Internal distribution code:

(A) [-] Publication in OJ
(B) [-] To Chairmen and Members
(C) [-] To Chairmen
(D) [X] No distribution

Datasheet for the decision of 21 June 2016

Case Number: T 1902/10 - 3.5.07

Application Number: 02012929.2

Publication Number: 1241594

IPC: G06F17/30

Language of the proceedings: EN

Title of invention:
System and method for locating pages on the world wide web and for locating documents from a network of computers

Applicant:
Yahoo! Inc.

Headword:
Web page indexing/YAHOO

Relevant legal provisions:
EPC Art. 56, 84, 123(2)

Keyword:
Amendments - allowable (no) (main request)
Claims - support in the description (no) (first auxiliary request)
Inventive step - (yes) (second auxiliary request)

Decisions cited:

Catchword:

Europäisches Patentamt
European Patent Office
Office européen des brevets

Beschwerdekammern
Boards of Appeal
Chambres de recours

European Patent Office
D-80298 MUNICH
GERMANY
Tel. +49 (0) 89 2399-0
Fax +49 (0) 89 2399-4465

Case Number: T 1902/10 - 3.5.07

D E C I S I O N
of Technical Board of Appeal 3.5.07
of 21 June 2016

Appellant: (Applicant)	Yahoo! Inc. 701 First Avenue Sunnyvale, CA 94089 (US)
Representative:	Boult Wade Tennant Verulam Gardens 70 Gray's Inn Road London WC1X 8BT (GB)
Decision under appeal:	**Decision of the Examining Division of the European Patent Office posted on 1 April 2010 refusing European patent application No. 02012929.2 pursuant to Article 97(2) EPC.**

Composition of the Board:

Chairman	R. Moufang
Members:	M. Rognoni
	R. de Man

Summary of Facts and Submissions

I. The former applicant, "Alta Vista Company", (former appellant) appealed against the decision of the Examining Division to refuse the European patent application No. 02012929.2.

1.1 In the contested decision the Examining Division arrived, *inter alia*, at the following conclusions:

- claims 1 and 6 of the main request filed during the oral proceedings on 25 February 2010 did not comply with Article 123(2) EPC;

- claims 1 and 6 according to the first auxiliary request filed during the oral proceedings lacked essential features (Article 84 EPC);

- the subject-matter of claims 1 and 6 according to the second auxiliary request filed during the oral proceedings lacked inventive step (Article 56 EPC) in view of the following prior art:

D1: Eichmann, D.: "The RBSE Spider - Balancing Effective Search Against Web Load", Computer Networks and ISDN Systems, May 1994;

D3: Fagin, R. *et al.*: "Extendible Hashing - A Fast Access Method for Dynamic Files", ACM Transactions on Database Systems, Vol. 4, No. 3, September 1979, pages 315 to 344.

II. With the statement of grounds of appeal, the former appellant filed three sets of claims for a main request and first and second auxiliary requests, respectively, and requested that a patent be granted in accordance

with the main request or, should the main request not be allowable, in the form of the first or second auxiliary request.

In the statement of grounds of appeal, the former appellant furthermore pointed out that its requests were the same as those on file at the end of the oral proceedings before the Examining Division, with the exception that claims 1 and 6 of each request had been amended to require that the web information file was stored in secondary storage.

III. A change of the applicant from "Alta Vista Company" to "Yahoo! Inc.", due to a transfer and a merger, was registered with effect from 3 December 2015. Yahoo! Inc. thereby acquired the status of appellant.

IV. With letter dated 2 March 2016, the appellant was summoned to oral proceedings scheduled to take place on 21 June 2016.

V. In a communication pursuant to Article 15(1) RPBA dated 23 March 2016, the Board drew the appellant's attention to the following prior art:

D6: US-A-5 204 958.

Furthermore, the Board expressed the following preliminary opinions:

- the feature *"wherein the fingerprint value is a compressed encoding of the URL of a corresponding web page"* in claim 1 of the main request extended beyond the content of the application as originally filed (Article 123(2) EPC;

- some essential features of the invention appeared not to be clearly specified in claim 1 of the first auxiliary request (Article 84 EPC);

- in the light of the prior-art documents D1 and D6 and of common general knowledge, it would appear obvious to a person skilled in the art to arrive at a system falling within the terms of claim 1 according to the second auxiliary request (Article 56 EPC).

VI. With letter dated 7 April 2016, the appellant requested that the oral proceedings scheduled for 21 June 2016 be postponed.

VII. In a communication dated 29 April 2016, the Board noted that the appellant had not provided sufficient justification for the postponement of the oral proceedings and had not complied with the requirement to file its request as soon as possible (cf. Notice in Special edition No. 3 OJ EPO 2007, 115 (point 2.)). Hence, the appellant's request could not be acceded to and the oral proceedings would be held as scheduled on 21 June 2016.

VIII. In response to the Board's preliminary opinion, the appellant, with letter dated 18 May 2016, maintained the previous requests and submitted three additional sets of claims as third, fourth and fifth auxiliary requests, respectively. The appellant did not comment on the Board's objections against the main request and the first auxiliary request, but argued in support of the inventive step of the subject-matter of claims 1 and 6 of the second auxiliary request, having regard in particular to the combination of documents D1 and D6.

IX. With letter dated 14 June 2016, the appellant, *inter alia*, informed the Board that it would not be attending the oral proceedings scheduled for 21 June 2016.

X. Oral proceedings were held as scheduled on 21 June 2016 in the absence of the appellant. At the end of the oral proceedings, the Chairman pronounced the Board's decision.

XI. The appellant requested that the decision under appeal be set aside and a patent be granted on the basis of the main request or, if that was not possible, on the basis of one of the first to fifth auxiliary requests.

Furthermore, in the letter dated 14 June 2016, the appellant requested that, if the Board found the appellant's arguments persuasive for at least one of the requests on file, but believed that there remained further issues, then the application be remitted to the Examining Division for further prosecution in writing.

XII. Claim 1 of the appellant's main request reads as follows:

"A system for locating web pages stored on remotely located computers connected by a network, each web page having a unique URL (universal resource locator), at least some of said web pages including URL links to other ones of the web pages, the system comprising:

a communications interface connected to the network for fetching specified ones of the web pages from said remotely located computers in accordance with corresponding URLs;

a web information file stored in secondary memory having a set of entries, each entry denoting, for a corresponding web page, a URL and fetch status information;

characterised in that

a web information table, stored in RAM (random access memory), having a set of entries, each entry denoting fetch status information for the corresponding web page, and a fingerprint value wherein the fingerprint value is a compressed encoding of the URL of a corresponding web page; and

a web procedure, executed by the system, for fetching and analyzing web pages, said web procedure including instructions for fetching web pages whose web information file entries meet predefined selection criteria based on said fetch status information, for determining for each URL link in each received web page whether a corresponding entry already exists in the web information table, and for each URL link which does not have a corresponding entry in the web information table adding a new entry in the web information table and a corresponding new entry in the web information file."

Claims 2 to 5 are directly or indirectly dependent on claim 1. Claim 6 is directed to a method of locating web pages stored on remotely located computers. Claims 7 to 10 are directly or indirectly dependent on claim 6.

Claim 1 according to the first auxiliary request differs from claim 1 of the main request in that the term "compressed" in the first feature of the characterising portion has been replaced with the term "hashed".

Claim 1 according to the second auxiliary request reads as follows:

"A system for locating web pages stored on remotely located computers connected by a network, each web page having a unique URL (universal resource locator), at least some of said web pages including URL links to other ones of the web pages, the system comprising:

a communications interface connected to the network for fetching specified ones of the web pages from said remotely located computers in accordance with corresponding URLs;

a web information file stored in secondary memory having a set of entries, each entry denoting, for a corresponding web page, a URL and fetch status information;

characterised in that

a web information table, stored in RAM (random access memory), having a set of entries, each entry denoting fetch status information for the corresponding web page, a fingerprint value wherein the fingerprint value is a hashed encoding of the URL of a corresponding web page and a file location value that indi[c]ates the location of a corresponding entry in the web information disk file; and

a web procedure, executed by the system, for fetching and analyzing web pages, said web procedure including instructions for fetching web pages whose web information file entries meet predefined selection criteria based on said fetch status information, for determining for each URL link in each received web page whether a corresponding entry already exists in the web information table by generating a fingerprint of each URL, determining where in the web information table an entry having the generated finger print value would be stored, determining if that entry is stored in the web information table, returning a failure value if the entry is not found, and returning a success value, a fetched flag and disk position value if the entry is found in the web information table, and for each URL link which does not have a corresponding entry in the web information table adding a new entry in the web information table by generating a fingerprint value of each URL, determining where in the web information table an entry having that fingerprint value should be stored, storing an entry in the web information table at the determined position with a fetch flag indicating the web page has not yet been fetched, the fingerprint value and a disk file position and a corresponding new entry in the web information file."

Claims 2 to 5 are dependent on claim 1.

Claim 6 of the second auxiliary request reads as follows:

"A method of locating web pages stored on remotely located computers connected by a network, each web page having a unique URL (universal resource locator), at least some of said web pages including URL links to other ones of the web pages, comprising the steps of:

storing a web information file in secondary memory having a set of entries, each entry denoting, for a corresponding web page, a URL and fetch status information;

characterised in that

storing in RAM (random access memory) a web information table having a set of entries, each entry denoting fetch status information for a corresponding web page, a fingerprint value, wherein the fingerprint value is a hashed encoding of the URL and a file location value that indi[c]ates the location of a corresponding entry in the web information disk file; and

executing a web procedure, system [*sic*] for fetching and analyzing web pages, including (A) sequentially scanning entries in the web information file to determine which of said entries meet predefined selection criteria, (B) fetching web pages whose web information file entries meet said predefined selection criteria, (C) determining for each URL link to another web page in each received web page whether a corresponding entry already exists in the web information table by generating a fingerprint of each URL, determining where in the web information table an entry having the generated finger print value would be stored, determining if that entry is stored in the web

information table, returning a failure value if the entry is not found, and returning a success value, a fetched flag and disk position value if the entry is found in the web information table, and (D) for each URL link which does not have a corresponding entry in the web information table adding a new entry in the web information table by generating a fingerprint value of each URL, determing [*sic*] where in the web information table an entry having that fingerprint value should be stored, storing an entry in the web information table at the determined position with a fetch flag indicating the web page has not yet been fetched, the fingerprint value and a disk file position and a corresponding new entry in the web information file."

Claims 7 to 10 are dependent on claim 6.

The third to fifth auxiliary requests are not relevant to the Board's decision.

XIII. The appellant's arguments relevant to the Board's decision are summarised in the "Reasons" (see paragraphs 4.1, 4.2, 7.1 and 18. to 18.6 below).

Reasons for the Decision

1. The appeal is admissible.

The invention

2. The present application relates to a system and a method for quickly locating and analysing web pages on the World Wide Web. As explained in the description (see page 2, lines 15 to 29 of the application as

originally filed), known systems for locating pages on the Web, that is "Web crawlers", generally start with a root set of known web pages and create a disk file with a distinct entry for every known web page. As additional web pages are fetched and their links to other pages are analysed, additional entries are made in the disk file to reference web pages not previously known to the Web crawler. The information about web pages already processed is generally stored in a disk file, because the amount of information in the disk file is too large to be stored in random access memory (RAM).

2.1 The first paragraph of page 3 of the description relates to the disk input/output operations incurred when processing one web page in order to find out if records for the references contained in this web page already exist in the web information disk file. Given an assumed limitation of 50 disk seeks per second, the present application concludes that only about one typical web page can be processed per second (*ibid.* page 3, lines 27 to 31). Network latency also tends to limit the number of pages that Web crawlers can process within a given time interval.

2.2 The present invention aims at providing a system and a method for quickly locating and making a directory of web pages on the World Wide Web.

2.3 The gist of the present invention consists essentially in providing a Web crawler system which includes a hash table stored in random access memory (RAM) and a sequential disk file (or web information disk file) stored in secondary memory.

For every web page known to the system, the Web crawler system stores an entry in the sequential disk file and a smaller entry in the hash table. The hash table entry includes a fingerprint value that is unique to the corresponding web page, a one bit "fetched flag" that indicates whether or not the corresponding page has been fetched and analysed, and a file location value that indicates the location of a corresponding entry in the web information disk file.

As every unique URL corresponding to a disk file entry is mapped into a similarly unique fingerprint value stored in RAM, a Web crawler needs only to check the table in RAM to verify whether a web page has already been processed.

Main request

3. Claim 1 according to the main request relates to a *"system for locating web pages stored on remotely located computers connected by a network".* Each web page has *"a unique URL (universal resource locator)"*, and *"at least some of said web pages includ*[e] *URL links to other ones of the web pages".*

The claimed system comprises the following features itemised by the Board:

(a) a communications interface connected to the network for fetching specified ones of the web pages from said remotely located computers in accordance with corresponding URLs;

(b) a web information file stored in secondary memory having a set of entries,

(i) each entry denoting, for a corresponding web page, a URL and fetch status information;

(c) a web information table, stored in RAM (random access memory), having a set of entries,
(i) each entry denoting fetch status information for the corresponding web page,
(ii) and a fingerprint value wherein the fingerprint value is a compressed encoding of the URL of a corresponding web page; and

(d) a web procedure, executed by the system, for fetching and analyzing web pages,
(i) said web procedure including instructions for fetching web pages whose web information file entries meet predefined selection criteria based on said fetch status information,
(ii) for determining for each URL link in each received web page whether a corresponding entry already exists in the web information table, and
(iii) for each URL link which does not have a corresponding entry in the web information table adding a new entry in the web information table and a corresponding new entry in the web information file.

Article 123(2)

4. In the contested decision, the Examining Division held that the term "*compressed*" in the expression "*compressed encoding of the URL of a corresponding web page*" (see feature (c)(ii) of claim 1) was an apparent generalisation of the disclosure on page 10 (mistakenly

identified as page 19 in the decision), line 20 of the original description which clearly indicated that the compressed encoding was a hashing and nothing else. In fact, general compressed encoding would not necessarily provide the technical effects of a hashing indexing method (fast retrieval) or even imply the features of a hashing indexing method. Hence, in the Examining Division's opinion, this amendment was not directly and unambiguously derivable from the original application.

4.1 In the statement of grounds of appeal, the appellant essentially argued that a purpose of encoding the URLs was to enable them to be stored in a primary storage for fast access and checking. The size of tables in prior-art Web crawlers forced them to be stored in a secondary memory, making access slow. The currently claimed invention solved this problem by compressing the URL, thereby reducing the information to a size that could be stored in the fast primary memory. This was explained at lines 26 to 30 of page 18 as being one way in which the invention improved the speed of prior art Web crawlers. A specific example of a compression method using a hash table was given from line 19 of page 9, but the skilled person understood that this was only one way in which compression of the URL could be achieved and that the invention extended to the principle of compressing URLs by any suitable means. In other words, performance was improved by the compression part of the exemplary hashing technique alone. Indexing, as provided by a hashing technique, was not required.

4.2 Hence, the skilled person reading the application would understand that the invention provided improved speeds by compressing the URLs to allow the web information table to be stored in a primary storage, and would

directly and unambiguously derive that the compressing could be achieved by any suitable means and was not restricted to hashed encoding.

5. According to the summary of the invention (see application as originally filed, page 4, line 35 to page 5, line 10), "*[t]he Web crawler system includes a hash table stored in random access memory (RAM) and a sequential file (herein called the "sequential disk file" or the "Web information disk file") stored in secondary memory, typically disk storage. For every Web page known to the system, the Web crawler system stores an entry in the sequential disk file as well as a smaller entry in the hash table. The hash table entry includes a fingerprint value, a fetched flag that is set true only if the corresponding Web page has been successfully fetched, and a file location indicator that indicates where the corresponding entry is stored in the sequential disk file. Each sequential disk file entry includes the URL of a corresponding Web page, plus fetch status information concerning that Web page*" (underlining added).

Furthermore, it is specified on page 9, line 33 to page 10, line 2 that, "*[w]hile the exact size of the hash table entries is not important, it is important that each hash table entry 160 is significantly smaller (e.g., at least 75% smaller on average) than the corresponding disk file entry*".

In the section "Alternative Embodiments" (page 18, lines 19 to 22), it is pointed out that "*[a]ny data structure that has the same properties of the Web information hash table 130, such as a balanced tree, a skip list, or the like, could be used in place of the*

hash table structure 130 of the preferred embodiment" (underlining added).

5.1 According to the Wikipedia definition, "*a **hash table** (**hash map**) is a data structure used to implement an associative array, a structure that can map keys to values*" (underlining added).

On the other hand, "***data compression, source coding,** or **bit-rate reduction** involves encoding information using fewer bits than the original representation*" (see Wikipedia - underlining added).

In other words, the application teaches that a hash table has web information entries that correspond to the web information entries of a disk file and that the former are considerably smaller than the latter, although the exact size of the hash table entries is said to be not important (original application, page 9, lines 33 to 34). In the preferred embodiment given on page 9, lines 30 to 32, a fingerprint value of a hash table entry is 63-bits long. In fact, rather than the "smaller" size, the salient characteristic of fingerprint values is their "uniqueness" which ensures a direct correspondence between fingerprint values and URLs of web pages. In this case, verification of the existence of a URL in a disk file is effectively performed by searching for the corresponding fingerprint value in a hash table stored in RAM.

Even if a table of *"compressed"* URLs of web pages could advantageously be stored in RAM, a generic compression algorithm would not generate entries with a "data structure" having the same properties of the web information table of the present invention and thus would not constitute an implementation of the disclosed

teaching. Furthermore, the primary purpose of data compression is to encode information using fewer bits than the original representation, while preserving the initial information content which, in principle, can be recovered through decoding. As, according to the present invention, the original representation of the URLs is stored in the disk information file, there is no reason to use a compression algorithm for generating fingerprint values for a table which would then require some additional indexing to be made easily and quickly searchable.

5.2 In summary, the Board concurs with the Examining Division that feature (c)(ii) of claim 1 in the Board's itemisation extends beyond the content of the application as originally filed (Article 123(2) EPC).

5.3 Thus, the appellant's main request is not allowable under Article 123(2) EPC.

First auxiliary request

6. Claim 1 according to the first auxiliary request differs from claim 1 of the main request in that feature (c)(ii) has been amended as follows:

(c) a web information table, [...]

(ii') wherein the fingerprint value is a ~~compressed~~ *hashed* encoding of the URL of a corresponding web page.

7. According to the Examining Division (see contested decision, point 14.1), some of the essential features of accessing the web information table in RAM by means of a hashed URL for checking the existence of a URL or

for storing information regarding a new URL were not present in the independent claims of the request then on file.

Furthermore, the Examining Division considered that the original application consistently disclosed that the web information file was stored in the "secondary memory" or disk storage. In the application, a hash indexing method was employed as a solution to the problem generated by the fact that the secondary memory access time was a performance bottleneck when it came to accessing large files, as in the case of Web crawlers. Thus, storing web information files in the secondary memory was essential to the clear definition of the technical context of the invention.

7.1 As pointed out by the appellant, the latter objection has been addressed by the new first auxiliary request.

Furthermore, the appellant has essentially argued that a technical problem was solved by storing the web information table in RAM, due to the increased access speed. This solution was achieved by the use of hashed encoding, and did not rely on any interaction between the web information table and the web information file. Further improvements could be achieved beyond the basic system by utilising the full set of features set out in the description. However, there was no requirement for those features to be used in order to solve a technical problem of improving speed of operation.

8. Claim 1 according to the first auxiliary request specifies a system comprising a web information file stored in secondary memory (cf. feature (b) in paragraph 3.) and a web information table stored in RAM (cf. feature (c)). Each entry of the web information

table relates to a URL and fetch status information of a web page. Each entry of the web information table denotes the fetch status information of a web page and a hashed encoding of its URL. Claim 1 does not specify any link between these two features of the claimed system. Nor does it imply that entries of the web information table are directly related to entries in the web information file.

8.1 The last feature of claim 1 (see feature (d) in paragraph 3. above) describes a web procedure for fetching and analyzing web pages. It includes unspecified "instructions" for fetching web pages according to predefined selection criteria satisfied by the corresponding web information file entries.

The web procedure comprises the step of determining whether each URL link in a received page has a corresponding entry in the web information table, and the step of adding a new entry in the web information table and in the web information file for each URL link which is not represented in the web information table.

9. As pointed out above (see 7.1), the appellant has argued that the claimed system improved access speed simply by storing the web information table in RAM and did not rely on any interaction between the web information table and the web information file.

10. It is true that *"a Web page directory table is stored in RAM with sufficient information to determine which Web pages links represent new Web pages not previously known to the Web crawler, enabling received Web pages to be analyzed without having to access a disk file"* (application as filed, page 18, lines 26 to 30). Thus, in principle, no "link" between the web information

table and the web information file would be required for an analysis of the received web pages.

However, the description makes clear that there is a "link" and that it is provided by a file location value which *"indicates the location of a corresponding entry in the Web information disk file"* (page 9, lines 26 to 28). In fact, as specified on page 18, lines 7 to 11, the web information table is used as an index into the web information disk file, so that *"an entry in the Web information disk file is accessed by first reading the disk file address in the corresponding entry in the Web information hash table and then reading the Web information disk file entry at that address"* (page 18, lines 12 to 16).

10.1 In summary, claim 1 according to the first auxiliary request covers systems which do not involve any interaction between the web information table and the web information file, whereas the description consistently shows that each entry of the web information table provides a link to a corresponding entry in the web information file so that the RAM-based table provides information as to which URLs are stored in the disk file. As claim 1 does not reflect this important aspect of the invention, its subject-matter does not find full support in the original description (Article 84 EPC).

11. As to the features of the web procedure (see features (d)(i) and (ii) in section 3. above), the Board finds that it is unclear what kind of web procedure is actually used to determine whether an entry for a certain web link exists in the web information table and, in particular, whether the fingerprint value is

used to locate the entry relating to a web page in the web information table.

11.1 On the other hand, the description clearly teaches that a search for an entry in the web information table relating to a received web page is performed on the basis of a "unique fingerprint value" obtained by hash encoding the web page URL (see for instance application as filed, page 11, lines 15 to 26).

11.2 Hence, also as far as the web procedure is concerned, claim 1 covers systems which do not find full support in the description.

12. As claim 1 does not comply with Article 84 EPC, the appellant's first auxiliary request is not allowable.

Second auxiliary request

13. Claim 1 according to the second auxiliary request differs from claim of the main request in that features (c) and (d) read as follows (additions with respect to the main request are in italics, deletions in strikethrough):

(c) a web information table, stored in RAM (random access memory), having a set of entries,
(i) each entry denoting fetch status information for the corresponding web page,
(ii) a fingerprint value wherein the fingerprint value is a ~~compressed~~ *hashed* encoding of the URL of a corresponding web page *and*
(iii) *a file location value that indicates the location of a corresponding entry in the web information disk file*; and

(d) a web procedure, executed by the system, for fetching and analyzing web pages,

(i) said web procedure including instructions for fetching web pages whose web information file entries meet predefined selection criteria based on said fetch status information,

(ii) for determining for each URL link in each received web page whether a corresponding entry already exists in the web information table *by*

- *generating a fingerprint of each URL,*
- *determining where in the web information table an entry having the generated finger print value would be stored,*
- *determining if that entry is stored in the web information table,*
- *returning a failure value if the entry is not found, and*
- *returning a success value, a fetched flag and disk position value if the entry is found in the web information table,* and

(iii) for each URL link which does not have a corresponding entry in the web information table adding a new entry in the web information table *by*

- *generating a fingerprint value for each URL,*
- *determining where in the web information table an entry having that fingerprint value should be stored,*
- *storing an entry in the web information table at the determined position with a fetch flag indicating the web page has not yet been fetched, the fingerprint value and a disk file position* and a

corresponding new entry in the web information file.

14. As to feature (c)(ii) and (iii), it is specified on page 10, lines 15 to 19 of the application as filed that a fingerprint of a URL is computed by a hash table manager and that the corresponding fingerprint function is designed to ensure that every unique URL is mapped into a similarly unique fingerprint value.

Furthermore, according to the list given on page 9, lines 15 to 28, each entry in the web information table includes a fingerprint value and a file location value that indicates the location of a corresponding entry in the web information disk file.

14.1 The procedure according to feature (d)(ii) for determining whether the entry for a particular web page already exists in the web information table finds support, for instance, in the application as filed on page 10, line 25 to page 11, line 2.

14.2 The procedure according to feature (d)(iii) for adding a new entry in the web information table corresponds essentially to the procedure specified on page 11, lines 15 to 26.

14.3 Hence, claim 1 according to the second auxiliary request does not include subject-matter extending beyond the content of the application as originally filed (Article 123(2) EPC).

15. The Board is also satisfied that claim 1 of the second auxiliary request overcomes the objection under Article 84 EPC raised with respect to the first auxiliary request.

16. Document D1, which represents the closest prior art, relates to a system for locating web pages stored on remotely located computers connected by a network (see Abstract).

17. According to the contested decision, the subject-matter of claim 1 differed from the system known from document D1 in that it further comprised a web information table according to feature (c) of the Board's itemisation and a web procedure according to features (d)(ii) and (d)(iii).

17.1 Starting from the system according to document D1, the Examining Division identified the problem to be solved as achieving fast data retrieval.

According to the Examining Division, the person skilled in the art of information retrieval, confronted with the above technical problem, would choose hash indexing and a hash table as standard design implementation choice among the data structures commonly known in the art.

18. The appellant has pointed out that the Examining Division rejected the claims on the basis of a combination of D1 with common general knowledge and presented document D3 as evidence of this common general knowledge. However, a single research publication was insufficient evidence of common general knowledge. Thus, it had not been shown that the skilled person could, and in particular would, find in the common general knowledge all the features required to arrive at the claimed invention.

18.1 The appellant has furthermore submitted that, if document D3 was held to be representative of common general knowledge, then it provided reasons why the skilled person would not consider the claimed solution. On page 323 of D3, it was stated that hashing had usually been confined to tables which fitted into the main memory, and whose size could be estimated reliably. Neither of these criteria was met by the system of document D1 to which, according to the Examining Division, the skilled person would apply the hashing techniques of document D3 so as to arrive at the claimed invention. The tables of document D1 did not fit into the main memory, and the size could not be estimated reliably, as it would grow to an undefined limit if more web pages were explored. The skilled person was therefore actually led away from applying hash techniques to document D1.

18.2 According to the appellant, there were many possible ways in which the problem of providing faster data retrieval in the context of Web crawlers could be addressed, since there were many aspects to the slow performance of data retrieval in existing Web crawlers, as set out in the background of the invention. For example, the speed of prior-art systems could be improved by utilising faster systems or by developing larger-capacity RAM memory. As all these systems addressed the issue of slow data retrieval, the skilled person, starting from the system according to document D1, could choose to investigate any of these areas to find a solution. There was nothing which could prompt the skilled person to consider addressing the problem by modifying the data handling and storage processes.

18.3 The appellant has pointed out that in document D1 only a single database was utilised, which was accessed by

the Web spider and the indexer to conduct the process of retrieving and analysing web pages. The database of document D1 was equivalent to the web information file of the current system.

In order to arrive at the system according to claim 1, an entirely new and separate (but connected) database was created. Furthermore the database was stored in RAM, in contrast to the database of document D1 which was entirely disk-based. To enable operation of the new dual structure, the behaviour of the system had to be modified such that the web information table was checked for the presence of links and appropriate action taken depending on whether linking already existed. There was no suggestion in document D1 of any of these features and therefore no reason for the skilled person to consider the current solution.

18.4 The appellant has furthermore stressed that the hashing techniques referred to by the Examining Division were generally utilised to provide an efficient look-up system which allowed rapid identification of data location by means of hash keys. The current invention solved a technical problem by encoding URLs to reduce their size and thereby allow storage of a second information table in a RAM.

18.5 In the appellant's view, the present solution was very different from the solution which the skilled person would arrive at by applying standard hashing techniques to the system of document D1. The skilled person would actually modify the disk-based database structure of document D1 to include a hash key field, such that more rapid look-up of particular pages could be performed. This approach would be functional and would lead to a more rapid operation of a Web crawler as the time taken

to handle data in the database improved. This solution would be different from the claimed invention which relied on a separate database which stored in a RAM encoded URLs. It was this combination of URL encoding and storing of encoded URLs in a RAM which gave the currently claimed solution its advantages over the prior art.

18.6 In summary, the appellant has essentially argued that the common general knowledge suggested by the Examining Division as leading to the claimed subject-matter actually pointed to a different solution comprising a disk-based database with a hash-key look-up table.

In fact, nothing in the general knowledge of hashing at the priority data suggested providing a separate database stored in RAM storage means, as required by claim 1.

19. In the Board's opinion, certain aspects of the present invention, such as storing entries in RAM in the form of a hash table, can indeed be regarded as generally known or at least obvious to the skilled person. However, the essential teaching of the present invention which consists in determining, in the context of web crawling, whether a URL already exists in a database of processed web pages by checking a RAM-based hash table of fingerprints of the stored URLs, is neither known from, nor suggested by, documents D1, D3 or their combination.

20. As to document D6, the Board notes that it relates to database management systems for storing large databases (see column 1, first paragraph). According to D6 (column 2, line 62 to column 3, line 3) *"[a] database index file is maintained by a computer system having*

primary random access memory and secondary memory. A record for each item added to the database is stored in a sequential file in secondary memory (disk storage) and an indexed pointer to the new record is stored in a small B-tree stored in primary random access memory. The full index file for the database is a second, large B-tree stored in secondary memory. Leaf-nodes of the full index file are stored in packed, indexed order".

As pointed out in D6, column 3, lines 38 to 43, the primary memory is a high speed, random access memory.

20.1 Furthermore, according D6 (column 11, line 64 to column 12, line 7), *"the B-tree data structures used in the preferred embodiment could be replaced by other data structures, so long as the replacement data structures define a sorted order for referencing the records in the main data base. For instance, hash tables could be used in place of the B-trees of the preferred embodiment for storing indexed pointers. Entries in the hash tables would be stored in hash index order instead of indexed order. A small, memory resident, hash table would be merged periodically into a larger hash table stored in secondary memory, with the merge procedure proceeding in hash index order"* (underlining added).

21. The appellant has acknowledged that document D6 discussed the use of a small-B-tree which was stored in RAM and a large B-tree which was stored on disk, and that according to D6 (column 11, line 64 to column 12, line 7) *"hash tables could be used in place of the B-trees of the preferred embodiments for storing indexed pointers".*

21.1 However, as convincingly shown by the appellant, the teachings underlying the use of RAM in the present application and in document D6 are quite different.

As explained in column 7, lines 18 to 58, of D6 the purpose of creating an index file ("small B-tree") stored in a RAM is to temporarily store a sufficiently large number of indexed pointers to enable efficient storage of these indexed pointers in the disk memory ("secondary memory") using a rolling merge type of procedure. After the indexed pointers ("small B-tree") are completely merged into the "large B-tree", all entries of the small B-tree stored in RAM are deleted (cf. Figure 4B, steps 316 and 318).

21.2 On the other hand, the web information table stored in a RAM according to claim 1 maintains an entry for each entry in the web information file stored in the secondary memory, since it is used to quickly identify whether there is a corresponding entry in the secondary memory.

21.3 Hence, in the Board's opinion, a person skilled in the art, wishing to implement a system for exploring the World Wide Web and locating web pages as described in document D1, would not regard the teaching of document D6 as relevant. However, even under the assumption that the skilled person might take D6 into account, the Board agrees with the appellant that the combination of the teachings of documents D1 and D6 would not lead to the claimed subject-matter.

22. Although the subject-matter of claim 1 is to be regarded as inventive with respect to the available prior art, the Board finds that some further issues

have to be dealt with before a patent can be granted on the basis of the second auxiliary request.

22.1 In particular, claim 1 specifies that one of the entries of the web information table stored in RAM denotes "fetch status information". However, according to the description of the original application (page 5, lines 6 to 10), the hash table includes, inter alia, *"a fetched flag that is set true only if the corresponding Web page has been successfully fetched"*. On the other hand, "fetch status information" is entered in the "sequential disk file" (page 5, lines 10 to 12). Other passages of the description (see page 6, line 4 to 8) use the term "*fetch flag*".

22.2 For the sake of clarity, a comma should be inserted between "a corresponding web page" and "and a file location value" in the first paragraph of the characterising part of claim 1. Furthermore, the clause "*characterised in that*" does not appear to be compatible with the following sentence structure and, in particular, verb forms (*"a web information table ... having")*. The clause *"characterised by"* appears more appropriate.

It is also noted that *"the web information disk file"* referred to in the characterising part of claim 1 (first paragraph) corresponds evidently to *"a web information file stored in secondary memory"* introduced in the preamble of claim 1. For the sake of clarity, it is suggested to specify in the characterising part that the web information file is a web information disk file.

22.3 Also to improve the clarity of the claim wording the term "value" should be repeated after "a fetched flag"

in line 10 of the last paragraph of claim 1 as submitted with the grounds of appeal (cf. application as filed, page 10, last line to page 11, line 2).

22.4 Independent claim 6 presents the same clarity issues and should be amended accordingly. The consistency between the independent claims and their respective dependent claims should also be checked. In particular, it seems that dependent claims 5 and 10 may now be redundant.

Finally, as highlighted in section XII. of the decision, the Board has noted some typographical errors in claims 1 and 6 of the second auxiliary request. Thus, proofreading of the claims is suggested.

23. As the Board considers that the appellant's second auxiliary request can provide a basis for granting a patent, provided that some minor clarity issues are overcome, there is no need to consider the lower ranking requests.

24. In view of the above, the Board comes to the conclusion that the decision under appeal is to be set aside and that the case is to be remitted to the department of first instance for further prosecution in order to give the appellant the opportunity to handle the outstanding clarity issues relating to the second auxiliary request.

Order

For these reasons it is decided that:

1. The decision under appeal is set aside.

2. The case is remitted to the department of first instance for further prosecution.

The Registrar: The Chairman:

I. Aperribay R. Moufang

Decision electronically authenticated