SYSTEM AND METHOD FOR GENERATING TRAINING DATA FOR FUNCTION APPROXIMATION OF AN UNKNOWN PROCESS SUCH AS A SEARCH ENGINE RANKING ALGORITHM

Inventor: Parashuram Kulkarni, New York, NY (US)

Correspondence Address:
DILWORTH & BARRESE, LLP
1000 WOODBURY ROAD, SUITE 405
WOODBURY, NY 11797 (US)

Appl. No.: 12/367,656
Filed: Feb. 9, 2009

Related U.S. Application Data
Provisional application No. 61/093,586, filed on Sep. 2, 2008.

Publication Classification
Int. Cl. G06F 17/30 (2006.01)
U.S. Cl. 707/5, 707/E17.108

ABSTRACT
A system and method for generating training data for a machine learning system. A training data generator server sends at least one keyword to a search engine. The training data generator server receives at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second rank, the first and second rank being based on the keyword. The training data generator server assigns a first label to the first page based on the first rank; and assigns a second label to the second page based on the second rank. The first web page, second page, first label and second label are forwarded to a machine learning server.
<table>
<thead>
<tr>
<th>Label</th>
<th>Web page</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td><img src="100.png" alt="Image" /></td>
</tr>
<tr>
<td>99</td>
<td><img src="99.png" alt="Image" /></td>
</tr>
<tr>
<td>98</td>
<td><img src="98.png" alt="Image" /></td>
</tr>
<tr>
<td>97</td>
<td><img src="97.png" alt="Image" /></td>
</tr>
<tr>
<td>...</td>
<td><img src="ellipsis.png" alt="Image" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="3.png" alt="Image" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="2.png" alt="Image" /></td>
</tr>
<tr>
<td>1</td>
<td><img src="1.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Fig. 3
Search engine queries a search engine index using keywords to produce web pages or process uses keywords to produce an output.

Search engine or process forwards keywords or inputs and ranked web pages or outputs to training data server.

Assign label to pages or outputs based on rank.

Labels and pages or outputs are used as training data.

Fig. 4
SYSTEM AND METHOD FOR GENERATING TRAINING DATA FOR FUNCTION APPROXIMATION OF AN UNKNOWN PROCESS SUCH AS A SEARCH ENGINE RANKING ALGORITHM


BACKGROUND OF THE INVENTION

[0002] 1. Field of the Invention
[0003] This disclosure relates to machine learning algorithms and, more particularly, to generation of training data for machine learning algorithms.
[0004] 2. Description of the Related Art
[0005] Referring to FIG. 1, the World Wide Web ("WWW") is a distributed database including literally billions of pages accessible through the Internet. Searching and indexing these pages to produce useful results in response to user queries is constantly a challenge. A search engine is typically used to search the WWW.
[0006] A typical prior art search engine 20 is shown in FIG. 1. Pages from the Internet or other source 22 are accessed through the use of a crawler 24. Crawler 24 aggregates pages from source 22 to ensure that these pages are searchable. Many algorithms exist for crawlers and in most cases these crawlers follow links in known hypertext documents to obtain other documents. The pages retrieved by crawler 24 are stored in a database 36. Thereafter, these pages are indexed by an indexer 26. Indexer 26 builds a searchable index of the pages in a database 34. For example, each web page may be broken down into words and respective locations of each word on the page. The pages are then indexed by the words and their respective locations.

[0007] In use, a user 32 sends a search query to a dispatcher 30. Dispatcher 30 compiles a list of search nodes in cluster 28 to execute the query and forwards the query to those selected search nodes. The search nodes in search node cluster 28 search respective parts of the index 34 and return search results along with a document identifier to dispatcher 30. Dispatcher 30 merges the received results to produce a final result set displayed to user 32 sorted by ranking scores based on a ranking function.

[0008] The ranking further is a function of the query itself and the type of page produced. Factors that are used for relevance include hundreds of features extracted, collected or identified for each page including: a static relevance score for the page such as link cardinality and page quality, superior parts of the page such as titles, metadata and page headers, authority of the page such as external references and the “level” of the references, the GOOGLE page rank algorithm, and page statistics such as query term frequency in the page, words on a page, global term frequency, term distances within the page, etc.

[0009] The use of search engines has become one of the most popular online activities with billions of searches being performed by users every month. Search engines are also a starting point for consumers for shopping and various day to day purchases and activities. With billions of dollars being spent by consumers online, it has become ever more important for web sites to organize and optimize their web pages in an effort to be more visible and accessible to users of a search engine.

[0010] As discussed above, for each web page, hundreds of features are extracted and a ranking function is applied to those features to produce a ranking score. A merchant with a web page would like his page to be ranked higher in a result set based on relevant search keywords with web pages of his competitor for the same keywords. For example, for a merchant selling telephones, that merchant would like his web page to acquire a higher ranking score, and appear higher in a result set produced by a search engine based on the keyword query “telephone” than the ranking scores of web sites of his competitors for the same keyword. There are some prior art solutions available to guess the ranking algorithm used by a search engine and to provide recommendations about improvements that can be made to web pages so that the ranking score for a web page relating to particular keywords may improve. However, most of these systems use manual human judgment and historical knowledge about search engines. Humans must be trained to perform this analysis. The basis for these judgments are mostly guesses or arrives at by trial and error. Consequently, most prior art solutions are inaccurate, time consuming, and require expensive human capital. Moreover, these solutions are available only for specific search engines and are not immune to changes in search or ranking algorithms used by known search engines nor do they have the ability to adapt to new search engines.

SUMMARY OF THE INVENTION

[0011] One embodiment of the invention is a method for generating training data for a machine learning system. The method comprises sending at least one keyword to a search engine; and receiving at a first processor at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second rank, the first and second rank being based on the keyword. The method further comprises assigning at the first processor a first label to the first page based on the first rank; assigning at the first processor a second label to the second page based on the second rank; and forwarding the first web page, second page, first label and second label to a machine learning processor.

[0012] Another embodiment of the invention is a method for generating training data for a machine learning system. The method comprises sending at least one input to a system effective to perform a process; and receiving at a first processor at least a first and a second output from the system in response to the input, the first output having a first rank, the second output having a second rank, the first and second rank being based on the input. The method further comprises assigning at the first processor a first label to the first output based on the first rank; assigning at the first processor a second label to the second output based on the second rank; and forwarding the first result, second result, first label and second label to a machine learning processor.

[0013] Yet another embodiment of the invention is a system for generating training data for a machine learning system. The system comprises a first processor effective to send at least one keyword to a search engine. The first processor is further effective to: receive at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second
rank, the first and second rank being based on the keyword; assign a first label to the first page based on the first rank; and assign a second label to the second page based on the second rank. The system further comprises a machine learning processor connected to the first processor, the machine learning processor effective to receive the first web page, second web page, first label and second label.

[0014] Still another embodiment of the invention is a computer readable storage medium including computer executable code effective to generate training data for a machine learning system. The code includes the steps of sending at least one keyword to a search engine; and receiving at a first processor at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second rank, the first and second rank being based on the keyword. The code further includes the steps of assigning at the first processor a first label to the first page based on the first rank; assigning at the first processor a second label to the second page based on the second rank; and forwarding the first web page, second page, first label and second label to a machine learning processor.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] The drawings constitute a part of the specification and include exemplary embodiments of the present invention and illustrate various objects and features thereof.

[0016] FIG. 1 is a system drawing a search engine in accordance with the prior art.

[0017] FIG. 2 is a system drawing of a machine learning system in accordance with an embodiment of the invention.

[0018] FIG. 3 is a schematic drawing of a database structure in accordance with an embodiment of the invention.

[0019] FIG. 4 is a flow chart of a process which could be used in accordance with an embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT(S)

[0020] Various embodiments of the invention are described hereinafter with reference to the figures. Elements of like structures or function are represented with like reference numerals throughout the figures. The figures are only intended to facilitate the description of the invention or as a limitation on the scope of the invention. In addition, an aspect described in conjunction with a particular embodiment of the invention is not necessarily limited to that embodiment and can be practiced in conjunction with any other embodiments of the invention.

[0021] When applying a ranking function, search engines receive as input: 1) at least one keyword and 2) a plurality of web pages in a result set produced based on keyword(s). With those inputs, the search engine produces as an output a ranking score for each web page. The inventors recognized this phenomenon and produced a system and algorithm to reverse engineer the function performed by search engines to produce that output. Stated another way, search engines perform the following ranking function to generate a ranking score for each page in a result set:

\[ \text{ranking score} = F(\text{input}) \]

[0022] Where the input is the search query in the form of keyword(s) and the extracted features of the pages in the result set. The present system and method determines training data that may be used to determine the function F used by a search engine.

[0023] In order to approximate the ranking function, training data may be sent to a machine learning system. Generating such training data is perhaps the most difficult and labor intensive part of any machine learning system. As discussed above, prior art techniques for generating training data include the use of teams of humans subjectively viewing selected portions of available data such as keywords and result sets. Even if collection of data may be automated, in the prior art, labeling of the data is performed manually. Such labeling techniques are often inaccurate as they are subject to human judgment of a complex system such as a search engine. A human being typically cannot judge by intuition whether he has collected all kinds of different search results to ensure that the training data is diverse and it is generally not possible to manually track or generate a diverse set of data. A diverse training set is desired for a machine learning algorithm to work well. Moreover, human labeling is not accurate because it is generally not possible to judge a label value by intuition.

[0024] Referring to FIG. 2, there is shown a system 80 in accordance with an embodiment of the invention. System 80 includes a training data generator server 60. Training data generator server or processor 60 sends keywords 62 over a network 64 (such as the Internet) to a search engine server 66. Keywords 62 could be virtually any set of keywords that, when input to a search engine, yield web pages in a result set. It is desirable to generate a number of different sets of keywords. Many techniques could be used to generate such sets. For example, keyword tools provided by search engines such as the MSN Keyword tool, or the GOOGLE ADWORDS tool could be used, third party tools which monitor and collect keywords based on popularity usage and other metrics may be used, or statistical analysis may be used to determine important keywords from web pages and web logs. For example, by collecting the frequency distribution of keywords from web pages and web logs, it may be possible to identify important keywords from pages. Keywords 62 are sent by search engine server 66 to a search engine index 68.

[0025] Search engine index 68 outputs web pages 70 that are responsive to a search query including keywords 62. Search engine server 66 receives web pages 70 and orders or ranks web pages 70 based on an unknown ranking algorithm to produce ranked web pages 76.

[0026] Ranked web pages 76 are sent over network 64 and fed to training data generator server 60. Training data generator server 60 stores ranked web pages 76 and labels 82 for those pages in a training data storage 84. A label 82 is associated with each ranked web page 76 corresponding to the rank of the ranked web page 76 based on keyword 62. Label 82 allows system 80 to represent the relevance of each ranked web page 76 to keywords 62. Prior art labeling techniques required manually intensive, inaccurate and expensive human capital. Humans would view each ranked web page 76 and provide an appropriate label. The inventors have determined that a linear distribution of the ranking scores is a good representation of those scores. Consequently, if L ranked web pages 76 are considered, the highest ranked web page is given a label L, the second highest is given a label L-1, etc.

[0027] Referring to FIGS. 2 and 3, there is shown an example of a training data structure 110 which may be stored in training data storage 84. As shown, for a keyword 112 ("telephone" is shown) training data structure 110 may include a label column 114 and a web page column 118. Label column 114 includes labels 116 for ranked web pages 76.
(FIG. 2). The web pages themselves may be stored in web page column 118. The contents of training data structure 110 may be forwarded and used as training data in a machine learning server or processor 74. Machine learning server 74 may use any known machine learning techniques on training data 110 to produce an approximated ranking function 88.

[0028] Referring to FIG. 4, there is shown a flow chart of a process which could be used in accordance with an embodiment of the invention. The process could be used with, for example, system 80 described with respect to FIG. 2. As shown at step S2, at least one input or keyword is sent to a search engine or any other system implementing a process. At step S4, the search engine queries a search engine index using the keyword to produce a result set including web pages or the process uses the keywords as input or produce an output. At step S6, the search engine ranks the web pages or the process ranks the output. At step S8, the search engine or process forwards the inputs or keywords and ranked web pages or outputs to a training data server or processor. At step S10, the training data server assigns a label to each page or output based on the rank. At step S12, the labels and pages or outputs are used as training data.

[0109] Clearly, although different servers are shown for various elements those servers could be combined in a single processor housing or location.

[0030] A system in accordance with that described above can be used to collect training data for any search engine. Moreover, the system can adapt automatically to changes in ranking functions of existing search engines and produce new training data accordingly. Prior art systems are significantly limited in that subjective, expensive human capital is used to analyze only samples of available data. A system in accordance with the invention could analyze one page or thousands of pages easily and efficiently.

[0031] The invention has been described with reference to an embodiment that illustrates the principles of the invention and is not meant to limit the scope of the invention. Modifications and alterations may occur to others upon reading and understanding the preceding detailed description. It is intended that the scope of the invention be construed as including all modifications and alterations that may occur to others upon reading and understanding the preceding detailed description insofar as they come within the scope of the following claims or equivalents thereof. Various changes may be made without departing from the spirit and scope of the invention.

[0032] Although the above description is focused on the search engine context, the inventive concepts may be applied to any function approximation system where the inputs and outputs are known.

[0033] As can be discerned, the system and process described above is more accurate than human labeling because, in part, results of the unknown process, such as search engine ranking, are used. As the system is automated, it is possible to easily collected large amounts of training data without manual intervention. Ranking algorithms produced in accordance with the invention are change resistant. This is because training data is based on search results. If any search engine changes its ranking algorithm the results will change and the training data will change. Prior art systems based on intuition and prior knowledge of humans cannot adapt as easily. The system works with known and to be developed search engines and can easily be applied to specific sites such as TRAVELOCITY.COM.

What is claimed is:

1. A method for generating training data for a machine learning system, the method comprising:
   sending at least one keyword to a search engine;
   receiving at a first processor at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second rank, the first and second rank being based on the keyword;
   assigning at the first processor a first label to the first page based on the first rank;
   assigning at the first processor a second label to the second page based on the second rank; and
   forwarding the first web page, second page, first label and second label to a machine learning processor.

2. The method as recited in claim 1, wherein the first and second labels are based on a linear distribution of a ranking of the first and second pages by the search engine.

3. The method as recited in claim 1, wherein the pages are web pages.

4. The method as recited in claim 1, wherein the keyword is generated using at least one of an MSN keyword tool, GOOGLE ADWORDS, and a statistical analysis of keywords from web pages.

5. A method for generating training data for a machine learning system, the method comprising:
   sending at least one input to a system effective to perform a process;
   receiving at a first processor at least a first and a second output from the system in response to the input, the first output having a first rank, the second output having a second rank, the first and second rank being based on the input;
   assigning at the first processor a first label to the first output based on the first rank;
   assigning at the first processor a second label to the second output based on the second rank; and
   forwarding the first result, second result, first label and second label to a machine learning processor.

6. The method as recited in claim 5, wherein the first and second labels are based on a linear distribution of a ranking of the first and second pages by the search engine.

7. The method as recited in claim 5, wherein the pages are web pages.

8. The method as recited in claim 5, wherein the keyword is generated using at least one of an MSN keyword tool, GOOGLE ADWORDS, and a statistical analysis of keywords from web pages.

9. A system for generating training data for a machine learning system, the system comprising:
   a first processor effective to send at least one keyword to a search engine;
   the first processor further effective to:
      receive at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second rank, the first and second rank being based on the keyword;
      assign a first label to the first page based on the first rank; and
      assign a second label to the second page based on the second rank; and
a machine learning processor connected to the first processor, the machine learning processor effective to receive the first web page, second web page, first label and second label.

10. The system as recited in claim 9, wherein the first and second labels are based on a linear distribution of a ranking of the first and second pages by the search engine.

11. The system as recited in claim 9, wherein the pages are web pages.

12. The system as recited in claim 9, wherein the keyword is generated using at least one of an MSN keyword tool, GOOGLE ADWORDS, and a statistical analysis of keywords from web pages.

13. A computer readable storage medium including computer-executable code effective to generate training data for a machine learning system, the code including the steps of:
   sending at least one keyword to a search engine;
   receiving at a first processor at least a first and a second page from the search engine in response to the keyword, the first page having a first rank, the second page having a second rank, the first and second rank being based on the keyword;
   assigning at the first processor a first label to the first page based on the first rank;
   assigning at the first processor a second label to the second page based on the second rank; and
   forwarding the first web page, second page, first label and second label to a machine learning processor.

14. The storage medium as recited in claim 13, wherein the first and second labels are based on a linear distribution of a ranking of the first and second pages by the search engine.

15. The storage medium as recited in claim 13, wherein the pages are web pages.

16. The storage medium as recited in claim 13, wherein the keyword is generated using at least one of an MSN keyword tool, GOOGLE ADWORDS, and a statistical analysis of keywords from web pages.

* * * * *