noise level or non-linear distortion, and match of the language, such as the type of accent or dialect in the audio files), and whether the putative hits correspond to true occurrences or if they correspond to false alarms.

[018]   In general, ranking putative hits according to their score results in true hits (detections of true occurrences of the query) being typically ranked more highly than false alarms (detections that do not correspond to true occurrences). A threshold can be set such that putative hits with scores above the threshold are presented to a user. Setting of the threshold (in conjunction with the distributions of the scores themselves) affects a number of characteristics of the set of putative hits that are presented to the user. These characteristics include a total number of putative hits presented to the user, the ratio of true hits to the number of putative hits presented (precision), the ratio of the number of true detections to the total number of true occurrences in the data (recall), the number of false alarms in the presented putative hits per unit of input data (e.g., per call, per unit of time, such as per hour or call time or active speech time), or the ratio of true detections versus false alarms. Note that some of these quantities may also be affected by the number of true query events that occur in the data being searched. For example, in an extreme in which there are no true occurrences of a query in the data, then quantities such as precision are necessarily zero.

[019]   The quality of a user experience is, in general, dependent on one or more of the quantities described above, which are in turn dependent on the setting of a score threshold for the presentation of putative hits.

[020]   In some implementations of the information retrieval system, the score threshold for presentation of putative hits is dynamically adjusted to maintain a consistent user experience across different active corpora, for example, by adjusting the thresholds for each presentation of search results. The system adjusts the threshold based on the query and/or the size and nature of the active corpus, in order to maximize the consistency of the user experience. The user experience of search results refers to user's perception of a presentation of a set of results to that user. For example, in a computer interface, a user experience of the search results occurs when the user is presented with a list of putative hits and/or when the putative hits are displayed as tick marks along a timeline representing an audio file. The active corpus refers to the set of documents currently part of the user experience, e.g., the subset of documents of the overall corpus that are active or visible at

any particular time. For example, when entering a new query, the set of active documents is often the entire corpus, but when the user selects a single document (e.g., during playback of an audio file), the set of active documents becomes a singleton set containing that document only. The consistency of the user experience refers to presenting a user with a result set that optimizes a particular criterion, such as constant absolute number of false positives, constant rate of false positives, constant absolute number of true positives, or constant rate of true positives.

[021]   Inside many audio corpora, there is an underlying structure. For example, a call center corpus will consist of individual phone calls, while a corpus of news broadcasts can be structured by shows, and further subdivided into stories within a broadcast. In many cases, search results often are clustered within these subdivisions. For this reason, one might search an entire corpus of thousands of hours with a high threshold (to ensure high precision) to find a few documents of interest. Then the user can analyze the individual documents, which will have much shorter durations. Because of this much shorter duration, the threshold can be lowered to improve the recall while still maintaining an acceptable number of false positives.

[022]   For example, if the consistency criterion is constant absolute number of false positives (e.g., the user can only tolerate up to 3 false alarms per experience of search results), when the user first enters a query to search the entire corpus, the system may set a high threshold (e.g., 0.975). But when the user is visualizing the search results for a single file, e.g., when the user selects that file to be played back, the threshold for that same query might be lowered (e.g., the threshold could be lowered to 0.834). The change in the threshold maintains consistency in the user experience because in both searches the user receives search results that contain up to 3 false alarms. When the large corpus is searched the threshold is higher because only up to 3 false alarms are allowed in the entire corpus. When the single file is searched the threshold is lowered because the search results can contain up to three false alarms for that single file.

[023]   While in the example above the threshold was decreased when the corpus was limited, in some examples reducing the active media set may result in an increase in the threshold. For example, if the selected file is of lesser audio quality than the entire corpus as a whole.

[024] In some implementations, if the consistency criterion is constant response time, the information retrieval system can change the search algorithm to reduce the computation required over a large corpus.

[025] Referring to FIG. 1, in an example of a wordspotting system is controlled by a user 110. The user provides a query 120, such as a key phase in text form, a Boolean query with multiple terms, or a voice input of a keyphrase. The user also may select a particular portion 115 of the available input. For example, an audio corpus 125 may include thousands of recorded telephone calls, but the user is only interested in a particular subset, for example, involving a particular agent or being associated with a particular inbound telephone number. Based on the query 120 and the selected portion 115 of the input corpus 125, a wordspotter 130 generates a set of putative hits 135, each associated with a score. A presentation interface 145 selects those putative hits that are above a threshold score and forms a presentation 150 that is perceived by the user 110.

[026] As introduced above, one aspect of providing a consistent user interface to the user relates to setting the threshold to be used by the presentation interface 145. In general, in various examples, a threshold calculation component 160 determines the appropriate threshold based on one or more factors. Some examples of these factors are as follows.

[027] A distribution of scores of low-scoring putative hits 135 can be a factor that provides the threshold calculation with a way to estimate the distribution of scores for false alarms for the selected part of the corpus. The input selection 115 can be a factor that is used to calculate the amount of input speech being processed, for example, in setting a consistent number of false alarms per unit input (e.g., per call, per hour, etc.). The query 120 itself can be a factor for the threshold calculation, for example, according to its phonetic content. The user can also be a factor, for example, according to explicit user preferences (e.g., a desired maximum false alarm rate) or according to implicit characteristics (e.g., a novice user may have lower tolerance for false alarms than an expert user).

[028] Note that in some examples, the process is iterative. For example, a user 110 may first select an entire corpus for analysis using a first query. Based on the results, the user may select a subset of the input, for use with that same first query or a new second query. A further factor that may be relevant to the threshold calculation component 160 is that the threshold may depend

where in such an iteration the user is. For example, as the user iterates and hones in on desired data, the threshold may be lowered to provide more and more information to the user on smaller and smaller subsets of the corpus.

[029]   In some examples, the input selection 115 is performed after the execution of the wordspotter 130 in the sense that the user may have a graphical display that permits "zooming in" on a particular part of the analyzed input. In such a case, the threshold calculation may be recomputed as the user zooms in, for example, lowering the threshold to maintain a constant expected number of false alarms per displayed time interval (e.g., 3 false alarms per graphical screen width). In some examples, displays with different resolutions may be provided concurrently to the user, and each display may have a different threshold. For example, an overall timeline may have a "magnifying glass" that shows more resolution, with that increased resolution may correspond to a different (e.g., lower) threshold.

[030]   In some implementations, threshold generation is based on finding a threshold to get a constant false alarm rate. Such a threshold can be difficult to determine since putative hit lists may or may not include an unknown number of true hits. However, in some instances the statistics of higher scoring false alarms can be predicted from the lower scores. This can be useful since one can assume for the more numerous lower scoring putative hits that the results are primarily false alarms. Using this data a statistical model of the distribution of scores of false alarms is constructed. With this model, one can calculate thresholds that will produce the desired characteristics.

[031]   In some examples, the approach is to take the top 140 scores per hour. The distribution of false alarm scores is then modeled by a shifted exponential distribution:

$$p\left(R; F_A, F_{Min}\right) = \begin{cases} 0 & , \quad R < F_{Min} \\ -F_A \exp\left(F_A\left(R - F_{Min}\right)\right) & , \quad R \geq F_{Min} \end{cases},$$

where R is the score, and the parameters F_A and F_{Min} are dependent on the audio and the search term.

[032]   In order to estimate these parameters for a given term, a search is performed on either the actual search data, or some audio that is representative of the speech of interest. Once these search results are collected, an estimation scheme (e.g., Maximum Likelihood (ML) estimation) is used to choose the model parameters.

[033] In an example using a Maximum Likelihood (ML) estimator the observed scores are represented as the vector **R**. The parameters of the score distribution are then given by:

$$\hat{F}_{Min} = \min\left(\mathbf{R}\right)$$

$$\hat{F}_A = \frac{N}{\sum\limits_{i=1}^{N}\left(\hat{F}_{Min} - R_i\right)} \cdot$$

[034] One issue with this estimator is that it is susceptible to outliers and numerous true hits. For these reason, we found that an estimator based on rank-order statistics to be more suitable:

$$\hat{F}_{Min} = \min\left(\mathbf{R}\right)$$

$$\hat{F}_A = \frac{\ln(2)}{\hat{F}_{Min} - median\left(\mathbf{R}\right)} \cdot$$

[035] We are interested in the number of false alarms above a threshold, denoted by λ, so this can be expressed as a function of the cumulative density function (CDF):

$$\begin{aligned}
\lambda\left(R; F_A, F_{Min}\right) &= L \cdot \left(1 - P\left(R; F_A, F_{Min}\right)\right) \\
&= L \cdot \left(1 - \left(1 - \exp\left(- F_A F_{Min}\right)\exp\left(F_A R\right)\right)\right) \\
&= L \cdot \exp\left(- F_A F_{Min}\right)\exp\left(F_A R\right) \\
&= \exp\left(F_A R + \ln(L) - F_A F_{Min}\right)
\end{aligned}$$

where L is the rate of putative hits sampled (typically 140 Results/hour), and P is the CDF. We then define a new variable F_B and write the false alarm function as:

$$F_B = F_A F_{Min} + \ln(L)$$

$$\lambda\left(R; F_A, F_{Min}\right) = \exp\left(F_A R + F_B\right)$$

[036] Given a desired false alarm rate λ, a new threshold (T) can be set according to:

$$T(\lambda) = \frac{\ln(\lambda) - F_B}{F_A}$$

[037] Although the method described above performed well in many examples, other implementations are can also be used. For example, some

other implementations may provide higher accuracy in the case of common terms in a query. The use of the rank-order statistics based estimator mitigates many of the problems associated with true hits occurring in the model training data. However, if there is an extremely common phrase (more than 30 occurrences per hour), then this can result in thresholds that are much higher than desired. However, in many cases these very common phrases can be detected since the results have dramatically different distributions from results that consist only of false alarms.

[038] In these implementations, we introduce $\hat{H}^{(k)}$, which is the estimated number of true hits in the audio. This value is estimated iteratively along with updates for $\hat{F}_A^{(k)}$, where the superscript k represents the iteration number. This procedure is repeated for a fixed number of iterations, which is 10 in our implementation.

[039] The model parameter update is given by:

$$\hat{F}_{Min} = \min(\mathbf{R}); \quad \hat{F}_A^{(k)} = \ln(2) / \left(\left(\mathbf{R}_{N-\hat{H}^{(k)}/2} \right) - \hat{F}_{Min} \right)$$

where N is the number of raw scores in R, and $\mathbf{R}_{N-\hat{H}^{(k)}/2}$ is an order statistic. The iterations are initialized with $\hat{H}^{(0)} = 0$. Estimation of $\hat{H}^{(k)}$ is given by the following pseudocode with the assumption that that the raw scores **R** have been sorted

for i=0:*N-1* **do**
$$S = \hat{F}_A^{(k)} \left(\mathbf{R}_i - \hat{F}_{Min} \right)$$
$$C = N - i$$
$$Diff = C - N \exp(S)$$
if $(S > 2)$ **then**
$$\hat{H}^{(k)} = \max\left(\hat{H}^{(k)}, Diff \right)$$

[040] This method simply looks at the higher scores and finds the point of maximum deviation from the statistical model. In this process, the underlying false alarm parameters are more robust and an estimate for the true hits is also generated.

[041] It is to be understood that the foregoing description is intended to illustrate and not to limit the scope of the invention, which is defined by the scope of the appended claims. Other embodiments are within the scope of the following claims.

What is claimed is:

1.   An information retrieval system for searching a corpus, the information
retrieval system being configured to:

      operate in a manner that optimizes the consistency of a user experience
          given a subset of a corpus and a search query.

2.   The method of claim 1, wherein the system is further configured to:

      generate a confidence score for putative hits from the retrieval system;
          and

      display results above a threshold to maintain consistency according to
          a consistency criterion.

3.   The method of claim 2, wherein the consistency criterion comprises a
constant number of false alarms.

4.   The method of claim 2, wherein the consistency criterion comprises a
constant false alarm rate.

5.   The method of claim 2, wherein the consistency criterion comprises a
constant number of true positives.

6.   The method of claim 2, wherein the consistency criterion comprises a
constant precision value.

7.   The method of claim 2, wherein the consistency criterion comprises a
constant number of misses.

8.   The method of claim 2, wherein the consistency criterion comprises a
constant recall value.

9.   The method of claim 1, wherein the information retrieval system
comprises a wordspotter.

10.   The method of claim 1, wherein the active corpus comprises a corpus defined by metadata.

11.   The method of claim 1, wherein the retrieval system is further configured to modify the search algorithm to optimize the consistency criterion.

12.   The method of claim 14, wherein the system is configured to account for an audio quality of the active corpus in optimizing the consistency criterion.

13.   The method of claim 14, wherein the consistency criterion comprises a constant computation time.

14.   The method of claim 14, wherein the consistency criterion comprises a constant amount of computation.

15.   The method of claim 14, wherein the consistency criterion comprises a constant amount memory used.

16.   A system configured to:

        receive a search of a corpus from a user;

        present the user with a list of putative hits above a certain threshold;

        receive a modification to the active subset of the corpus from the user;

        dynamically adjust the threshold; and

        present to the user a new list of putative hits based on the adjusted
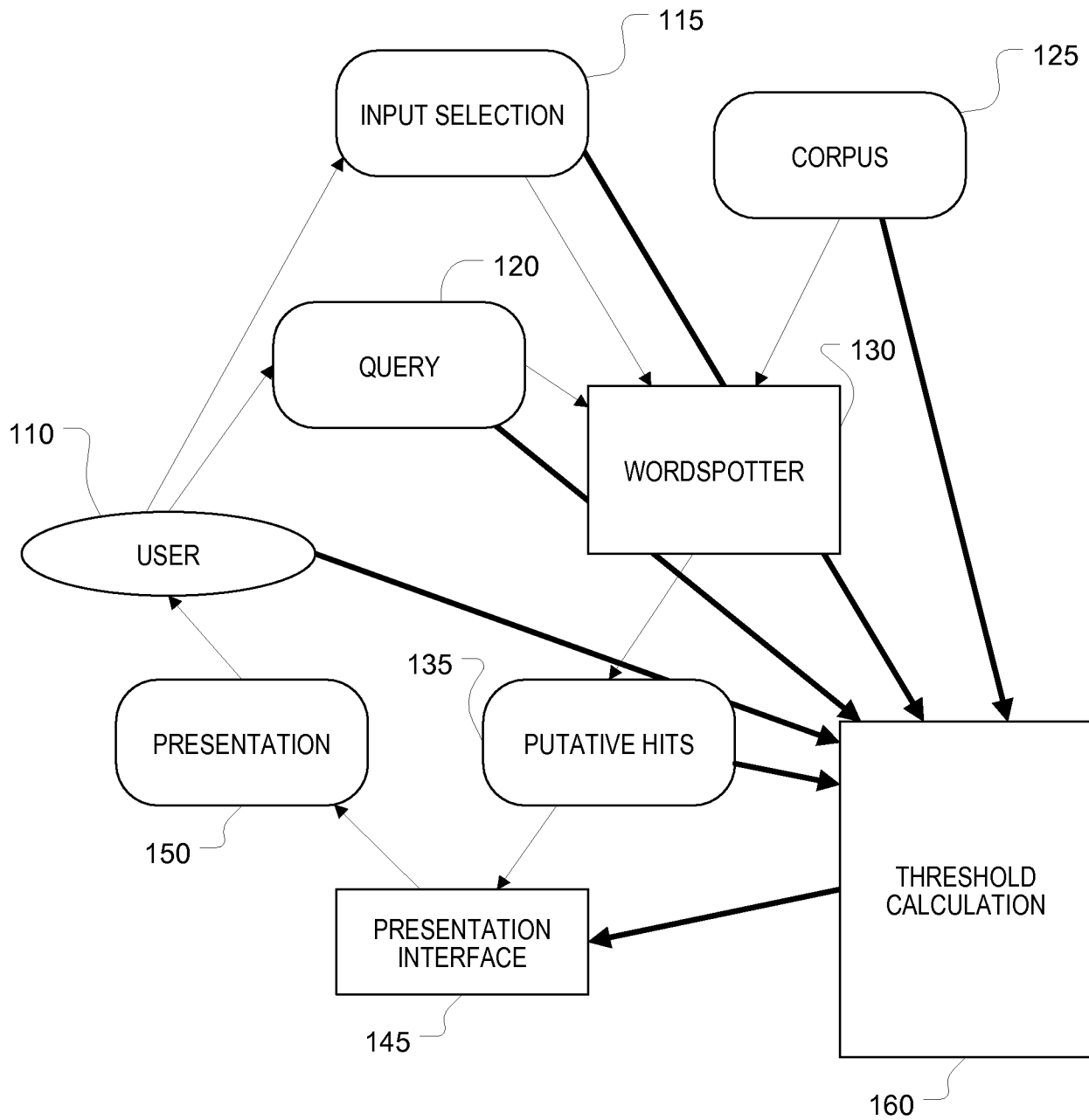   threshold.

48103

FIG. 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 08/73564

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 17/30 (2008.04)
USPC - 707/3

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)
USPC: 707/3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 707/1, 3, 6, 7, 10; 700/1, 90, 91, 92

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic Databases: PubWEST (PGPB,USPT,USOC,EPAB,JPAB); GoogleScholar
Search Terms Used: putative hits, search results, wordspotter, wordspotting, sub-corpus, information retrieval system, query, consistent user interface, threshold, metadata, meta-data, metatag, true dectection, false alarm/positive

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2007/0112740 A1 (GEVA) 17 May 2007 (17.05.2007) Entire document, especially Abstract, para [0023]-[0024]	16
Y	US 2007/0033003 A1 (MORRIS) 08 February 2007 (08.02.2007) Entire document, see Abstract, para [0012], [0042], [0050], [0053]	1-15
Y	US 2006/0206324 A1 (SKILLING et al.) 14 September 2006 (14.09.2006) Entire document, especially para [0016], [0032], [0075]	1-15
A	US 2007/0185831 A1 (CHURCHER) 09 August 2007 (09.08.2007)	1-16

☐ Further documents are listed in the continuation of Box C. ☐

*	Special categories of cited documents:	"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
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier application or patent but published on or after the international filing date	"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
"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)	"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
"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	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 October 2008 (23.10.2008)	05 NOV 2008

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)